King Fahd University of Petroleum & Minerals

DEPARTMENT OF MATHEMATICAL SCIENCES

Technical Report Series

TR 283

October 2002

Inequalities Involving Sample Means, Median and Extreme Observations

A. Laradji, A.H. Joarder

DHAHRAN 31261 ● SAUDI ARABIA ● www.kfupm.edu.sa/math/ ● E-mail: mathdept@kfupm.edu.sa

Inequalities Involving Sample Means, Median and Extreme Observations

A. LARADJI and A. H. JOARDER

Department of Mathematical Sciences, King Fahd University of Petroleum and Minerals Dhahran 31261, Saudi Arabia. Emails: alaradji@kfupm.edu.sa; anwarj@kfupm.edu.sa

Abstract Some inequalities involving sample means, sample median, the smallest and the largest observations are established. An upper bound of the absolute difference between the sample mean and median are also derived. Interesting inequalities obtained for cases when all the observations have the same sign.

1. Introduction

Inequalities involving sample means, median and extreme observations are not generally known. This pedagogical note is inspired by Shiffer and Harsha (1980) and Macleod and Henderson (1984) who worked on the bounds of sample standard deviation. Some inequalities involving sample means, sample median, the smallest and the largest observations are established. An upper bound of the absolute difference between the sample mean and median are also derived. Interesting inequalities are deduced for cases when all the observations are nonnegative or have the same sign. We believe that the inequalities will, in particular, provide additional information to students in statistics, and, in general, open a new direction of further research to refine inequalities on other sample statistics along the line of Shiffer and Harsha (1980), Macleod and Henderson (1984) and Eisenhauer (1983).

Both sample mean and median are popular measures of central tendency. However there are situations when one is preferred to the other. The sample mean is rigidly defined, fairly and easily calculated and quite intelligible to a layman. It also utilizes all the data and is highly amenable to mathematical treatment. Advanced statistical theories related to sample mean appear to be very elegant while with sample median, sample mode or any other measures of central tendency they are in most cases intractable. The main drawback of the sample mean is that it gives equal weight to all observations, and as such is affected by extreme observations.

Though the sample median is not rigidly defined for samples, it is easily computed, readily comprehensible and is not affected by extreme observations. It did not get as much popularity among statisticians because it is not well suited for algebraic manipulations. However it is preferred to sample mean in the situations (a) when the relative frequency distribution of the sample is highly asymmetrical (b) when there is an open class interval at one end or both ends of the relative frequency distribution (c) when it is difficult to measure the variable numerically (d) when the characteristic under study is population median (e) when the population under study is not symmetrical rather highly asymmetrical. Statistical theories related to sample median is related to Order Statistics, an area of statistics that has slowly developed in the last four decades.

Let $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ be the order statistics corresponding to the sample $(x_1, x_2, \cdots, x_n)$ with median $\tilde{x} = \frac{1}{2}\left(x_{([n/2+1/2])} + x_{([n/2+1])}\right)$ where $[m]$ is the bracket function denoting largest integer not exceeding m. Also let the arithmetic, geometric and harmonic mean be denoted by $a(x_1, x_2, \cdots, x_n) = \bar{x}$, $g(x_1, x_2, \cdots, x_n)$ and $h(x_1, x_2, \cdots, x_n)$ respectively. In this paper we establish interesting inequalities involving some of the sample characteristics, namely, $\bar{x}$, $g(x_1, x_2, \cdots, x_n)$, $h(x_1, x_2, \cdots, x_n)$, $\tilde{x}$, $x_{(1)}$ and $x_{(n)}$.

2. Main Results

Lemma 2.1 Let $x_i \le y_i$ $(1 \le i \le n)$. Then

(*i*) $\sum_{i=1}^{n} x_i \le \sum_{i=1}^{n} y_i$

(*ii*) $\prod_{i=1}^{n} x_i \le \prod_{i=1}^{n} y_i$ if $x_{(1)} \ge 0$.

(*iii*) $\sum_{i=1}^{n} \frac{1}{x_i} \ge \sum_{i=1}^{n} \frac{1}{y_i}$ if $x_{(1)} > 0$.

Consider the three sequences $A = \{a_1, a_2, \cdots, a_{2n}\}$, $B = \{b_1, b_2, \cdots, b_{2n}\}$ and $C = \{c_1, c_2, \cdots, c_{2n}\}$ each having $2n$ numbers defined by

$$a_k = \begin{cases} x_{(1)} & \text{if } 1 \le k \le n \\ \tilde{x} & \text{if } n+1 \le k \le 2n \end{cases},$$

$$b_k = x_{([k/2+1/2])} \text{ and } c_k = \begin{cases} \tilde{x} & \text{if } 1 \le k \le n \\ x_{(n)} & \text{if } n+1 \le k \le 2n \end{cases}$$

These sequences are then

$A = \{x_{(1)}, x_{(1)}, \cdots, x_{(1)}, \tilde{x}, \tilde{x}, \cdots, \tilde{x}\}$, $B = \{x_{(1)}, x_{(1)}, x_{(2)}, x_{(2)}, \cdots, x_{(n)}, x_{(n)}\}$ and $C = \{\tilde{x}, \tilde{x}, \cdots, \tilde{x}, x_{(n)}, x_{(n)}, \cdots, x_{(n)}\}$ where A and C contain n medians $(\tilde{x})$. For $1 \le k \le n$,

$a_k = x_{(1)} \le x_{([k/2+1/2])} = b_k \le \frac{1}{2}\left(x_{([n/2+1/2])} + x_{([n/2+1])}\right) = \tilde{x} = c_k$ and for $n+1 \le k \le 2n$,

$a_k = \tilde{x} \le \frac{1}{2}\left(x_{([n/2+1/2])} + x_{([n/2+1])}\right) \le x_{([k/2+1/2])} = b_k \le x_{(n)} = c_k$. Since the elements of the three sets satisfy the conditions of Lemma 2.1, we have the following theorem .

Theorem 2.1 For any sample of $n \ge 2$ observations $x_1, x_2, \cdots, x_n$ with $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, the following inequalities hold:

(*i*) $\dfrac{x_{(1)} + \tilde{x}}{2} \le \bar{x} \le \dfrac{\tilde{x} + x_{(n)}}{2}$

(*ii*) $\sqrt{x_{(1)}\, \tilde{x}} \le g(x_1, x_2, \cdots, x_n) \le \sqrt{\tilde{x}\, x_{(n)}}$ if $x_{(1)} \ge 0$ and

(*iii*) $\dfrac{2}{\dfrac{1}{x_{(1)}} + \dfrac{1}{\tilde{x}}} \le h(x_1, x_2, \cdots, x_n) \le \dfrac{2}{\dfrac{1}{\tilde{x}} + \dfrac{1}{x_{(n)}}}$ if $x_{(1)} > 0$ and

where $g(x)$ and $h(x)$ are the geometric and harmonic means of a sample of n observations.

Proof.

(*i*) Applying Lemma 2.1 (*i*) to the sets A and B, and then to B and C we have $nx_{(1)} + n\tilde{x} \le 2n\bar{x}$ and $2n\bar{x} \le nx_{(n)} + n\tilde{x}$ so that
$nx_{(1)} + n\tilde{x} \le 2n\bar{x} \le nx_{(n)} + n\tilde{x}$. Dividing this inequality by $2n$ we obtain the desired inequality.

(*ii*) Applying Lemma 2.1 (*ii*) to the above three sets of numbers A, B and C we have

$\left(x_{(1)}\, \tilde{x}\right)^n \le \left(x_1 x_2 \cdots x_n\right)^2 \le \left(\tilde{x}\, x_{(n)}\right)^n$ if $x_{(1)} \ge 0$. Since $g(x_1, x_2, \cdots, x_n) = \left(x_1 x_2 \cdots x_n\right)^{1/n}$ is the geometric mean of the observations, the proof is thus complete.

(*iii*) Applying Lemma 2.1 (*iii*) to the above three sets of numbers we have

$$\frac{2}{\dfrac{n}{x_{(1)}} + \dfrac{n}{\tilde{x}}} \le \frac{1}{\dfrac{2}{x_{(1)}} + \dfrac{2}{x_{(2)}} + \cdots + \dfrac{2}{x_{(n)}}} \le \frac{2}{\dfrac{n}{\tilde{x}} + \dfrac{n}{x_{(n)}}} \text{ if } x_{(1)} \ge 0.$$

Since $h(x_1, x_2, \cdots, x_n) = \dfrac{n}{\dfrac{1}{x_{(1)}} + \dfrac{1}{x_{(2)}} + \cdots + \dfrac{1}{x_{(n)}}}$ is the harmonic mean of the observations, the proof is complete.

Corollary 2.1 If $x_{(1)} > 0$, then $\frac{1}{2}h(x_1, x_2, \cdots, x_n) \le \tilde{x} \le 2\bar{x}$

Proof. By Theorem 2.1 (iii) we have

$$\frac{1}{2}h(x_1, x_2, \cdots, x_n) \le \frac{1}{1/\tilde{x} + 1/x_{(n)}} \le \frac{1}{1/\tilde{x}} = \tilde{x} \le x_{(1)} + \tilde{x}$$

and by Theorem 2.1 (i) we have $x_{(1)} + \tilde{x} \le 2\bar{x}$.

Theorem 2.2 For any sample of $n \ge 2$ observations $x_1, x_2, \cdots, x_n$ with $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, the following inequalities hold:

(*i*) $x_{(i)} + (n-i)x_{(i+1)} \le n\bar{x}$, for $1 \le i \le n$, $x_{(1)} = 0$

(*ii*) $\frac{1}{2}\left[\left(1-\frac{1}{n}\right)x_{(1)} + \left(1+\frac{1}{n}\right)\tilde{x}\right] \le \bar{x} \le \frac{1}{2}\left[\left(1+\frac{1}{n}\right)\tilde{x} + \left(1-\frac{1}{n}\right)x_{(n)}\right]$,

(*iii*) $|\tilde{x} - \bar{x}| \le \frac{n-1}{n+1} \max(\bar{x} - x_{(1)}, x_{(n)} - \bar{x})$

Proof. (i) Assume first that $x_{(1)} = 0$. Then, for $1 \le i \le n$, we have

$$\begin{aligned} n(x_{(i)} - \bar{x}) &= (x_{(i)} - x_{(1)}) + (x_{(i)} - x_{(2)}) + \cdots + (x_{(i)} - x_{(i-1)}) + 0 \\ &\quad + (x_{(i)} - x_{(i+1)}) + (x_{(i)} - x_{(i+2)}) + \cdots + (x_{(i)} - x_{(n)}) \\ &\le (i-1)x_{(i)} - [(x_{(i+1)} - x_{(i)}) + (x_{(i+2)} - x_{(i)}) + \cdots + (x_{(n)} - x_{(i)})] \\ &\le (i-1)x_{(i)} - (n-i)(x_{(i+1)} - x_{(i)}) \end{aligned}$$

so that

$$x_{(i)} + (n-i)x_{(i+1)} \le n\bar{x}. \tag{2.1}$$

(ii) For odd n and $i = (n+1)/2$, it follows from (2.1) that

$$\tilde{x} + \frac{n-1}{2} x_{((n+3)/2)} \le n\bar{x}$$

or, $\tilde{x} + \frac{n-1}{2} \tilde{x} \le n\bar{x}$

where we have used the fact that $\tilde{x} = x_{((n+1)/2)} \le x_{((n+3)/2)}$ for odd n. Hence for any odd n, we finally have

$$(n+1)\tilde{x} \le 2n\bar{x}\,. \tag{2.2}$$

When n is even, letting $i = n/2$ and $i = n/2+1$, it follows from (2.1) that

$$x_{(n/2)} + \frac{n}{2}\, x_{(n/2+1)} \le n\bar{x} \text{ and } x_{(n/2+1)} + \left(\frac{n}{2}-1\right) x_{(n/2+2)} \le n\bar{x}$$

$$\text{or, } \frac{n}{2}\, x_{(n/2+1)} + \left(\frac{n}{2}-1\right) x_{(n/2+2)} + 2\tilde{x} \le 2n\bar{x}$$

$$\text{or, } \frac{n}{2}\, \tilde{x} + \left(\frac{n}{2}-1\right) \tilde{x} + 2\tilde{x} \le 2n\bar{x} \tag{2.3}$$

where we have used the fact that $\tilde{x} \le x_{(n/2+1)} \le x_{(n/2+2)}$ for even n, so that the inequality (2.2) also follows from (2.3). Hence for any sample of size $n \ge 2$ with $x_{(1)} = 0$, we have

$$(n+1)\tilde{x} \le 2n\bar{x}\,. \tag{2.4}$$

Suppose now that $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, where $x_{(1)}$ is no longer assumed to be zero, and let $y_i = x_{(i)} - x_{(1)}$ $(1 \le i \le n)$ so that $\tilde{y} = \tilde{x} - x_{(1)}$, $\bar{y} = \bar{x} - x_{(1)}$. Then it follows from (2.4) that $(n+1)\tilde{y} \le 2n\bar{y}$ i.e. $(n+1)(\tilde{x} - x_{(1)}) \le 2n(\bar{x} - x_{(1)})$ so that $(n-1)x_{(1)} + (n+1)\tilde{x} \le 2n\bar{x}$.

Next from $-x_{(n)} \le -x_{(n-1)} \le \cdots \le -x_{(1)}$, similarly we obtain
$(n-1)(-x_{(n)}) + (n+1)(-\tilde{x}) \le 2n(-\bar{x})$ or, $(n+1)\tilde{x} + (n-1)x_{(n)} \ge 2n\bar{x}$.
The proof is thus complete.

(iii) By writing $2\bar{x} = \left(1-\frac{1}{n}\right)\bar{x} + \left(1+\frac{1}{n}\right)\bar{x}$, it follows from Theorem 2.2 (ii) that

$$\left(1-\frac{1}{n}\right)x_{(1)} + \left(1+\frac{1}{n}\right)\tilde{x} \le \left(1-\frac{1}{n}\right)\bar{x} + \left(1+\frac{1}{n}\right)\bar{x} \le \left(1+\frac{1}{n}\right)\tilde{x} + \left(1-\frac{1}{n}\right)x_{(n)}$$

$$\text{or, } \left(1-\frac{1}{n}\right)(\bar{x} - x_{(n)}) \le \left(1+\frac{1}{n}\right)(\tilde{x} - \bar{x}) \le \left(1-\frac{1}{n}\right)(\bar{x} - x_{(1)})$$

$$\text{or, } -\frac{n-1}{n+1}\left(x_{(n)} - \bar{x}\right) \le \tilde{x} - \bar{x} \le \frac{n-1}{n+1}\left(\bar{x} - x_{(1)}\right)$$

$$\text{or, } |\tilde{x} - \bar{x}| \le \frac{n-1}{n+1} \max(\bar{x} - x_{(1)}, x_{(n)} - \bar{x})\,.$$

It is worth noting that the inequalities

$$\frac{1}{2}(x_{(1)}+\tilde{x}) \le \bar{x} \le \frac{1}{2}(\tilde{x}+x_{(n)})$$

in Theorem 2.1 (i) can be deduced from Theorem 2.2 (ii) in the following way:

$$\frac{1}{2}(x_{(1)}+\tilde{x}) \le \frac{1}{2}\left(x_{(1)}+\tilde{x}+\frac{1}{n}(\tilde{x}-x_{(1)})\right)=\frac{1}{2}\left(\left(1-\frac{1}{n}\right)x_{(1)}+\left(1+\frac{1}{n}\right)\tilde{x}\right) \le \bar{x}$$

$$\le \frac{1}{2}\left[\left(1+\frac{1}{n}\right)\tilde{x}+\left(1-\frac{1}{n}\right)x_{(n)}\right]=\frac{1}{2}\left[\tilde{x}+x_{(n)}-\frac{1}{n}(x_{(n)}-\tilde{x})\right] \le \frac{1}{2}(\tilde{x}+x_{(n)})$$

Corollary 2.2 The following inequalities hold for any sample of $n \ge 2$ observations:

(i) $2|\bar{x}| \ge |\tilde{x}|$, if the observations have the same sign. (2.5)

(ii) $x_{(1)} \le \frac{2n}{n-1}\,\bar{x}+\left(1-\frac{2n}{n-1}\right)\tilde{x} \le x_{(n)}$ (2.6)

Proof. (i) If $x_{(1)} \ge 0$, then both $\bar{x}$ and $\tilde{x}$ are nonnegative, and $\tilde{x}/2 \le (x_{(1)}+\tilde{x})/2$ which cannot exceed $\bar{x}$ by (2.5). If $x_{(n)} \le 0$ then both $\bar{x}$ and $\tilde{x}$ are nonpositive and $\bar{x} \le (\tilde{x}+x_{(n)})/2$ which cannot exceed $\tilde{x}/2$ by (2.5) again. Taking absolute values we have the inequality in (i).

(ii) The inequalities follow directly from Theorem 2.2 (ii).

Remarks

(i) If the observations $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$ have the same sign then $2|\bar{x}| = |\tilde{x}|$ occurs exactly when all x's are equal to 0. If $x_{(1)} \ge 0$, then $2|\bar{x}| = |\tilde{x}|$ implies $2\bar{x} = \tilde{x}$ so that we have $0 \le \left(1-\frac{1}{n}\right)x_{(1)}+\left(1+\frac{1}{n}\right)\tilde{x} \le \tilde{x}$ by Theorem 2.2 (ii) and hence $\frac{1}{n}\,\tilde{x}+\left(1-\frac{1}{n}\right)x_{(1)} = 0$ which happens only if $\tilde{x} = 0$ i.e. if $2\bar{x} = 0$ and so all observations are 0's. A similar argument applies when $x_{(n)} \le 0$.

(ii) If $x_{(1)} > 0$, then $2\bar{x} \ge \frac{n+1}{n}\tilde{x} > \tilde{x}$ by Theorem 2.2 (ii). Similarly, $2\bar{x} < \tilde{x}$ if $x_{(n)} < 0$.

(iii) In case not all the observations have the same sign, an example of a sample showing $2\bar{x} = \tilde{x}$ may be: $n = 3$, $x_{(1)} = -10$, $x_{(2)} = 10$, $x_{(3)} = 15$ which could be average temperatures of three days in a city.

(iv) If all the observations are nonnegative, then for a negatively skewed distribution we have $\tilde{x}/2 \le \bar{x} \le \tilde{x}$, but for a positively skewed distribution we have $\tilde{x} \le \bar{x} \le (\tilde{x} + x_{(n)})/2$

Corollary 2.3 If $n \ge 2$ observations have the same sign, then $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| \le 1$.

Proof. Since x's have the same sign, it follows from (2.5) that

$\frac{\tilde{x}}{\bar{x}} = \left|\frac{\tilde{x}}{\bar{x}}\right| \le 2$. Now if $\frac{\tilde{x}}{\bar{x}} \ge 1$, then $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| = \frac{\tilde{x}}{\bar{x}} - 1 \le 1$ and if $\frac{\tilde{x}}{\bar{x}} < 1$, then $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| = 1 - \frac{\tilde{x}}{\bar{x}} < 1$. Thus, for any case, $\left|\frac{\tilde{x}}{\bar{x}} - 1\right| \le 1$.

Acknowledgements

The authors gratefully acknowledge the excellent research facilities available at King Fahd University of Petroleum and Minerals, Saudi Arabia.

References

Eisenhauer, J. G. (1993). A measure of relative dispersion. *Teaching Statistics*, **15**(2), 37-39.

Mcleod, A.J. and Henderson, G.R. (1984). Bounds for the sample standard deviation. *Teaching Statistics*, **6**(3), 72-76.

Shiffler, R.E. and Harsha, P.D. (1980). Upper and lower bounds for the sample standard deviation. *Teaching Statistics*, **2**(3), 84-86.