A Flexible, Likelihood-Based Method for Estimating Underlying Distributions, Mean and SD, Based on Reported Quantiles Such as Median and Inter-Quartile Range

Kim Rand PhD, Anabel Estévez-Carrillo MSc
Maths in Health B.V., The Netherlands

INTRODUCTION

Evidence synthesis using methods including meta- and network meta-analysis rely on reported sample size, mean and variance indicators. However, studies commonly report other statistics, such as the median, inter-quartile range (IQR), range, confidence interval, percentiles, or various combinations thereof.

Several methods have been developed to estimate mean and standard deviation based on reported median + IQR or median + range. However, these methods are limited to cases including the median + IQR, range, or both IQR and range, limiting usefulness where atypical quantiles are reported, such as e.g. mean and 90% confidence interval. Furthermore, the existing methods work best in situations where the underlying distribution approximates normal. Unfortunately, the use of e.g. median and IQR in reporting tends to be in cases where the data is *not* normally distributed, meaning that the methods are best suited where they are not needed.

Table 1. Distributional Cases

Name	Distribution
Var1	Normal, Mean = 5, SD = 1
Var2	Lognormal, meanlog = 5, SDlog = 0.25
Var3	Lognormal, meanlog = 5, SDlog = 0.5
Var4	Lognormal, meanlog = 5, SDlog = 1
Var5	Normal, Mean = 50, SD = 17
Var6	Lognormal, meanlog = 4, SDlog = 0.3
Var7	Exponential, rate = 10
Var8	Beta, shape1 = 9, shape2 = 4
Var9	Weibull, shape = 2, scale = 35

METHODS

CURRENT METHODS FOR ESTIMATING MEAN AND SD

The literature on meta-analyses is focused on three kinds of statistics often reported, usually referred to as scenarios 1, 2, and 3:

- **S1**: Min, median, max
- **S2**: 25%, median, 75%
- **S3**: Min, 25%, median, 75%, max

In cases where a normal distribution can be assumed, the method presented by Luo et al. 2018[1] is considered superior for the mean, and the method by Wan et al. 2014[2], is considered best for estimating SD. However, these quantiles tend to be reported when the distribution is not normal.

Luo et al. method for estimating means

$$\hat{x} = \left(\frac{4}{4+n^{0.75}}\right)\frac{Q_{min}+Q_{max}}{2} + \left(\frac{n^{0.75}}{5+n^{0.75}}\right)Q_2 \quad in\ S1$$

$$\hat{x} = \left(0.7\frac{0.39}{n}\right)\frac{Q_1+Q_3}{2} + \left(0.3-\frac{0.39}{n}\right)Q_2 \quad in\ S2$$

$$\hat{x} = \left(\frac{2.2}{2.2+n^{0.75}}\right)\frac{Q_{min}+Q_{max}}{2} + \left(0.7\frac{0.72}{n^{0.55}}\right)\frac{Q_1+Q_3}{2} + \left(0.3-\frac{0.72}{n^{0.55}}-\frac{2.2}{2.2+n^{0.75}}\right)Q_2\left(\frac{n^{0.75}}{5+n^{0.75}}\right)Q_2 \quad in\ S3$$

Wan et al. method for estimating SD

$$\hat{s} = \frac{Q_{max}-Q_{min}}{2\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)} \quad in\ S1$$

$$\hat{s} = \frac{Q_3-Q_1}{2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)} \quad in\ S2$$

$$\hat{s} = \frac{Q_{max}-Q_{min}}{4\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)} + \frac{Q_3-Q_1}{4\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)} \quad in\ S3$$

McGrath 2020[3] suggests two new methods, referred to as quantile estimation (QE) and box-cox (BC). The QE approach minimizes the squared difference between the observed and expected quantiles, while the BC approach uses a box-cox transformation combined with the methods by Luo (mean) and Wan (SD).

THE INTERVAL REGRESION METHOD

We developed a flexible, likelihood-based approach to estimating underlying distributions from reported quantile information, the interval regression (IR) approach.

The procedure is implemented in R and combines interval regression with the density function for observation i in a ranked order of N observations. Let φ and Φ denote the probability and cumulative density function for the distribution in question, respectively (not necessarily limited to the normal). We generalize the likelihood of a singular observation with known rank in a known number of observations drawn from a continuous distribution:

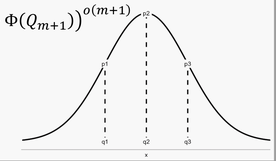
$$p(x_i) = \Phi(x_i)^{i-1}\phi(x_i)\left(1-\Phi(x_i)\right)^{N-i}$$

Let q1, q2, ... qm denote the reported quantiles, N be the number of observations, and n = N-m. We get the general expression for the loglikelihood to be maximized:

$$p(q) = \Phi(Q_1)^{o(1)}\phi(Q_1)\left(\Phi(Q_2)-\Phi(Q_1)\right)^{o(2)}\phi(Q_2)\ldots\left(\Phi(Q_m)-\Phi\left(Q_{(m-1)}\right)\right)^{o(m)}\phi(Q_m)\left(1-\Phi(Q_{m+1})\right)^{o(m+1)}$$

Or, on a log scale

$$log\left(p(q)\right) = \sum_{i=1}^{m}\left[log\left(\phi(Q_i)\right)\right] + \sum_{i=1}^{m+1}\left[o(i)\log\left(\Phi(Q_i)-\Phi\left(Q_{(i-1)}\right)\right)\right]$$



PERFORMANCE EVALUATION

The IR method was compared to the quantile estimation and box-cox approaches developed by McGrath 2020 using simulated data as in McGrath 2020, drawn randomly nine different distributions, with between 25 and 1000 observations.

- The distributional cases reported in the literature are presented in **Table 1**.
- Sample sizes were varied between 25 and 1000
- 1000 samples of each type of each sample size

The methods were compared in terms of Average Relative Error $\left(ARE(\hat{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\hat{x_i}-\bar{x_i}}{\bar{x_i}}\right)$, and Absolute Relative Error $\left(AbsRE(\hat{x}) = \frac{1}{n}\sum_{i=1}^{n}\frac{abs(\hat{x_i}-\bar{x_i})}{\bar{x_i}}\right)$

Figure 2. Graphical Presentation of Results

RESULTS

As shown in **Figure 1**, the likelihood-based approach performed as well or better than the best available alternative methods in all the simulated cases, is applicable to other situations.

Figure 1. Performance Evaluation

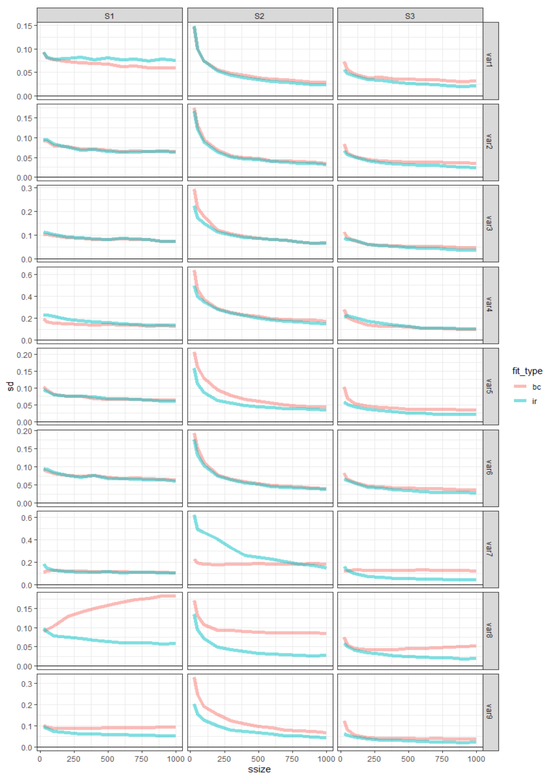


Figure 2 illustrates the presentation of results using the interval regression approach.

CONCLUSIONS

The likelihood-based distribution estimation procedure outperforms methods currently used in meta and network-meta-analyses and allows for estimation of underlying distributions in cases not previously supported. Application of this method will allow inclusion of previously excluded studies in evidence synthesis.

REFERENCES

1. Luo, Dehui, Xiang Wan, Jiming Liu, and Tiejun Tong. "Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range." *Statistical methods in medical research* 27, no. 6 (2018): 1785-1805.
2. Wan, Xiang, Wenqian Wang, Jiming Liu, and Tiejun Tong. "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.
3. McGrath, S., Zhao, X., Steele, R., Thombs, B.D., Benedetti, A. and 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), pp.2520-2537.

Interested in exploring more about our IR method?
Test it through our interactive web application.
Scan the QR code or visit:
https://apps.mathsinhealth.com/IRMethod