Upper bounds on deviations from the mean and the mean absolute deviation

Arya Aghili-Ashtiani
ORCID 0000-0002-7581-3415
Department of Electrical Engineering
Tafresh University
Tafresh, 39518-79611
Iran
arya.aghili@tafreshu.ac.ir

Abstract. In this paper, a number of upper bounds are introduced for the mean absolute deviation (MAD) from the mean when there is some information about the number of the data that are above or below the average. The upper bounds are compared with each other according to their tightness. A unified structure is found to be useful to prove all of the proposed upper bounds as well as the other existing upper bounds. In the path to formulate that structure, also, an upper bound is introduced for the individual deviations from the mean. The results are clarified and verified by a few examples.

Keywords: statistical dispersion, upper bound, partial knowledge, mean absolute deviation, individual deviations.

1. Introduction

Let $\{x_1, \dots, x_n\}$ be an arbitrary data-set, where $x_i \in \mathbb{R}$, and define $\mu_x = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, i.e., the (arithmetic) mean of the data, $x_{\min} = \min x_i$ and $x_{\max} = \max x_i$. Deviations from the mean are the distances of the data from their mean, i.e., $|x_i - \bar{x}|$, which is used in defining both the mean absolute deviation (MAD) and the standard deviation (SD). If $x_1 = x_2 = \cdots = x_n$, all deviations are zero and the data-set is trivial. Throughout the paper, *non-trivial* data-sets are considered in which at least one data is different from the others.

Three most common measures for statistical dispersion of a data-set $\{x_i\}$ are: (1) range: $r_x = x_{\max} - x_{\min}$; (2) MAD: $\rho_x = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$; (3) SD: $\sigma_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. For the MAD, we have

$$(1) \qquad \rho_x = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \leqslant \frac{1}{n}\sum_{i=1}^{n}(x_{\max} - x_{\min}) = r_x.$$

For the variance and the SD, we can write

$$(2) \qquad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \leqslant \frac{1}{n}\sum_{i=1}^{n}(x_{\max} - x_{\min})^2 = r_x^2$$

which means

$$\sigma_x \leqslant r_x. \tag{3}$$

Equations (1), (2), and (3), respectively, represent very simple upper bounds for MAD, variance, and SD of any arbitrary data-set $\{x_i\}$. Having an upper bound for a statistical measure is of analytical importance (e.g., to normalize the measure or to yield other useful inequalities) as well as practical importance (e.g., to verify the calculated measure for a given data-set). The tighter an upper bound, the more information it represents. Therefore, it is desirable to tighten the upper bounds as much as possible.

In Section 2, at first, some tight upper bounds are introduced for the individual deviations from the mean. Then, some upper bounds are introduced for the MAD which are tighter than (1). Simple examples are presented to clarify and justify the results after comparing the proposed upper bounds in Section 3.

Regarding the MAD as a statistical measure of dispersion, it is worth mentioning that some closed-form expressions have been introduced for the MAD in [3] for several probability distributions. The MAD can be used in real-world applications, e.g., to model the risk in a portfolio optimization problem, [4], [5]. The main benefit of using MAD-based optimization is that it allows for linear optimization which leads to a drastic reduction in the computational cost; See [10] to find more about this benefit and some others. The MAD is one of the measures that can be used to characterize the probability distributions; See [6] for example. The MAD per se may be considered as a random variable; See [7] where the probability distribution of the MAD and its estimation from data for some specific distributions are studied. Also, some other uses of MAD are introduced in [7], e.g., sample size determination for data with normal distribution.

2. Looking for appropriate upper bounds

Deviations from the mean of the data and their mean are considered in this section and a few somehow tight upper bounds are found for them.

2.1 Individual deviations from the mean

For every individual deviation from the mean we obviously know that $|x_i - \bar{x}| < x_{\max} - x_{\min} = r_x$. But let us introduce a tighter upper bound.

Theorem 2.1. *Let $\{x_1, \dots, x_n\}$ be a non-trivial data-set, where $x_i \in \mathbb{R}$. Then:*

$$\begin{array}{lll} (i) & \frac{1}{n} r_x \leqslant |x_i - \bar{x}| \leqslant \frac{n-1}{n} r_x & \text{for} \quad x_i \in \{x_{\min}, x_{\max}\}; \\ (ii) & |x_i - \bar{x}| \leqslant \frac{n-2}{n} r_x & \text{for} \quad x_i \notin \{x_{\min}, x_{\max}\}. \end{array}$$

Proof. Let us sort $\{x_i\}$ in non-decreasing order and denote the ordered data-set by $\{y_j\}$. So,

$$y_1 \leqslant \ldots \leqslant y_j \leqslant \ldots \leqslant y_n, \tag{4}$$

where $y_1 = x_{\min}, y_n = x_{\max}$, and $r_y = r_x = y_n - y_1$.
Case (i): For $x_i = x_{\max}$, we should show that $\frac{1}{n}r_x \leqslant x_{\max} - \bar{x} \leqslant \frac{n-1}{n}r_x$ which is equivalent to $\frac{1}{n}r_y \leqslant y_n - \bar{y} \leqslant \frac{n-1}{n}r_y$. By expanding $\bar{y}$ as $\frac{1}{n}(y_1 + \sum_{j=2}^{n-1} y_j + y_n)$ we can write

$$y_n - \bar{y} = -\frac{1}{n}y_1 - \frac{1}{n}\sum_{j=2}^{n-1} y_j + (1 - \frac{1}{n})y_n. \tag{5}$$

Using (4) we know that $(n-2)y_1 \leqslant \sum_{j=2}^{n-1} y_j \leqslant (n-2)y_n$. Therefore,

$$-\frac{1}{n}y_1 - \frac{n-2}{n}y_n + \frac{n-1}{n}y_n \leqslant y_n - \bar{y} \leqslant -\frac{1}{n}y_1 - \frac{n-2}{n}y_1 + \frac{n-1}{n}y_n$$

which leads to

$$\frac{1}{n}r_y \leqslant y_n - \bar{y} \leqslant \frac{n-1}{n}r_y. \tag{6}$$

Likewise, for $x_i = x_{\min}$, by writing

$$\bar{y} - y_1 = \frac{1-n}{n}y_1 + \frac{1}{n}\sum_{j=2}^{n-1} y_j + \frac{1}{n}y_n. \tag{7}$$

we obtain

$$\frac{1-n}{n}y_1 + \frac{n-2}{n}y_1 + \frac{1}{n}y_n \leqslant \bar{y} - y_1 \leqslant \frac{1-n}{n}y_1 + \frac{n-2}{n}y_n + \frac{1}{n}y_n$$

which leads to

$$\frac{1}{n}r_y \leqslant \bar{y} - y_1 \leqslant \frac{n-1}{n}r_y. \tag{8}$$

Case (ii): Now suppose that $x_i \in (x_{\min}, x_{\max})$ and x_i corresponds with y_j. So, for any $y_j \geqslant \bar{y}$ with $j \neq n$, we should show that $y_j - \bar{y} \leqslant \frac{n-2}{n}r_y$. To this end, we can expand $\bar{y}$ as $\frac{1}{n}(y_1 + \sum_{k=2,k\neq j}^{n-1} y_k + y_j + y_n)$ to write

$$y_j - \bar{y} = -\frac{1}{n}y_1 - \frac{1}{n}\sum_{k=2,k\neq j}^{n-1} y_k + (1 - \frac{1}{n})y_j - \frac{1}{n}y_n. \tag{9}$$

Using (4) we know that $y_1 \leqslant y_j \leqslant y_n$ and $(n-3)y_1 \leqslant \sum_{k=2,k\neq j}^{n-1} y_k \leqslant (n-3)y_n$ which leads to

$$-\frac{n-2}{n}r_y \leqslant y_j - \bar{y} \leqslant \frac{n-2}{n}r_y. \tag{10}$$

Likewise, for any $y_j \leqslant \bar{y}$ with $j \neq 1$, we obtain

$$-\frac{n-2}{n}r_y \leqslant \bar{y} - y_j \leqslant \frac{n-2}{n}r_y \tag{11}$$

by the same expansion of $\bar{y}$. This completes the proof. □

Remark 1. Beside presenting an upper bound, item (i) of Theorem 2.1 presents a lower bound on $(x_{\max} - \bar{x})$ and $(\bar{x} - x_{\min})$.

Example 1. Consider an ascendingly-sorted data-set $\{y_1, y_2, y_3, y_4\}$. According to Theorem 2.1, we have $\frac{1}{4}(y_4 - y_1) \leqslant \bar{y} - y_1 \leqslant \frac{3}{4}(y_4 - y_1)$, etc.

2.2 Mean Absolute Deviation (MAD)

As presented in (1), we know that $\rho_x \leqslant r_x$. However, in this section, we are going to find some tighter upper bounds for MAD.

To this end, let us sort $\{x_i\}$ in non-decreasing order and denote the ordered data-set by $\{y_j\}$. So, in comparison with the mean value $\bar{y} = \bar{x}$, we will have a *lower* data segment $(y_j < \bar{y})$, an *upper* data segment $(\bar{y} < y_j)$, and a *boundary* data segment $(y_j = \bar{y})$. Hence

$$y_1 \leqslant y_2 \leqslant \ldots \leqslant y_l < \bar{y} < y_{n-u+1} \leqslant \ldots \leqslant y_{n-1} \leqslant y_n, \tag{12}$$

where l and u are the number of data in the lower and upper data segments respectively. For a non-trivial data-set, l and u are in $[1, n-1]$ and

$$l + u \leqslant n. \tag{13}$$

Also, it is clear that $y_1 = x_{\min}, y_n = x_{\max}$, $r_x = r_y = y_n - y_1$, and $\rho_x = \rho_y = \frac{1}{n}\sum_{i=1}^{n} |y_i - \bar{y}|$.

It should be noticed that $n\bar{y} = \sum_{i=1}^{n} y_i$, so $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$. For the lower data segment, $y_i - \bar{y} < 0$, and for the upper data segment, $y_i - \bar{y} > 0$. Defining two positive indices $L := \sum_{j=1}^{l}(\bar{y} - y_j)$ and $U := \sum_{j=n-u+1}^{n}(y_j - \bar{y})$, we have

$$L = U. \tag{14}$$

On the other hand, $n\rho_y = \sum_{j=1}^{l}(\bar{y}-y_j)+\sum_{j=n-u+1}^{n}(y_j-\bar{y}) = L+U$, Therefore,

$$\rho_y = \frac{2}{n}L = \frac{2}{n}U. \tag{15}$$

Lemma 2.1. *Let $\{y_j\}$ be a data-set in non-decreasing order. B is an upper bound of the MAD ρ_y, if $L \leqslant \frac{n}{2}B$ (or equivalently $U \leqslant \frac{n}{2}B$).*

Proof. The proof is obvious by (15). □

Let us define the arithmetic mean of the lower and upper data segments respectively as $\bar{y}_l := \sum_{j=1}^{l} y_j/l$ and $\bar{y}_u := \sum_{i=n-u+1}^{n} y_j/u$. Therefore

$$\begin{cases} L = l\bar{y} - \sum_{j=1}^{l} y_j = l\left(\bar{y} - \bar{y}_l\right), \\ U = \sum_{j=n-u+1}^{n} y_j - u\bar{y} = u\left(\bar{y}_u - \bar{y}\right) \end{cases}. \tag{16}$$

Obviously, $y_1 \leqslant \bar{y}_l \leqslant \bar{y} \leqslant \bar{y}_u \leqslant y_n$.

Lemma 2.2. *Let $l, u, n \in \mathbb{N}$, such that $l + u \leqslant n$. Then,*

$$\frac{1}{l} + \frac{1}{u} \geqslant \frac{4}{n}. \tag{17}$$

Proof. We know that $u \leqslant n - l$. Thus, $\frac{1}{u} \geqslant \frac{1}{n-l}$, since $u \geqslant 1$ and $n - l \geqslant 1$. Therefore,

$$\frac{1}{l} + \frac{1}{u} \geqslant \frac{1}{l} + \frac{1}{n-l}$$

and the proof is complete since we know that $\frac{1}{l} + \frac{1}{n-l} \geqslant \frac{4}{n}$. □

Theorem 2.2. *Let $\{x_1, \ldots, x_n\}$ be a non-trivial data-set, where $x_i \in \mathbb{R}$. Then,*

$$\rho_x \leqslant \frac{1}{2} r_x. \tag{18}$$

Proof. We can equivalently prove $\rho_y \leqslant r_y/2$. By Lemma 2.1, it suffices to prove $L \leqslant nr_y/4$ or $U \leqslant nr_y/4$. Let us proceed by contradiction. Knowing $L = U$, suppose that L and U are greater than $nr_y/4$. By (16) we can write

$$\begin{cases} \bar{y} - \bar{y}_l > nr_y/4l, \\ \bar{y}_u - \bar{y} > nr_y/4u \end{cases}$$

which leads to $\bar{y}_u - \bar{y}_l > nr_y \left(\frac{1}{l} + \frac{1}{u}\right)/4$. Then, by Lemma 2.2, $\bar{y}_u - \bar{y}_l > r_y$ which is a contradiction. □

The upper bound presented by Theorem 2.2 is independent of the size of the data-set (n) and only uses the range (r_x). Even tighter upper bounds may be found for the MAD, when more information is available.

If the mean of the data is known in addition to $x_{\min}$ and $x_{\max}$, a tighter upper bound may be found. Considering the fact that the MAD is the arithmetic mean of the absolute value of individual deviations, we immediately find

$$\rho_x \leqslant \max\left(x_{\max} - \bar{x}, \bar{x} - x_{\min}\right) \tag{19}$$

which seems to represent an appropriate upper bound but a better one is presented in Theorem 2.3.

Lemma 2.3. *Let $\{x_1, \ldots, x_n\}$ be a non-trivial data-set, where $x_i \in \mathbb{R}$. Then,*

$$4lu \leqslant n^2. \tag{20}$$

Proof. Knowing $4lu \leqslant 4l(n - l)$ it suffices to prove $4l(n - l) \leqslant n^2$ which can be rewritten as $0 \leqslant n^2 - 4nl + 4l^2 = (n - 2l)^2$. □

Theorem 2.3. *Let $\{x_1, \ldots, x_n\}$ be a non-trivial data-set, where $x_i \in \mathbb{R}$. Then,*

$$\rho_x \leqslant \sqrt{(x_{\max} - \bar{x})(\bar{x} - x_{\min})}. \tag{21}$$

Proof. By Lemma 2.1, it suffices to prove that either L or U is less than or equal to $n\sqrt{(y_n-\bar{y})(\bar{y}-y_1)}/2$. By contradiction and by (16) we can write

$$\begin{cases}\bar{y}-\bar{y}_l > n\sqrt{(y_n-\bar{y})(\bar{y}-y_1)}/2l,\\ \bar{y}_u-\bar{y} > n\sqrt{(y_n-\bar{y})(\bar{y}-y_1)}/2u\end{cases}$$

to result in

$$\frac{(\bar{y}_u-\bar{y})}{(y_n-\bar{y})}\frac{(\bar{y}-\bar{y}_l)}{(\bar{y}-y_1)} > n^2/4lu$$

which is a contradiction, since the left hand side is less than or equal to one, but by Lemma 2.3, the right hand side is greater than or equal to one. □

It is interesting to notice that the MAD and the SD are respectively the arithmetic mean and the quadratic mean of the absolute value of individual deviations. Thus, knowing the fact that the quadratic mean is always greater than or equal to the arithmetic mean, *every upper bound for the SD is an upper bound for the MAD too.* In [2], an upper bound that uses the mean of the data-set (known as Bhatia-Davis) is obtained for the SD which is the same as (21) and is tighter than (3), since $(x_{\max}-\bar{x})(\bar{x}-x_{\min}) < (x_{\max}-x_{\min})^2$. It is worth mentioning that the Bhatia-Davis upper bound can be used directly when $\bar{x}$ is known. If $\bar{x}$ is not known, $(x_{\max}-\bar{x})(\bar{x}-x_{\min})$ should take its maximum value which happens when $x_{\max}-\bar{x}=\bar{x}-x_{\min}$ or $\bar{x}=(x_{\max}+x_{\min})/2$. Therefore, we obtain

$$\sigma_x \leqslant r_x/2, \tag{22}$$

which represents the Popoviciu's upper bound for SD, introduced in [8]. Popoviciu's upper bound is looser than Bhatia-Davis (because of not using the value of $\bar{x}$) but tighter than (3) (while using the same amount of information). Besides, note that the Bhatia-Davis upper bound is indeed the geometric mean of $(x_{\max}-\bar{x})$ and $(\bar{x}-x_{\min})$. The harmonic mean of $(x_{\max}-\bar{x})$ and $(\bar{x}-x_{\min})$, represented in (23), too, is an upper bound for MAD which has been introduced firstly in [1] and used in [9]. Theorem 2.4 presents a new straightforward proof for it.

Theorem 2.4. *Let $\{x_1,\dots,x_n\}$ be a non-trivial data-set, where $x_i\in\mathbb{R}$. Then,*

$$\rho_x \leqslant 2(x_{\max}-\bar{x})(\bar{x}-x_{\min})/r_x. \tag{23}$$

Proof. According to Lemma 2.1, it is sufficient to prove that $n(y_n-\bar{y})(\bar{y}-y_1)/r_y$ is an upper bound for L and U. We will proceed by contradiction. So, suppose that $L=U$ is greater than $n(y_n-\bar{y})(\bar{y}-y_1)/r_y$. Then, knowing $r_y>0$, by (16), we can write

$$\begin{cases}l(\bar{y}-\bar{y}_l)r_y > n(y_n-\bar{y})(\bar{y}-y_1),\\ u(\bar{y}_u-\bar{y})r_y > n(y_n-\bar{y})(\bar{y}-y_1)\end{cases}$$

which can be rewritten as

$$\begin{cases} l(\bar{y}-y_1)r_y + l(y_1-\bar{y}_l)r_y > n(y_n-\bar{y})(\bar{y}-y_1), \\ u(\bar{y}_u-y_n)r_y + u(y_n-\bar{y})r_y > n(y_n-\bar{y})(\bar{y}-y_1) \end{cases} . \tag{24}$$

On one hand, we know that $y_1 - \bar{y}_l < 0$ and $\bar{y}_u - y_n < 0$. Therefore, (24) can be reduced to

$$\begin{cases} l(\bar{y}-y_1)r_y > n(y_n-\bar{y})(\bar{y}-y_1), \\ u(y_n-\bar{y})r_y > n(y_n-\bar{y})(\bar{y}-y_1) \end{cases} . \tag{25}$$

On the other hand, we know that $\bar{y} - y_1 > 0$ and $y_n - \bar{y} > 0$. Therefore, (25) can be reduced to

$$\begin{cases} lr_y > n(y_n-\bar{y}), \\ ur_y > n(\bar{y}-y_1) \end{cases} . \tag{26}$$

Summing the inequalities (26) leads to $l + u > n$ which is a contradiction and the proof is complete. □

In what follows we are going to introduce some tighter upper bounds for the MAD when some information is available about the number of the data below or above the average.

Lemma 2.4. *Let s, l, u, n be natural numbers such that $l + u \leqslant n$ and s belongs to $[\min(l, n-l, u, n-u), \max(l, n-l, u, n-u)]$. Then,*

$$lu \leqslant \min\big(l(n-l), u(n-u)\big) \leqslant s(n-s). \tag{27}$$

Proof. The left inequality is correct, since l and u are positive, $u \leqslant n - l$, and $l \leqslant n - u$. For the right inequality, let us define $f(s) := s(n-s) = -s^2 + ns$ which is a parabolic function of s as depicted in Fig. 1. Obviously, $f(s) > \min\big(f(l), f(u)\big)$ for every s in the specified interval, since f is a convex function. This completes the proof. □

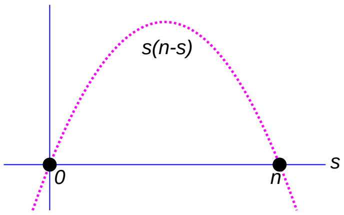

Figure 1: $f(s) = s(n-s)$

Theorem 2.5. *Let $\{x_1, \ldots, x_n\}$ be a non-trivial data-set, where $x_i \in \mathbb{R}$. Then,*

$$\rho_x \leqslant \frac{\sqrt{lu}}{n} r_x \leqslant \frac{1}{n} \min\left(\sqrt{l(n-l)}, \sqrt{u(n-u)}\right) r_x \leqslant \frac{\sqrt{s(n-s)}}{n} r_x. \tag{28}$$

Proof. Regarding the left inequality, according to Lemma 2.1, it suffices to prove $L \leqslant r_y\sqrt{lu}/2$ or $U \leqslant r_y\sqrt{lu}/2$. We can prove it by contradiction. Knowing $L = U$, suppose that L and U are greater than $r_y\sqrt{lu}/2$. By (16) we can write

$$\begin{cases} \bar{y} - \bar{y}_l > r_y\sqrt{\frac{u}{l}}/2, \\ \bar{y}_u - \bar{y} > r_y\sqrt{\frac{l}{u}}/2 \end{cases}$$

which leads to $\bar{y}_u - \bar{y}_l > r_y(\sqrt{\frac{u}{l}} + \sqrt{\frac{l}{u}})/2 \geqslant r_y$ which is a contradiction. The middle and the right inequality are obtained directly from Lemma 2.4. □

Lemma 2.5. *Let $l, u, n \in \mathbb{N}$, such that $l + u \leqslant n$. Then,*

$$\frac{lu}{l+u} \leqslant \frac{1}{n} \min\big(l(n-l), u(n-u)\big) \leqslant \frac{s(n-s)}{n}. \tag{29}$$

Proof. First consider the left inequality. We know that $l \geq 0$ and $l + u \leq n$. So, $0 \leq -l(l + u - n)$. By adding nu to both sides we have $nu \leq -l^2 - lu + nl + nu = (n-l)(l+u)$. By multiplying l to both sides and rearranging we have $\frac{lu}{l+u} \leq \frac{l(n-l)}{n}$. Likewise, we derive $\frac{lu}{l+u} \leq \frac{u(n-u)}{n}$ and the left inequality is proved. The proof of the right inequality is the same as the proof of the right inequality in Lemma 2.4. □

Theorem 2.6. *Let $\{x_1, \ldots, x_n\}$ be a non-trivial data-set, where $x_i \in \mathbb{R}$. Then,*

$$\rho_x \leqslant \frac{2lu}{l+u}.\frac{r_x}{n} \leqslant \frac{2}{n^2} \min\left(l(n-l), u(n-u)\right) r_x \leqslant \frac{2s(n-s)}{n^2} r_x, \tag{30}$$

for every $s \in [\min(l, n-l, u, n-u), \max(l, n-l, u, n-u)]$.

Proof. First consider the left inequality. According to Lemma 2.1, it suffices to prove $L \leqslant \frac{lu}{l+u} r_y$ or $U \leqslant \frac{lu}{l+u} r_y$. Let us proceed by contradiction. Having $L = U$, suppose that L and U are greater than $\frac{lu}{l+u} r_y$. So, by (16) we have

$$\begin{cases} \bar{y} - \bar{y}_l > \frac{u}{l+u} r_y, \\ \bar{y}_u - \bar{y} > \frac{l}{l+u} r_y \end{cases}$$

which leads to $\bar{y}_u - \bar{y}_l > r_y$ which is a contradiction. The middle and the right inequality are obtained directly from Lemma 2.5. □

Table 1: Summary of the upper bounds and their required information

Label	Upper Bound	Required Information	Reference
B_0	r_x	r_x	(obvious)
B_1	$r_x/2$	r_x	(18) [8]
B_2	$\sqrt{(x_{max}-\bar{x})(\bar{x}-x_{\min})}$	$x_{\min}, x_{\max}, \bar{x}$	(21) [2]
B_3	$2(x_{\max}-\bar{x})(\bar{x}-x_{\min})/r_x$	$x_{\min}, x_{\max}, \bar{x}$	(23) [1]
B_4	$\frac{\sqrt{s(n-s)}}{n} r_x$	r_x, n, s	(28)
B_5^l	$\frac{\sqrt{l(n-l)}}{n} r_x$	r_x, n, l	(28)
B_5^u	$\frac{\sqrt{u(n-u)}}{n} r_x$	r_x, n, u	(28)
B_5	$\min\left(B_5^l, B_5^u\right)$	r_x, n, l, u	(28)
B_6	$\frac{\sqrt{lu}}{n} r_x$	r_x, n, l, u	(28)
B_7	$2\frac{s(n-s)}{n^2} r_x$	r_x, n, s	(30)
B_8^l	$2\frac{l(n-l)}{n^2} r_x$	r_x, n, l	(30)
B_8^u	$2\frac{u(n-u)}{n^2} r_x$	r_x, n, u	(30)
B_8	$\min\left(B_8^l, B_8^u\right)$	r_x, n, l, u	(30)
B_9	$2lur_x/(l+u)n$	r_x, n, l, u	(30)

3. Summary and examples

Let us assign a label to each upper bound for ease of reference. Table 1 shows the assigned labels along with the information required to calculate the upper bounds.

It is obvious that $B_1 \leq B_0$. Also, as B_1, B_2 and B_3 can be considered respectively as arithmetic, geometric, and harmonic means of $(x_{\max} - \bar{x})$ and $(\bar{x} - x_{\min})$, it is obvious that $B_3 \leq B_2 \leq B_1$.

Note that by $0 \leqslant n^2 - 4sn + 4s^2 = (n-2s)^2$, we have $4s(n-s) \leqslant n^2$ and then $\frac{2}{n}\sqrt{s(n-s)} \leqslant 1$. So, we can write

$$2\frac{s(n-s)}{n^2} \leqslant \frac{\sqrt{s(n-s)}}{n}, \tag{31}$$

which leads to $B_7 \leqslant B_4$, $B_8^l \leqslant B_5^l$, $B_8^u \leqslant B_5^u$, and $B_8 \leqslant B_5$.

Furthermore, for a fixed n, both $2\frac{s(n-s)}{n^2}$ and $\frac{\sqrt{s(n-s)}}{n}$ take their maximum value when s is the nearest integer to $n/2$, So, we can write

$$\frac{\sqrt{s(n-s)}}{n} \leqslant \frac{1}{2} \tag{32}$$

which means $B_4 \leqslant B_1$ and $B_7 \leqslant B_1$.

We can categorize the upper bounds with respect to the required information, as shown in Table 2. Also in this table, the best upper bound is introduced for each information group.

Table 2: Information groups and suitable upper bounds for MAD

Group Label	Available Information	Possible Bounds	Tightest Bound
I_1	r_x	B_0,B_1	B_1
I_2	$x_{\min}, x_{\max}, \bar{x}$	B_2,B_3	B_3
I_3^s	r_x, n, s	B_4,B_7	B_7
I_3^l	r_x, n, l	B_5^l,B_8^l	B_8^l
I_3^u	r_x, n, u	B_5^u,B_8^u	B_8^u
$I_3^{l,u}$	r_x, n, l, u	B_5,B_6,B_8,B_9	B_9

Table 3: The actual MAD versus some upper bounds for each data-set in Example 2: B_1, B_2, and B_3 respectively from [8], [2], and [1]; B_5, B_6, B_8, and B_9 proposed in this paper.

Set	l	u	MAD	B0	B1	B2	B3	B5	B6	B8	B9
A	50	50	5.000	10	5	5.000	5.000	5.000	5.000	5.000	5.000
B	75	25	3.750	10	5	4.330	3.750	4.330	4.330	3.750	3.750
C	90	10	1.800	10	5	3.000	1.800	3.000	3.000	1.800	1.800
D	99	1	0.198	10	5	0.995	0.198	0.995	0.995	0.198	0.198
E	25	25	2.500	10	5	5.000	5.000	4.330	2.500	3.750	2.500
F	10	10	1.000	10	5	5.000	5.000	3.000	1.000	1.800	1.000
G	1	1	0.100	10	5	5.000	5.000	0.995	0.100	0.198	0.100

Example 2. Consider the following data-sets, all with $n = 100$, $x_{\min} = 0$, and $x_{\max} = 100$:

Data-set A Fifty “0” and fifty “10”, i.e., $\{0 \times 50, 10 \times 50\}$.

Data-set B Seventy five “0” and twenty five “10’, i.e., $\{0 \times 75, 10 \times 25\}$.

Data-set C Ninety “0” and ten “10”, i.e., $\{0 \times 90, 10 \times 10\}$.

Data-set D Ninety nine “0” and one “10”, i.e., $\{0 \times 99, 10 \times 1\}$.

Data-set E Twenty five “0”, Fifty “5”, and twenty five “10”, i.e., $\{0 \times 25, 5 \times 50, 10 \times 25\}$.

Data-set F Ten “0”, eighty “5”, and ten “10”, i.e., $\{0 \times 10, 5 \times 80, 10 \times 10\}$.

Data-set G One “0”, ninety eight “5”, and one “10”, i.e., $\{0 \times 1, 5 \times 98, 10 \times 1\}$.

The actual MAD for each data-set is compared to the proposed upper bounds in Table 3. The upper bounds B_4 and B_7 are not present in Table 3 because they depend on s and s is not unique in general.

Example 3. Let $\{x_i\}$ be a confidential data-set which means that we are not aware of exactly what data does it contain. Only a limited amount of information is available to us such as $x_{\min} = 0$, $x_{\max} = 10$, and $n = 100$ in addition to only one of the following statements:

a. We know Nothing else.

b. We know that the mean of the data is 7.0.

c. We know that exactly 50 data are above average.

d. We know about 40 data instances that they are not above average.

e. We know that exactly 30 data are below average.

f. We know that exactly 50 data are above average and exactly 30 data are below average.

Let us find an appropriate upper bound for the MAD, i.e., $\rho \leqslant B$. Choosing the best upper bound for each case we have:

a. $B_1 = r_x/2 = 5.000$.

b. $B_3 = 2(x_{\max} - \bar{x})(\bar{x} - x_{\min})/r_x = 4.2$.

c. $B_8^u = 2u(n-u)r_x/n^2 = 5.000$.

d. $B_7 = 2s(n-s)r_x/n^2 = 4.800$.

e. $B_8^l = 2l(n-l)r_x/n^2 = 4.200$.

f. $B_9 = 2lur_x/n(l+u) = 3.750$.

Finally, note that the results of this paper can be applied to any data-set with finite number of elements and the probability distribution of the data is not needed to be known.

Acknowledgements

The author acknowledges Tafresh University for financial support.

References

[1] A. Ben-Tal, E. Hochman, *More bounds on the expectation of a convex function of a random variable*, Journal of Applied Probability, 9 (1972), 803-812.

[2] R. Bhatia, C. Davis, *A better bound on the variance*, American Mathematical Monthly, 107 (2000), 353-357

[3] P. Diaconis, S. Zabell, *Closed form summation for classical distributions: variations on theme of de moivre*, Statistical Science, 6 (1991), 284-302.

[4] A.A. Kamil, K. Ibrahim, *Mean-absolute deviation portfolio optimization problem*, Journal of Information and Optimization Sciences, 28 (2007), 935-944.

[5] H. Konno, A. Wijayanayake, *Mean-absolute deviation portfolio optimization model under transactions costs*, Journal of the Operations Research, 42 (1999), 422-435.

[6] R.M. Korwar, *On characterizations of distributions by mean absolute deviations and variance bounds*, Ann. Inst. Statist. Math., 43 (1991), 287-295.

[7] T. Pham-Gia, T.L. Hung, *The mean and median absolute deviations*, Mathematical and Computer Modelling, 34 (2001), 921-936.

[8] T. Popoviciu, *Sur les équations algébriques ayant toutes leurs racines réelles (in French)*, Mathematica (Cluj), 9 (1935), 129-145.

[9] K. Postek, A. Ben-Tal, D. den Hertog, B. Melenberg, *Robust optimization with ambiguous stochastic constraints under mean and dispersion information*, Operations Research, 66 (2018).

[10] G. Rehnman, N. Tesch, *Application of mean absolute deviation optimization in portfolio management,* Dissertation, KTH Royal Institute of Technology, Stockholm, Sweden, 2018.

Accepted: 18.10.2019