The *a priori* procedure (APP) for estimating median under skew normal settings with applications in economics and finance

Liqun Hu
New Mexico State University, Las Cruces, New Mexico, USA
Tonghui Wang
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico, USA
David Trafimow
Department of Psychology, NMSU, Las Cruces, New Mexico, USA
S.T. Boris Choy
Discipline of Business Analytics, The University of Sydney, Sydney, Australia
Xiangfei Chen
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico, USA
Cong Wang
Mathematical and Statistical Sciences, University of Nebraska Omaha, Omaha, Nebraska, USA, and
Tingting Tong
Department of Mathematical Sciences, New Mexico State University, Las Cruces, New Mexico, USA

Received 11 September 2023
Revised 20 October 2023
Accepted 23 October 2023

Abstract

Purpose – The authors' conclusions are based on mathematical derivations that are supported by computer simulations and three worked examples in applications of economics and finance. Finally, the authors provide a link to a computer program so that researchers can perform the analyses easily.

Design/methodology/approach – Based on a parameter estimation goal, the present work is concerned with determining the minimum sample size researchers should collect so their sample medians can be trusted as good estimates of corresponding population medians. The authors derive two solutions, using a normal approximation and an exact method.

Findings – The exact method provides more accurate answers than the normal approximation method. The authors show that the minimum sample size necessary for estimating the median using the exact method is

The authors would like to thank three anonymous reviewers for their valuable suggestions and comments, which improved the quality of the manuscript.

Asian Journal of Economics and Banking
Emerald Publishing Limited
e-ISSN: 2633-7991
p-ISSN: 2615-9821
DOI 10.1108/AJEB-09-2023-0087

substantially smaller than that using the normal approximation method. Therefore, researchers can use the exact method to enjoy a sample size savings.

Originality/value – In this paper, the *a priori procedure* is extended for estimating the population median under the skew normal settings. The mathematical derivation and with computer simulations of the exact method by using sample median to estimate the population median is new and a link to a free and user-friendly computer program is provided so researchers can make their own calculations.

Keywords The *a priori* procedure, Minimum sample size, Skew normal population, Sample median, Confidence level, Precision

Paper type Research paper

1. Introduction

The *a priori* procedure (APP) is designed as a predata procedure where the goal is to estimate the sample size needed for sample statistics to be good estimates of corresponding population parameters. The researcher specifies how close she wants the sample statistic of concern to be to the corresponding population parameter, which is the precision issue. And she specifies the desired probability that the sample statistic will be within the range to which the precision specification refers, which is the confidence issue. For example, suppose the population distribution is normal, and the researcher wishes to have a 0.95 probability of obtaining a sample mean within 0.10 standard deviations of the population mean. In that case, the necessary sample size to meet the confidence and precision specifications is 385 (see Trafimow, 2017; Trafimow *et al.*, 2019).

Of course, researchers might not wish to assume normal distributions; for example, they could assume skew normal distributions. However, APPs exist there, too, with respect to locations (Trafimow *et al.*, 2019), scales (Wang *et al.*, 2019b) and shapes (Wang *et al.*, 2019a). There have been additional advances too (e.g. Cao *et al.*, 2021, 2022; Chen *et al.*, 2021; Tong *et al.*, 2022; Trafimow *et al.*, 2020; Wang *et al.*, 2021, 2022; Wei *et al.*, 2020; Wilson *et al.*, 2022). However, as the APP literature continues to expand, there is a simple issue that somehow has escaped investigation. Specifically, the APP literature has bypassed the humble median as a topic. And that is the present issue. Given researcher-provided specifications for precision and confidence, what sample size does the researcher need to collect so that the sample median estimates the population median within the limits of the specifications? For example, what sample size does a researcher need to collect to have a 0.95 probability of obtaining a sample median that is within 0.10 standard deviations of the population median? The subsequent section and an appendix provide the mathematical derivation. This is followed by computer simulations and two worked examples related to economics and finance, that support the derivation. We also provide a link to a free and user-friendly computer program so researchers can make their own calculations.

2. Sampling distribution of the median under skew normal settings

Suppose that X is a continuous random variable with probability density function (PDF), $f(x)$. The median, denoted by $\tilde{\mu}$, is a value of X satisfying

$$P\left(X \leq \tilde{\mu}\right) = \int_{-\infty}^{\tilde{\mu}} f(x)dx = \frac{1}{2} = \int_{\tilde{\mu}}^{\infty} f(x)dx = P\left(X \geq \tilde{\mu}\right).$$

Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with PDF $f(x)$ and Y_j be the jth order statistic of the sample, $j = 1, 2, \ldots, n$, i.e.

$$\min\{X_1, \ldots X_n\} = Y_1 \leq Y_2 \leq \cdots \leq Y_n = \max\{X_1, X_2, \ldots, X_n\}.$$

The sample median $\tilde{X}$ is defined by

$$\tilde{X} = \begin{cases} Y_{\frac{n+1}{2}} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(Y_{\frac{n}{2}} + Y_{\frac{n}{2}+1}\right) & \text{if } n \text{ is even.} \end{cases}$$

Consider a random sample of size $n = 2m - 1$ taken from the standard skew normal distribution $SN(\alpha)$ and let $\tilde{Z}$ be the sample median. According to order statistics, the PDF of $\tilde{Z}$ is given by

$$\begin{aligned} f_{\tilde{Z}}(z) &= \frac{n!}{(m-1)!(n-m)!} f_Z(z)[F_Z(z)]^{m-1}[1-F_Z(z)]^{n-m} \\ &= \frac{(2m-1)!}{[(m-1)!]^2} f_Z(z)\{[F_Z(z)[1-F_Z(z)]\}^{m-1}, \quad z \in \Re, \end{aligned}$$

where $f_Z(z)$ and $F_Z(z)$ are the PDF and cumulative distribution function (CDF) of $Z \sim SN(\alpha)$.

For a normal population, the sample mean $\overline{X}$ is the most efficient estimator of the population mean μ with the smallest variance. The efficiency of the sample median $\tilde{X}$, measured by the ratio of the variance of $\overline{X}$ to the variance of $\tilde{X}$, is $\frac{4m}{\pi(2m-1)}$ where $m = \frac{1}{2}(n+1)$. For large n, the efficiency is approximately equal to $\frac{2}{\pi}$.

In literature, the sampling distribution of the sample median from any continuous population with PDF $f(x)$ is asymptotically normal with mean $\mu = \tilde{\mu}$ and variance $\sigma^2 = 1/(4nf(\tilde{\mu})^2)$. See Chu, 1955. For small sample size, Rider (1960) discussed how well or how badly the variance of the asymptotically normal represents the true variance.

In this paper, we propose an APP for estimating the population median $\tilde{\mu}$ based on a random sample from a skew normal distribution. Skew normal distributions differ from normal distributions in three ways. The mean is replaced by the location, the standard deviation is replaced by the scale and there is the addition of a third parameter, the shape parameter. A random variable $Z \in \Re$ is said to follow a standard skew normal distribution if its PDF is given by (see Azzalini, 1985):

$$f_Z(z) = 2\phi(z)\Phi(\alpha z), \qquad z \in \Re, \tag{1}$$

where α is the shape parameter that controls the skewness of the distribution, $\phi(\cdot)$ and $\Phi(\cdot)$ are the PDF and the CDF of the standard normal distribution, respectively. For simplicity, we write $Z \sim SN(\alpha)$. The skew normal distribution is positively (negatively) skewed if $\alpha > 0$ $(\alpha < 0)$.

Note that the density curves of Z are positively skewed if $\alpha > 0$ and negatively skewed if $\alpha < 0$. This new class of distribution shares similar properties with the normal distribution. A location-scale extension to the non-standard skew normal distribution is given below.

Definition 1. Let $Z \sim SN(\alpha)$. The random variable $X = \xi + \omega Z$ follows a skew normal distribution with location $\xi \in \Re$ and scale $\omega^2 > 0$ and its PDF is given by

$$f_X(x) = \frac{2}{\omega}\phi\left(\frac{x-\xi}{\omega}\right)\Phi\left(\alpha\,\frac{x-\xi}{\omega}\right), \tag{2}$$

denoted by $X \sim SN(\xi, \omega^2, \alpha)$.

Two remarks are addressed below.

Remark 1. *The main purpose of this paper is to use the APP to obtain the minimum sample size required for estimating the population median, $\tilde{\mu}$, using the sample median, $\tilde{X}$. Therefore, without loss of generality, we assume that the sample size $n = 2m - 1$ is an odd number with a positive integer m.*

Remark 2. *Let $\tilde{\mu}_z$ and $\tilde{\mu}$ be the population medians of $Z \sim SN(\alpha)$ and $X \sim SN(\xi, \omega^2, \alpha)$, respectively. The linear relationship between X and Z gives $\tilde{\mu} = \xi + \omega\tilde{\mu}_z$. Similarly, if $X_1, X_2, \ldots, X_n$ form a random sample from the $SN(\xi, \omega^2, \alpha)$ distribution and $Y_1, Y_2, \ldots, Y_n$ be the corresponding order statistics. Then $\tilde{X} = Y_m = \xi + \omega\tilde{Z}$, where $\tilde{Z}$ is the sample median of the random sample, $Z_1, Z_2, \ldots Z_n$, from the $SN(\alpha)$ distribution, where $Z_i = (X_i - \xi)/\omega$, $i = 1, 2, \ldots, n$. Thus it suffices to find the minimum sample size for estimating $\tilde{\mu}_z$ using the sample median $\tilde{Z}$.*

The density curves of $\tilde{Z}$ for various values of n and α are displayed in Figures 1 and 2, respectively. Figure 1 shows that, for $n = 101$ and a fixed α value, the distribution of $\tilde{Z}$ is symmetric. Moreover, the location of the distribution increases and the scale of the distribution decreases as α increases. For $\alpha = 1$, Figure 2 shows that the distribution of $\tilde{Z}$ is symmetric with the same value for the location for different sample sizes. The scale of the distribution decreases as the sample size increases. We can see how the skewness parameter α affects the density curves in Figure 1, and how the sample size n affects the density curves in Figure 2.

Note that the relationship between the median $\tilde{\mu}_z$ and the skewness parameter α of the standard skew normal distribution is

$$\int_{-\infty}^{\tilde{\mu}_z} f_Z(z;\alpha)dz = \frac{1}{2} = \int_{\tilde{\mu}_z}^{\infty} f_Z(z;\alpha)dz.$$

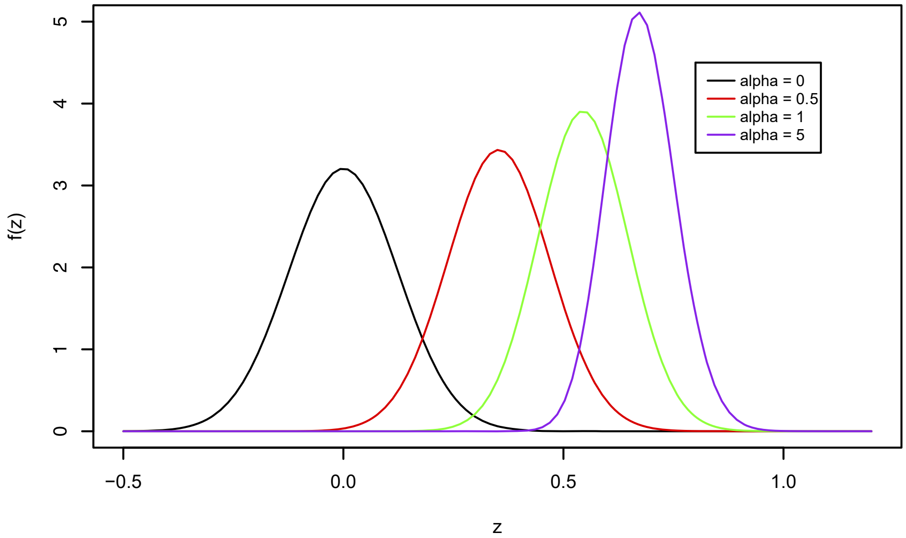
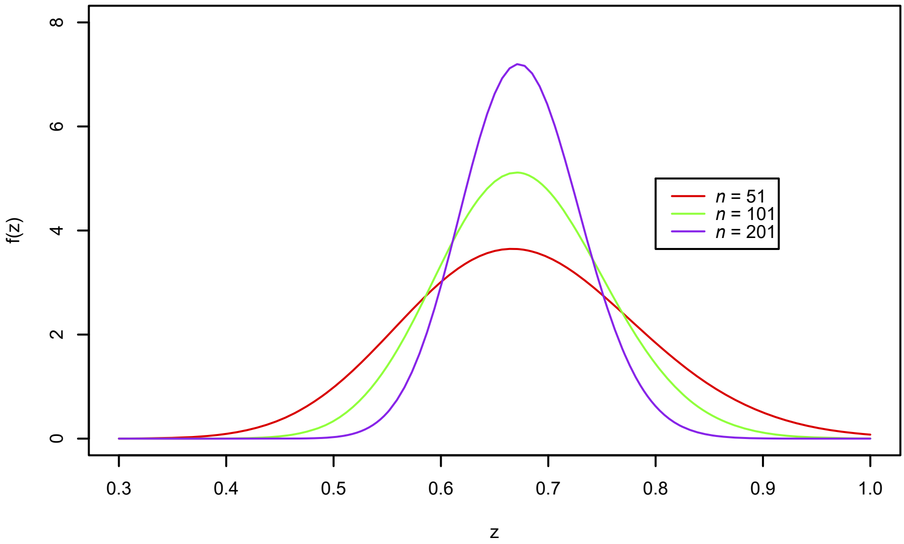


Figure 1.
Density curves of $\tilde{Z}$ with $\alpha = 0, 0.5, 1, 5$, and $n = 101$

Source(s): Figure by authors



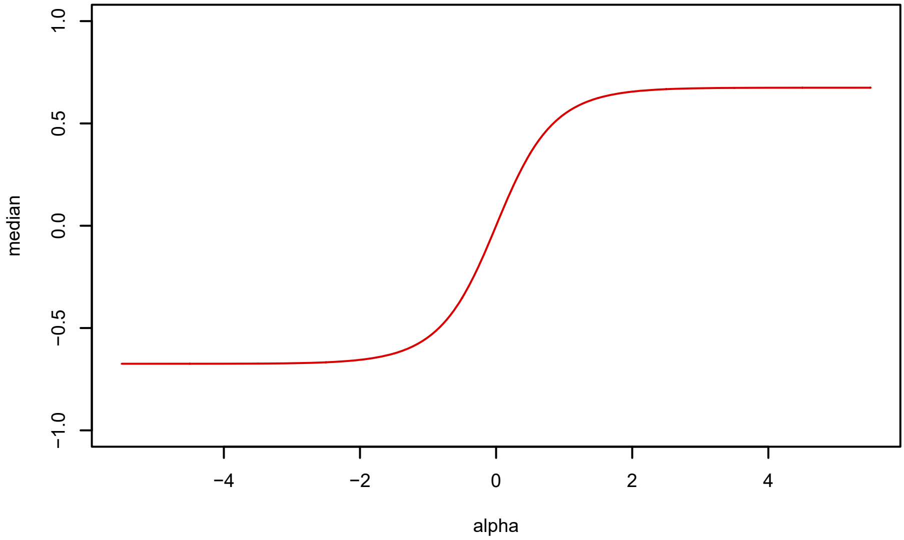
Source(s): Figure by authors

Figure 2. Density curves of $\tilde{Z}$ with $n = 51, 101, 201,$ and $\alpha = 5$

As shown in Figure 3 for $n = 101$, $\tilde{\mu}_z$ increases with α and $\tilde{\mu}_z$ will converge to the median of the standard half normal distribution as $\alpha \rightarrow \infty$.

3. The APP for estimating the population median

In this section, we will using two methods (exact and normal approximation) to find the minimum sample sizes for estimating the population median.



Source(s): Figure by authors

Figure 3. The relationship between $\tilde{\mu}_z$ and α in $SN(\alpha)$ for $n = 101$

3.1 The APP using an exact method

Theorem 1. Let $\tilde{Z}$ be the sample median of a random sample of size $n = 2m - 1$ from the standard skew normal population with skewness parameter α and $\tilde{\mu}_z$ be the population median. Let c be the confidence level and f be the precision, which is specified such that

$$P\left(\left|\tilde{Z} - \tilde{\mu}_z\right| \leq f\sigma\right) = c,$$

where σ is the standard deviation of the SN(α) distribution. Since the PDF of $\tilde{Z}$ is asymmetric, the above equation can be rewritten as

$$P\left(f_1 \leq \frac{\tilde{Z} - \tilde{\mu}_z}{\sigma} \leq f_2\right) = c, \tag{3}$$

where f_1 and f_2 are selected such that $\max\{|f_1|, f_2\} \leq f$*. Under APP, the required sample size n can be obtained from*

$$\int_{f_1}^{f_2} f_W(w)dw = c \tag{4}$$

subject to the length of the confidence interval $\ell = f_2 - f_1$ being the minimum, where $f_W(w)$ is the PDF of $W = (\tilde{Z} - \tilde{\mu}_z)/\sigma$ and is given by

$$f_W(w) = \frac{(2m-1)!}{[(m-1)!]^2} f_Z(u)\{F_Z(u)[1 - F_Z(u)]\}^{m-1}, \qquad u = \tilde{\mu}_z + \sigma z.$$

Proof: Note that the variance of $Z \sim SN(\alpha)$ is $\sigma^2 = 1 - \frac{2\delta^2}{\pi}$, where $\delta = \frac{\alpha}{\sqrt{1+\alpha^2}}$. For a given confidence level c and precision f, we set up Equation (4) and find the sample size n required such that the length of the confidence interval $\ell = (f_2 - f_1)$ is minimized. Let $W = (\tilde{Z} - \mu_{\tilde{z}})/\sigma$. It can be shown that the length of the confidence interval ℓ is minimized when $f_W(f_2) = f_W(f_1)$ so that the ratio $f_W(f_2)/f_W(f_1) = 1$ (The proof is given in Appendix: A). Thus the required sample size n, together with f_1 and f_2 such that $\max\{|f_1|, f_2\} \leq f$ can be determined simultaneously. □

Remark 3. To illustrate the results in Theorem 1, the required minimum sample size n with $\min\{|f_1|, f_2\} \leq f = 0.2$ *and the ratio $f_W(f_2)/f_W(f_1)$ are listed in Table 1 for c =* 0.95 *and values of* $\alpha = 1, 2, 5, 10$. *From this table, we can see that the required*

Table 1. The ratio $f_W(f_2)/f_W(f_1)$, f_1, f_2, and the required minimum sample size n for $c = 0.95$ and $f = 0.2$, and the values of $\alpha = 1$, 2, 5, 10

α	n	f_1	f_2	$f_W(f_2)/f_W(f_1)$
0	151	−0.2000	0.2000	1.0000
1	149	−0.1990	0.2000	0.9985
2	145	−0.1963	0.2000	0.9995
5	153	−0.1926	0.2000	1.0005
10	161	−0.1926	0.2000	0.9991

Source(s): Table by authors

sample sizes n obtained satisfy our assumptions in Theorem 1 *numerically. Also, the density curves of W for $n = 51, 101, 201$ and $\alpha = 1$ are given in Figure 4. These density curves are slightly skewed to right since $\alpha = 1$.*

Figure 4 shows that the distribution of W for $\alpha = 1$ and $n = 51, 101$ and 201, respectively. From the graph, one can see that the distribution is slightly skewed to right (or symmetric) with the same location while the scale of the distribution decreases with increasing sample size n.

3.2 The APP using the normal approximation

In the last subsection, we provide an exact method to obtain the minimum sample size for the estimation of the population median under the skew normal setting. In this subsection, we propose a normal approximation method to simplify the mathematical calculations. We will set up the APP procedure for estimating population median $\tilde{\mu}_z$ with sample median $\tilde{Z}$.

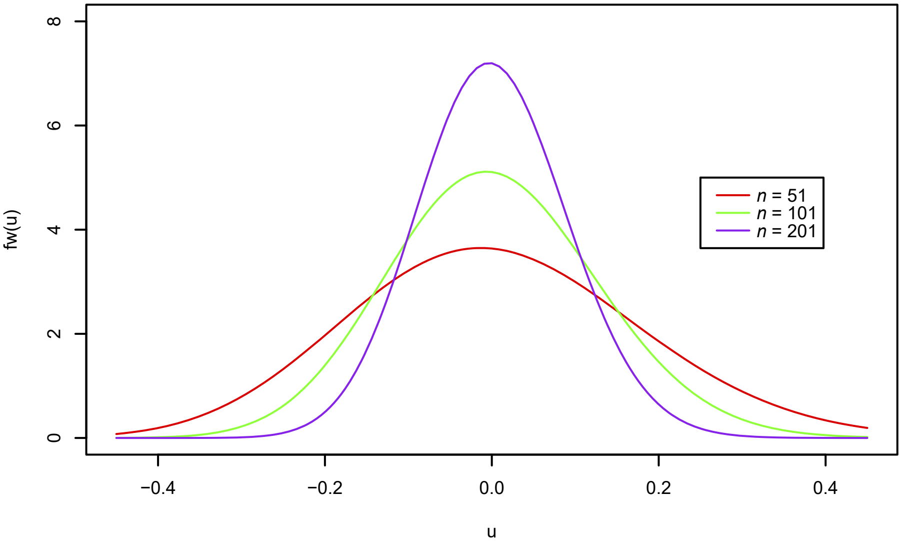
Theorem 2. *Suppose that $Z_1, Z_2, \ldots, Z_n$ form a random sample of size $n = 2m - 1$ from the standard skew normal distribution $SN(\alpha)$. Let c be a confidence level and f be the precision such that the error associated with the median estimator, $\tilde{Z}$, is $f\sigma_1$ with conference c, i.e.*

$$P\left(-f\sigma_1 \leq \tilde{Z} - \tilde{\mu}_z \leq f\sigma_1\right) = c, \tag{5}$$

where $\sigma_1^2 = [8f_Z(\tilde{\mu}_z)]^{-1}$ with $n = 2$ ($n = 2$ is the minimum sample size needed for the existence of median) and $f_Z(z)$ is given in Equation (1). *Then the required minimum sample size n is given by*

$$n = 2\left(\frac{z_{(1-c)/2}}{f}\right)^2 - 1, \tag{6}$$

where $z_{(1-c)/2}$ is the value of the standard normal random variable Z_0 such that $P(Z_0 > z_{(1-c)/2}) = (1-c)/2$.



Source(s): Figure by authors

Figure 4.
Density curves of W with $n = 51, 101, 201$ and $\alpha = 5$

Then the required sample size $n = 2m - 1$ (m is the middle number of sample size n, which is given in Equation (6)).

Proof: Since $\tilde{Z}$ can be approximated by the normal distribution with mean $\tilde{\mu}_z$ and variance $\sigma^2 = \left[4nf_Z(\tilde{\mu}_z)^2\right]^{-1}$, the confidence interval of $\tilde{\mu}_z$ with confidence c and precision f is given by

$$P\left(\tilde{Z} - z_{(1-c)/2}\cdot\sigma \leq \tilde{\mu}_z \leq \tilde{Z} + z_{(1-c)/2}\cdot\sigma\right) = c.$$

Note that $\sigma_1 = \sigma\sqrt{\frac{n}{2}}$. Thus the above confidence interval is equivalent to

$$P\left(-z_{(1-c)/2}\ \sigma_1\sqrt{\frac{2}{n}} \leq \tilde{Z} - \tilde{\mu}_z \leq z_{(1-c)/2}\ \sigma_1\sqrt{\frac{2}{n}}\right) = c. \tag{7}$$

From Equations (5) and (7), we obtain $f = z_{(1-c)/2}\sqrt{\frac{2}{n}}$, which is equivalent to Equation (6) and the desired result follows. □

> *Remark 4. Note that this normal approximation method applies to any continuous distribution with PDF f(x) and the minimum sample size n required is free from the skewness parameter α of the SN(α) distribution.*

The required sample size n needed for the given confidence levels $c = 0.90$ and $c = 0.95$, and various values of precision f for α are shown in Table 2. Note that the minimum sample size n needed with confidence level $c = 0.95$ and precision $f = 0.1$ is $769 = 2(385) - 1$ and 385 is the sample size required for estimating the population mean using the sample mean with the same c and f. See Trafimow, 2017; Trafimow *et al.* (2019) for details.

4. Simulation study

In this section, we perform a simulation study to evaluate the performance of the proposed APP. To compute the required minimum sample size n for estimating the population median $\tilde{\mu}_z$ of a standard skew normal distribution using the exact method, we provide an online calculator at: https://apprealization.shinyapps.io/nformedian/.

The minimum sample sizes n for various values of f and α are given in Table 3 for $c = 0.95$ and Table 4 for $c = 0.90$, respectively. Our programs were created using R software, and they are available upon request. For a fixed value of precision f, skewness α and confidence level c, 10,000 random samples with the corresponding minimum sample sizes (obtained in Tables 1 and 2) are based on the standard skew normal distribution. Then 10,000 confidence intervals

Table 2. Required sample sizes for estimating $\tilde{\mu}_z$ for the confidence level c = 0.95, 0.90 and precision f = 0.10, 0.15, 0.20, 0.25, and 0.30 using the normal approximation method

Precision (f)	Confidence (c)	m	$n = 2m - 1$
$f = 0.1$	0.95	385	769
	0.9	271	541
$f = 0.15$	0.95	171	341
	0.9	121	341
$f = 0.2$	0.95	97	193
	0.9	68	135
$f = 0.25$	0.95	62	123
	0.9	44	87
$f = 0.30$	0.95	43	85
	0.9	30	59

Source(s): Table by authors

for $\tilde{\mu}_z$ are constructed and coverage rates(cr) are obtained, which are given in Tables 3 and 4, respectively. For $c = 0.95$ (0.90), all coverage rates are near 0.95 (0.90).

Tables 3 and 4 show that at confidence level $c = 0.95$, the minimum sample sizes n needed for estimating the population median $\tilde{\mu}_z$ using the sample median $\tilde{Z}$ for given α and precision f. The tables also show the coverage rate(cr) with $M = 10000$ (Simulate 10,000 random samples from standard skew normal population with required sample size).

For $\alpha = 0$, the change in the minimum sample size n given the precision f using the exact and normal approximations methods are displayed in Figure 5. Figure 5 shows that n decreases as f increases for both $c = 0.95$ and $c = 0.90$ and the discrepancy in n under the two methods also reduces. Figure 5 indicates the comparison of the sample sizes n needed for using normal approximation and the sample sizes n needed for using exact method (for $\alpha = 0$).

Here, with the same minimum sample size n presented in Tables 3 and 4 for $c = 0.90$, we compare the coverage rate of estimating the parameter $\tilde{\mu}_z$ using the exact method with using the normal approximation and the results are shown in Table 5. Similarly, with the same minimum sample size n found in Table 1 for $c = 0.95$, we can also compare the coverage rate for estimating the parameter $\tilde{\mu}_z$ using the exact method with the coverage rate of estimating $\tilde{\mu}_z$ using the normal approximation. These comparisons are shown in Table 6.

In Tables 5 and 6, the coverage rates of using exact method are larger than those of using the normal approximation, which indicates that using the exact method is more efficient than using the normal approximation method.

5. Real data analyses

In this section, we analyze two real data sets to assess the performance of our APP methods for estimating the median under skew normal settings.

Example 5.1. This data set contains the salaries for San Francisco city employees in 2014 (with unit of 10,000 dollars). The total number of observations is 29773, and the median salary is 8.4680. This data set can be downloaded from: https://www.kaggle.com/datasets/kaggle/sf-salaries?resource=download

	$\alpha = 0$		$\alpha = 1$		$\alpha = 2$		$\alpha = 5$		$\alpha = 10$	
f	n	cr	n	cr	n	cr	n	cr	n	cr
0.10	605	0.9508	597	0.9476	577	0.9484	613	0.9509	643	0.9529
0.15	269	0.9473	265	0.9519	257	0.9453	273	0.9504	285	0.9492
0.20	151	0.9491	149	0.9523	145	0.9499	153	0.9517	161	0.9501
0.25	97	0.9519	95	0.9473	93	0.9482	99	0.9505	103	0.9511
0.30	67	0.9442	67	0.9475	65	0.9586	69	0.9488	71	0.9525

Source(s): Table by authors

Table 3.
For $c = 0.95$

	$\alpha = 0$		$\alpha = 1$		$\alpha = 2$		$\alpha = 5$		$\alpha = 10$	
f	n	cr	n	cr	n	cr	n	cr	n	cr
0.10	425	0.8953	421	0.8959	407	0.9020	433	0.8990	453	0.8992
0.15	189	0.8973	187	0.8954	181	0.8982	193	0.9032	201	0.8945
0.20	107	0.8945	105	0.8932	103	0.9026	109	0.9003	113	0.8959
0.25	69	0.8956	67	0.8938	65	0.8929	69	0.8975	69	0.9011
0.30	47	0.8885	47	0.8926	45	0.8912	49	0.9037	51	0.9005

Source(s): Table by authors

Table 4.
For $c = 0.90$

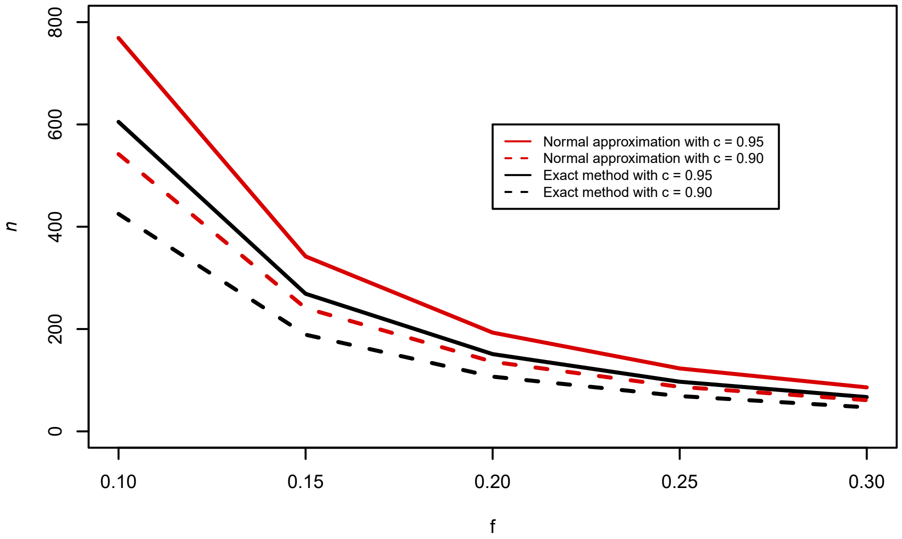


Figure 5. Sample size n ranges along the vertical axis as a function of precision f along the horizontal axis

Note(s): The curves indicate whether using normal approximation (red curves) or using exact method (black curves) and whether the confidence level is set at 0.95 or 0.90

Source(s): Figure by authors

Using the method of moment estimation, the fitted distribution is $SN(3.8718, 6.9955^2, 7.3712)$. The histogram and its fitted kernel and skew normal density curves of salaries are given, respectively, in Figure 6.

Now, if we choose $f = 0.10$ and $c = 0.95$, the required minimum sample size is $n = 635$ (exact method). We then draw a random sample of size $n = 635$. The sample median is 8.4945 and the 95% confidence interval for using exact method is (8.0714, 8.9176). The estimated median of the fitted model is 8.5902. The 95% confidence interval constructed using the normal approximation is (8.0662, 8.9228). Note that the population median 8.4680 falls into both confidence intervals. The length of the 95% confidence interval using the exact method

Table 5. Coverage rate (cr) of population median falling within confidence interval for using exact method, coverage rate (cr_N) of population median falling within confidence interval using normal approximation, population median ($\tilde{\mu}_z$), average, absolute bias ($|Bias|$) and standard deviation ($Std.Dev$) of sample median with $M = 10,000$ and n found in Tables 3 and 4

f	α	n	cr	cr_N	$\tilde{\mu}_z$	$\tilde{Z}_{average}$	$\|Bias\|$	$Std.Dev$
0.1	0	605	0.9508	0.9191	0.0000	−0.0043	0.0043	0.0508
	1	597	0.9476	0.9172	0.5449	0.5415	0.0034	0.0419
	2	577	0.9484	0.9103	0.6553	0.6528	0.0025	0.0357
	5	613	0.9509	0.9189	0.6744	0.6729	0.0015	0.0317
	10	643	0.9529	0.9290	0.6745	0.6730	0.0015	0.0309
0.2	0	151	0.9491	0.9156	0.0000	−0.0178	0.0178	0.1026
	1	149	0.9523	0.9105	0.5449	0.5308	0.0141	0.0856
	2	145	0.9499	0.9082	0.6553	0.6440	0.0113	0.0719
	5	153	0.9517	0.9218	0.6744	0.6662	0.0082	0.0630
	10	161	0.9501	0.9305	0.6745	0.6661	0.0084	0.0619

Source(s): Table by authors

is 0.8462, which is shorter than that (0.8566) using the normal approximation. This result confirms that the exact method is more efficient than the normal approximation method in the estimation of population median.

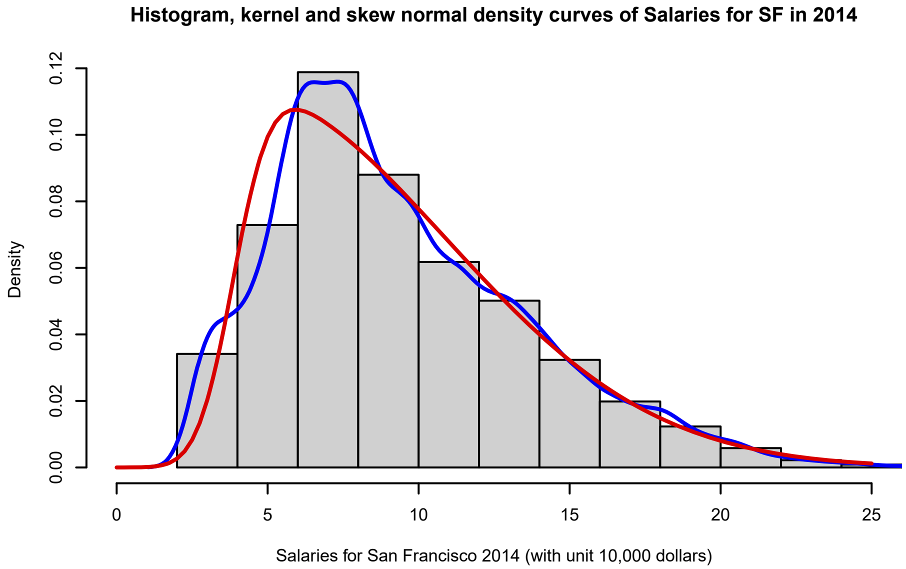
Example 5.2. This data set contains the market's opening price for wheat future in 2000–2023 (Futures are financial contracts obligating the buyer to purchase and the seller to sell a specified amount of a particular grain at a predetermined price on a future date). The total number of observations is 5,778, and the median is 5.0225 (The unit is dollars/bushel). The data set can be downloaded from: https://www.kaggle.com/datasets/guillemservera/grains-and-cereals-futures.

Using the method of moment estimation, the fitted skew normal distribution is $SN(3.1606, 3.2044^2, 3.9325)$. The histogram and its fitted kernel and skew normal density curves of market's opening price are given, respectively, in Figure 7.

Table 6. Coverage rate (cr) of population median falling within confidence interval for using exact method, coverage rate (cr_N) of population median falling within confidence interval using normal approximation, population median ($\tilde{\mu}_z$), average, absolute bias ($|Bias|$) and standard deviation ($Std.Dev$) of sample median with $M = 10,000$ and n found in Table 1 (for $c = 0.95$)

f	n	α	cr	cr_N	$\tilde{\mu}_z$	$\tilde{Z}_{average}$	$\vert Bias\vert$	$Std.Dev$
0.1	769	0	0.9731	0.9511	0.0000	−0.0037	0.0037	0.0450
		1	0.9755	0.9495	0.5449	0.5419	0.0030	0.0368
		2	0.9773	0.9489	0.6553	0.6532	0.0021	0.0310
		5	0.9728	0.9538	0.6744	0.6730	0.0014	0.0283
		10	0.9680	0.9516	0.6745	0.6730	0.0015	0.0287
0.2	193	0	0.9737	0.9524	0.0000	−0.0145	0.0145	0.0901
		1	0.9702	0.9465	0.5449	0.5332	0.0117	0.0764
		2	0.9770	0.9524	0.6554	0.6463	0.0091	0.0619
		5	0.9718	0.9486	0.6744	0.6669	0.0075	0.0573
		10	0.9685	0.9474	0.6745	0.6678	0.0067	0.0569

Source(s): Table by authors



Source(s): Figure by authors

Figure 6. The histogram, kernel density (blue) and skew normal density curve of $SN(3.8718, 6.9955^2, 7.3712)$ (red) of salaries for San Francisco city employees in 2014

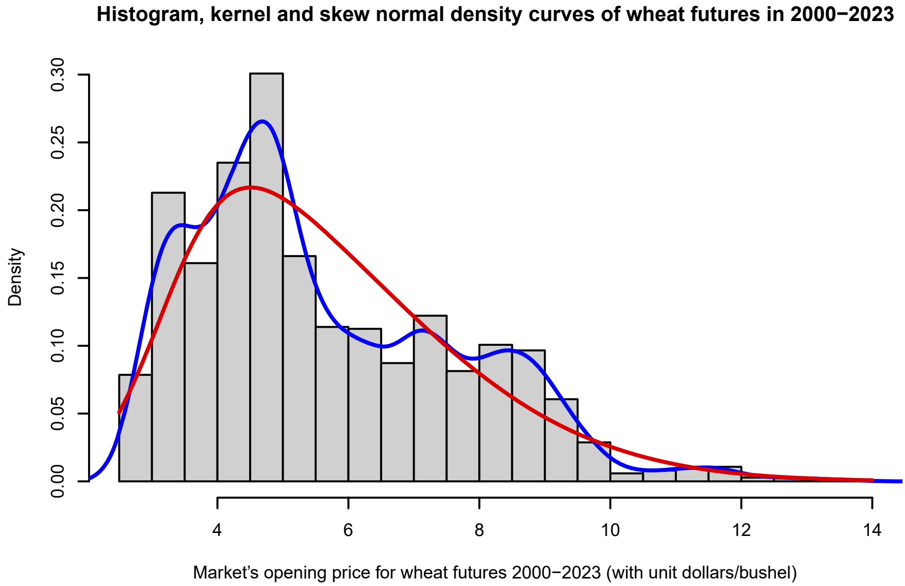


Figure 7. The histogram, kernel density (blue) and skew normal density curve of $SN(3.1606, 3.2044^2, 3.9325)$ (red) of market's opening price for wheat future in 2000–2023

Source(s): Figure by authors

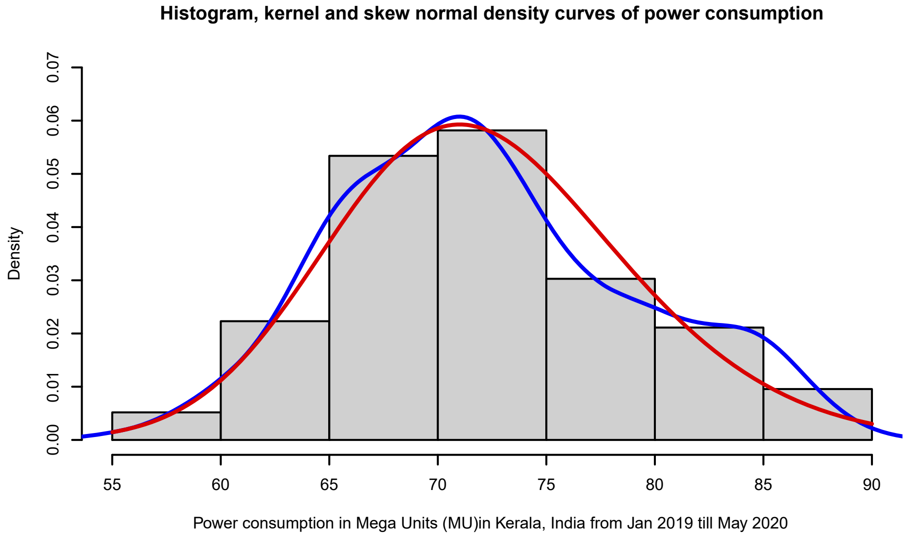
Now, if we use $f = 0.15$ and $c = 0.95$, the required minimum sample size is $n = 265$ (exact method). We then draw a random sample of size $n = 265$. The sample median is 5.2150, and the 95% confidence interval for using exact method is (4.9255, 5.5045). The estimated median of the fitted model is 5.3210. The 95% confidence interval constructed using the normal approximation is (4.9102, 5.5197). Note that the population median 5.0225 falls into both confidence intervals. The length of the 95% confidence interval using the exact method is 0.5790, which is shorter than that (0.6095) using the normal approximation. This result confirms that the exact method is more efficient than the normal approximation method in the estimation of population median.

Example 5.3. Consider the data set on power consumption in Mega Units (MU) in Kerala, India, from January 2019 till May 2020, given in https://www.kaggle.com/datasets/twinkle0705/state-wise-power-consumption-in-india/.

The total number of observations is 502, and the median is 71.45.

Using the method of moment estimation, the fitted skew normal distribution is $SN(65.9654, 9.2893^2, 1.5483)$. The histogram of the data set and its fitted kernel and skew normal density curves of power consumption are given, respectively, in Figure 8.

Now, if we use $f = 0.20$ and $c = 0.95$, the required minimum sample size is $n = 147$ (exact method). We then draw a random sample of size $n = 147$ from the whole data set and the sample median is 71.3000. The constructed 95% confidence interval for population median using exact method is (69.9333, 72.6667). The estimated median of the fitted model is 71.8059. The 95% confidence interval constructed using the normal approximation is (69.9212, 72.6788). Note that the population median 71.45 falls into both confidence intervals. The length of the 95% confidence interval using the exact method is 2.7333, which is shorter than that (2.7576) using the normal approximation. This result confirms that the exact method is more efficient than the normal approximation method in the estimation of population median.



Source(s): Figure by authors

Figure 8. The histogram, kernel density (blue) and skew normal density curve of *SN*(65.9654, 9.2893^2, 1.5483) (red) of power consumption in Kerala, India, from January 2019 till May 2020

6. Conclusions, limitations and future research

Although the exact method for obtaining the required minimum sample size provides better estimates than the normal approximation method, the latter is easier for researchers to use. However, the link to the program we provided adequately addresses the difficulty of using the exact method.

Two findings are noteworthy. The exact method results in smaller minimum sample sizes necessary for using sample medians to estimate corresponding population medians. Thus, the exact method provides sample size savings relative to the normal approximation method. Second, unlike the population mean, the population median is not a parameter of the skew normal distribution. Consequently, the required minimum sample size to meet specifications for precision and confidence should be larger for the median than the mean and our findings confirm this. However, an unexpected and unprecedented finding pertains to the effect of skewness on the determination of the minimum sample size necessary to meet specifications for precision and confidence. For the location, it has been well-documented that as the shape parameter increases, the required minimum sample size necessary to meet specifications decreases. For example, in Trafimow *et al.* (2019), suppose the specifications for precision and confidence are 0.10 and 0.95 for a random sample taken from the standard skew normal population. For location estimation, the minimum sample sizes needed when the shape parameter is 0 (normality), 0.5, 1, 2 and 5 are 385, 158, 146, 140 and 138, respectively. Note the monotonically decreasing sample size trend; if we remain with locations, the larger the shape parameter, the smaller the minimum sample size required to meet specifications. In contrast, for median estimation, the analogous sample sizes are 605, 603, 597, 577 and 613. Thus, the median differs dramatically from the mean in that median estimation does not follow decreasing monotonicity.

Researchers have often been advised to use the median, as opposed to the mean, when there is significant skewness. However, the present work qualifies the recommendation. To have impressive precision and confidence for using the sample median to estimate the population median, necessitates larger than typical sample sizes. Nor does increasing skewness provide sample size savings as when using the sample location to estimate the corresponding population location. This is not to say that researchers should never use the median, as there are times when the median is a useful statistic necessary for either theoretical or applied purposes. For skew normal population distributions, researchers should consider using the location instead of the median, if this is consistent with the researcher's theoretical or applied goals. If only the median can fit researcher goals, then the researcher should be prepared to either collect a large sample size or suffer a precision or confidence penalty.

For our future research, we will attempt to extend our APP for estimating parameters from skew normal distributions to other distributions, such as skew inverse Gaussian, log-skew normal, generalized gamma distributions, etc.

References

Azzalini, A. (1985), "A class of distributions which includes the normal ones", *Scandinavian Journal of Statistics*, Vol. 12 No. 2, pp. 171-178, available at: https://www.jstor.org/stable/4615982

Cao, L., Wang, C., Wang, T. and Trafimow, D. (2021), "The APP for estimating population proportion based on skew normal approximations and the Beta-Bernoulli process", *Communications in Statistics - Simulation and Computation*. doi: 10.1080/03610918.2021.2012192.

Cao, L., Tong, T., Trafimow, D., Wang, T. and Chen, X. (2022), "The a priori procedure for estimating the mean in both log-normal and gamma populations and robustness for assumption violations", *Methodology*, Vol. 18 No. 1, pp. 24-43, doi: 10.5964/meth.7321.

Chen, X., Trafimow, D., Wang, T., Tong, T. and Wang, C. (2021), "The APP procedure for estimating the cohen's effect size", *Asian Journal of Economics and Banking*, Vol. 5 No. 3, pp. 289-306, doi: 10.1108/AJEB-08-2021-0095.

Chu, J.T. (1955), "On the distribution of the sample median", *The Annals of Mathematical Statistics*, Vol. 26 No. 1, pp. 112-116, doi: 10.1214/aoms/1177728598, available at: https://www.jstor.org/stable/2236761

Rider, P.R. (1960), "Variance of the median of small samples from several special populations", *Journal of the American Statistical Association*, Vol. 55 No. 289, pp. 148-150, doi: 10.1080/01621459.1960.10482056?journalCode=uasa20.

Tong, T., Trafimow, D., Wang, T., Wang, C., Hu, L. and Chen, X. (2022), "The a priori procedure (APP) for estimating regression coefficients in linear models", *Methodology*, Vol. 18 No. 3, pp. 203-220, doi: 10.5964/meth.8245.

Trafimow, D. (2017), "Using the coefficient of confidence to make the philosophical switch from a posteriori to a priori inferential statistics", *Educational and Psychological Measurement*, Vol. 77 No. 5, pp. 831-854, doi: 10.1177/0013164416667977.

Trafimow, D., Wang, T. and Wang, C. (2019), "From a sampling precision perspective, skewness is a friend and not an enemy!", *Educational and Psychological Measurement*, Vol. 79 No. 1, pp. 129-150, doi: 10.1177/0013164418764801.

Trafimow, D., Wang, C. and Wang, T. (2020), "Making the a priori procedure work for differences between means", *Educational and Psychological Measurement*, Vol. 80 No. 1, pp. 186-198, doi: 10.1177/0013164419847509.

Wang, C., Wang, T., Trafimow, D. and Myüz, H.A. (2019a), "Desired sample size for estimating the skewness under skew normal settings", in *Structural Changes and Their Econometric Modeling 12*, Vol. 808, pp. 152-162, doi: 10.1007/978-3-030-04263-9_11.

Wang, C., Wang, T., Trafimow, D. and Zhang, X. (2019b), "Necessary sample size for estimating the scale parameter with specified closeness and confidence", *International Journal of Intelligent Technologies and Applied Statistics*, Vol. 12 No. 1, doi: 10.6148/IJITAS.201903_12(1).0002.

Wang, C., Wang, T., Trafimow, D., Li, H., Hu, L. and Rodriguez, A. (2021), "Extending the a priori procedure (APP) to address correlation coefficients", *Data Science for Financial Econometrics*, Vol. 898, pp. 141-149, doi: 10.1007/978-3-030-48853-6_10.

Wang, C., Wang, T., Trafimow, D. and Myüz, H.A. (2022), "Necessary sample sizes for specified closeness and confidence of matched data under the skew normal setting", *Communications in Statistics-Simulation and Computation*, Vol. 51 No. 5, pp. 2083-2094, doi: 10.1080/03610918.2019.1661473.

Wei, Z., Wang, T., Trafimow, D. and Talordphop, K. (2020), "Extending the a priori procedure to normal bayes models", *International Journal of Intelligent Technologies and Applied Statistics*, Vol. 13 No. 2, pp. 169-183, doi: 10.6148/IJITAS.202006_13(2).0004.

Wilson, C., Trafimow, D., Wang, T. and Wang, C. (2022), "Evaluating the sampling precision of social identity related published research", *Graduate Student Journal of Psychology*, Vol. 19, doi: 10.52214/gsjp.v19i.10053.

Appendix

In this Appendix, we are going to show the length of the confidence interval $\ell = f_2 - f_1$ is minimized subject to

$$\int_{f_1}^{f_2} f_W(w)dw = c \tag{A1}$$

if $f_W(f_2) = f_W(f_1)$. Indeed, treating f_2 as a function of f_1, we obtain, by differentiating both ℓ and Equation (A1) with respect to f_1,

$$\frac{\partial \ell}{\partial f_1} = \frac{\partial f_2}{\partial f_1} - 1 \qquad \text{and} \qquad f_W(f_2) \cdot \frac{\partial f_2}{\partial f_1} - f_W(f_1) = 0.$$

Solving these two equations, we obtain

$$f_W(f_2)\left(\frac{\partial \ell}{\partial f_1} + 1\right) - f_W(f_1) = 0.$$

To guarantee that ℓ is minimized, we should have $\frac{\partial \ell}{\partial f_1} = 0$, which is equivalent to

$$f_W(f_2) = f_W(f_1),$$

the desired result follows.

Corresponding author

Tonghui Wang can be contacted at: twang@nmsu.edu