

# Meta-analysis of Medians

Charles T. Gray, Luke Prendergast, and Hien Nguyen\*

December 18, 2019

## Abstract

## 1 Medians pose a problem in meta-analyses

Software tools for meta-analysis, such as Cochrane’s RevMan or the R package `metafor::`, 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 the range or interquartile range. This leads to the omission of studies that report medians from the meta-analysis. 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.

do this later  
RevMan citation

do this later  
metafor citation

**more words** people would like to meta-analyse medians, but they can’t

**more words** expand this section, demonstrate how to do a meta-analysis

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

$$\hat{\delta}_k = \delta + \gamma_k + \varepsilon_k \quad (1)$$

where  $\hat{\delta}_k$  is the estimated effect from the  $k$ th study,  $\delta$  is the population effect of interest,  $\varepsilon_k$  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

---

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

a known (or estimated) variance, a meta-analysis can be carried out to estimate  $\delta$  and the random effect variance  $\tau^2$ . 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., [?] and [?].

## 1.2 A solution.

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 [?], and extensions by [?, ?], 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 [?]).*

**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 quantiles measures to moment-based measures. The method that we propose can be easily adapted to both single-study and meta-analysis contexts. In the case of meta-analyses, meta-regression may also be conducted when some of the studies report means and standard deviations and others report medians and interquartile ranges. This can be done by introducing a moderator factor variable to explain the differences between an effect based on means and an effect based on medians.

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 ??, we return to the motivating example, discussed in Section 2, to demonstrate how our method can be applied. Concluding remarks are provided in Section ??.

do this later  
find Disad-  
vantages and  
work into text

## 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 ?? 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. [?]. 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 interquartile 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  $\bar{x}$  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 [?]. 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.

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

Table 1: Data from a meta-analysis of d-dimer levels in pre-eclampsia presented by [?], 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,  $\bar{x}$ , and standard deviation,  $s$ . Two studies (Bellart and Heilmann) provide the sample median,  $m$ , and interquartile range,  $iqr$ . One study (Catarino) 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 [?, 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 [?]. 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. [Hozo2005] who suggest estimating the mean and standard deviation from a reported median, minimum and maximum, as well as sample size, i.e from  $S_1 = \{a, m, b; n\}$ . This provides a way to calculate the variance of the effect, as required by contemporary meta-analysis tools, although there are some limitations. Firstly,  $S_1$  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  $\{a, q_1, m, q_2, b; n\}$  where the minimum, maximum, median, as well as first and third quartiles are reported[?]. Wan et al. improve on Hozo and Bland’s solutions, as well as providing a solution for the set  $\{q_1, m, q_3; n\}$  where the interquartile range is provided as an interval along with the median[?]. A nice review of the methods, including an improvement, can be found in [?]. 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.

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 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  $\nu$  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 [?])

$$\text{var}(M) \approx \frac{1}{n} \cdot \frac{1}{4 [f(\nu)]^2}. \quad (2)$$

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,

$$\text{var}(M_1 - M_2) = \text{var}(M_1) + \text{var}(M_2). \quad (3)$$

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

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

In practice we do not know the true population median  $\nu$ , 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 ??, 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)

$$\text{varmed}(M) := \frac{1}{4n \left[ g \left( M; \hat{\theta} \right) \right]^2} \quad (5)$$

where  $g$  is a pre-specified density and  $\hat{\theta}$  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).

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  $\theta$ ,  $g(\nu; \theta) \approx f(\nu)$  for a diverse range of densities,  $f$ .

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

## 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

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

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.

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

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

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

## 5.2 simulation results

# 6 Meta-analysis of medians

Our motivating problem was meta-analysis of medians. 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.

**more words** Now we can walk through a meta-analysis

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

**more words** metaresearch context