Asian Journal of Economics and Banking
ISSN 2588-1396
http://ajeb.buh.edu.vn/Home

Extending A Priori Procedure to Two Independent Samples Under Skew Normal Settings

Cong Wang[1], Tonghui Wang[1] †, David Trafimow[2], Jing Chen[3]

[1]Department of mathematical sciences, New Mexico State University, USA

[2]Department of Psychology, New Mexico State University, USA

[3]Graduate College, Jose Rizal University, Philippines

Article Info

Received: 9/12/2018
Accepted: 15/01/2019
Available online: In Press

Keywords
Skew normal, sampling precision, confidence level, independent samples

JEL, MSC2010 classification
.....

Abstract

It is useful for researchers to be able to estimate the sample size necessary to have an impressive probability of obtaining a difference between sample locations of two independent groups that is close to the difference between corresponding population locations. The present manuscript provides the necessary equations. To increase generality, the derivations presented are not based on the typical assumption that the samples come from the family of normally distributed populations but rather from the much larger family of skew normal distributions. In addition, counter to many researchers' intuitions, we demonstrate that greater sampling precision ensues from skewness than from normality, all else being equal, with simulation results. Finally a real data example on faculty salaries of New Mexico State University is given for the illustration of our main results.

†Corresponding author: Tonghui Wang, Department of mathematical sciences, New Mexico State University, USA. Tel.: (575)646-2507. Email address: twang@nmsu.edu

1. INTRODUCTION

The starting point for the present work is a proposal that it is useful for researchers to consider, prior to performing experiments, how close they wish their sample statistics to be to corresponding population parameters, and at what probability. Trafimow [7] showed how to estimate the number of participants needed to meet specifications for closeness and probability in the context of a single group, where the population parameter of interest is the group mean. Trafimow and MacDonald [8] expanded this to include multiple means; but both contributions assumed normal distributions. In contrast, Trafimow et al. [9] and Wang et al. [11] showed that it is possible to perform similar calculations under the larger family of skew normal distributions, and for locations rather than means. Nevertheless, there remains an important limitation. Researchers often wish to compare differences in locations between independent samples, where levels of population skewness might be the same or different, and where the sample sizes might be the same or different. What sample sizes are necessary to attain specifications for closeness and probability in such cases involving differences between locations of control versus experimental conditions? The derivations to be proposed address this question.

For data that do not follow a normal distribution, it is natural to consider the skew normal distribution introduced by Azzalini [3]. A random variable Z is said to be a standard skew normal random variable with shape parameter λ if its probability density function is given by

$$f_Z(z) = 2\phi(z)\Phi(\lambda z), \qquad (1.1)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution, respectively. Since then, this kind of distribution and its multivariate form have been studied by many researchers including Azzalini [2], Azzalini and Dalla Valle [1], Gupta and Chang [4], Gupta et al. [5], Vernic [10], Wang et al. [12], Ye and Wang [14], and Ye et al. [15].

Now suppose that we have a population and want to construct the confidence interval for the location parameter. We start from the question about how many participants we need so that we can be confident the sample and the population locations are close. For the normal case, Trafimow [7] provided the answer for the one sample case by fixing the probability of the difference of sample mean and population mean within some precision f standard deviation at confidence level c. In this paper, we consider the difference of the location parameters from two independent skew normal populations. The goal is to determine the sample size needed to meet specifications for closeness and confidence, for using the sample difference in locations to estimate the population difference in locations.

The paper is organized as follows. Some properties of the family of multivariate skew normal distributions are presented in Section 2. In Section 3, we consider how to determine the least required sample size. The simulation work is provided in Section 4 for the validity of the derived equations and an

example with real data application is given for illustration of our main results in Section 5.

2. BRIEF REVIEW OF THE FAMILY OF MULTIVARIATE SKEW NORMAL DISTRIBUTIONS

In this paper, $\mathcal{M}_{n\times p}$ will denote the set of all $n \times p$ matrices over the real field $\mathbb{R}$, $\mathbb{R}^n$ will denote $\mathcal{M}_{n\times 1}$. For any $T \in \mathcal{M}_{n\times n}$, T' is the transpose of T. For any positive definite matrix $T \in \mathcal{M}_{n\times n}$ and $c > 0$, T^c and T^{-c} will be the c-th root of T and T^{-1}, respectively.

Definition 2.1. *The random vector* $\mathbf{X} = (X_1, \cdots, X_n)' \in \mathbb{R}^n$ *is said to have a multivariate skew normal distribution with location parameter* $\boldsymbol{\mu}$*, scale parameter* Σ*, and skewness (or shape) parameter* $\boldsymbol{\alpha}$*, denoted by* $\mathbf{X} \sim SN_n(\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha})$*, if its density function is given by*

$$f_{\mathbf{X}}(\mathbf{x}) = 2\phi_n(\mathbf{x}; \boldsymbol{\mu}, \Sigma)\Phi\left(\boldsymbol{\alpha}'\Sigma^{-1/2}(\mathbf{x} - \boldsymbol{\mu})\right), \tag{2.1}$$

where $\phi_n(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ *is the n-dimensional normal probability density function (pdf) with mean vector* $\boldsymbol{\mu}$ *and covariance matrix* Σ *and* $\Phi(z)$ *be the cumulative distribution function (cdf) of the standard normal random variable* Z*.*

The proof of the following lemma is given in Wang et al.[13].

Lemma 2.1. *Suppose that* $\mathbf{X} = (X_1, X_2, ..., X_n)' \sim SN_n(\boldsymbol{\mu}, \Sigma, \boldsymbol{\alpha})$ *with* $\boldsymbol{\mu} = \xi\mathbf{1}_n$, $\Sigma = \omega^2 I_n$ *and* $\boldsymbol{\alpha} = \lambda\mathbf{1}_n$*, where* $\mathbf{1}_n = (1, 1, \ldots, 1)'$*. Let* $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *and* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$ *be the sample mean and sample variance, respectively. Then*

(a) $\bar{X} \sim SN(\xi, \frac{\omega^2}{n}, \sqrt{n}\lambda)$*.*

(b) Each $X_i \sim SN(\xi, \omega^2, \lambda^*)$ *where* $\lambda^* = \lambda/\sqrt{1 + (n-1)\lambda^2}$*,* $i = 1, ..., n$*.*

(c) $\bar{X}$ *and* S^2 *are independent.*

(d) Let $T = \frac{\sqrt{n}(\bar{X}-\xi)}{S}$*. Then* T *has the skew t distribution with skewness parameter* $\sqrt{n}\lambda$ *and* $n - 1$ *degrees of freedom.*

Remark 2.1. For the sake of simplicity, statisticians often assume the independent sampling. But from Lemma 2.1, this assumption is not necessary as long as the data are from the same population so that we can assume that they are identically distributed.

3. THE SAMPLE SIZE NEEDED FOR ESTIMATING THE DIFFERENCE OF POPULATION LOCATION PARAMETERS

Consider two independent samples of unknown sample sizes n and m such that

$$\mathbf{X} = (X_1, \cdots, X_n)' \sim SN_n(\boldsymbol{\mu}_1, \Sigma_1, \boldsymbol{\alpha_1}) \tag{3.1}$$

and

$$\mathbf{Y} = (Y_1, \cdots, Y_m)' \sim SN_m(\boldsymbol{\mu}_2, \Sigma_2, \boldsymbol{\alpha_2}), \tag{3.2}$$

where $\boldsymbol{\mu}_1 = \xi_1\mathbf{1}_n$, $\boldsymbol{\mu}_2 = \xi_2\mathbf{1}_m$, $\boldsymbol{\alpha}_1 = \lambda_1\mathbf{1}_n$, $\boldsymbol{\alpha}_2 = \lambda_2\mathbf{1}_m$, $\Sigma_1 = \omega_1^2 I_n$ and $\Sigma_2 = \omega_2^2 I_m$. Without loss of generality, we can assume that $n \leq m$ and their ratio $k = m/n$ is assumed to be known.

Remark 3.1. In this paper, we will focus on obtaining the minimum sample size n required for estimating $\xi_d = \xi_1 - \xi_2$ with known λ_1 and λ_2. For the estimation of the shape parameters λ under skew normal settings, see Wang et

al. [11], Zhu et al. [16] and Ma et al.[6].

Now consider the linear transformation of X_i's and Y_j's given below.

$$U_i = X_i - \sqrt{\frac{n}{m}}Y_i + \frac{1}{\sqrt{nm}}\sum_{j=1}^{n} Y_j - \frac{1}{m}\sum_{k=1}^{m} Y_k \tag{3.3}$$

where $i = 1, \cdots, n$. This is given by Scheffé (1943) in the univariate normal case. Then the following result holds.

Theorem 3.1. *Consider two independent samples given in (3.1) and (3.2), and let U_i be defined in (3.3) and $\bar{U} = \frac{1}{n}\sum_{i=1}^{n} U_i$. Then*

(i) The density of $\bar{U} = \bar{X} - \bar{Y}$ is given by

$$f_{\bar{U}}(u) = 4\phi(u; \xi_d,\ \omega^2)E_Z[g(Z,\ u)] \quad \text{with } g(z,u) = \Phi(h_1 z + k_1)\Phi(h_2 z + k_2), \tag{3.4}$$

where $\omega^2 = \frac{\omega_1^2}{n} + \frac{\omega_2^2}{m}$, $Z \sim N(0,\ 1)$, $h_1 = \sqrt{\frac{n}{m}}\frac{\omega_2\lambda_1}{\omega}$, $h_2 = \sqrt{\frac{m}{n}}\frac{\omega_1\lambda_2}{\omega}$, $k_1 = \frac{(u-\xi_d)\omega_1\lambda_1}{\omega^2}$, and $k_2 = -\frac{(u-\xi_d)\omega_2\lambda_2}{\omega^2}$.

(ii) The mean and variance of $\bar{U}$ is given by

$$E(\bar{U}) = \xi_d + \delta_1\omega_1 - \delta_2\omega_2,$$

$$Var(\bar{U}) = \omega^2 - (\delta_1^2\omega_1^2 + \delta_2^2\omega_2^2).$$

where

$$\delta_1 = \sqrt{\frac{2}{\pi}}\frac{\lambda_1}{\sqrt{1+n\lambda_1^2}},$$

$$\delta_2 = \sqrt{\frac{2}{\pi}}\frac{\lambda_2}{\sqrt{1+m\lambda_2^2}}.$$

Proof. Rewrite $\mathbf{U} = (U_1, \cdots, U_n)'$ as $\mathbf{U} = \mathbf{X} - (A_1 - A_2 + A_3)\mathbf{Y}$ where $A_1, A_2, A_3 \in \mathcal{M}_{n\times m}$ with $A_1 = \sqrt{\frac{n}{m}}(I_n,\ \mathbf{0})$, $A_2 = \frac{1}{\sqrt{mn}}(\mathbf{1}_n\mathbf{1}_n',\ \mathbf{0})$, and $A_3 = \frac{1}{m}\mathbf{1}_n\mathbf{1}_m'$. Then it is easy to see $\bar{U} = \bar{X} - \bar{Y}$. By Lemma 2.1 and 2.2, we know that $\bar{X} \sim SN(\xi_1,\ \frac{\omega_1^2}{n},\ \sqrt{n}\lambda_1)$ and $\bar{Y} \sim SN(\xi_2,\ \frac{\omega_2^2}{m},\ \sqrt{m}\lambda_2)$. Note that $\bar{X}$ and $\bar{Y}$ are independent. Then, by Lamma 2.3, the density of $\bar{U}$ is

$$f_{\bar{U}}(u) = \int_{-\infty}^{\infty} f_{\bar{X}}(u+v)f_{\bar{Y}}(v)dv = 4\phi(u;\xi_d,\omega^2)\int_{-\infty}^{\infty}\phi(v;a,b^2)p(v)dv,$$

where $\omega^2 = \frac{\omega_1^2}{n} + \frac{\omega_2^2}{m}$, $a = \frac{\omega_2^2(\xi_d-u)}{m\omega^2}$, $b^2 = \frac{\omega_1^2\omega_2^2}{nm\omega^2}$, and

$$p(v) = \Phi\left(n\lambda_1\frac{v-(\xi_d-u)}{\omega_1}\right)\Phi\left(m\lambda_2\frac{v}{\omega_2}\right).$$

Let $Z = \frac{V-a}{b}$. Then the pdf of $\bar{U}$ is reduced to

$$f_{\bar{U}}(u) = 4\phi(u;\xi_d,\ \omega^2)E_Z[g(Z,\ u)],$$

$$\text{with} \quad g(z) = \Phi(h_1 z + k_1)\Phi(h_2 z + k_2),$$

where $h_1 = \sqrt{\frac{n}{m}}\frac{\omega_2\lambda_1}{\omega}$, $h_2 = \sqrt{\frac{m}{n}}\frac{\omega_1\lambda_2}{\omega}$, $k_1 = \frac{(u-\xi_d)\omega_1\lambda_1}{\omega^2}$ and $k_2 = -\frac{(u-\xi_d)\omega_2\lambda_2}{\omega^2}$. Thus, the density of $\bar{U}$ is obtained.

Corollary 3.1. *In Theorem 3.1 (i),*
(a) if $\lambda_2 = 0$, then $\bar{U} \sim SN(\xi_d,\ \omega^2,\ \lambda_{1*})$ with $\lambda_{1*} = \frac{\omega_1\lambda_1}{\sqrt{(1+n\lambda_1^2)\omega^2 - \omega_1^2\lambda_1^2}}$;
(b) if $\lambda_1 = 0$, then $\bar{U} \sim SN(\xi_d,\ \omega^2,\ \lambda_{2*})$ with $\lambda_{2*} = -\frac{\omega_2\lambda_2}{\sqrt{(1+m\lambda_2^2)\omega^2 - \omega_2^2\lambda_2^2}}$;
and
(c) if $\lambda_1 = \lambda_2 = 0$, then $\bar{U} \sim N(\xi_d,\ \omega^2)$.

Remark 3.2 By the definition of the close skew normal given in Gupta et al. [5] and Zhu et al. [17], both U_i and $\bar{U}$ are closed skew normally distributed.

Specifically, we can show that

$$\bar{U} \sim CSN_{1,2}(\xi_d,\ \omega^2,\ D,\ \mathbf{0},\ \Delta),$$

where $D = \left(\frac{\omega_1\lambda_1}{\omega^2},\ -\frac{\omega_2\lambda_2}{\omega^2}\right)'$, and

$$\Delta = \begin{bmatrix} 1+(n-\frac{\omega_1^2}{\omega^2})\lambda_1^2 & \frac{\lambda_1\lambda_2\omega_1\omega_2}{\omega^2} \\ \frac{\lambda_1\lambda_2\omega_1\omega_2}{\omega^2} & 1+(m-\frac{\omega_2^2}{\omega^2})\lambda_2^2 \end{bmatrix}.$$

The pdf of $\bar{U}$ given in (3.4) can be written as

$$f_{\bar{U}}(u) = 4\phi(u;\ \xi_d,\ \omega^2)\Phi_2[D(u-\xi_d;\ \mathbf{0},\ \Delta)],$$

where and ϕ and Φ_2 are the pdf and two-dimensional cdf of standard normal distribution.

Let $r = \frac{m}{n}$, $\xi_d = 0$, and $\omega_1 = \omega_2 = 1$. The density curves of $\bar{U}$ with different k are given in Figure 1 and Figure 2.

Remark 3.3. The density curves of $\bar{U}$ given in (3.4) are plotted in Figure 1 for $n = 97$ and $n = 40$ respectively. From Figure 1, we know that the variation of the ratio $k = m/n = 1, 1.2$, and 1.5 does not importantly change the shapes of densities either under normal or skew normal settings. Therefore, we may assume that the two sample sizes are equal to (say) n.

3.1. The Sample Size Needed for Estimating ξ_d with Known ω_1 and ω_2

In order to determine the minimum sample size n needed to be $c \times 100\%$ confident for the given sampling precision, we consider the distribution of $\bar{U}$ given in Theorem 3.1, for known ω_1 and ω_2 and $m = n$ by Remark 3.3.

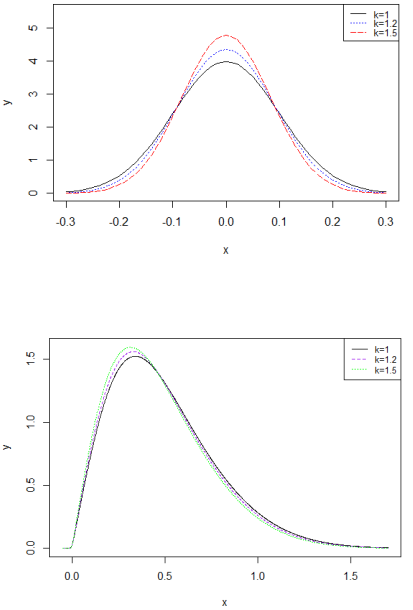

Figure 1: The density functions of $\bar{U}$ with $k = 1, 1.2, 1.5$. for $n = 97$, $\lambda_1 = \lambda_2 = 0$ (left), and $n = 40$, $\lambda_1 = -\lambda_2 = 5$ (right), respectively.

Theorem 3.2. *Let c be the confidence level and f be the precision which are specified such that the error associated with estimator $\bar{U}$ is $E = f\omega_0$ where $\omega_0^2 = \omega_1^2 + \omega_2^2$. More specifically, if*

$$P[f_1\omega_0 \le \bar{U} - E(\bar{U}) \le f_2\omega_0] = c, \quad (3.5)$$

where f_1 and f_2 are restricted by $max(|f_1|,\ f_2) \le f$, and $E(\bar{U})$ is the mean of $\bar{U}$. Then the minimum sample size n required can be obtained by

$$\int_L^U 4\phi(v)E_Z[h(Z,\ v)]dv = c \quad (3.6)$$

such that the length of the confidence interval is the shortest, where $L = \sqrt{n}f_1 + \frac{\gamma}{\omega}$ and $U = \sqrt{n}f_2 + \frac{\gamma}{\omega}$ with $\gamma = \omega_1\delta_1 -$

$\omega_2\delta_2$ *and* $Z \sim N(0,\ 1)$. *Here*

$$h(z, v) = \Phi(s_1 v + t_1 z)\Phi(s_2 v + t_2 z)$$

with $s_1 = \frac{\omega_2\lambda_1}{\omega}$, $s_2 = \frac{\omega_1\lambda_2}{\omega}$, $t_1 = \frac{\omega_1\lambda_1}{\omega}$ *and* $t_2 = -\frac{\omega_2\lambda_2}{\omega}$.

Proof. From Theorem 3.1, $E(\bar{U}) = E(\bar{X}-\bar{Y}) = \xi_d+\gamma$ with $\gamma = \omega_1\delta_1-\omega_2\delta_2$. Let $V = \frac{\bar{U}-\xi_d}{\omega}$. Then the pdf of V is

$$f_{_V}(v) = 4\phi(v)E_Y[q(Y,\ v)],$$

where $Y \sim N(\frac{-\omega_2^2 v}{n\omega}, b^2)$, and

$$q(y,\ v) = \Phi\left(n\lambda_1\frac{y+\omega v}{\omega_1}\right)\Phi(n\lambda_2 y/\omega_2).$$

By standardizing the distribution of Y,

$$f_{_V}(v) = 4\phi(v)E_Z[h(Z,\ v)],$$

where

$$h(z,\ v) = \Phi(s_1 v + t_1 z)\Phi(s_2 v + t_2 z)$$

with $s_1 = \frac{\omega_2\lambda_1}{\omega}$, $s_2 = \frac{\omega_1\lambda_2}{\omega}$, $t_1 = \frac{\omega_1\lambda_1}{\omega}$ and $t_2 = -\frac{\omega_2\lambda_2}{\omega}$. Then Equation (3.5) is equivalent to

$$P\left(\sqrt{n}f_1 + \frac{\gamma}{\omega} \le V \le \sqrt{n}f_2 + \frac{\gamma}{\omega}\right) = c.$$

So, (3.6) is obtained with $L = \sqrt{n}f_1 + \frac{\gamma}{\omega}$ and $U = \sqrt{n}f_2 + \frac{\gamma}{\omega}$. Then the required n can be solved through the integral equation (3.6).

From Theorem 3.2, we have the following remark.

Remark 3.4 The specified value of n, f_1 and f_2 are obtained simultaneously, given that f and the $c\times100\%$ confidence interval have been specified. Also, if the conditions in Theorem 3.2 are satisfied, we can construct the $c \times 100\%$ confidence interval for ξ_d, given by

$$[\bar{U} - \omega U,\ \bar{U} - \omega L],$$

and the length of the confidence interval is decreased by the increase of k, and the confidence interval of ξ_d for $k > 1$ is a subset of that of $k = 1$ under the assumptions in Theorem 3.2.

Corollary 3.2. *In Theorem 3.2,*
(a) if $\lambda_2 = 0$, *then the least* n *can be obtained by*

$$\int_{L_0}^{U_0} 2\phi(z)\Phi(\lambda_{1*}z) = c,$$

where the L_0 *and* U_0 *are as in the Theorem 3.2 under* $\delta_2 = 0$ *and* λ_{1*} *from Corollary 3.1(a). Then the* $c \times 100\%$*confidence interval for* ξ_d *is*

$$[\bar{U} - U_0\omega,\ \bar{U} - L_0\omega],$$

and
(b) if $\lambda_1 = \lambda_2 = 0$, *then the least* n *can be obtained and* $n = (\frac{z}{f})^2$, *where* z *is the* z*-score corresponding to the confidence level* c. *Also, the* $c \times 100\%$ *confidence interval for* ξ_d *is*

$$\left[\bar{U} - f\sqrt{\sigma_1^2+\sigma_2^2},\ \bar{U} + f\sqrt{\sigma_1^2+\sigma_2^2}\right].$$

3.2. The Sample Size Needed for Estimating ξ_d with Unknown ω_1 and ω_2

In this part, we will assume that ω_1 and ω_2 are unknown but equal, denoted as ω. Then we have the following result when the ratio of m and n is 1.

Theorem 3.3. *Let*

$$T = \frac{\sqrt{n}(\bar{U} - \xi_d)}{S_p}, \text{ with } \quad S_p^2 = \frac{S_1^2 + S_2^2}{2}.$$

Then the pdf of T is give by

$$f_T(t) = 4T(t;\ 2n-2)E_X\{E_Z[(G(Z,\ X,\ T)|x,\ t)]\}, \tag{3.7}$$

where $X|_{T=t} \sim \chi^2(2n-1)$, $Z|_{X=x,\ T=t} \sim N(0,1)$, the $T(t;\ 2n-2)$ being the pdf of t-distribution with $2n-2$ degrees of the freedom and

$$G(z,\ x,\ t) = \Phi(\frac{\lambda_1}{2}z + h_{1*})\Phi(\frac{\lambda_2}{2}z + h_{2*}),$$

where $h_{i*} = \frac{\lambda_i t\sqrt{nx}}{\sqrt{2(2n-2+t^2)}}$ *for* $i = 1,\ 2$.

Proof. Let $Z = \frac{\sqrt{n}(\bar{U}-\xi_d)}{\omega}$. Then by Theorem 3.2, the pdf of Z

$$f_Z(z) = 4\phi(z)E(h(Y,\ Z)|z).$$

Since $(n-1)S_i^2/\omega^2 \sim \chi^2(n-1)$ for $i = 1$ and 2, where S_1^2 and S_2^2 are independent,

$$V = (2n-2)S_p^2/\omega^2 \sim \chi^2(2n-2).$$

Note that $\bar{U}$ and S_p^2 are independent. Then the joint distribution of (T, V)

$$f_{T,\ V}(t,\ v) = f_{Z,\ V}\left(t\sqrt{\frac{v}{2n-2}}, v\right)\sqrt{\frac{v}{2n-2}}.$$

Thus, the pdf of T is

$$f_T(t) = c(t)\int_0^\infty f(x)E[(G(Y)|x,\ t)]dx$$

where $f(x)$ is the density function of χ^2 with $(2n-1)$ degrees of freedom, and

$$c(t) = \frac{4}{\sqrt{\pi(2n-2)}}\frac{\Gamma((2n-1)/2)}{\Gamma(2n-2)/2)}\times \left(1 + \frac{t^2}{2n-2}\right)^{-(2n-1)/2}.$$

Note that $c(t)$ is the density of t distribution of degree of freedom $2n-2$. So,

$$f_T(t) = 4T(t;\ 2n-2)E_X\{E[(G(Y)|x,\ t)]\}.$$

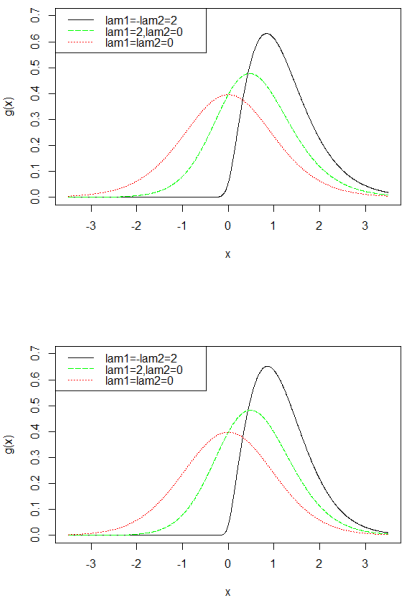

Figure 2: The density functions of T with different λ_1 (lam1) and λ_2 (lam2) for $n = 20$ (left), and $n = 40$ (right), respectively.

Remark 3.5 If we let $\lambda_1 \neq 0$ and $\lambda_2 = 0$ in Theorem 3.3, then the distribution of T is reduced to skew t distribution with $2n-2$ degrees of freedom and skewness $\sqrt{n}\lambda_1/\sqrt{2+n\lambda_1^2}$. More specially, if $\lambda_1 = \lambda_2 = 0$ in Theorem 3.3, then the distribution of T is reduced to

the t distribution with $2n-2$ degrees of freedom. The Density curves of T given in (3.7) are plotted in Figure 2 for $n=20$ and $n=40$, respectively. From the Figure 2, we know that variations of skewness parameters λ_1 and λ_2 do affect the shapes of densities.

Theorem 3.4. *Suppose the conditions in Theorem 3.2 hold. Then the least n can be obtained by solving the integration equation*

$$\int_{L_*}^{U_*} f(t)dt = c,$$

where $f_T(t)$ is given in Theorem 3.3, with $L_* = \frac{\sqrt{n}f_1+q}{S}$ and $U_* = \frac{\sqrt{n}f_2+q}{S}$, in which $q = \sqrt{\frac{n}{2}}(\delta_1-\delta_2)$ and $S = \sqrt{1-\delta_1^2}\frac{S_p}{S_1}$ for $S_1 = \sqrt{S_1^2}$ and $S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$. And then, the $c\times 100\%$ confidence interval for ξ_d will be

$$\left[\bar{U}-\sqrt{\frac{2}{n}}S_pU_*,\quad \bar{U}-\sqrt{\frac{2}{n}}S_pL_*\right]. \tag{3.8}$$

The proof is similar as Theorem 3.2.

Remark 3.6 The average length of the $c\times 100\%$ confidence interval for ξ_d given (3.8) is

$$EL = \frac{\sqrt{2}\omega(f_2-f_1)}{\sqrt{1-\delta_1^2}}.$$

Note that the confidence interval for ξ_d when $k=m/n>1$ is a subset of the interval given in (3.8) under the assumptions in Theorem 3.2 and hence its average length is shorter than EL.

Corollary 3.3. *From Theorem 3.4, we can obtain the following results.*

(a) If $\lambda_2=0$, then the least n can be obtained by

$$\int_{L_1}^{U_1} f_T(t)dt = c,$$

where $L_1 = L_*$ and $U_1 = U_*$ in (3.8) under $\delta_2=0$ so that the $c\times 100\%$ confidence interval for ξ_d is

$$\left[\bar{U}-\sqrt{\frac{2}{n}}S_pU_1,\ \bar{U}-\sqrt{\frac{2}{n}}S_pL_1\right]. \tag{3.9}$$

(b) If $\lambda_1=\lambda_2=0$, then the $c\times 100\%$ confidence interval for ξ_d is

$$[\bar{U}-f\sqrt{S_1^2+S_2^2},\quad \bar{U}+f\sqrt{S_1^2+S_2^2}]. \tag{3.10}$$

4. SIMULATION

We perform simulation results to support our derivations in Section 3. We assume that $\omega_1=\omega_2$ and the confidence $c=0.95,\ 0.9$. We will obtain the minimum n needed for precision $f=0.2$, which is listed in Table 1 and Table 2. From Table 1, we know that the required n is decreasing as λ_1 increases. Similar result is obtained from Table 2 when $\lambda_1=\lambda_2$.

Using the Monte Carlo simulations, we account relative frequency for the difference of location parameters $\xi_d=1,\ 2$, scale parameters $\omega_1=\omega_2=\omega_*=1,\ 2$, and different skewness parameter λ_1 with $\lambda_2=0$. The summary of relative frequencies is given in Table 3. From Table 3, we use "r.f." to denote the relative frequency for 90% confidence intervals given precision $f=0.2$. Also we use "p.e." to be the point estimate average of ξ_d. All results are illustrated simulation runs $M=10000$.

λ_1	n	f_1	f_2
0	72	-0.2	0.2
0.1	60	-0.1994	0.1991
0.2	53	-0.2	0.1996
0.3	51	-0.1998	0.1984
0.4	49	-0.1996	0.1984
0.5	48	-0.1999	0.1987
1	47	-0.1994	0.1982

Table 1: The minimum value of sample size n for different λ_1 with precision $f = 0.2$, $\lambda_2 = 0$ and confidence level $c = 0.9$.

λ	n	f_1	f_2
0	97	-0.2	0.2
0.1	72	-0.1995	0.1995
0.2	56	-0.1986	0.1986
0.3	49	-0.1982	0.1982
0.4	45	-0.1995	0.1995
0.5	43	-0.1988	0.1988
1	40	-0.1992	0.1992

Table 2: The minimum sample size n for different λ with $f = 0.2$, and $c = 0.95$ where $\lambda = \lambda_1 = \lambda_2$.

λ_1	n	$\xi_d = 1,\ \omega_* = 1$		$\xi_d = 1,\ \omega_* = 2$		$\xi_d = 2,\ \omega_* = 1$		$\xi_d = 2,\ \omega_* = 2$	
		r.f.	p.e.	r.f.	p.e.	r.f.	p.e.	r.f.	p.e.
0.1	60	0.8944	1.0010	0.8983	1.0628	0.8977	1.9992	0.8982	2.0637
0.2	53	0.8893	1.0010	0.8910	1.0788	0.8851	1.9978	0.8892	2.0825
0.3	51	0.8992	1.0011	0.8857	1.0906	0.8944	2.0001	0.8890	2.0707
0.4	49	0.8824	0.9982	0.8966	1.0881	0.8830	1.9971	0.8980	2.0876
0.5	48	0.8922	0.9976	0.8903	1.0896	0.8984	1.9979	0.8927	2.0934
1	47	0.8995	0.9990	0.9021	1.0114	0.9001	1.9997	0.8913	2.0963

Table 3: The relative frequency (r.f.) and the corresponding average point estimate of different value of ξ_d (p.e.) and λ for $f = 0.2$, $c = 0.9$ and $\omega_* = 1,\ 2$.

Density curves and their corresponding histograms of 95% confidence interval for ξ are given in Figure 3. The curve on the left is for for $\lambda_1 = \lambda_2 = 0$, with $\xi_d = 0$, $\omega_* = 1$, and $f = 0.2$(normal case), and the curve on the right is for $\lambda_1 = 5$, $\lambda_2 = -5$ with $\xi_d = 0$, $\omega_* = 1$, and $f = 0.2$. Also the 95% confidence intervals are listed in Figure 3.

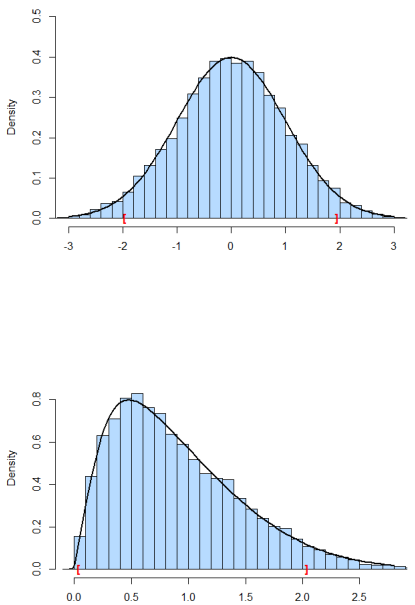

Figure 3: Density functions and histogram of 95% confidence interval for

$\xi_d = 0$ with skewness parameters $\lambda_1 = \lambda_2 = 0$, $f = 0.2$ (left), and $\lambda_1 = 5$, $\lambda_2 = -5$, $f = 0.2$ (right), respectively.

5. AN EXAMPLE WITH REAL DATA

We provide an example for illustration of our main results obtained. The data sets contain the salaries from Departments of Sciences (DS) and the remaining departments in the College of Arts and Sciences (RD), New Mexico State University, which are obtained from the Budget Estimate (2018/19)[18]. By the method of moment estimation, the estimated distribution based on the data sets are $SN(3.4131,\ 3.5768^2,\ 3.8487)$ for DS and $SN(4.0995,\ 3.8794^2,\ 2.1394)$ for RD, respectively (with unit \$10000). The histogram and its corresponding curve of $\bar{U}$ given by (3.3) are shown in Figure 4.

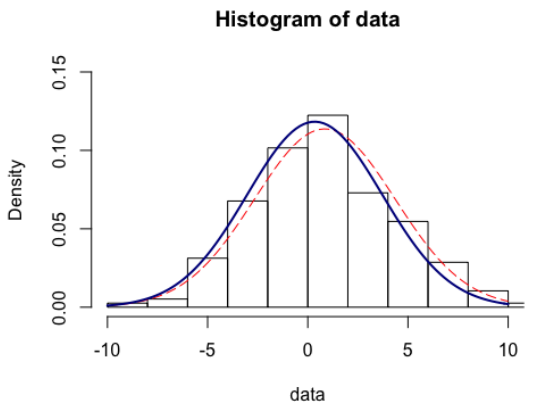

Figure 4: The histogram and its curves of the difference given by (3.3) for the real data sets both under normal(the dish line) and skew normal (the solid line) cases.

Now we suppose that the skewness parameters and the scale parameters are known to be $\lambda_1 = 3.8487$, $\lambda_2 = 2.1394$, $\omega_1 = 3.5768$, $\omega_2 = 3.8794$ and the ratio of the sample sizes $k = 1.02$ are known. If we consider the precision $f = 0.2$ and confidence level $c = 0.95$, then the minimum sample size needed is 43. Randomly choose the samples of the same size 43, from both populations, we obtain $\bar{U} = -0.6530$. Then by the Remark 3.4, the 95% confidence intervals for ξ_d are $[-0.9602, -0.3481]$ under skew normal assumptions, and $[-0.9808, -0.3252]$ under normal population assumptions given in Corollary 3.2 (b). Similarly if scale parameters are assumed to be unknown, by Theorem 3.4 and Corollary 3.3 (b), the 95% confidence intervals for ξ_d are $[-1.3115, 0.0008]$ under skew normal assumptions, and $[-1.3559, 0.0499]$ under normal assumptions, respectively. Note that in both cases, the lengths of confidence intervals under skew normal settings are shorter than those under normal assumptions.

References

[1] Azzalini, A. and Dalla Valle, A. (1996). *The multivariate skew-normal distribution*, Biometrica 83(4), 715-726.

[2] Azzalini, A. and Capitanio, A. (1999). *Statistical application of the multivariate skew normal distribution*, J. Roy. Statist. Soc. B 83, 579-602.

[3] Azzalini, A. (2013). *The skew-normal and the related families*, Combridge University Press, volume 3.

[4] Gupta, A. K. and Chang, F. C. (2003). *Multivariate skew symmetric distributions*, Appl. Math. Lett. 16, 643-646.

[5] Gupta, A. K., Gouzalez, G. and Dominguez-Molina, J. A. (2004). *A multivariate skew normal distribution*, J. Multivariate Anal. 82, 181-190.

[6] Ma, Z., Chen, Y., Wang, T., and Peng, W.(2019). *The Inference on the location parameters under multivariate skew normal settings*, Beyond Traditional Probabilistic Methods in Economics. ECONVN 2019. Studies in Computational Intelligence, vol 809. Springer, Cham.

[7] Trafimow, D. (2016). *Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics*, Educational and Psychological Measurement.

[8] Trafimow, D. and MacDonald, J. A. (2017). *Performing inferential statistics prior to data collection*, Educational and Psychological Measurement, 77(2), 204-219.

[9] Trafimow, D., Wang, T, and Wang, C. (2018). *From a sampling precision perspective, skewness is a friend and not an enemy!* Educational and Psychological Measurement (DOI: 10.1177/0013164418764801), pp. 1-22.

[10] Vernic, R. (2006). *Multivariate skew-normal distributions with applications in insurance*, Insurance: Mathematics and Economics 38. 413-426.

[11] Wang, C., Wang, T., Trafimow, D. and Myuz, H. (2019). *D*esired sample size for estimating the skewness parameter under skew normal settings in "Structural Changes and their Economic Modeling" (V. Kreinovich and S, Sriboonchitta Eds.), Springer-Verlag, Switzerland, pp. 152-162.

[12] Wang, T, Li, B. and Gupta, A. K. (2009). *Distribution of quadratic forms under skew normal settings*, Journal of Multivariate Analysis, 100, 533-545.

[13] Wang, Z., Wang, C. and Wang, T. (2016). *Estimation of location parameter on the skew normal setting with known coefficient of variation and skewness*, International Jurnal of Intelligent Technology and Applied Statistics, Vol.9, no.3, 45-63.

[14] Ye, R. and Wang, T. (2015). *Inferences in linear mixed models with skew-normal random effects*, Acta Mathematica Sinica, English Series, Volume 31, No. 4, pp. 576 - 594.

[15] Ye, R., Wang, T., and Gupta, A.K. (2014). *Distribution of matrix quadratic forms under skew normal settings*, Journal of Multivariate Analysis, 131, 229-239.

[16] Zhu, X., Ma, Z., Wang, T. and T. Teetranont (2017). *P*lausibility regions on the skewness parameter of skew normal distributions based on inferential models in "Robustness in Econometrics" (V. Krennovich, S. Sriboonchitta, and V. Huynh Eds.), Springer-Verlag, Switzerland, pp. 267-286.

[17] Zhu, X., Li, B., Wang, T. and Gupta, A. K. (2019). *Sampling distributions of skew normal populations associated with closed skew normal distributions*, Random Operators and Stochastic Equations. DOI:10.1515/rose-2018-2007.

[18] New Mexico State University (2018/19). *Budget estimate. [Salaries] (2018/19)*, Las Cruces, N.M.: The University 1994/95-.