Recovering Mean and Standard Deviation from Boxplot Data: A Novel Bayesian Approach

THE UNIVERSITY OF TEXAS AT ARLINGTON

Wenqisi (Lydia) Pan[1], Zeyu Lu[1], Lin Xu[2], Xinlei Wang[1]

The University of Texas at Arlington, Arlington, Texas.[1] University of Texas Southwestern Medical Center, Dallas, Texas.[2]

Introduction

Many clinical studies report summary statistics such as the median, quartiles, and range in boxplots, rather than providing sample means and standard deviations, which are crucial for meta-analyses.

To bridge this gap, we propose a novel Bayesian approach that utilizes the joint distribution of order statistics and weakly informative priors to **estimate the mean** and **standard deviation** while also **quantifying uncertainty**. Ultimately, our method enhances the reliability of quantitative synthesis in clinical research.

Problem Setup

n = sample size	S_i = Scenario i
a = the minimum value	
q_1 = the first quartile	$S_0 = \{\bar{x}, s; n\}$
m = the median	$S_1 = \{a, m, b; n\}$
q_3 = the third quartile	$S_2 = \{q_1, m, q_3; n\}$
b = the maximum value	$S_3 = \{a, q_1, m, q_3, b; n\}$

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from the normal distribution $N(\mu, \sigma^2)$, and $X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ be the ordered statistics of $x_1, x_2, \ldots, x_n$. For simplicity, let $\boldsymbol{n = 4q+1}$, with q being positive integer.

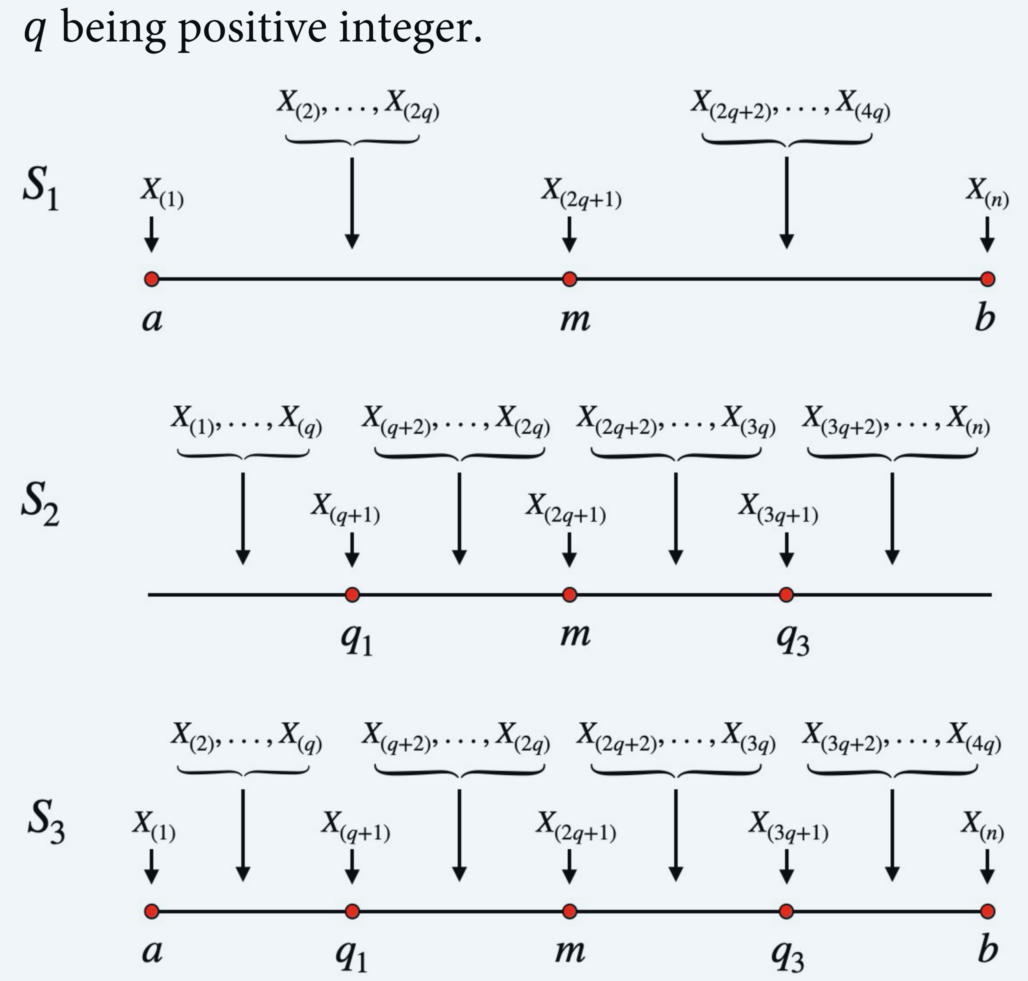

Figure 1 Illustration of the data structure for the case n=4q+1 across all three scenarios.

Method

$S_1 = \{a, m, b; n\}$

$$p(\mu,\sigma^2 \mid X) \propto p(X \mid \mu,\sigma^2)\cdot\pi(\mu)\cdot\pi(\sigma^2)$$

$$\begin{aligned}
&\propto N(X_{(1)} \mid \mu,\sigma^2)\cdot N(X_{(2q+1)} \mid \mu,\sigma^2)\cdot N(X_{(4q+1)} \mid \mu,\sigma^2)\\
&\cdot\left[\Phi\left(\frac{X_{(2q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(1)}-\mu}{\sigma}\right)\right]^{2q-1}\\
&\cdot\left[\Phi\left(\frac{X_{(4q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(2q+1)}-\mu}{\sigma}\right)\right]^{2q-1}\\
&\cdot\pi(\mu)\cdot\pi(\sigma^2)
\end{aligned}$$

$S_2 = \{q_1, m, q_3; n\}$

$$p(\mu,\sigma^2 \mid X) \propto p(X \mid \mu,\sigma^2)\cdot\pi(\mu)\cdot\pi(\sigma^2)$$

$$\begin{aligned}
&\propto N(X_{(q+1)} \mid \mu,\sigma^2)\cdot N(X_{(2q+1)} \mid \mu,\sigma^2)\cdot N(X_{(3q+1)} \mid \mu,\sigma^2)\\
&\cdot\left[\Phi\left(\frac{X_{(q+1)}-\mu}{\sigma}\right)\right]^{q}\\
&\cdot\left[\Phi\left(\frac{X_{(2q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(q+1)}-\mu}{\sigma}\right)\right]^{q-1}\\
&\cdot\left[\Phi\left(\frac{X_{(3q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(2q+1)}-\mu}{\sigma}\right)\right]^{q-1}\\
&\cdot\left[1-\Phi\left(\frac{X_{(3q+1)}-\mu}{\sigma}\right)\right]^{q}\\
&\cdot\pi(\mu)\cdot\pi(\sigma^2)
\end{aligned}$$

$S_3 = \{a, q_1, m, q_3, b; n\}$

$$p(\mu,\sigma^2 \mid X) \propto p(X \mid \mu,\sigma^2)\cdot\pi(\mu)\cdot\pi(\sigma^2)$$

$$\begin{aligned}
&\propto N(X_{(1)} \mid \mu,\sigma^2)\cdot N(X_{(q+1)} \mid \mu,\sigma^2)\cdot N(X_{(2q+1)} \mid \mu,\sigma^2)\\
&\cdot N(X_{(3q+1)} \mid \mu,\sigma^2)\cdot N(X_{(4q+1)} \mid \mu,\sigma^2)\\
&\cdot\left[\Phi\left(\frac{X_{(q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(1)}-\mu}{\sigma}\right)\right]^{q-1}\\
&\cdot\left[\Phi\left(\frac{X_{(2q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(q+1)}-\mu}{\sigma}\right)\right]^{q-1}\\
&\cdot\left[\Phi\left(\frac{X_{(3q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(2q+1)}-\mu}{\sigma}\right)\right]^{q-1}\\
&\cdot\left[\Phi\left(\frac{X_{(4q+1)}-\mu}{\sigma}\right)-\Phi\left(\frac{X_{(3q+1)}-\mu}{\sigma}\right)\right]^{q-1}\\
&\cdot\pi(\mu)\cdot\pi(\sigma^2)
\end{aligned}$$

We specify the prior of μ and σ^2 as:

$$\pi(\mu) \propto I\left(L_\mu \le \mu \le U_\mu\right)$$

$$\pi(\sigma^2) \propto (\sigma^2)^{-\alpha-1} e^{-\frac{\beta}{\sigma^2}}$$

The lower and upper bound was set for μ to a safe range and $\alpha = \beta = 0.01$ for σ^2. With 2-dimensional grid approximation, we can get our estimates for μ and σ, denoted as $\hat{x}$ and $\hat{s}$.

Results

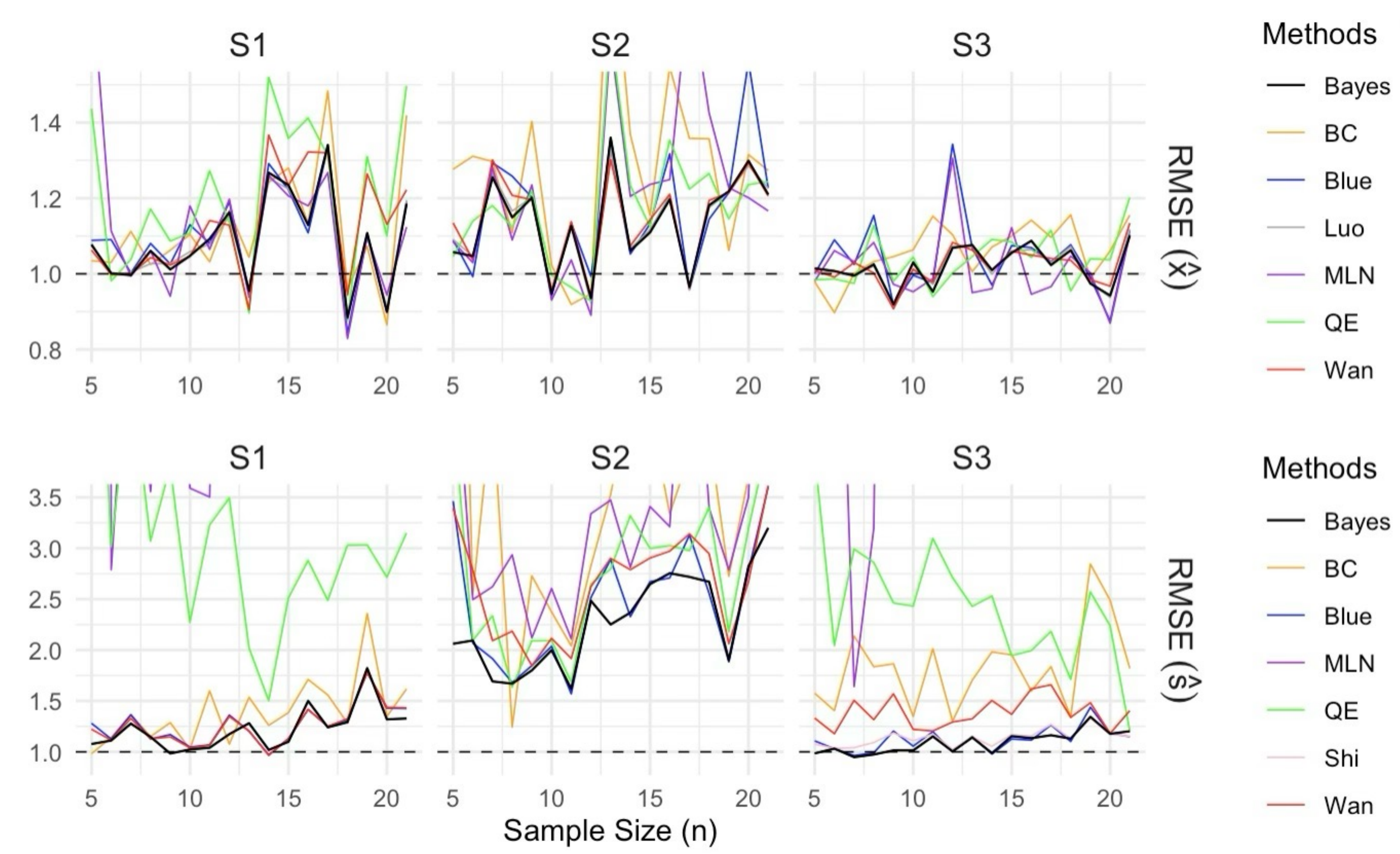

Figure 2 **Relative Mean Square Error** of the sample mean and sample standard deviation estimation for normal data. $\boldsymbol{RMSE(\hat{x})} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\hat{x}-\mu)^2}{\frac{1}{T}\sum_{t=1}^{T}(\bar{x}-\mu)^2}$, $\boldsymbol{RMSE(\hat{s})} = \frac{\frac{1}{T}\sum_{t=1}^{T}(\hat{s}-\sigma)^2}{\frac{1}{T}\sum_{t=1}^{T}(s-\sigma)^2}$, *smaller RMSE value indicates better performance.*

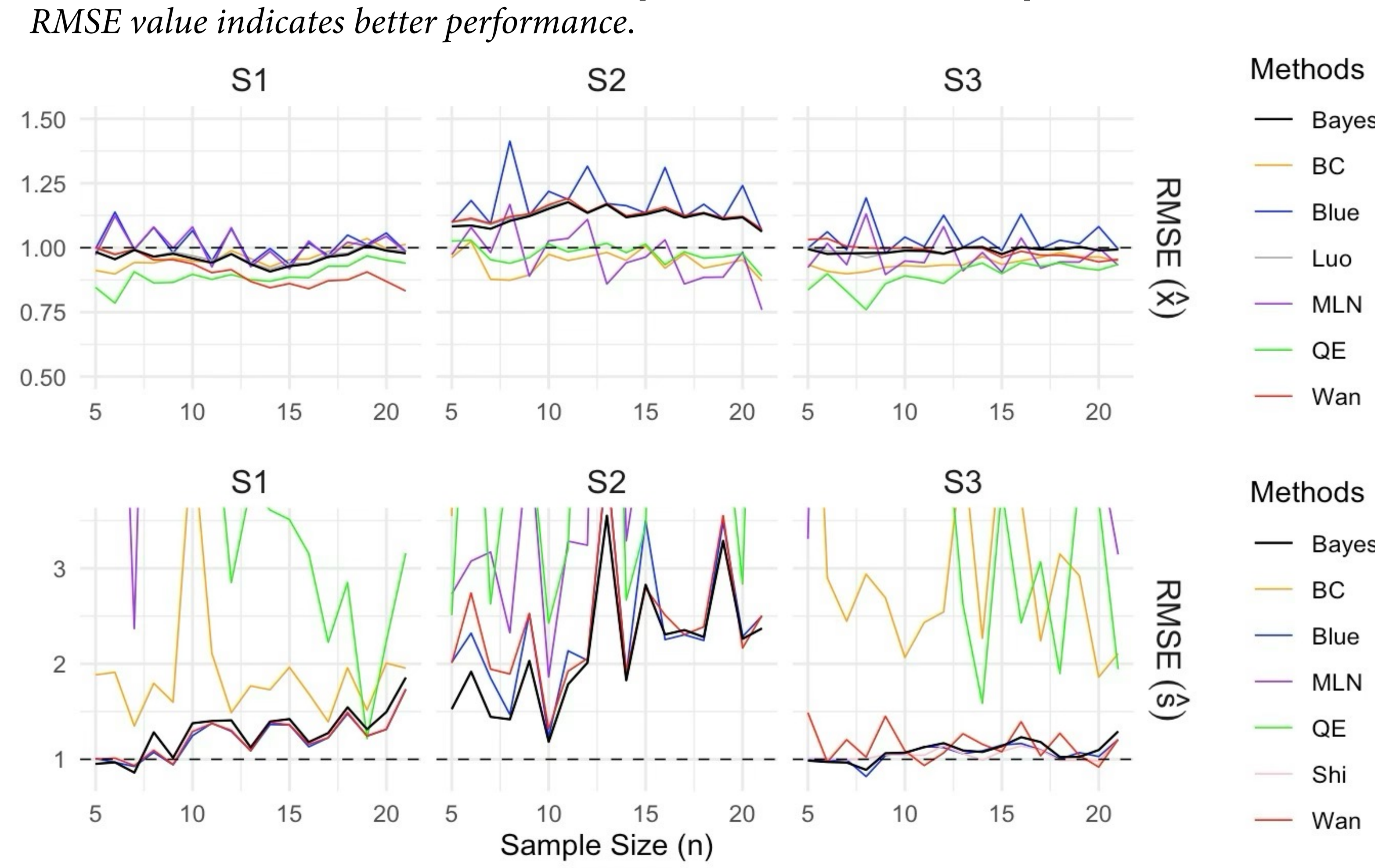

Figure 3 **Relative Mean Square Error** of the sample mean and sample standard deviation estimation for skewed data (Robustness Testing).

Conclusion

Estimation Accuracy (RMSE)

- ✓ Our method remains competitive among top methods across all scenarios (S1–S3)
- ✓ Shows consistently low RMSE, especially for estimating σ

Robustness

- ✓ Stable performance across small sample sizes, skewed or heavy-tailed distributions, and the presence of outliers
- ✓ Robust estimation of σ across all scenarios and competitive performance for μ

Uncertainty Quantification

- ✓ Provides credible intervals for both μ and σ, which is not available in any other methods
- ✓ Coverage rate consistently greater than 0.95 across all scenarios, ensuring reliable interval estimates — critical for clinical data applications

Bayesian Framework

- ➢ Bayesian approach enables the integration of new clinical datasets as priors
- ➢ Posterior updating without re-running all simulations
- ➢ Efficient incorporation of historical evidence

Future Work

- ➢ Extend to varying **treatment effect** measures in meta-analyses
- ➢ Develop customized methods for the other **skewed distributions**
- ➢ Enhance real-world applicability when data deviate from normality

References

1. Wan, Xiang, et al. "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range." *BMC medical research methodology* 14 (2014): 1-13.
2. Luo, D., Wan, X., Liu, J., & Tong, T. (2018). Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. Statistical methods in medical research, 27(6), 1785-1805.
3. Shi, Jiandong, et al. "Optimally estimating the sample standard deviation from the five-number summary." Research synthesis methods 11.5 (2020): 641-654.
4. McGrath, S., Zhao, X., Steele, R., Thombs, B. D., Benedetti, A., & DEPRESsion Screening Data (DEPRESSD) Collaboration. (2020). Estimating the sample mean and standard deviation from commonly reported quantiles in meta-analysis. Statistical methods in medical research, 29(9), 2520-2537.
5. Cai, Siyu, Jie Zhou, and Jianxin Pan. "Estimating the sample mean and standard deviation from order statistics and sample size in meta-analysis." *Statistical methods in medical research* 30.12 (2021): 2701-2719.
6. Yang, B., Huang, X., Gao, H., Leung, N. H., Tsang, T. K., & Cowling, B. J. (2022). Immunogenicity, efficacy, and safety of SARS-CoV-2 vaccine dose fractionation: a systematic review and meta-analysis. BMC medicine, 20(1), 409.