Supplementary Material: Does the generalized mean have the potential to control outliers?

Soumalya Mukhopadhyay

Department of Statistics, Siksha Bhavana, Visva Bharati, Santiniketan 731235, India

Amlan Jyoti Das

Indian Statistical Institute, Kolkata, India

Ayanendranath Basu

Interdisciplinary Statistical Research Unit, Indian Statistical Institute, Kolkata 700108, India

Aditya Chatterjee

Department of Statistics, University of Calcutta, Kolkata 700019, India

Sabyasachi Bhattacharya*

Agricultural and Ecological Research Unit, Indian Statistical Institute, Kolkata 700108, India

1 Technical Proofs

In this section we present the technical proofs of Lemmas 2-4.

Lemma 2 : *Suppose $X_1, X_2, \ldots X_n$ are random samples from a population which consists of two components. The dominating population has mean μ_1 and variance ${\sigma_1}^2$, while μ_2 and ${\sigma_2}^2$ are the mean and variances of the outlying population. Let $\pi(> 1/2)$ be the mixing proportion and let c_1, c_2 be the coefficients of variation of*

*Corresponding author. e-mail:sabyasachi@isical.ac.in

the dominating and outlying population with the assumption $c_1, c_2 < 1/3$. *Then*

$$\frac{\sqrt{n}\left({M_\alpha}^\alpha - \theta^*\right)}{\sqrt{V^*}} \quad \rightarrow Z \sim \quad N(0,1), as \;\; n \rightarrow \infty$$

where

$$\begin{aligned}
\theta^* &= \pi{\mu_1}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_1^2}{2\mu_1^2}\right] + (1-\pi){\mu_2}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_2^2}{2\mu_2^2}\right] \\
V^* &= \left[\pi{\mu_1}^{2\alpha}\left[1 + \frac{\alpha(2\alpha-1)\sigma_1^2}{\mu_1^2}\right] + (1-\pi){\mu_2}^{2\alpha}\left[1 + \frac{\alpha(2\alpha-1)\sigma_2^2}{\mu_2^2}\right]\right] \\
&\quad -\left[\pi{\mu_1}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_1^2}{2\mu_1^2}\right] + (1-\pi){\mu_2}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_2^2}{2\mu_2^2}\right]\right]^2.
\end{aligned}$$

Proof: If $X_1, X_2, \ldots X_n$ are random observations from the distribution $F = \pi F_1 + (1-\pi)F_2, 0 < \pi < 1$, then

$$E_F\left(\frac{1}{n}\sum_{i=1}^{n}{X_i}^\alpha\right) = \pi E_{F_1}\left(\frac{1}{n}\sum_{i=1}^{n}{X_i}^\alpha\right) + E_{F_2}\left(\frac{1}{n}\sum_{i=1}^{n}{X_i}^\alpha\right).$$

Using the results of Lemma (1), we have

$$E_F\left(\frac{1}{n}\sum_{i=1}^{n}{X_i}^\alpha\right) \approx \pi{\mu_1}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_1^2}{2\mu_1^2}\right] + (1-\pi){\mu_2}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_2^2}{2\mu_2^2}\right] = \theta^* \;\; (say)$$

and

$$V_F\left(\frac{1}{n}\sum_{i=1}^{n}{X_i}^\alpha\right) = \frac{V^*}{n}$$

where

$$\begin{aligned}
V^* &= V_F\left(X_1^\alpha\right) \\
&= E_F\left(X_1^{2\alpha}\right) - \left[E_F\left(X_1^\alpha\right)\right]^2 \\
&= \left[\pi E_{F_1}\left(X_1^{2\alpha}\right) + (1-\pi)E_{F_2}\left(X_1^{2\alpha}\right)\right] - \left[\pi E_{F_1}\left(X_1^\alpha\right) + (1-\pi)E_{F_2}\left(X_1^\alpha\right)\right]^2 \\
&\approx \left[\pi{\mu_1}^{2\alpha}\left[1 + \frac{\alpha(2\alpha-1)\sigma_1^2}{\mu_1^2}\right] + (1-\pi){\mu_2}^{2\alpha}\left[1 + \frac{\alpha(2\alpha-1)\sigma_2^2}{\mu_2^2}\right]\right] \\
&\quad -\left[\pi{\mu_1}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_1^2}{2\mu_1^2}\right] + (1-\pi){\mu_2}^\alpha\left[1 + \frac{\alpha(\alpha-1)\sigma_2^2}{2\mu_2^2}\right]\right]^2.
\end{aligned}$$

Here ${M_\alpha}^\alpha = T = \frac{1}{n}\sum_{i=1}^{n}{X_i}^\alpha$. Then by Central limit theorem, we can say that

$$\frac{\sqrt{n}({M_\alpha}^\alpha - \theta^*)}{\sqrt{V^*}} \quad \rightarrow Z \sim \quad N(0,1), as \;\; n \longrightarrow \infty.$$

Lemma 3: Under the assumptions stated in Lemma 2, we have

$$\frac{(M_\alpha - \theta^{**})}{\sqrt{V^{**}}} \quad \to Z_1 \sim \quad N(0,1), as \;\; n \to \infty$$

where

$$\begin{aligned} \theta^{**} &= \theta^{*\frac{1}{\alpha}} \\ &= \left[\pi{\mu_1}^\alpha\left(1+\frac{\alpha(\alpha-1)\sigma_1^2}{2\mu_1^2}\right)+(1-\pi){\mu_2}^\alpha\left(1+\frac{\alpha(\alpha-1)\sigma_2^2}{2\mu_2^2}\right)\right]^{\frac{1}{\alpha}} \end{aligned}$$

and

$$V^{**} \;\; = \;\; \frac{1}{n\alpha^2}\theta^{*\frac{2}{\alpha}-2}V^*.$$

Proof:

Let T_n be a statistic based on $X_1, X_2, \ldots X_n$ such that

$$\frac{(T_n - \theta_0)}{\sigma_0} \quad \to Z \sim \quad N(0,1), as \;\; n \to \infty$$

and $g(T_n)$ is a continuous function of T_n. Then by delta method, the asymptotic distribution of $g(T_n)$ is also Normal with mean $g(\theta_0)$ and variance $\sigma_0^2[g'(\theta_0)]^2$, i.e.

$$\frac{(g(T_n) - g(\theta_0))}{\sigma_0 g'(\theta_0)} \quad \to Z_1 \sim \quad N(0,1), as \;\; n \to \infty.$$

Here $T_n = {M_\alpha}^\alpha$ and $g(T_n) = M_\alpha = {T_n}^{\frac{1}{\alpha}}$. Thus, $g'(T_n) = \dfrac{1}{\alpha}{T_n}^{\frac{1}{\alpha}-1}$ Again, ${\sigma_0}^2 = \dfrac{V^*}{n}$ and $\theta_0 = \theta^*$. Hence the desired result follows.

Lemma 4: Suppose we have a mixture of two populations (without any reference to a dominating or an outlying population) with respective means μ_1 and μ_2, standard deviations σ_1 and σ_2 and coefficients of variation c_1 and c_2. Without loss of generality, let $c_1 < c_2$. If c_p denotes the pooled coefficient of variation of the combined set, then either ${c_p}^2 > \max\{c_1, c_2\}$ or ${c_1}^2 < {c_p}^2 < {c_2}^2$. This indicates that ${c_p}^2$ is higher than c^2.

Proof: Let π be the proportion of observations from the first population. Then the mean and variance of the combined population are given respectively by

$$\begin{aligned} \mu &= \pi\mu_1 + (1-\pi)\mu_2, \\ \sigma^2 &= \pi{\sigma_1}^2 + (1-\pi){\sigma_2}^2 + \pi(1-\pi)(\mu_1-\mu_2)^2. \end{aligned}$$

So,

$$\begin{aligned} {c_p}^2 &= \frac{\sigma^2}{\mu^2} \\ &= \frac{\pi{\sigma_1}^2 + (1-\pi){\sigma_2}^2 + \pi(1-\pi)(\mu_1-\mu_2)^2}{(\pi\mu_1 + (1-\pi)\mu_2)^2} \\ &= \frac{\frac{{c_1}^2{\mu_1}^2}{1-\pi} + \frac{{c_2}^2{\mu_2}^2}{\pi} + (\mu_1-\mu_2)^2}{(\frac{\sqrt{\pi}}{\sqrt{1-\pi}}\mu_1 + \frac{\sqrt{1-\pi}}{\sqrt{\pi}}\mu_2)^2}. \end{aligned}$$

Therefore, after some algebraic manipulations, we have

$$\begin{aligned} {c_p}^2 - {c_1}^2 &= \frac{{c_1}^2\,({\mu_1}^2 + {\mu_2}^2 - 2\mu_1\mu_2) + \frac{{\mu_2}^2}{\pi}({c_2}^2 - {c_1}^2) + (\mu_1-\mu_2)^2}{(\frac{\sqrt{\pi}}{\sqrt{1-\pi}}\mu_1 + \frac{\sqrt{1-\pi}}{\sqrt{\pi}}\mu_2)^2} \\ &= \frac{(\mu_1-\mu_2)^2(1+{c_1}^2) + \frac{{\mu_2}^2}{\pi}({c_2}^2 - {c_1}^2)}{(\frac{\sqrt{\pi}}{\sqrt{1-\pi}}\mu_1 + \frac{\sqrt{1-\pi}}{\sqrt{\pi}}\mu_2)^2}. \end{aligned}$$

Since, $c_1 < c_2$, it follows that $c_p > c_1$, since the remaining terms are all positive.

Similarly, we found

$${c_2}^2 - {c_p}^2 = \frac{-(\mu_1-\mu_2)^2 + 2\mu_1\mu_2{c_2}^2 + \frac{{\mu_1}^2}{1-\pi}({c_2}^2 - {c_1}^2) - {c_2}^2({\mu_1}^2 + {\mu_2}^2)}{(\frac{\sqrt{\pi}}{\sqrt{1-\pi}}\mu_1 + \frac{\sqrt{1-\pi}}{\sqrt{\pi}}\mu_2)^2}.$$

Thus, ${c_2}^2 > {c_p}^2$ according as

$$\begin{aligned} \frac{{\mu_1}^2}{1-\pi}({c_2}^2 - {c_1}^2) &> (\mu_1-\mu_2)^2(1+{c_2}^2) \\ or,\;\; (\mu_1-\mu_2)^2 &< \frac{\frac{{\mu_1}^2}{1-\pi}({c_2}^2 - {c_1}^2)}{(1+{c_2}^2)}. \end{aligned}$$