Characterizations based on certain regression assumptions of adjacent order statistics

Wen-Jang Huang* and Nan-Cheng Su†

*Department of Applied Mathematics, National University of Kaohsiung, Kaohsiung 81148, Taiwan;
†Department of Statistics, National Cheng-Kung University, Tainan 70101, Taiwan

Abstract

Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the order statistics from independent and identically distributed random variables $\{X_i, 1 \leq i \leq n\}$ with a common absolutely continuous distribution function. We investigate characterizations of distributions by using equality and linearity of $E(X_{(1)}^2 - (\eta X_{(2)} + \theta)X_{(1)}|X_{(2)})$ and $E(X_{(n)}^2 - (\eta X_{(n-1)} + \theta)X_{(n)}|X_{(n-1)})$, respectively, where η and θ are constants. It turns out a large class of distributions can be characterized. In particular, many important distributions, such as normal, gamma, exponential, inverse gamma, student t, and uniform distributions can be characterized correspondingly. Similar characterizations by using analogous regressional properties within the class of sample processes can also be obtained.

Keywords: Characterization; exponential distribution; gamma distribution; inverse gamma distribution; normal distribution; order statistics; regression function; student t distribution; uniform distribution.
AMS 2000 Subject Classifications: Primary: 62E10; Secondary: 62G30

1 Introduction

Throughout this work, for a fixed $n \geq 1$, let $\{X_i, 1 \leq i \leq n\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with a common absolutely continuous distribution function F and probability density function (p.d.f.) f. Also assume that F has support (a, b), where $-\infty \leq a < b \leq \infty$. Let $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the order statistics based

*Corresponding author. E-mail: huangwj@nuk.edu.tw; Tel: 886-7-5919169; Fax: 886-7-5919360. Support for this research was provided in part by the National Council of the Republic of China, Grant No. NSC 98-2118-M-390-001-MY2.

†E-mail: sunanchen@gmail.com; Tel: 886-6-2757575 ext 53623; Fax: 886-6-2342469. Support for this research was provided in part by the National Council of the Republic of China, Grant No. NSC 98-2118-M-006-010

on $\{X_i,\ 1 \le i \le n\}$. The properties and characterizations related to order statistics have been widely studied and some excellent reviews can be found in books such as Arnold *et al.* (1992) and David and Nagaraja (2003).

Recently, characterizations of F, especially the family of student t distribution, based on some simple properties of certain regression functions associated with the order statistics have been investigated by many authors. Here for $v > 0$, the p.d.f. of the t_v distribution is

$$f_v(u) = \frac{\Gamma((v+1)/2)}{\sqrt{\pi v}\Gamma(v/2)} \frac{1}{(1+u^2/v)^{(v+1)/2}},\ \ -\infty < u < \infty. \tag{1}$$

Nevzorov *et al.* (2003) showed that, for $n = 3$, the regression relation

$$E((X_{(1)} + X_{(3)})/2 | X_{(2)} = x) = x,\ a < x < b, \tag{2}$$

characterizes the t_2 distribution. Other interesting results can be found in Balakrishnan and Akhundov (2003), Akhundov *et al.* (2004), Nevzorova *et al.* (2007), and Akhundov and Nevzorov (2010), etc. Among them, Akhundov and Nevzorov (2010) rewrote (2) as

$$E(X_{(2)} - X_{(1)} | X_{(2)} = x) = E(X_{(3)} - X_{(2)} | X_{(2)} = x),\ a < x < b, \tag{3}$$

then used the regression relation

$$E((X_{(2)} - X_{(1)})^2 | X_{(2)} = x) = E((X_{(3)} - X_{(2)})^2 | X_{(2)} = x),\ a < x < b, \tag{4}$$

to characterize the t_3 distribution.

To begin with note that the conditional p.d.f.'s of $X_{(1)}$ given $X_{(2)} = x$, and $X_{(n)}$ given $X_{(n-1)} = x$ are

$$f_{X_{(1)}|X_{(2)}=x}(u) = \frac{f(u)}{F(x)},\ a < u < x < b, \tag{5}$$

and

$$f_{X_{(n)}|X_{(n-1)}=x}(u) = \frac{f(u)}{1-F(x)},\ a < x < u < b, \tag{6}$$

respectively, which nevertheless are independent of n. The facts that these two conditional distributions are the same for all n make the extensions of Akhundov and Nevzorov (2010) naturally to the general case with sample size n.

Motivated by the above observations, in this work, first in Section 2 we characterize F by using a more general regression assumption

$$\begin{aligned} & E(X_{(1)}^2 - (\eta X_{(2)} + \theta) X_{(1)} | X_{(2)} = x) \\ = \ & E(X_{(n)}^2 - (\eta X_{(n-1)} + \theta) X_{(n)} | X_{(n-1)} = x),\ a < x < b, \end{aligned} \tag{7}$$

where $n \ge 2$, and η, θ are constants, or equivalently,

$$\begin{aligned} & E((\frac{\eta}{2} X_{(2)} - X_{(1)} + \frac{\theta}{2})^2 | X_{(2)} = x) \\ = \ & E((X_{(n)} - \frac{\eta}{2} X_{(n-1)} - \frac{\theta}{2})^2 | X_{(n-1)} = x),\ a < x < b. \end{aligned} \tag{8}$$

Note that when $n = 3$, $\eta = 2$, and $\theta = 0$, (8) reduces to (4). We show that for a fixed n, under different η and θ, not only t distribution, many common distributions such as normal, gamma, exponential, inverse gamma and uniform distributions, which are all important in both theoretical and applied work in statistics, can be characterized respectively.

Next in Section 3, by using either the left or the right side of (7) is a linear function of x, similar characterizations can be obtained.

Finally, a point process version of our results will be given in the end.

2 Main results

From now on, assume the p.d.f. f is differentiable on (a, b). Our main results are based on the following simple yet useful lemma, which gives the solutions of a special first-order linear differential equation.

Lemma 1. Consider the equation:

$$(p_1x^2 + p_2x + p_3)f'(x) = (q_1x + q_2)f(x),\ a < x < b, \tag{9}$$

where p_1, p_2, p_3, q_1 and q_2 are real constants, and $p_1^2 + p_2^2 + p_3^2 \neq 0$. Then the general solutions of (9) are given by

(i) $p_1 = 0$, $p_2 = 0$, $p_3 \neq 0$ and

$$f(x) = c_1 \exp\left\{\frac{q_1}{2p_3}x^2 + \frac{q_2}{p_3}x\right\}; \tag{10}$$

(ii) $p_1 = 0$, $p_2 \neq 0$ and

$$f(x) = c_2 \left|p_2x + p_3\right|^{\frac{p_2q_2 - p_3q_1}{p_2^2}} \exp\left\{\frac{q_1}{p_2}x\right\}; \tag{11}$$

(iii) $p_1 \neq 0$, $4p_1p_3 > p_2^2$ and

$$f(x) = c_3 \left|p_1x^2 + p_2x + p_3\right|^{\frac{q_1}{2p_1}} \exp\left\{\frac{2p_1q_2 - p_2q_1}{p_1\sqrt{4p_1p_3 - p_2^2}} \arctan\frac{2p_1x + p_2}{\sqrt{4p_1p_3 - p_2^2}}\right\}; \tag{12}$$

(iv) $p_1 \neq 0$, $4p_1p_3 < p_2^2$ and

$$f(x) = c_4 \left|p_1x^2 + p_2x + p_3\right|^{\frac{q_1}{2p_1}} \left|\frac{2p_1x + p_2 - \sqrt{p_2^2 - 4p_1p_3}}{2p_1x + p_2 + \sqrt{p_2^2 - 4p_1p_3}}\right|^{(2p_1q_2 - p_2q_1)/(2p_1\sqrt{p_2^2 - 4p_1p_3})}; \tag{13}$$

(v) $p_1 \neq 0$, $4p_1p_3 = p_2^2$ and

$$f(x) = c_5 \left|p_1x^2 + p_2x + p_3\right|^{\frac{q_1}{2p_1}} \exp\left\{\frac{p_2q_1 - 2p_1q_2}{p_1(2p_1x + p_2)}\right\}, \tag{14}$$

where $c_1, \cdots, c_5$ are positive constants such that $\int_a^b f(x)dx = 1$.

It is easy to see that (10) contains the p.d.f.'s of exponential and normal distributions, (11) contains the p.d.f.'s of gamma distribution, and (14) contains the p.d.f.'s of inverse gamma distribution. The above five cases are not exclusive. For example, when $2p_1q_2 = p_2q_1$, (12)-(14) reduce to the same form

$$f(x) = c\left|p_1x^2 + p_2x + p_3\right|^{\frac{q_1}{2p_1}},\ a < x < b, \tag{15}$$

where $c > 0$ is a constant, which includes the p.d.f.'s of the t_v distribution, $v > 0$. Also when $q_1 = q_2 = 0$, (10)-(14) reduce to $f(x) = 1/(b-a),\ a < x < b$, the p.d.f. of the uniform distribution if both a and b are finite. Note that when a, b or both a and b are finite, the solutions of the p.d.f. f also include some truncated distributions. For example, the p.d.f.'s of doubly truncated normal distribution belongs to (10).

We now present our main results. Let $E(X_1) = \mu_1 < \infty$, $E(X_1^2) = \mu_2 < \infty$. As X_1 is nondegenerate, $\mu_2 - \mu_1^2 = \mathrm{Var}(X_1) > 0$. Assume for some fixed $n \geq 2$, and constants η and θ,

$$\begin{aligned} & E(X_{(1)}^2 - (\eta X_{(2)} + \theta)X_{(1)}|X_{(2)} = x) \\ = \ & E(X_{(n)}^2 - (\eta X_{(n-1)} + \theta)X_{(n)}|X_{(n-1)} = x),\ a < x < b. \end{aligned} \tag{16}$$

In view of (5) and (6), (16) implies

$$\begin{aligned} & \frac{1}{F(x)}\left(\int_a^x u^2 f(u)du - (\eta x + \theta)\int_a^x uf(u)du\right) \\ = \ & \frac{1}{1-F(x)}\left(\int_x^b u^2 f(u)du - (\eta x + \theta)\int_x^b uf(u)du\right),\ a < x < b. \end{aligned} \tag{17}$$

As

$$\int_x^b uf(u)du = \mu_1 - \int_a^x uf(u)du, \tag{18}$$

and

$$\int_x^b u^2 f(u)du = \mu_2 - \int_a^x u^2 f(u)du, \tag{19}$$

it follows that

$$\int_a^x u^2 f(u)du - (\eta x + \theta)\int_a^x uf(u)du = F(x)(\mu_2 - \theta\mu_1 - \eta\mu_1 x),\ a < x < b. \tag{20}$$

Taking the second derivatives of both sides of (20) with respect to x, and after some manipulations, we obtain

$$\begin{aligned} & ((\eta-1)x^2 + (\theta - \eta\mu_1)x + (\mu_2 - \theta\mu_1))f'(x) \\ = \ & ((2-3\eta)x + 2\eta\mu_1 - \theta)f(x),\ a < x < b, \end{aligned} \tag{21}$$

a differential equation with the form of (9). Hence the solutions of (21) can be obtained by using Lemma 1.

As applications, in the following table, we list some widely used distributions which can be characterized by using the assumption (16).

Table 1

Characterization of distributions by using assumption (16) for certain (η, θ, a, b).

(η, θ, a, b)	Distribution	$f(x)$
$(1, \mu_1, -\infty, \infty)$	Normal $\mathcal{N}(\mu_1, \sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma^2}\right\}$, where $\sigma^2 = \mu_2 - \mu_1^2$.
$(1, \frac{\mu_2}{\mu_1}, 0, \infty)$	Gamma $\mathcal{G}amma(\alpha, \beta)$	$\frac{1}{\Gamma(\alpha)\beta^\alpha}x^{\alpha-1}\exp\left\{-\frac{x}{\beta}\right\}$, where $\alpha = \frac{\mu_1^2}{\mu_2-\mu_1^2}$ and $\beta = \frac{\mu_2-\mu_1^2}{\mu_1}$.
$(1, 2\mu_1, 0, \infty)$	Exponential $\mathcal{E}(\lambda)$	$\lambda e^{-\lambda x}$, where $\lambda = \frac{1}{\mu_1}$.
$(\eta, (2-\eta)\mu_1, -\infty, \infty)$, where $\eta > 1$.	Student t $t_v(\mu_1, \rho)$	$\frac{\Gamma((v+1)/2)}{\rho\sqrt{\pi v}\Gamma(v/2)}\left(\left(\frac{x-\mu_1}{\rho}\right)^2\frac{1}{v}+1\right)^{-(v+1)/2}$, where $v = \frac{2\eta-1}{\eta-1}$ and $\rho = \sqrt{\frac{\mu_2-\mu_1^2}{2\eta-1}}$.
$(\frac{2}{3}, \frac{4}{3}\mu_1, a, b)$	Uniform $\mathcal{U}(a, b)$	$\frac{1}{b-a}$
$(\frac{\mu_2}{\mu_1^2}, \frac{\mu_2}{\mu_1}, 0, \infty)$	Inverse Gamma $\mathcal{IG}amma(\alpha, \beta)$	$\frac{\beta^\alpha}{\Gamma(\alpha)}x^{-\alpha-1}\exp\left\{-\frac{\beta}{x}\right\}$, where $\alpha = \frac{2\mu_2-\mu_1^2}{\mu_2-\mu_1^2}$ and $\beta = \frac{\mu_1\mu_2}{\mu_2-\mu_1^2}$.

Note that in the case $(\eta, (2-\eta)\mu_1, -\infty, \infty)$, if $\eta = 2$ or $\mu_1 = 0$, then $\theta(= (2-\eta)\mu_1) = 0$, and condition (16) is equivalent to

$$E((\frac{\eta}{2}X_{(2)} - X_{(1)})^2 | X_{(2)} = x) = E((X_{(n)} - \frac{\eta}{2}X_{(n-1)})^2 | X_{(n-1)} = x), \ -\infty < x < \infty. \quad (22)$$

Hence the corresponding characterization covers the result reported by Akhundov and Nevzorov (2010). That is the $t_3(\mu_1, \rho)$ distribution can be determined by (22) with $\eta = 2$ and $n = 3$. On the other hand, it can be seen easily that (16) may not always has a proper solution of the p.d.f. f, the case $\eta = \theta = 0$ is an obvious example.

3 Characterizations by linearity of regression

It is worth noting that (20) can be rewritten as

$$E(X_{(1)}^2 - (\eta X_{(2)} + \theta)X_{(1)} | X_{(2)} = x) = -\eta\mu_1 x + (\mu_2 - \theta\mu_1), \ a < x < b.$$

Inspired by this, in this section first we characterize distributions by the weaker assumption

$$E(X_{(1)}^2 - (\eta X_{(2)} + \theta)X_{(1)}|X_{(2)} = x) = \gamma x + \delta,\ a < x < b, \tag{23}$$

where η, θ, γ and δ are constants. Now (23) implies

$$\int_a^x u^2 f(u)du - (\eta x + \theta)\int_a^x uf(u)du = (\gamma x + \delta)\, F(x),\ a < x < b. \tag{24}$$

Taking the second derivatives of both sides of (24) with respect to x, and after some manipulations, we have

$$((\eta - 1)x^2 + (\theta + \gamma)x + \delta)f'(x) = ((2 - 3\eta)x - (\theta + 2\gamma))f(x),\ a < x < b. \tag{25}$$

Again (25) has the form of (9). Hence the solutions of (25) can be obtained by Lemma 1. We present some parallel characterization results of the previous section in the following table.

Table 2

Characterization of distributions by using assumption (23) for certain $(\eta, \theta, \gamma, \delta, a, b)$.

$(\eta, \theta, \gamma, \delta, a, b)$	Distribution
$(1, \theta, -\theta, \delta, -\infty, \infty)$, where $\delta > 0$.	$\mathcal{N}(\theta, \delta)$
$(1, \theta, \gamma, 0, 0, \infty)$, where $\theta + \gamma > 0$ and $2\theta + 3\gamma > 0$.	$\mathcal{G}amma(\alpha, \beta)$, where $\alpha = \frac{2\theta+3\gamma}{\theta+\gamma}$ and $\beta = \theta + \gamma$.
$(1, \theta, -\theta/2, 0, 0, \infty)$, where $\theta > 0$.	$\mathcal{E}(2/\theta)$
$(\eta, \frac{\gamma(\eta-2)}{\eta}, \gamma, \delta, -\infty, \infty)$, where $\eta > 1$ and $\delta > \frac{\gamma^2(\eta-1)}{\eta^2}$.	$t_v(\mu, \rho)$, where $v = \frac{2\eta-1}{\eta-1}$, $\mu = -\frac{\gamma}{\eta}$ and $\rho = \frac{1}{\sqrt{(\delta - \gamma^2(\eta-1)/\eta^2)(2\eta-1)}}$.
$(\frac{2}{3}, \frac{2(a+b)}{3}, -\frac{a+b}{3}, -\frac{ab}{3}, a, b)$,	$\mathcal{U}(a, b)$
$(\eta, \theta, -\theta, 0, 0, \infty)$, where $\eta > 1$ and $\theta > 0$.	$\mathcal{IG}amma(\alpha, \beta)$, where $\alpha = \frac{2\eta-1}{\eta-1}$ and $\beta = \frac{\theta}{\eta-1}$.

Next, (16) also leads to

$$\int_x^b u^2 f(u)du - (\eta x + \theta)\int_x^b uf(u)du = (\mu_2 - \theta\mu_1 - \eta\mu_1 x)(1 - F(x)),\ a < x < b,$$

which in turn is equivalent to

$$E(X_{(n)}^2 - (\eta X_{(n-1)} + \theta)X_{(n)}|X_{(n-1)} = x) = -\eta\mu_1 x + (\mu_2 - \theta\mu_1)\,,\ a < x < b.$$

Therefore, the distribution of X_1 can also be characterized based on the two largest order statistics. More precisely, assume

$$E(X_{(n)}^2 - (\eta X_{(n-1)} + \theta)X_{(n)}|X_{(n-1)} = x) = \gamma x + \delta,\ a < x < b, \tag{26}$$

where η, θ, γ and δ are constants. By using (6), (26) yields

$$\int_x^b u^2 f(u)du - (\eta x + \theta)\int_x^b uf(u)du = (\gamma x + \delta)(1 - F(x)),\ a < x < b. \tag{27}$$

Differentiating both sides of (27) with respect to x twice to arrive at

$$((\eta - 1)x^2 + (\theta + \gamma)x + \delta)f'(x) = ((2 - 3\eta)x - (\theta + 2\gamma))f(x),\ a < x < b. \tag{28}$$

Again (28) has a form of (9), using Lemma 1 yields the solutions of (28) and characterization results follows.

For applications, a table which is exactly the same as Table 2 can be obtained. Hence it is omitted.

4 Conclusions

As there are many characterizations of distributions by the sample process generated by a sequence of i.i.d. random variables. It is natural to ask whether our results can be extended to that of point processes.

Let $\{M(t),\ a < t < b\}$ denote the sample process generated by $\{X_i,\ 1 \le i \le n\}$ with the common distribution function F. That is $M(t)$ is the number of $X_{(i)} \le t$, $1 \le i \le n$. For $a < t < b$ and $1 \le k \le n - 1$, along the lines of Lemma 1 of Huang and Su (1999), where the process is defined in $[0, \infty)$, it can be shown

$$f_{X_{(1)},\cdots,X_{(k)}|X_{(k+1)}=t} = f_{X_{(1)},\cdots,X_{(k)}|M(t)=k}\ , \tag{29}$$

and

$$f_{X_{(k+1)},\cdots,X_{(n)}|X_{(k)}=t} = f_{X_{(k+1)},\cdots,X_{(n)}|M(t)=k}\ . \tag{30}$$

By (29), (30), and our previous results, F can be characterized immediately by using one of the following conditions:

$$\begin{aligned} &E(X_{(1)}^2 - (\eta t + \theta)X_{(1)}|M(t) = 1) \\ &= \quad E(X_{(n)}^2 - (\eta t + \theta)X_{(n)}|M(t) = n - 1),\ a < t < b, \end{aligned} \tag{31}$$

$$E(X_{(1)}^2 - (\eta t + \theta)X_{(1)}|M(t) = 1) = \gamma t + \delta,\ a < t < b, \tag{32}$$

or

$$E(X_{(n)}^2 - (\eta t + \theta)X_{(n)}|M(t) = n-1) = \gamma t + \delta,\ a < t < b, \tag{33}$$

which corresponds to (16), (23) and (26), respectively. Consequently, each of the characterization results in Examples 1-15 has a process version. We omit the details here.

On the other hand, it is interesting to characterize the common distribution of $\{X_i,\ 1 \le i \le n\}$ by using the relation

$$E(g(X_{(1)}, \cdots, X_{(n)})|X_{(k)} = x) = h(x),\ a < x < b, \tag{34}$$

where $g : R^n \to R$ and $h : R \to R$ and fixed $1 \le k \le n$. Yet as mentioned by Balakrishnan and Akhundov (2003), there is no solution for this problem in general. In this work, we have offered some functions g and h, where there are solutions. Other possible g and h, which may lead (34) have solutions will be studied in the future.

Acknowledgements.

The authors are very grateful to the referee for the valuable comments and suggestions that have considerably improved this work.

References

[1] Akhundov, I. and Nevzorov, V.B. (2010). A simple characterization of Student's t_3 distribution. *Statist. Prob. Lett.* 80, 293-295.

[2] Akhundov, I.S., Balakrishnan, N. and Nevzorov, V.B. (2004). New characterizations by properties of midrange and related statistics. *Commun. Statist. Theory Meth.* 33, 3133-3143.

[3] Arnold, B.C., Balakrishnan, N. and Nagaraja, H.N. (1992). *A First Course in Order Statistics.* Wiley, New York.

[4] Balakrishnan, N. and Akhundov, I.S. (2003). A characterization by linearity of the regression function based on order statistics. *Statist. Prob. Lett.* 63, 435-440.

[5] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics.* Wiley, New York.

[6] Huang, W.J. and Su, J.C. (1999). On certain problems involving order statistics - a unified approach through order statistics property of point processes. *Sankhyā A* 61, 36-49.

[7] Nevzorov, V.B., Balakrishnan, N., and Ahsanullah, M. (2003). Simple characterizations of Student's t_2 distribution. *Statistician* 52, 395-400.

[8] Nevzorova, L., Nevzorov, V.B. and Akhundov, I. (2007). A simple characterization of Student's t_2 distribution. *Metron-International J. Statist.* LXV (1), 53-57.