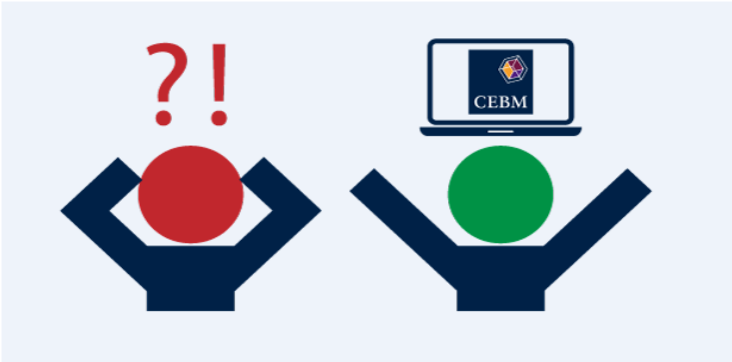
Tip for data extraction for meta-analysis – C3



What if the data I want are reported for the wrong group?

Kathy Taylor

Previously, in post C1, 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 focus on the 2nd case - **the summary statistic you want is reported, but it's for the wrong group.**

Wanting summary data for the combined group

Sometimes you may find that sample sizes, means and standard deviations (SDs) are reported for subgroups based on patient characteristics (e.g. hypertensive or normotensive), or treatment (e.g. same drug at different doses), but you want these summary statistics for the combined population. Provided the means, SDs and the numbers in each subgroup are reported, you can derive the summary statistics for the combined group (the 'right' group) by using a bit of maths (see below if you're interested):

Summary statistic	*Group 1*	*Group 2*	*Combined group*
N	n_1	n_2	$n_1 + n_2$
Mean	m_1	m_2	$\frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$
SD	sd_1	sd_2	$\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2 + \frac{n_1 n_2}{n_1 + n_2}(m_1^2 + m_2^2 - 2m_1 m_2)}{n_1 + n_2 - 1}}$
Percentage	p_1	p_2	$\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$

Remember that 'A multiplied B' is shown as 'AB'.

An approximation (and slight underestimate) of the pooled standard deviation (SD) is the usual pooled SD

$$\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}$$

The usual pooled SD will be useful in cases where you know the mean of the combined population but you don't know the means of the subgroups.

If there are more than two subgroups, the combined group equations may be applied sequentially. For example, if the subgroups were normotensive (group 1), hypertensive (group 2) and hypotensive (group 3), the above equations could be used to calculate the summary data for groups 1 and 2, and then summary data for all patients could be derived by pooling the data for the combined group 1+2 and group 3.

Let me show you an example using trial data where two intervention groups have been assigned two doses of the same intervention drug. Grant et al report the effects of high and medium dose of metformin on the cardiovascular risk in a population with type II diabetes. They report a drop (mean±SD) in HbAiC levels in the high-dose group (n=14) of 1.2%±1.3% and in the medium-dose group (n=13) of 0.9% ±1.0%. Using the combined group equations the drop in HbAiC in the combined population (n=27) is calculated as 1.1%±1.2%.

Summary statistic	*Group 1 (high dose)*	*Group 2 (low dose)*	*Combined group (Metformin)*
N	*14*	*13*	27
mean	*1.2*	*0.9*	1.1
SD	*1.3*	*1.1*	1.2

If there are incomplete data, the group equations cannot be applied. For example, if only the total number of patients is reported, the numbers in each treatment group cannot be derived. I will deal with this and other cases of incomplete data in a future blog post.

Wanting summary data for a particular subgroup

Another situation of 'wrong' data being reported is where you're interested in extracting subgroup summary statistics (e.g. for people with hypertension), and although the study has considered that population in a subgroup analysis, and reports some summary subgroup data, the particular data you want have not been reported. If complete summary data are provided for all patients and the complementary subgroup to what you want (in this case, it might be people who are normotensive), you can calculate the 'correct' summary data by using another set of equations, which are simply the group equations rearranged.

Summary statistic	*Combined group*	*Group 2*	*Group 1*
N	n_c	n_2	$n_1 = n_c - n_2$
mean	m_c	m_2	$m_1 = \frac{n_c m_c - n_2 m_2}{n_c - n_2}$
SD	sd_c	sd_2	$\sqrt{\frac{(n_c - 1)sd_c^2 - (n_2 - 1)sd_2^2 - \frac{n_c n_2}{n_c - n_2}(m_c^2 + m_2^2 - 2m_c m_2)}{n_c - n_2 - 1}}$
*Percentage**	p_c	p_2	$\frac{n_c p_c - n_2 p_2}{n_c - n_2}$

Remember that A multiplied by B is shown as AB.

Let me show you an example of data are reported where the subgroups are according the level of albumin in the urine - normoalbuminuria (normal levels) and microalbuminuria (elevated levels) in a trial of an anti-hypertensive, Losartan, verses usual care. Sawaki et al report the baseline urinary albumin creatinine ratio of the intervention group (n=14) at 61.7±79.9 mg/g, and at 130.3±81.5mg/g in the intervention group with microalbuminuria (n=6). Using the rearranged group equations, the baseline data for the intervention group (n=8) with normoalbuminuria are derived as 10.3±7.3mg/g.

Summary statistic	*Losartan combined group*	*Losartan group 2 (microalbuminuria)*	*Losartan group 1 (normoalbuminuria)*
N	*14*	*6*	*8*
mean	*61.7*	*130.3*	*10.3*
SD	*79.9*	*81.5*	*7.3*

Here's a tip…
You can use the group equations to pool summary statistics of subgroups or the rearranged group equations to calculate data for a particular subgroup.

In my next post I'll show a worked example to illustrate how the change score and endpoint equations given previously in post C2 can complement the group and rearranged group equations when calculating summary statistics that are not reported.

Where did the equations come from?

The mean of a sample is the average value $\frac{\sum_1^n x_i}{n}$

where

$\sum x_i$ represents the sum of all the values in the sample $x_1, x_2, x_3 \dots x_n$

n is the total number of observations.

The standard deviation (SD) of the sample is

$$\sqrt{\frac{\sum_1^n (x_i-\mu)^2}{n-1}} = \sqrt{\frac{\sum_1^n (x_i^2+\mu^2-2x_i\mu)}{n-1}} = \sqrt{\frac{\sum_1^n x_i^2-n\mu^2}{n-1}}$$ **equation 0**

where

μ is the mean of the sample.

Group 1 has n_1 values $x_1, x_2, x_3 \dots x_{n_1}$ and mean m_1

Group 2 has n_2 values $y_1, y_2, y_3 \dots y_{n_2}$ and mean m_2

To derive the equation for the combined mean and SD		
Mean of group 1	$m_1 = \frac{\sum_1^{n_1} x_i}{n_1}$	**equation 1**
Mean of group 2	$m_2 = \frac{\sum_1^{n_2} y_i}{n_2}$	**equation 2**
Mean of combined group	$m_c = \frac{\sum_1^{n_1} x_i + \sum_1^{n_2} y_i}{n_1 + n_2}$	**equation 3**
From equation 1	$m_1 n_1 = \sum_1^{n_1} x_i$	**equation 4**
From equation 2	$m_2 n_2 = \sum_1^{n_2} y_i$	**equation 5**
Substitute equations 4 and 5 into equation 3	$m_c = \frac{m_1 n_1 + m_2 n_2}{n_1 + n_2}$	**equation 6**
Use equation 0 to calculate the SD of group 2 and square this SD	$sd_1^2 = \frac{\sum_1^{n_1} x_i^2 - n_1 m_1^2}{n_1 - 1}$	
Rearrange for $\sum_1^{n_1} x_i^2$	$\sum_1^{n_1} x_i^2 = (n_1 - 1) sd_1^2 + n_1 m_1^2$	**equation 7**
Square the SD of group 2	$sd_1^2 = \frac{\sum_1^{n_1} x_i^2 - n_1 m_1^2}{n_1 - 1}$	
Rearrange for $\sum_1^{n_2} y_i^2$	$\sum_1^{n_2} y_i^2 = (n_2 - 1) sd_2^2 + n_2 m_2^2$	**equation 8**

Use equation 0 to calculate the SD of the combined group and square this SD	$sd_c^2 = \frac{\sum_1^{n_1} x_i^2 + \sum_1^{n_2} y_i^2 - (n_1 + n_2)m_c^2}{n_1 + n_2 - 1}$	**equation 9**
Substitute equations 6, 7 and 8 in equation 9. Tidy up (several terms cancel out).	$sd_c^2 =$ $\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2 + \frac{n_1 n_2}{n_1 + n_2}(m_1^2 + m_2^2 - 2m_1 m_2)}{n_1 + n_2 - 1}}$	**equation 10**

To derive the combined probability equation		
Probability of event in group 1	$p_1 = \frac{e_1}{n_1}$	
Rearrange	$p_1 n_1 = e_1$	**equation 11**
Probability of event in group 2	$p_2 = \frac{e_2}{n_2}$	
Rearrange	$p_2 n_2 = e_2$	**equation 12**
Probability of the combined group	$p_c = \frac{total\ events}{total\ number}$	**equation 13**
Substitute equations 11 and 12 in equation 13	$p_c = \frac{p_1 n_1 + p_2 n_2}{n_1 + n_2}$	**equation 14**

To derive the equations for the mean and SD of group 1		
Number in group 1	$n_1 = n_c - n_2$	**equation 15**
Substitute equation 15 in equation 6 and rearrange	$m_1 = \frac{n_c m_c - n_2 m_2}{n_c - n_2}$	**equation 16**
Substitute equations 15 and 16 in equation 10 and rearrange	$sd_1^2 =$ $\sqrt{\frac{(n_c - 1)sd_c^2 - (n_2 - 1)sd_2^2 - \frac{n_c n_2}{n_c - n_2}(m_c^2 + m_2^2 - 2m_c m_2)}{n_c - n_2 - 1}}$	**equation 17**
To derive equations for the mean and SD of group 2		
Swop the subscripts "1" and "2" in equation 17	$sd_2^2 =$ $\sqrt{\frac{(n_c - 1)sd_c^2 - (n_1 - 1)sd_1^2 - \frac{n_c n_1}{n_c - n_1}(m_c^2 + m_1^2 - 2m_c m_1)}{n_c - n_1 - 1}}$	
To derive the equation for the probability of group 1		
Substitute equation 15 in equation 14 and rearrange	$p_1 = \frac{n_c p_c - n_2 p_2}{n_c - n_2}$	**equation 18**
To derive the equation for the probability of group 1		
Swop the subscripts "1" and "2" in equation 18	$p_2 = \frac{n_c p_c - n_2 p_1}{n_c - n_1}$	

Dr Kathy Taylor teaches data extraction in Meta-analysis. This is a short course that is also available as part of our MSc in Evidence-Based Health Care, MSc in EBHC Medical Statistics, and MSc in EBHC Systematic Reviews.

Follow updates on this blog, related news, and to find out about other examples of statistics being made more broadly accessible on Twitter @dataextips