Journal of Statistical Computation and Simulation
Vol. 74, No. 7, July 2004, pp. 487–494

Taylor & Francis
Taylor & Francis Group

THE DENSITY OF THE SKEW NORMAL SAMPLE MEAN AND ITS APPLICATIONS

JOHN T. CHEN,* ARJUN K. GUPTA and TRUC T. NGUYEN

Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, Ohio 43403, USA

(Received 15 October 2002; In final form 11 May 2003)

This paper focuses on the distribution of the skew normal sample mean. For a random sample drawn from a skew normal population, we derive the density function and the moment generating function of the sample mean. The density function derived can be used for statistical inference on the disease occurrence time of twins in epidemiology, in which the skew normal model plays a key role.

Keywords: Skew normal distribution; Independence; Skew normal sample mean; Confidence interval; Disease occurring time in twins; Leukemia

AMS 1999 Subject classification: Primary 62H10, Secondary 62E17

1 INTRODUCTION

Azzalini (1985) discussed the following class of random variables: A random variable Z is a skew normal random variable if, there exists a real $\lambda \in R$, so that Z has a density function

$$\phi(z; \lambda) = 2\phi(z)\Phi(\lambda z) \quad -\infty < z < \infty \tag{1}$$

where $\phi(z)$ and $\Phi(z)$ are pdf and cdf of N(0, 1) respectively. He denoted this distribution family by $Z \sim \mathrm{SN}(\lambda)$. The random variable Z retains many statistical properties of the standard normal distribution. For example, if $Z \sim \mathrm{SN}(\lambda)$, then $Z^2 \sim \chi_1^2$.

Compared with normal distribution, skew normal distribution is flexible in the way that it removes the assumption of symmetry for the density function. It describes the distribution of a random variable in the following way: when $\lambda > 0$ the density function is positively skewed; when $\lambda < 0$, the density function is negatively skewed. The skew factor λ essentially services as another dimension to shape the distribution of the underlying population. When $\lambda = 0$, Eq. (1) becomes $\phi(\cdot)$, which is the density of the standard normal random variable. The moment generating function of $\mathrm{SN}(\lambda)$ can be expressed as:

$$M(t) = 2\exp\left(\frac{t^2}{2}\right)\Phi\left(\frac{\lambda t}{(1+\lambda^2)^{1/2}}\right). \tag{2}$$

* Corresponding author. E-mail: jchen@bgnet.bgsu.edu

ISSN 0094-9655 print; ISSN 1563-5163 online
DOI: 10.1080/0094965031000147687

Now that the distribution is not symmetric, to describe the distribution, we need quantities such as the measure of skewness

$$\frac{\mu_3}{\mu_2^{3/2}} = \operatorname{sign}(\lambda)\left(2-\frac{\pi}{2}\right)\left(\frac{\lambda^2}{(\pi/2)+((\pi/2)-1)\lambda^2}\right)^{3/2} \tag{3}$$

and kurtosis

$$\frac{\mu_4}{\mu_2^2} = 2(\pi-3)\left(\frac{\lambda^2}{(\pi/2)+((\pi/2)-1)\lambda^2}\right)^{2}. \tag{4}$$

Once λ returns to zero, all the quantities return to the ones for normal distribution.

There is an expanding literature dealing with theory and applications of skew normal distribution. For example, Aigner *et al.* (1977), Johnson *et al.* (1988), Genton *et al.* (2001), Gupta and Chen (2001) as well as Gupta and Brown (2001), to list just a few. However, given a set of random sample from a skew normal distribution family, there is no study yet regarding the distribution and statistical properties of the sample mean statistic. Note that in the skew normal family, if the density of X reads

$$f(x) = 2\phi(x, \mu, \sigma)\Phi(\lambda x, \mu, \sigma)$$

where $\phi(x, \mu, \sigma)$ and $\Phi(y, \mu, \sigma)$ are the pdf and cdf of a normal population with mean μ and variance σ, respectively, then the expectation of X is

$$E(X) = \mu + \sqrt{\frac{2}{\pi}}\frac{\lambda}{\sqrt{1+\lambda^2}}\sigma \tag{5}$$

With the fact that $\bar{X}$ is an unbiased estimator of $E(X)$, one can see obviously that the sample mean does not converge to the mean of the corresponding skew normal density, μ. However, μ combined with the skew factor and the variance σ measures the long term average of skew normal random samples.

In what follows, we shall derive the skew normal sample mean distribution, which is followed by the moment generating function of $\bar{X}$, as well as the quadratic form of a skew normal sample. We then conclude the article with a discussion on the occurrence time of disease in twins in epidemiology studies.

2 THE DISTRIBUTION OF SKEW NORMAL SAMPLE MEAN

One of the conventional methods in investigating a statistic is to figure out its probability distribution, moment generating function and moments. We start with the density function of skew normal sample mean in this section.

THEOREM 1 *Let $X_1, \ldots, X_n$ be independently and identically distributed random variables following* SN(λ). *The density function of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ reads*

$$f(x) = 2^n\sqrt{\frac{n}{2\pi}}e^{-(nx^2/2)}\Phi_n\left(\lambda x\left(\frac{1}{1+\lambda^2}\mathbf{I}_n + \frac{1}{n(1+\lambda^2)}\mathbf{1}\mathbf{1}'\right)^{1/2}\mathbf{1}\right)$$

where $\mathbf{1} = (1, \ldots, 1)'$, $\mathbf{I}_n$ *is the identity matrix and* $\Phi_n(\cdot)$ *is the n-variate cdf of the standard normal random vector.*

Proof To obtain the probability density function of the sample mean, we start from the joint distribution of the vector $\mathbf{x} = (x_1, \ldots, x_n)'$. Notice that

$$\begin{aligned} f(\mathbf{x}) &= 2^n \prod_{i=1}^{n} \phi(x_i)\Phi(\lambda x_i) \\ &= 2^n \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-\mathbf{x}'\mathbf{x}/2} \prod_{i=1}^{n} \int_{-\infty}^{\lambda x_i} \phi(y)\,\mathrm{d}y \end{aligned} \tag{6}$$

Let $\mathbf{y} = \mathbf{A}\mathbf{x}$ where matrix $\mathbf{A}$ is constructed as follows.

$$\mathbf{A}' = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n) \quad \text{and} \quad \mathbf{A} = \begin{pmatrix} \mathbf{a}_1' \\ \mathbf{a}_2' \\ \vdots \\ \mathbf{a}_n' \end{pmatrix}$$

where

$$\mathbf{a}_1 = n^{-1/2}(1, \ldots, 1)'$$

and for $\mathbf{a}_j, j = 2, \ldots, n$, the first $j-1$ elements take value 1, and the jth element takes value $1-j$, *i.e.*

$$\mathbf{a}_j = [j(j-1)]^{-1/2}(1, 1, \ldots, 1, 1-j, 0, \ldots, 0)'.$$

Then, since $\mathbf{A}$ is an orthogonal matrix, $\mathbf{A}'\mathbf{A} = \mathbf{I}_n$ and $\mathbf{A}\mathbf{A}' = \mathbf{I}_n$, also

$$\sqrt{n}y_1 = \sum_{i=1}^{n} x_i \quad \text{with} \quad \mathbf{y} = (y_1, y_2, \ldots, y_n)'.$$

For $i = 1, \ldots, n$, denote $\mathbf{y} = (y_1, \mathbf{z}')'$ and $\mathbf{a}_i = (a_{i1}, \mathbf{b}_i')'$ where

$$\mathbf{z} = \begin{pmatrix} y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} \quad \text{and} \quad \mathbf{b}_i = \begin{pmatrix} a_{i2} \\ a_{i3} \\ \vdots \\ a_{in} \end{pmatrix}$$

we have

$$\begin{aligned} \mathbf{x}'\mathbf{x} &= y'\mathbf{A}'\mathbf{A}y \\ &= y_1^2 + \mathbf{z}'\mathbf{z} \end{aligned}$$

and

$$\mathbf{a}_i'\mathbf{y} = a_{i1}y_1 + \mathbf{b}'\mathbf{z}.$$

Let

$$\mathbf{B}' = (\mathbf{b}_1, \ldots, \mathbf{b}_n)_{(n-1)\times n}$$

and

$$\mathbf{c} = (a_{11}, \ldots, a_{n1})',$$

we have

$$\mathbf{A}'\mathbf{A} = \begin{pmatrix} \mathbf{c}' \\ \mathbf{B}' \end{pmatrix} \begin{pmatrix} \mathbf{c}' \\ \mathbf{B}' \end{pmatrix}'.$$

Since $\mathbf{A}$ is orthogonal, the special structure of $\mathbf{A}$ yields

$$\mathbf{B}'\mathbf{B} = \mathbf{I}_{n-1}. \tag{7}$$

Therefore the density function of $\mathbf{y}$ by (7) can be expressed as

$$\begin{aligned} f(\mathbf{y}) &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-\mathbf{y}'\mathbf{y}/2} \prod_{i=1}^{n} \int_{-\infty}^{\lambda a_i'y} \phi(t)\,\mathrm{d}t \\ &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-(y_1^2+\mathbf{z}'\mathbf{z})/2} \prod_{i=1}^{n} \int_{-\infty}^{\lambda(a_{i1}y_1+\mathbf{b}_i'\mathbf{z})} \phi(t)\,\mathrm{d}t \\ &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-y_1^2/2}\mathrm{e}^{-(\mathbf{z}'\mathbf{z})/2} \prod_{i=1}^{n} \int_{-\infty}^{\lambda(a_{i1}y_1+\mathbf{b}_i'\mathbf{z})} \phi(t)\,\mathrm{d}t. \end{aligned} \tag{8}$$

We shall integrate out vector $\mathbf{z}$ to get the density of $y_1 = \sqrt{n}\bar{X}$, to this end, by (8) we have

$$\begin{aligned} f(y_1) &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-y_1^2/2} \int_{\mathbf{z}} \mathrm{e}^{-(\mathbf{z}'\mathbf{z})/2} \prod_{i=1}^{n} \int_{-\infty}^{\lambda(a_{i1}y_1+\mathbf{b}_i'\mathbf{z})} \phi(t)\,\mathrm{d}t\,\mathrm{d}\mathbf{z} \\ &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-y_1^2/2} \int_{\mathbf{z}} \mathrm{e}^{-(\mathbf{z}'\mathbf{z})/2} \int_{-\infty}^{\lambda(a_{11}y_1+\mathbf{b}_1'\mathbf{z})} \cdots \int_{-\infty}^{\lambda(a_{n1}y_1+\mathbf{b}_n'\mathbf{z})} \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-t't/2}\,\mathrm{d}t\,\mathrm{d}\mathbf{z} \\ &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)\mathrm{e}^{-y_1^2/2} \int_{\mathbf{z}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-1} \mathrm{e}^{-(\mathbf{z}'\mathbf{z})/2} \int_{R^*} \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-(t-\lambda\mathbf{B}\mathbf{z})'(t-\lambda\mathbf{B}\mathbf{z})/2}\,\mathrm{d}t\,\mathrm{d}\mathbf{z} \end{aligned} \tag{9}$$

where

$$R^* = (-\infty, \lambda a_{11}y_1) \times \cdots \times (-\infty, \lambda a_{n1}y_1).$$

Therefore (9) can be expressed as

$$\begin{aligned} f(y_1) &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)\mathrm{e}^{-y_1^2/2} \int_{R^*} \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-t't/2} \int_{\mathbf{z}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-1} \mathrm{e}^{-(2t'\lambda\mathbf{B}\mathbf{z}+\lambda^2\mathbf{z}'\mathbf{B}'\mathbf{B}\mathbf{z}+\mathbf{z}'\mathbf{z})/2}\,\mathrm{d}\mathbf{z}\,\mathrm{d}t \\ &= 2^n\left(\frac{1}{\sqrt{2\pi}}\right)\mathrm{e}^{-y_1^2/2} \int_{R^*} \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-t'(\mathbf{I}-\lambda^2\mathbf{B}(\mathbf{I}_{n-1}+\lambda^2\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}')t/2} \\ &\quad\times \int_{\mathbf{z}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-1} \mathrm{e}^{-[(\mathbf{z}-(\mathbf{I}+\lambda^2\mathbf{B}'\mathbf{B})^{-1}\lambda\mathbf{B}'t)'(\mathbf{I}+\lambda^2\mathbf{B}'\mathbf{B})(\mathbf{z}-(\mathbf{I}+\lambda^2\mathbf{B}'\mathbf{B})^{-1}\lambda\mathbf{B}'t)]/2}\,\mathrm{d}\mathbf{z}\,\mathrm{d}t. \end{aligned} \tag{10}$$

By (7) $\mathbf{B}'\mathbf{B} = \mathbf{I}_{n-1}$, we have

$$\mathbf{I}_{n-1} + \lambda^2 \mathbf{B}'\mathbf{B} = (1+\lambda^2)\mathbf{I}_{n-1}. \tag{11}$$

Substituting (11) into (10) yields

$$\int_{\mathbf{z}} \left(\frac{1}{\sqrt{2\pi}}\right)^{n-1} \mathrm{e}^{-[(\mathbf{z}-(\mathbf{I}+\lambda^2\mathbf{B}'\mathbf{B})^{-1}\lambda\mathbf{B}'t)'(\mathbf{I}+\lambda^2\mathbf{B}'\mathbf{B})(\mathbf{z}-(\mathbf{I}+\lambda^2\mathbf{B}'\mathbf{B})^{-1}\lambda\mathbf{B}'t)]/2}\, \mathrm{d}\mathbf{z} = (1+\lambda^2)^{-(n-1)/2}, \tag{12}$$

therefore, putting (12) into (10) results in

$$\begin{aligned} f(y_1) &= 2^n \left(\frac{1}{\sqrt{2\pi}}\right) \mathrm{e}^{-y_1^2/2} \int_{R^*} \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-t'(\mathbf{I}-(\lambda^2/1+\lambda^2)\mathbf{BB}')t/2} (1+\lambda^2)^{-(n-1)/2}\, \mathrm{d}t \\ &= 2^n \left(\frac{1}{\sqrt{2\pi}}\right) \mathrm{e}^{-y_1^2/2} \int_{-\infty}^{\lambda a_{21} y_1} \cdots \int_{-\infty}^{\lambda a_{n1} y_1} \left(\frac{1}{\sqrt{2\pi}}\right)^n \mathrm{e}^{-t'(\mathbf{I}-(\lambda^2/1+\lambda^2)\mathbf{BB}')t/2} (1+\lambda^2)^{-(n-1)/2}\, \mathrm{d}t \\ &= 2^n \left(\frac{1}{\sqrt{2\pi}}\right) \mathrm{e}^{-y_1^2/2} \Phi_n(\mathbf{T}^{1/2}\mathbf{w}). \end{aligned} \tag{13}$$

In (13) we denote

$$\mathbf{T} = \mathbf{I}_n - \frac{\lambda^2}{1+\lambda^2}\mathbf{BB}' \tag{14}$$

with $\mathbf{BB}' = \mathbf{I} - \mathbf{cc}'$, thus

$$|\mathbf{T}| = (1+\lambda^2)^{-(n-1)}. \tag{15}$$

Since $a_{i1} = 1/\sqrt{n}$,

$$\mathbf{w} = \lambda y_1 \frac{1}{\sqrt{n}} \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}. \tag{16}$$

Therefore, by (13)–(16), the density of $\bar{X} = (1/\sqrt{n})y_1$ reads

$$f(x) = 2^n \frac{\sqrt{n}}{\sqrt{2\pi}} e^{-nx^2/2} \Phi_n \left(\lambda x \left(\frac{\lambda^2}{1+\lambda^2} \mathbf{I}_n + \frac{1}{n(1+\lambda^2)} \mathbf{11}' \right)^{1/2} \mathbf{1} \right)$$

where $\mathbf{1} = (1, \ldots, 1)'$. ■

3 THE MOMENT GENERATING FUNCTION AND THE QUADRATIC FORM

In this section, we shall discuss the moment generating function of the skew normal sample mean, as well as the quadratic form of a skew normal random sample. For the first one, with Theorem 1, the moment generating function of $\bar{X}$ can be obtained by using the general method of integration

$$M(t) = \int_u \mathrm{e}^{tu} 2^n \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-nu^2/2} \Phi_n \left(\lambda u \left(\frac{\lambda^2}{1+\lambda^2} \mathbf{I} + \frac{1}{n(1+\lambda^2)} \mathbf{11}' \right)^{1/2} \mathbf{1} \right) \mathrm{d}u.$$

We have

THEOREM 2 *Let $X_1, \ldots, X_n$ be an i.i.d random sample from* SN(λ). *The moment generating function of $\bar{X}$ is*

$$M(t) = 2^n e^{t^2/2n} \Phi^n \left(\frac{\lambda}{\sqrt{1+\lambda^2}} \frac{t}{n} \right). \tag{17}$$

Remark By the fact that $X_1, \ldots, X_n$ are independent, in conjunction with the moment generating function of each X_i given in (1), the moment generating function of the skew normal sample mean can also be found by the following method.

$$\begin{aligned} M_{\bar{X}}(t) &= E(e^{t\bar{X}}) = E(e^{t/n \sum_{i=1}^{n} X_i}) = \prod_{i=1}^{n} E(e^{t/n X_i}) = \prod_{i=1}^{n} 2e^{(t^2/2n^2)} \Phi \left(\frac{\lambda}{\sqrt{1+\lambda^2}} \frac{t}{n} \right) \\ &= 2^n e^{t^2/2n} \Phi^n \left(\frac{\lambda}{\sqrt{1+\lambda^2}} \frac{t}{n} \right). \end{aligned} \tag{18}$$

The result in (18) coincides with the moment generating function directly calculated from the density function in Theorem 1.

For the quadratic form of a skew normal random sample, the result is obvious according to Azzalini (1985). For convenience of reference in the sequel, we state it as a theorem.

THEOREM 3 *If $X_1, \ldots, X_n$ constitute a random sample from a skew normal population* SN(λ), *then $\sum_{j=1}^{n} X_j^2 \sim \chi_n^2$.*

Proof By Azzalini (1985), the distribution of X_i^2 follows a χ_1^2 distribution, therefore, $\sum_{j=1}^{n} X_j^2$ follows χ_n^2 since $X_1, \ldots, X_n$ are independent. ∎

4 THE APPLICATION IN EPIDEMIOLOGY

Gupta and Chen (2001) studied methods of goodness of fit for skew normal models. They used a set of data of the time when one of the twins first caught a cold. The data set can be found in Roberts (1966). For this set of data, as shown in Gupta and Chen (2001), the skew normal model provides a uniformly better fit than the normal distribution model.

This observation agrees with the work of Roberts (1966). Under the assumption that X and Y follow the same distribution N(μ, σ^2), Roberts (1966) showed that the density of the time of one of the twins got a cold is,

$$f(w) = \frac{2}{(2\pi)^{1/2}} \Phi \left(-\sqrt{\frac{1-\rho}{1+\rho}} w \right) \exp \left(\frac{-w^2}{2} \right) \tag{19}$$

where $W = (Z - \mu)/\sigma$ with $Z = \min(X, Y)$ as the minimum age of occurrence of a disease in twins, and X, Y are the ages of the occurrence of a disease in the twins respectively, and

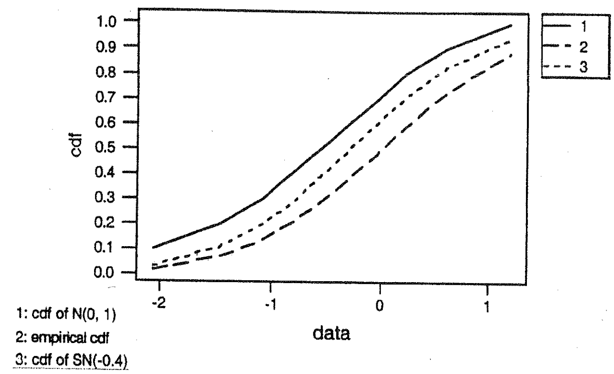

FIGURE 1 Comparison of fits.

ρ is the correlation coefficient between X and Y. Equation (18) is, in fact, a skew normal distribution with skew parameter

$$\lambda = -\sqrt{\frac{1-\rho}{1+\rho}}.$$

BY Theorem 3, when $\mu = \mu_0$, the confidence interval for the variance σ^2 is

$$\left(\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{a}, \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{b}\right)$$

where constants a and b satisfy

$$P(\chi_n^2 \geq a) = \frac{\alpha}{2} \quad \text{and} \quad P(\chi_n^2 \leq b) = \frac{\alpha}{2}.$$

For the unknown population mean, when λ and σ are available, μ can be estimated as follows. Since $(X_i - \mu)/\sigma \sim \mathrm{SN}(\lambda)$, we have the $(1-\alpha)100\%$ confidence interval for μ as $(\bar{X} - \sigma s_u, \bar{X} + \sigma s_l)$ with s_l and s_u satisfying

$$\int_{s_l}^{s_u} f(t)\,\mathrm{d}t = 1 - \alpha$$

where $f(x)$ is the density function of the skew normal sample mean in Theorem 1.

For other studies on twins in epidemiology, the underlying distribution of the time when a disease occurs in the twins, may also be modeled as a skew normal population. And the corresponding parameters can be estimated similarly. For example, if it is expected that both twins will get leukemia, as discussed in Roberts (1966), the time between the onset of the disease in the first until the onset of the disease in the second can be estimated using the skew normal model when the other parameters can be clinically assumed by experience.

Acknowledgement

The Research of Chen and Gupta was partially supported by Technology Innovation Enhancement Grant from OBOR.