22 META-ANALYSIS

INTRODUCTION

Anytime someone thinks a question is important enough to invest time and money in answering it, it's a safe bet that others think the same thing. The clearest example of this point is seen in political polling. As an election approaches, we are bombarded with polls that report the candidate preferences of voters. Even if polls are conducted properly, the estimates will vary from sample to sample because of sampling error. How do we make sense of these variable estimates of the same parameter? In this chapter, we'll see that many individual estimates can be averaged to produce a more precise or accurate estimate of the population parameter.

Although political polling may be the first example that comes to mind, it is not the only situation in which we might wish to average point estimates from different sources. For example, many studies may have estimated (i) the proportion of adults that exhibit psychiatric symptoms, (ii) the average level of depression of Scandinavians, (iii) the average number of years required to complete a graduate degree in psychology, or (iv) the average area of the primary visual cortex in humans. Meta-analysis is the name given to a family of methods that average information from several different sources.

In previous chapters, we considered the estimation of a single parameter from a single sample. In most cases, the procedure required calculating a point estimate of the parameter (i.e., a statistic) and its estimated standard error. We computed $(1-\alpha)100\%$ confidence intervals by putting together the point estimate and estimated standard error with the level of confidence we wished to have. In this chapter, we will see that combining results from two or more studies is very much like combining the results of two or more scores. Rather than treating the scores of individuals as the data to be analyzed, meta-analysis treats the statistics of samples (e.g., means or effect sizes) as the data to be analyzed.

BASICS OF META-ANALYSIS

Primary and Secondary Literature

Research journals publish two kinds of papers. Most papers report original research. The term "original" means that new data have been collected and reported in the paper. The data may have been collected to address a question that has been of interest to others or a novel question that has not been previously studied. The focus here is not on whether papers address an original question but rather whether the papers contain original data. Papers reporting original data are considered part of the **primary literature**.

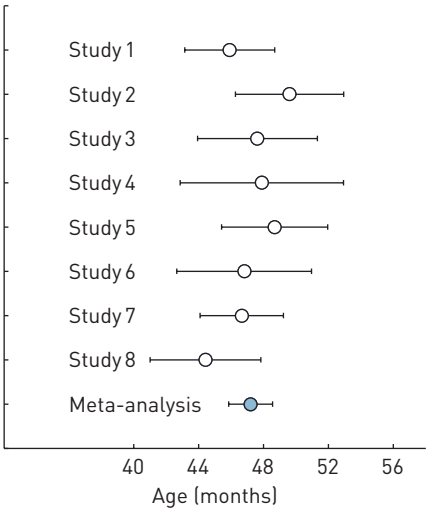
Papers that review the primary literature in one way or another make up the **secondary literature**. **Meta-analysis** is a quantitative method within the secondary literature that combines results from the primary literature to provide an answer to a question of interest

The **primary literature** comprises papers reporting original data.

The **secondary literature** comprises papers that combine the results of reports in the primary literature.

Meta-analysis is a quantitative approach that combines several results from the primary literature.

FIGURE 22.1 ■ An Example of Meta-Analysis



The means of eight studies are shown along with their 95% confidence intervals. The blue dot shows the mean of the eight study means; i.e., the meta-mean, $M = \Sigma m/k$. The confidence interval around the meta-mean is $M \pm t_{\alpha/2}(s_M)$. A meta-mean will fall closer to the population mean, on average, than the means of individual studies. In addition, the confidence interval around a meta-mean will be narrower, on average, than confidence intervals around the means of individual studies.

that is a more precise, or accurate, than the answers provided by individual papers.

An Example Combining Means

We will begin with an example to illustrate the basic features of meta-analysis, using the following question: "At what age do children first count to 10?" One could pose this question to a random sample of parents and use their responses to estimate the population parameter. Of course, retrospective self-report measures of this sort are prone to error (e.g., memory) and bias (e.g., parents wishing to have their child seem special). Despite such methodological difficulties, it's easy to see that this might be a developmental marker of interest to psychologists. Therefore, to illustrate meta-analysis, we will imagine that the primary literature contains many studies that have estimated the mean age at which children in their samples first count to 10.

Later in this chapter, we will address important issues of how studies are identified and selected for analysis from the primary literature. First, however, we will discuss a simple statistical conceptualization underlying meta-analysis. This conceptualization (or statistical model) will seem familiar because it is the same model used in previous chapters. This simplified approach to meta-analysis assumes that the children in each of the studies selected for analysis are simply different random samples from the same population. If several samples have been drawn from the same population, then the means of these samples can be averaged, and a confidence interval can be placed around the *mean of means*. A mean of means will be a more precise estimate of the mean of the population being sampled; therefore, the confidence interval around a mean of means will be narrower, on average, than the confidence intervals placed around individual sample means.

The **meta-mean** (M) is a statistic that is the mean of a number of sample means.

Figure 22.1 illustrates the ideas described above. Each white dot represents a sample mean, and the intervals around the sample means are 95% confidence intervals. The blue dot is the mean of the eight sample means. We refer to the mean of means as a **meta-mean** and denote it with a capital M. The arms around M represent the 95% confidence interval. The confidence interval around M shows that it is a more precise estimator than any of the individual sample means. Furthermore, meta-means will fall closer to the population mean μ, on average, than will sample means. We next turn to the computation of M and the confidence interval around it.

Computing a Confidence Interval Around *M*

Table 22.1 shows the results of the eight hypothetical studies depicted in Figure 22.1. Each study estimates the mean age (in months) at which children first count to 10, and each is given an arbitrary number from 1 to 8, shown in the first column. The second column shows the mean obtained in each study. Finally, columns 3 and 4 show sample variances and sample sizes, respectively.

We can compute a confidence interval around M using exactly the same calculations we used to place a confidence interval around m. The only difference is that means are the units of analysis in the case of meta-analysis, whereas scores were the units of analysis in previous chapters.

The calculation of a meta-mean (M) is identical to the calculation of a sample mean:

$$M = \sum_{i=1}^{k} m_i / k. \tag{22.1}$$

This equation defines a mean exactly as in Chapter 2. However, a few frills have been added here. We saw summation signs like these in Chapters 18 and 19, but we will take a moment to review them here. The k at the top of the summation sign is the number of means being averaged. The letter i is called an index, and it takes on integer values from 1 to k. The values of i provide a way to refer to each of the k sample means ($m_1, m_2, \ldots, m_k$) being averaged. (The motivation for using indexes will become clearer as we move along.) For the data in Table 22.1, $k = 8$; however, in other cases, k could be any number greater than 1. You can read equation 22.1 as follows: Sum the k sample means $m_1, m_2, \ldots, m_k$ and divide this sum by k to create a mean of sample means. Call this mean of means M.

TABLE 22.1 ■ Hypothetical Data

i	m_i	s_i^2	n_i
1	45	98	20
2	50	70	60
3	48	70	30
4	48	63	40
5	49	84	20
6	47	112	10
7	46	56	10
8	43	21	10
	M		
	47	8.47	

The capital letter M emphasizes that the mean in question is a mean of means. M estimates μ, the mean of the population of scores.

Just as we can compute the variance for a collection of scores, we can compute the variance for a collection of means:

$$S^2 = \frac{\sum_{i=1}^{k} (m_i - M)^2}{k-1}. \tag{22.2}$$

This variance is computed in the same way as the variance of a sample. The only difference is that the squared differences are computed from $(m_i - M)$, rather than $(y_i - m)$. Note that the S has been capitalized in S^2 to emphasize that it was computed from means rather than scores.

The mean of means (M) is a statistic, which means that it is different for each sample of k means, just as m is different for each sample of n scores. Therefore, as a statistic, M has a sampling distribution with a mean and a variance. The mean of the sampling distribution of M is μ, the population mean. The variance of the distribution of M is σ_M^2, which is estimated as follows:

$$s_M^2 = S^2/k. \tag{22.3}$$

So, just as $s_m^2 = s^2/n$ estimates the variance of the distribution of m (σ_m^2), s_M^2 estimates the variance of the distribution of M (σ_M^2). If we take the square root of s_M^2, we obtain the estimated standard error of M:

$$s_M = \sqrt{s_M^2}. \tag{22.4}$$

Once we have the estimated standard error of M, we can compute the confidence interval as

$$M \pm t_{\alpha/2}(s_M). \tag{22.5}$$

In this case, we have $k-1 = 7$ degrees of freedom. Our confidence in this interval comes from knowing that $(1-\alpha)100\%$ of all such intervals will capture μ.

An Example Calculation

Let's now step through the calculation of $M \pm t_{\alpha/2}(s_M)$ for the data in Table 22.1. When we apply equation 22.1 to these means we obtain

$$M = \sum_{i=1}^{k} m_i/k = 47.$$

TABLE 22.2 ■ Computing the Sum of Squares

i	m_i	$m_i - M$	$(m_i - M)^2$
1	45	−2	4
2	50	3	9
3	48	1	1
4	48	1	1
5	49	2	4
6	47	0	0
7	46	−1	1
8	43	−4	16
Sum	376	0	36
Mean	47		

You should be able to confirm this using a calculator or Excel. Next, we compute the variance of the sample means using equation 22.2. To do this, we first compute the sum of squared deviations about M as shown in Table 22.2, and we find in this case that $ss = 36$. To compute the variance, we simply divide the sum of squares by $k-1$, where k is the number of sample means. The calculation is as follows:

$$S^2 = \frac{\sum_{i=1}^{k}(m_i - M)^2}{k-1} = \frac{36}{7} = 5.1429.$$

Now that the routine variance calculations are completed, we need only compute the estimated standard error of M. First we compute the square of the estimated standard error of M according to equation 22.3:

$$s_M^2 = S^2/k = 5.1429/8 = 0.6429.$$

We then compute the estimated standard error itself according to equation 22.4:

$$s_M = \sqrt{s_M^2} = \sqrt{0.6429} = 0.8018.$$

The last step is to compute a confidence interval around M. As is customary, we will compute the 95% confidence interval, using equation 22.5. There are $k-1 = 7$ df; so from the t-table, we find that $t_{\alpha/2} = 2.365$. Therefore, the 95% confidence interval is

$$\text{CI} = M \pm t_{\alpha/2}(S_M) = 47 \pm 2.365(0.8018) = [45.10, 48.90].$$

The confidence limits have been rounded to two decimal places.

So that's all there is to this simple form of meta-analysis. We've used exactly the same calculations that were used for confidence intervals around sample means. The only change is that means have replaced scores in our calculations. As a consequence, some of the parameters estimated change, even though the structure and logic of the calculations are the same. Table 22.3 summarizes the parallels between confidence intervals for sample means and confidence intervals in a meta-analysis.

Although we ignored the sample sizes and sample variances in the example above, there is an alternative method that makes use of these quantities when computing s_M. However, this method involves assumptions that are almost never true in practice. The method

TABLE 22.3 ■ The Parallels Between Confidence Intervals for *m* and *M*

CI for a Mean: $m \pm t_{\alpha/2}(s_M)$		CI for a Meta-Mean: $M \pm t_{\alpha/2}(s_M)$	
Statistic and Formula	**Parameter**	**Statistic and Formula**	**Parameter**
$m = \sum_{i=1}^{n} y_i / n$	μ	$M = \sum_{i=1}^{k} m_i / k$	μ
$s^2 = \sum_{i=1}^{n} (y_i - m)^2 / (n-1)$	σ^2	$S^2 = \sum_{i=1}^{k} (m_i - M)^2 / (k-1)$	*
$s_m^2 = s^2 / n$	σ_m^2	$s_M^2 = S^2 / k$	σ_M^2
$s_m = \sqrt{s_m^2}$	σ_m	$s_M = \sqrt{s_M^2}$	σ_M

*The parameter estimated by S^2 is discussed in Appendix 22.1.

discussed above does not involve these implausible assumptions and is therefore more general. There is a brief discussion of the alternative method in Appendix 22.1. In the next section, we will make use of sample sizes in our calculation of the confidence interval around *M*, while continuing to ignore the sample variances.

LEARNING CHECK 1

1. State whether the following statements are true or false.
   (a) A report of three original experiments is part of the primary literature.
   (b) A report of three experiments replicating a classic study in psychology is part of the primary literature.
   (c) A report that combines 25 replications of a classic study in psychology is part of the primary literature.
   (d) *M* is a parameter.
   (e) The expected value of *M* is μ.
2. $m = \{45, 60, 75\}$ is a collection of means drawn from the primary literature. Compute *M* and the 95% confidence interval around *M*.

Answers

1. (a) True. (b) True. (c) False. (d) False. (e) True.
2. $M = 60$, $s^2 = 225$, $s_M = \sqrt{S^2/k} = 8.6603$. $M \pm t_{\alpha/2}(s_M) = 60 \pm 4.303(8.6603) = [22.73, 97.27]$.

META-ANALYSIS FOR SAMPLES OF DIFFERENT SIZES

Because meta-analysts don't collect the data reported in the primary literature, sample size is certain to vary from study to study. So, does this make any difference? The answer depends on how different the sample sizes are and how many sample means are being combined. Because larger samples provide more precise point estimates of μ, it would seem sensible to (somehow) give more weight to the means of large samples when computing

M. (Remember that the same logic was used to compute s^2_{pooled} in Chapter 11.) If there are many samples of roughly the same size, then weighting means according to sample size probably makes little difference. If there are few means, computed from samples that differ greatly in size, then weighting means according to sample size will make a big difference. Therefore, the safest course of action is to always take sample size into account. But how?

Computing Means as Weighted Sums

Let's review our definition of *M*:

$$M = \sum_{i=1}^{k} m_i / k. \tag{22.6a}$$

By now we're familiar with subscripts, and in the following paragraphs we'll see why they can be useful. Although we haven't remarked on this before, there are actually two ways to read equation 22.6a. The first says sum $m_1, m_2, \ldots, m_k$ and then divide this sum by k. To make this explicit, we can put the summation within parentheses as follows:

$$M = \left(\sum_{i=1}^{k} m_i\right) / k. \tag{22.6b}$$

On the other hand, we could read equation 22.6a to say divide each of the means $m_1, m_2, \ldots, m_k$ by k and then sum these fractions. To make this explicit, we can put the division within parentheses as follows:

$$M = \sum_{i=1}^{k} (m_i / k). \tag{22.6c}$$

If you do an example for yourself (e.g., using the numbers [3, 6, 27]), you will see that these two interpretations of equation 22.6a produce exactly the same result, so the ambiguity makes no difference. For present purposes, however, equation 22.6c is very useful. Remember that dividing by k is the same thing as multiplying by $1/k$. Therefore, the equation

$$M = \sum_{i=1}^{k} \frac{1}{k} * m_i \tag{22.6d}$$

produces exactly the same result as the previous three equations. These points might seem trivial, but they allow us to define *M* as a **weighted sum** and thus deal with the common situation in which sample sizes are different.

A **weighted sum** is a sum of numbers that have been multiplied by a weight. The sum of all weights equals 1.

We refer to the following quantity as *weight*:

$$w_i = \frac{1}{k}.$$

When each of the k sample means is multiplied by w_i, these weighted means can be summed to define *M* as follows:

$$M = \sum_{i=1}^{k} w_i * m_i\,. \tag{22.7a}$$

Equation 22.7a tells us to multiply each mean (m_i) by its corresponding weight (w_i) and then sum the products.

Weights can also be described more generally as

$$w_i = \frac{n_i}{\sum_{i=1}^{k} n_i}.$$

For example, if there are $k = 8$ means with $n = 5$ scores contributing to each mean, then

$$w_i = \frac{n_i}{\sum_{i=1}^{k} n_i} = \frac{5}{5*8} = \frac{1}{8} = \frac{1}{k}.$$

In this case, a weight can be seen to represent the number of scores associated with the ith mean (n_i) as a proportion of the total number of scores associated with the k means ($\sum_{i=1}^{k} n_i$). We can see from this perspective that a weight can be defined very simply even if sample sizes are different.

There is no need for a separate step to compute weights, which are then applied to the mean. Instead, computing a weighted mean can be accomplished in one simple step as follows:

$$M = \frac{\sum_{i=1}^{k} n_i * m_i}{\sum_{i=1}^{k} n_i}. \tag{22.7b}$$

That is, we multiply each sample mean by its corresponding sample size and then divide the sum of these products by the total number of scores in all samples combined. Therefore, we can use equation 22.7b to compute M as a weighted sum when sample sizes are different (or the same). This way of computing the mean as a weighted sum should feel familiar, because this is exactly how we computed GPA in the appendices of Chapter 1.

Computing *M* as a Weighted Sum for Unequal Sample Sizes

Table 22.4 illustrates the calculation of M as a weighted sum. The sample means (in column 2) and sample sizes (column 3) have been taken from Table 22.1. At the bottom of column 3, we see that the sum of the sample sizes is $\sum_{i=1}^{k} n_i = 200$, so M will be based on 200 scores. The fourth column shows the product of each mean with its associated sample size ($n_i * m_i$). The sum of these products is $\sum_{i=1}^{k} n_i * m_i = 9600$. Therefore, using equation 22.7b, we find that

$$M = \frac{\sum_{i=1}^{k} n_i * m_i}{\sum_{i=1}^{k} n_i} = \frac{9600}{200} = 48.$$

Before moving on, I want to reiterate that equations 22.7a and 22.7b produce exactly the same results. In the calculations above, we used equation 22.7b to compute M. If we had used equation 22.7a, we would first compute weights as $w_i = n_i / \sum_{i=1}^{k} n_i$, as shown in the fifth column of Table 22.4. In this case, we can see that the $n_1 = 20$ scores associated with m_1 represent 10% of the total number of scores in the meta-analysis; i.e., $w_i = n_i / \sum_{i=1}^{k} n_i = 20/200 = .1$. When we multiply each mean by its corresponding weight, we obtain the products shown in the last column of Table 22.4. The sum of these products is $M = \sum_{i=1}^{k} w_i * m_i = 48$, as shown at the bottom of the last column. Therefore, equation 22.7b is just a computationally simpler version of equation 22.7a.

We've seen that $M = 47$ when computed without considering sample size (bottom of column 2) and $M = 48$ when computed as a weighted sum (bottom of column 6). So, why are these two means different? Consider the second mean in Table 22.4, $m_2 = 50$. This is the

TABLE 22.4 ■ Computing the Mean as a Weighted Sum When Sample Sizes Are Unequal

i	m_i	n_i	$n_i * m_i$	$w_i = n_i / \sum_{i=1}^{k} n_i$	$w_i * m_i$
1	45	20	900	0.10	4.50
2	50	60	3000	0.30	15.00
3	48	30	1440	0.15	7.20
4	48	40	1920	0.20	9.60
5	49	20	980	0.10	4.90
6	47	10	470	0.05	2.35
7	46	10	460	0.05	2.30
8	43	10	430	0.05	2.15
	$\sum_{i=1}^{k} n/k$	$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i * m_i$	$\sum_{i=1}^{k} w_i$	$\sum_{i=1}^{k} w_i * m_i$
	47	200	9600	1	48

largest mean in the sample of eight means, and it also has the largest sample size, $n_2 = 60$. In contrast, the smallest mean $m_8 = 43$ is associated with the smallest sample size, $n = 10$. Therefore, when M is computed as a weighted sum, more weight is given to m_2 than to m_8. As a consequence, the mean computed as a weighted sum (48) is larger than when all means are given equal weight (47). Because means of large samples are more precise estimators of μ than means of small samples, the mean of means based on the weighted sum will be closer to μ, on average, than a mean of means that gives equal weight to all samples.

Computing S^2 as a Weighted Sum for Unequal Sample Sizes

We now turn to the question of computing our estimate of σ_M^2 when sample sizes are different. Previously we used equation 22.2 to compute the variance about M:

$$S^2 = \frac{\sum_{i=1}^{k} (m_i - M)^2}{k-1}.$$

This quantity can be computed as a weighted sum as follows:

$$S^2 = \frac{\sum_{i=1}^{k} n_i (m_i - M)^2}{\sum_{i=1}^{k} n_i} \frac{k}{k-1}. \tag{22.8}$$

In equation 22.8, squared deviations from M [i.e., $(m_i - M)^2$] are multiplied by their corresponding sample sizes and then the sum of these squared deviations is divided by the number of scores ($\sum_{i=1}^{k} n_i$). This quantity is a biased estimator and so it is multiplied by $k/(k-1)$ to correct the bias. (Please see Appendix 22.1 for a discussion of the parameter estimated by S^2.)

Table 22.5 illustrates the calculation of $\sum_{i=1}^{k} n_i * (m_i - M)$. The first two columns identify the studies (i) and study means (m_i). Column three shows the squared deviation of each

sample mean (m_i) from $M = 48$, which was computed as a weighted sum. Column 4 shows sample sizes, and column 5 shows the squared differences multiplied by sample size. Finally, the bottom of column 5 shows the sum of weighted squared differences (740), which is the first term in equation 22.8.

The calculation of s_M^2 continues as follows. First, calculate

$$S^2 = \frac{\sum_{i=1}^{k} n_i (m_i - M)^2}{\sum_{i=1}^{k} n_i} \frac{k}{k-1} = \frac{740}{200} * \frac{8}{7} = 4.229.$$

Then compute the square of the estimated standard error using equation 22.3:

$$s_M^2 = S^2/k = 4.229/8 = 0.529.$$

Finally, compute the estimated standard error itself using equation 22.4:

$$s_M = \sqrt{s_M^2} = \sqrt{0.529} = 0.727.$$

TABLE 22.5 ■ Computing S^2

i	m_i	$(m_i - M)^2$	n_i	$n_i * (m_i - M)^2$
1	45	9	20	180
2	50	4	60	240
3	48	0	30	0
4	48	0	40	0
5	49	1	20	20
6	47	1	10	10
7	46	4	10	40
8	43	25	10	250
			$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i * (m_i - M)^2$
			200	740

The last step is to compute a confidence interval around M. As before, we will compute the 95% confidence interval, using equation 22.5. There are k–1 df; from the t-table, we find that $t_{\alpha/2} = 2.365$. Therefore, the 95% confidence interval is

$$\text{CI} = M \pm t_{\alpha/2}(s_M) = 48 \pm 2.365(0.727) = [46.28, 49.72].$$

We have 95% confidence in this interval because we know that 95% of all intervals computed this way will capture μ. An example of these calculations performed in Excel is given in Appendix 22.2.

A Computational Shortcut

When computing S^2 using equation 22.8, we multiplied the weighted sum by $k/(k-1)$ to eliminate bias; from this unbiased quantity, we computed $s_M^2 = S^2/k$. We can compute s_M^2 more simply using the following formula:

$$s_M^2 = \frac{\sum_{i=1}^{k} n_i (m_i - M)^2}{(k-1) * \sum_{i=1}^{k} n_i}.$$

We won't step through why this equation is equivalent to $s_M^2 = S^2/k$, but if you substitute in the numbers computed above, you will see that they are.

Discussion

We already noted that point estimates in meta-analysis depend on whether sample sizes are taken into account. Because larger samples are better estimates of μ, M computed as a weighted sum will fall closer to μ, on average, than M computed as a simple (unweighted) mean. That is, M computed as a weighted sum is a better point estimate of μ.

Similar comments hold for confidence intervals. Assuming that all samples are drawn from the same population, the estimated standard errors will be smaller, on average, when

sample size is taken into account. This is because the largest samples will produce means that fall closer to M, on average. Because large samples make a greater contribution to M, the squared deviations of the large sample means from M will be smaller, and these smaller deviations will be given greater weight. For these reasons, as we noted above, sample size should be taken into account when conducting a meta-analysis. That is, M and s_M^2 should be computed using weighted sums.

As a final point, the meta-analysis described above can be used to combine many different statistics. We will reuse the symbols M and S^2 in all of these contexts to avoid a proliferation of symbols. This will keep things simpler in one sense but will also require some memory work to keep track of what is being estimated in each case. We will see an example of this point in a later section.

*Alternative Methods of Meta-Analysis

The method of meta-analysis described above is a modest variant of a method described by Hunter and Schmidt (1990), which is widely used in the meta-analysis literature. An alternative method that is at least as widely used was described by Hedges and Olkin (1985). The main difference between the Hunter-Schmidt and Hedges-Olkin methods is that Hedges and Olkin make use of the sample variances to compute the weights. In this method, the weights applied to the sample means are given by

$$w_i = \frac{1}{s_i^2/n_i}.$$

The denominator of this equation is simply the square of the estimated standard error of the mean for the ith sample mean; i.e., $s_{m_i}^2 = s_i^2/n_i$. The logic is that samples with smaller variances are more precise estimates of μ and are therefore given greater weight. So, both sample size and sample variance affect the weight given to a sample mean.

A more complete description of the Hedges and Olkin (1985) method is given in Appendix 22.3. In the following sections, we will continue with the Hunter-Schmidt method because it is a simple extension of the confidence intervals used in all previous chapters, and because it seems to provide confidence intervals that are generally similar to those provided by the Hedges-Olkin method (e.g., Hafdahl & Williams, 2009).

LEARNING CHECK 2

1. Why should we compute M and S^2 as weighted sums?
2. If m = [10, 20, 20, 50, 30] are the means of five samples, and n = [60, 40, 60, 20, 20] are the sample sizes, compute M and S^2 as weighted sums.
3. Compute the 95% confidence interval around M.

Answers

1. M estimates μ. Because the means of larger samples are better estimators of μ than the means of smaller samples, weighting samples means according to sample size will make M a better estimator of μ. Similar comments apply to S^2, which will be smaller, on average, when computed as a weighted sum.
2. $M = 21$, $S^2 = 161.25$.
3. CI $= M \pm t_{\alpha/2}(s_M) = 21 \pm 2.776(5.6789) = 21 \pm 15.7672 =$ [5.24, 36.76].

FIXED-EFFECTS VERSUS RANDOM-EFFECTS MODELS

The model described above assumed that all samples were drawn from *the same population of scores*; hence, all sample means estimate the same population mean. This conceptualization is sometimes referred to as an instance of the **fixed-effects model**. It is particularly simple to think about, and one might even wonder whether another conceptualization is possible.

The **fixed-effects model,** in its simplest form, assumes that all sample means in a meta-analysis represent different random samples from the same distribution.

There is an alternative to the fixed-effects model, however. It is called the **random-effects model**. Rather than assume that all samples were drawn from the same population, the random-effects model permits the possibility that samples were drawn from different populations. Although the random-effects model is slightly more complicated theoretically, it is more realistic than the fixed-effects model. Fortunately, despite this additional theoretical complexity, the Hunter-Schmidt meta-analysis is performed exactly as with the fixed-effects model. (This is not true for the Hedges-Olkin model described in Appendix 22.3.)

The **random-effects model** allows for the possibility that sample means in a meta-analysis represent random samples drawn from different distributions.

If we continue with the counting-to-10 example, then the means combined in a meta-analysis are very likely to have come from research labs in different parts of the world. The essence of the random-effects model is that we don't assume that the scores obtained in each part of the world come from distributions having identical parameters.

Figure 22.2 illustrates the concepts underlying the random-effects model. There are 16 distributions of age-of-counting-to-10 scores. To make things concrete, each distribution is associated with an American city. Although only 16 distributions are shown (because of space limitations), it's possible for there to be an unlimited number of such distributions. Each population of ages has a mean and a standard deviation.

In the random-effects model, we think of there being a distribution of population means. In the 16 distributions in Figure 16.2, the mean of each population is shown. This distribution of population means itself has a mean, which we call μ_{Meta}, and variance, which we call σ^2_{Meta}. These are defined as

$$\mu_{\text{Meta}} = \frac{\sum_{i=1}^{N} \mu_i}{N}$$

and

$$\sigma^2_{\text{Meta}} = \frac{\sum_{i=1}^{N} (\mu_i - \mu_\mu)^2}{N}.$$

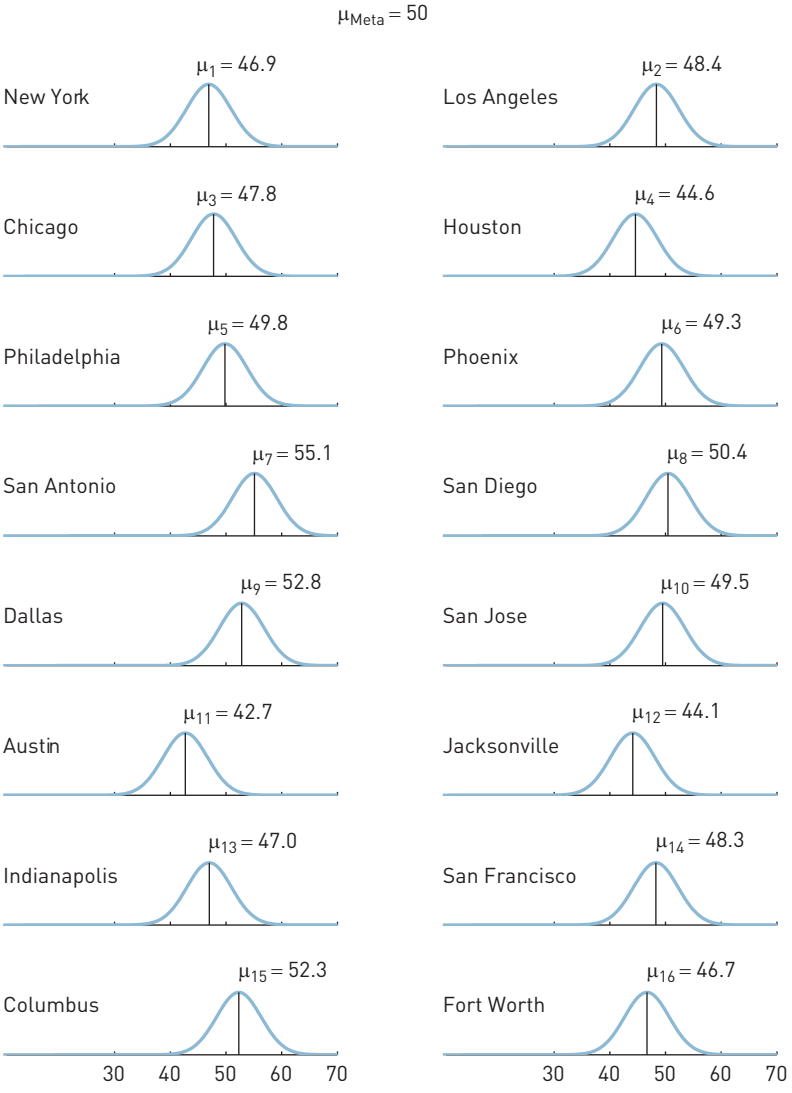
In these formulas, N is the number of populations.

In the random-effects model, the mean of sample means, M, no longer estimates the mean of a single population, as it does in the fixed-effects model. Rather, M estimates μ_{Meta}, the mean of all population means. (This is an example of the symbol M being used to denote a statistic that estimates different parameters.)

The concepts behind the random-effects model are actually very similar to the concepts underlying estimates of population means from sample means. Each score contributing to a population mean is subject to measurement error. For example, in Chapter 1 we noted that each time we measure a person's height, we will come up with a slightly different measurement, no matter how accurate our measuring device is. It is just the nature of measurement to have measurement error. So, when we measure a person's height, this measurement is really an estimate of the mean of all possible measurements we could take. Therefore, when we average the scores of individuals, we are really averaging estimates of each individual's mean score. The random-effects model is very similar. Each sample mean estimates its own population mean. Therefore, when we average the means of samples, we are really averaging estimates of each population's mean.

Furthermore, when we draw a sample of scores from a population, we don't assume that we've drawn every single score from the population. This is true by definition because a sample is a subset of a population. The same point applies to meta-analysis. The means that

FIGURE 22.2 ■ Sixteen Distributions of Scores



In all distributions, the scores are age-of-counting-to-10, in months. The mean of each population is show (above the peak of the distribution). These 16 distributions are a subset of a much larger number of distributions. (Imagine similar distributions for each city in the United States, North America, or the world.) The mean of all population means is $\mu_{Meta} = 50$.

we've drawn are a sample of the possible means that could have been drawn. Therefore, when we think of the sampling distribution of *M*, we do not assume that we are repeatedly sampling from the same populations, such as the 16 cities shown in Figure 22.2. Rather, the sampling distribution of *M* represents means of a theoretically infinite number of populations whose means vary about μ_{Meta} with standard deviation σ_{Meta}.

Although the fixed- and random-effects models estimate different parameters (μ in the case of the fixed-effects model, and μ_{Meta} in the case of the random-effects model), the calculation in both cases is $CI = M \pm t_{\alpha/2}(s_M)$. If the fixed-effects model correctly describes the underlying situation, then our confidence interval would be an interval estimate of μ. If the random-effects model correctly describes the underlying situation, then our confidence interval would be an interval estimate of μ_{Meta}.

SIGNIFICANCE TESTS

Our focus has been, as always, on estimating a population parameter. As we've seen in previous chapters, confidence intervals can be used to test any null hypothesis of interest. In the meta-analysis described in the previous section, the 95% CI for M was [46.28, 49.72]. Let's now assume that M estimates μ_{Meta}, which represents the mean of population means for age-of-counting-to-10 scores in the United States. If one happened to know that μ_{Meta} in Canada was 45, then we could set up a hypothesis test to assess whether or not 45 is a plausible value for μ_{Meta} in the United States. Our hypotheses would be

$$H_0: \mu_{Meta} = 45;$$

$$H_1: \mu_{Meta} \neq 45.$$

Because the interval [46.28, 49.72] does not include $\mu_{\mu} = 45$, we would reject H_0 and conclude that 45 is an implausible value for μ_{Meta} in the United States. We would say that the result is statistically significant. Once again, statistically significant just means that if H_0 were true, it would be very unusual for M to be so far from μ_{Meta} that the confidence interval around M does not capture it. In our example, if $\mu_{USA} = \mu_{Canada}$, it would be very unusual for M to be so far from μ_{Canada} that the 95% confidence interval around M does not capture it.

What do these results mean? If the difference between μ_{Canada} and μ_{USA} is statistically significant, is this a good thing, a bad thing, or what? If the age of counting to 10 is greater in the United States than Canada, should there be a call to action to get children counting earlier in the United States, or does it reflect a healthy attitude toward child rearing, in which children are not pushed to attain counting skills too early? Furthermore, is the difference between $\mu_{Meta} = 45$ and $M = 48$ really that important? Maybe yes, maybe no. These are questions that have nothing to do with statistics. Statistics simply provide guidance for human decision makers.

If educators in the United States interpret these results to mean that more efforts should be devoted to getting children to count earlier, then the question becomes how to do this. A variety of methods can be tried and compared. Judgments about which of these methods works best and, perhaps, is most cost-effective require further study. Questions like these (i.e., which of two methods yields the greater change in age-of-counting-to-10 scores) were the subject of Chapters 11 and 12.

LEARNING CHECK 3

1. State whether the following statements are true or false.
   (a) The assumptions of the random-effects model and fixed-effects model are identical.
   (b) Meta-analyses are performed the same way using the random-effects model and the fixed-effects model.
   (c) The fixed-effects model assumes a potentially infinite number of population means.
   (d) M estimates μ_{Meta} in the fixed-effects model.
   (e) $\mu_{Meta} = \sum_{i=1}^{N} \mu_i / N$.
   (f) $\sigma^2_{Meta} = \sum_{i=1}^{N} (\mu_i - \mu_{Meta})^2$.
2. m = [10, 20, 20, 50, 30] and n = [60, 40, 60, 20, 20]. Compute the 95% confidence interval around M. Test the null hypothesis H_0: $\mu_{Meta} = 35$.

Answers

1. (a) False. (b) True. (c) False. (d) False. (e) True. (f) False.
2. CI = $M \pm t_{\alpha/2}(s_M) = 21 \pm 2.776(5.6789)$
   $= 21 \pm 15.7672 = [5.24, 36.76]$.
   We retain H_0 because the 95% confidence interval includes 35.

META-ANALYSIS COMBINING EFFECT SIZES

The preceding discussion showed a very general approach to meta-analysis in which we combined estimates of a single population mean (fixed-effects model) or the mean of population means (random-effects model). However, as noted at the beginning of Part 3, researchers are most often interested in the association between variables; this could be the difference between two population means or the correlation between two variables. We will now apply meta-analysis to effect sizes derived from the difference between two means.

The simplest question we can ask in clinical science is "How well does some treatment work?" This question typically takes the form of asking how much better a treatment works than a placebo. Turner, Matthews, Linardatos, Tell, and Rosenthal (2008) considered the weight of evidence relating to the effectiveness of pharmacological treatments for depression. They performed a meta-analysis of the effect sizes found in 74 relevant studies, some of which were published and others of which were not. Turner et al. (2008) made many important observations about what does and does not make it into the scientific literature, and these points will be addressed later. For the moment, however, we will think about the meta-analyses they conducted on effect sizes. We will see that meta-analyses that combine effect sizes can be conducted in exactly the same way as meta-analyses that combine means.

Many of the studies considered by Turner et al. (2008) made use of the Hamilton Depression Rating Scale (HDRS); see Figure 22.3. This is a 17-item, clinician-administered rating scale. For each of the 17 items, a clinician rates the patient on a 3- to 5-point scale to indicate the severity of the associated symptom of depression. The minimum score is 0 and the maximum score is 55. Scores between 0 and 7 are considered to be in the normal range and scores above 20 indicate moderate to severe depression.

At the end of Chapter 11, we noted that dependent variables and statistics may differ from study to study. Therefore, other measures of depression are available including the Beck Depression Inventory (BDI), the Montgomery-Åsberg Depression Rating Scale (MADRS), or the number of days of work missed following the intervention. However, for simplicity, we will consider eight hypothetical studies that employed the HDRS as the primary measure of depression. In these eight hypothetical studies, the mean of the placebo group is expected to be higher than the mean of the treatment group; i.e., the prediction is that $\mu_{placebo} - \mu_{treatment} > 0$, or $\delta > 0$. Therefore, for this illustration, positive estimates of δ (i.e., d) would mean an improvement, because the treated group(s) would show lower HDRS scores.

Figure 22.4 follows the same format as Figure 22.1 and summarizes the results of the eight hypothetical studies. The white and light blue dots represent the results of unpublished and published studies, respectively, along with their 95% confidence intervals, and the dark blue dot shows the weighted mean (M) of the eight individual effect sizes, along with its 95% confidence interval. As always, M is, on average, a more precise estimate of the effect than any of the individual estimates.

Estimating δ or δ_{Meta}

The distinction between the fixed- and random-effects models of meta-analysis applies equally to estimating population means and effect sizes. In the fixed-effects model, all values of d estimate the same δ. This means that the eight placebo samples were all drawn from the same population and the eight treatment samples were all drawn from the same population. In the random-effects model, all values of d estimate a different δ. This means that each study drew scores from different treatment and placebo populations. The meta-analysis estimates the mean

FIGURE 22.3 ■ Hamilton Depression Rating Scale

1. DEPRESSED MOOD (Sadness, hopeless, helpless, worthless)
0 = Absent
1 = These feelings are indicated only on questioning
2 = These feelings are spontaneously reported verbally
3 = Communicates feelings non-verbally i.e., through facial expression, posture, voice, and tendency to weep
4 = Patient reports VIRTUALLY ONLY these feelings in his spontaneous verbal and non-verbal communication

2. FEELINGS OF GUILT
0 = Absent
1 = Self reproach, feels he has let people down
2 = Ideas of guilt or rumination over past errors or sinful deed
3 = Present illness is a punishment. Delusions of guilt
4 = Hears accusatory or denunciatory voices and/or experiences threatening visual hallucinations

3. SUICIDE
0 = Absent
1 = Feels life is not worth living
2 = Wishes he were dead or any thoughts of possible death to self
3 = Suicide ideas or gesture
4 = Attempts at suicide (any serious attempt rates)

4. INSOMNIA EARLY
0 = No difficulty falling asleep
1 = Complains of occasional difficulty falling asleep - more than 1/2 hour
2 = Complains of nightly difficulty falling asleep

5. INSOMNIA MIDDLE
0 = No difficulty
1 = Patient complains of being restless and disturbed during the night
2 = Waking during the night - any getting out of bed (except for purposes of voiding)

6. INSOMNIA LATE
1 = No difficulty
2 = Waking in early hours of the morning but goes back to sleep
3 = Unable to fall asleep again if he gets out of bed

7. WORK AND ACTIVITIES
0 = No difficulty
1 = Thoughts and feelings of incapacity, fatigue or weakness related to activities (work or hobbies)
2 = Loss of interest in activities (hobbies or work) - either directly reported by patient, or indirectly in listlessness, indecision and vacillation (feels he has to push himself to work or do activities)
3 = Decrease in actual time spent in activities or decrease in productivity. In hospital, if patient does not spend at least three hours a day in activities (hospital job or hobbies) exclusive of ward chores
4 = Stopped working because of present illness. In hospital, if patient engages in no activities except ward chores, or if patient fails to perform ward chores unassisted

8. RETARDATION: PSYCHOMOTOR (Slowness of thought and speech; impaired ability to concentrate; decreased motor activity)
0 = Normal speech and thought
1 = Slight retardation at interview
2 = Obvious retardation at interview
3 = Interview difficult
4 = Complete stupor

9. AGITATION
0 = None
1 = Fidgetiness
2 = Playing with hands, hair, etc.
3= Moving about, can't sit still
4 = Hand wringing, nail biting, hair-pulling, biting of lips

10. ANXIETY: PSYCHIC
0 = No difficulty
1 = Subjective tension and irritability
2 = Worrying about minor matters
3 = Apprehensive attitude apparent in face or speech
4 = Fears expressed without questioning

11. ANXIETY: SOMATIC (Physiological concomitants of anxiety, such as - Gastro-intestinal: dry mouth, wind, indigestion, diarrhea, cramps, belching. - Cardio-vascular: palpitations, headaches. - Respiratory: hyperventilation, sighing. - Urinary frequency - Sweating)
0 = Absent
1 = Mild
2 = Moderate
3 = Severe
4 = Incapacitating

12. SOMATIC SYMPTOMS: GASTROINTESTINAL
0 = None
1 = Loss of appetite but eating without staff encouragement. Heavy feelings in abdomen
2 = Difficulty eating without staff urging. Requests or requires laxatives or medication for bowels or medication for gastro-intestinal symptoms

13. SOMATIC SYMPTOMS: GENERAL
0 = None
1 = Heaviness in limbs, back or head. Backaches, headache, muscle aches. Loss of energy and fatigability
2 = Any clear-cut symptom

14. GENITAL SYMPTOMS (loss of libido, menstrual disturbances)
0 = Absent
1 = Mild
2 = Severe

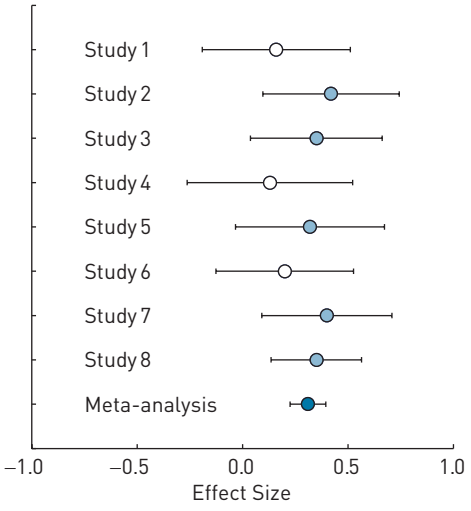
15. HYPOCHONDRIASIS
0 = Not present
1 = Self-absorption (bodily)
2 = Preoccupation with health
3 = Frequent complaints, requests for help, etc.
4 = Hypochondriacal delusions

16. LOSS OF WEIGHT
0 = No weight loss
1 = Probable weight loss associated with present illness (>500g/week)
2 = Definite weight loss(>1kg/week)

17. INSIGHT
0 = Not depressed (based on above items) OR Acknowledges being depressed and ill
1 = Acknowledges illness but attributes cause to bad food, climate, overwork, virus, need for rest, etc.
2 = Denies being ill at all

The Hamilton Depression Rating Scale (HDRS) is a clinician-administered rating scale. The clinician rates the patient on each item to indicate the severity of the associated symptom of depression. The minimum score is 0 and the maximum score is 55. Scores between 0 and 7 are considered to be in the normal range and scores above 20 indicate moderate to severe depression.

FIGURE 22.4 ■ Hypothetical Data



Effect sizes from eight studies submitted to a meta-analysis. White dots correspond to effect sizes from unpublished studies, light blue dots correspond to effect sizes from published studies, and the dark blue dot represents the result of a meta-analysis that combines the results of all eight studies.

of many different population δs, which we call δ_{Meta}. As in the case of meta-analysis for a single mean, the calculations are the same whether we are estimating δ or δ_{Meta}.

We will use exactly the same terminology as before to conduct the meta-analysis. Equation 22.9 computes the average effect size (M) as a weighted sum:

$$M = \frac{\sum_{i=1}^{k} n_i * d_i}{\sum_{i=1}^{k} n_i}. \tag{22.9}$$

Equation 22.10 computes s_M as a weighted sum:

$$s_M = \sqrt{\frac{\sum_{i=1}^{k} n_i (d_i - M)^2}{(k-1)\sum_{i=1}^{k} n_i}}. \tag{22.10}$$

Converting Statistics to *d*

Table 22.6 summarizes the data plotted in Figure 22.4. It was noted in Chapter 11 that studies addressing the same general question may use different dependent variables (e.g., the HDRS, the BDI, the MADRS, or number of days of work missed) and different statistics (e.g., d, t_{obs}, r^2, or confidence intervals). We will assume in this example that all eight studies used the same dependent variable. However, to provide an illustration of the real-world problem of combining different statistics, Table 22.6 shows four different statistics (d, t_{obs}, r^2, and 95% confidence intervals) that may be computed from the same dependent variable. In this section, we will review how to convert t_{obs} and r^2 to d. These conversions were introduced in Chapter 11. We will also see how to extract information from confidence intervals to then compute d.

TABLE 22.6 ■ Hypothetical Data Combined in a Meta-Analysis

Study	Statistic	Value	$n_{treatment}$	$n_{placebo}$	Expressed as *d*
Study 1*	d	0.16	62	63	0.16
Study 2	t_{obs}	2.57	75	75	0.42
Study 3	r^2	0.0301	82	78	0.35
Study 4*	95% CI	[–2, 3.93]	48	52	0.13
Study 5	d	0.32	65	60	0.32
Study 6*	t_{obs}	1.93	71	74	0.20
Study 7	r^2	0.0389	84	81	0.40
Study 8	95% CI	[1, 4.12]	165	175	0.35

*Denotes unpublished studies.

Converting t_{obs} to d

It was shown in Chapter 11 that t_{obs} is converted to d as follows:

$$d = t_{obs}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}. \tag{22.11a}$$

An alternative version that avoids the need to compute two separate fractions is

$$d = t_{obs}\sqrt{\frac{n_1+n_2}{n_1 n_2}}. \tag{22.11b}$$

Equation 22.11b allows one to defer rounding until the last step, and thus produces a more accurate result. When applied to the results of Study 2 in Table 22.6, $t_{obs} = 2.57$ is converted to d as follows:

$$d = t_{obs}\sqrt{\frac{n_1+n_2}{n_1 n_2}} = 2.57\sqrt{\frac{75+75}{75*75}} = 0.42.$$

Converting r^2 to d

In Chapter 11, it was also shown that r^2 is converted to d as follows:

$$d = \sqrt{df_{within}\left(\frac{r^2}{1-r^2}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}. \tag{22.12a}$$

An alternative version of this equation also avoids the need to compute and then add two fractions:

$$d = \sqrt{df_{within}\left(\frac{r^2}{1-r^2}\right)\left(\frac{n_1+n_2}{n_1 n_2}\right)}. \tag{22.12b}$$

When applied to the results of Study 3 in Table 22.6, $r^2 = 0.0301$ is converted to d as follows:

$$\begin{aligned} d &= \sqrt{df_{within}\left(\frac{r^2}{1-r^2}\right)\left(\frac{n_1+n_2}{n_1 n_2}\right)} \\ &= \sqrt{158\left(\frac{.0301}{.9699}\right)\left(\frac{82+78}{82*78}\right)} \\ &= .35. \end{aligned}$$

Converting $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1-m_2})$ to d

We've noted many times that confidence intervals are the most useful and most general method of data reporting. Unfortunately, if you haven't been given the two sample means and variances, then converting a confidence interval to d requires deconstructing the confidence interval to recover the quantities needed to compute d. Although there are several steps in this process, they are not complicated. Stepping through this process helps to consolidate our understanding of the components of the confidence interval. The definition of d in this case is

$$d = \frac{m_1 - m_2}{s_{pooled}}.$$

To compute this quantity, we will have to recover $m_1 - m_2$ and s_{pooled} from a confidence interval.

We start by assuming we've been given only the confidence limits $(m_1 - m_2) \pm t_{\alpha/2}(s_{m_1 - m_2})$. We denote the upper limit as upper and the lower limit as lower. The difference between the two sample means is simply the middle point of this interval. So, if we take the average, we find

$$m_1 - m_2 = [\text{upper} + \text{lower}]/2. \tag{22.13}$$

Recovering s_{pooled} from the confidence interval requires three steps. We start by recovering the margin of error (*moe*), which is defined as $t_{\alpha/2}(s_{m_1 - m_2})$. The *moe* is simply the difference between $m_1 - m_2$ and one of the confidence limits. It can be computed as follows:

$$moe = \text{upper} - (m_1 - m_2). \tag{22.14}$$

Because $moe = t_{\alpha/2}(s_{m_1 - m_2})$, we can recover $s_{m_1 - m_2}$ by dividing *moe* by $t_{\alpha/2}$ as follows:

$$s_{m_1 - m_2} = \frac{moe}{t_{\alpha/2}}. \tag{22.15}$$

Finding $t_{\alpha/2}$ is not complicated. Assuming we know the sizes of the two samples, we can compute df_{within} as $n_1 + n_2 - 2$. Assuming also that we know the level of confidence (e.g., 95%), we can find $t_{\alpha/2}$ using **T.INV.2T** in Excel (see Appendix 10.1). Using Excel, we find that $t_{\alpha/2} = \textbf{T.INV.2T}(\alpha, \textbf{df}_{\textbf{within}})$. The penultimate step is to compute s_{pooled}, which is accomplished as follows:

$$s_{\text{pooled}} = s_{m_1 - m_2}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}. \tag{22.16}$$

Having deconstructed the confidence interval to recover $m_1 - m_2$ and s_{pooled}, we can compute the estimated effect size as

$$d = \frac{m_1 - m_2}{s_{\text{pooled}}}.$$

The steps just outlined will be applied to the confidence interval shown in row 4 of Table 22.6. There we see that the 95% confidence interval for $m_1 - m_2$ is [–2, 3.93], based on samples of size 48 and 52. Using this information, the following steps allow us to recover *d*.

Step 1. Compute $m_1 - m_2$. This requires knowing the upper and lower limits of the confidence interval.

$$m_1 - m_2 = [3.93 + -2]/2 = 1.93/2 = .965.$$

Step 2. Compute the margin of error (*moe*). This requires knowing the upper limit of the confidence interval and the center of the interval, $m_1 - m_2$.

$$moe = upper - (m_1 - m_2) = 3.93 - 0.965 = 2.965.$$

Step 3. Compute the estimated standard error of $m_1 - m_2$ ($s_{m_1 - m_2}$). This requires knowing the *moe* and $t_{\alpha/2}$. In this case, there are $48 + 52 - 2 = 98$ degrees of freedom. Using Excel, we can find that $t_{\alpha/2} = 1.984$. Therefore,

$$s_{m_1 - m_2} = \frac{moe}{t_{\alpha/2}} = \frac{2.965}{1.984} = 1.494.$$

Step 4. Compute s_{pooled} from $s_{m_1 - m_2}$ as follows:

$$s_{\text{pooled}} = s_{m_1 - m_2} \sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 1.494 \sqrt{\frac{48 * 52}{48 + 52}} = 7.464.$$

Step 5. Compute d from $m_1 - m_2$ and s_{pooled} in the usual way:

$$d = \frac{m_1 - m_2}{s_{\text{pooled}}} = \frac{0.965}{7.464} = 0.13.$$

Calculating the Confidence Interval

Once all statistics have been converted to d, it is straightforward to compute the mean effect size (M) as a weighted sum. Table 22.7 shows these calculations. The first column shows indexes that denote the eight studies. The second column shows the estimated effect sizes, and the third column shows the number of scores in the corresponding study; i.e., $n_i = n_{\text{treatment}} + n_{\text{placebo}}$. The last column shows the products of n_i and d_i. The sum of these products is $\sum_{i=1}^{k} n_i * d_i = 406$ and the total number of scores is $\sum_{i=1}^{k} n_i = 1310$. Therefore, $M = 406/1310 = 0.3099$, which we'll round to 0.31.

Once M has been computed, equation 22.8b (the computational shortcut) can be used to compute s_M^2. In Table 22.8, the study indexes and estimated effect sizes are shown in the first

TABLE 22.7 ■ Computing the Mean Effect Size as a Weighted Sum

i	d_i	n_i	$n_i * d_i$
1	0.16	125	20.00
2	0.42	150	63.00
3	0.35	160	56.00
4	0.13	100	13.00
5	0.32	125	40.00
6	0.20	145	29.00
7	0.40	165	66.00
8	0.35	340	119.00
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i * d_i$
		1310	406.000

TABLE 22.8 ■ Computing the Mean Squared Deviation From $M = 0.31$

i	d_i	$(d_i - M)^2$	n_i	$n_i * (d_i - M)^2$
1	0.16	0.0225	125	2.813
2	0.42	0.0121	150	1.815
3	0.35	0.0016	160	0.256
4	0.13	0.0324	100	3.240
5	0.32	0.0001	125	0.013
6	0.20	0.0121	145	1.755
7	0.40	0.0081	165	1.337
8	0.35	0.0016	340	0.544
			$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i * (d_i - M)^2$
			1310	11.7710

and second columns. The third column computes the squared deviation of each d_i from $M = 0.31$; i.e., $(d_i - M)^2$. The fourth column shows the sample sizes and the fifth column shows the products of n_i and $(d_i - M)^2$. The sum of these products is $\sum_{i=1}^{k} n_i * (d_i - M)^2 = 11.77$ and the total number of scores is $\sum_{i=1}^{k} n_i = 1310$. From these numbers, we can compute s_M as follows:

$$s_M = \sqrt{\frac{\sum_{i=1}^{k} n_i (d_i - M)^2}{(k-1) * \sum_{i=1}^{k} n_i}} = \sqrt{\frac{11.77}{7 * 1310}} = 0.0361.$$

The last step is to compute a 95% confidence interval around M using equation 22.5. There are $k-1 = 8 - 1 = 7$ *df*, and from the *t*-table, we find that $t_{\alpha/2} = 2.365$. Therefore, the 95% confidence interval is

$$\text{CI} = M \pm t_{\alpha/2}(s_M) = 0.31 \pm 2.365(0.0361) = [0.22, 0.40].$$

We have 95% confidence in this interval because we know that 95% of all such intervals will capture μ_δ.

The meta-analysis procedures described above are very general. We could go on to show how to conduct a meta-analysis for correlation coefficients, as discussed in Chapter 15. However, the procedure is exactly like the two examples already shown, so nothing new would be covered.

The Validity of Meta-Analysis Depends on the Available Data

It was mentioned earlier that Turner et al. (2008) made some important observations about what studies are published in the research literature. In fact, investigating which studies are published and which aren't was the purpose of their meta-analysis. To understand their study, we must first review the testing process that drugs must undergo before they are approved for sale in the United States.

In the United States, the Food and Drug Administration (FDA) is responsible for assessing whether drugs are safe and effective. As a final stage of the approval process, drugs must undergo clinical trials. According to Turner et al. (2008), FDA approval requires, in essence, statistical superiority of a candidate drug over a placebo in two well-controlled studies. That is, there must be a statistically significant improvement over a placebo.

Clinical trials must be registered with the FDA. A contract of sorts is made in which the nature of the experiment is agreed on, as is the nature of the primary outcomes. For example, scores on the HDRS might be primary outcomes and scores on other measures such as the BDI might be secondary outcomes. The FDA maintains records of all clinical trials and their results, which can be obtained through freedom of information requests. In this way, Turner et al. (2008) were able to obtain the results of 74 clinical trials related to pharmacological treatments for depression.

Turner et al. (2008) wondered if the FDA decision about drug efficacy determined whether the study would be published. The FDA classifies the outcomes of clinical trials as positive, negative, or questionable. Figure 22.5 shows the association between the FDA decision about the drug and whether a report of the study was published in an academic journal. Of the 74 studies, 38 were judged positive by the FDA. Of these 38 positive outcomes, 37 were published and one was not.

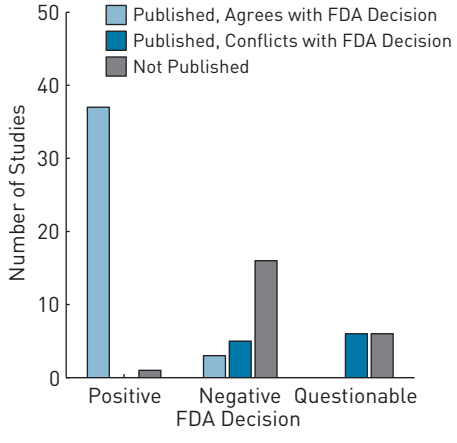
Of the 74 studies, 24 were judged negative by the FDA. Three of the negative studies were published and written in a way that agreed with the FDA decision and 16 were not published. However, 5 of these 24 studies were published but written in a way that conflicted with the FDA conclusion. That is, "the highlighted finding conflicted with the FDA-defined primary outcome" (Turner et al., 2008, p. 255).

The remaining 12 studies were judged questionable, meaning that the FDA did not consider them positive or negative. Questionable studies did not show statistical significance on the primary outcome measures but did show statistical significance on one or more of the secondary outcomes. Of these 12 questionable studies, six were not published and six were but were written in a way that conflicted with the FDA conclusion.

Figure 22.5 summarizes the fates of the 74 studies in the meta-analysis. Almost all of the studies with positive outcomes were published (37 of 38), but only three of the negative or questionable outcomes were published and written in a way the agreed with the FDA conclusion (3 of 24). Of 12 questionable studies, six were published but drew conclusions that conflicted with those of the FDA. Therefore, anybody reading the literature hoping to assess the effectiveness of these antidepressants (your doctor, for example) would conclude that the evidence overwhelmingly supported the conclusion that the drugs were effective. This reader would not be aware of the many negative outcomes that were not published or the conflicts with the FDA conclusions.

The meta-analysis conducted by Turner et al. (2008) showed that the mean effect size was 0.37 for the FDA studies that were eventually published and 0.15 for the unpublished FDA studies. This is a clear example of the file-drawer problem. In general, statistically significant results are published and those failing to attain statistical significance are not. An example of this problem is illustrated in Figure 22.4. In our hypothetical meta-analysis, the light blue dots represent studies that were published and the white dots represent studies that weren't. In our example, as in the Turner et al. study, the unpublished effect sizes were smaller than the published effect sizes.

FIGURE 22.5 ■ FDA Decisions and Publication



The relationship between FDA decisions and publication status for the 74 studies considered by Turner et al. (2008).

Turner et al. (2008) were not able to say why the negative outcomes were not published. It could be that the authors of the studies chose not to submit them for publication. Or it could be that journal editors rejected the papers because the results were not statistically significant. Whatever the reason, the paper by Turner et al. highlights the fact that a meta-analysis is only as good as the available data. If only positive outcomes and positive replications make it into the literature, then the literature will have little value. Turner et al. were very fortunate to have data on unpublished studies. Most meta-analysts are not so fortunate.

LEARNING CHECK 4

1. Convert the statistics in studies 6, 7, and 8 in Table 22.6 to *d*. That is,
   (a) Convert t_{obs} from row 6 to *d*.
   (b) Convert r^2 from row 7 to *d*.
   (c) Convert the 95% confidence interval from row 8 to *d*.
2. Table 22.6 shows five hypothetical studies that were published and three that were not.
   (a) Compute the 95% confidence interval around the mean effect size for the five published studies.
   (b) Compute the 95% confidence interval around the mean effect size for the three unpublished studies.

Answers

1. Converting statistics to *d*.

   (a) $t_{obs} = 1.93$, $n_1 = 71$, $n_2 = 74$.

   $$d = t_{obs}\sqrt{\frac{n_1 + n_2}{n_1 n_2}} = 1.93\sqrt{\frac{71+74}{71*74}}$$
   $$= 1.93 * 0.1661 = 0.32.$$

   (b) $r^2 = 0.0389$, $n_1 = 84$, $n_2 = 81$.

   $$d = \sqrt{df_{within}\left(\frac{r^2}{1-r^2}\right)\left(\frac{n_1+n_2}{n_1 n_2}\right)}$$
   $$= \sqrt{163\left(\frac{0.0389}{0.9611}\right)\left(\frac{84+81}{84*81}\right)}$$
   $$= \sqrt{163(0.0405)(0.0243)} = 0.40.$$

   (c) lower = 1, upper = 4.12, $n_1 = 165$, $n_2 = 175$

   $m_1 - m_2 = [4.12 - 1]/2 = 2.56$

   $moe = upper - (m_1 - m_2) = 4.12 - 2.56 = 1.56$

   $s_{m_1 - m_2} = moe/t_{\alpha/2} = 1.56/1.967 = 0.7931$

   $$s_{pooled} = s_{m_1 - m_2}\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 0.7931\sqrt{\frac{165*175}{165+175}} = 7.3089$$

   $$d = \frac{m_1 - m_2}{s_{pooled}} = \frac{2.56}{7.3089} = 0.35$$

2. Meta-analyses for published and unpublished studies.

   (a) Published

   $M = 0.366$, $s_M = 0.0167$, $df = 4$, $t_{\alpha/2} = 2.776$

   $M \pm t_{\alpha/2}(s_M) = [0.32, 0.41]$

   (b) Unpublished

   $M = 0.168$, $s_M = 0.0202$, $df = 2$, $t_{\alpha/2} = 4.303$

   $M \pm t_{\alpha/2}(s_M) = [0.08, 0.25]$

STEPS IN CONDUCTING A META-ANALYSIS

The Turner et al. (2008) study is an excellent meta-analysis. However, it differs in some ways from most meta-analyses in the literature. One of the biggest differences is that the studies selected for inclusion by Turner et al. were determined by the studies that registered for clinical trials with the FDA. This is a very clearly defined source of studies that provided a wealth of data. Most meta-analyses do not have such a clearly defined source for studies nor access to such extensive data. In the following paragraphs, we will review the elements of a meta-analysis and comment briefly on the challenges associated with each.

Formulating a Question

The first step in a meta-analysis is to formulate a clear question or questions to be answered. In the case of Turner et al. (2008), the questions were about the effectiveness of pharmacological treatments of depression and the relationship between FDA decisions and publication. The question about effectiveness is a well-formed question, and many studies have addressed it over the years. Similar questions could ask about the health risks associated with sugar, salt, or coffee, or the health benefits of exercise, reading, or social engagement. One could ask about the relative merits of phonics versus whole language approaches to reading instruction, the costs and benefits of working in open offices, the effects of television exposure on cognitive development, or the changes in fluid cognition with aging.

However, each of these questions might be too broad to be manageable. For example, does health risk mean cardiovascular health, muscle tone, mental health, or something else? Therefore, it is necessary to narrow the question to something more manageable, such as the association between salt consumption and cardiovascular health.

Collect Studies

Collecting studies to be included in a meta-analysis is a two-step process. First, potential studies for inclusion must be identified, and then these must be reduced to those meeting more stringent inclusion criteria.

Searching for Studies

Once a question has been formulated, the search for relevant research begins. The first part of this search is easier than ever. There are many science databases that can be searched to find articles addressing the question of interest. Two of the databases relevant to psychology are PubMed (pubmed.gov) and PsycINFO. Your university undoubtedly provides access to many more such databases, and you will be introduced to these as part of your training in psychology. These databases can be searched to find research articles with (i) specific keywords in the titles or abstracts, (ii) publication years, or (iii) authors' names.

As you know from using Google or other search engines, choosing appropriate search terms is critical to obtaining useful results. For example, I just did a Google search with the terms "salt" and "cardiovascular health," and Google returned 434,000 results. You can be sure that most of these links do not provide high-quality scientific results pertaining to this question. However, the same kind of problem exists with scientific databases. When I searched PubMed with the same terms, links to 2329 articles were returned. These articles are quite heterogeneous. For example, the first result returned had the title "Early-Stage Heart Failure With Preserved Ejection Fraction in the Pig: A Cardiovascular Magnetic Resonance Study." The sixth was titled "Land Use, Transport, and Population Health: Estimating the Health Benefits of Compact Cities." These two articles may or may not be relevant to the meta-analyst interested in the association between salt consumption and cardiovascular health.

As we saw with the Turner et al. (2008) study, the file-drawer problem can be a major limitation in a meta-analysis. There is always the strong possibility that the only studies in the literature are those that made it through the $p < .05$ filter. Therefore, a challenge to performing a meta-analysis is to find unpublished articles pertaining to the question that has been formulated. A landmark in the history of meta-analysis was a report by Smith and Glass (1977) that examined the effectiveness of psychotherapy. Smith and Glass searched through published bibliographies (i.e., precursors to the electronic versions of databases we use today) and the reference lists of the papers themselves.

By searching Dissertation Abstracts (a database of masters and doctoral theses), Smith and Glass (1977) were able to locate some relevant studies that were unpublished. However, many other unpublished studies have left no trace anywhere and are therefore unavailable to the meta-analyst. This is a huge vulnerability of the entire psychological literature, and it may take years to eliminate the damaging effects that the $p < .05$ filter has had.

Excluding Studies

Although locating articles relevant to the meta-analysis may have its challenges, the more difficult part is to examine each one to determine whether it should be part of the analysis. This means that the analyst must specify inclusion criteria. For example, a well-formulated question will have a clearly specified target population (e.g., North American adults) and a well-specified methodology (e.g., clinical trials with a control and treatment group).

The number of possible inclusion criteria is extremely large, but these must be specified before the analysis is conducted. If a meta-analyst has a hypothesis about the effect under study (e.g., a hypothesis that psychotherapy is effective or ineffective), then it would obviously be a huge (unconscious) temptation to include only those studies that are consistent with the hypothesis. This can be a tricky point, because an important inclusion criterion might be the quality of the study. For example, one might judge a study to be of poor quality if data collection were sloppy or if it were not run double blind, meaning that neither the subjects nor researchers knew the experimental hypothesis. It is certainly easy to imagine the temptation to judge a study as low quality when its results conflict with one's hypothesis. In the study by Turner et al. (2008), some of the inclusion criteria were as follows:

> From the FDA reviews of submitted clinical trials, we extracted efficacy data on all randomized, double-blind, placebo-controlled studies of drugs for the short-term treatment of depression. We included data pertaining only to dosages later approved as safe and effective; data pertaining to unapproved dosages were excluded. (p. 253)

As you can see, many choices made by the meta-analyst will determine which results are combined. Therefore, the meta-analyst must explain which databases were searched and which search terms were used, and must provide a list of inclusion criteria. The reader must be informed about what the inclusion criteria were, because he or she might disagree with these. Just as a paper in the primary literature has a method section that describes the study in sufficient detail to allow it to be replicated, the same is true for a meta-analysis. Without such information, the meta-analysis is of no value.

Coding Studies

Once studies have been selected, their results must be coded in two different ways. The first is to put the dependent variables onto a common scale (e.g., d or r), and the second is to code aspects of the study (e.g., participants, methods, materials) that might have a systematic effect on the study outcome. We discuss these in turn.

Coding the Dependent Variables

We've seen several times that standardized effect sizes are useful because they provide a common scale on which to express statistics arising from many different dependent variables. In Chapter 11, we saw how to convert t, F, and r^2 to d. In Chapter 15, we saw how to convert the regression slope (b) to r, and in this chapter we saw how to convert a confidence interval to d.

It is straightforward to transform many different measures to *d* or *r* when sufficient information has been presented to permit the transformation. Unfortunately, because of the dominance of significance testing, it has been common for researchers to report the result of a contrast as simply $p > .05$, which tells us only that the result was not statistically significant. In such situations, one may need to contact the authors of the original papers to try to retrieve this information.

Coding Moderating Variables

An important component of meta-analysis that has not been covered explicitly to this point involves the search for variables that affect the distribution of effect sizes in the literature. We saw an example of this in the Turner et al. (2008) study when they recorded (among other things) whether the study had been published or not. They found that the average effect size was larger in published studies than in unpublished studies. Variables that affect the distribution of effect sizes are called *moderator variables*.

Smith and Glass (1977), in their classic meta-analysis of the effectiveness of psychotherapy, recorded many variables that might function as moderators. For example, they recorded the type of therapy employed by the therapist, the number of years that the therapist had been practicing, the education level of the therapist, and many other variables. Some of these variables *could* have influenced the distribution of effect sizes. That is, one type of therapy might have proven more effective than others. If there had been large differences between the outcomes for three treatment types, then we would have said that treatment type moderates the difference in outcomes between control and treatment conditions. Therefore, the meaning of the term "moderation" in meta-analysis is essentially the same as its meaning in multiple regression.

As it turned out, treatment type did not act as a moderator in this case. The estimated effect sizes for psychodynamic, desensitization, and behavior modification treatments of phobias were 0.92, 1.05, and 1.12, respectively. The estimated effect sizes for the treatment of neuroticism were 0.64, 0.52, and 0.85, respectively. In fact, very few of the other variables that Smith and Glass (1977) examined seemed to be systematically related to treatment outcomes. They thus concluded the following:

> Despite volumes devoted to the theoretical differences among different schools of psychotherapy, the results of research demonstrate negligible differences in the effects produced by different therapy types. Unconditional judgments of superiority of one type or another of psychotherapy, and all that these claims imply about treatment and training policy, are unjustified. (Smith & Glass, 1977, p. 760)

It should be noted that subsequent meta-analyses suggest that cognitive behavioral therapy (CBT) is more effective than others in the treatment of anxiety disorders (Butler, Chapman, Forman, & Beck, 2006; Tolin, 2010).

The point here is not to draw conclusions about the effectiveness of psychotherapy. Rather, it is to note that the literature may (or may not) reveal that effect sizes depend on one or more moderator variables. Therefore, when coding studies to be included in a meta-analysis, one must also code the characteristics that distinguish the studies and which might explain the distribution of effect sizes in the literature.

Compute the Mean Effect Size and CI

We covered the calculation of mean effect sizes and confidence intervals in the preceding sections of this chapter. As noted, the method described is very similar to one proposed by

Hunter and Schmidt (1990), but the method of Hedges and Olkin (1985) is also widely used. Once again, this method is described in Appendix 22.3.

Interpret the Results

As with any research project, there are two kinds of questions that may be addressed. The first questions are those formulated at the outset of the analysis. For example, "How effective is psychotherapy?" or "How effective are antidepressants?" Such questions are answered by computing a confidence interval around the effect size of interest. As always, interpreting these results means being able to explain their importance in the context of the relevant research literature. No statistical rule can pronounce the results important or otherwise.

Other questions emerge in the course of analysis that were not formulated at the outset. For example, one might discover a moderating effect showing that one type of psychotherapy is more effective for adolescents and another more effective for adults. This is the kind of discovery one may make anytime measurements are made. This is exactly what happens in the primary literature itself. A research question is formulated and an experiment is designed and run. The analysis of the experiment was established at the outset, but once data have been collected, there may be unexpected patterns in them. One could treat these as noise (random variations), or one might see important clues to why they arose.

In a lecture in 1854, French chemist Louis Pasteur remarked that "in the fields of observation, chance favors only the prepared mind." By this, Pasteur meant that a serendipitous finding (a finding that one wasn't explicitly looking for) will be recognized as important only if you have the necessary background knowledge (preparation). Of course, such results require verification, through either meta-analyses or further primary research. It is the discovery of unexpected and meaningful patterns in data that drive new research questions and open up new lines of research.

SUMMARY

Meta-analysis provides a way to overcome the limitations of individual studies by combining evidence from many different studies. In this chapter, we saw how to compute a confidence interval around a mean of sample means. We compute the confidence interval as follows:

$$M \pm t_{\alpha/2}(s_M).$$

The most general way to compute M is as follows:

$$M = \sum_{i=1}^{k} w_i * m_i,$$

where

$$w_i = n_i / \sum_{i=1}^{k} n_i.$$

This method applies equally if sample sizes are the same or different. We compute S^2 as

$$S^2 = \sum_{i=1}^{k} w_i (m_i - M)^2 \frac{k}{k-1},$$

from which we obtain

$$s_M = \sqrt{\frac{S^2}{k}}.$$

The random-effects model assumes that M estimates μ_{Meta}, which is the mean of many population parameters; e.g., M may estimate the mean value of μ, the mean value of $\mu_1 - \mu_2$, the mean value of δ, or the mean value of ρ. The simplest version of the fixed-effects model assumes that M estimates a parameter (μ, $\mu_1 - \mu_2$, δ, or ρ) of a single population. The assumptions of the random-effects model are far more realistic than the assumptions of the fixed-effects model.

KEY TERMS

fixed-effects model 11
meta-analysis 1
meta-mean (M) 2
primary literature 1
random-effects model 11
secondary literature 1
weighted sum 6

EXERCISES

*Exercises marked with an asterisk rely on material covered in appendices.

Definitions and Concepts

1. What is the difference between primary and secondary literature?
2. What is meta-analysis?
3. What is a weighted sum?
4. What is a meta-mean?
5. What is the difference between the fixed-effects model and the random-effects model?

True or False

State whether the following statements are true or false.

6. M can be used to estimate μ.
7. M can be used to estimate μ_{Meta}.
8. M is a biased statistic.
9. If sample sizes differ, then $S^2 = \sum_{i=1}^{k}(m_i - M)^2 \frac{k}{k-1}$.
10. If sample sizes differ, then $M = \sum_{i=1}^{k}(m_i) * 1/k$.

Calculations

11. For the following numbers, compute the mean and variance in the usual way and as weighted sums. In each case, explain why the two means are the same or different.
    (a) $m = [10, 20, 50, 40, 30]$, $n = [50, 40, 20, 10, 30]$
    (b) $m = [10, 20, 50, 40, 30]$, $n = [5, 5, 5, 5, 5]$
    (c) $m = [16, 32, 64, 8, 128]$, $n = [2, 2, 4, 8, 8]$
    (d) $m = [16, 32, 64, 8, 128]$, $n = [128, 64, 32, 16, 8]$
12. For each part the preceding question, compute the 95% confidence interval around M, taking sample size into account.

Scenarios

13. A researcher at an ophthalmology clinic reported in a journal article that 5% of male patients visiting his clinic had some sort of color deficiency (what many people incorrectly call color blindness). A researcher at another ophthalmology clinic reported in a journal article that 8% of her male patients had some sort of color deficiency. Which of these papers contributed to the primary literature and which to the secondary literature?
14. A researcher examined a large number of research papers that tested the null hypothesis that the average digit span of university students was 7. She reported in a publication that 12 studies rejected the null hypothesis and 10 retained it. Is her publication part of the primary or secondary literature? Does her finding mean that the mean digit span of university students is 7?
15. If the fixed-effects model correctly describes the means submitted to a meta-analysis, then M will fall closer to μ, on average, than individual means. The 95% confidence interval around M will be narrower, on average, than 95% confidence intervals around individual means. Why are these things true? Why is the term "on average" so important in these statements?
16. The primary visual cortex in the occipital lobe contains mechanisms that are critical to visual perception. It is a very large brain region, although its size differs from person to person. The following table summarizes mean area (in square centimeters) from seven published studies. Please perform a meta-analysis on these data by computing a point estimate and the 95% confidence interval around the point estimate.

i	m_i	n_i
1	45.37	5
2	95.72	2
3	81.02	5
4	66.41	9
5	51.56	6
6	72.24	8
7	86.86	3

17. In the past 2 years, pharmaceutical companies have commissioned 16 studies of low-density lipoprotein (LDL) cholesterol levels in the population of North American adolescents. High levels of LDL are considered bad. The units used to measure LDL are mg/dL (milligrams per deciliter). Of these 16 studies, eight were published in medical journals.

    (a) The mean LDL levels in the eight published studies are shown below, along with the number of participants in each study. Calculate M and the 95% confidence interval around M. (I recommend doing these calculations in Excel.)

i	m_i	n_i
1	135.19	27
2	146.95	38
3	132.04	30
4	134.48	28
5	132.56	41
6	128.62	34
7	131.35	34
8	124.83	25

    (b) The following table shows the results of the eight studies that, for one reason or another, we not published. Calculate M and the 95% confidence interval around M.

i	m_i	n_i
9	112.64	53
10	117.93	51
11	116.40	33
12	123.38	38
13	119.28	45
14	109.35	30
15	117.98	46
16	122.28	38

    (c) Do you notice any differences in the results of your meta-analyses for the published and unpublished studies? Explain what these differences are and try to think of reasons that might explain them.

    (d) Calculate M and the 95% confidence interval around M for all 16 studies.

18. *If the fixed-effects model correctly describes the means submitted to a meta-analysis, what does S^2 estimate? Please describe another way that the same quantity can be estimated.

19. *If the random-effects model correctly describes the means submitted to a meta-analysis, what does S^2 estimate?

ANSWERS

Definitions and Concepts

1. The primary literature comprises papers reporting original data, whereas the secondary literature comprises papers that combine results from the primary literature.

2. Meta-analysis is a quantitative approach that combines several results from the primary literature.

3. A weighted sum is a sum of numbers that have been multiplied by a weight. The sum of all weights equals 1.

4. The meta-mean (M) is a statistic that is the mean of a number of sample means.

5. The fixed-effects model assumes that all sample means in a meta-analysis represent different random samples from the same distribution, whereas the random-effects model does not make this assumption.

True or False

6. True.

7. True.

8. False. M is an unbiased estimate of μ or μ_{Meta}.
9. True.
10. False. $M = \sum_{i=1}^{k}(m_i) * 1/k$ does not weight means according to sample size.

Calculations

11. (a) $m = [10, 20, 50, 40, 30]$, $n = [50, 40, 20, 10, 30]$.

Simple mean = 30, simple variance = 250, weighted mean = 24, weighted variance = 230. The weighted mean is smaller because sample size is inversely related to the magnitude of the mean. That is, smaller values (e.g., 10) are associated with larger sample sizes.

Calculations: Question 11 (a)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	10	50	500	9800
2	20	40	800	640
3	50	20	1000	13520
4	40	10	400	2560
5	30	30	900	1080
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		150	3600	27600
	Unweighted		Weighted	
M	30.0		24.00	
S^2	250.0		230.00	

(b) $m = [10, 20, 50, 40, 30]$, $n = [5, 5, 5, 5, 5]$.

Simple mean = 30, simple variance = 250, weighted mean = 30, weighted variance = 250. Because the sample sizes are the same, the weights applied in the case of the weighted sum are identical.

(c) $m = [16, 32, 64, 8, 128]$, $n = [2, 2, 4, 8, 8]$.

Simple mean = 49.6, simple variance = 2380.8, weighted mean = 60, weighted variance = 3340. The weighted mean is larger because sample size is directly related to the magnitude of the mean. The average of the two largest samples is 68 and the average of the two smallest samples is 24. Consequently, the weighted sum exceeds the simple average.

Calculations: Question 11 (b)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	10	5	50	2000
2	20	5	100	500
3	50	5	250	2000
4	40	5	200	500
5	30	5	150	0
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		25	750	5000
	Unweighted		Weighted	
M	30.0		30.00	
S^2	250.0		250.00	

Calculations: Question 11 (c)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	16	2	32	3872
2	32	2	64	1568
3	64	4	256	64
4	8	8	64	21632
5	128	8	1024	36992
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		24	1440	64128
	Unweighted		Weighted	
M	49.6		60.00	
S^2	2380.8		3340.00	

(d) $m = [16, 32, 64, 8, 128]$, $n = [128, 64, 32, 16, 8]$.

Simple mean = 49.6, simple variance = 2380.8, weighted mean = 29.42, weighted variance = 740.06. The weighted mean is smaller because sample size is inversely related to the magnitude of the mean. That is, smaller means (e.g., 16) are associated with larger sample sizes than are larger values (e.g., 128).

Calculations: Question 11 (d)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	16	128	2048	23050.12
2	32	64	2048	426.22
3	64	32	2048	38266.27
4	8	16	128	7340.62
5	128	8	1024	77745.15
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		248	7296	146828.39
	Unweighted		Weighted	
M	49.6		29.42	
S^2	2380.8		740.06	

12. For each part of the preceding question, compute the 95% confidence interval around M, taking sample size into account.
    (a) $M = 24$, $S^2 = 230$, $s_M = 6.78$, $t_{\alpha/2} = 2.776$, 95% CI = [5.17, 42.83].
    (b) $M = 30$, $S^2 = 250$, $s_M = 7.07$, $t_{\alpha/2} = 2.776$, 95% CI = [10.37, 49.63].
    (c) $M = 60$, $S^2 = 3340$, $s_M = 28.85$, $t_{\alpha/2} = 2.776$, 95% CI = [−11.75, 131.75].
    (d) $M = 29.42$, $S^2 = 740.06$, $s_M = 12.17$, $t_{\alpha/2} = 2.776$, 95% CI = [−4.35, 63.19].

Scenarios

13. Both papers contribute to the primary literature because both present original data.
14. It is part of the secondary literature because it combines the results of studies in the primary literature. Her finding does not mean that the mean digit span of university students is 7. The mean digit span of university students would be better estimated from a meta-analysis of the 22 means reported in the studies reviewed.
15. Ultimately, M is based on more scores than any individual m. Statistics based on larger numbers (e.g., M) have a smaller sampling error, meaning that, on average, they fall closer to the parameter (μ in this case) than statistics based on smaller samples (e.g., m). Similarly, the standard errors associated with large samples are smaller than the standard errors associated with small samples. Therefore, M falls closer to μ on average than m. For the same reason, confidence intervals will be narrower. The term "on average" is important here because it's possible for a given m to fall closer to μ than a given M.
16. The table below shows the calculation of M and S^2.

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	45.37	5	227	2471.68
2	95.72	2	191	1581.05
3	81.02	5	405	899.99
4	66.41	9	598	12.82
5	51.56	6	309	1544.40
6	72.24	8	578	171.96
7	86.86	3	261	1112.42
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		38	2569	7794.33
	Weighted			
M	67.60			
S^2	239.30			

To compute the confidence interval, we first compute s_M, which can be computed in one step as

$$s_M = \sqrt{\frac{S^2}{k}} = \sqrt{\frac{239.3}{7}} = 5.847.$$

Because there are 7 − 1 = 6 degrees of freedom, $t_{\alpha/2} = 2.447$. The confidence interval can now be computed as

$$\text{CI} = M \pm t_{\alpha/2}(s_M) = 67.6 \pm 2.447(5.847) = [53.29, 81.91].$$

17. (a) The table shows the calculation of M and S^2.

To compute the confidence interval, we first compute s_M, which can be computed in one step as

$$s_M = \sqrt{\frac{S^2}{k}} = \sqrt{\frac{44.10}{8}} = 2.348.$$

Calculations: Question 17 (a)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	135.19	27	3650	61.63
2	146.95	38	5584	6692.32
3	132.04	30	3961	80.61
4	134.48	28	3765	17.95
5	132.56	41	5435	51.36
6	128.62	34	4373	870.25
7	131.35	34	4466	184.46
8	124.83	25	3121	1957.72
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		257	34356	9916.30
	Weighted			
M	133.68			
S^2	44.10			

Because there are $8 - 1 = 7$ degrees of freedom, $t_{\alpha/2} = 2.365$. The confidence interval can now be computed as

$$\text{CI} = M \pm t_{\alpha/2}(s_M) = 133.68 \pm 2.365(2.348) = [128.13, 139.23].$$

(b) The table shows the calculation of M and S^2.

To compute the confidence interval, we first compute s_M, which can be computed in one step as

$$s_M = \sqrt{\frac{S^2}{k}} = \sqrt{\frac{19.26}{8}} = 1.552.$$

Because there are $8 - 1 = 7$ degrees of freedom, $t_{\alpha/2} = 2.365$. The confidence interval can now be computed as

$$\text{CI} = M \pm t_{\alpha/2}(s_M) = 117.47 \pm 2.365(1.552) = [113.80, 121.14].$$

Calculations: Question 17 (b)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
9	112.64	53	5970	1237.70
10	117.93	51	6014	10.68
11	116.40	33	3841	37.96
12	123.38	38	4688	1326.15
13	119.28	45	5368	147.02
14	109.35	30	3281	1979.24
15	117.98	46	5427	11.85
16	122.28	38	4647	878.26
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		334	39236	5628.86
	Weighted			
M	117.47			
S^2	19.26			

(c) Yes. The mean of the published studies ($M = 133.68$) is greater than the mean of the unpublished studies ($M = 117.47$). Perhaps these hypothetical pharmaceutical companies would like the evidence to show that the average LDL level is greater than it is.

(d) The table shows the calculation of M and S^2.

Calculations: Question 17 (d)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
1	135.19	27	3650	3073.87
2	146.95	38	5584	19117.84
3	132.04	30	3961	1696.47
4	134.48	28	3765	2777.60
5	132.56	41	5435	2650.25
6	128.62	34	4373	571.52
7	131.35	34	4466	1586.02
8	124.83	25	3121	2.40
9	112.64	53	5970	7480.23

(Continued)

Calculations: Question 17 (d) (Continued)

i	m_i	n_i	$n_i m_i$	$n_i(m_i - M)^2$
10	117.93	51	6014	2214.89
11	116.40	33	3841	2175.88
12	123.38	38	4688	49.39
13	119.28	45	5368	1235.63
14	109.35	30	3281	6903.94
15	117.98	46	5427	1967.54
16	122.28	38	4647	190.68
		$\sum_{i=1}^{k} n_i$	$\sum_{i=1}^{k} n_i m_i$	$\sum_{i=1}^{k} n_i(m_i - M)^2$
		591	73591	53694.17
	Weighted			
M	124.52			
S^2	96.91			

To compute the confidence interval, we first compute s_M, which can be computed in one step as

$$s_M = \sqrt{\frac{S^2}{k}} = \sqrt{\frac{96.91}{16}} = 2.461.$$

Because there are 16 – 1 = 15 degrees of freedom, $t_{\alpha/2} = 2.131$. The confidence interval can now be computed as

$$\text{CI} = M \pm t_{\alpha/2}(s_M) = 12.52 \pm 2.131(2.461) = [119.27, 129.77].$$

18. If the fixed-effects model correctly describes the situation, then the variance among the sample means is an estimate of the variance of the distribution of means (σ_m^2).

19. *If the random-effects model correctly describes the means submitted to a meta-analysis, what does S^2 estimate? If the random-effects model correctly describes the situation, then S^2 estimates $(\sigma_{\text{Meta}}^2 + \sigma_w^2)/n$, where σ_{Meta}^2 is the variance of the population means, σ_m^2 is the average of all population variances, and n is the size of the individual samples.

APPENDIX 22.1: THE PARAMETERS ESTIMATED BY S^2

In Chapter 22, we did not discuss what S^2 estimates because doing so would have broken the flow of our discussion. We can now address this question. The answer depends on whether the fixed- or random-effects model correctly describes the nature of the population(s) involved. To simplify this discussion, we will make two assumptions: (i) we will assume that in all cases, distributions of scores have the same variance (σ^2); and (ii) we will assume that all samples involve the same number of scores (n). With these assumptions, we can compute S^2 more simply as

$$S^2 = \frac{\sum_{i=1}^{k}(m_i - M)^2}{k-1} \qquad (22.A1.1)$$

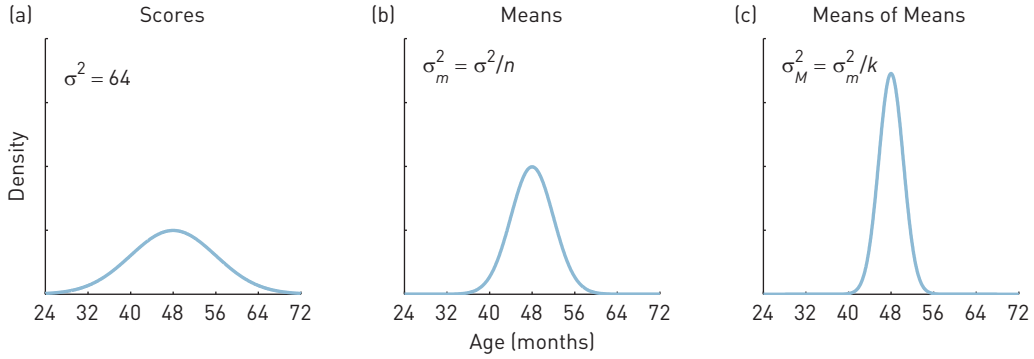
Let's first assume that the *fixed-effects model* correctly describes the population of scores under study. If this is so, then all samples have been drawn from exactly the same population of scores. In this case, S^2 estimates the variance of the distribution of means, σ_m^2. This might not seem obvious, so let's step through the logic.

Figure 22.A1.1a shows a distribution of scores that has a mean of $\mu = 48$ and variance of $\sigma^2 = 64$. Figure 22.A1.1b shows the sampling distribution of the mean for sample size n. (Remember, we're assuming that all samples are the same size.) The variance of this distribution of means is, as always, $\sigma_m^2 = \sigma^2/n$.

A meta-analysis assumes that k means have been drawn from the distribution of means. So, just as we can draw n scores from the left distribution and compute s^2 to estimate σ^2, we can draw k means from the center distribution and compute the variance among the means (S^2) to estimate σ_m^2. Therefore, the variance among the sample means is an estimate of the variance of the distribution of means. That is, when the fixed-effects model is correct, S^2 estimates σ_m^2.

Let's now assume that the *random-effects model* correctly describes the populations of scores under study. This situation is a little more complex but can be understood by referring back to Figure 22.2. Figure 22.2 shows that under the random-effects model, we imagine a very large number of populations of scores,

FIGURE 22.A1.1 ■ The Fixed Effects Model



The distributions underlying the fixed-effects model of meta-analysis. Scores (a), means (b), and means of means (c). See the text for further details.

each of which has a different mean, μ_i. The distribution of these population means has a mean (μ_{Meta}) and variance (σ^2_{Meta}). We continue to assume that each distribution of scores has the same variance, σ^2. In a meta-analysis, we have selected k of these populations at random and drawn a sample of size n from each one; from these k samples, we computed k sample means. (M is computed from these k sample means.)

If the random-effects model is correct, then there are two contributions to the variability among sample means. One source is the variance within distributions (σ^2_m) and the other is the variance between the distribution means (σ^2_{Meta}). As a consequence, the variance of the distribution of means would be

$$\sigma^2_m + \sigma^2_{\text{Meta}}, \tag{22.A1.2}$$

which is to say that S^2 now estimates $\sigma^2_m + \sigma^2_{\text{Meta}}$. (The variance of the distribution of M would be

$$\sigma^2_M = \frac{\sigma^2_m + \sigma^2_{\text{Meta}}}{k},$$

but that's not our current focus.)

Is it possible to determine whether the fixed or random-effects model correctly describes the means that have been combined? The answer is yes, and I will sketch out a simplified explanation. The important point is that if $\sigma^2_{\text{Meta}} = 0$, then the fixed-effects model is correct. In this case, there is only one source of variance contributing to the distribution of M. That is, if $\sigma^2_{\text{Meta}} = 0$, then equation 22.A1.2 reduces to σ^2_m. Therefore, when $\sigma^2_{\text{Meta}} = 0$, equation 22.A1.1 estimates σ^2_m.

Now, a second way to estimate σ^2_m involves the average sample variance, as follows:

$$\bar{s}^2 = \frac{\sum_{i=1}^{k} s_i^2}{k}. \tag{22.A1.3}$$

In equation 22.A1.3, the mean sample variance is called $\bar{s}^2$. The line over s indicates that this is the mean and distinguishes it from the sample variance used everywhere else in the book. This new quantity ($\bar{s}^2$) estimates σ^2, which is the variance common to all distributions of scores. Therefore, if the fixed-effects model is correct ($\sigma^2_{\text{Meta}} = 0$), then $\bar{s}^2/n$ estimates $\sigma^2/n = \sigma^2_m$.

We now have two ways to estimate σ^2_m. If $\sigma^2_{\text{Meta}} = 0$, we would expect the ratio of these two quantities to be close to 1, on average. That is,

$$\frac{S^2}{\bar{s}^2/n}$$

should be close to 1. However, if $\sigma^2_{\text{Meta}} > 0$, then this ratio should be greater than 1. One could use a significance test to determine whether the ratio is larger than would be expected by chance on the assumption that $\sigma^2_{\text{Meta}} = 0$.

It's not clear what value there is in testing whether $\sigma^2_{\text{Meta}} = 0$, per se. If we had a multimodal distribution of our effect size (e.g., d or r), then this might suggest the existence of a moderating variable. However, evidence that $\sigma^2_{\text{Meta}} > 0$ is not in itself evidence of a moderating variable.

APPENDIX 22.2: USING EXCEL TO CONDUCT A META-ANALYSIS

The data in Figure 22.A2.1 have been taken from Table 22.6. The effect sizes and sample sizes from eight studies are given in rows 1 to 8 of columns **A** and **B**, respectively. To compute M as a weighted sum, we first multiply effect sizes by sample sizes in rows 1 to 8 of column **C**. The sum of these products is divided by the sum of all sample sizes in cell **B13** to yield M.

To compute S^2 as a weighted sum, we first multiply sample sizes by the squared deviations of the effect sizes from M [i.e., $n_i(d_i - M)^2$] in rows 1 to 8 of column **E**. S^2 is the sum of these squared deviations divided by the total number of scores (N, shown in cell **B12**) and then corrected for bias by multiplying $k/(k-1)$. The result is shown in cell **B16**.

The estimated standard error of M is computed in cell **B17** as $s_M = \sqrt{S^2/k}$. The **T.INV** function is used in cell **B18** to compute $t_{\alpha/2}$, where α is .05. Finally, the lower and upper limits of the confidence interval are computed in cells **B20** and **B21**, respectively.

FIGURE 22.A2.1 ■ Meta-Analysis in Excel

	A	B	C	D	E	F
1	d	n	n*d	Formula	n*(d-M)^2	Formula
2	0.16	125	20	=B2*A2	2.8096	=B2*(A2-B13)^2
3	0.42	150	63	=B3*A3	1.8175	=B3*(A3-B13)^2
4	0.35	160	56	=B4*A4	0.2570	=B4*(A4-B13)^2
5	0.13	100	13	=B5*A5	3.2373	=B5*(A5-B13)^2
6	0.32	125	40	=B6*A6	0.0127	=B6*(A6-B13)^2
7	0.20	145	29	=B7*A7	1.7521	=B7*(A7-B13)^2
8	0.40	165	66	=B8*A8	1.3388	=B8*(A8-B13)^2
9	0.35	340	119	=B9*A9	0.5461	=B9*(A9-B13)^2
10						
11	Quantity	Values	Formulas			
12	N	1310	=SUM(B2:B9)			
13	M	0.310	=SUM(C2:C9)/B12			
14	k	8	=COUNT(B2:B9)			
15	df	7	=B14-1			
16	S^2	0.0103	=SUM(E2:E9)/B12*B14/B15			
17	s_M	0.0358	=SQRT(B16/B14)			
18	$t_{\alpha/2}$	2.3646	=T.INV(.975,B15)			
19						
20	lower	0.2252	=B13-B18*B17			
21	upper	0.3946	=B13+B18*B17			

These data are taken from Table 22.6. The effect sizes from eight studies are given in column **A** and the sample sizes are given in column **B**. The analysis is described in the text.

APPENDIX 22.3: AN ALTERNATIVE APPROACH TO META-ANALYSIS

In Chapter 22, we described a method of meta-analysis closely linked to that described by Hunter and Schmidt (1990). In this case, the mean and the variance of the means (or other statistics) are computed as weighted sums. The alternative method proposed by Hedges and Olkin (1985) also computes the mean of means as a weighted sum but includes the sample variance as part of the weight. We will first describe the Hedges and Olkin fixed-effects model for estimating μ, and then the random-effects model for estimating μ_{Meta}. We will then describe how the random-effects model is used to estimate $\mu_1 - \mu_2$, δ, and ρ. Finally, a few things will be said about the relative merits of the Hunter-Schmidt and Hedges-Olkin models.

Fixed-Effects Meta-Analysis for μ

In the Hedges-Olkin fixed-effects model, the meta-mean, M, is computed as a weighted sum, as before:

$$M = \frac{\sum_{i=1}^{k} w_i m_i}{\sum_{i=1}^{k} w_i}. \qquad (22.A3.1)$$

In contrast to what we used in Chapter 22, the weights are defined as follows:

$$w_i = \frac{1}{s_i^2/n_i}. \qquad (22.A3.2a)$$

Equation 22.A3.2a can also be written as

$$w_i = \frac{n_i}{s_i^2}. \qquad (22.A3.2b)$$

In equation 22.A3.2a, s_i^2 is the variance associated with m_i and n_i is the sample associated with m_i. The denominator of equation 22.A3.2a, s_i^2/n_i, is the squared

estimated standard error for the mean for m_i, i.e., $s^2_{m_i}$. Equation 22.A3.2b makes clear that w_i will *increase* as s_i^2 *decreases* and n_i *increases*. The logic here is to give greater weight to large samples and samples with smaller variance; samples with small variance are assumed to be more precise estimates of μ.

The estimated standard error of M is

$$s_M = \sqrt{\frac{1}{\sum_{i-1}^{k} w_i}}. \qquad (22.A3.3)$$

When the s_i^2 are small and the n_i are large, then the sum of w_i (i.e., $\sum_{i-1}^{k} w_i$) will be large, and its reciprocal (equation 22.A3.3) will be small. Conversely, when the s_i^2 are large and the n_i are small, the sum of w_i (i.e., $\sum_{i-1}^{k} w_i$) will be small and its reciprocal will be large. In general, there will be a mixture of large and small w_i. From M and s_M, a 95% confidence interval is computed as

$$\text{CI} = M \pm 1.96(s_M). \qquad (22.A3.4)$$

The data in Table 22.A3.1 show eight sample means and variances and their associated sample sizes. All samples were drawn from a normal population with μ = 48 and σ = 8. Because all samples were drawn from the same distribution, the fixed-effects model applies to these data. Table 22.A3.1 also shows the quantities required for the calculation of a confidence interval around M using equations 22.A3.1 to 22.A3.4. For example, the weight associated with m_1 is

$$w_1 = \frac{1}{s_1^2/n_1} = \frac{1}{40.3/20} = 0.496,$$

which when multiplied by the sample mean yields

$$w_1 m_1 = 0.496 * 48.6 = 24.119.$$

When the weights are calculated for each of the eight sample means, M is computed as follows:

$$M = \frac{\sum_{i=1}^{k} w_i m_i}{\sum_{i=1}^{k} w_i} = \frac{182.182}{3.818} = 47.72.$$

Using equation 22.A3.3, we compute s_M as follows:

$$s_M = \sqrt{\frac{1}{\sum_{i-1}^{k} w_i}} = \sqrt{\frac{1}{3.818}} = 0.51.$$

Using equation 22.A3.4, we find that the 95% confidence interval around M is

$$\text{CI} = M \pm 1.96(s_M) = 47.72 \pm 1.96(0.51)$$
$$= [46.72, 48.72].$$

We have 95% confidence that this interval contains the population mean, μ.

Random-Effects Meta-Analysis for μ

The random-effects model assumes that scores are drawn from populations with different means. Let's assume that these population means are normally distributed with $\mu_{\text{Meta}} = 48$ (as before) and $\sigma_{\text{Meta}} = 5$. For simplicity, we will further assume that each population of scores has the same standard deviation, σ = 8. Imagine choosing eight of these populations at random and drawing a random sample from each one. The number of scores in each sample is shown in column n_i of Table 22.A3.2. To the left of the samples sizes are the sample means (m_i) and variances (s_i^2).

Although it's not appropriate, we *could* use the data in Table 22.A3.2 to compute a confidence interval around M for a fixed-effects model just as we did in the preceding section. Doing so, we find that

$$M = \frac{\sum_{i=1}^{k} w_i m_i}{\sum_{i=1}^{k} w_i} = \frac{175.303}{3.583} = 49.93$$

TABLE 22.A3.1 ■ Fixed-Effects Meta-Analysis

i	m_i	s_i^2	n_i	w_i	$w_i m_i$
1	48.6	40.3	20	0.496	24.119
2	47.8	53.4	60	1.124	53.708
3	47.0	65.8	30	0.456	21.429
4	49.1	48.1	40	0.832	40.832
5	44.8	102.7	20	0.195	8.724
6	48.7	43.0	10	0.233	11.326
7	44.8	25.5	10	0.392	17.569
8	49.1	109.7	10	0.091	4.476
				$\sum_{i-1}^{k} w_i$	$\sum_{i=1}^{k} w_i m_i$
				3.818	182.182

and

$$s_M = \sqrt{\frac{1}{\sum_{i-1}^{k} w_i}} = \sqrt{\frac{1}{3.583}} = 0.53.$$

From this, we find the 95% confidence interval to be

$$\text{CI} = M \pm 1.96(s_M) = 49.93 \pm 1.96(0.53) = [48.89, 50.97].$$

Of course, this confidence interval is inappropriate because it assumes the fixed-effects model. As a consequence, s_M (equation 22.A3.3) is based on the sample variances only. This means that s_M does not include variability attributable to σ_{Meta}, as would be needed for the random-effects model.

Equations 22.A3.1 to 22.A3.4 can be modified to include an estimate of the variability attributable to the population means (σ^2_{Meta}), in addition to the variability within populations (σ). Equation 22.A3.5 shows that the weights in the random-effects model add the term s^2_{Meta} to the weights of the fixed-effects model as follows:

$$w_i^* = \frac{1}{s_i^2/n_i + s^2_{\text{Meta}}}. \qquad (22.A3.5)$$

In this case, s^2_{Meta} is an estimate of σ^2_{Meta} (Hedges, 1992). (Make sure you see how equation 22.A3.5 relates to equation 22.A3.2a.) These new weights are used to compute M^*,

$$M^* = \frac{\sum_{i=1}^{k} w_i^* m_i}{\sum_{i=1}^{k} w_i^*}, \qquad (22.A3.6)$$

and s_M^*,

$$s_M^* = \sqrt{\frac{1}{\sum_{i-1}^{k} w_i^*}}. \qquad (22.A3.7)$$

s_M^* is the estimated standard error of M^* that includes an estimate of σ^2_{Meta}. In equations 22.A3.5, 22.A3.6, and 22.A3.7, the asterisk (*) indicates the random effects model and is used to distinguish the weights, M, and standard error in the fixed-effects model from the corresponding quantities in the random-effects model.

The calculation of s^2_{Meta} is the new part here. It involves first computing the weighted sum of squared deviations of the sample means from M using the weights of the fixed-effects model as follows:

$$Q = \sum_{i=1}^{k} w_i (m_i - M)^2. \qquad (22.A3.8a)$$

TABLE 22.A3.2 ■ Random-Effects Meta-Analysis for Sample Means From Eight Studies

i	m_i	s_i^2	n_i	w_i	$w_i m_i$	$w_i m_i^2$	w_i^2	w_i^*	$w_i^* m_i^2$
1	50.0	38.0	20	0.526	26.316	1315.789	0.277	0.049	2.428
2	41.9	67.7	60	0.886	37.134	1555.932	0.785	0.050	2.113
3	50.8	48.9	30	0.613	31.166	1583.215	0.376	0.049	2.499
4	52.0	60.9	40	0.657	34.154	1776.026	0.431	0.049	2.572
5	49.5	63.5	20	0.315	15.591	771.732	0.099	0.046	2.263
6	52.7	35.7	10	0.280	14.762	777.952	0.078	0.045	2.367
7	55.1	78.2	10	0.128	7.046	388.237	0.016	0.038	2.078
8	51.7	56.6	10	0.177	9.134	472.242	0.031	0.041	2.123
				$\sum_{i=1}^{k} w_i$	$\sum_{i=1}^{k} w_i m_i$	$\sum_{i=1}^{k} w_i m_i^2$	$\sum_{i=1}^{k} w_i^2$	$\sum_{i=1}^{k} w_i^*$	$\sum_{i=1}^{k} w_i^* m_i$
				3.583	175.303	8641.126	2.095	0.367	18.442
				M	Q	c	s^2_{Meta}	s_M^*	M^*
				48.933	63.046	2.998	18.697	1.651	50.246

A computational shortcut is to compute

$$Q = \sum_{i=1}^{k} w_i m_i^2 - \frac{\left(\sum_{i=1}^{k} w_i m_i\right)^2}{\sum_{i=1}^{k} w_i}. \qquad (22.A3.8b)$$

To compute s^2_{Meta}, we use the following formula:

$$s^2_{\text{Meta}} = \frac{Q - df}{c} \qquad (22.A3.9)$$

where

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i}, \qquad (22.A3.10)$$

and $df = k-1$.

When σ^2_{Meta} is 0 or very small, sampling error makes it possible for equation 22.A3.9 to yield a negative value for s^2_{Meta}. Of course, this makes no sense because a variance cannot be negative. Therefore, if equation 22.A3.9 produces a negative s^2_{Meta}, we simply set $s^2_{\text{Meta}} = 0$. (Note: Q and c are arbitrary symbols typically used in descriptions of the Hedges-Olkin model.)

Although there a several steps required to compute s^2_{Meta}, equations 22.A3.8b and 22.A3.10 show that most of the quantities are relatively easy to compute from information shown in Table 22.A3.2. First,

$$\begin{aligned} Q &= \sum_{i=1}^{k} w_i m_i^2 - \frac{\left(\sum_{i=1}^{k} w_i m_i\right)^2}{\sum_{i=1}^{k} w_i} \\ &= 8641.126 - \frac{175.303^2}{3.583} \\ &= 63.046. \end{aligned}$$

Second,

$$c = \sum_{i=1}^{k} w_i - \frac{\sum_{i=1}^{k} w_i^2}{\sum_{i=1}^{k} w_i} = 3.583 - \frac{2.095}{3.583} = 2.998.$$

Finally,

$$s^2_{\text{Meta}} = \frac{Q - df}{c} = \frac{63.046 - 7}{2.998} = 18.697.$$

Once s^2_{Meta} has been computed, the random-effects weights w_i^* are computed. These are shown in the second to last column of Table 22.A3.2, and the weighted means ($w_i^* m_i$) are shown in the last column. From these weighted means, we find that

$$M^* = \frac{\sum_{i=1}^{k} w_i^* m_i}{\sum_{i=1}^{k} w_i^*} = \frac{18.442}{0.367} = 50.246,$$

and

$$s^*_M = \sqrt{\frac{1}{\sum_{i-1}^{k} w_i^*}} = \sqrt{\frac{1}{0.367}} = 1.651.$$

Using M^* and s^*_M, we can compute

$$\begin{aligned} \text{CI} &= M^* \pm 1.96(s^*_M) = 50.246 \pm 1.96(1.651) \\ &= [47.01,\ 53.48]. \end{aligned}$$

We saw earlier that applying the fixed-effects calculations to the data in Table 22.A3.2 produced a 95% confidence interval of $49.93 \pm 1.96(0.53)$, whereas we've just seen that the random-effects model produces a 95% confidence interval of $50.25 \pm 1.96(1.65)$. Therefore, $M^* > M$ and $s^*_M > s_M$. The fact that $s^*_M > s_M$ makes sense because s^*_M was computed using equation 22.A3.7 and includes an estimate of σ^2_{Meta}. Clearly the fixed-effects model in this case produces a confidence interval that is narrower than it should be.

The fact that $M^* > M$ is not immediately obvious, but looking at columns w_i and w_i^* provides an explanation. Column w_i shows weights $[w_i = 1/(s_i^2/n_i)]$ that are highly variable because both sample sizes and sample variances are highly variable. By contrast, column w_i^* shows weights $[w_i^* = 1/(s_i^2/n_i + s^2_{\text{Meta}})]$ that are more homogeneous. The reason is that s^2_{Meta} is greater than s_i^2/n_i, making w_i^* more dependent on s^2_{Meta} than s_i^2/n_i. Because of the homogeneity of w_i^*, $M^* = 50.246$ is closer to the simple mean of means (50.46) than $M = 49.93$.

Fixed- and Random-Effects Meta-Analysis for $\mu_1 - \mu_2$, δ, and ρ

The Hedges (1992) approach to meta-analysis can be used to perform meta-analyses to estimate many other parameters including $\mu_1 - \mu_2$, δ, and ρ. The only difference in these cases is the quantity used to compute the weights. In the case of a simple mean described above, w_i was defined as $w_i = 1/(s_i^2/n_i)$. As noted earlier, s_i^2/n_i is the square of the estimated standard error of m_i. Therefore, the square of the estimated standard error is used to compute w_i for all other statistics

Weights for Estimating $\mu_1 - \mu_2$

When we perform a fixed-effects meta-analysis for the difference between two population means, we assume the existence of many pairs of population

means ($\mu_1 - \mu_2$), and for each there is a statistic is $m_1 - m_2$ to estimate this difference. Each $m_1 - m_2$ is associated with an estimate of the variance of its sampling distribution; i.e.,

$$s^2_{m_1 - m_2} = \frac{s^2_{\text{pooled}_i}}{n_{1_i}} + \frac{s^2_{\text{pooled}_i}}{n_{2_i}}.$$

For the fixed-effects meta-analysis, the weights are

$$w_i = \frac{1}{\frac{s^2_{\text{pooled}_i}}{n_{1_i}} + \frac{s^2_{\text{pooled}_i}}{n_{2_i}}}. \qquad (22.A3.11)$$

The random-effects weights are

$$w^*_i = \frac{1}{\frac{s^2_{\text{pooled}_i}}{n_{1_i}} + \frac{s^2_{\text{pooled}_i}}{n_{2_i}} + s^2_{\text{Meta}}}, \qquad (22.A3.12)$$

where s^2_{Meta} estimates the variance among the population $\mu_1 - \mu_2$. s^2_{Meta} is computed as in equation 22.A3.9. The estimated standard error (s_M) is computed as in equation 22.A3.7.

Weights for Estimating δ

When we perform a meta-analysis for the *standardized difference* between two population means (δ), the statistic d is computed for many pairs of means. Each d_i is associated with an estimate of the variance of its sampling distribution; i.e.,

$$s^2_{d_i} = \frac{d^2_i}{2df_i} + \frac{1}{n_{1_i}} + \frac{1}{n_{2_i}}.$$

For the fixed-effects meta-analysis, the weights are

$$w_i = \frac{1}{\frac{d^2_i}{2df_i} + \frac{1}{n_{1_i}} + \frac{1}{n_{2_i}}}. \qquad (22.A3.13)$$

For the random-effects meta-analysis, the weights are

$$w^*_i = \frac{1}{\frac{d^2_i}{2df_i} + \frac{1}{n_{1_i}} + \frac{1}{n_{2_i}} + s^2_{\text{Meta}}}, \qquad (22.A3.14)$$

where s^2_{Meta} estimates the variance among the population δ_i. s^2_{Meta} is computed as in equation 22.A3.9. The estimated standard error (s_M) is computed as in equation 22.A3.7.

Weights for Estimating ρ

When we perform a meta-analysis for the population correlation coefficient (ρ), the Fisher transform is used to first convert sample correlation coefficients to z_{r_i}; see Chapter 15. For each z_{r_i}, the associated variance is

$$\sigma^2_{z_{r_i}} = \frac{1}{n-3}. \qquad (22.A3.15)$$

For the fixed-effects meta-analysis, the weights are

$$w_i = \frac{1}{\frac{1}{n-3}} = n-3. \qquad (22.A3.16)$$

For the random-effects meta-analysis, the weights are

$$w^*_i = \frac{1}{\frac{1}{n-3} + s^2_{\text{Meta}}} \qquad (22.A3.17)$$

where s^2_{Meta} estimates the variance among the population ρ_i. s^2_{Meta} is computed as in equation 22.A3.9. The estimated standard error (s^*_M) is computed as in equation 22.A3.7. Once the meta-analysis is performed on the z_{r_i} values, the point estimate and confidence limits are transformed back to r values using the inverse Fisher transform.

Which Method to Use?

Many researchers have wondered about which method of meta-analysis is more accurate. This question arises because neither model is perfect in the sense that neither guarantees that nominal 95% confidence intervals will capture the parameter of interest exactly 95% of the time.

When we compute $M^* \pm 1.96(s^*_M)$, for example, we say that our *nominal confidence* level is 95%, which means that 95% of all intervals computed this way *should* capture the population parameter under ideal conditions. However, the percentage of intervals that *actually* capture the parameter of interest is called the *empirical coverage*. Because of the complexity of meta-analysis, the empirical coverage is not guaranteed to match the nominal confidence.

To illustrate this distinction, I can use a computer to simulate the circumstances that gave rise to the data in Table 22.A3.2. That is, I can draw eight samples of size 20, 60, 30, 40, 20, 10, 10, and 10 from populations, each of which has a standard deviation of $\sigma = 10$. The means of these eight populations are randomly drawn from a normal distribution with $\mu_{Meta} = 48$ and $\sigma_{Meta} = 5$. Each selection of eight samples is a sampling experiment. For each sampling experiment, I can compute a nominal 95% confidence interval around *M* (computed with the Hunter-Schmidt method) or *M** (computed with the Hedges-Olkin method).

The question is, what proportion of times will these two confidence intervals capture $\mu_{Meta} = 48$? To answer this question, I repeated the above sampling experiment 1,000,000 times and for each sample I computed *M* and *M**. I found that the Hunter-Schmidt method captures μ_{Meta} 90.9% of the time and the Hedges-Olkin method captures μ_{Meta} 92.3% of the time. Therefore, the *empirical coverage* in these two cases is 90.9 and 92.3. In neither case does the empirical coverage equal the nominal confidence of 95%.

These results might suggest that the Hedges-Olkin method does better than the Hunter-Schmidt method. However, there are many possible variations on the situation we've been considering. Rather than eight samples of size 20, 60, 30, 40, 20, 10, 10, and 10, we could consider 32 samples with sizes randomly drawn from a normal distribution of sizes with mean 30 and standard deviation 12. In this case, I find that the Hunter-Schmidt method captures μ_{Meta} 93.7% of the time and the Hedges-Olkin method captures μ_{Meta} 92.5% of the time. When there are 32 samples with sizes randomly drawn from a normal distribution of sample sizes with mean 100 and standard deviation 12, the empirical coverage for the Hunter-Schmidt method is 95% and the empirical coverage for the Hedges-Olkin method is 92.2%.

You should be able to see the enormous range of possible variations that must be considered when comparing the two methods. For example, one has to consider (i) the number of samples, (ii) the sample sizes, (iii) the values of σ and μ_{Meta}, (iv) whether all σ are equal, and (v) the parameter being estimated ($\mu_1 - \mu_2$, δ, and ρ, just to name a few). Comparing these methods is an ongoing area of research, and not all studies agree with each other (see, for example, Field, 2005, versus Hafdahl & Williams, 2009). However, my sense is that the Hunter-Schmidt method described in Chapter 22 generally works at least as well as the Hedges-Olkin model. For example, Table 6 in Hafdahl and Williams (2009) shows that in most cases the Hunter-Schmidt method produces empirical coverage that is at least as good as or better than the Hedges-Olkin model when the parameter being estimated is ρ.

References

Butler, A. C., Chapman, J. E., Forman, E. M., & Beck, A. T. (2006). The empirical status of cognitive-behavioral therapy: A review of meta-analyses. *Clinical Psychology Review*, *26*(1), 17–31. doi:10.1016/j.cpr.2005.07.003

Field, A. P. (2005). Is the meta-analysis of correlation coefficients accurate when population correlations vary? *Psychological Methods*, *10*(4), 444–467. doi:10.1037/1082-989X.10.4.444

Hafdahl, A. R. (2009). Improved Fisher *z* estimators for univariate random-effects meta-analysis of correlations. *British Journal of Mathematical and Statistical Psychology*, *62*(Pt 2), 233–261. doi:10.1348/000711008X281633

Hafdahl, A. R., & Williams, M. A. (2009). Meta-analysis of correlations revisited: Attempted replication and extension of Field's (2001) simulation studies. *Psychological Methods*, *14*(1), 24–42. doi:10.1037/a0014697

Hedges, L. V. (1992). Meta-analysis. *Journal of Educational Statistics*, 17(4), 279–296.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London, England: Academic Press.

Hunter, J. E., & Schmidt, F. (1990). *Methods of meta-analysis*. Newbury Park, CA: SAGE.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, CA: SAGE.

Smith, M. L., & Glass, G. V. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. doi:10.1037//0003-066X.32.9.752

Tolin, D. F. (2010). Is cognitive–behavioral therapy more effective than other therapies? A meta-analytic review. *Clinical Psychology Review*, *30*(6), 710–720. doi:10.1016/j.cpr.2010.05.003

Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, *358*(3), 252–260. doi:10.1056/NEJMsa065779