Deep Meta Tool: GUI tool to obtain Mean and Standard Deviation (SD) from Median and Interquartile range (IQR)

Deepanshu Sharma
Amity University Rajasthan https://orcid.org/0000-0003-1862-0606
Surya Priya Ulaganathan
Amity University Rajasthan
Vinay Sharma
Amity University Rajasthan
Sakshi Piplani
Amity University Rajasthan
Ravi Ranjan Kumar Niraj (rrkniraj@gmail.com)
Amity University Rajasthan

Research Article

Keywords: Mean, Median, Interquartile range (IQR), Standard Deviation (SD), Metaanalysis

DOI: https://doi.org/10.21203/rs.3.rs-828102/v1

Abstract

Background and objectives

Meta-analysis is a statistical procedure which enables the researcher to integrate the results of various studies that were conducted for the same purpose. However, more often than not, researchers find themselves in a position unable to proceed further due to the complexity of the mathematics involved and unavailability of raw data. To alleviate the said difficulty, we are presenting a tool that will enable researchers to process raw data.

Methods

The GUI tool is written in python. The tool offers an automated conversion and obtainment of mean and standard deviation (SD) from median and interquartile range, utilizing the methods offered by Hozo et al. 2005 and Bland 2015.

Results

The tool is tested on some sample data and validation is performed for Bland method on the data provided in the Bland method publication (14).

Conclusions

The provided tool is an easy alternative for the preparation of input data required for clinical meta-analysis in the required format.

1. Introduction

Meta-analysis is a systematic approach incorporating unequivocally concrete mathematics that gives out overall estimates along with measurable precision, inter-relating and combining the numerous studies of the same objectives, without the reviewer's bias. It provides a standard modus operandi to enable and maintain a uniform threshold and benefit an effective process to establish statistically significant results (3, 5). Starting from the period around 1904 when Pearson introduced data 'pooling' in his studies, the definite term of 'meta-analysis' was coined and introduced in 1976 (6). Prior to the foundation of meta-analysis, the two major routines of elucidating the conclusions were either through anecdotal voting or via finding the source of 'error' in the antagonistically deduced studies (9).

For delving into the course of action of statistically amalgamating research summaries, one must have the knowledge of the raw data or the defining parameters of the reported data. Irrespective of the analytical model to be applied, the researcher must be aware of the statistical terms Mean and Standard Deviation (SD). However, on a practical note, it has been observed and reported that several research studies do not disclose these parameters. In many cases, the aforesaid terms are elusive (7, 8, and 10). This conundrum, presented in the early data processing and analysing, is overcome using various methods that derive these variables. Buck's method for estimating mean and standard error of mean is explained by Piggott, T. D. (1994) (13). Review by Wiebe et al. explains multiple imputations ranging from algebraic calculations, study-level imputations and no-imputation or exclusion cases. The review has also alongside reported the shortcomings. Additionally, the review has discussed the scenarios of where the respected method would be appropriate, either as missing completely at random, missing at random or non-ignorable missing data. It recommends multiple imputation as compared against single or nonparametric imputation (5). Another majorly explored method is using Bayesian models, among which Kwon et. al have proposed and explained the Approximate Bayesian Computation (ABC). The method utilizes an inherent function to statistically compile the other known parameters and simulates the data for estimating mean and SD (2, 12).

Nonetheless, it has not been brought to our comprehension of literature for the existence of an ingenious and straightforward easy-to-use tool that would implement statistics for the data ascription. As aforementioned, in order for the research studies to be qualified for the inclusion of meta-analysis, evaluation of missing data is essential (11). Hence, the stated problem paves way for the motivation of the formulation of our software Deep Meta Tool. This toolbox, first of its kind, will be suitable for deriving the mean and standard deviation from the existing parameters of continuous data reported by the researcher. The methods adopted for the same are proposed by Hozo et. al (2005) and Bland (2015).

Hozo et. al (2005) have presented a simple yet competent approach of determining the mean and SD by employing the media, size of the data, maximum and minimum. Depending on the sample size, Hozo et. al provide with alternate formula for deducing mean and SD. The study also makes use of various simulations under differed circumstances of data distributions to validate the procedure. Under Normal distribution, the formula gives a very close estimate of within 4% and for larger datasets; the median itself is an excellent estimator as the mean. The simulations for normal and skewed distribution laid the foundation for demonstrating the accurate variants of the formula for estimating mean and SD (1).

Published more recently, Bland (2015) takes the method by Hozo et. al a step forward by incorporating the first and third quartile along with the median, maximum and minimum and sample size. The scheme given by Bland takes up a few inequalities as assumptions and applied a few amendments to generate a new set of formulae. Similar simulation studies corroborate the suggested procedure and comparison between Hozo et. al (2005) and the extended version including quartiles by Bland (2015) states the conditions under which the former offers a better estimate than the latter and vice versa (14).

After meticulous reviewing of both the methods, the software was developed specifying the disparate stipulations which would yield the best valuations of mean and SD.

2. Methods

2.1. Problems and Background

Pooling data, an initial step in the process of meta-analysis, often requires the mean value and the SD. Withal, in many reported cases, the publisher only reveals the median, sample size, range and/or interquartile ranges (IQR) (1, 14). This poses a challenge to the researcher, who many at times might not clearly apprehend the underlying mathematics to derive the required values of mean and variation (or SD). Hence, the rationale behind the software development is to make this approach of deduction easier for non-statisticians.

Based on the data usually available on the scientific article, and the accuracy during different sample size and range, the toolbox has encompassed the two methods devised by Hozo et. al (2005) and Bland (2015). The former requires the input values to be median, size of sample, minimum and maximum of the range of data. It involves certain fundamental inequalities to be applied and infer the mean and SD, pertinent to all types of distribution of data. The final conceived formula for determining mean is as follows:

Formula 1. *Mean calculation by Hozo et al., method.*

$$Mean = \frac{a + 2\,m + b}{4}$$

**If sample size (n) exceeds 25, then median itself is the best estimator.*

We observe two contexts of data, where the respective estimate of mean is based on the sample size. Understood as when the value of n (sample size) is lesser than 25, then the approximated value of mean would be given by the formula (1) where "a" is smallest value of IQR (minimum), "b" is the largest value of IQR (maximum), and "m" is the median. Whereas, when the value of n exceeds 25, it has been appraised that the median itself is the best estimate [Formula 1] (1).

For the evaluation of SD, three formulae based on n (sample size) have been framed. It uses the same parameters of "a", "b" and "m". The formula given for determining variance (or SD) is shown below [Formula 2]:

Formula 2. *Standard Deviation calculation by Hozo et al., method.*

$$Standard\ Deviation = \begin{cases} \frac{1}{12}\left(\frac{(1-2m-b)^2}{4} + (b-a)^2\right)^{\frac{1}{2}} & \boldsymbol{n \le 15} \\ \frac{b-a}{4} & \boldsymbol{15 < n \le 70} \\ \frac{b-a}{6} & \boldsymbol{n > 70} \end{cases}$$

The latter method, given by Bland (2015) demands First quartile and Third quartile along with the median, sample size, minimum and maximum of the range of data as the input. It is an extended version of the method proposed by Hozo et. al. Similar to the original method, Bland (2015) also exploits the usage of inequalities for each observation. But here, using minimum, maximum and the three quartiles, the inequalities have been generated. The culminating formula is as follows [Formula 3]:

Formula 3. *Mean calculation by Bland method.*

$$Mean = \frac{a + 2q_1 + 2m + 2q_3 + b}{8}$$

We observe that the variables necessitated are "a", "q1", "m", "q3", and "b"; which are the minimum (smallest value of range), first quartile, median, third quartile and maximum (largest value of range) respectively. As for the evaluation of SD, using the same parameters, the formula is given below [Formula 4]:

Formula 4. *Standard Deviation calculation Bland method.*

$$\begin{aligned} Standard\ Deviation\ &= \frac{1}{16}\left(a^2 + 2q_1^2 + 2m^2 + 2q_3^2 + b^2\right) \\ &+ \frac{1}{8}\left(aq_1 + q_1 m + mq_3 + q_3 b\right) \\ &- \frac{1}{64}\left(a + 2q_1 + 2m + 2q_3 + b\right)^2 \end{aligned}$$

2.2. Software Framework

2.2.1 Software Architecture

Deep Meta Tool (DMT) is a tool designed to support data preparation for meta-analysis. The software provides a quick and user-friendly platform for the calculation of Mean and SD from median and IQR, essential to conduct a meta-analysis, combining two different statistical methods.

Our tool is written in Python (3.8.3). The base window and GUI features are given by Tkinter module. Pandas, OS and math module are also used to carry out the inner functioning of the tool.

We have used pyinstaller module to convert our python file to windows executable. Current version of our tool is supported on Windows and Linux operating systems. We are expecting user feedback to enhance *and modify the tool in newer versions. We would make it available for multiple operating systems including Android.*

2.2.2 Software Functionalities

DMT is an open-source, free software, after whose installation you would obtain the following files at the installation directory –

1. *Deep Meta Tool Version 1.0 folder, which has all the system files.*
2. *An application file to launch the software. (Make a desktop shortcut for ease.)*
3. *Sample_Data.csv file, which is useful to understand input data format required for the Median Calculator section.*
4. *A User Manual.pdf.*
5. *All other are system files required to run the software.*

As elaborated in the previous sections, the software presents both the methods on its user interface (displayed in Fig. 1). The first option of the software utilizes Hozo et al., 2005 method to calculate and the data format required as input for the prediction is Median (IQR). The software enables a very straight-forward functionality demanding inputs required for the formula such as IQR minimum and maximum, median, etc. and a clear “Calculate Mean” and “Calculate SD” button for the generation of output as well.

The data collection and processing prior to meta-analysis involves various datasets. Conceiving this precondition, the software includes a “Store Value” option which enables the user to store their calculated mean and SD data. An Excel sheet compiling all the computed and stored values as the output file named as "Hozo_Method_Output.csv" by default, gets automatically saved in the same directory as the software once “Generate Output” is clicked. It is to be noted that no value must be left empty, and the user must be aware of the sample size and utilize the corresponding formula accordingly.

As for the Bland method, the software has incorporated another input category of Raw Data. DMT is pre-programmed to extract the required parameters for the enumeration of SD and mean, from the Raw Data input and directly inputted into the software. The procured parameters can be found under the name "Bland_Method_Parameters.txt" in the DTM Version 1.0 folder. A similar functionality of storing values and generating an output file is in-built in Bland method too, stored by default as "Bland_Method_Output.csv". It has been observed that the bigger the sample size, more accurate is the estimation of mean and SD. The format of the raw data is as shown in Table 1. and must be named as MEDIAN_DATA.

The application has a “About author” section to exhibit a QR code accessing the author’s website for contact, report a bug or feedback. On the Help toolbar, a detailed guide to usage of software is specified.

MEDIAN_DATA
2.85
2.85
2.98
3.04
3.1
3.1
3.9
3.3
3.54

Table 1. *Sample data format for bland method.*

3. Results

3.1. Implementation

The application is programmed in such a way that an uncomplicated manner of calculation is feasible. The implementation of the software using the above-mentioned functionalities is simplified to a maximum and the toolbox comes along with a guide to supplement the working procedure.

The estimated data can be put into the Meta-analysis datasheet format (provided with the software in the Sample Files directory) for RStudio/MetaXL. MetaXL in an excel add-in to perform meta-analysis. The user can also execute the R script offered in the same folder to lead with meta-analysis on the data and generate forest and funnel plot by R.

The empirical evidence pertinent to the proper functioning of DMT (Bland method) is shown in Table 2 and 3. The dataset employed for the same was acquired from the Bland 2015 (14) paper. The raw data was utilized to extract the required variables and mean & SD were estimated. The results, when compared to the manual method proposed in the original paper, are shown to be congruous.

Parameters	In paper	Through DMT
Minimum	2.85	2.85
First Quartile	3.54	3.54
Median	4.1	4.1
Third Quartile	4.5	4.529
Maximum	5.43	5.43
Mean	**4.07**	**4.077**
Standard Deviation	**0.68**	**0.6547**

Table 2. *Bland method data-1comparison*

Parameters	In paper	Through DMT
Minimum	0.15	0.15
First Quartile	0.35	0.35
Median	0.46	0.46
Third Quartile	0.6	0.6
Maximum	1.66	1.66
Mean	**0.58**	**0.57875**
Standard Deviation	**0.34**	**0.3333611**

Table 3. *Bland method data-2 comparison*

3.2. Illustrative Examples

The toolbox can be well comprehended by the visualization of the workflow given in Fig. 2. And for further cognizance of the software, two arbitrary examples, one for each method included, is explained. Fig. 3 shows Hozo method and uses the illustrative dataset comprising of a total of 10 values, median 4.2, minimum and maximum of 3.6 and 5.8 respectively to obtain the mean and SD as 4.45 and 0.65. By the same token, Fig. 4 demonstrates the Bland method integrated in the application. This instance explicates how the software utilizes raw data and drawn-out parameters and estimates the mean and SD to be as 3.16 and 0.336 respectively. Both these examples, when performed in actuality, will generate designated output files in Excel sheets containing two columns, for mean and SD, in the order of the stored values.

4. Discussion

The working theme during the development of the toolbox resonated with the idea and purpose of equipping researchers belonging to any background to operate and accomplish data pooling and perform meta-analysis.

This robust application will ensure a quick and hassle-free initial processing of raw data to abate the cumbersome calculations and overcome inconsistencies of human error. As very commonly and unfortunately discerned, most of the research papers do not impart the exact values of crucial parameters such as mean and SD, which are of utmost importance to get going with the meta-analysis. The frequently given variables are mean, IQR, sample size, etc. which when applied with the befitting formulae can yield desired estimates. The motive of DMT is to enable this appropriate assessment of data to become easy and simple. The habitual indecision amongst especially non-mathematical researchers of opting the formulae or derivations of data can lead to miscalculations and disparities among data ultimately eliciting an erroneous conclusion or cause-and-effect relationship. The complicated equations and statistics could mentally wear out the researcher resulting in inaccuracies. Since this is part of the preliminary step, an error in this stage would amplify to great extents at the culmination of the meta-analysis. The subjective biases are also eradicated with the automated projection of mean and SD. The elegant and elementary interface and guide to operate permits an audience of various background including students to unravel the basics of data processing and subsequent meta-analysis.

While the current version of the toolbox encompasses two methods of enumeration, the feedback will be deployed to expand into other methods and include other degrees of meta-analysis and amend any practical difficulties.

DMT provides a user-friendly interfaced application to enable data pooling, an essential phase of meta-analysis. This paper presents what, and why of the toolbox and how it shall impact the broad spectrum of users. The software has made use of two popular and accurate procedures of estimation given by Hozo et. al (1) and Bland (14) and facilitate users to run the program time-effectively without having to learn the underlying ramifications nor manually performing the statistics. Along with the parameter extraction, the obtained data can be directly implemented to RStudio/ MetaXL and further conduct meta-analysis. The upcoming related labour will be dedicated to the improved development of the software, in addition, extending into other levels of meta-analysis.

Declarations

Funding: We do not received any funding for this work.

Conflicts of interest: The authors declare that they have no known conflict of interest, financial or personal, in any way influencing the work of this paper.

Availability of data and material: Yes

Software availability: Yes

Authors' contributions: RRKN conceived the idea; DS & SPU executed the work, VS & SP manuscript editing and analysing results.

Ethics approval: Not applicable

Consent to participate: All authors have their consent for publication.

References

1. Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Medical Research Methodology. 2005;5:1–10. doi:10.1186/1471-2288-5-13
2. Kwon D, Reis IM. Simulation-based estimation of mean and standard deviation for meta-analysis via Approximate Bayesian Computation (ABC). BMC Medical Research Methodology. 2015;15(1):1–12. doi:10.1186/s12874-015-0055-5
3. Wan X, Wang W, Liu J, Tong T. Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. BMC Medical Research Methodology. 2014;14(1):1–13. doi:10.1186/1471-2288-14-135
4. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled clinical trials. 1986; Sep 1;7(3):177 – 88.
5. Wiebe N, Vandermeer B, Platt RW, Klassen TP, Moher D, Barrowman NJ. A systematic review identifies a lack of standardization in methods for handling missing variance data. Journal of clinical epidemiology. 2006 Apr 1;59(4):342 – 53.
6. Jones DR. Meta-analysis: weighing the evidence. Statistics in medicine. 1995 Jan 30;14(2):137–49.
7. Weir, C.J., Butcher, I., Assi, V. *et al.* Dealing with missing standard deviation and mean values in meta-analysis of continuous outcomes: a systematic review. *BMC Med Res Methodol* 18, 25 (2018). https://doi.org/10.1186/s12874-018-0483-0
8. Bond Jr CF, Wiitala WL, Richard FD. Meta-analysis of raw mean differences. Psychological Methods. 2003 Dec;8(4):406.
9. Schmidt FL. What do data really mean? Research findings, meta-analysis, and cumulative knowledge in psychology. American psychologist. 1992 Oct;47(10):1173.
10. Streiner, D. L., &Joffe, R. (1998). The Adequacy of Reporting Randomized, Controlled Trials in the Evaluation of Antidepressants. *The Canadian Journal of Psychiatry*, *43*(10), 1026–1030.

https://doi.org/10.1177/070674379804301008

11. Hartung J, Knapp G, Sinha BK. Statistical meta-analysis with applications. John Wiley & Sons; 2011 Sep 20.
12. Stangl DK, Berry DA. Meta-analysis: past and present challenges. Marcel Dekker, New York; 2000 Apr 20.
13. Pigott TD. Methods for handling missing data in research synthesis. The handbook of research synthesis. 1994;4(10).
14. Bland M. Estimating mean and standard deviation from the sample size, three quartiles, minimum, and maximum. International Journal of Statistics in Medical Research. 2014 Jan 27;4(1):57–64.

Metadata

***Table 4**. Software metadata.*

Current software version	*1.0*
Permanent link to executables of this version	*https://github.com/DeepanshuSharma-BNB/Deep_Meta_Tool*
Computing platform / Operating System	*Windows*
Installation requirements & dependencies	*Turn off Windows firewall/antivirus software or provide admin access to Deep Meta Tool Version 1.0.exe while installation*
If available Link to user manual - if formally published include a reference to the publication in the reference list	*https://github.com/DeepanshuSharma-BNB/Deep_Meta_Tool*
Support email for questions	*deepanshu.bnb@gmail.com*

Figures

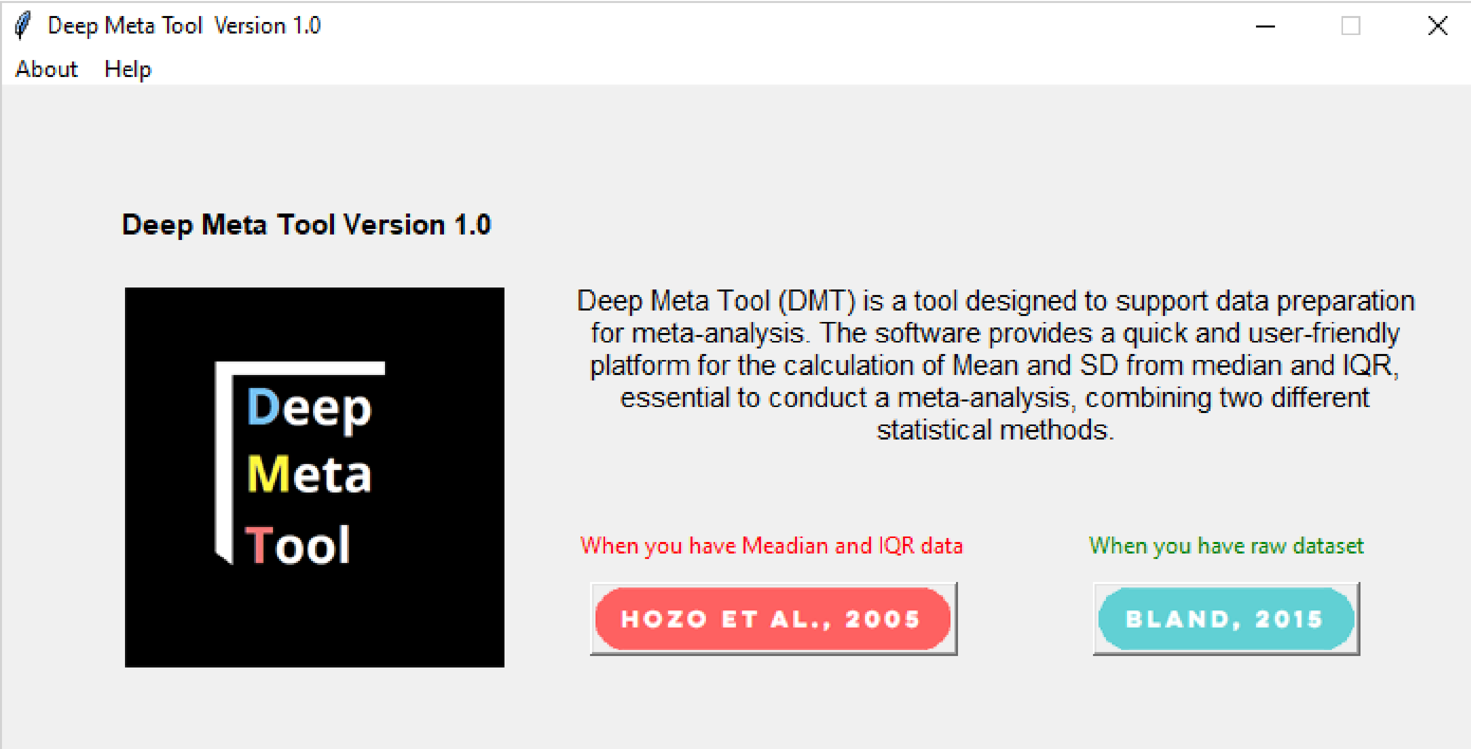


Figure 1

Deep Meta Tool version 1.0 main screen.

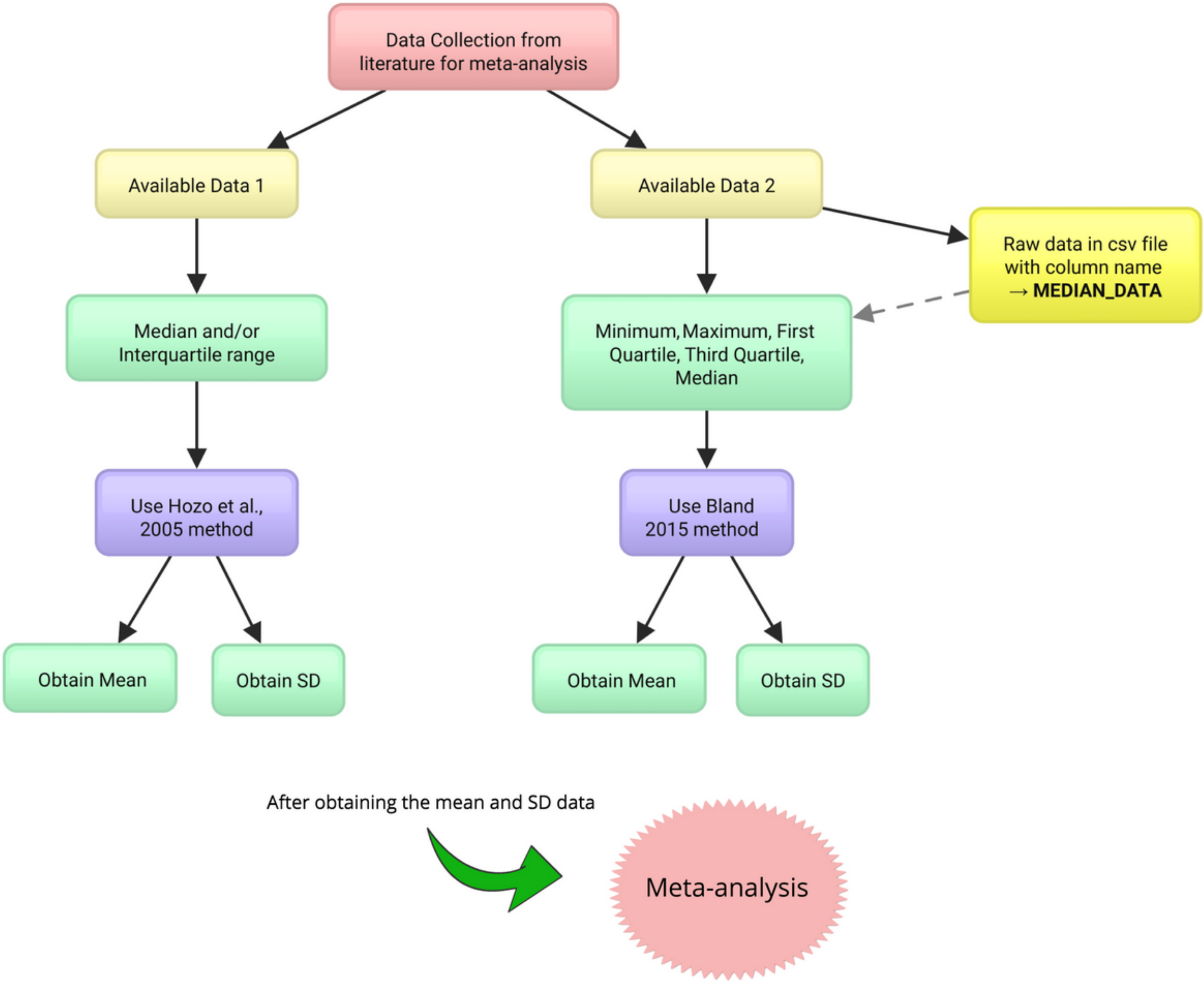


Figure 2

Flowchart of the process.

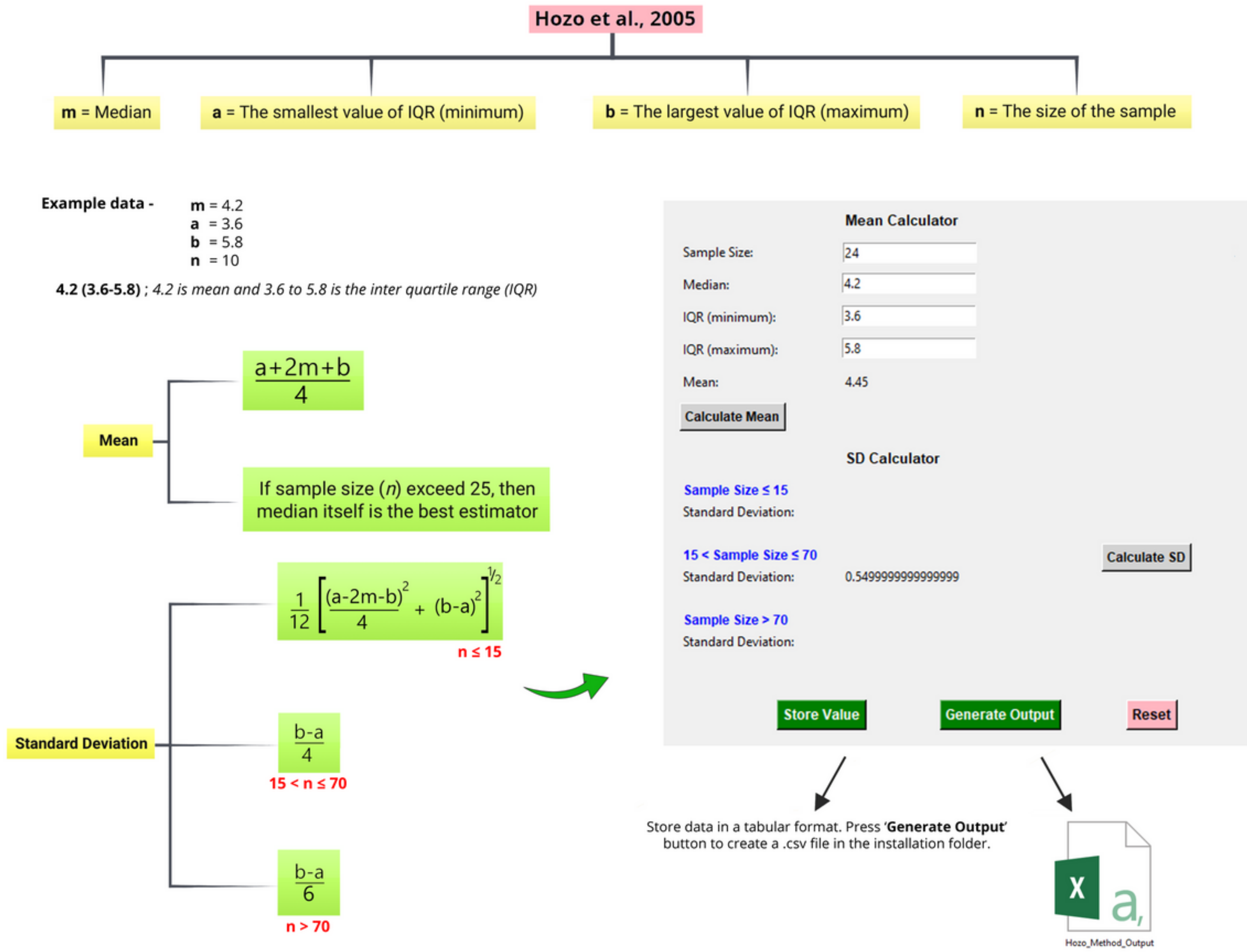
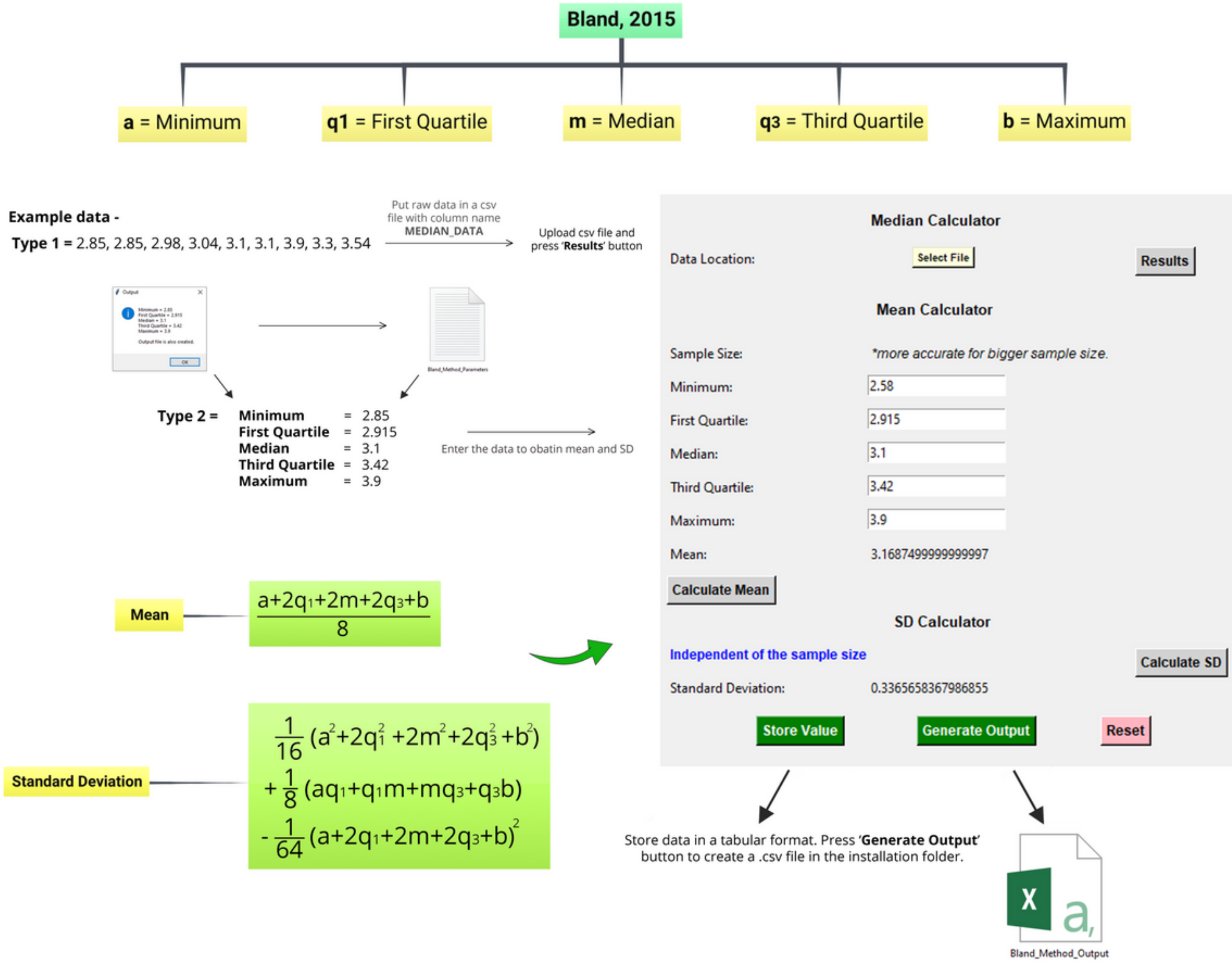


Figure 3

An illustrative example displaying the working of Hozo method.



$$\text{Mean} = \frac{a+2q_1+2m+2q_3+b}{8}$$

$$\text{Standard Deviation} = \frac{1}{16}(a^2+2q_1^2+2m^2+2q_3^2+b^2) + \frac{1}{8}(aq_1+q_1m+mq_3+q_3b) - \frac{1}{64}(a+2q_1+2m+2q_3+b)^2$$

Figure 4

An illustrative example displaying the working of Bland method.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Software.zip