BMC Medical Research Methodology

Research article

Estimating the mean and variance from the median, range, and the size of a sample

Stela Pudar Hozo[1], Benjamin Djulbegovic[2] and Iztok Hozo*[1]

Address: [1]Indiana University Northwest, Department of Mathematics, Gary, IN, 46408 USA and [2]Interdisciplinary Oncology Program, H. Lee Moffitt Cancer Center and Research Institute at the University of South Florida, Tampa, FL, USA

Email: Stela Pudar Hozo - spudar@iun.edu; Benjamin Djulbegovic - DjulbeBM@moffitt.usf.edu; Iztok Hozo* - ihozo@iun.edu

* Corresponding author

Published: 20 April 2005

BMC Medical Research Methodology 2005, **5**:13 doi:10.1186/1471-2288-5-13

This article is available from: http://www.biomedcentral.com/1471-2288/5/13

Received: 26 September 2004
Accepted: 20 April 2005

Abstract

Background: Usually the researchers performing meta-analysis of continuous outcomes from clinical trials need their mean value and the variance (or standard deviation) in order to pool data. However, sometimes the published reports of clinical trials only report the median, range and the size of the trial.

Methods: In this article we use simple and elementary inequalities and approximations in order to estimate the mean and the variance for such trials. Our estimation is distribution-free, i.e., it makes no assumption on the distribution of the underlying data.

Results: We found two simple formulas that estimate the mean using the values of the median (m), low and high end of the range (a and b, respectively), and n (the sample size). Using simulations, we show that median can be used to estimate mean when the sample size is larger than 25. For smaller samples our new formula, devised in this paper, should be used. We also estimated the variance of an unknown sample using the median, low and high end of the range, and the sample size. Our estimate is performing as the best estimate in our simulations for very small samples ($n \leq 15$). For moderately sized samples ($15 < n \leq 70$), our simulations show that the formula range/4 is the best estimator for the standard deviation (variance). For large samples ($n > 70$), the formula range/6 gives the best estimator for the standard deviation (variance).

We also include an illustrative example of the potential value of our method using reports from the Cochrane review on the role of erythropoietin in anemia due to malignancy.

Conclusion: Using these formulas, we hope to help meta-analysts use clinical trials in their analysis even when not all of the information is available and/or reported.

Background

To perform meta-analysis of continuous data, the meta-analysts need the mean value and the variance (or standard deviation) in order to pool data. However, sometimes, the published reports of clinical trials only report the median, range and the size of the trial. In this article we use simple and elementary inequalities in order to estimate the mean and the variance for such trials. Our estimation is distribution-free, i.e., it makes no assumption on the distribution of the underlying data. In fact, the value of our approximation(s) is in giving a method for estimating the mean and the variance exactly when there is no indication of the underlying distribution of the data. In current practice, the median is often substituted for the mean, and the Range/4 or Range/6 for the standard deviation. However, it has not been shown that median can indeed be used to replace mean values, nor when the range-formulas are appropriate.

Methods

Assumptions

Suppose a clinical trial reports the following summary measures for a certain event:

m = Median

a = The smallest value (minimum)

b = The largest value (maximum)

n = The size of the sample.

In this article, we want to estimate the mean, and the standard deviation of this sample of size n. First we will order this sample by size:

$a = x_1 \leq x_2 \leq x_3 \leq \ldots x_{M-1} \leq x_M = m \leq x_{M+1} \leq \ldots \leq x_{n-1} \leq x_n = b$,

where the M^{th} number is the median, and $M = \frac{n+1}{2}$ (for the sake of simplicity, we will assume that n is an odd number).

Results

Estimating the sample mean $\bar{x}$

We begin with several simple inequalities:

$$\begin{array}{l} a \leq x_1 = a \leq a \\ \left.\begin{array}{l} a \leq x_2 \leq m \\ a \leq x_3 \leq m \\ \vdots \\ a \leq x_{M-1} \leq m \end{array}\right\} (M-2)\_inequalities \\ m \leq x_M = m \leq m \\ \left.\begin{array}{l} m \leq x_{M+1} \leq b \\ m \leq x_{M+2} \leq b \\ \vdots \\ m \leq x_{n-1} \leq b \end{array}\right\} (M-2)\_inequalities \\ b \leq x_n = b \leq b \end{array} \qquad (1)$$

Adding up and diving by n, the middle column is exactly the sample mean, $\bar{x}$.

Adding up and diving by n for all three columns, we get the following inequality:

$$\frac{(M-1)a+(M-1)m+b}{n} \leq \bar{x} \leq \frac{a+(M-1)m+(M-1)b}{n}$$

After replacing $M = \frac{n+1}{2}$ into the inequality above, we get:

$$LB = \frac{\left(\frac{n-1}{2}\right)a+\left(\frac{n-1}{2}\right)m+b}{n} = \frac{a+m}{2}+\frac{b-\frac{a+m}{2}}{n}$$

$$\leq \bar{x} \leq$$

$$UB = \frac{a+\left(\frac{n-1}{2}\right)m+\left(\frac{n-1}{2}\right)b}{n} = \frac{m+b}{2}+\frac{a-\frac{m+b}{2}}{n}$$

Therefore, the lower bound for the sample mean is

$$LB = \frac{a+m}{2}+\frac{b-\frac{a+m}{2}}{n} \qquad (2)$$

The upper bound for the sample mean is

$$UB = \frac{m+b}{2}+\frac{a-\frac{m+b}{2}}{n} \qquad (3)$$

The sample mean can than be estimated as

$$\bar{x} \approx \frac{LB+UB}{2} = \frac{\frac{a+m}{2}+\frac{b-\frac{a+m}{2}}{n}+\frac{m+b}{2}+\frac{a-\frac{m+b}{2}}{n}}{2}, \text{ or}$$

$$\bar{x} \approx \frac{a+2m+b}{4}+\frac{a-2m+b}{4n} \qquad (4)$$

When the size of the sample is fairly large, the second fraction becomes negligible and the estimate can be written in a simplified form:

$$\bar{x} \approx \frac{a+2m+b}{4} \qquad (5)$$

We can use this simple expression even if we do not know the size of the sample. The length of the interval which contains the sample mean (the interval [LB, UB]), is approximately

$$l = \left(UB - LB \right) = \frac{n-3}{2n}\left(b - a \right)$$

Estimating the sample variance

Even when the only information we have about a set of data is it's range: $R = b - a$, we can still estimate the standard deviation. If our data are normally distributed, then $P[-2\sigma < X - \mu < 2\sigma] = 0.95$, and therefore, the range covers approximately 4σ, i.e., $\sigma \approx \frac{R}{4}$.

When the data we are dealing with are not normally distributed, we can still use the Chebyshev's inequality [1,2] $\left(P\left(\left|X-\mu\right| < k\sigma\right) \geq 1 - \frac{1}{k^2}\right)$, and obtain the following for $k = 3$: $P\left[-3\sigma < X - \mu < 3\sigma\right] \geq 1 - \frac{1}{9} = .89$. Therefore, the range covers approximately 6σ, i.e., $\sigma \approx \frac{R}{6}$.

On the other hand, if the summary results for a clinical trial include the median and the size of the sample, we can presumably do better than the two range approximations above. Next section deals with that situation.

The Variance S^2 – distribution free inequalities

Using the inequalities (1) and taking in consideration that all the data are non-negative, we can multiply each row i with the value x_i ($i = 1, 2, 3, ..., n$). We obtain the following inequalities:

$$\begin{array}{l}
ax_1 \leq {x_1}^2 = a^2 \leq ax_1 \\
\left.\begin{array}{l}
ax_2 \leq {x_2}^2 \leq mx_2 \\
ax_3 \leq {x_3}^2 \leq mx_3 \\
\vdots \\
ax_{M-1} \leq {x_{M-1}}^2 \leq mx_{M-1}
\end{array}\right\}\left(M-2 \right)\_inequalities \\
mx_M \leq {x_M}^2 = m^2 \leq mx_M \\
\left.\begin{array}{l}
mx_{M+1} \leq {x_{M+1}}^2 \leq bx_{M+1} \\
mx_{M+2} \leq {x_{M+2}}^2 \leq bx_{M+2} \\
\vdots \\
mx_{n-1} \leq {x_{n-1}}^2 \leq bx_{n-1}
\end{array}\right\}\left(M-2 \right)\_inequalities \\
bx_n \leq {x_n}^2 = b^2 \leq bx_n
\end{array} \qquad (6)$$

Adding up by columns, we have the following:

$$\begin{array}{l}
LB = a^2 + a\left(x_2 + x_3 + \cdots + x_{M-1} \right) + m^2 + m\left(x_{M+1} + x_{M+2} + \cdots + x_{n-1} \right) + b^2 \leq \\
\qquad \leq {x_1}^2 + {x_2}^2 + {x_3}^2 + \cdots + {x_{n-1}}^2 + {x_n}^2 \leq \\
\leq UB = a^2 + m\left(x_2 + x_3 + \cdots + x_{M-1} \right) + m^2 + b\left(x_{M+1} + x_{M+2} + \cdots + x_{n-1} \right) + b^2
\end{array} \qquad (7)$$

Using the inequalities (1) again, we estimate the sums in LB and UB as

$$\begin{array}{l}
x_2 + x_3 + \cdots + x_{M-1} \approx \left(M-2 \right)\frac{a+m}{2} \\
x_{M+1} + x_{M+2} + \cdots + x_{n-1} \approx \left(M-2 \right)\frac{m+b}{2}
\end{array} \qquad (8)$$

Therefore, the expressions in (7) can be estimated as

$$\begin{array}{l}
LB \approx a^2 + a\left(M-2 \right)\frac{a+m}{2} + m^2 + m\left(M-2 \right)\frac{m+b}{2} + b^2 \\
\quad = a^2 + m^2 + b^2 + \left(M-2 \right)\frac{a^2 + am + mb + m^2}{2} \\
\quad \textit{and} \\
UB \approx a^2 + m\left(M-2 \right)\frac{a+m}{2} + m^2 + b\left(M-2 \right)\frac{m+b}{2} + b^2 \\
\quad = a^2 + m^2 + b^2 + \left(M-2 \right)\frac{am + m^2 + bm + b^2}{2}
\end{array} \qquad (9)$$

The sum of squares can be therefore estimated as

$$\begin{array}{l}
{x_1}^2 + {x_2}^2 + {x_3}^2 + \cdots + {x_{n-1}}^2 + {x_n}^2 \\
\approx \frac{LB + UB}{2} \\
\approx a^2 + m^2 + b^2 + \left(\frac{n+1}{2} - 2 \right)\frac{a^2 + 2am + m^2 + m^2 + 2mb + b^2}{4} \\
= a^2 + m^2 + b^2 + \left(\frac{n-3}{2} \right)\frac{\left(a+m \right)^2 + \left(m+b \right)^2}{4}
\end{array} \qquad (10)$$

The sample variance can be evaluated from the computational formula

$$S^2 = \frac{n\sum_{i=1}^{n} {X_i}^2 - \left(\sum_{i=1}^{n} X_i \right)^2}{n\left(n-1 \right)} = \frac{\sum_{i=1}^{n} {X_i}^2 - n\left(\frac{1}{n}\sum_{i=1}^{n} X_i \right)^2}{n-1} \qquad (11)$$

We can estimate $\sum_{i=1}^{n} {X_i}^2$ using (10) and $\frac{1}{n}\sum_{i=1}^{n} X_i$ using (4). Therefore, after simplifying:

$$S^2 \approx \frac{1}{n-1}\left(a^2 + m^2 + b^2 + \left(\frac{n-3}{2} \right)\frac{\left(a+m \right)^2 + \left(m+b \right)^2}{4} - n\left(\frac{a+2m+b}{4} + \frac{a-2m+b}{4n} \right)^2 \right) \qquad (12)$$

Note that if we let n grow without bound, the expression (12) becomes the well-known range formula $S \approx \frac{b-a}{4} = \frac{R}{4}$.

The Variance S^2 – equidistantly spaced data

The formula (4) can also be obtained by dividing the range [*a*, *b*] into two parts: [*a*, *m*), and [*m*, *b*]. We then subdivide each of these two parts into subintervals using equally spaced partition points. In other words, we are estimating each of the data points (except for *a*, *m*, and *b*) with uniformly spaced approximate points:

$$x_i = a + (i-1)\cdot\frac{m-a}{M-1}\ \ (\text{where } i = 1,2,\ldots M-1) \qquad (13)$$

and

$$y_i = m + (i-1)\cdot\frac{b-m}{M-1}(\text{ where } i = 1,2,\ldots M) \qquad (14)$$

Therefore our sample is approximately given as

$a = x_1 \le x_2 \le x_3 \le \ldots \le x_{M-1} \le y_1 = m \le y_2 \le \ldots \le y_{M-1} \le y_M = b.$

We can use this partition to estimate the sample variance $S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{M-1}(x_i - \bar{x})^2 + \sum_{i=1}^{M}(y_i - \bar{x})^2\right)$ (and standard deviation, *S*). After a little algebra, the sample variance can be estimated by

$$S^2 = \frac{(n+1)}{48n(n-1)^2}\left((n^2+3)(a-2m+b)^2 + 4n^2(b-a)^2\right) \qquad (15)$$

If we let the number of estimation points increase without bounds, i.e., assume that *n* in the expression (15) is very large, we obtain a simplified version of the expression above:

$$S^2 \approx \frac{1}{12}\left(\frac{(a-2m+b)^2}{4} + (b-a)^2\right) \qquad (16)$$

Discussion

Analysis and performance of estimates

In order to verify the accuracy of these estimates, we ran several simulations using the computer package Maple where the data were variously distributed, and obtained the tables below.

We drew samples from five different distributions, Normal, Log-normal, Beta, Exponential and Weibull. The size of the sample ranged from 8 to about 100. In the first subsection we present the results of our estimation for a normal distribution, which is what meta-analysts would commonly assume. We also show the results of simulations where the data were selected from a skewed distributions. In each case we compared the relative error made by estimating the sample mean with the approximation given by formulas (4) and (5), as well as by the median, and the relative error made by estimating the sample variance by the formulas (12) and (16), as well as the well-known standard deviation estimators Range/4 and Range/6.

Normal distribution

We drew 200 random samples of sizes ranging from 8 to 100 from a Normal Distribution with a population mean 50 and standard deviation 17. Then we graphed the average relative error vs. the sample size. Both estimators for the mean, formulas (4) and (5), are very close to the sample mean (within 4%). For sample sizes smaller than 29, formula (5) is actually outperforming the median as a mean estimator. For larger sample sizes, however, the median is more consistent estimator for a normally distributed sample.

The variance estimators however show greater distinction. For a very small sample size (up to 15) the formula (16) is performing the best (within 10% of the real sample standard deviation). When the sample size is between 16 and 70, the formula Range/4 is the best estimator of the sample standard deviation, with a relative error between 10–15%. However, for larger sample sizes, the formula Range/6 performs the best for this distribution. To compare the precision of these estimates on average, we collected the results of our simulation in the Additional file 1.

Simulation with a skewed distribution (Log-Normal, Beta, Exponential and Weibull)

We also decided to run a simulation where the algorithm selects a sample from a skewed distribution. We decided to use Log-Normal distribution with parameters μ = 4, and σ = 0.3, Beta distribution with parameters *a* = 9 and *b* = 4, Exponential distribution with the parameter λ = 10 and Weibull distribution *a* = 2 and *b* = 35. These parameters were chosen arbitrarily, and the simulation results did not differ when we used different parameters (naturally, larger variance translates into larger relative error for mean estimators for any distribution). Just like in the case of Normal distribution, we ran our algorithm 200 times for each sample size ranging from 8 to 100. For each of the estimation formulas we then calculated the average relative error. We will summarize the best formula for estimation in Table 1.

Therefore, counter intuitively, even for the skewed distributions we tested, it seems like that for a larger sample size (usually more than 25) simply replacing sample mean with the reported median is the best estimate of the sample mean. This is an interesting result and we are not aware that it was previously demonstrated. It gives assurance to meta-analysts that simple replacement of mean with medians in meta-analysis is a viable option. Formula

Table 1: The best formula for estimation by distribution.

Best Formula for Sample size (n)	Mean Estimation		Standard Deviation Estimation		
	Formula (5)	Median	Formula (16)	Range/4	Range/6
Log-Normal	$n \leq 23$	23 <n	$n \leq 15$	15 <$n \leq 64$	64 <n
Beta	$n \leq 30$	30 <n	$n \leq 15$	15 <$n \leq 100$	100 <n
Exponential	$n \leq 21$	21 <n	$n \leq 15$	15 <$n \leq 66$	66 <n
Weibull	$n \leq 25$	25 <n	$n \leq 16$	16 <$n \leq 110$	110 <n

(5), even though taking more parameters into account (the range and the sample size), on average only outperforms the median for small sample sizes. However, a large number of trials used in meta-analyses do have very small number of patients for each arm (as small as 10–15). For these trials, formula (5) seems to give an alternative to just using the median.

When estimating the standard deviation, formula (16) is the best estimate for very small sample sizes (less than 16), after which the range formulas (Range/4 and Range/6) are better. Range/4 formula works best for samples of moderate size (between 16 and about 70), while for really large samples, Range/6 is the best estimator.

Detailed results of each simulation with a skewed distribution are given in the Additional file 2, Additional file 3, Additional file 4, and Additional file 5.

If the reader wants to try these formulas with a different set of data, we have provided an Excel spreadsheet file with the formulas at http://www.iun.edu/~mathiho/medmath/Estimating.xls

Effect on the mean difference in meta-analysis

In this section we will discuss the use of these estimating formulas on the effect size for the meta-analysts. When pooling the means from various sources for a meta-analysis, the usual procedure is to calculate differences in the means between the experimental arm of a study and the control arm, $m_p = m_c - m_e$, and the combined variance for each study, $v_p = \frac{v_c}{n_c} + \frac{v_e}{n_e}$ (for example, see [3]). The pooled mean difference is then calculated by using weighted sum of these differences, where the weight is the reciprocal of the combined variance for each study.

To determine whether our estimates make a huge difference when compared to the actual mean difference and variance, we drew two samples of the same size from a same distribution. We applied our methods to the Log-Normal [4, 0.3] distribution since this skewed distribution is frequently encountered in biology and medicine.

First we ran a test-case meta-analysis. After drawing fifteen samples of random sizes (between 8 and 100) from our distribution, we used our estimation formulas to estimate the mean and the variance from the median and the range. Then we performed meta-analysis using STATA, treating the samples as one subgroup and their estimates as another subgroup to determine the pooled means and heterogeneity. Our results for the weighted mean difference, WMD (see Figure 1) are presented in Table 2.

In order to capture a more consistent measure of the effect of our estimation on pooled mean difference, we repeated this process by varying the number of trials in the meta-analysis from 8 to 100. In particular we are interested in the difference between the real pooled weighted mean difference in the sample group and the pooled weighted mean difference from a meta-analysis using estimated means and variances.

The actual population mean from which we drew samples is 57.11 and the standard deviation is 17.53 (Log-Normal [4, 0.3]). The actual average pooled sample mean difference between two samples (one was control, the other experimental group) was 0.031. Using the medians and range, we estimated the means for each sample, and performed the meta-analysis using these estimates. The average pooled (estimated) mean difference was 0.002, making the difference between the two methods 0.029 (on average). Individually, the pooled means (both, the real sample pooled means, and the estimated pooled means) differed a little more. In Figure 2 the black diamonds represent the actual pooled mean difference using actual sample means. The red circles represent the same pooled mean differences using our estimation formulas (we connected the corresponding symbols for clarity). The horizontal axis represents the number of trials in the meta-analysis (from 8 to 100).

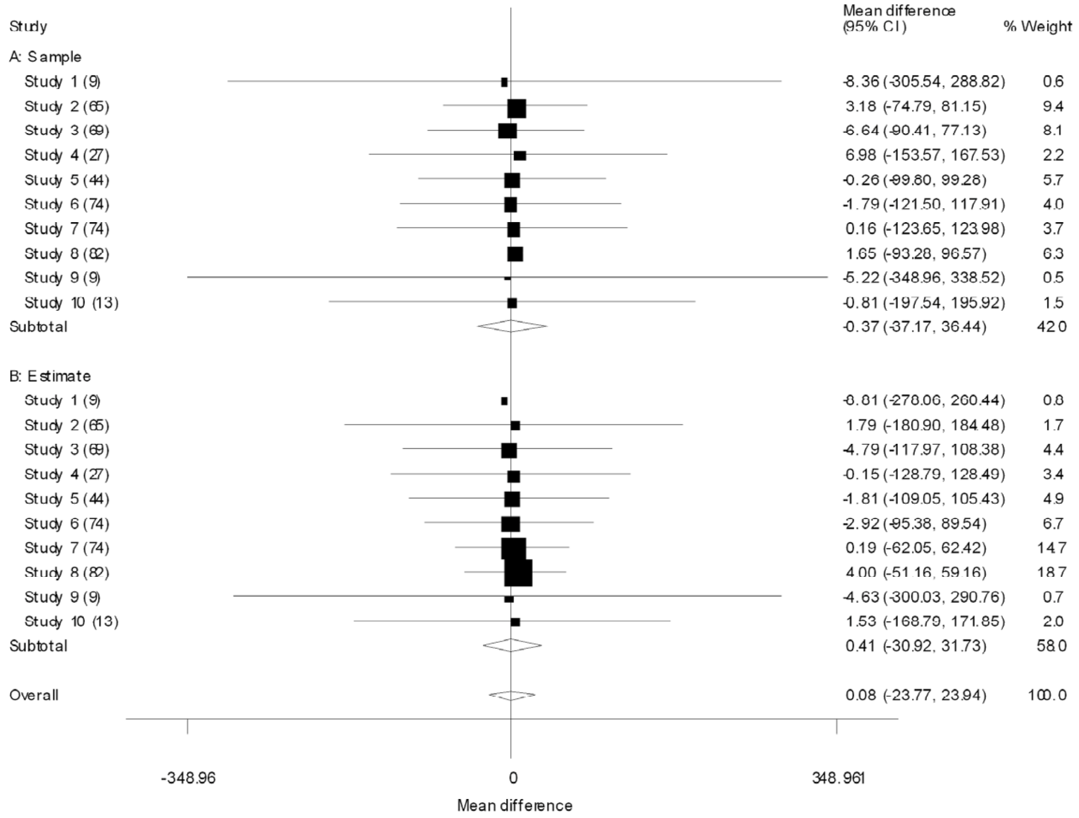

Figure 1

Meta-Analysis of random data. After drawing fifteen samples of random sizes (between 8 and 100) from the Log-Normal [4, 0.3] distribution, we used our estimation formulas to estimate the mean and the variance from the median and the range. Then we performed meta-analysis using STATA, treating the real samples as one subgroup and their estimates as another subgroup to determine the results and heterogeneity.

As seen from the Figure 2, the estimates of the mean were fairly accurate and useful. On the other hand, the estimates for the variance were a lot less precise, missing the actual value of the variance by 10 % – 20% (see the Additional Files 1, 2, 3, 4, 5). However, in some situations, using these estimates might still be better than the alternative – excluding the trials which reported the wrong summary data (median instead of mean). Using our estimation method, we can see the effect of such trials on pooled summary measures. In the next section we will illustrate our method in an actual systematic review.

An illustrative example of the potential value of our methods

American Society of Hematology/ American Society of Oncology (ASH/ASCO) developed practice guidelines for the use of erythropoietin (Epo), a drug whose annuals sales exceed several billions of dollars in the US alone, based on the systematic review of the effects of Epo on various clinical outcomes of interest including improvement of anemia by increase of hemoglobin[4]. The results were expressed as the mean increase in hemoglobin in Epo arm compared with the control. However, a number of the papers reported median increase instead of mean increase and standard deviation. Due to lack of available methods to use median values, the authors of this important review, decided not to use these papers in their meta-analysis. Recently, the Cochrane review was published attempting to provide more updated analysis of the effects of Epo in anemia related to malignancy [5]. The Cochrane reviewers did meta-analyze data to calculate an average weighted mean increase in hemoglobin as the result of Epo treatment. However, the Cochrane investigators could not include the totality of evidence in relation to

Table 2: Results of our meta-analysis with the real sample data as one subgroup, and our estimates of the sample as the second subgroup.

	Actual Sample		Our Estimate	
	WMD [95% CI]	% Weight	WMD [95% CI]	% Weight
Pooled WMD	-0.37 [-37.17, 36.44]	42.00	0.41 [-30.92, 31.73]	58.00
Overall pooled WMD			0.08	100.00
	Heterogeneity statistic	degrees of freedom	P	I-squared
Sample	0.04	9	1.000	0.0%
Estimate	0.04	9	1.000	0.0%
Overall	0.08	19	1.000	0.0%
Overall Test for heterogeneity between sub-groups				
	0.00	1	0.975	
Significance test(s) of WMD = 0		Sample	z = 0.02	p = 0.984
		Estimate	z = 0.03	p = 0.980
		Overall	z = 0.01	p = 0.995

[1] I-squared: the variation in WMD attributable to heterogeneity

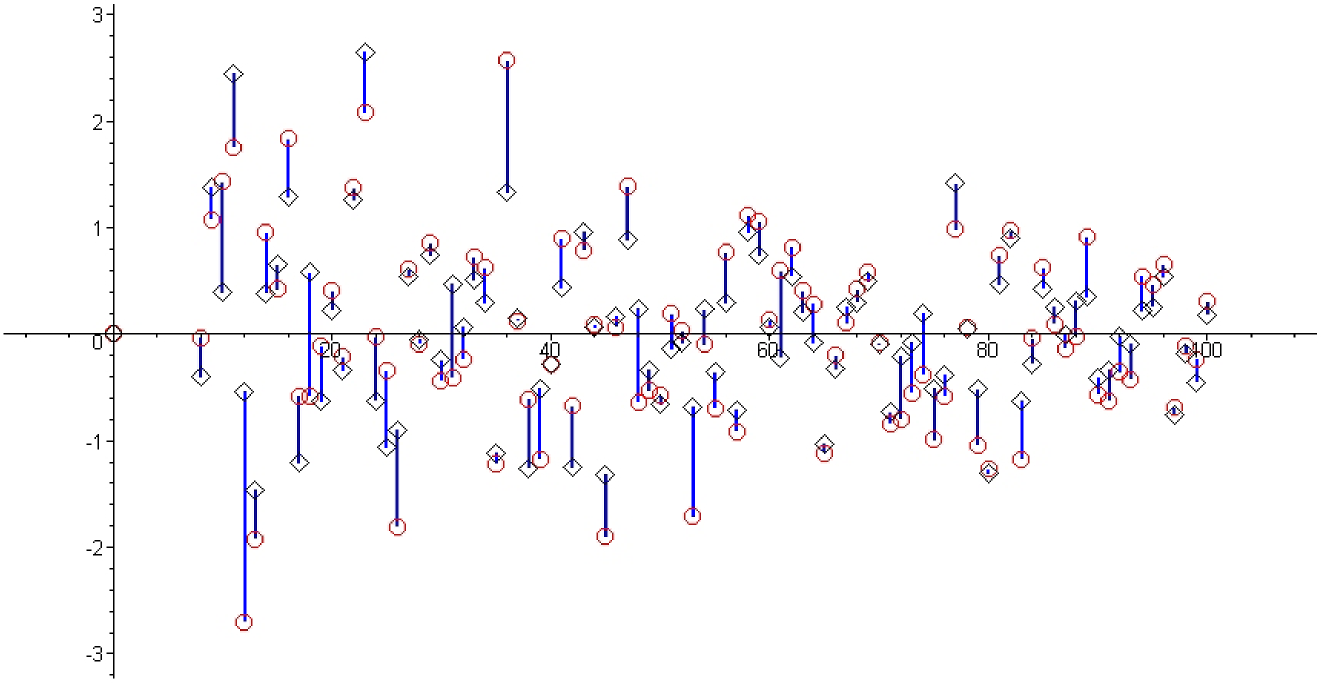

Figure 2

Actual pooled mean difference and estimated pooled mean difference. The black diamonds represent the actual pooled mean difference using sample means. The red circles represent the pooled mean differences for the same samples using our estimation formulas (we connected the corresponding symbols for clarity). The horizontal axis represents the number of trials in the meta-analysis (from 8 to 100).

this outcome since a number of the trials reported data as medians instead of means. Therefore, published meta-analyses related to the effect of Epo in anemia due to malignancy suffer from the phenomena akin to the outcome reporting bias [6] simply due to fact that methods are not yet developed to allow researches to use data medians.

Here we illustrate that it is actually possible to use medians and pool, and improve inclusiveness of meta-analyses. For example, the Cochrane investigators were only able to pool 2 studies [7,8] to evaluate the effect of Epo on change in hemoglobin in the patients with the baseline level of hemoglobin >12 g/dl who underwent chemotherapy. Their results show that on average Epo increases hemoglobin by 2.05 g/dl. However, the Cochrane investigators could not pool data from other available studies in the literature with similar eligibility. ASH/ASCO guidelines listed two other studies that were eligible for the meta-analysis (and two that were not).

For the first of these studies [9], by Welch at al, the ASH/ASCO guidelines paper reported the mean hemoglobin change for each of the two arms, the experimental and the control. However, they did not report the data for the standard deviation of these means. Since the size of each arm is 15 patients, our formula (16) provides the best estimate of the standard deviation using the median and the range. We used Figure 1 on page 263 in Welch at al. [9] to estimate the range of the hemoglobin change for each arm and used formula (16) to determine the standard deviation. The ASH/ASCO guidelines paper also reported the difference in medians of hemoglobin response for the largest study eligible for the meta-analysis conducted by Thatcher at al [10]. Thatcher et al do report in their paper ranges of hemoglobin for patients treated by Epo and control. This trial was a three-arm study, in which two doses of Epo were compared against the control. For the purpose of this analysis, we separated the data from each of the Epo arms and compared them against one half of the control group (just like the rest of the studies in the Cochrane review). Using the methods described here, we were able to estimate mean increase (using formula (5)) and standard deviation (using Range/4 formula in both comparisons). When we incorporated these results into the Cochrane meta-analysis, we found that the effect of Epo on mean increase in hemoglobin significantly changed: the pooled estimate decreased from an average of 2.05 g/dl in hemoglobin increase to 1.22 g/dl, i.e., a decrease of approximately 40% (see Figure 3)!

Our estimates come with some uncertainty. To see what effect this uncertainty has on the outcome of our meta-analysis, we varied the estimated means in Thatcher at al by 4% and the estimated standard deviation in both, Thatcher at al and Welch at al, by 10% to 15% (according to sample sizes, as indicated in the Additional Files 1, 2, 3, 4, 5). The summary pooled estimate now ranged from the low of 1.09 to the high of 1.32, which represents a decrease between 36% and 47%.

This example outlines how our method can be potentially useful for meta-analysts. It is important to realize that this example is provided only to illustrate our method. Our goal here is not to challenge the Cochrane review or ASH/ASCO guidelines. Nevertheless, we believe that this example is a good illustration of the potential of our method. While it is common practice that the investigators simply pool what is available to them it is actually not known how often studies are excluded because of reporting a different summary statistic. In future we will attempt to systematically address this issue and evaluate, for example, how often the Cochrane reviews did not pool data from the available median values when they pooled data on continuous outcomes. We hope that availability of our methods to the wider meta-analytic audience may further improve the inclusiveness of all relevant studies for the Cochrane and other meta-analyses.

Conclusion

We found that a simple formula (5): $\bar{x} \approx \frac{a+2m+b}{4}$ can be used to estimate the mean using the values of the median (m), low and high end of the range (a and b , respectively).

Using simulation methods we were able to determine that formula (5) is a best estimator for the mean when dealing with a small sample size. As soon as sample size exceeds 25, the median itself is the best estimator.

The variance can be estimated using the formula (16)

$$S^2 \approx \frac{1}{12}\left(\frac{(a-2m+b)^2}{4}+(b-a)^2\right).$$

Together with the well-known estimators (Range/4 for a normal distribution, and Range/6 for any random distribution) this formula provides a useful tool for meta-analysts. Using simulations, we determined that for very small samples (up to 15) the best estimator for the variance is the formula (16). When the sample size increases, Range/4 is the best estimator for the standard deviation (and variance) until the sample sizes reach about 70. For large samples (size more than 70) Range/6 is actually the best estimator for the standard deviation (and variance).

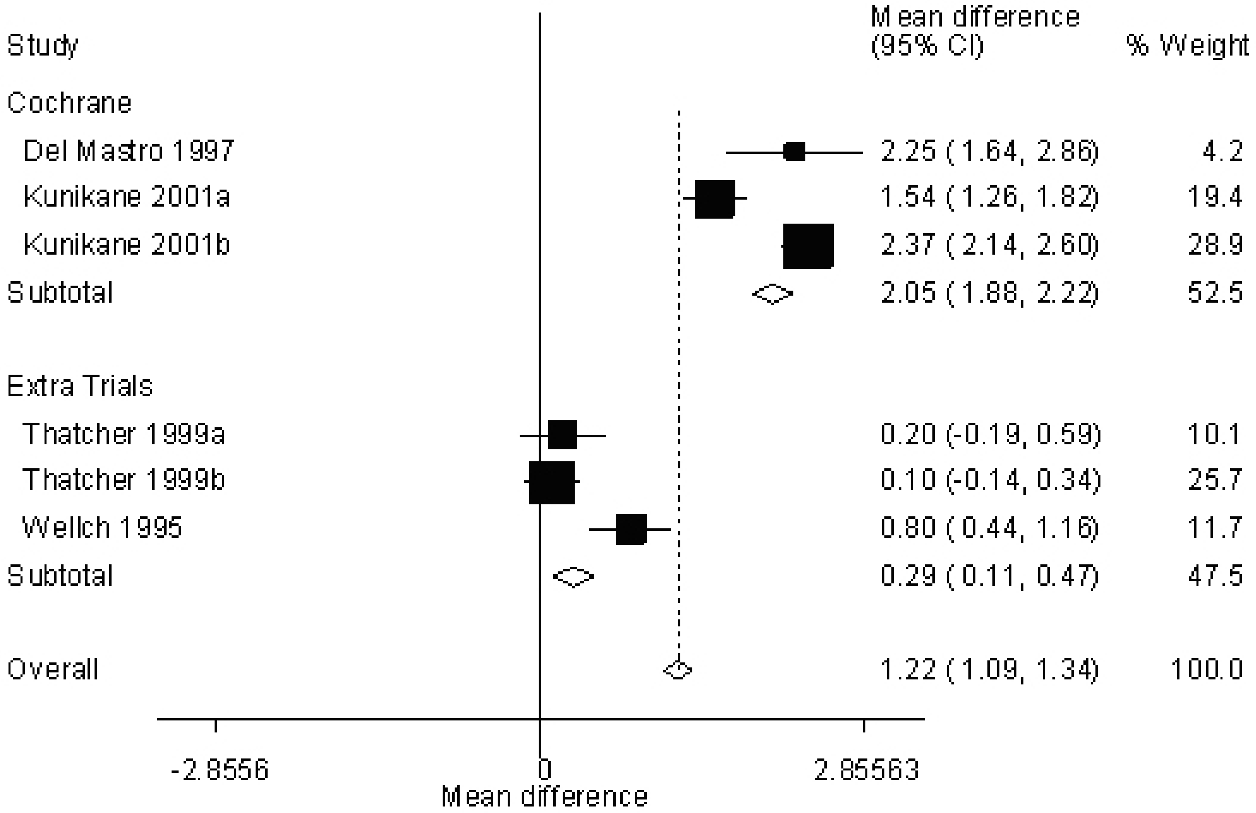

Figure 3

An example: Meta Analysis with all eligible trials included. Cochrane investigators [5] were only able to pool two studies to evaluate the effect of Epo on change in hemoglobin in the patients with the baseline level of hemoglobin >12 g/dl who underwent chemotherapy. Their results show that on average Epo increases hemoglobin by 2.05 g/dl. Using our estimation formulas, we were able to include two other studies eligible for this meta-analysis ([9, 10]). The pooled estimate decreased to 1.22 g/dl, i.e., a decrease of approximately 40%.

Table 3: The best estimating formula for an unknown distribution.

Sample Size:	$n \leq 15$	$15 < n \leq 25$	$25 < n \leq 70$	$70 < n$
Estimate Mean	Formula (5)		Median	
Estimate Standard Deviation	Formula (16)	Range/4		Range /6

In summary, the best estimators for the mean and the standard deviation for different sample sizes are given in Table 3.

Using these formulas, we hope to enable meta-analysts use clinical trials even when not all of the information is available and/or reported.

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

SPH developed most of the formulas for estimation. IH conducted the simulations using Maple. BD conceived of the problem, participated in its framing and coordination, and helped SPH and IH draft the manuscript. All authors read and approved the final manuscript.

Additional material

Additional File 1

Normal distribution. *The top row of the table displays the results of estimating the mean, while the second row displays the results of estimating the standard deviation. Each number in this table represents the average relative error of 200 samples from a Normal distribution.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-5-13-S1.pdf]

Additional File 2

Log-Normal distribution. *The top row of the table displays the results of estimating the mean, while the second row displays the results of estimating the standard deviation. Each number in this table represents the average relative error of 200 samples from a Log-Normal distribution.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-5-13-S2.pdf]

Additional File 3

Beta Distribution. *The top row of the table displays the results of estimating the mean, while the second row displays the results of estimating the standard deviation. Each number in this table represents the average relative error of 200 samples from a Beta distribution.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-5-13-S3.pdf]

Additional File 4

Exponential Distribution. *The top row of the table displays the results of estimating the mean, while the second row displays the results of estimating the standard deviation. Each number in this table represents the average relative error of 200 samples from a Exponential distribution.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-5-13-S4.pdf]

Additional File 5

Weibull Distribution. *The top row of the table displays the results of estimating the mean, while the second row displays the results of estimating the standard deviation. Each number in this table represents the average relative error of 200 samples from a Weibull distribution.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2288-5-13-S5.pdf]

References

1. Hogg RV, Craig AT: **Introduction to mathematical statistics.** 5th edition. New York Toronto , Macmillan College Pub. Co. ; Maxwell Macmillan Canada ; Maxwell Macmillan International; 1995:xi, 564.
2. Mood AMF, Graybill FA, Boes DC: **Introduction to the theory of statistics.** 3d edition. New York, , McGraw-Hill; 1974:xvi, 564.
3. Petiti DB: **Meta-analysis, decision analysis and cost-effectiveness analysis. Methods for quantitative synthesis in medicine. 2nd ed.** New York , Oxford press; 2000.
4. Rizzo JD, Lichtin AE, Woolf SH, Seidenfeld J, Bennett CL, Cella D, Djulbegovic B, Goode MJ, Jakubowski AA, Lee SJ, Miller CB, Rarick MU, Regan DH, Browman GP, Gordon MS: **Use of epoetin in patients with cancer: evidence-based clinical practice guidelines of the American Society of Clinical Oncology and the American Society of Hematology.** *J Clin Oncol* 2002, **20(19):**4083-4107.
5. Bohlius J, Langensiepen S, Schwarzer G, Seidenfeld J, Piper M, Bennet C, Engert A: **Erythropoietin for patients with malignant disease.** *Cochrane Database Syst Rev* 2004:CD003407..
6. Chan AW, Hrobjartsson A, Haahr MT, Gotzsche PC, Altman DG: **Empirical Evidence for Selective Reporting of Outcomes in Randomized Trials: Comparison of Protocols to Published Articles.** *JAMA* 2004, **291(20):**2457-2465.
7. Del Mastro L, Venturini M, Lionetto R, Garrone O, Melioli G, Pasquetti W, Sertoli MR, Bertelli G, Canavese G, Costantini M, Rosso R: **Randomized phase III trial evaluating the role of erythropoietin in the prevention of chemotherapy-induced anemia.** *J Clin Oncol* 1997, **15(7):**2715-2721.
8. Kunikane H, Watanabe K, Fukuoka M, Saijo N, Furuse K, Ikegami H, Ariyoshi Y, Kishimoto S: **Double-blind randomized control trial of the effect of recombinant human erythropoietin on chemotherapy-induced anemia in patients with non-small cell lung cancer.** *Int J Clin Oncol* 2001, **6(6):**296-301.
9. Welch RS, James RD, Wilkinson PM, Fb: **Recombinant Human Erythropoietin and Platinum-Based Chemotherapy In Advanced Ovarian Cancer.** *Cancer J Sci Am* 1995, **1(4):**261.
10. Thatcher N, De Campos ES, Bell DR, Steward WP, Varghese G, Morant R, Vansteenkiste JF, Rosso R, Ewers SB, Sundal E, Schatzmann E, H. S: **Epoetin alpha prevents anaemia and reduces transfusion requirements in patients undergoing primarily platinum-based chemotherapy for small cell lung cancer.** *Br J Cancer* 1999, **80(3-4):**396-402.

Pre-publication history

The pre-publication history for this paper can be accessed here:

http://www.biomedcentral.com/1471-2288/5/13/prepub