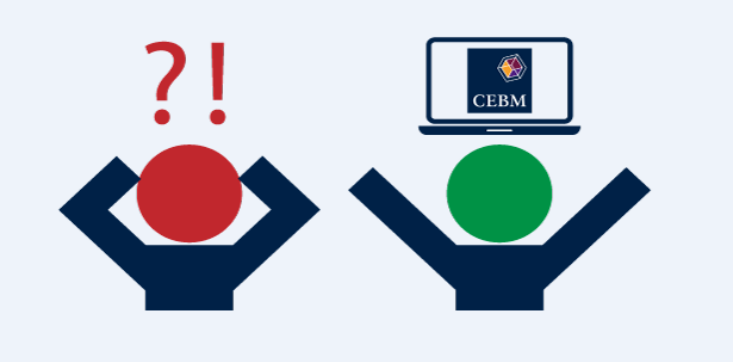
Tip for data extraction for meta-analysis - 25



What if you're missing a mean and only a similar statistical statistic is given?

Kathy Taylor

Previously, I highlighted a list of ways where, when extracting data for meta-analysis of continuous outcomes, you might find that a summary statistic that you want is missing. In this post I'll give some examples of the 3rd way - **a similar summary statistic is reported, but it's not the statistical measure that you want** - when you have missing means.

Finding a median reported

You may find that instead of a mean, a median is reported. A median is a different type of average. The reporting of medians indicates that the distribution of outcome data is skewed. The median and mean are equal if the distribution of the data is perfectly symmetrical (Figure 1). When the distribution is skewed, the mean and median will differ, and the difference between them will depend on the degree of skewness.

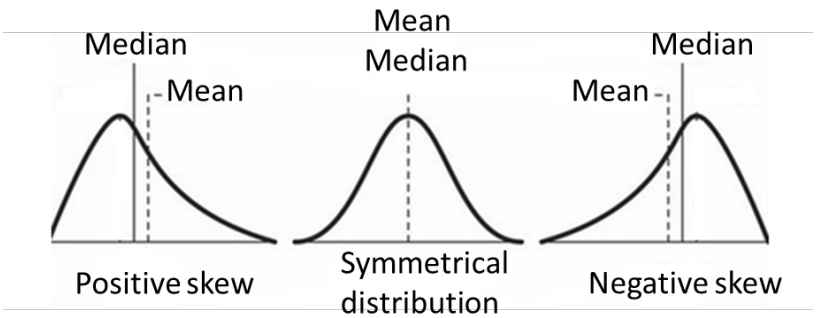


Figure 1. The relative position of the mean and median depending on the data's distribution

The Cochrane Handbook (section 6.5.2.9) highlights three papers which provide equations for estimating means from other summary statistics.

One is the paper by Hozo et al who concluded that even with skewed data, the sample mean can be estimated by the median. They provided estimates for the mean based on the sample size and range (min and max are the bounds of the range):

$mean \approx median$ if n ≥25

$mean \approx \frac{min+2median+max}{4}$ if n<25

The curved equal sign means 'approximately equal'. The estimates were tested using simulation, and drawing samples from normal and skewed distributions.

The second paper is by Bland who uses more information by providing estimates of the sample mean based on the median, range, sample size and interquartile range (q_1 to q_2):

$$\frac{min(n+3)+2(n-1)(q_1+median+q_3)+\max(n+3)}{8n}$$

When n is large, this equation simplifies to

$$\frac{min+2(q_1+median+q_3)+\max}{8}$$

He tested his estimates on three real data sets with simulated data drawing samples from normal and skewed distributions.

The formulae of Bland and Hozo et al both work better with small samples.

The third paper is by Wan et al who provide estimates of the sample mean based on the median and interquartile range. This has the advantage of not being influenced by extreme values.

$$\frac{q_1+median+q_3}{3}$$

Using simulation, they also tested their estimates, drawing samples from normal and skewed distributions and found smaller relative errors compared to Bland's approach. Wan et al also provide a very useful spreadsheet which you can use to calculate and compare their estimated means with those of Bland and Hozo et al.

Another paper by Luo et al provide improved estimates of the sample mean based on the sample size, median, range and interquartile range. Their estimates use a weighted formulation. They consider three scenarios.

Firstly, when the sample size, median and range are reported.

$$\left(\frac{4}{4+n^{0.75}}\right)\left(\frac{min+max}{2}\right)+\left(\frac{n^{0.75}}{4+n^{0.75}}\right)median$$

Secondly, when the sample size, median and interquartile range are reported.

$$\left(0.7+\frac{0.39}{n}\right)\left(\frac{q_1+q_3}{2}\right)+\left(0.3-\frac{0.39}{n}\right)median$$

Thirdly, when the sample size, median, range and interquartile range are reported.

$$\left(\frac{2.2}{2.2+n^{0.75}}\right)\left(\frac{min+max}{2}\right)+\left(0.7-\frac{0.72}{n^{0.55}}\right)\left(\frac{q_1+q_3}{2}\right)+\left(0.3+\frac{0.72}{n^{0.55}}-\frac{2.2}{2.2+n^{0.75}}\right)median$$

These are approximations of their more complicated formulae that are reported in their paper. They demonstrate the accuracy of their estimates using simulation and they provide an excel spreadsheet.

Finding a geometric mean reported

Sometimes a geometric mean is reported. This is another type of average, which arises from the analysis of skewed data which have been log-transformed and then back-transformed (using the exponential function) when presenting results. With small samples, skewed data is often log-transformed, before analysis, because standard inferences on the means of skewed data is only acceptable for large samples. With large samples we assume that the means of outcome measurements are approximately normally distributed due to the central limit theorem.

So, instead of means (which are more formally known as arithmetic means) and standard deviations (SD), geometric means are reported, either with confidence intervals (CIs), the exponential of the SD of the log-transformed values (often referred to as the tolerance factor or the inappropriately named as the 'SD of the geometric mean'), or the exponential of the standard error (SE) of the log-transformed values. Geometric means and arithmetic means should not be pooled. If most of your studies report arithmetic means, you will want to convert geometric mean summary data to arithmetic mean summary data. Pooling is possible by using the conversion equations of Higgins et al. It's a two-stage process (Figure 2) as the geometric mean data has to be log-transformed first.

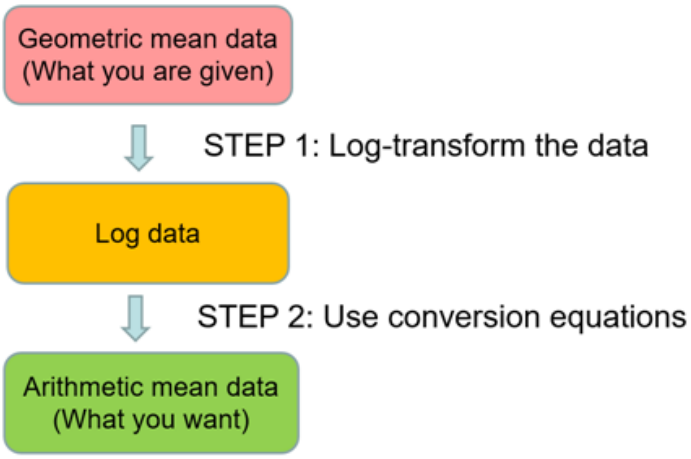


Figure 2. From geometric means to arithmetic means

Using the following notation:

g is the geometric mean of a treatment arm
(g_{lower} to g_{upper}) is the confidence interval of g
e^{Sz} is the incorrectly named SD of g
e^{SE_z} is the incorrectly named SE of g

STEP 1

Calculate the log-transformed measurements $(\bar{z}\ and\ s_z)$ from the geometric mean data.

$\bar{z} = ln(g)$ and

$s_z = \frac{(\ln(g_{upper}) - \ln(g_{lower}))\sqrt{n}}{2t}$ OR

$s_z = ln(e^{Sz})$ in cases where e^{Sz} has been reported OR

$s_z = ln\left(e^{\sqrt{n} \times SE_z}\right)$

where t is the 97.5 percentage point of the t-distribution with (n-1) degrees of freedom.

STEP 2

Apply the conversion equations to the log-transformed data to calculate the arithmetic mean summary data $(\bar{x}\ and\ s_x)$. There are two sets of equations depending on the similarities between the SDs of the two treatment arms. Higgins et al recommend comparing the SDs on the log scale as it's more plausible. If the SDs are different, use Method 1. If the SDs are similar, use Method 2.

Method 1

For each treatment arm, calculate

$$\bar{x} = exp\left(\bar{z} + \frac{s_z^2}{2}\right)$$

$$s_x = \sqrt{(exp(s_z^2) - 1)exp(2\bar{z} + s_z^2)}$$

Method 2

First calculate

$$s_{z,pooled} = \sqrt{\frac{(n_1 - 1)s_{z,1}^2 + (n_2 - 1)s_{z,2}^2}{n_1 + n_2 - 2}}$$

Then, for each treatment arm, calculate

$$\bar{x} = exp\left(\bar{z} + \frac{s_{z,pooled}}{2}\right)$$

$$s_x = \sqrt{(exp(s_{z,pooled}^2) - 1)exp(2\bar{z} + s_{z,pooled}^2)}$$

Higgins et al also provide equations to convert the other way, from arithmetic means to geometric means (Figure 3). You might want this if the majority of your included studies report geometric means.

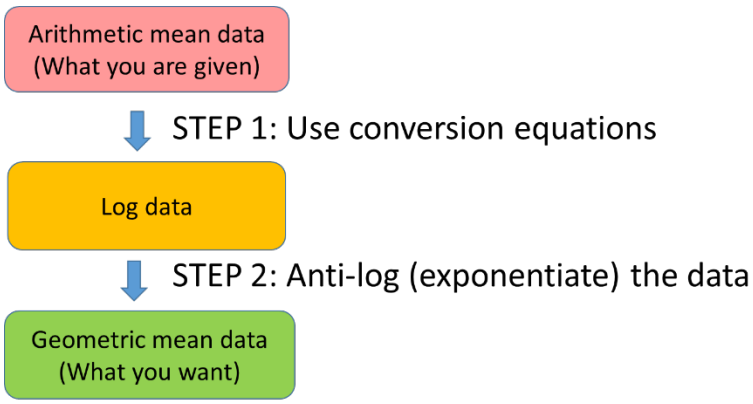


Figure 3. From arithmetic means to geometric means

STEP 1

Convert the arithmetic mean summary data to log summary data.

$\bar{x}$ and s_x are the arithmetic mean and SD

$\bar{z}$ and s_z are the mean and SD of the log data

If the SDs are different, use Method 1. If the SDs are similar, use Method 2.

Method 1

For each treatment arm, calculate

$$\bar{z} = ln(\bar{x}) - \frac{1}{2} ln\left(\frac{s_x^2}{\bar{x}^2} + 1\right)$$

$$s_z = \sqrt{ln\left(\frac{s_x^2}{\bar{x}^2} + 1\right)}$$

Method 2

For each treatment arm, calculate

$$s_z = \sqrt{ln\left(\frac{s_x^2}{\bar{x}^2} + 1\right)}$$

Then calculate

$$s_{z,pooled} = \sqrt{\frac{(n_1 - 1)s_{z,1}^2 + (n_2 - 1)s_{z,2}^2}{n_1 + n_2 - 2}}$$

Then for each treatment arm, calculate

$$\bar{z} = ln(\bar{x}) - \frac{1}{2} s_{z,pooled}^2$$

STEP 2

Back-transforming (exponentiating) the log data calculates the geometric mean data.

$\bar{z}$ and s_z are the log data.

g $= e^{\bar{z}}$ and e^{s_z} are the geometric mean data.

Meta-analysis can be carried out on the log scale and SD for the log values can be calculated using the following equation:

$$SD = \frac{(upper\ CI - lower\ CI)}{3.92}\sqrt{n}$$

Let me show you an example of converting geometric mean data to arithmetic mean data.

For the review that I worked on we used these conversion equations when we extracted data from the study by Romero et al. They reported microalbuminuria (albumin excretion rate) at 6 months with geometric mean (95% CI). Data for the intervention group, treated with

Captopril, was converted from 60 (35 to 104) mg/24hr to mean (SD) of 90 (101) mg/24hr, and data for the untreated group was converted from 91 (58 to 141) mg/24hr was converted to 119 (101) mg/24hr (Table). I applied the equations from Method 1.

Table. Calculating arithmetic mean data

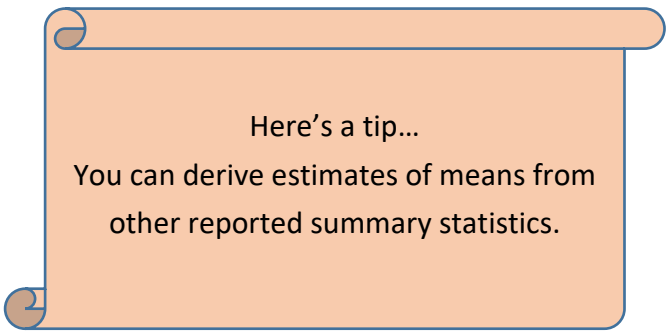
	Captopril	Untreated
n	13	13
gm	60	91
g lower	35	58
g upper	104	141
z=Ln(g)	4.09	4.51
dof	12	12
t-value	2.18	2.18
Sz	0.90	0.74
x	90.05	119.22
Sx	100.77	100.91

See below to find out where the t-value came from. Higgins et al highlight that their estimates are likely to be biased in small sample studies. As this study had small samples, it was removed as part of a sensitivity analysis.

Other methods

Other approaches of dealing with missing means highlighted by the review of Weir et al include the simulation-based approximate Bayesian computation (ABC) approach of Kwon and Reis.

If a large proportion of studies have missing means, pooling is not recommended.



In my next blog post I'll give some more examples of **a similar summary statistic is reported, but it's not the statistical measure that you want** when you have missing SDs.

Where did the equations and t-value come from?

Converting geometric means to arithmetic means and the reverse

Higgins et al derive their equations in their paper.

Calculating means from medians, range and interquartile range

Hozo et al, Bland, Wan et al and Luo et al also derive their equations in their respective papers.

Calculating an SD from a 95% confidence interval:

This was derived earlier and in my next blog post I will give more details.

What about the t-value?

This came from a t-distribution table (Figure 4).

dof	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
130	0.254	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Figure 4. t-distribution table

The first column of the table shows the degrees of freedom (dof) and the area probabilities (also known as percentages or p-values) are shown in the first row. As indicated in the t-

distribution curve above the table, the p-values represent the area under the t-distribution curve in the tail, from the t-value to infinity (shaded black) for different dofs.

	p values
dof	t-values

In the example I gave, the Captopril group had n=13 so dof=n-1=12.
The area probability for the 97.5 percentage point of the t-distribution using the above table =1-0.975=0.025. The corresponding t-value is 2.179 (Figure 5).

dof	0.4	0.25	0.1	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.599
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.215	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
70	0.254	0.678	1.294	1.667	1.994	2.381	2.648	2.899	3.211	3.435
130	0.254	0.676	1.288	1.657	1.978	2.355	2.614	2.856	3.154	3.367
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Figure 5. t-distribution table for dof=12 and p-value=0.025

The same t-value can also be calculated using EXCEL by typing in an EXCEL cell =ABS(T.INV(one sided p-value, dof)). With my example this is =ABS((T.INV(0.025,12).

Note that I'm using EXCEL 2016. Earlier versions use the term TINV(two-sided p-value, dof) i.e. TINV(0.05,12)=ABS(T.INV(0.025,12)=2.178813.

Dr Kathy Taylor teaches data extraction in Meta-analysis, https://www.conted.ox.ac.uk/courses/meta-analysis</link> This is a short course that is also available as part of our MSc in Evidence-Based Health Care https://www.conted.ox.ac.uk/about/msc-in-evidence-based-health-care, MSc in Medical Statistics https://www.conted.ox.ac.uk/about/msc-in-ebhc-medical-statistics, and MSc in Systematic Reviews https://www.conted.ox.ac.uk/about/msc-in-ebhc-systematic-reviews

Follow me on Twitter @dataextips for updates on my blog, related news, and to find out about further examples where others, like me, are trying to make statistics more broadly accessible.