Abstract

The era of big data is coming, and meta-analysis is attracting increasing attention to analytically combine the results from several similar clinical trials to provide an overall estimation of a treatment effectiveness. The sample mean and standard deviation are two commonly used statistics in meta-analysis but some trials use the median, the minimum and maximum values, or sometimes the first and third quartiles to report the results. Thus, to pool results in a consistent format, researchers need to transform those information back to the sample mean and standard deviation. In this paper, we investigate the optimal estimation of the sample mean for meta-analysis from both theoretical and empirical perspectives. A major drawback in the literature is that the sample size, needless to say its importance, is either ignored or used in a stepwise but somewhat arbitrary manner, e.g., the famous method proposed by Hozo et al. (2005). We solve this issue by incorporating the sample size in a smoothly changing weight in the estimators to reach the optimal estimation. Our proposed estimators not only improve the existing ones significantly but also share the same virtue of the simplicity. The real data application indicates that our proposed estimators is capable to serve as “rules of thumb” and could be widely applied in meta-analysis.

Keywords: Median, Meta-analysis, Mid-range, Mid-quartile range, Optimal weight, Sample mean, Sample size

*Co-corresponding author. E-mail: xwan@comp.hkbu.edu.hk
†Co-corresponding author. E-mail: tongt@hkbu.edu.hk
# Introduction

Meta-analysis plays a crucial role in medical research for helping researchers summarize data comprehensively and make proper decisions to cure patients (Abuabara et al., 2012). In such an analysis, the researchers select several independent studies based on the same clinical question, and then use statistical techniques to combine the results together so that a reliable conclusion could be drawn eventually (Cook et al., 1997). To statistically combine data from multiple studies, the first step is to determine a summary measure as well as its corresponding statistics. In most of the studies, the sample mean and standard deviation are two commonly used statistics in the data analysis. However, instead of directly reporting the sample mean and standard deviation, the median, the first and third quartiles, the minimum and maximum values are often recorded in clinical trial studies. As a result, when proceeding meta-analysis, people need to transform the reported information such as the median and the minimum and maximum values to the sample mean in order to conduct further analysis.

To transform the data, Hozo et al. (2005) used inequalities to establish some estimators for the sample mean and variance. They were the first to suggest methodology for this estimation problem. Their proposed method is simple and has been widely adopted in the literature. In the article, they discussed how to estimate the sample mean and variance in the scenario where only the sample median, extremum and sample size are reported. Recently, Wan et al. (2014) pointed out that Hozo et al.’s method has some serious drawbacks and, in particular, is less accurate for the estimation of the sample variance. In view of this, they introduced a quartile method to improve the sample standard deviation estimation. They have further extended the new methodology to two other frequently encountered scenarios in reporting clinical trial results. Through simulation studies, they have demonstrated that their newly proposed methods greatly outperform the existing methods including Hozo et al. (2005) and Bland (2015).

Wan et al. (2014) had fully discussed and consummated the approaches in estimating the sample standard deviation under different conditions. For the estimation of sample mean, they simply followed the estimation methods in Hozo et al. (2005) and Bland (2015). These existing methods, however, suffer from some major limitations due to insufficient use of the information in the sample size. In this paper, we propose some new methods by incorporating the sample size in a smoothly changing weight in the proposed
estimators to reach the optimal estimation. The proposed methods not only improve the existing estimators significantly but also share the same virtue of the simplicity. Our proposed estimators are able to serve as “rules of thumb” for the sample mean estimation in meta-analysis.

2 Sample mean estimation

For the sake of consistency, we follow essentially the same notations as those in Wan et al. (2014). Specifically, we let \( n \) be the sample size and denote the 5-number summary for the data as

\[
a = \text{the minimum value}, \quad q_1 = \text{the first quartile}, \quad m = \text{the median}, \quad q_3 = \text{the third quartile}, \quad b = \text{the maximum value}.
\]

In clinical trial reports, the 5-number summary for the data may not be provided in full. We consider the three scenarios that are most frequently encountered:

\[
S_1 = \{a, m, b; n\}, \quad S_2 = \{q_1, m, q_3; n\}, \quad S_3 = \{a, q_1, m, q_3, b; n\},
\]

where \( S_1, S_2 \) and \( S_3 \) represent \( C_1, C_3 \) and \( C_2 \) in Wan et al. (2014), respectively. According to Triola (2009), we refer to \( b - a \) as the range, \( (a + b)/2 \) as the mid-range, \( q_3 - q_1 \) as the interquartile range, and \( (q_1 + q_3)/2 \) as the mid-quartile range. As a common practice, the range and the interquartile range are often used to measure the standard deviation, whereas the mid-range and the mid-quartile range are used to measure the center (or mean) of the population.
2.1 Existing methods

2.1.1 Hozo et al.’s method for $S_1 = \{a, m, b; n\}$

Scenario $S_1$ represents the situation where the median, the minimum, the maximum and the sample size are given in a study. Hozo et al. (2005) were the first to address the sample mean estimation problem. By inequalities, they proposed the following estimator for the sample mean:

$$\bar{X} \approx \begin{cases} \frac{(a + 2m + b)}{4} & n \leq 25, \\ m & n > 25. \end{cases} \quad (1)$$

Although very easy to implement, we note that the estimator (1) may not be sufficiently accurate as it incorporates the sample size in a stepwise manner. The sample mean estimation has a sudden change from $m$ to $(a + 2m + b)/4$ when the sample size reduces to 25. This change might lead to a less precise estimation. For example, when the sample size is changed from 26 to 25, the estimated sample mean might be a lot different than the actual one because of the “jump” in the estimator. In contrast, within the respective interval of $n > 25$ or $n \leq 25$, the sample mean estimation is independent of the sample size and the information in the sample size is completely ignored. As a consequence, such an estimate may not be reliable for practical use. This motivates us to consider an improved estimation of the sample mean by incorporating the sample size in a smoothly changing manner.

2.1.2 Wan et al.’s method for $S_2 = \{q_1, m, q_3; n\}$

Scenario $S_2$ reports the first and third quartiles instead of the minimum and the maximum, together with the median and the sample size. Other than the sample range, i.e. the difference between the minimum and the maximum, the interquartile range is usually less sensitive to outliers and hence is also popularly reported in clinical trial studies. For scenario $S_2$, Wan et al. (2014) proposed to estimate the sample mean by

$$\bar{X} \approx \frac{q_1 + m + q_3}{3}. \quad (2)$$

It is evident that the sample size information is not used in their proposed estimation. We also note that an equal weight is assigned to each summary statistic in the estimator (2). In particular, the weight for the median is $1/3$ in scenario $S_2$ compared to $1/2$ in scenario $S_1$. Hence, it would also be of interest to investigate if a smoothly changing manner is
needed for assigning the appropriate weights to the median and the two quartiles with respect to the sample size.

2.1.3 Bland’s method for $S_3 = \{a, q_1, m, q_3, b; n\}$

Scenario $S_3$ is a combination of the scenarios $S_1$ and $S_2$. It assumes that the 5-number summary of the data are all given for further analysis. Following the same idea in Hozo et al. (2005), Bland (2015) proposed the following estimator for the sample mean:

$$
\bar{X} \approx \frac{a + 2q_1 + 2m + 2q_3 + b}{8}.
$$

(3)

Once again, the sample size information is not used in the estimation of the sample mean. The estimator assigns an equal weight to $q_1$, $m$ and $q_3$, respectively, and another equal weight to $a$ and $b$, respectively. Similar to the other two scenarios, we will investigate if a smoothly changing manner is needed for assigning the appropriate weights to the 5-number summary of the data with respect to the sample size.

2.2 Improved methods

Let $X_1, X_2, \ldots, X_n$ be a random sample of size $n$ from the normal distribution $N(\mu, \sigma^2)$, and $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ be the ordered statistics of the sample. Also for simplicity, we assume that the sample size $n = 4Q + 1$ with $Q \geq 1$ being a positive integer. With the above notations, we have $a = X_{(1)}$, $q_1 = X_{(Q+1)}$, $m = X_{(2Q+1)}$, $q_3 = X_{(3Q+1)}$, and $b = X_{(n)} = X_{(4Q+1)}$. In addition, we let $Z_1, Z_2, \ldots, Z_n$ be independent and identically distributed (i.i.d.) random variables from the standard normal distribution $N(0, 1)$, and $Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(n)}$ be the ordered statistics of $\{Z_1, \ldots, Z_n\}$. Then accordingly, we have $X_i = \mu + \sigma Z_i$ and $X_{(i)} = \mu + \sigma Z_{(i)}$ for $i = 1, \ldots, n$.

2.2.1 Improved estimation of the sample mean in $S_1 = \{a, m, b; n\}$

Following the discussion in Section 2.1.1, we propose to estimate the sample mean by

$$
\bar{X}(w) = w \left( \frac{a + b}{2} \right) + (1 - w)m,
$$

(4)

where $w$ is the weight assigned to the mid-range $(a+b)/2$, and the remaining weight $1-w$ is assigned to the median $m$. It is evident that the proposed estimator is a weighted average of the mid-range and the median, where both quantities are the measures of
center for the population. In the special case if we take $w = \frac{1}{2}$ for $n \leq 25$ and $w = 0$ for $n > 25$, the proposed estimator reduces to the estimator (1) proposed by Hozo et al. (2005). Such an allocation of the weight is somewhat arbitrary and can be less reliable.

We consider to solve the issue by incorporating the sample size in a smoothly changing manner. That is, we consider the weight $w = w(n)$ as a function of the sample size. Then from the decision-making point of view, we define the optimal weight $w_{opt} = w_{opt}(n)$ to be the weight that minimizes the expected loss function of the estimator. In this paper, we consider the squared loss function $L(\bar{X}(w), \mu) = (\bar{X}(w) - \mu)^2$, then accordingly, the expected loss function is the commonly used mean squared error (MSE) of the estimator.

By Lemma 3 in Appendix A.3, it is shown that the proposed estimator (4) is an unbiased estimator of the true mean $\mu$. Hence, we have $\text{MSE}(\bar{X}(w)) = (w^2/4)\text{Var}(a + b) + (1 - w)^2\text{Var}(m) + w(1 - w)\text{Cov}(a + b, m)$. Note that the MSE function is a quadratic function of $w$ and has a unique minimum value on $[0, 1]$.

To derive the optimal weight, we take the first derivative of MSE with respect to $w$ and set the result equal to zero. It gives the the optimal weight as

$$w_{opt}(n) = \frac{4\text{Var}(m) - 2\text{Cov}(a + b, m)}{\text{Var}(a + b) + 4\text{Var}(m) - 4\text{Cov}(a + b, m)}.$$ 

Recall that $a = \mu + \sigma Z_{(1)}$, $b = \mu + \sigma Z_{(n)}$ and $m = \mu + \sigma Z_{(2Q+1)}$. Together with the symmetry of the standard normal distribution, we can represent the optimal weight as

$$w_{opt}(n) = \frac{K(n)}{K(n) + 1},$$

where $K(n) = 2[E(Z_{(2Q+1)}^2) - E(Z_{(1)}Z_{(2Q+1)})] / [E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) - 2E(Z_{(1)}Z_{(2Q+1)})]$. The derivation of (5) is given as Proof 5 in Appendix A.3. It is clear that the optimal weight $w_{opt}(n)$ is independent of $\mu$ and $\sigma^2$ and is only a function of $n$.

To explore the behavior of the optimal weight, we use the statistical software $R$ to numerically compute the values of $w_{opt}(n)$ for $n$ from 5 to 101 and plot them in Figure 1. We observe that $w_{opt}(n)$ decreases as $n$ increases, in particular, the optimal weight reduces to about 0.1 when $n = 101$. When the sample size is large or very large, the estimator will assign most of the weight to the median as it provides a more robust estimate for the measure of center compared to the mid-range. In fact, as mentioned in Triola (2009), the mid-range is rarely used in practice as, from an asymptotic point of view, it lacks efficiency and robustness as an estimator. When the sample size is small, however, a well designed combination of the mid-range and the median may provide a better estimation.
of the sample mean compared to only using the median. In addition, we note that the optimal weight \( w_{\text{opt}}(n) \) is about 0.25 when \( n = 25 \). This explains why in a stepwise manner with \( w = 0 \) and \( w = 0.5 \) being the only two options, \( \text{Hozo et al. (2005)} \) suggested to take \( w = 0.5 \) when \( n \leq 25 \) and \( w = 0 \) when \( n > 25 \).

Note that the optimal weight \( w_{\text{opt}}(n) \) in (5) may not be readily accessible for practitioners as it involves some complicated statistical computation. In what follows, we propose an approximation formula for \( w_{\text{opt}}(n) \) and then display the final estimator of the sample mean as an “rule of thumb” for practical use. To approximate \( K(n) \), we consider the simple power function \( K(n) = an^b \). Using the observed true weights in Figure 1, we figure out that the best coefficients are about \( a = 4 \) and \( b = -0.75 \). This leads to the approximated optimal weight as

\[
\tilde{w}_{\text{opt}}(n) \approx \frac{4}{4 + n^{0.75}}.
\] (6)

For comparison, we also display the approximated optimal weights (6) and the weights proposed by Hozo et al. in Figure 1. It is evident that the approximated optimal weights provide a nearly perfect match to the true optimal weights, in particular for the sample size ranging from 5 to 101.

Finally, by plugging the approximation formula (6) into the estimator (4), we propose the “rule of thumb” for Scenario \( S_1 \) as

\[
\bar{X} \approx \left( \frac{4}{4 + n^{0.75}} \right) \frac{a + b}{2} + \left( \frac{n^{0.75}}{4 + n^{0.75}} \right) m.
\] (7)

The performance of (7) is evaluated in Section 3.1, together with its numerical comparison with the estimation method in \( \text{Hozo et al. (2005)} \).

2.2.2 Improved estimation of the sample mean in \( S_2 = \{q_1, m, q_3; n\} \)

For scenario \( S_2 \), we propose the new estimator for the sample mean as

\[
\bar{X}(w) = w \left( \frac{q_1 + q_3}{2} \right) + (1 - w)m,
\] (8)

where \( w \) and \( 1 - w \) are the weights assigned to the mid-quartile range \((q_1 + q_3)/2\) and the median \( m \). The new estimator is a weighted average of the mid-quartile range and the median. It is worth mentioning that the mid-quartile range is also a measure of the population center, which is the numerical value midway between the first and third
Figure 1: The true optimal weights (the green empty circles, simulated using the statistical software R), the approximated optimal weights (the red line, computed using Eq. (6)), and the weights in Hozo et al. (the black line, computed using Eq. (1)) for scenario $S_1$. 

Scenario 1
quartiles (Johnson and Kuby, 2007). In addition, by taking \( w = 2/3 \) for all \( n \), the new estimator reduces to the estimator (2) proposed by Wan et al. (2014).

Following the similar arguments as in Section 2.2.1, the optimal weight for the mid-quartile range that minimizes the MSE function is obtained as

\[
w_{\text{opt}}(n) = \frac{4 \text{Var}(m) - 2 \text{Cov}(q_1 + q_3, m)}{\text{Var}(q_1 + q_3) + 4 \text{Var}(m) - 4 \text{Cov}(q_1 + q_3, m)},
\]

where \( q_1 = \mu + \sigma Z_{(Q+1)} \), \( q_3 = \mu + \sigma Z_{(3Q+1)} \). The detailed derivation of (9) is provided as Proof 9 in Appendix B.1. The numerical values of \( w_{\text{opt}}(n) \) for \( n \) from 5 to 101 are displayed in Figure 2. It is evident that \( w_{\text{opt}}(n) \) is again a decreasing function of \( n \). We further observe that the optimal weight reduces to around 0.7 but not 0 as \( n \) increases. That is, when the sample size is large, we will assign the mid-quartile range a weight about 0.7, or equivalently, assign a weight of 0.35 to the first and third quartiles respectively.

In statistics, the first and third quartiles are also robust summary statistics and they are treated to be equally important as the median in practical research.

Noting that the optimal weight \( w_{\text{opt}}(n) \) in (9) is rather complicated for practitioners, we now propose an approximation formula for \( w_{\text{opt}}(n) \). In view of Figure 2, we take the baseline to be 0.7 and then approximate the remaining part to be a power function. That is, we consider the approximation form as \( 0.7 + an^b \). Next, we use the observed true weights and figure out that the best \( a \) and \( b \) values are given as \( a = 0.4 \) and \( b = -1 \). This leads to the approximated optimal weight as

\[
\tilde{w}_{\text{opt}}(n) \approx 0.7 + \frac{0.4}{n}.
\]

To assess the accuracy of the approximation, we also display the values of \( \tilde{w}_{\text{opt}}(n) \) in Figure 2. It is evident that the approximated optimal weights fit well the true optimal weights, in particular for the sample size ranging from 5 to 101.

Finally, by plugging the approximation formula (10) into the estimator (8), we propose the “rule of thumb” for Scenario \( S_2 \) as

\[
\bar{X} \approx \left( 0.7 + \frac{0.4}{n} \right) \frac{q_1 + q_3}{2} + \left( 0.3 - \frac{0.4}{n} \right) m.
\]

The performance of (11) is evaluated in Section 3.2, together with its numerical comparison with the estimation method in Wan et al. (2014).
Figure 2: The true optimal weights (the green empty circles, simulated using the statistical software R) and the approximated optimal weights (the red line, computed using Eq. [10]) for scenario $S_2$. 
2.2.3 Improved estimation of the sample mean in \( S_3 = \{a, q_1, m, q_3, b; n\} \)

For scenario \( S_3 \), following the same spirit, we propose to estimate the sample mean by

\[
\bar{X} = w_1 \left( \frac{a + b}{2} \right) + w_2 \left( \frac{q_1 + q_3}{2} \right) + (1 - w_1 - w_2)m,
\]

where \( w_1, w_2 \) and \((1 - w_1 - w_2)\) are the weights assigned to the mid-range \((a + b)/2\), the mid-quartile range \((q_1 + q_3)/2\) and the median \(m\), respectively. Taking \( w_1 = 0.25 \) and \( w_2 = 0.5 \), the proposed estimator reduces to the Bland estimator in (3). In Appendix C, we show that (12) is an unbiased estimator of \( \mu \). We further derive, by minimizing the derived MSE(\( \bar{X} \)) in Lemma 12, the optimal weights of \( w_1 \) and \( w_2 \) in (17) and (18), respectively.

To explore the behavior of the optimal weights, we plot the true values of \( w_{1,\text{opt}}(n) \) and \( w_{2,\text{opt}}(n) \) in Figure 3. From the figure, we note that \( w_{1,\text{opt}}(n) \) is a decreasing function of \( n \), and \( w_{2,\text{opt}}(n) \) is an increasing function of \( n \). In particular, when the sample size is large, \( w_{1,\text{opt}}(n) \) reduces to 0 and \( w_{2,\text{opt}}(n) \) increases to about 0.7. This shows that scenario \( S_3 \) converges to scenario \( S_2 \) as \( n \) tends to infinity. Noting also that (17) and (18) are rather complicated for practical use, we now suggest approximation formulas for both the true optimal weights. Following the structures in (6) and (10), we propose to estimate the two optimal weights by \( a/(a + n^b) \) and \( 0.7 - cn^d \), respectively. Then by the observed true weights, the best values of the coefficients are \( a = 2.05, b = 0.72, c = 1 \) and \( d = 0.65 \). This leads to the approximated optimal weights as

\[
\tilde{w}_{1,\text{opt}}(n) \approx \frac{2.05}{2.05 + n^{0.72}} \quad \text{and} \quad \tilde{w}_{2,\text{opt}}(n) \approx 0.7 - \frac{1}{n^{0.65}}. \tag{13}
\]

To assess the accuracy of the approximation, we also display the values of \( \tilde{w}_{1,\text{opt}}(n) \) and \( \tilde{w}_{2,\text{opt}}(n) \) in Figure 3. It is evident that the approximated optimal weights match precisely their respective values of the true optimal weights.

Finally, by plugging the approximation formula (13) into the estimator (12), we propose the “rule of thumb” for scenario \( S_3 \) as

\[
\bar{X} \approx \left( \frac{2.05}{2.05 + n^{0.72}} \right) a + \frac{b}{2} \left( 0.7 - \frac{1}{n^{0.65}} \right) \frac{q_1 + q_3}{2} + \left( 0.3 + \frac{1}{n^{0.65}} - \frac{2.05}{2.05 + n^{0.72}} \right) m. \tag{14}
\]

The performance of (14) is evaluated in Section 3.3, together with its numerical comparison with the estimation method in Bland (2015).
Figure 3: The true optimal weights $w_1$ and $w_2$ (the green circles and the red circles, respectively), and the approximated optimal weights $\tilde{w}_{1,\text{opt}}(n)$ and $\tilde{w}_{2,\text{opt}}(n)$ (the orange line and the blue line, respectively).
3 Simulation studies

To compare the performance between existing methods and our newly proposed methods, we conduct some simulation studies. Using the same settings as in Hozo et al. (2005), five different distributions are taken into consideration: the normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 17$, the log-normal distribution with location parameter $\mu = 4$ and scale parameter $\sigma = 0.3$, the beta distribution with shape parameters $\alpha = 9$ and $\beta = 4$, the exponential distribution with rate parameter $\lambda = 10$, and the Weibull distribution with shape parameter $k = 2$ and scale parameter $\lambda = 35$. The simulation results will be displayed in Figures 4 to 8 for the three scenarios, respectively.

For each simulation, we first generate a random sample with $n$ observations and calculate the actual sample mean using the whole sample. Next, we use the information of the sample median, minimum and maximum values and/or interquartiles to estimate the sample mean by the existing methods and our newly proposed methods, respectively. The relative error of each method is defined as below to evaluate the accuracy of the two estimates:

$$\text{Relative error of } \bar{X} = \frac{\text{the estimated } \bar{X} - \text{the true value of } \bar{X}}{\text{the true value of } X}. \quad (15)$$

The number of repeats for each distribution is 10,000 and the mean relative error will be reported for each method.

3.1 Simulation study for $\mathcal{S}_1 = \{a, m, b; n\}$

In this simulation study, we compare Hozo et al.’s estimator in Eq. (1) and our proposed estimator in Eq. (4). Figure 4 reports the average relative errors of 10,000 simulations for normal distribution with the sample size ranging from 5 to 200. It is worth to mention that for normal distribution, we compare the absolute average error between the new method and Hozo et al.’s method for better demonstrating the results. As it is clearly shown in the figure, the new estimator always has a smaller error than Hozo et al.’s estimator. Although these two methods seems to have similar performance when sample size becomes larger than 200, it is still obvious that the new estimator is a better choice for practical research because unlike Hozo et al.’s method, the weights on the mid-range and the median in the new method have a smoother change, which makes the new estimator more adaptive and more stable. The stability of the new method can be further shown in the simulation
studies using different distributions, specifically, skewed distributions.

Figure 4: Average relative error of the sample mean estimation for data from normal distribution. The pink lines with solid circles represent Hozo et al.’s method, and the light blue lines with empty circles represent the new estimator.

Figure 5 provides the simulation results of the other four distributions as we mentioned before. From Figure 5, we can easily observe that for each of four skewed distributions, Hozo et al.’s method has a huge breakpoint at sample size of 25 (based on the Eq. (1)). Moreover, Hozo et al.’s method has a fluctuated relative error for different distributions. Especially for exponential distribution, Hozo et al.’s estimator displayed a relative error as high as 30% of the true sample mean. On the other hand, the new estimator still have a good performance in estimating sample means even for skewed distribution. As
Figure 5: Average relative error of the sample mean estimation for data from non-normal distributions. The pink lines with solid circles represent Hozo et al.’s method, and the light blue lines with empty circles represent the new estimator.
it is shown in the figure, the average error for each distribution produced by the newly proposed method is close to zero for small sample size and always within 5% of the true sample mean. Although the relative error from the new method seems to have an increasing trend in Figure 3, it is worth to mention that the relative error tends to converge and no longer increases as the sample size increases. Hence, we think that the relative error of the new estimator is acceptable and the new method is reliable for application. To sum up, the new estimator has a stable performance for both skewed and non-skewed data with acceptable relative errors in scenario $S_1$.

3.2 Simulation results for $S_2 = \{q_1, m, q_3; n\}$

In this simulation study, we compare Wan et al.’s estimator in Eq. (2) and our proposed method in Eq. (8). Figure 6 compares the average relative errors of Wan et al.’s method and the proposed estimator for normal distribution. In general, these two estimators perform equally well. Figure 7 shows the estimation results from the two estimators for the other four distributions. It is easily to observe that for skewed distributions, when average errors are negative, all the light blue points (from the new estimator) lay above the pink solid points (from Wan et al.’s estimator), and when average errors are positive, all the light blue points lay below the pink solid points. Based on these results, we conclude that the new estimator has lower error and it is capable to provide a more accurate estimation for skewed distribution for scenario $S_2$. In conclusion, the proposed estimator has a better performance for skewed data with acceptable relative errors in scenario $S_2$.

3.3 Simulation study for $S_3 = \{a, q_1, m, q_3, b; n\}$

In this simulation study, we compare the performance of Bland’s estimator in Eq. (3) and our proposed estimator in Eq. (12). The simulation results are shown in Figures 8 and 9. Figure 8 shows that the relative error generated from our proposed method is much smaller than Bland’s method for normal distribution. Bland’s method seems to return an average error close to 1.8%, while our new method basically has an error less than 1%. Although the error from our proposed method had an increasing trend when sample size is less than 15, it reduced to zero when sample size become larger. That is, the newly proposed estimator is able to provide a more accurate estimation of the sample mean in comparison to to Bland’s method. Figure 9 shows the performance of those two methods...
Figure 6: Average relative error of the sample mean estimation for data from normal distribution. The pink lines with solid circles represent Wan et al.’s method, and the light blue lines with empty circles represent the new estimator.
Figure 7: Average relative error of the sample mean estimation for data from non-normal distributions. The pink lines with solid circles represent Wan et al.’s method, and the light blue lines with empty circles represent the new estimator.
on the other four skewed distributions. It is obvious that the error generated from our method is much closer to zero than Bland’s approach. Therefore, the new estimator is capable to provide reliable estimations of the sample mean for both symmetric and skewed distributions in scenario $S_3$.

![Normal Distribution](image)

Figure 8: Average relative error of the sample mean estimation for data from normal distribution. The pink lines with solid circles represent Bland’s method, and the light blue lines with empty circles represent the new estimator.
Figure 9: Average relative error of the sample mean estimation for data from non-normal distributions. The pink lines with solid circles represent Bland’s method, and the light blue lines with empty circles represent the new estimator. The horizontal line is $error = 0$. 
4 Real Data Analysis

To illustrate the potential value of our method in real data analysis, we collect some real data and compare the estimations using our methods with the ones using the existing method. The collected data is from a systemic review and meta-analysis of the association between low serum vitamin D and risk of active tuberculosis in humans (Nnoaham and Clarke, 2008).

For the illustrative purpose, we test the performance of our methods in reality, we decided to use real data to estimate the sample mean. In this section, we only illustrate the analysis results of two data set. One of the chosen data is from a systematic review of and, the other is from Aloor (2015). In Nnoaham and Clarke (2008), there are three studies reported the sample median, minimum and maximum values. In Aloor (2015), three studies reported the median, the first and the third quartiles and three studies reported the median, minimum and maximum values. Hence, we are able to use our proposed methods for scenarios $S_1$ and $S_2$ to estimate the sample mean and compare the results with Hozo et al.’s method and Wan et al.’s method, respectively.

4.1 Data description

In Nnoaham and Clarke (2008), the summary statistics reported from seven studies were used to conduct the meta-analysis. Among those seven studies, three of them only reported the sample median and range, which is exactly the case of Scenario $S_1$. For these three studies, the sample mean and standard deviation need to be estimated from the sample median and range in order to calculate the pooled effect size. The summary statistics are presented in Table II in which Studies 1 to 3 reported the median, minimum and maximum values, Study 4 and 5 reported the mean values and standard deviations, Study 6 reported the odds ratio for vitamin D deficiency in tuberculosis cases compared to controls, and Study 7 reported the mean value and range (i.e., difference between the minimum and maximum).

4.2 Results and comparison

To conduct a random effect meta-analysis, Nnoaham and Clarke first used Hozo et al.’s method to estimate the sample mean and standard deviation for the first three studies
Table 1: Summary of included studies

<table>
<thead>
<tr>
<th>Index</th>
<th>Study</th>
<th>Size (Cases)</th>
<th>Size (Controls)</th>
<th>Results (serum Vitamin D levels)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Davies et al. (1985)</td>
<td>40</td>
<td>40</td>
<td>Median (range) in: Cases-16.0 nmol/L (2.25-74.25 nmol/L), Controls-27.25 nmol/L (9.0-132.5 nmol/L)</td>
</tr>
<tr>
<td>2</td>
<td>Grange et al. (1985)</td>
<td>40</td>
<td>38</td>
<td>Median (range) in: Cases-65.75 nmol/L (43.75-130.5 nmol/L), Controls-69.5 nmol/L (48.5-125 nmol/L)</td>
</tr>
<tr>
<td>3</td>
<td>Davies et al. (1987)</td>
<td>15</td>
<td>15</td>
<td>Median (range) in: Cases-39.75 nmol/L (16.75-89.25 nmol/L), Controls-65.5 nmol/L (26.25-114.75 nmol/L)</td>
</tr>
<tr>
<td>4</td>
<td>Davies et al. (1988)</td>
<td>51</td>
<td>51</td>
<td>Mean (SD) in: Cases-69.5 nmol/L (24.5 nmol/L), Controls-95.5 nmol/L (29.25 nmol/L)</td>
</tr>
<tr>
<td>5</td>
<td>Chan et al. (1994)</td>
<td>24</td>
<td>24</td>
<td>Mean (SD) in: Cases-46.5 nmol/L (18.5 nmol/L), Controls-52.25 nmol/L (15.75 nmol/L)</td>
</tr>
<tr>
<td>6</td>
<td>Wilkinson et al. (2000)</td>
<td>103</td>
<td>42</td>
<td>Odds ratio (CI) of cases compared to controls-2.9 (1.3-6.5)</td>
</tr>
<tr>
<td>7</td>
<td>Sasidharan et al. (2002)</td>
<td>35</td>
<td>16</td>
<td>Mean (range) in: Cases-26.75 nmol/L (2.5-75 nmol/L), Controls-48.5 nmol/L (22.5-145 nmol/L)</td>
</tr>
</tbody>
</table>

and the sample standard deviation for Study 7 in Table 1. Next, the mean difference (i.e. the Cohen’s $d$ value \cite{Cohen2013}) is computed as the effect size. The odds ratio in Study 6 is directly converted to the effect size by Chinn \cite{Chinn2000}. Finally, the pooled effect size is computed and the heterogeneity between studies will be assessed using the $\chi^2$ statistic and the $I^2$ index (the amount of variation due to heterogeneity). Their results are presented in Table 2. It is worth to mention that in Nnoaham and Clarke \cite{Nnoaham2008}, they mistakenly reported the estimated effect size of Study 2.

Following the same aforementioned procedure, we use Eq. (7) to estimate the sample mean for the first studies and the method of Wan et al. \cite{Wan2014} to estimate the sample deviation for the first three studies and Study 7. Then we further use the same method as Nnoaham and Clarke did to compute the pooled effect size, the $\chi^2$ statistic, and the $I^2$ index with our estimations. The new results are reported in Table 3.

Comparing Table 2 and Table 3 we can observe some significant differences between the old results and the new results. The most noticeable one is the effect size difference for Study 1. The effect size using Hozo et al.’s method is within the range of large effect level while it only reaches the median effect level using our method (i.e., 0.8656 vs. 0.6622). The effect size different for Study 2, 3 and 7 is also non-trivial. Although the pooled effect
Table 2: Effect sizes of low serum vitamin D in tuberculosis (using Hozo et al.’s method)

<table>
<thead>
<tr>
<th>Index</th>
<th>Study</th>
<th>Size (Cases)</th>
<th>Size (Control)</th>
<th>Effect Size (SE)</th>
<th>Weight %</th>
<th>95% CI of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Davies et al. (1985)</td>
<td>40</td>
<td>40</td>
<td>0.8656 (0.0562)</td>
<td>17.79</td>
<td>[0.4043, 1.3218]</td>
</tr>
<tr>
<td>2</td>
<td>Grange et al. (1985)</td>
<td>40</td>
<td>38</td>
<td>0.0824 (0.0527)</td>
<td>18.97</td>
<td>[-0.3621, 0.5263]</td>
</tr>
<tr>
<td>3</td>
<td>Davies et al. (1987)</td>
<td>15</td>
<td>15</td>
<td>0.9190 (0.1590)</td>
<td>6.29</td>
<td>[0.1569, 1.6664]</td>
</tr>
<tr>
<td>4</td>
<td>Davies et al. (1988)</td>
<td>51</td>
<td>51</td>
<td>0.9637 (0.0447)</td>
<td>22.35</td>
<td>[0.5511, 1.3719]</td>
</tr>
<tr>
<td>5</td>
<td>Chan et al. (1994)</td>
<td>22</td>
<td>23</td>
<td>0.3353 (0.0944)</td>
<td>10.59</td>
<td>[-0.2554, 0.9221]</td>
</tr>
<tr>
<td>6</td>
<td>Wilkinson et al. (2000)</td>
<td>103</td>
<td>42</td>
<td>0.5882 (0.0352)</td>
<td>28.40</td>
<td>[0.2220, 0.9525]</td>
</tr>
<tr>
<td>7</td>
<td>Sasidharan et al. (2002)</td>
<td>35</td>
<td>16</td>
<td>0.9584 (0.1045)</td>
<td>9.57</td>
<td>[0.3329, 1.5749]</td>
</tr>
</tbody>
</table>

Total (95% CI) 306 225 0.6732 100.00 [0.4961, 0.8498]

Test for heterogeneity: $\chi^2 = 11.6594, df = 6 (P = 0.07), I^2 = 48.539\%$

Table 3: Effect sizes of low serum vitamin D in tuberculosis (using the new method)

<table>
<thead>
<tr>
<th>Index</th>
<th>Study</th>
<th>Size (Cases)</th>
<th>Size (Control)</th>
<th>Effect Size (SE)</th>
<th>Weight %</th>
<th>95% CI of Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Davies et al. (1985)</td>
<td>40</td>
<td>40</td>
<td>0.6622 (0.0542)</td>
<td>18.46</td>
<td>[0.2098, 1.1105]</td>
</tr>
<tr>
<td>2</td>
<td>Grange et al. (1985)</td>
<td>40</td>
<td>38</td>
<td>0.1588 (0.0528)</td>
<td>18.93</td>
<td>[-0.2864, 0.6030]</td>
</tr>
<tr>
<td>3</td>
<td>Davies et al. (1987)</td>
<td>15</td>
<td>15</td>
<td>0.9852 (0.1614)</td>
<td>6.19</td>
<td>[0.2171, 1.7378]</td>
</tr>
<tr>
<td>4</td>
<td>Davies et al. (1988)</td>
<td>51</td>
<td>51</td>
<td>0.9637 (0.0447)</td>
<td>22.35</td>
<td>[0.5511, 1.3719]</td>
</tr>
<tr>
<td>5</td>
<td>Chan et al. (1994)</td>
<td>22</td>
<td>23</td>
<td>0.3353 (0.0944)</td>
<td>10.59</td>
<td>[-0.2554, 0.9221]</td>
</tr>
<tr>
<td>6</td>
<td>Wilkinson et al. (2000)</td>
<td>103</td>
<td>42</td>
<td>0.5882 (0.0352)</td>
<td>28.40</td>
<td>[0.2220, 0.9525]</td>
</tr>
<tr>
<td>7</td>
<td>Sasidharan et al. (2002)</td>
<td>35</td>
<td>16</td>
<td>0.9084 (0.1036)</td>
<td>9.66</td>
<td>[0.2861, 1.5223]</td>
</tr>
</tbody>
</table>

Total (95% CI) 306 225 0.6574 100.00 [0.4961, 0.8498]

Test for heterogeneity: $\chi^2 = 9.2091, df = 6 (P = 0.162), I^2 = 34.847\%$
size from both methods seem to be close to each other, we found that the \( I^2 \) indices for heterogeneity between studies are quite different (i.e., 48.54\% from Hozo et al.’s method and 34.85\% from our method). According to Higgins et al. (2003), the value 48.54\% of \( I^2 \) is very close to moderate heterogeneity level while the value 34.85\% is close to little heterogeneity level. It is obvious that using our method in this study may eventually lead to a different conclusion.

To sum up, in Section 2.1.1, we have pointed out that Hozo et al.’s estimators may not be reliable in practical use and shown that our method can give rise to better estimations. After conducting the real data meta-analysis mentioned above, we are confident that the newly proposed methods could help researchers to make a more reliable conclusion.

5 Conclusion

Meta-analysis is a popular way to provide an overall estimation of a treatment effectiveness from a set of similar clinical trials. The sample mean and the standard deviation are often used in meta-analysis but sometimes the results are recorded using the median, the minimum and maximum values, or maybe the first and third quartiles. Searching for a reliable approximation method to obtain the sample mean and deviation and then conduct further research has been becoming a popular topic. The estimation of the sample deviation has been thoroughly discussed and significantly improved in Wan et al. (2014). But the current estimation of the sample mean adopts either the famous method proposed by Hozo et al. (2005) or the extension based on Hozo et.al’s method. Its obvious limitation is that the information of the sample size is not fully used or even ignored in the sample mean estimation. Therefore, in this work, we focus on improving the existing methods in the sample mean estimation from both theoretical and empirical perspectives.

Among three frequently encountered scenarios, the simulation studies show that our newly proposed methods, which incorporates the sample size via a smoothly changing weight in the estimation, greatly improve the existing method. For all scenarios, we provide both theoretical and empirical computations for optimal weights. The simulation results show that the empirical computation of optimal weight not only matches the theoretical computation with high accuracy but has almost the same simplicity as the existing methods. Here we provide a summary table of the new estimators of the sample mean in different scenarios, which may serve as a comprehensive guidance for researchers.
when performing meta-analysis. To help the researchers to utilize the proposed mean estimators, an Excel spread sheet containing all estimators in Table 4 is provided as the additional file. Using the spread sheet, users could easily obtain the sample mean values by inputting the corresponding information for appropriate scenario such as the sample size, median and extremum values. We also provide the formulas for Hozo et al.’s, Bland’s and Wan et al.’s methods in the Excel spread sheet for comparison purpose.

Table 4: **Summary table for estimating $\bar{X}$ under different scenarios**

<table>
<thead>
<tr>
<th>Scenario $S_1$</th>
<th>Scenario $S_2$</th>
<th>Scenario $S_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hozo et al. (2005)</td>
<td>Eq. (1)</td>
<td>–</td>
</tr>
<tr>
<td>Wan et al. (2014)</td>
<td>–</td>
<td>Eq. (2)</td>
</tr>
<tr>
<td>Bland (2015)</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>New methods</td>
<td>Eq. (7)</td>
<td>Eq. (11)</td>
</tr>
</tbody>
</table>

To further illustrate the performance of the newly proposed methods, a real data meta-analysis were conducted using seven studies from a systemic review and meta-analysis of the association between low serum vitamin D and risk of active tuberculosis in humans (Nnoaham and Clarke, 2008). We compared the effect sizes obtained from our proposed methods with those from the existing methods. It is evident that there are some significant differences between the new results and the old ones. Since the simulation studies indicate that the new methods could improve the estimation performance, we believe they could help the researchers to make a more convincing conclusion when conducting meta-analysis in real-world settings.

In conclusion, the new methods with respect to the three commonly used scenarios have better performance in the sample mean estimation in comparison to the existing literature. If some other scenarios with additional information appear in the future, we believe our new method could also be further extended. Additionally, since the estimators we established in this paper were based on the normality assumption, some might question the reliability of the estimator, as in meta-analysis, the median and the first and third quartiles were mainly reported because the data do not follow a normal distribution. In this case, in the future research, it is reasonable to consider developing a pre-test for the hypothesis that whether the underlying distribution is symmetric. If the underlying
distribution of the data is very asymmetric, one might assume that the information they contain could have large bias. Hence, the pre-test could help researchers to decide whether or not we should include the asymmetric data in the meta-analysis.

References


A Scenario $S_1$

Recall that $X_i = \mu + \sigma Z_i$, $i = 1, 2, \ldots, n$ and $n = 4Q + 1$, for the mean estimator:

$$\bar{X} = w\frac{a + b}{2} + (1 - w)m.$$  

where $w$ is the weight assigned to the mid-range $(a+b)/2$, and the remaining weight $1-w$ is assigned to the median $m$.

A.1 Introduction of standard normal distribution

Normal distribution, also written as $N(\mu, \sigma^2)$, is commonly used in statistics, which has a symmetric bell-shaped curve. The probability density function (PDF) is defined as:

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}.$$  

where $\mu$ is the mean value and $\sigma^2$ is the variance. When $\mu = 0$ and $\sigma^2 = 1$, we have the standard normal distribution, denoted by $N(0, 1)$.

For the standard normal distribution, notation $\phi(\cdot)$ is referred to its PDF and $\Phi(\cdot)$ is referred to its cumulative density function (CDF). In addition, as we assume the data follows normal distribution, it is essential to state the symmetric properties of standard normal distribution.

Suppose random variable $z \sim N(0, 1)$, then the following properties state:

$$\phi(z) = \phi(-z),$$
$$\Phi(z) = 1 - \Phi(-z),$$

Median = $\mu$ = Mode.  \hfill (16)

The above properties will be applied in the following proofs and derivations.

A.2 Basic definition

Definition 1. Suppose $Z_i = z_i$, $i = 1, \ldots, n$ and in specific, $-\infty < z_r < z_s < \infty$, then the expected values in the above equation defined by David and Nagaraja (1970) can be expressed in the following terms: (note: $n = 4Q + 1$)

$$E(Z_{(r)}) = \frac{n!}{(r-1)!(n-r)!} \int_{-\infty}^{\infty} z_r^k [\Phi(z_r)]^{r-1} [1 - \Phi(z_r)]^{n-r} \phi(z_r) dz_r,$$
$$E(Z_r Z_s) = C_{rs} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} z_r z_s [\Phi(z_r)]^{r-1} [\Phi(z_s) - \Phi(z_r)]^{s-r-1} [1 - \Phi(z_s)]^{n-s} dz_r dz_s,$$

where $C_{rs} = n! / [(r-1)!(s-r-1)!(n-s)!]$.
In this paper, we have $-\infty < z_1 \leq z_{Q+1} \leq z_{2Q+1} \leq z_{3Q+1} \leq z_n < \infty$.

### A.3 Important lemmas

**Lemma 1.** The expected values $E(Z_{(1)}) = -E(Z_{(n)})$.

**Lemma 2.** The expected value $E(Z_{(2Q+1)}) = 0$.

**Lemma 3.** Eq. (1) is an unbiased estimator.

**Lemma 4.** The expected value $E(Z_{(1)}Z_{(2Q+1)}) = E(Z_{(n)}Z_{(2Q+1)})$.

**Lemma 5.** Eq. (5) can be obtained by letting the first derivative of MSE with respect to $w$, equals to zero.

**Proof 1.** Based on Definition (1) and symmetric properties of standard normal distribution, we have

$$E(Z_{(1)}) = n \int_{-\infty}^{\infty} z_1 [1 - \Phi(z_1)]^{n-1} \phi(z_1) dz_1,$$

$$E(Z_{(n)}) = n \int_{-\infty}^{\infty} z_n [\Phi(z_n)]^{n-1} \phi(z_n) dz_n = n \int_{-\infty}^{\infty} (z_1) [1 - \Phi(z_1)]^{n-1} \phi(z_1) dz_1.$$

Therefore, we can conclude that $E(Z_{(1)}) = -E(Z_{(n)})$.

**Proof 2.** In order to prove that $E(Z_{(2Q+1)}) = 0$, we need to prove $E(Z_{(2Q+1)}) = E(-Z_{(2Q+1)})$. Based on Definition (1),

$$E(Z_{(2Q+1)}) = \frac{4Q+1}{[(2Q)!]^2} \int_{-\infty}^{\infty} z_{2Q+1} \Phi(z_{2Q+1})^{2Q} [1 - \Phi(z_{2Q+1})]^{2Q} \phi(z_{2Q+1}) dz_{2Q+1},$$

$$E(-Z_{(2Q+1)}) = \frac{4Q+1}{[(2Q)!]^2} \int_{-\infty}^{\infty} (-z_{2Q+1}) \Phi(-z_{2Q+1})^{2Q} [1 - \Phi(-z_{2Q+1})]^{2Q} \phi(-z_{2Q+1}) d(-z_{2Q+1}).$$

According to the symmetric properties of standard normal distribution, we could easily conclude that $E(Z_{(2Q+1)}) = E(-Z_{(2Q+1)}) = 0$. 
Proof 3. The expected value of the estimator can be obtained as following:

\[ E(\bar{X}) = \frac{w}{2} E(a) + \frac{w}{2} E(b) + (1 - w) E(m) \]

\[ = \frac{w}{2} E(X_{(1)}) + \frac{w}{2} E(X_{(n)}) + (1 - w) E(X_{(2Q+1)}) \]

\[ = \frac{w}{2} E(\mu + \sigma Z_{(1)}) + \frac{w}{2} E(\mu + \sigma Z_{(n)}) + (1 - w) E(\mu + \sigma Z_{(2Q+1)}) \]

\[ = \frac{w}{2} \left( \mu + \sigma E(Z_{(1)}) \right) + \frac{w}{2} \left( \mu + \sigma E(Z_{(n)}) \right) + (1 - w) \left( \mu + \sigma E(Z_{(2Q+1)}) \right) \]

\[ = w\mu + (1 - w)\mu + \sigma \left[ E(Z_{(1)}) + E(Z_{n}) \right] + \sigma E(Z_{(2Q+1)}). \]

By Lemma 1, the expected value of \( \bar{X} \) becomes \( \mu + \sigma E(Z_{(2Q+1)}) \). Consequently, by applying Lemma 2, the expected value of the mean estimator \( \bar{n} \) now equals to

\[ E(\bar{X}) = \mu \]

which indicates that the estimator is unbiased.

Proof 4. According to Definition 1, \( E(Z_{(1)}Z_{(2Q+1)}) \) and \( E(Z_{(n)}Z_{(2Q+1)}) \) have the following formulations (Let \( M = 2Q + 1 \)):

\[ E(Z_{(1)}Z_{(M)}) = C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{1}z_{M} \left[ \Phi(z_{M}) - \Phi(z_{1}) \right]^{2Q-1} \left[ 1 - \Phi(z_{M}) \right]^{2Q} \phi(z_{1})\phi(z_{M})dz_{1}dz_{M}, \]

\[ E(Z_{(n)}Z_{(M)}) = C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{n}z_{M} \left[ \Phi(z_{M}) - \Phi(z_{n}) \right]^{2Q-1} \left[ \Phi(z_{M}) \right]^{2Q} \phi(z_{n})\phi(z_{M})dz_{n}dz_{M}, \]

where \( C = (4Q + 1)!/[((2Q - 1)!(2Q)!]. \)

Then by the symmetric properties of standard normal distribution, \( z_{1} = -z_{n} \) and \( 1 - \Phi(z_{M}) = \Phi(z_{M}) \), and thus,

\[ E(Z_{(n)}Z_{(M)}) = C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{n}z_{M} \left[ \Phi(z_{n}) - \Phi(z_{M}) \right]^{2Q-1} \left[ \Phi(z_{M}) \right]^{2Q} \phi(z_{M})\phi(z_{n})dz_{n}dz_{M} \]

\[ = C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} (-z_{n})z_{M} \left[ \Phi(z_{n}) - \Phi(z_{M}) \right]^{2Q-1} \left[ 1 - \Phi(z_{M}) \right]^{2Q} \phi(z_{M})\phi(-z_{n})dz_{n}dz_{M} \]

\[ = C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{1}z_{M} \left[ (1 - \Phi(z_{1})) - (1 - \Phi(z_{M})) \right]^{2Q-1} \left[ 1 - \Phi(z_{M}) \right]^{2Q} \phi(z_{M})\phi(z_{1})dz_{1}dz_{M} \]

\[ = C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{1}z_{M} \left[ \Phi(z_{M}) - \Phi(z_{1}) \right]^{2Q-1} \left[ 1 - \Phi(z_{M}) \right]^{2Q} \phi(z_{M})\phi(z_{1})dz_{1}dz_{M} \]

\[ = E(Z_{(1)}Z_{(M)}). \]

That is, we have proved that \( E(Z_{(1)}Z_{(M)}) = E(Z_{(n)}Z_{(M)}). \)
**Proof 5.** Recall the MSE in Section 2.2.1 and simplifies it to the following equation (again, let $M = 2Q + 1$):

$$
\text{MSE}(\bar{X}(w)) = \frac{w^2}{2} \sigma^2 \left[ E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) \right] \\
+ w(1-w) \sigma^2 \left[ E(Z_{(1)}Z_{(M)}) + E(Z_{(n)}Z_{(M)}) \right] \\
+ (1-w)^2 \sigma^2 \left[ E(Z_{(M)}^2) - E(Z_{(M)}^2) \right].
$$

By taking the first derivative of MSE with respect to $w$ and let the result to be zero, the optimal weight of the sample midrange can be obtained by:

$$
w_{\text{opt}}(n) = \frac{2 \left[ E(Z_{(1)}^2) - E(Z_{(1)}Z_{(M)}) \right] - \left[ E(Z_{(1)}Z_{(M)}) + E(Z_{(n)}Z_{(M)}) \right]}{E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) + 2 \left[ E(Z_{(M)}^2) - E(Z_{(M)}^2) - E(Z_{(1)}Z_{(M)}) - E(Z_{(n)}Z_{(M)}) \right]}.
$$

However, according to Lemma 2 and Lemma 4, we can eliminate or combine some terms in the above equation and finally, we have,

$$
w_{\text{opt}}(n) = \frac{2 \left[ E(Z_{(M)}^2) - E(Z_{(1)}Z_{(M)}) \right]}{E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) + 2 \left[ E(Z_{(M)}^2) - 2E(Z_{(1)}Z_{(M)}) \right]}.
$$

If we let $K(n) = 2[E(Z_{(M)}^2) - E(Z_{(1)}Z_{(M)})]/[E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) - 2E(Z_{(1)}Z_{(M)})]$, the above equation is exactly the optimal weight formula, Eq. (5).

**B Scenario $S_2$**

Recall the estimator of the sample mean for scenario $S_2$,

$$
\bar{X}(w) = w \left( \frac{q_1 + q_3}{2} \right) + (1-w)m,
$$

where $w$ and $1-w$ are the weights assigned to the mid-quartile range $(q_1 + q_3)/2$ and the median $m$, respectively.

**B.1 Important lemmas**

To simplify the expression of the expected values, let positive integers $Q_1 = Q + 1$, $M = 2Q + 1$ and $Q_3 = 3Q + 1$.

**Lemma 6.** The expected values $E(Z_{(Q_1)}) = -E(Z_{(Q_3)})$.

**Lemma 7.** Eq. (5) is an unbiased estimator.
Lemma 8. The expected values \( E(Z_{Q_1}Z_{M}) = E(Z_{M}Z_{Q_3}) \). 

Lemma 9. Eq. (9) can be obtained by letting the first derivative of MSE with respect to \( w \), equals to zero.

Proof 6. By Definition (11) and symmetric properties of standard normal distribution, we have
\[
E(Z_{Q_1}) = C \int_{-\infty}^{\infty} z_{Q_1} [\Phi(z_{Q_1})]^Q [1 - \Phi(z_{Q_1})]^{3Q} \phi(z_{Q_1}) dz_{Q_1},
\]
\[
E(Z_{Q_3}) = C \int_{-\infty}^{\infty} z_{Q_3} [\Phi(z_{Q_3})]^Q [1 - \Phi(z_{Q_3})]^{3Q} \phi(z_{Q_3}) dz_{Q_3}
\]
\[= C \int_{-\infty}^{\infty} (-z_{Q_1}) [\Phi(z_{Q_1})]^Q [1 - \Phi(z_{Q_1})]^{3Q} \phi(z_{Q_1}) dz_{Q_1},
\]
where \( C = (4Q + 1)!/[Q!(3Q)!] \).

Therefore, we proved that \( E(Z_{Q_1}) = -E(Z_{Q_3}) \).

Proof 7. Since the mean estimator for scenario \( S_2 \) has similar form as the estimator for \( S_1 \), we can easily obtain that
\[
E(\bar{X}) = \frac{w}{2} E(q_1) + \frac{w}{2} E(q_3) + (1 - w)E(m)
\]
\[= w\mu + (1 - w)\mu + \sigma \left[ E(Z_{Q_1}) + E(Z_{Q_3}) \right] + \sigma E(Z_{M}).
\]

By Lemma 2 and Lemma 6 the above equation becomes \( E(\bar{X}) = \mu \), which implies that the proposed estimator for this scenario is unbiased.

Proof 8. According to Definition (11)
\[
E(Z_{Q_1}Z_{M}) = C \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{Q_1} z_{M} [\Phi(z_{Q_1})]^Q [1 - \Phi(z_{M})]^{2Q} \frac{[\Phi(z_{M}) - \Phi(z_{Q_1})]^{Q-1} \phi(z_{Q_1}) \phi(z_{M}) dz_{Q_1} dz_{M}},
\]
\[
E(Z_{Q_3}Z_{M}) = C \int_{-\infty}^{\infty} \int_{-\infty}^{z_{M}} z_{Q_3} z_{M} [\Phi(z_{M})]^{2Q} [1 - \Phi(z_{M})]^Q \frac{[\Phi(z_{M}) - \Phi(z_{Q_3})]^{Q-1} \phi(z_{M}) \phi(z_{Q_3}) dz_{M} dz_{Q_3}},
\]
where \( C = (4Q + 1)!/[Q!(2Q)!](Q - 1)! \)
By the symmetric properties of standard normal distribution, \( z_{Q_3} = -z_{Q_3} \) and 1 – \( \Phi(z_M) = \Phi(-z_M) \). Thus,

\[
E(Z_{(Q_3)}Z_{(M)}) = C \int_{-\infty}^{\infty} \int_{z_{Q_3}}^{z_M} \Phi(z_M)^{2Q} [1 - \Phi(z_{Q_3})]^Q dz_M dz_{Q_3},
\]

\[
= C \cdot \int_{-\infty}^{\infty} \int_{-\infty}^{z_M} (-z_{Q_3})z_M [1 - \Phi(z_M)]^{2Q} [\Phi(z_{Q_1})]^Q dz_M dz_{Q_3},
\]

\[
= C \cdot \int_{-\infty}^{\infty} \int_{z_{Q_1}}^{z_M} z_M [1 - \Phi(z_M)]^{2Q} [\Phi(z_{Q_1})]^Q dz_M dz_{Q_1},
\]

\[
= E(Z_{(Q_1)}Z_{(M)}).
\]

**Proof 9.** For scenario \( S_2 \), by letting \(-\infty < z_1 \leq z_{2Q+1} \leq z_n < \infty \), the MSE of the proposed estimator (Eq. (3)) can be expressed as:

\[
\text{MSE}(\bar{X}(w)) = \text{Var}\left[w \left( \frac{q_1 + q_3}{2} \right) + (1 - w)m\right]
\]

\[
= \frac{w^2}{4} \text{Var}(q_1 + q_3) + (1 - w)^2 \text{Var}(m) + w(1 - w)\text{Cov}(q_1 + q_3, m).
\]

It is obvious that Eq. (3) for the current scenario and Eq. (4) for the previous scenario has similar form of formulation. Hence, the MSE should also has a similar simplified formulation as in Proof[5]

\[
\text{MSE}(\bar{X}(w)) = \frac{w^2}{2} \sigma^2 \left[ E(Z_{(Q_1)}^2) + E(Z_{(Q_1)}Z_{(Q_3)}) \right]
\]

\[
+ w(1 - w)\sigma^2 \left[ E(Z_{(Q_1)}Z_{(M)}) + E(Z_{(Q_3)}Z_{(M)}) \right]
\]

\[
+ (1 - w)^2 \sigma^2 \left[ E(Z_{(Q_1)}^2) - E(Z_{(M)}^2) \right].
\]

By taking the first derivative of the above MSE equation with respect to \( w \) and let the result to be zero, the optimal weight of the sample mid-quartile range can be obtained by:

\[
w_{\text{opt}}(n) = \frac{2E(Z_{(M)}^2) - [E(Z_{(Q_1)}Z_{(M)}) + E(Z_{(M)}Z_{(Q_3)})]}{E(Z_{(Q_1)}^2) + E(Z_{(Q_1)}Z_{(Q_3)}) + 2 \left[ E(Z_{(M)}^2) - E(Z_{(Q_1)}Z_{(M)}) - E(Z_{(M)}Z_{(Q_3)}) \right]}.
\]

By Lemma[2] and Lemma[8] \( E(Z_{(Q_1)}Z_{(M)}) \) and \( E(Z_{(M)}Z_{(Q_3)}) \) can be combined and \( E(Z_{(M)}) \) can be eliminated. Eventually, \( w_{\text{opt}}(n) \) has below formulation:

\[
w_{\text{opt}}(n) = \frac{2 \left[ E(Z_{(M)}^2) - E(Z_{(Q_1)}Z_{(M)}) \right]}{E(Z_{(Q_1)}^2) + E(Z_{(Q_1)}Z_{(Q_3)}) + 2 \left[ E(Z_{(M)}^2) - 2E(Z_{(Q_1)}Z_{(M)}) \right]}.
\]
If we use the notations of variance (Var) and covariance (Cov) instead of expected values, the above equation is exactly Eq. (11).

C Scenario $S_3$

Recall the estimator of the sample mean for scenario $S_3$,

\[ \bar{X} = w_1 \left( \frac{a + b}{2} \right) + w_2 \left( \frac{q_1 + q_3}{2} \right) + (1 - w_1 - w_2)m, \]

where $w_1$, $w_2$ and $(1 - w_1 - w_2)$ are the weights assigned to the mid-range $(a + b)/2$, the mid-quartile range $(q_1 + q_3)/2$ and the median $m$, respectively.

C.1 Important lemmas

Following the same simplified notation in Appendix B, let positive integers $Q_1 = Q + 1$, $M = 2Q + 1$ and $Q_3 = 3Q + 1$.

**Lemma 10.** Eq. (12) is an unbiased estimator.

**Lemma 11.** The expected values $E(Z(1)Z(Q_1)) = E(Z(n)Z(Q_3))$ and $E(Z(1)Z(Q_3)) = E(Z(n)Z(Q_1))$.

**Lemma 12.** Eq. (18) can be obtained by taking the partial derivatives of MSE with respect to $w_1$ and $w_2$ and let them equal to zero, respectively.

**Proof 10.** The expected value of the estimator can be obtained as

\[
E(\bar{X}) = \frac{w_1}{2} E(a) + \frac{w_1}{2} E(b) + \frac{w_2}{2} E(q_1) + \frac{w_2}{2} E(q_3) + (1 - w_1 - w_2) E(m) \\
= \frac{w_1}{2} E(X(1)) + \frac{w_1}{2} E(X(n)) + \frac{w_2}{2} E(X_{(Q+1)}) + \frac{w_2}{2} E(X_{3Q+1}) + (1 - w_1 - w_2) E(X_{2Q+1}) \\
= \frac{w_1}{2} E \left( \mu + \sigma Z(1) \right) + \frac{w_1}{2} E \left( \mu + \sigma Z(n) \right) + \frac{w_2}{2} E \left( \mu + \sigma Z_{(Q+1)} \right) \\
+ \frac{w_2}{2} E \left( \mu + \sigma Z_{(3Q+1)} \right) + (1 - w_1 - w_2) E \left( \mu + \sigma Z_{(2Q+1)} \right) \\
= \frac{w_1}{2} \left[ \mu + \sigma E(Z(1)) \right] + \frac{w_1}{2} \left[ \mu + \sigma E(Z(n)) \right] \\
+ \frac{w_2}{2} \left[ \mu + \sigma E(Z_{(Q+1)}) \right] + \frac{w_2}{2} \left[ \mu + \sigma E(Z_{(3Q+1)}) \right] + (1 - w_1 - w_2) \left[ \mu + \sigma E(Z_{(2Q+1)}) \right] \\
= \mu + \frac{w_1}{2} \sigma \left[ E(Z(1)) + E(Z(n)) \right] \\
+ \frac{w_2}{2} \sigma \left[ E(Z_{(Q+1)}) + E(Z_{3Q+1}) \right] + \sigma E \left( Z_{(2Q+1)} \right)
\]
By using Lemma 1, Lemma 2 and Lemma 3, the expected value of $\bar{X}$ simply equals to $\mu$. Therefore, the mean estimator for scenario $S_3$ is also unbiased.

**Proof 11.** Based on Definition 1

\[
E(Z_{(1)}Z_{(Q_1)}) = C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{z_{Q_1}} z_1 z_{Q_1} [1 - \Phi(z_{Q_1})]^{3Q} \Phi(z_{Q_1}) - \Phi(z_1)]^{Q-1} \phi(z_{Q_1})\phi(z_1)dz_1dz_{Q_1},
\]

\[
E(Z_{(n)}Z_{(Q_3)}) = C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{z_{Q_3}} z_{Q_3} z_n [\Phi(z_{Q_3})]^{3Q} \Phi(z_n) - \Phi(z_{Q_3})]^{Q-1} \phi(z_n)\phi(z_{Q_3})dz_n dz_{Q_3},
\]

\[
E(Z_{(1)}Z_{(Q_3)}) = C_2 \int_{-\infty}^{\infty} \int_{-\infty}^{z_{Q_3}} z_1 z_{Q_3} [1 - \Phi(z_{Q_3})]^{Q} \Phi(z_{Q_3}) - \Phi(z_1)]^{3Q-1} \phi(z_{Q_3})\phi(z_1)dz_1dz_{Q_3},
\]

\[
E(Z_{(n)}Z_{(Q_1)}) = C_2 \int_{-\infty}^{\infty} \int_{-\infty}^{z_{Q_1}} z_{Q_1} z_n [\Phi(z_{Q_1})]^{Q} \Phi(z_n) - \Phi(z_{Q_1})]^{3Q-1} \phi(z_n)\phi(z_{Q_1})dz_n dz_{Q_1},
\]

where $C_1 = (4Q + 1)!/[3Q!Q!]$, $C_2 = (4Q + 1)!/[Q!3Q!]$.

Since $E(Z_{(1)}Z_{(Q_1)})$ and $E(Z_{(n)}Z_{(Q_3)})$, $E(Z_{(1)}Z_{(Q_3)})$ and $E(Z_{(n)}Z_{(Q_1)})$ have the similar formulations, our first task is to prove $E(Z_{(1)}Z_{(Q_1)}) = E(Z_{(n)}Z_{(Q_3)})$. By the symmetric properties of standard normal distribution, it is easy to obtain that:

\[
E(Z_{(n)}Z_{(Q_3)}) = C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{z_{Q_3}} z_{Q_3} z_n [\Phi(z_{Q_3})]^{3Q} \Phi(z_n) - \Phi(z_{Q_3})]^{Q-1} \phi(z_n)\phi(z_{Q_3})dz_n dz_{Q_3},
\]

\[
= C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{z_{Q_3}} z_{Q_3} z_n [\Phi(z_{Q_3})]^{3Q} \Phi(z_n) - \Phi(z_{Q_3})]^{Q-1} \phi(z_n)\phi(z_{Q_3})dz_n dz_{Q_3},
\]

\[
= C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{-z_{Q_3}} (-z_{Q_3})(-z_n) [1 - \Phi(z_{Q_3})]^{3Q} \Phi(-z_{Q_3})\phi(z_{Q_3})dz_{Q_3} dz_n,
\]

\[
= C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{-z_{Q_3}} (1 - \Phi(z_{Q_3}))^{Q-1} \phi(-z_{Q_3})\phi(-z_n)d(-z_{Q_3})d(-z_n),
\]

\[
= C_1 \int_{-\infty}^{\infty} \int_{-\infty}^{-z_{Q_3}} (1 - \Phi(z_{Q_3}))^{Q-1} \phi(z_{Q_3})\phi(z_n)d(-z_{Q_3})d(-z_n),
\]

\[
= E(Z_{(1)}Z_{(Q_1)}).
\]
Following the same logic as above, $E(Z_{(1)}Z_{(Q)}) = E(Z_{(n)}Z_{(Q)})$ can also be proved easily.

**Proof 12.** The MSE of the sample mean estimator (Eq. 12) as

$$\text{MSE}(\bar{X}(w)) = \text{Var}(\bar{X})$$

$$= \text{Var} \left( \frac{w_1}{2} a + \frac{w_1}{2} b + \frac{w_2}{2} q_1 + \frac{w_2}{2} q_3 + (1 - w_1 - w_2) m \right)$$

$$= \left[ \frac{w_1^2}{4} \text{Var}(a) + \frac{w_1^2}{4} \text{Var}(b) \right] + \left[ \frac{w_2^2}{4} \text{Var}(q_1) + \frac{w_2^2}{4} \text{Var}(q_3) \right] + (1 - w_1 - w_2)^2 \text{Var}(m)$$

$$+ \frac{w_1^2}{2} \text{Cov}(a, b) + w_1(1 - w_1 - w_2) \text{Cov}(a, m) + w_1(1 - w_1 - w_2) \text{Cov}(b, m)$$

$$+ \frac{w_2^2}{2} \text{Cov}(q_1, q_3) + w_2(1 - w_1 - w_2) \text{Cov}(q_1, m) + w_2(1 - w_1 - w_2) \text{Cov}(q_3, m)$$

$$+ \frac{w_1w_2}{2} \left[ \text{Cov}(a, q_1) + \text{Cov}(b, q_1) + \text{Cov}(a, q_3) + \text{Cov}(b, q_3) \right] .$$

$$= \frac{w_1^2}{2} \sigma^2 \left[ E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) \right] + \frac{w_2^2}{2} \sigma^2 \left[ E(Z_{(Q+1)}^2) + E(Z_{(Q+1)}Z_{(3Q+1)}) \right]$$

$$+ w_1(1 - w_1 - w_2) \sigma^2 \left[ E(Z_{(1)}Z_{(2Q+1)}) + E(Z_{(n)}Z_{(2Q+1)}) \right]$$

$$+ w_2(1 - w_1 - w_2) \sigma^2 \left[ E(Z_{(Q+1)}Z_{(2Q+1)}) + E(Z_{(3Q+1)}Z_{(2Q+1)}) \right]$$

$$+ (1 - w_1 - w_2)^2 \sigma^2 E(Z_{(2Q+1)}^2)$$

$$+ \frac{w_1w_2}{2} \sigma^2 \left[ E(Z_{(1)}Z_{(Q+1)}) + E(Z_{(n)}Z_{(Q+1)}) + E(Z_{(1)}Z_{(3Q+1)}) + E(Z_{(n)}Z_{(3Q+1)}) \right] .$$

However, the above formula is still very complex and our next task is to simplify it.

By applying Lemma 2, Lemma 4, Lemma 8 and Lemma 11 the MSE equation could be further simplified as:

$$\text{MSE}(\bar{X}(w)) = \frac{w_1^2}{2} \sigma^2 \left[ E(Z_{(1)}^2) + E(Z_{(1)}Z_{(n)}) \right]$$

$$+ \frac{w_2^2}{2} \sigma^2 \left[ E(Z_{(Q+1)}^2) + E(Z_{(Q+1)}Z_{(3Q+1)}) \right]$$

$$+ (1 - w_1 - w_2)^2 \sigma^2 E(Z_{(2Q+1)}^2)$$

$$+ 2(1 - w_1 - w_2)(w_1 + w_2) \left[ E(Z_{(1)}Z_{(m)}) + E(Z_{(q_1)}Z_{(m)}) \right]$$

$$+ w_1w_2 \sigma^2 \left[ E(Z_{(1)}Z_{(Q+1)}) + E(Z_{(1)}Z_{(3Q+1)}) \right] .$$

As it is mentioned in Section 2.2.3, to find the optimal weights that minimize the MSE of $\bar{X}$, we need to let the partial derivatives of MSE formula with respect to $w_1$ and $w_2$, to be zero. However, as Eq. 13 is derived based on the initial MSE equation, which does not use the expected values in the formulation, we will just follow the same notations in Section 2.2.3.
Thus, the partial derivatives have the following formulation (for ease of display, characters \( A \) to \( F \) are already defined in Section 2.2.3).

\[
\begin{align*}
\frac{\partial \text{MSE}(\bar{X})}{\partial w_1} &= w_1(A + 4C - 4E) + w_2(4C + D - 2E - 2F) - (4C - 2E) = 0, \\
\frac{\partial \text{MSE}(\bar{X})}{\partial w_2} &= w_1(4C + D - 2E - 2F) + w_2(B + 4C - 2F) - (4C - 2F) = 0.
\end{align*}
\]

By solving the above equations, the optimal weights \( w_{1,\text{opt}}(n) \) and \( w_{2,\text{opt}}(n) \) are given as

\[
w_{1,\text{opt}}(n) = \left[ \frac{4C - 2E}{4C + D - 2E - 2F} - \frac{4C - 2F}{B + 4C - 4F} \right] \left[ \frac{A + 4C - 4E}{4C + D - 2E - 2F} - \frac{4C + D - 2E - 2F}{B + 4C - 4F} \right]^{-1}
\]

(17)

and

\[
w_{2,\text{opt}}(n) = \frac{(4C - 2F) - w_1(4C + D - 2E - 2F)}{B + 4C - 4F},
\]

(18)

where \( A = \text{Var}(a + b), B = \text{Var}(q_1 + q_3), C = \text{Var}(m), D = \text{Cov}(a + b, q_1 + q_3), E = \text{Cov}(a + b, m), \) and \( F = \text{Cov}(q_1 + q_3, m). \)