# Estimating Mean and Standard Deviation from the Sample Size, Three Quartiles, Minimum, and Maximum

Martin Bland*

*Prof. of Health Statistics, University of York, York YO10 5DD, England*

**Abstract:** *Background*: We sometimes want to include in a meta-analysis data from studies where results are presented as medians and ranges or interquartile ranges rather than as means and standard deviations. In this paper I extend a method of Hozo *et al.* to estimate mean and standard deviation from median, minimum, and maximum to the case where quartiles are also available.

*Methods*: Inequalities are developed for each observation using upper and lower limits derived from the minimum, the three quartiles, and the maximum. These are summed to give bounds for the sum and hence the mean of the observations, the average of these bounds in the estimate. A similar estimate is found for the sum of the observations squared and hence for the variance and standard deviation.

*Results*: For data from a Normal distribution, the extended method using quartiles gives good estimates of sample means but sample standard deviations are overestimated. For data from a Lognormal distribution, both sample mean and standard deviation are overestimated. Overestimation is worse for larger samples and for highly skewed parent distributions. The extended estimates using quartiles are always superior in both bias and precision to those without.

*Conclusions*: The estimates have the advantage of being extremely simple to carry out. I argue that as, in practice, such methods will be applied to small samples, the overestimation may not be a serious problem.

**Keywords:** Quartile, minimum, maximum, mean, standard deviation, systematic review.

## THE NEED FOR A METHOD

In systematic review and meta-analysis, we sometimes want to combine data from studies where results are presented in different forms. To combine results from several studies, we need to have results in a common format. When the outcome is a quantitative variable, this format is preferably as means and standard deviations. Sometimes research results are published as medians and ranges or interquartile ranges. If possible, reviewers should try to obtain the original data so that means and standard deviations can be computed. Often it is not possible to contact study authors or raw data have been lost. Reviewers may then have to omit these studies from the quantitative part of the review, or try to salvage what they can from publications which give only very limited summaries of data. Reviewers may be able to back-calculate means and standard deviations from confidence intervals or P values. Sometimes we have to take a ruler to graphs to extract data. If the original results are not obtainable, how can we estimate means and standard deviations from what we do have?

For a systematic review, we had two papers which had published the medians and the limits of the ranges and the interquartile ranges. We wanted to estimate means and standard deviations from the available information. Hozo *et al.* [1] published a method for estimating the mean and variance from the median, range, and the size of a sample. I asked whether we could improve on this by including the first and third medians?

## DEVELOPING THE STATISTICAL METHOD

### Notation

sample size = $n$

minimum = $a$

first quartile = $b$

median = $c$

third quartile = $d$

maximum = $e$

Suppose we have a sample where all quartiles are actual observations, so that there are $k$ observations between minimum and first quartile, $k$ observations between first quartile and median, $k$ observations between median and third quartile, and $k$ observations between third quartile and maximum. Then $n = 4k + 5$.

Following Hozo e*t al.*, we can set up a series of inequalities for the observations $x_i$.

*Address correspondence to this author at the Prof. of Health Statistics, University of York, York YO10 5DD, England; Tel: +44 (0)1904 321334; Fax: +44 (0)1904 321382; E-mail: martin.bland@york.ac.uk

E-ISSN: 1929-6029/15 

$a \le x_1 = a \le a$

$a \le x_2 \le b$ }

…… }

$a \le x_i \le b$ } $k$ inequalities

…… }

$b \le x_{k+2} = b \le b$

…… }

$b \le x_i \le c$ } $k$ inequalities

…… }

$c \le x_{2k+3} = c \le c$

…… }

$c \le x_i \le d$ } $k$ inequalities

…… }

$d \le x_{3k+4} = d \le d$

…… }

$d \le x_i \le e$ } $k$ inequalities

…… }

$e \le x_{4k+5} = e \le e$

We can sum these inequalities to give a single inequality:

$$a + ka + b + kb + c + kc + d + kd + e \le \Sigma x_i \le a + kb + b + kc + c + kd + d + ke + e$$
$$(k + 1)(a + b + c + d) + e \le \Sigma x_i \le a + (k + 1)(b + c + d + e)$$

This gives us upper and lower limits for the sum of $x_i$ and hence, dividing by $n$, limits for the mean. The estimate of Hozo *et al.* [1] is the average of these limits:

$$\frac{(k+2)a + 2(k+1)(b+c+d) + (k+2)e}{2n}$$

If we replace $4k + 5$ by $n$, the estimate of the mean becomes:

$$\frac{(n+3)a + 2(n-1)(b+c+d) + (n+3)e}{8n}$$

Following the same procedure, without the first and third quartiles, Hozo *et al.* [1] obtained:

$$\frac{a+2c+e}{4} + \frac{a-2c+e}{4n} = \frac{(n+1)a + 2(n-1)c + (n+1)e}{4n}$$

For the variance, we find similar inequalities for $x_i^2$:

$ax_1 \le x_1^2 = a^2 \le ax_1$

$ax_2 \le x_2^2 \le bx_2$ }

…… }

$ax_i \le x_i^2 \le bx_i$ } $k$ inequalities

…… }

$bx_{k+2} \le x_{k+2}^2 = b^2 \le bx_{k+2}$

…… }

$bx_i \le x_i^2 \le cx_i$ } $k$ inequalities

…… }

$cx_{2k+3} \le x_{2k+3}^2 = c^2 \le x_{2k+3}$

…… }

$cx_i \le x_i^2 \le dx_i$ } $k$ inequalities

…… }

$dx_{3k+4} \le x_{3k+4}^2 = d^2 \le x_{3k+4}^2$

…… }

$dx_i \le x_i^2 \le ex_i$ } $k$ inequalities

…… }

$e^2 \le x_{4k+5}^2 = e^2 \le e^2$

We can sum these inequalities to give a single inequality:

$$\begin{aligned} a^2 + a(x_2 + \ldots + x_{k+1}) \quad &\le \Sigma x_i^2 \le \quad a^2 + b(x_2 + \ldots + x_{k+1}) \\ + b^2 + b(x_{k+3} + \ldots + x_{2k+2}) \quad &\qquad\quad + b^2 + c(x_{k+3} + \ldots + x_{2k+2}) \\ + c^2 + c(x_{2k+3} + \ldots + x_{3k+3}) \quad &\qquad\quad + c^2 + d(x_{2k+3} + \ldots + x_{3k+3}) \\ + d^2 + d(x_{3k+5} + \ldots + x_{4k+4}) \quad &\qquad\quad + d^2 + e(x_{3k+5} + \ldots + x_{4k+4}) \\ + e^2 \quad &\qquad\quad + e^2 \end{aligned}$$

In a similar position, Hozo *et al.* [1] replace $(x_2 + \ldots + x_{M-1})$ by the estimate $(M - 2)(a + c)/2$, where $M$ is the number of observations up to but not including the median. In the same way, we can replace $(x_2 + \ldots + x_{k+1})$ by the estimate $k(a + b)/2$, $(x_{k+3} + \ldots + x_{2k+2})$ by the estimate $k(b + c)/2$, and so on. The inequality becomes

$$a^2 + ak(a + \mathrm{b})/2 + b^2 + bk(\mathrm{b} + \mathrm{c})/2 + c^2 + ck(\mathrm{c} + d)/2 + d^2 + dk(d + e)/2 + e^2 \le \Sigma x_i^2 \le a^2 + bk(a + \mathrm{b})/2 + b^2 + ck(\mathrm{b} + \mathrm{c})/2 + c^2 + dk(\mathrm{c} + d)/2 + d^2 + ek(d + e)/2 + e^2$$

which becomes

$$[(k + 2)(a^2 + b^2 + c^2 + d^2) + k(ab + bc + cd + de) + 2e^2]/2 \le \Sigma x_i^2 \le [(k + 2)(b^2 + c^2 + d^2 + e^2) + k(ab + bc + cd + de) + 2a^2]/2$$

If we replace $4k+5$ by $n$, we get

$$[(n + 3)(a^2 + b^2 + c^2 + d^2) + (n - 5)(ab + bc + cd + de) + 8e^2]/8 \leq \Sigma x_i^2 \leq [(n + 3)(b^2 + c^2 + d^2 + e^2) + (n - 5)(ab + bc + cd + de) + 8a^2]/8$$

If we average these limits to get an estimate of $\Sigma x_i^2$ we get

$$[2(n + 3)(b^2 + c^2 + d^2) + 2(n - 5)(ab + bc + cd + de) + (n + 11)(a^2 + e^2)]/16$$

Ignoring the first and third quartiles, Hozo *et al.* [1] get:

$$(a^2 + c^2 + e^2) + (n+3)[(a + c)^2 + (c + e)^2]/8$$

We then subtract the estimated mean squared multiplied by $n$ and divide the result by $(n - 1)$ to get the estimated variance; the square root gives us the estimated standard deviation.

## APPLICATION TO A PRACTICAL EXAMPLE

Table **1** shows a sample of measurements of forced expiratory volume from a sample of 57 medical students [2]. The quantiles for these data are:

minimum = 2.85 = *a*

1st quartile = 3.54 = *b*

median = 4.1 = *c*

3rd quartile = 4.5 = *d*

maximum = 5.43 =*e*

If we apply the formula for the mean we get estimated mean = 4.07 litre, compared to a directly calculated mean = 4.06 litre. If we apply the formula for standard deviation the estimate is 0.68 litre, compared to directly calculated standard deviation = 0.67 litre. Thus the approximation is quite good.

In contrast, Table **2** [2] shows a data set which clearly has a highly skewed distribution, vitamin D measured in the blood of 26 men [2]. The quantiles for these data are:

minimum = 14 = *a*

1st quartile = 25 = *b*

median = 31 = *c*

3rd quartile = 48 = *d*

maximum = 83 =*e*

If we apply the formula for the mean we get estimated mean = 38.5, compared to a directly calculated mean = 36.9 mmol/litre. For standard deviation the estimate is 19.2 mmol/litre, compared to directly calculated standard deviation = 17.2 mmol/litre. Thus the approximation is not quite so good, being slightly too high for both mean and standard deviation.

Table **3** shows a larger data set which also clearly has a highly skewed distribution, serum triglyceride from 282 babies [2]. The quantiles for these data are:

minimum = 0.15 = *a*

1st quartile = 0.35 = *b*

median = 0.46 = *c*

3rd quartile = 0.60 = *d*

maximum = 1.66 =*e*

If we apply the formula for the mean we get estimated mean = 0.58 mmol/litre, compared to a directly calculated mean = 0.51 mmol/litre. For standard deviation the estimate is 0.34 mmol/litre, compared to directly calculated standard deviation = 0.22 mmol/litre. Thus the approximation is relatively poor, being slightly too high for the mean and much too high for the variability.

When we have an outcome variable which has a highly skew distribution, we would usually prefer to carry out meta-analysis on the logarithmic scale if possible. Whether we can do this would depend on what can be extracted from the studies in the review.

**Table 1: FEV (Litres) for 57 Male Medical Students**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2.85 | 3.19 | 3.50 | 3.69 | 3.90 | 4.14 | 4.32 | 4.50 | 4.80 | 5.20 |
| 2.85 | 3.20 | 3.54 | 3.70 | 3.96 | 4.16 | 4.44 | 4.56 | 4.80 | 5.30 |
| 2.98 | 3.30 | 3.54 | 3.70 | 4.05 | 4.20 | 4.47 | 4.68 | 4.90 | 5.43 |
| 3.04 | 3.39 | 3.57 | 3.75 | 4.08 | 4.20 | 4.47 | 4.70 | 5.00 | |
| 3.10 | 3.42 | 3.60 | 3.78 | 4.10 | 4.30 | 4.47 | 4.71 | 5.10 | |
| 3.10 | 3.48 | 3.60 | 3.83 | 4.14 | 4.30 | 4.50 | 4.78 | 5.10 | |

**Table 2: Vitamin D Levels Measured in the Blood of 26 Healthy Men**

| | | | | | | |
|---|---|---|---|---|---|---|
| 14 | 22 | 26 | 31 | 42 | 52 | 67 |
| 17 | 24 | 26 | 31 | 43 | 54 | 83 |
| 20 | 25 | 27 | 32 | 46 | 54 | |
| 21 | 26 | 30 | 35 | 48 | 63 | |

**Table 3: Serum Triglyceride (mmol/Litre) Measured in Cord Blood from 282 Babies**

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.15 | 0.29 | 0.32 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.56 | 0.60 | 0.70 | 0.86 |
| 0.16 | 0.29 | 0.33 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.56 | 0.60 | 0.72 | 0.87 |
| 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.42 | 0.47 | 0.52 | 0.56 | 0.60 | 0.72 | 0.88 |
| 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.61 | 0.74 | 0.88 |
| 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.62 | 0.75 | 0.95 |
| 0.20 | 0.29 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.62 | 0.75 | 0.96 |
| 0.21 | 0.30 | 0.33 | 0.36 | 0.40 | 0.44 | 0.47 | 0.52 | 0.56 | 0.63 | 0.76 | 0.96 |
| 0.22 | 0.30 | 0.33 | 0.36 | 0.40 | 0.44 | 0.48 | 0.52 | 0.56 | 0.64 | 0.76 | 0.99 |
| 0.24 | 0.30 | 0.33 | 0.37 | 0.40 | 0.44 | 0.48 | 0.52 | 0.56 | 0.64 | 0.78 | 1.01 |
| 0.25 | 0.30 | 0.34 | 0.37 | 0.40 | 0.44 | 0.48 | 0.53 | 0.57 | 0.64 | 0.78 | 1.02 |
| 0.26 | 0.30 | 0.34 | 0.37 | 0.40 | 0.44 | 0.48 | 0.54 | 0.57 | 0.64 | 0.78 | 1.02 |
| 0.26 | 0.30 | 0.34 | 0.37 | 0.40 | 0.44 | 0.48 | 0.54 | 0.58 | 0.64 | 0.78 | 1.04 |
| 0.26 | 0.30 | 0.34 | 0.38 | 0.40 | 0.45 | 0.48 | 0.54 | 0.58 | 0.65 | 0.78 | 1.08 |
| 0.27 | 0.30 | 0.34 | 0.38 | 0.40 | 0.45 | 0.48 | 0.54 | 0.58 | 0.66 | 0.78 | 1.11 |
| 0.27 | 0.30 | 0.34 | 0.38 | 0.41 | 0.45 | 0.48 | 0.54 | 0.58 | 0.66 | 0.80 | 1.20 |
| 0.27 | 0.31 | 0.34 | 0.38 | 0.41 | 0.45 | 0.48 | 0.54 | 0.59 | 0.66 | 0.80 | 1.28 |
| 0.28 | 0.31 | 0.34 | 0.38 | 0.41 | 0.45 | 0.48 | 0.55 | 0.59 | 0.66 | 0.82 | 1.64 |
| 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.45 | 0.48 | 0.55 | 0.59 | 0.66 | 0.82 | 1.66 |
| 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.46 | 0.48 | 0.55 | 0.59 | 0.67 | 0.82 | |
| 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.46 | 0.49 | 0.55 | 0.60 | 0.67 | 0.82 | |
| 0.28 | 0.32 | 0.35 | 0.39 | 0.41 | 0.46 | 0.49 | 0.55 | 0.60 | 0.68 | 0.83 | |
| 0.28 | 0.32 | 0.35 | 0.39 | 0.42 | 0.46 | 0.49 | 0.55 | 0.60 | 0.70 | 0.84 | |
| 0.28 | 0.32 | 0.35 | 0.40 | 0.42 | 0.46 | 0.50 | 0.55 | 0.60 | 0.70 | 0.84 | |
| 0.28 | 0.32 | 0.36 | 0.40 | 0.42 | 0.46 | 0.50 | 0.55 | 0.60 | 0.70 | 0.84 | |

Using either the proposed method or the original Hozo method, it is simple to estimate the mean and standard deviation of the log transformed data, because the logarithm is a monotonic function and so the logs of the quantiles will be the quantiles of the logs. For the vitamin D data, the mean and standard deviation for $\log_e$(vitamin D) estimated from the quantiles are 3.51 and 0.49, compared to the directly calculated mean and standard deviation of 3.51 and 0.44. The mean and standard deviation for $\log_e$ (triglyceride) estimated from the quantiles are −0.76 and 0.54, compared to the directly calculated mean and standard deviation of −0.76 and 0.39. Again, the standard deviation is overestimated for this large sample.

## SIMULATION STUDIES

To explore these estimates further, Table **4** shows the result of simulations of a Normal sample with different sample sizes. These results are for single samples, but they are typical. As Hozo *et al.* [1] noted, their formulae work better for smaller samples. Table **5** shows the result of 1000 simulation runs for the Normal distribution at each of these six sample sizes. The estimates of the mean are unbiased at all sample sizes

**Table 4: Estimation of Mean and Standard Deviation for a Normal Sample, Mean = 5, SD = 1**

| Sample size | Actual | | Hozo *et al.* [1] method | | Extended method using quartiles | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| 10 | 4.90 | 0.87 | 4.98 | 0.87 | 4.82 | 0.94 |
| 20 | 5.00 | 0.96 | 5.15 | 0.93 | 5.05 | 0.96 |
| 30 | 4.80 | 0.90 | 4.81 | 0.91 | 4.78 | 0.89 |
| 50 | 5.01 | 0.96 | 4.92 | 1.20 | 4.93 | 1.10 |
| 100 | 5.02 | 0.98 | 4.97 | 1.17 | 5.01 | 1.11 |
| 500 | 4.97 | 1.03 | 4.93 | 1.32 | 4.92 | 1.20 |

**Table 5: Deviations of Estimate from Actual Sample Mean and Standard Deviation in 1000 Estimations of Mean and Standard Deviation for a Normal Sample, Mean = 5, SD = 1**

| Parameter estimated | Sample size | Actual minus Hozo *et al.* [1] method | | Actual minus extended method using quartiles | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Sample mean | 10 | −0.001 | 0.122 | 0.001 | 0.054 |
| | 20 | 0.004 | 0.138 | 0.001 | 0.062 |
| | 30 | −0.002 | 0.142 | −0.001 | 0.065 |
| | 50 | −0.004 | 0.139 | −0.003 | 0.065 |
| | 100 | 0.004 | 0.139 | 0.001 | 0.068 |
| | 500 | 0.002 | 0.128 | 0.001 | 0.063 |
| Sample standard deviation | 10 | −0.026 | 0.099 | −0.041 | 0.052 |
| | 20 | −0.086 | 0.117 | −0.069 | 0.069 |
| | 30 | −0.133 | 0.124 | −0.095 | 0.076 |
| | 50 | −0.199 | 0.122 | −0.131 | 0.079 |
| | 100 | −0.296 | 0.132 | −0.192 | 0.088 |
| | 500 | −0.532 | 0.126 | −0.348 | 0.087 |

and the standard deviation of the mean estimates is always less for the extended estimates. The estimated standard deviation is biased, being too large, and for all but the smallest sample the bias is greater for the Hozo *et al.* [1] estimates than for the extended estimates. The bias gets bigger as the sample size increases. For the samples of size 100, the bias in the extended estimate is 0.19 or 19%. The standard deviations of the estimates do not decrease with increasing sample size, which is curious. The standard deviation of the extended estimates is always less than the standard deviation of the estimates made without the quartiles, so the extended estimates are less biased and more precise than those of Hozo *et al.* [1].

Table **6** shows a similar simulation for samples from a Lognormal distribution. For the small samples, the estimates are quite good by both methods but as the sample size increases both means and standard deviations are over-estimated. Table **7** shows the results of sets of 1000 simulations using the same Lognormal distribution. At all sample sizes, the estimates of both mean and standard deviation are too big, with the bias increasing with the sample size. The standard deviations of the estimates also increase with the sample size, so they become more biased and less precise. It seems rather counter-intuitive that bigger samples produce less reliable estimates. The extended estimate is always better than the Hozo *et al.* [1] estimate, in both bias and standard deviation. With samples of size 100, the extended estimate gives bias 4.36 or 3% in the estimate of the mean and 8.7 or 22% in the estimated standard deviation. This Lognormal distribution is not particularly skew. The shape of the parent distribution is shown in Figure **1**.

Table **8** shows the results of sets of 1000 simulations using a more highly skew Lognormal

**Table 6: Estimation of Mean and Standard Deviation for a Lognormal Sample, μ = 5, σ = 0.25, Mean = 153, SD = 39**

| Sample size | Actual | | Hozo *et al*. [1] method | | Extended method using quartiles | |
|---|---|---|---|---|---|---|
| | Mean | SD | Mean | SD | Mean | SD |
| 10 | 166 | 52 | 163 | 51 | 164 | 51 |
| 20 | 161 | 46 | 155 | 47 | 161 | 47 |
| 30 | 155 | 51 | 168 | 59 | 157 | 53 |
| 50 | 150 | 37 | 157 | 41 | 152 | 41 |
| 100 | 151 | 39 | 164 | 53 | 157 | 47 |
| 500 | 151 | 39 | 182 | 73 | 165 | 65 |

**Table 7: Deviations of Estimate from Actual Sample Mean and Standard Deviation in 1000 Estimations of Mean and Standard Deviation for a Lognormal Sample, μ = 5, σ = 0.25, Mean = 153, SD = 39**

| Parameter estimated | Sample size | Actual minus Hozo *et al.* [1] method | | Actual minus extended method using quartiles | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Sample mean | 10 | −2.67 | 5.64 | −0.92 | 2.38 |
| | 20 | −4.60 | 6.31 | −3.36 | 4.82 |
| | 30 | −6.76 | 7.13 | −2.52 | 3.25 |
| | 50 | −8.44 | 7.49 | −3.26 | 3.44 |
| | 100 | −10.99 | 7.33 | −4.36 | 3.51 |
| | 500 | −18.03 | 7.44 | −7.78 | 3.67 |
| Sample standard deviation | 10 | −1.04 | 3.90 | −1.69 | 2.11 |
| | 20 | −1.63 | 2.72 | −2.77 | 2.94 |
| | 30 | −5.84 | 5.58 | −4.13 | 3.85 |
| | 50 | −8.65 | 6.18 | −5.93 | 4.32 |
| | 100 | −12.90 | 6.59 | −8.73 | 4.88 |
| | 500 | −23.69 | 7.01 | −16.39 | 5.40 |

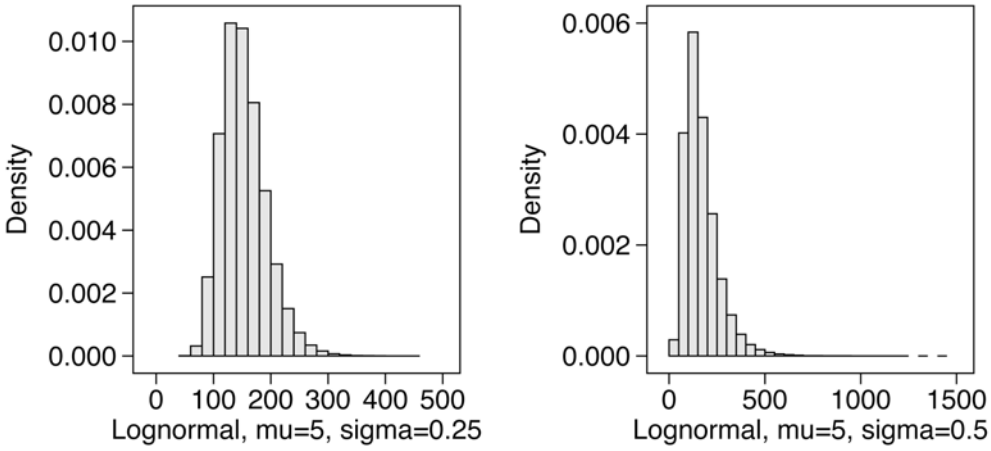


**Figure 1:** Histograms of Lognormal distributions used in the simulations.

distribution, with μ = 5, σ = 0.5. The shape of the distribution can be seen in Figure **1**. At all sample sizes the estimates are too big, with the bias increasing with the sample size. The standard deviations of the

**Table 8: Deviations of Estimate from Actual Sample Mean and Standard Deviation in 1000 Estimations of Mean and Standard Deviation for a Lognormal Sample, μ = 5, σ = 0.5, Mean = 168, SD = 90**

| Parameter estimated | Sample size | Actual minus Hozo *et al.* [1] method | | Actual minus extended method using quartiles | |
|---|---|---|---|---|---|
| | | Mean | SD | Mean | SD |
| Sample mean | 10 | −11.1 | 16.0 | −3.9 | 6.2 |
| | 20 | −22.4 | 22.8 | −8.0 | 9.7 |
| | 30 | −27.4 | 22.5 | −10.1 | 10.1 |
| | 50 | −37.6 | 26.5 | −14.5 | 12.0 |
| | 100 | −52.3 | 30.6 | −21.3 | 14.6 |
| | 500 | −85.9 | 34.1 | −37.6 | 16.8 |
| Sample standard deviation | 10 | −2.7 | 9.0 | −3.8 | 5.3 |
| | 20 | −10.4 | 14.5 | −8.1 | 10.1 |
| | 30 | −15.5 | 16.0 | −11.4 | 11.7 |
| | 50 | −24.1 | 19.3 | −17.4 | 14.8 |
| | 100 | −38.9 | 25.0 | −28.3 | 20.0 |
| | 500 | −75.0 | 31.7 | −56.7 | 26.2 |

estimates increased markedly with the sample size. These effects are more marked than in Table **7**. The extended estimate is always better than the Hozo *et al.* [1] estimate, in both bias and standard deviation. With samples of size 100, the extended estimate gives bias 21.3 or 13% in the mean and 28.3 or 17% in the standard deviation.

## DISCUSSION

The extended estimates are clearly superior to those of Hozo *et al*. [1], having less bias and smaller standard deviations. This might be expected, because they make use of more information. However, they still produce estimates of standard deviations which are too large, particularly in larger samples, and estimates of means which are also biased if the parent distribution is skew.

Why does the performance deteriorate with increasing sample size and with increasing skewness? The estimates use the maximum and minimum of the distribution. If the sample is large, extreme values are more likely to occur and an extreme outlying point can have a big effect on the estimates. Even for a symmetrical distribution, the estimates become more variable as the sample size increases, though they are unbiased. For a positively skew distribution, extreme values are likely to be high and so produce mean estimates which are too high. The other problem with the estimates is the replacement of sums of consecutive observations, e.g. $x_{3k+5} + \ldots + x_{4k+4}$, by the average of the two limits of that data segment multiplied by $k$, e.g. $k(d + e)/2$. This is not a particularly good approximation because, in a unimodal distribution with tails, the average value of such a set of observations is likely to be closer to the centre of the distribution than is the average of the limits. This will inflate the estimate of $\Sigma x_i^2$ and will therefore inflate the estimated variance and standard deviation. Skewness will make this problem worse.

The main application in systematic reviews is likely to be in extracting information from fairly small studies. It would be quite unusual for a large study to have quantitative outcome data without a mean and standard deviation and also for the data to be inaccessible. It is small studies where the authors become uncontactable and where data are lost. Fortunately, both the Hozo *et al.* [1] and the extended estimates have better performances in small studies than in large ones.

The bias in the estimated means may not be a great problem for meta-analysis, as we usually have two means and use the difference between them. The bias would be present in both estimates. The bias in the standard deviation will result in inflated sample errors for these differences in means and they will hence have reduced weight compared to what they would have if the means had been calculated directly. This is at least better than having a standard error which is too small and a weight that is too great.

There is plenty of room for improvement in these estimates and other methods will be developed which give better meta-analyses than these will provide. I have not been able to find a way to adjust for the overestimation of the variability, particularly for large samples, and it is to be hoped that one can be found in the future. However, they have the great advantage of being very easy to use and could be applied by any systematic reviewer, without the need for specialised software.

## CONCLUSIONS

The extended estimate is clearly to be preferred over the Hozo *et al.* [1] estimate and if we have the quartiles we should use them. The main problem is that the standard deviations are overestimated. The bias in sample mean might not be too important in a meta-analysis, where it would be expected to be present in both intervention and control groups. It is better to have too large a standard error than too small a standard error and we might accept this as reflecting the inferior data quality produced by using the indirect estimate.

## ACKNOWLEDGEMENTS

I thank Sally Bell-Syer for drawing my attention to this problem, Lesley Fairley for checking my algebra, and Simon Crouch for discussions on the draft.

## REFERENCE

[1] Hozo SP, Djulbegovic B, Hozo I. Estimating the mean and variance from the median, range, and the size of a sample. BMC Med Res Methodol 2005; 5: 13. http://dx.doi.org/10.1186/1471-2288-5-13

[2] Bland M. An Introduction to Medical Statistics, 3rd ed. Oxford, University Press 2000.

Received on 28-10-2014 Accepted on 18-12-2014 Published on 27-01-2015

http://dx.doi.org/10.6000/1929-6029.2015.04.01.6