Meta-analysis of Medians

Charles T. Gray, Luke Prendergast, Emily Kothe, and Hien Nguyen*

January 3, 2020

Abstract

1 Medians pose a problem in meta-analyses

Software tools for meta-analysis, such as Cochrane's RevMan [12], newly superseded by the cloud-based RevMan Web [13] or the R package `metafor::` [3], require estimates of both effect and variance of that effect. However, the sample variance for the reported effect of interest is not always available. When the reported statistics are medians, measure of spreads commonly provided are in the form of quartiles, as opposed to the required variance of the effect of interest. This leads to the omission of studies that report medians from the meta-anaysis. In this manuscript we present a method for estimating the variance of the sample median so that studies reporting medians may be included in meta-analyses.

This manuscript is a component of the research compendium created to solve this problem. In this case, the research compendium comprises not just a manuscript, but a pair of software packages, `varameta::` [11], which translates estimators presented in this manuscript to code, and `simeta::` [10], for simulating meta-analysis data, to see how the estimators in `varameta::` perform. Thus, in addition to intrinsic questions regarding meta-analysing medians, this manuscript considers on the ideas presented in the companion papers [8, 9] that ruminate, with this analysis as a case study, on why and how we may build reproducible research compendia.

1.1 What's the problem?

In published research, skewed data is often summarised by reporting the median and interquartile range. While this may be useful in a descriptive single-study sense, the lack of reported estimator variability poses a challenge in the context

*The authors are appreciative for the insights and comments from Kerrie Mengersen, and Kate Smith-Miles.

of meta-analysis. Software for performing meta-analyses, such as the widely-used `R` package `metafor` [3], require an estimate of the variance of the reported effects to conduct the meta-analysis under the assumed model

$$\widehat{\delta}_k = \delta + \gamma_k + \varepsilon_k \tag{1}$$

where $\widehat{\delta}_k$ is the estimated effect from the kth study, δ is the population effect of interest, ε_k is the error allowing for sampling variability and $\gamma_k \sim N(0, \tau^2)$ is the random effect to allow for differences in the true effects between studies. Given the estimated effects for K studies, all assumed to be normally distributed with a known (or estimated) variance, a meta-analysis can be carried out to estimate δ and the random effect variance τ^2. Our focus is on meta-analysis of three different effects involving the median. The first is simply the median itself when there is only one group of interest in each study. The second is the difference of two medians when there is two groups to be compared within each study (such as a case and control group). The third, which may be more suitable than the difference in medians when measurements of scale differ between studies, is the ratio of medians. For more on meta-analysis see, e.g., [**borenstein2008introduction**] and [**kulinskaya2008meta**].

1.2 Why propose a new method?

In this paper we propose a method for meta-analyses of studies whose effects are reported in the form of median and interquartile range (or range). The previously proposed method of [1], and extensions by [5, 6], solve this problem by estimating the mean and standard deviation from the provided summary statistics. For some applications there may be two noticeable drawbacks to this approach.

Disadvantage 1 *The methods to convert to a mean and standard deviation perform well when the underlying distribution is symmetric, and in some cases more specifically when it is normal. However, results have shown that performance can be poor in the presence of skew (e.g., see [7]).*

Disadvantage 2 *Those initially who published the summary measures, may have chosen to report medians and ranges because they had decided that moment-based measures such as the mean were not suitable descriptors.*

In the presence of underlying skewed distributions, both Disadvantages 1 and 2 may cause a real threat to the validity of any inference following conversion from

do this later find Disadvantages and work into text

To illustrate this problem, we begin with an example meta-analysis from medical research. We then briefly touch on how our method contributes to the existing solutions for this problem. In Section **??**, we define our estimator for the variance of the sample median and consider alternatives. We show how this estimator can be used in meta-analysis, and extend to meta-regression if the difference between the means and medians are of interest. Simulation results

are provided in Section **??** that assess the performance of our estimator in both the single-study and meta-analysis setting. Finally, in Section 6.1, we return to the motivating example, discussed in Section 2, to demonstrate how our method can be applied. Concluding remarks are provided in Section **??**.

2 A motivating example

To motivate our method, we provide an example of the variety of summary statistics that can arise in meta-analyses. We shall return to this example in Section 6.1 to see how our method facilitates a meta-regression of all studies, rather than just the three studies originally considered which were based on means and standard deviations.

We choose notations similar to those used by Wan *et al.* [6]. Define: a, the minimum value; q_1, the first quartile; m, the median; q_3, the third quartile; b, the maximum value; n, the sample size. iqr denotes the interquantile range and this may be reported as an interval, i.e. (q_1, q_3), or a width, i.e. $q_3 - q_1$. We also let $\overline{x}$ and s denote the sample mean and sample standard deviation respectively. Later, we subscript such measures appropriately to identify different groups within studies, and in the example below this equates to, e.g., m_c denoting the median for a control group and m_t for the treatment group.

As an example of how studies collated in meta-analyses report these summary statistics differently, consider the dataset presented in Table 1, taken from a systematic review of d-dimer in pre-eclampsia [4]. Ideally, one would want to perform a meta-analysis using the effects from all studies. However, estimator variance is only reported for three of the seven studies presented with the remaining studies reporting medians and ranges.

		Control group			Treatment group			
study	year	location	scale	n_c	location	scale	n_t	**reported**
Dusse	2003	1146.6	311.2	28	1263.8	411.9	43	$\boldsymbol{\overline{x}, s}$
Schjtlein	1997	1390.0	559.0	97	1545.0	849.5	200	$\boldsymbol{\overline{x}, s}$
Terao	1991	221.52	179.9	80	347.87	460.5	13	$\boldsymbol{\overline{x}, s}$
Catarino	2008	538.2	(391.2, 822.8)	42	448.5	(313.0, 1091.3)	44	$\boldsymbol{m, (q_1, q_3)}$
Bellart	1998	545.0	225.0	65	2090.0	1800.0	12	$\boldsymbol{m}$, **iqr**
Heilmann	2007	1149.0	456.0	33	1623.6	932.9	111	$\boldsymbol{m}$, **iqr**
He	1997	183.0	(110.0, 340.0)	24	315.0	(145.0, 1150.0)	30	$\boldsymbol{m, (a, b)}$

Table 1: Data from a meta-analysis of d-dimer levels in pre-eclampsia presented by [4], measures of location and scale are varied: there are means and standard deviations; medians and interquartile ranges; quartiles; and medians and ranges. The types of estimates reported are listed in the final column denoted 'reported'.

Three studies (first authors Dusse, Schjtlein, and Terao) detailed in the table provide the sample mean, $\overline{x}$, and standard deviation, s. Two studies (Bellart and Heilmann) provide the sample median, m, and interquartile range, iqr.

One study (Catrino) provides the sample median, as well as the first and third quartiles, q_1 and q_3. Finally, one more study (He) provides the sample median and the minimum a and maximum b observed values. All studies provide the sample size n and their respective estimates of location and scale for both the control and the pre-eclamptic groups.

In order to perform a meta-analysis via conventional methods, we require, at minimum, the studies' effect estimates, associated variances, and sample sizes. While full access to the raw data of each study would enable researchers to calculate the necessary sample variance for each study, there are many practical reasons, such as the time it would take to gather the data, that reduce the practicality of this approach, a point that is well made by others [5, e.g., p. 57]. Since only three studies presented in Table 1 report sample variance, Pinheiro *et al.*'s meta-analysis was restricted to these three datasets [4]. This paper provides a method that allows meta-analyses to be performed over studies reporting a variety of summary statistics, such as those outlined in Table 1.

3 Existing solutions to this problem

A potential solution is offered by Hozo *et al.* [1], who suggest estimating the mean and standard deviation from a reported median, minimum and maximum, as well as sample size, i.e from $C_1 := \{a, m, b; n\}$. This provides a way to calculate the variance of the effect, as required by contemporary meta-analysis tools, although there are some limitations. Firstly, C_1 does not cover all cases of reported medians. In our example considered in Table 1 to see that there is only applicable study (He).

Bland extends on Hozo *et al.*'s solution, but for the set $C_2 := \{a, q_1, m, q_2, b; n\}$ where the minimum, maximum, median, as well as first and third quartiles are reported[5]. Wan *et al.* improve on Hozo and Bland's solutions, as well as providing a solution for the set $C_3 := \{q_1, m, q_3; n\}$ where the interquartile range is provided as an interval along with the median[6]. A nice review of the methods, including an improvement, can be found in [7]. However, it is noted that underlying normality appears to be the motivation for all methods and below we details some limitations.

Firstly, note that the summary statistics sets $\{a, m, b; n\}$, $\{a, q_1, m, q_2, b; n\}$, and $\{q_1, m, q_3; n\}$ do not cover all of the presentations of summary statistics seen in Table 1. Thus, even if Pinheiro *et al.* had access to all methods, the meta-analysers would still have work ahead of them to include all studies presented here.

do this later Flip the fractions in Table to the a/b format Hien said to use. Will also be easier to read.

Secondly, and more importantly, to convert medians and interquartile ranges (or ranges) to means and standard deviations ignores the implicit information conveyed by the reported summary statistics; that is, that the study's authors perceived an asymmetry in the data. Our motivation is to provide a solution that enables meta-analyses to retain this information and, in addition, provide a method of comparing the studies that reported means with the studies that

Source	$\overline{X}$	S	C
Hozo [1]	$\frac{a+2m+b}{4}$	$S \approx \begin{cases} \frac{1}{\sqrt{12}}\left[(b-a)^2 + \frac{(a-2m+b)^2}{4}\right]^{\frac{1}{2}} & n \leqslant 15 \\ \frac{b-a}{4} & 15 < n \leqslant 70 \\ \frac{b-a}{6} & n > 70. \end{cases}$	$C_1 = \{a, m, b; n\}$
Bland [5]	$\frac{a+2q_1+2m+2q_3+b}{8}$	$[\frac{1}{16}(a^2+2q_1^2+2m^2+2q_3+b^2)+ \frac{1}{8}(aq_1+q_1m+mq_3+q_3b)- \frac{1}{64}(q+2q_1+2m+2q_3+b)^2]^{\frac{1}{2}}$	$C_2 = \{a, q_1, m, q_3, b; n\}$
Wan [6]	$\frac{a+2m+b}{4}$	$\frac{b-a}{2\Phi^{-1}\left(\frac{n-0.375}{n}+0.25\right)}$	$C_1 = \{a, m, b; n\}$
Wan [6]	$\frac{a+2q_1+2m+2q_3+b}{8}$	$\frac{b-a}{4\Phi^{-1}\left(\frac{n-0.375}{n+0.25}\right)} + \frac{q_3-q_1}{4\Phi^{-1}\left(\frac{0.75n-0.1215}{n+0.25}\right)}$	$C_2 = \{a, q_1, m, q_3, b; n\}$
Wan [6]	$\frac{q_1+m+q_3}{3}$	$\frac{q_3-q_1}{2\Phi^{-1}\left(\frac{0.75n-0.125}{n+0.25}\right)}$	$C_3 = \{q_1, m, q_3; n\}$

Table 2: This table is a rephrasing and detail of Table 3 from Wan *et al.* [6], and guides the simulation discussed in Section **??**. Here we have details of various estimators for sample mean and sample standard deviation, for different data sets. These estimators are defined in terms of sample summary statistics: minimum a, maximum b, median m, first quartile q_1, third quartile q_3, and sample size n. Source indicates the first author of the paper the equations are found in. The sample mean and sample standard deviation estimators are presented by columns $\overline{X}$ and S, respectively. The column C presents the sample summary statistics required as parameters in the coupled estimators.

reported medians.

4 Estimating the variance of the sample median

We first focus on providing expressions for the variance of a single median, a difference in two independent median estimators and the log ratio of two medians.

Consider a population median denoted ν with corresponding estimator M, taken to be the middle order statistic from a sample with n observations. Let f denote the probability density function for the underlying population. Then the median estimator, M, is asymptotically normal with approximate variance (see, e.g. Ch.7 of [2])

$$\operatorname{var}(M) \approx \frac{1}{n} \cdot \frac{1}{4\left[f(\nu)\right]^2}. \tag{2}$$

Using this approximated variance, we can then extend to the variance of the difference and the variance of the ratio of two sample medians. For the difference of two sample medians, we have, assuming that the estimators are independent,

$$\operatorname{var}(M_1 - M_2) = \operatorname{var}(M_1) + \operatorname{var}(M_2). \tag{3}$$

Using the delta method, the variance of the log ratio of two sample medians is given by

$$\operatorname{var}\left[\log\left(\frac{M_1}{M_2}\right)\right] \approx \frac{\operatorname{var}(M_1)}{\nu_1^2} + \frac{\operatorname{var}(M_2)}{\nu_2^2}. \tag{4}$$

In practice we do not know the true population median ν, nor the true population density f, so estimates are required . It is common to only have access to the sample median and interquartile range (or range) from a single study. Or, in the case of the comparison of two samples, we may have two sample medians and associated interquartile ranges (or ranges).

However, as we shall explore in Section 4.2, both the log-normal and the normal densities provide surprisingly close approximations of the true densities evaluated at the median. In this paper, we propose the following adaptation of equation (2)

$$\mathrm{v}(M) := \frac{1}{4n\left[g\left(M;\hat{\boldsymbol{\theta}}\right)\right]^2} \tag{5}$$

where g is a pre-specified density and $\hat{\boldsymbol{\theta}}$ is a vector of parameter estimates for g where the estimates arise from the limited information in the reported median and interquartile range (or range).

Remark 1 *The choice of g does not need to be similar to the true underlying distribution. Instead, it only need be close to the density evaluated at the median. It turns out that there are excellent choices for g that for appropriately chosen $\boldsymbol{\theta}$, $g(\nu;\boldsymbol{\theta}) \approx f(\nu)$ for a diverse range of densities, f.*

We now derive each of these parameter sets, for the normal, log-normal, exponential, and Cauchy distributions, before comparing the estimators derived in Section 4.2.

4.1 Approximating the variance of the median from limited information

Given that the true population density f of equation (2) is unknown, we propose replacing f with a nominated density g whose parameters are estimated from the information available. In doing so we obtain our approximated variance in (5) by choosing a suitable $g\left(M;\hat{\boldsymbol{\theta}}\right)$.

4.1.1 Using the normal distribution

For the normal density with parameters μ and σ, the quantile function is $G^{-1}(p) = \mu + \sigma\Phi^{-1}(p)$ where Φ is the standard normal cumulative distribution function. Using the symmetry of Φ and assuming that interquartile range has been reported, we know that the true interquartile range is given by $2\sigma\Phi^{-1}(0.25)$. Thus we have estimators $\hat{\mu} := M$ and

$$\hat{\sigma}^{(1)} := \frac{\mathrm{iqr}}{2\Phi^{-1}(0.25)}.$$

If the sample range is reported, as was the case with one study in the motivating example provided in Table 1 then we need a different estimate of σ. For $x_{[i]}$ denoting the ith order statistic for a sample of size n, $x_{[i]}$ is an estimate to approximately the $n^{-1}(i-0.5)$th population quantile. In particular, the maximum, or nth order statistic $x_{[n]}$, is an estimate to approximately the $[(n-0.5)/n]$th population quantile. Thus, by a similar argument, we have

$$\hat{\sigma}^{(2)} := \frac{x_{[n]} - x_{[1]}}{2\Phi^{-1}\left[(n-0.5)/n\right]}$$

where $x_{[n]} - x_{[1]}$ simply the reported range.

From above, if we choose the normal density for g, then $\hat{\boldsymbol{\theta}} = [\hat{\mu}, \hat{\sigma}^{(i)}]$ $(i = 1, 2)$ depending on whether the interquartile range or range is reported.

4.1.2 Using the log-normal distribution

If we were to choose the log-normal density with parameters μ and σ, then since the true median of a lognormal density is given by e^{μ}, we have an estimator for μ, given by

$$\hat{\mu} := \log(M).$$

We obtain our estimator for σ similarly when the interquartile range is reported. We know that the true interquartile range of the lognormal density is given by $G^{-1}\left(\frac{3}{4}\right) - G^{-1}\left(\frac{1}{4}\right)$ where G is the cumulative distribution function for the lognormal density. We have the associated quantile function

$$G^{-1}(p; \mu, \sigma) = \exp(\sigma\Phi^{-1}(p) + \mu).$$

Using this information, along with the symmetry of Φ and our estimate $\hat{\mu}$, we have

$$\hat{\sigma}^{(1)} := \frac{1}{\Phi^{-1}\left(\frac{3}{4}\right)} \log\left(\frac{\mathrm{iqr}\, e^{-\hat{\mu}} \pm \sqrt{\mathrm{iqr}^2\, e^{-2\hat{\mu}} + 4}}{2}\right).$$

By similar argument to the deriviations for the normal density in Section 4.1.1, if the range is reported then our estimate to σ is

$$\hat{\sigma}^{(2)} := \frac{1}{\Phi^{-1}\left(\frac{n-\frac{1}{2}}{n}\right)} \log\left[\frac{(x_{[n]} - x_{[1]})e^{-\hat{\mu}} \pm \sqrt{(x_{[n]} - x_{[1]})^2 e^{-2\hat{\mu}} + 4}}{2}\right].$$

Again, if we choose the log-normal density for g, then $\hat{\boldsymbol{\theta}} = [\hat{\mu}, \hat{\sigma}^{(i)}]$ $(i = 1, 2)$ depending on whether the interquartile range or range is reported.

4.1.3 Using the exponential distribution

For the exponential density, we need only estimate the rate parameter λ. Since the true median of the exponential density is given by $\log(2)/\lambda$, we can estimate $\hat{\lambda} := \log(2)/M$. Here, $\hat{\boldsymbol{\theta}}$ takes the single parameter estimate $\hat{\lambda}$.

4.1.4 Using the Cauchy distribution

We need to estimate two parameters for the Cauchy density: a location parameter η and a scale parameter θ. From the quantile function for the Cauchy distribution $G^{-1}(p) = \eta + \theta \tan[\pi\,(p - 0.5)]$, we know that the true median is the location parameter η and that the interquartile range is equal to 2θ. Hence, we can estimate $\hat{\eta} := M$ and if the interquartile range is reported $\hat{\theta}^{(1)} := \text{iqr}\,/2$. If the range is reported then similar to previous arguments, we can estimate

$$\hat{\theta}^{(2)} = \frac{x_{[n]} - x_{[1]}}{2\tan\left[\pi\left(\frac{n-0.5}{n} - \frac{1}{2}\right)\right]}.$$

When using the Cauchy, we then have $\hat{\boldsymbol{\theta}} = [\hat{\eta}, \hat{\theta}^{(i)}]$ $(i = 1, 2)$ depending on which range is reported.

4.2 Comparison between four choices of g

more words provide definition

5 Performance of estimator in coverage probability simulations

Now that we have defined an estimator for meta-analysing medians, let us explore the efficacy of this estimator under simulation, for different numbers of studies, distributions, and different assumptions about variation between studies and efficacy of intervention.

5.1 Simulation methodology

One approach for exploring the efficacy of a statistical estimator is to simulate *coverage probability*. In a coverage probability simulation, each trial requires randomly generated data.

more words Mathematics need to be incorporated.

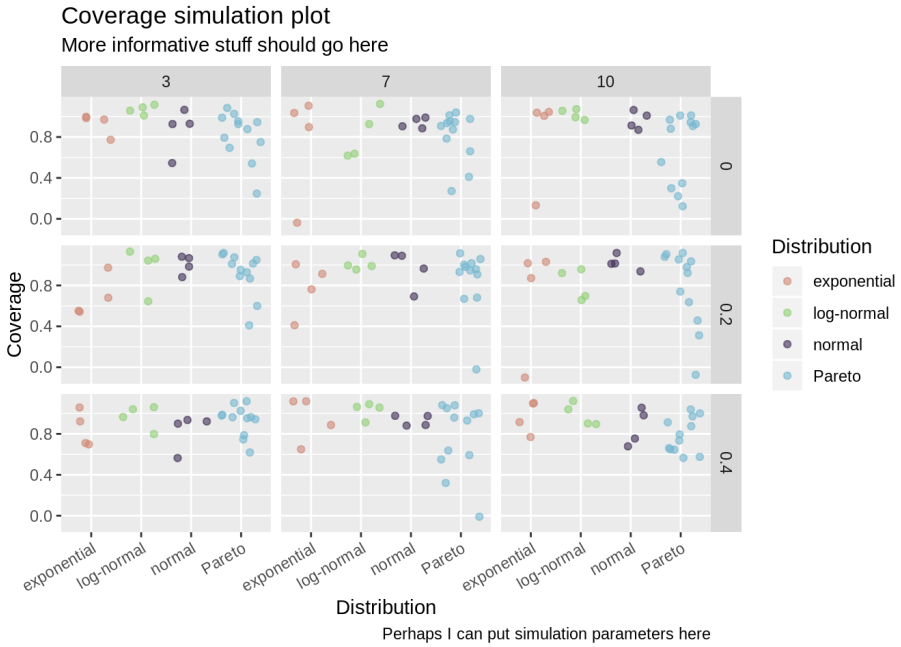

Figure 1: Simulation results.

$$\begin{aligned} & \log(m_k^I/m_k^C) && = \log(\nu^I/\nu^C) + \gamma_k \\ \implies & \log(\lambda_k^C) - \log(\lambda_k^I) && = \log(\lambda^C) - \log(\lambda^I) + \gamma_k \\ \implies & \log(\lambda_k^C) - \log(\lambda_k^I) = (\log(\lambda^C) + \gamma_k/2) - (\log(\lambda_k^I) - \gamma_k/2) \end{aligned}$$

$$\begin{aligned} \lambda_k^C &= \lambda^C \exp(\gamma_k/2) \\ \lambda_k^I &= \lambda^I \exp(-\gamma_k/2) \end{aligned}$$

more words See `simeta::` and `varameta:::`

5.2 simulation results

more words paused for writing tracking

Figure **??** presents results of

6 Meta-analysis of medians

more words Find a meta-analysis of medians where this estimator elicits a difference in results. Different example than tired old Pinheiro.

more words Go through each method: what set of summary statistics does each do? Look at table of equations in original overleaf, perhaps.

Our motivating problem was meta-analysis of medians, and we have followed the *toolchain walkthrough* for computationally developing a statistical estimator. This process took us from mathematical derivations, to estimator functions provided by *research compendium* `varameta::`, to *coverage probability* simulations provided by the package `simeta::` to explore the efficacy of this estimator under different sampling conditions. In so doing, this manuscript raises the question if examining research software engineering methodology, of exploring the efficacy of an estimator, in the context of rapidly evolving statistical tools for simulation and analysis, is of research merit in its own right. We begin by revisiting our motivating example, and turn to metaresearch observations from this analysis.

do this later footnote

do this later footnote

do this later footnote

6.1 Revisiting the motivating example

With a method for incorporating medians, we revisit the meta-analysis presented by Pinheiro *et al.* .

more words Figure: forest plot

more words paused for writing tracking

do this later Remember to replace the tired old Pinheiro example

6.2 Components of research for computational science

more words It is not immediately apparent what the best way to confer analyses.

more words metaresearch context

This manuscript derives mathematical estimators and presents simulation results of programmed instantiations of the algorithmic solution to the meta-analysis of medians. Rapid advances in the adoption software engineering provide an auxillary context of metaresearch; these are explored in companion manuscripts that describe the theoretical underpinnings of reproducible computing [9] and the practical steps in preparing this analysis as a reproducible research compendium [8]. Which is to say, metaresearch questions from this project have generated more products of research than the question of meta-analysing medians itself.

This manuscript is one product of a research question. However, against the backdrop of technological revolutions in data collection and code sharing, contemporary researchers face a constant challenge of upskilling in computational tools, in addition to the challenges of the discipline in which the researcher is working. This manuscript used two packages, `varameta::` and `simeta::` to perform analysis; this manuscript is but one research component of the compendium of research to assess estimators for meta-analysis of medians. This

splintered approach has the advantage that each component can exist for a different utility, individually, but together form a compendium of research.

Sharing and reflecting on how we do this can surely be of benefit to others, especially for those with the leisure to learn new techniques, such as graduate students.

References

[1] Stela Pudar Hozo, Benjamin Djulbegovic, and Iztok Hozo. "Estimating the Mean and Variance from the Median, Range, and the Size of a Sample". In: *BMC Medical Research Methodology* 5.1 (Apr. 2005), p. 13. ISSN: 1471-2288. DOI: 10.1186/1471-2288-5-13.

[2] Anirban DasGupta. *Asymptotic theory of statistics and probability*. Springer Science & Business Media, 2008.

[3] Wolfgang Viechtbauer. "Conducting Meta-Analyses in R with the metafor Package". In: *Journal of Statistical Software* 36.3 (2010), pp. 1–48.

[4] Melina de Barros Pinheiro et al. "D-Dimer in Preeclampsia: Systematic Review and Meta-Analysis". en. In: *Clinica Chimica Acta* 414 (Dec. 2012), pp. 166–170. ISSN: 0009-8981. DOI: 10.1016/j.cca.2012.08.003.

[5] Martin Bland. "Estimating Mean and Standard Deviation from the Sample Size, Three Quartiles, Minimum, and Maximum". en. In: *International Journal of Statistics in Medical Research* 4.1 (Jan. 2014), pp. 57-64-64. ISSN: 1929-6029.

[6] Xiang Wan et al. "Estimating the Sample Mean and Standard Deviation from the Sample Size, Median, Range and/or Interquartile Range". In: *BMC Medical Research Methodology* 14.1 (Dec. 2014), p. 135. ISSN: 1471-2288. DOI: 10.1186/1471-2288-14-135.

[7] Jiandong Shi et al. "How to Estimate the Sample Mean and Standard Deviation from the Five Number Summary?" In: *arXiv preprint arXiv:1801.01267* (2018).

[8] Charles T. Gray. "Code::Proof: Prepare for Most Weather Conditions". In: (2019). arXiv: 1910.06964 [stat.OT].

[9] Charles T. Gray and Ben Marwick. "Truth, Proof, and Reproducibility: There's No Counter-Attack for the Codeless". In: *arXiv:1907.05947 [math]* (July 2019). arXiv: 1907.05947 [math].

[10] Charles Gray. *Simeta: Simulate Meta-Analysis Data*. 2020.

[11] Charles Gray. *Varameta: Estimators for the Variance of the Sample Median*. 2020.

[12] Review Manager. *RevMan 5*.

[13] Review Manager. *RevMan Web*.