Ann. Inst. Statist. Math.
Vol. 46, No. 4, 769–775 (1994)

ON THE JOINT DISTRIBUTION OF GRUBBS' STATISTICS

TEA-YUAN HWANG[1] AND CHIN-YUAN HU[2]

[1] *Institute of Statistics, National Tsing Hua University, Hsinchu 30043, Taiwan, R.O.C.*
[2] *Department of Business Education, National Changhua University of Education, Changhua 50058, Taiwan, R.O.C.*

(Received March 5, 1993; revised April 15, 1994)

Abstract. In this paper, the joint pdf's of Grubbs' statistics for normal and exponential populations are obtained; and relationship between the two pdf's is established. New formulations of the first marginal pdf of Grubbs' statistics for these two populations are given; the formulation of the exponential population case is a new one. Iterative formulas for the pdf of Grubbs' statistics are also obtained.

Key words and phrases: Exponential distribution, Grubbs' statistics, normal distribution.

1. Introduction and definitions

The problem of testing outlying observations is of considerable importance in applied statistics for quite some time. Grubbs (1950) proposed a test statistic for testing whether the first i smallest observations in a sample of size n are outliers. In this paper, we are concerned with the exact distribution of this statistic for the normal and exponential populations. A historical survey can be found in the works of Barnett and Lewis (1984) and David (1981) and the references therein.

Let $Y_1, \ldots, Y_n$ be a random sample of size n from an absolutely continuous distribution, and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be their order statistics. For $0 \leq i \leq n-1$, let $\bar{Y}_{(i)}$ and $S^2_{(i)}$ be the reduced mean and the reduced sum of squares of the $n-i$ largest order statistics, respectively; that is

$$\bar{Y}_{(i)} = \frac{1}{n-i} \sum_{k=i+1}^{n} Y_{(k)}, \qquad S^2_{(i)} = \sum_{k=i+1}^{n} (Y_{(k)} - \bar{Y}_{(i)})^2. \tag{1.1}$$

Note that for $i = n-1$, $S^2_{(n-1)} = 0$ by the definition. The total sample mean $\bar{Y}_{(0)}$ and total sum of squares $S^2_{(0)}$ are denoted by $\bar{Y}_n$ and S^2_n respectively.

To test whether $Y_{(1)}, \ldots, Y_{(i)}$ are outliers, Grubbs (1950) proposed the following statistic, denoted by Z_i:

$$Z_i = \frac{\text{reduced sum of squares}}{\text{total sum of squares}} = \frac{S^2_{(i)}}{S^2_n}, \qquad 1 \leq i \leq n-2. \tag{1.2}$$

He obtained the exact distribution of Z_1 for the normal population which was obtained for the first time by Grubbs. Some related works can be found in Dixon (1950), Grubbs (1969), McMillan (1971), Tietjen and Moore (1972), Grubbs and Beck (1972), Fieller (1976) and an extensive account in Barnett and Lewis (1984). In this paper, we will derive the exact joint pdf of Grubbs' statistics for $i > 1$ which is still unknown even for the normal population.

In Section 2, the joint pdf's of Grubbs' statistics for both normal and exponential populations are derived; and the relationship between their pdf's is also established. In Section 3, new formulations of the marginal pdf's of Grubbs' statistic Z_1 for both the normal and exponential populations are given; the new formulation of the normal population case (see, Theorem 3.1 below) is essentially different from that of the previously published results mentioned earlier. Furthermore, the iterative formulas for the pdf's of Grubbs' statistics for these two populations are also obtained.

Finally, the approach given here can be extended to deal with a larger class of distributions; for example, when the original random vector follows the spherically symmetric distributions (see, Muirhead (1982), p. 37) or uniformly distributed over a positive simplex in $\mathbb{R}^n$ (see, Aitchison (1982, 1985)).

2. The joint pdf of Grubbs' statistics

For convenience, the following notations will be used throughout the rest of this article. Define statistics T_i, $1 \le i \le n-1$, as follows

$$T_i = \left[\frac{n-i+1}{n-i}\right]^{1/2} \cdot \left[\frac{Y_{(i)} - \bar{Y}_n}{S_n} + \frac{1}{n-i+1}\sum_{k=1}^{i-1} \frac{Y_{(k)} - \bar{Y}_n}{S_n}\right], \tag{2.1}$$

where the summation in (2.1) is taken as zero for $i = 1$. Let

$$A = \left\{(t_1, \ldots, t_{n-1}) : t_1^2 + \cdots + t_{n-1}^2 = 1, \left(\frac{n-k+2}{n-k}\right)^{1/2} \cdot t_{k-1} \le t_k \le 0, 2 \le k \le n-1\right\} \tag{2.2}$$

be a subset of the unit sphere in $\mathbb{R}^{n-1}$, thus the region of the set A looks like a hyperspheric cap, and its boundary is a closed curve with $(n-1)$ vertices and $(n-1)$ arcs (see Hwang and Hu (1993)). Define

$$\sigma_{n-1}(A) = \text{the uniform distribution over the subset } A.$$

In order to obtain the main results of this paper we need the following lemmas.

LEMMA 2.1. (i) *For the normal population, the joint pdf of* $T_1, \ldots, T_{n-1}$ *given by* (2.1) *is* $d\sigma_{n-1}(A)$. (ii) *For the exponential population, their joint pdf is*

$$b_n^* \cdot (-t_1)^{-(n-1)} \cdot d\sigma_{n-1}(A) \tag{2.3}$$

where

$$b_n^* = \frac{2 \cdot n! \cdot \pi^{(n-1)/2}}{n^{(n+2)/2} \cdot (n-1)^{(n+1)/2} \cdot \Gamma\left(\frac{n-1}{2}\right)}.$$

The proof of Lemma 2.1 can be found in Hwang and Hu ((1994), Corollary 3.2). Note that our present notation S_n^2 is $(n-1)$ times the S_n^2 used in that paper.

The following Lemma 2.2 establishes the exact relationship between the statistics $T_1, \ldots, T_{n-1}$ and the Grubbs' statistics.

LEMMA 2.2. *Let $Z_1, \ldots, Z_{n-2}$ be Grubbs' statistics defined in* (1.2) *and T_i be as given in* (2.1). *For $1 \le i \le n-2$,*

$$Z_i = 1 - \sum_{k=1}^{i} T_k^2.$$

PROOF. From (1.1), we have

$$\begin{aligned} S_n^2 &= S_{(i)}^2 + (n-i)(Y_{(i)} - \bar{Y}_n)^2 + \sum_{k=1}^{i} (Y_{(k)} - \bar{Y}_n)^2 \\ &= S_{(i)}^2 + \frac{1}{n-i}\left[\sum_{k=1}^{i} (Y_{(k)} - \bar{Y}_n)\right]^2 + \sum_{k=1}^{i} (Y_{(k)} - \bar{Y}_n)^2. \end{aligned}$$

Or

$$1 - Z_i = \sum_{k=1}^{i} \left[\frac{Y_{(k)} - \bar{Y}_n}{S_n}\right]^2 + \frac{1}{n-i}\left[\sum_{k=1}^{i} \frac{Y_{(k)} - \bar{Y}_n}{S_n}\right]^2, \qquad 1 \le i \le n-2. \tag{2.4}$$

From Hwang and Hu ((1994), Lemma 2.4), it follows that the right side of (2.4) is nothing but $\sum_{k=1}^{i} T_k^2$. Thus the lemma is established. □

THEOREM 2.1. (Normal case) *Let $Y_1, \ldots, Y_n$ be a random sample from the normal population and Z_i, $1 \le i \le n-2$, be the Grubbs' statistics. Then, the joint pdf of $(Z_1, \ldots, Z_{n-2})$ is*

$$a_n \cdot [(1 - z_1)(z_1 - z_2) \cdots (z_{n-3} - z_{n-2}) \cdot z_{n-2}]^{-1/2} \cdot I_{R(z)}$$

where

$$a_n = \frac{n! \cdot \Gamma\left(\frac{n-1}{2}\right)}{2^{n-1} \cdot \pi^{(n-1)/2}}, \tag{2.5}$$

$$R(\boldsymbol{z}) = \left\{ \boldsymbol{z} : \begin{array}{l} 0 \le z_1 \le n(n-2)/(n-1)^2, \quad z_0 = 1 \\ \max\left\{0, \dfrac{2(n-k+1)}{n-k} \cdot z_{k-1} - \dfrac{n-k+2}{n-k} \cdot z_{k-2}\right\} \\ \qquad \le z_k \le \dfrac{(n-k)^2 - 1}{(n-k)^2} \cdot z_{k-1} \\ 2 \le k \le n-2 \end{array} \right\}, \tag{2.6}$$

and I stands for the indicator function.

PROOF. Note that the random variables $T_1, \ldots, T_{n-1}$ satisfy the nontrivial constraint $T_1^2 + \cdots + T_{n-1}^2 = 1$. Now use the first assertion of Lemma 2.1 and apply the transformation $(t_1, \ldots, t_{n-1}) \to (z_1, \ldots, z_{n-2})$, where $z_i = 1 - t_1^2 - \cdots - t_i^2$, $1 \le i \le n-2$. The absolute value of Jacobian is

$$|J| = 2^{-(n-2)} \cdot [(1-z_1)(z_1 - z_2) \cdots (z_{n-3} - z_{n-2})]^{-1/2}.$$

From Lemma 2.2, it then follows that the joint pdf of Grubbs' statistics $Z_1, \ldots, Z_{n-2}$ is proportional to the absolute value of Jacobian $|J|$. The set (2.6) in the indicator function I can be obtained from the set A as given in (2.2) by using the inverse relationship $t_i = -(z_{i-1} - z_i)^{1/2}$, $1 \le i \le n-2$, where we define $z_0 = 1$. Finally, a_n given in (2.5) is equal to the inverse of the surface area A times 2^{-n+2}, and the derivation of this surface area can be found in Hwang and Hu ((1994), Corollary 3.1). Hence we have established Theorem 2.1. □

As an illustration of Theorem 2.1, let $Y_1, \ldots, Y_4$ be iid random variables from the normal distribution. Then the joint pdf of Grubbs' statistics Z_1 and Z_2 for $n = 4$ is distributed over a triangular region R_4 with three vertices $(0,0)$, $(\frac{2}{3}, 0)$, $(\frac{8}{9}, \frac{2}{3})$, that is $R_4 = \{(z_1, z_2) : 0 < z_1 < 8/9, \max\{0, 3z_1 - 2\} < z_2 < 3z_1/4\}$, and the joint pdf is given by

$$f(z_1, z_2) = \frac{3}{2\pi} \cdot [(1-z_1)(z_1 - z_2) \cdot z_2]^{-1/2} \cdot I_{R_4}.$$

Thus, the marginal pdf's of Z_1 and Z_2 can be obtained from its joint pdf respectively as follows:

$$f_{Z_1}(z_1) = \begin{cases} (1-z_1)^{-1/2} & 0 < z_1 < \dfrac{2}{3} \\ (1-z_1)^{-1/2} \left[\dfrac{1}{4} - 3\sin^{-1}\left(5 - \dfrac{4}{z_1}\right)/(2\pi) \right] & \dfrac{2}{3} \le z_1 < \dfrac{8}{9} \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

and

$$f_{Z_2}(z_2) = \begin{cases} (3/2\pi) \cdot z_2^{-1/2} \cdot [\sin^{-1}(1/3) - \sin^{-1}(5z_2 - 3)/(3 - 3z_2)] & \\ & 0 < z_2 < \dfrac{2}{3} \\ 0 & \text{otherwise.} \end{cases} \tag{2.8}$$

THEOREM 2.2. (Exponential case) *Let $Y_1, \ldots, Y_n$ be a random sample from the exponential population. Then, the joint pdf of Grubbs' statistics $(Z_1, \ldots, Z_{n-2})$ is given by*

$$b_n \cdot (1-z_1)^{-n/2} \cdot [(z_1 - z_2) \cdots (z_{n-3} - z_{n-2}) \cdot z_{n-2}]^{-1/2} \cdot I_{R(z)}$$

where the set $R(\boldsymbol{z})$ is given as in (2.6) and the constant is

$$b_n = 2^{-(n-2)} \cdot (n!)^2 \cdot n^{-(n+2)/2} \cdot (n-1)^{-(n+1)/2}.$$

PROOF. This theorem follows from Lemma 2.1 and its proof is omit since it is similar to that of Theorem 2.1. □

The following Corollary 2.1 establishes an exact relationship between the joint pdf's of Grubbs' statistics for the normal and exponential populations.

COROLLARY 2.1. *Let $f^{(N)}$ and $f^{(E)}$ be the joint pdf's of Grubbs' statistics $Z_1, \ldots, Z_{n-2}$ respectively for the normal and exponential populations. Then, we have*

$$f^{(E)}(z_1, \ldots, z_{n-2}) = b_n^* \cdot (1 - z_1)^{-(n-1)/2} \cdot f^{(N)}(z_1, \ldots, z_{n-2}) \tag{2.9}$$

where the constant b_n^* is given by (2.3).

PROOF. This corollary follows from Theorems 2.1 and 2.2. □

3. The iterative formulas for the marginal pdf of Z_1

For the normal population, exact sampling distribution of Z_1 already provided by Grubbs (1950); some other alternative formulations of this distribution can be found in Barnett and Lewis (1984) and references therein. Because of its importance in the test for outlying observations, we will derive iterative formulas for this sampling distribution (see, Theorem 3.1 below).

For the exponential population, $Z_1 = 1 - W_E$, where W_E has been suggested by Shapiro and Wilk (1972) for testing various composite and simple hypotheses of exponentiality. Its exact null distribution is still unknown (see, Stephens (1978)), and we will derive its iterative formulas (Theorem 3.2) in this section.

First, we note that the sampling distribution of Z_1 for the exponential population case can be derived directly from the normal population case by means of the following relationship,

$$f_{Z_1}^{(E)}(z_1) = b_n^* \cdot (1 - z_1)^{-(n-1)/2} \cdot f_{Z_1}^{(N)}(z_1) \tag{3.1}$$

where $f_{Z_1}^{(N)}$ and $f_{Z_1}^{(E)}$ are the pdf of Z_1 for the normal and exponential populations respectively, and the constant b_n^* is given by (2.3). The relationship (3.1) follows from Corollary 2.1 by integrating out $z_2, \ldots, z_{n-2}$ in (2.6).

In order to obtain the main results of this section, we define a function H_n as follows: Let $H_3(z_1) = [z_1(1 - z_1)]^{-1/2}$ for $0 < z_1 < 3/4$ and zero otherwise. For $n \geq 4$,

$$H_n(z_1) = \int \cdots \int_{R^*(z_1)} [z_{n-2}(1 - z_1) \cdots (z_{n-3} - z_{n-2})]^{-1/2} dz_2 \cdots dz_{n-2} \tag{3.2}$$

for $0 < z_1 < n(n-2)/(n-1)^2$ and zero otherwise, where $R^*(z_1)$ is the set obtained by integrating out $z_2, \ldots, z_{n-2}$ over the set $R(\boldsymbol{z})$, given by (2.6), for each fixed z_1. The following Lemma 3.1 gives an iterative relationship for the function H_n.

LEMMA 3.1. *Let H_n be given by* (3.2). *Then, for $n \geq 3$*

$$H_{n+1}(z_1) = (1-z_1)^{-1/2} \cdot z_1^{(n-3)/2} \cdot \int_{h_n(z_1)}^{n(n-2)/(n-1)^2} H_n(x)dx$$

for $0 < z_1 < (n^2-1)/n^2$ and zero otherwise, where

$$h_n(z_1) = \max\left\{0, \frac{2n}{n-1} - \frac{n+1}{(n-1)z_1}\right\}. \tag{3.3}$$

PROOF. Use the definition of H_{n+1} and apply the transformation $(z_1, z_2, \ldots, z_{n-1}) \to (z_1, x_2, \ldots, x_{n-1})$, where $x_i = z_i/z_1$, $2 \leq i \leq n-1$, which has the Jacobian $|J| = z_1^{n-2}$. The result follows by an application of Fubini's theorem and a convenient change of variables. □

From Theorem 2.1 and the definition of H_n, the pdf of Z_1 for the normal population can be rewritten as

$$f_{Z_1}^{(N)}(z_1) = a_n \cdot H_n(z_1) \tag{3.4}$$

where the constant a_n is given by (2.5). For notational convenience, the suffix in $f_{Z_1}^{(N)}$ and $f_{Z_1}^{(E)}$ will be omitted in the following Theorems 3.1 and 3.2.

THEOREM 3.1. (Normal case) *Assume the population is normal and let f_n be the pdf of Z_1 with sample size n where $n \geq 3$. Then*

$$f_{n+1}(z_1) = \frac{(n+1)\cdot\Gamma\left(\frac{n}{2}\right)}{2\cdot\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \cdot \frac{z_1^{(n-3)/2}}{(1-z_1)^{1/2}} \cdot \int_{h_n(z_1)}^{n(n-2)/(n-1)^2} f_n(x)dx$$

for $0 < z_1 < (n^2-1)/n^2$ and zero otherwise, where $h_n(z_1)$ is given by (3.3). *The initial pdf for $n = 3$ is $f_3(z_1) = \frac{3}{2\pi}[z_1(1-z_1)]^{-1/2}$ for $0 < z_1 < \frac{3}{4}$ and zero otherwise.*

PROOF. This theorem follows from Lemma 3.1 and (3.4). □

Theorem 3.1 is true for $n \geq 3$, and the initial pdf for $n = 3$ can be obtained from Theorem 2.1 by setting $n = 3$ in (2.5) and (2.6). As an illustration, the pdf of Z_1 for the normal population with the sample size $n = 4$ can be derived from Theorem 3.1. It coincides with the expression for $f_{Z_1}(z_1)$ given in (2.7).

THEOREM 3.2. (Exponential case) *Let f_n be the pdf of Z_1 from a sample of size n from an exponential population. Then, for $n \geq 3$,*

$$f_{n+1}(z_1) = \frac{(n-1)^{(n+1)/2}}{2\cdot(n+1)^{(n-1)/2}} \cdot \frac{z_1^{(n-3)/2}}{(1-z_1)^{(n+1)/2}} \cdot \int_{h_n(z_1)}^{n(n-2)/(n-1)^2} (1-x)^{(n-1)/2} \cdot f_n(x)dx$$

for $0 < z_1 < (n^2-1)/n^2$ *and zero otherwise, where* $h_n(z_1)$ *is given by* (3.3). *The initial pdf for* $n = 3$ *is* $f_3(z_1) = \frac{1}{2\sqrt{3}} \cdot z_1^{-1/2} \cdot (1-z_1)^{-3/2}$ *for* $0 < z_1 < 3/4$ *and zero otherwise.*

PROOF. This theorem follows from (3.1) and Theorem 3.1. □

On using Theorem 3.2, we obtain the pdf of Z_1 for the exponential population, when $n = 4$, as: $f_{Z_1}(z) = \pi \cdot (1-z_1)^{-2}/(6\sqrt{3})$ for $0 < z_1 < 2/3$; $(1-z_1)^{-2} \cdot [\pi/6 - \sin^{-1}(5-4/z_1)]/(4\sqrt{3})$ for $2/3 \leq z_1 < 8/9$ and zero otherwise.

Finally, we note that Theorem 2.1, 2.2, 3.1 and 3.2 still hold, when the original random vector follows a spherical distribution and the uniform distribution over a positive simplex in $\mathbb{R}^n$, respectively. The proofs are essentially the same as given there. Some results obtained for these spherically symmetric situations by using these theorems can be found in Hu (1990, 1994).

Acknowledgements

The authors would like to thank the referees for careful reading of the original version, which resulted in substantial improvements.

REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion), *J. Roy. Statist. Soc. Ser. B*, **44**, 139–177.

Aitchison, J. (1985). A general class of distributions on the simplex, *J. Roy. Statist. Soc. Ser. B*, **47**, 136–146.

Barnett, V. and Lewis, T. (1984). *Outliers in Statistical Data*, 2nd ed., Wiley, New York.

David, H. A. (1981). *Order Statistics*, 2nd ed., Wiley, New York.

Dixon, W. J. (1950). Analysis of extreme values, *Ann. Math. Statist.*, **21**, 488–506.

Fieller, N. R. J. (1976). Some problems related to rejection of outlying observations, Ph.D. Thesis, Department of Statistics, University of Hull, U.K.

Grubbs, F. E. (1950). Sample criteria for testing outlying observations, *Ann. Math. Statist.*, **21**, 27–58.

Grubbs, F. E. (1969). Procedures for detecting outlying observations, *Technometrics*, **11**, 1–21.

Grubbs, F. E. and Beck, G. (1972). Extension of sample sizes and percentage points for significance tests of outlying observations, *Technometrics*, **14**, 847–854.

Hu, C. Y. (1990). On signal to noise ratio statistic SN_T, Ph.D. Thesis (under the supervision of Prof. Hwang), National Tsing Hua University, Taiwan.

Hu, C. Y. (1994). The most powerful location and scale invariant test under the assumption of symmetry, *Comm. Statist. A—Theory Methods*, **23**(1), 11–26.

Hwang, T. Y. and Hu, C. Y. (1993). The best lower bound of sample correlation coefficient with ordered restriction, *Statist. Probab. Lett.*, **19**, 195–198.

Hwang, T. Y. and Hu, C. Y. (1994). On the joint distribution of studentized order statistics, *Ann. Inst. Statist. Math.*, **46**, 165–177.

McMillan, R. G. (1971). Tests for one or two outliers in normal samples with unknown variance, *Technometrics*, **13**, 87–100.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*, 32–40, Wiley, New York.

Shapiro, S. S. and Wilk, M. B. (1972). An analysis of variance test for exponential distribution (complete samples), *Technometrics*, **14**, 355–370.

Stephens, M. A. (1978). On the W test for exponential with origin known, *Technometrics*, **20**, 33–35.

Tietjen, G. L. and Moore, R. H. (1972). Some Grubbs-type statistics for the detection of several outliers, *Technometrics*, **14**, 583–597.