Power Statistics for Meta-analysis: Tests for Mean Effects and Homogeneity

Marc J. Lajeunesse

A common justification for meta-analysis is the increased statistical power to detect effects over what is obtained from individual studies (Miller and Pollock 1994, Arnqvist and Wooster 1995a). A classic example in the medical sciences illustrates this advantage. Multiple independent studies were conducted to determine the effect of a blood-clot medication on the likelihood of surviving a heart attack; however, only 6 of 33 studies detected a statistically significant effect of this medication on patients. Pooling these studies with meta-analysis, however, Lau et al. (1992) found a significant and important overall effect: patients treated with this medication had a 20% reduction in the odds of dying. This is an impressive example of the value of meta-analysis given that 81% of the studies were unable to detect an effect. Individually, they lacked statistical power.

The statistical power of meta-analysis is also important for ecologists and evolutionary biologists because effect sizes are usually relatively small in these fields (Møller and Jennions 2002), and experimental sample sizes are often limited for logistic reasons (Arnqvist and Wooster 1995b). Consequently, many studies lack sufficient power to detect an experimental effect should it exist (Jennions and Møller 2003). Given that many studies lack power individually, how does a meta-analysis of these studies achieve greater statistical power? It is often assumed that the statistical power of meta-analysis is mainly a function of the number of studies included in the review; including more studies in meta-analysis results in a greater likelihood of detecting an effect should it exist (Hedges and Olkin 1985). However, recent advances in calculating the power of pooled effect sizes and homogeneity tests (Hedges and Pigott 2001, 2004), and subsequent surveys and simulations applying these calculations (e.g., Cohn and Becker 2003, Sutton et al. 2007), indicate a much more nuanced complexity to the statistical power afforded by meta-analysis.

Here I provide a brief overview of the factors that determine the statistical power of meta-analysis, and present statistics for calculating the power of pooled effect sizes to evaluate nonzero effects, and the power of within- and between-study homogeneity tests. With these statistics, I emphasize a “soft” retrospective philosophy where predetermined hypotheses about effect sizes and magnitudes of heterogeneity are used to estimate power. Finally, I survey ways to improve the statistical power of meta-analysis, and end with a discussion on the overall utility of power statistics for meta-analysis.
STATISTICAL POWER AND SAMPLING ERROR IN EXPERIMENTS AND META-ANALYSIS

The power ($\rho$) of a statistical test is the probability of finding a significant result when it exists (Cohen 1988, 1992). It is explicitly defined as $\rho = 1 - \beta$; the complement to the probability of failing to detect this existing result ($\beta$ or the type II error). Power is tied directly to the standard error of the data ($SE$), the degree to which the biological phenomenon measured through experimentation exists (known as the effect size or $\delta$), and the significance criterion of the statistical test (the alpha level or $\alpha$). Statistical power varies directly as a function of $SE$, $\delta$, and $\alpha$, and a way of illustrating how these parameters interact is:

$$\rho \propto \frac{\delta \cdot \alpha}{SE}. \quad (22.1)$$

Here, the statistical power of an experiment will be high when (1) the effect size is large, (2) the data are not variable and have a small standard error, and (3) the stringency of the statistical test is lenient (i.e., $\alpha > .05$). Given these interactions, if $\delta$, $SE$, and $\alpha$ are known, then statistical power can be calculated (for a more extensive discussion, see Cohen 1988).

The parameter of interest here for meta-analysis is the standard error. The standard error estimates the variability in sampling error, and is determined by the sample size of a study ($n$) and the variance of the data ($\sigma^2$):

$$SE = \sqrt{\frac{\sigma^2}{n}}. \quad (22.2)$$

When $n$ becomes large, the variability in sampling error becomes very small. A simpler interpretation of this relationship is that larger studies are more “precise” and have greater statistical power than smaller studies (Hedges and Olkin 1985). The simulation presented in Figure 22.1 illustrates the increased power of significance tests in larger studies. This is due to studies with a large $n$ sampling a greater fraction of the population; with fewer data, random sampling can yield a fraction of the data that under- or overestimates the true population (predicted) effect. In effect, having a large sample size reduces the error associated with random sampling. This results in a more sensitive hypothesis test with greater statistical power.

This within-study sensitivity of statistical tests to sampling error is the primary justification for using meta-analysis over vote-count methods (Chapter 1). Vote counting relies heavily on average counts of significant and nonsignificant studies to provide summaries of research. This overlooks statistical issues within studies that help distinguish between true “biological” null results and erroneous null results due to low statistical power. For example, should a vote-count review pool the significance tests found in Figure 22.1, it would have counted 370 of 1000 simulated studies as null—even though a good proportion of these statistical tests were “null” because of small sample sizes. Vote counting will thus yield a biased synthesis because a greater number of studies will be treated as supporting the null hypothesis, resulting in a summary that underestimates the overall experimental effect across studies should it exist (Hedges and Olkin 1980). It is also important to note that increasing the number of studies pooled by vote counting (or simply having a large review sample size) does not improve statistical power, but in fact further increases the probability of making a review-level type II error (e.g., concluding a null outcome when one actually exists; see Hedges and Olkin 1985, Hunter and Schmidt 1990).
Although this seems counterintuitive to sampling theory, this loss of power when pooling many studies is due to the likelihood of vote-count reviews including many more “null” studies that were incorrectly counted as null because of low within-study power.

Meta-analysis has greater statistical power than vote-count methods because it attempts to provide a synthesis free from the sampling error within studies. Meta-analysis achieves this in two ways. First, it quantifies each study using an effect size—an estimate of the magnitude and
direction of a study outcome (e.g., Hedges’ \(d\)). Having effect sizes (as opposed to vote counts) as the unit of review has the advantage that the significance tests of studies are not used to quantify results. This avoids the overrepresentation of “null” studies (incorrectly specified as “null” due to low statistical power) as in vote-count approaches. Second, meta-analysis takes advantage of the predictable differences in sampling error that occur among studies with dissimilar sample sizes; it does this by applying a weighting scheme that downweights studies with large sampling error. For example, Figure 22.1 illustrates a simulation of 1000 studies measuring the same “predicted” effect in a given population (here \(d = 0.5\)), but each study differs slightly from this “expected” effect because of sampling error. Sampling theory predicts that these slight differences in sampling error are normally distributed with a mean of zero (Hedges and Olkin 1985). The outcome of this distribution is the characteristic funnel shape of effect sizes—where smaller studies show greater variation than larger studies (Fig. 22.1). Here meta-analysis compensates for the low statistical power of individual studies because it downweights these studies with a large sampling error (using the inverse variance of each effect size as the weight; see Chapters 8 and 9). This weighting scheme is one way in which meta-analysis can achieve greater statistical “power,” because it pools the actual experimental outcomes of numerous independent studies relative to their sampling error; it is therefore less sensitive to false negative or false positive outcomes due to low statistical power within each individual study.

However, similar to primary research (e.g., experiments, correlative studies), meta-analyses are also subject to sampling error. Here, each study (quantified as an effect size) is treated as a separate data point, and a meta-analysis with too few data points can have a biased synthesis of these data. This review-level error is referred to as second-order sampling error (Hunter and Schmidt 1990), and influences the ability of meta-analysis to detect effects when review-level sample sizes are low. Figure 22.2A shows how the number of studies included in a meta-analysis \(K\) relates to this review-level sensitivity to sampling error. A more formal way of illustrating this relationship and the statistical power of meta-analysis to detect an overall (pooled) effect \((\hat{d})\) that deviates from zero \((d_0)\) is:

\[
\rho \propto \frac{(\hat{d} - d_0) \cdot \alpha}{SE(\hat{d})},
\]

where \(\alpha\) is stringency (alpha level) of the nonzero test, and \(SE\) is the standard error of the pooled effect size \((\hat{d})\). The meta-analysis \(SE\) is defined as:

\[
SE(\hat{d}) = \sqrt{\sigma^2(\hat{d})} = \sqrt{\frac{1}{\sum_{i=1}^{K} \left[\sigma^2(\hat{d}) + \tau^2\right]^{-1}}},
\]

where \(\sigma^2(\hat{d})\) is the sample variance of \(\hat{d}\), and \(\tau^2\) is the between-study variance. What should be apparent from this review-level \(SE\) is that the number of studies included in a meta-analysis \((K)\) does not directly influence the statistical power of meta-analysis. For comparison, see the contribution of the sample size \(n\) of an experiment in Equation 22.2. Instead, \(K\) affects \(SE\) indirectly because more studies included in the meta-analysis results in a greater sum of the weights used to penalize each study when pooling their effect sizes (e.g., the sum of all the inverse within-study variances). Thus, with every additional study included in a meta-analysis, another inverse variance (or weight) is included in the overall variance estimation. Figure 22.2B shows how the inclusion of additional studies in a review significantly decreases \(\sigma^2(\hat{d})\) and consequently the 95% CI (confidence interval) of the pooled effect size also decreases. Thus, the addition of numerous studies to a meta-analysis results in a more sensitive review-level hypothesis test (greater statistical power to detect an effect) because the 95% CI used to evaluate nonzero effects (e.g., \(\delta - \delta_0\)) becomes narrow.
Figure 22.2. A simulation on how within-study sampling error influences the between-study sampling error of 1000 meta-analyses. Each data point is a simulated meta-analysis pooling $K$ number of effect sizes ($\bar{\mu}$) as generated in Figure 22.1 under a fixed-effects model (Chapter 9), where A depicts the second-order sampling error of these meta-analyses, and B indicates the width of estimated 95% confidence intervals of each pooled effect size. The only parameter modified in this simulation was the sample size ($K$) of each meta-analysis. What is apparent from this simulation is that meta-analyses based on few studies are more likely to have pooled effect sizes that deviate from the expected effect (here 0.5) and have difficulty detecting this nonzero effect (e.g., broad 95% CI).
Another important parameter affecting the SE of meta-analysis is the between-study variance, $\tau^2$. The between-study variance estimates the amount of heterogeneity observed among effect sizes, and is applied in meta-analysis when a random-effects model is assumed (see Chapters 8 and 9). In a fixed-effects model, $\tau^2 = 0$ because it is assumed that sampling error is the only source of variation among studies (Chapter 8). The random-effects model is important for meta-analysis in ecology and evolutionary biology, because it assumes that there are multiple population effects contributing to variation among effect sizes. For example, visualize three overlaid funnel distributions (like the one found in Figure 22.1), but with each centered about different population effects (e.g., 0.2, 0.5, and 0.8). It has been argued that the random-effects model is the most biologically relevant model for ecological meta-analysis (Gurevitch and Hedges 1999).

However, assuming a random-effects model (where $\tau^2 \neq 0$) will yield a meta-analysis with less statistical power relative to the fixed-effects model (Cohn and Becker 2003). This decrease in power is due to each individual study carrying less weight in the overall estimation of $SE(\hat{d})$ (see Equation 22.4). This change in the weighting scheme will increase the width of the 95% CI because the weight of each study (e.g., the inverse variance) becomes smaller with the addition of $\tau^2$. For example, the weight of a fixed-effects model will always be larger than a random-effects model when $\tau^2 \neq 0$:

$$\frac{1}{\sigma^2(\hat{d}_k)} > \frac{1}{\sigma^2(\hat{d}_k) + \tau^2},$$

and taking the inverse of the sum of these random-effects weights across studies will yield a larger SE than the sum based on a fixed-effects model (see Equation 22.4). However, the magnitude of this decrease in power depends on the size of the between-study variance. Thus, for a random-effects model, the potential statistical power gained by including numerous studies in a meta-analysis might be offset by a large amount of heterogeneity ($\tau^2$) among these pooled effect sizes (Cohn and Becker 2003, Sutton et al. 2007).

This difference in statistical power between fixed- and random-effects models should not, however, be used as justification for assuming a priori only a fixed-effects analysis, or used as sole justification for considering one effect-size model over another. The efficiency of the weighting scheme used by a fixed-effects model is dependent on whether the data represent a homogenous collection of effect sizes (e.g., whether they share a common effect, as depicted in Fig. 22.1), and any violation of this assumption will result in a loss of statistical power (Chapter 8). This is one reason why Gurevitch and Hedges (1999) have argued that the random-effects model may be more appropriate for ecological meta-analysis because it is unclear how ecological data meet this assumption of homogeneity—the conservatism of the random-effects model is meant to balance this uncertainty (Chapter 8).

Explicitly testing the assumption of homogeneity for a fixed-effects model using a within-group $Q$-test (Chapters 8 and 9) has been a justifiable way to evaluate its appropriateness for meta-analysis (Hedges and Olkin 1985). This homogeneity test evaluates whether effect sizes vary beyond the predicted sampling error, as seen in Figure 22.1, and whether moderator variables or covariates should be explored to explain this variation. Not detecting significant homogeneity, however, is not evidence that it does not exist entirely—homogeneity tests themselves will have low statistical power to detect heterogeneity when review-level sample sizes are small (Hunter and Schmidt 1990). A loss of power is also an issue for the between-group $Q$-test when few studies are further parsed into subgroups to test hypotheses (Hedges and Olkin 1985; Hunter and Schmidt 1990). Alternatively, although uncommon in ecological meta-analyses, if a random-effects model must be assumed a priori, then a statistical test can be performed to evaluate whether $\tau^2 \neq 0$ (Hedges and Olkin 1985).
The issue as to how the between-study variance influences the statistical power of meta-analysis can be further exacerbated by how the between-study variance is itself estimated. The most common approach in ecological meta-analyses to date, and the one advocated by Hedges and Olkin (1985) and adopted by MetaWin (Rosenberg et al. 2000), is the DerSimonian-Laird estimator of $\tau^2$ (see Chapters 8 and 9). However, now there is evidence that other estimators, such as the restricted maximum-likelihood (REML) based alternative, may provide more efficient and unbiased estimates of $\tau^2$ under elaborate statistical modeling (e.g., meta-regression models; Viechtbauer 2006). Again, given the strong influence of $\tau^2$ for determining the power of meta-analysis, an accurate and reliable estimate of $\tau^2$ is important to minimize the potential for type I and type II errors.

In summary, the statistical power of meta-analysis increases when there are large nonzero population effects, when the statistical significance levels of the hypothesis test are lenient (i.e., $\alpha > 0.05$), and when there are many studies included in the review. However, statistical power decreases if there is large sampling error within each study and/or if the between-study variance among these studies is large.

**INTERPRETATION AND APPLICATION OF POWER ANALYSIS**

Before beginning with this section, I urge the reader to consider the vast literature on the use and utility of power analyses in primary studies (e.g., Hoenig and Heisey 2001, Jennions and Møller 2003, Nakagawa and Cuthill 2007). The main points of these papers are also germane for meta-analysis—given that null review-level results occur and need explanation (Peterman 1995). This is because there are pitfalls in power analysis that the reader should be aware of before attempting power analyses with their data. For example, below I outline one of the most grievous pitfalls in power analyses—this is the “hard” power analysis where the observed effect size of a statistical test is used to calculate its power. In the remainder of the section, I outline the “retrospective” power statistics for the “classic” statistical tests of meta-analysis; these are tests for nonzero effects and the within- and between-study homogeneity tests. These statistics are called “retrospective” because they are calculated after an experiment was designed and analyzed (or in our case after an analysis with meta-analysis), and should be distinguished from the “prospective” power analyses commonly used in the design of experiments (e.g., estimating adequate sample sizes).

**Retrospective power analysis**

In response to Arnqvist and Wooster’s (1995a) influential review of meta-analysis in ecology and evolution, Peterman (1995) raised the following issue in regard to the power of inference of meta-analysis: How should researchers interpret a meta-analysis that fails to reject a null hypothesis? This question was timely given the already considerable debate in primary research on how to distinguish between a true biological null (zero effect) and a null outcome due to low statistical power (see Peterman 1990a, 1990b). The problem relates to the application and interpretation of tests for statistical power; there is a null (nonsignificant) statistical outcome, and a power test is applied to evaluate whether this outcome is false negative (type II error) or a true “biological” null. One way to achieve this is to use the observed (calculated) effect size of the study as the basis for estimating statistical power (Gerodette 1987, Petterman 1990a, Taylor and Gerodette 1993). This approach is referred to as a “hard” retrospective power analysis. However, it is now known that this type of power analysis is uninformative because it will always indicate that a null study has low power. This is because the study’s $P$-value and power
are both statistically dependent on the observed effect size (Colegrave and Ruxton 2003; also see Equation 22.1). A “hard” power analysis simply restates the statistical significance of the test (Thomas and Juanes 1996, Foster 2001). In other words, a nonsignificant statistical test will always have low power when the observed effect size is used to calculate power.

To avoid this problem, a “soft” retrospective approach is preferred where a set of hypothesized population effect sizes (e.g., small, medium, and large effects) is used to evaluate the power of the study (Cohen 1988). This small, medium, large “t-shirt” approach can also be used to evaluate the power of meta-analysis by using expected variances given the number of studies synthesized by meta-analysis (see below; and Hedges and Pigott 2001). However, hypothesizing small, medium, and large effects is also not without criticism because without prior knowledge of the potential magnitude of actual population effects (which are largely unknown for most ecological systems), the above “t-shirt” hypotheses might not provide a meaningful frame of reference.

Another way to generate more meaningful population effect size (or variances) estimates for power analysis is to use the pooled effects from previously published meta-analyses (Hedges and Pigott 2004). For example, there have already been five meta-analyses synthesizing studies testing for local adaptation in parasites (Van Zandt and Mopper 1998, Lajeunesse and Forbes 2002, Lively et al. 2004, Greischar and Koskella 2007, Hoeksema and Forde 2008). Future meta-analyses (and primary research) in this field should use the pooled effect sizes from these previous meta-analyses as a range of potential hypotheses to evaluate statistical power. Alternatively, a more continuous range of population effects could be explored analytically or with simulations to evaluate statistical power—however, for the purpose of the chapter, I focus entirely on the “t-shirt” approach because of its simplicity and ease of calculation.

**When is there sufficient power?**

The convention for many fields is that there is sufficient power when \( \rho \geq 0.8 \). This is based on Cohen’s (1977) initial recommendation that a type II error rate of 0.2 is acceptable for experiments in the social sciences (remember, \( \rho = 1 - 0.2 \)). For meta-analysis, Hunter and Schmidt (1990) argue for a type II error rate of 0.25 (\( \rho \geq 0.75 \)). This more conservative error rate is justified due to the inability of a meta-analyst to experimentally manipulate the number of studies available for synthesis (or the representation of individual studies within moderator groups to test hypotheses), thus recognizing a limited means to improve statistical power. This of course assumes that the meta-analyst conducted a rigorous search of the literature for all the relevant studies (Chapters 3 to 5). Meta-analysts in applied ecology might want to consider even more conservative type II error rates given how the high cost of not rejecting a false null hypothesis can adversely influence policy decisions and management practices (Peterman 1990a, Di Stefano 2003; Chapter 26).

**STATISTICAL METHODS FOR EVALUATING POWER**

This section outlines the statistical methods for calculating the power of meta-analysis, and illustrates their application using Torres-Vila and Jennions’ (2005) meta-analysis on female fecundity of moths and butterflies when mated with virgin or nonvirgin males. For additional examples and further details on the derivation of these statistics, see Hedges and Pigott (2001, 2004). It is also important to note that these power statistics are applicable to any effect size metric (e.g., Hedges’ \( d \), response ratios, correlation coefficients, etc.), and that these equations apply to both fixed- and random-effects models for pooling effect sizes. Finally, many of these
tests of statistical power for meta-analysis are available in a SAS macro (Cafri and Kromrey 2008); however, for the more simple statistical tests, I report how to calculate these using Microsoft Excel equations.

**Power statistics for mean effect sizes**

A frequent test in meta-analysis is to determine whether a mean (pooled) effect size ($\mu_r$) differs from a null effect ($\mu_0$). Typically, for most effect size metrics this null effect is zero ($\mu_0 = 0$). Hedges and Olkin (1985) describe a direct way of testing whether $\mu_r = \mu_0$ using the following two-tailed statistic:

$$Z = \frac{\bar{\mu} - \mu_0}{\sigma_r(\bar{\mu})}, \text{ where if } |Z| > c_{a/2} \text{ then } \bar{\mu} \neq \mu_0. \quad (22.6)$$

Here, $\sigma^2(\bar{\mu})$ is the variance of $\bar{\mu}$, and $c_{a/2}$ is the critical value where $Z$ rejects the hypothesis that $\mu = \mu_0$. This critical value is defined as the $100(1 - \alpha)$ percentile of the inverse standard normal distribution, and equals $c_{0.05/2} = 1.96$ when the conventional significance level ($\alpha$) of 0.05 is assumed. For other significance levels, $c_{a/2}$ can be calculated in Microsoft Excel using the following equation: $= \text{NORMSINV} (1 - \alpha/2)$. Finally, the $P$-value of this $Z$ test is calculated in Excel using: $= 2 \times (1 - \text{NORMSDIST} (Z))$. The one-tailed version of this $Z$ test assumes a priori that the difference between the mean and null effect size has a known and predicted direction. For example, $\bar{\mu}$ will always be greater or equal to $\mu_0$. The critical value for this one-tailed test of $Z > c_\alpha$ becomes $c_{0.05} = 1.645$.

Assuming a significance level of 0.05, the statistical power ($\rho$) for the above $Z$ test is:

$$\rho^{2}_{\text{two-tailed}} = 2 - \Phi(1.96 - \bar{Z}) - \Phi(1.96 + \bar{Z}). \quad (22.7)$$

Here, $\Phi(x)$ is the standard normal cumulative function, and is calculated in Excel using: $= \text{NORMSDIST} (x)$. Whereas the statistical power for the one-tailed $Z$-test, with a priori knowledge on the direction of $\bar{\mu}$ relative to $\mu_0$, is

$$\rho^{2}_{\text{one-tailed}} = 1 - \Phi(1.645 - \bar{Z}). \quad (22.8)$$

Note that Equations 22.7 and 22.8 do not use the $Z$ value calculated directly from Equation 22.6. This would result in a “hard” retrospective power analysis (see the previous section for why this should be avoided). Instead, “soft” estimates of $Z$ (referred to here as $\bar{Z}$), that are independent from the observed pooled effect size $\bar{\mu}$, are hypothesized.

Estimating $\bar{Z}$ to perform a “soft” retrospective power-analysis requires both the observed sampling variance of the pooled effect $\sigma^2(\bar{\mu})$ and a hypothesis about the true population effect ($\mu$). Estimating the sampling variance is straightforward, and can be calculated directly using meta-analysis (Chapters 8 and 9). Hypothesizing the population effect $\bar{\mu}$ is less straightforward but can be derived from published information or by surveying a range of magnitudes of effect when calculating power (see previous section; Muncer et al. 2002, 2003). When published information on $\bar{\mu}$ is missing, a reviewer can hypothesize small, medium, and large values of $\bar{\mu}$ to evaluate the sensitivity in power of meta-analysis under these different magnitudes of effect. Cohen (1988) describes an effect size (measured as a standardized mean difference) of 0.2 as small, 0.5 as medium, and 0.8 as large. Although these values seem arbitrary, effect sizes like Hedges’ $d$ are measured in units of standard deviations, such that an effect of 0.2 actually indicates a difference of a fifth of a standard deviation. To integrate $\bar{\mu}$ into power analyses, $\bar{\mu}$ replaces $\mu$ from Equation 22.6; this yields the appropriate $\bar{Z}$ for Equations 22.7 and 22.8. Alternatively, a meta-analyst can use the range of observed effect sizes as good hypotheses on
TABLE 22.1. The statistical power of the pooled effect sizes (\( \tilde{\mu} \)) from Torres-Vila and Jennions’ (2005) meta-analysis on the fecundity of female moths and butterflies when mated with virgin or nonvirgin males. This analysis also includes mating system (e.g., polyandrous or monandrous) as a moderator variable to explain variation in research outcomes among the moths and butterflies used to test this hypothesis. Indicated in bold are the tests with low power (e.g., \( P_{\text{two-tailed}}^z < 0.75 \)).

<table>
<thead>
<tr>
<th>Grouping</th>
<th>( K )</th>
<th>Effect size</th>
<th>Statistical power (( P_{\text{two-tailed}}^z ))</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( \tilde{\mu} )</td>
<td>95% CI</td>
</tr>
<tr>
<td>Fixed effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all studies</td>
<td>25</td>
<td>0.36</td>
<td>0.27–0.45</td>
</tr>
<tr>
<td>polyandrous</td>
<td>12</td>
<td>0.48</td>
<td>0.35–0.61</td>
</tr>
<tr>
<td>monandrous</td>
<td>13</td>
<td>0.25</td>
<td>0.13–0.37</td>
</tr>
<tr>
<td>Random effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>all studies</td>
<td>25</td>
<td>0.33</td>
<td>0.21–0.46</td>
</tr>
<tr>
<td>polyandrous</td>
<td>12</td>
<td>0.46</td>
<td>0.28–0.64</td>
</tr>
<tr>
<td>monandrous</td>
<td>13</td>
<td>0.22</td>
<td>0.05–0.39</td>
</tr>
</tbody>
</table>

small and large effects—where for example, the smallest and largest observed effect size from the meta-analysis are treated as hypotheses of \( \tilde{\mu} \) (Hedges and Pigott 2001).

Table 22.1 shows the results of hypothesizing Cohen’s small, medium, and large values of \( \tilde{\mu} \) when estimating the statistical power of Torres-Vila and Jennions’ (2005) meta-analysis on the fecundity of female moths and butterflies when mated with virgin or nonvirgin males. This study had reasonable power to detect nonzero effects using a fixed-effects model, but under a random-effects model the tests had lower power to detect small effects. This result is expected given that a random-effects model will always have lower power relative to a fixed-effects model (Cohn and Becker 2003), and also given that the 95% CI nearly overlapped the null effect. A significant nonzero effect with low power should be interpreted cautiously. It offers only partial evidence that there is a true effect. Here, more data are needed to fully justify that the observed effect has a biological basis and that it is not due to sampling error.

Confidence intervals for assessing power: Advantages and limitations

Confidence intervals (CI’s) are a measure of statistical power—they estimate study uncertainty and provide information on the underlying population parameters, such as variance (Hayes and Steidl 1997, Hoenig and Heisey 2001, Colegrave and Ruxton 2002). However, more typically CI’s are used to test whether mean effect sizes (\( \tilde{\mu} \)) differ from zero (the null prediction or \( \mu_0 \)). If the 95% CI does not include the value specified by the null hypothesis (usually zero), then \( \mu_0 \) can be rejected (Steidl and Thomas 2001). The breadth of the CI around the mean effect size (\( \tilde{\mu} \)) indicates the statistical power of the test, because estimates of CI are related to estimates of power (Nakagawa and Cuthill 2007). For example, Figure 22.2B shows the rapid reduction in breadth of the 95% CI as the review sample size increases. This is also why confidence intervals, much like power analysis, can be used \textit{prospectively} in primary research to determine ideal sample sizes for experimentation (see Goodman and Berlin 1994). Unfortunately, an equivalent statistic to confidence intervals is not available for homogeneity tests, so the meta-analyst must apply the “soft” approach to assess the power of these tests.
Power statistics for homogeneity tests

Homogeneity tests, such as $Q$ statistics, are used to determine whether a group of studies share a common effect size. These analyses are important because they help assess the appropriateness of the fixed-effects model, and whether studies should be parsed among different moderator groups (Chapter 8). Below, I present the power statistics for the homogeneity test of fixed-effects ($\rho^Q$) and random-effects ($\rho^{Q_r}$) models. I only briefly describe the theory behind these homogeneity tests. For further details of these models, see Chapter 8. Homogeneity tests can also be applied in an ANOVA style meta-analysis where studies are parsed among moderator groups to evaluate within- and between-group variation (Hedges and Olkin 1985). Here, between-study homogeneity ($Q_B$) is calculated to test whether the pooled effect sizes among these groupings differ. These moderator groupings are important to test hypotheses and serve to explain variation among effect sizes beyond what is expected due to sampling error.

**Fixed-effects model: Within-study homogeneity**

Homogeneity statistic $Q_W$ tests whether the effect sizes ($\delta$) across a group of $K$ studies share a common effect (e.g., $\delta_1 = \delta_2 = \ldots = \delta_k$) and is calculated as follows for the fixed-effects model:

$$Q_W = \sum_{i=1}^{K} \frac{(\delta_i - \bar{\delta})^2}{v_i}$$

where $Q_W > \chi^2_{K-1}$ then $\delta_1 \neq \delta_2 \neq \ldots \neq \delta_k$.

(22.9)

Here, $v_i$ is the observed variance of each $i$th effect size. A significant $Q_W$ indicates that there is variation (heterogeneity) among effect sizes that is not explained solely by sampling error. The power of this homogeneity test is

$$\rho^{Q_w} = 1 - F(c_a|K-1; Q_W)$$

(22.10)

Power is calculated using the noncentral chi-square ($\chi^2$) cumulative distribution function, described here as $F(x | b; \lambda)$. The parameters of this function are as follows: $x$ is $c_a$, the 100(1-$\alpha$) percent point of the central $\chi^2$ distribution with $b$ degrees of freedom (here $b = K - 1$); and the noncentrality parameter ($\lambda$) is estimated with $Q_W$. Below, I describe how to hypothesize different values of $Q_W$ to evaluate power. Microsoft Excel can be used to estimate $c_a$: CHIINV ($\alpha$, $K - 1$). Unfortunately, Excel does not have a noncentral $\chi^2$ cumulative distribution function, but statistical software like SAS and SPSS provide such functions; these are PROBCHI ($c_a$, $K - 1$, $Q_W$) and NCDF.CHISQ ($c_a$, $K - 1$, $Q_W$), respectively. This function can also be found on Casio Computer Company online calculator (see http://keisan.casio.com/).

Estimating $Q_W$ requires making hypotheses about plausible levels of heterogeneity. Again we can apply a “rule of thumb” developed in the social sciences for possible values for these hypotheses. Hedges and Pigott (2001) describe a convention of three magnitudes of heterogeneity:

$$\hat{Q}_W = \begin{cases} 
\text{small} = .33(K - 1) \\
\text{medium} = .66(K - 1) \\
\text{large} = K - 1.
\end{cases}$$

(22.11)

These values are based on a review of meta-analyses which found that the ratio of between- and within-study variance is typically 0.33, but rarely exceeds 1 (Schmidt 1992). The $Q_W$ test can be seen as a crude evaluation of this ratio; where the top half of Equation 22.9 is the between-studies variance ($\delta - \bar{\delta}$) and the bottom $v$ is within-study variance. An application of these hypotheses of $\hat{Q}_W$ to evaluate the statistical power of homogeneity tests is found in Table 22.2. In general, Torres-Vila and Jennions’ (2005) $\hat{Q}_W$ tests had low power to detect heterogeneity, and in particular the low power of the $\hat{Q}_W$ test among monandrous taxa indicates that
its nonsignificance is susceptible to a type II error. Of course, low power to detect significant heterogeneity does not necessarily indicate that such heterogeneity exists, only that the test remains inconclusive given the available data.

**Fixed-effects model: Between-study homogeneity**

The between-study homogeneity statistic ($Q_B$) evaluates whether a collection of pooled effect sizes parsed among $m$ moderator groups differs (e.g., $\bar{\mu}_1 = \bar{\mu}_2 = \ldots = \bar{\mu}_m$). Note that $Q_B$ is an omnibus test and needs only one of the grouped pooled effects to differ in order to be significant. When $m > 2$, post hoc contrasts are needed to evaluate which moderator groups differ (see Hedges and Pigott 2004). The between-study homogeneity test for the fixed-effects model is calculated as

$$Q_B = \sum_{j=1}^{m} \frac{(\bar{\mu}_j - \bar{\mu})^2}{\sigma^2(\bar{\mu}_j)},$$

and $\bar{\mu}$ is the pooled effect size across all studies. The statistical power of $Q_B$ is evaluated with:

$$\rho^{\Omega_2} = 1 - F(c_x|m - 1; Q_B),$$

where $F(x \mid v; \lambda)$ is the same noncentral $\chi^2$ cumulative distribution function used in Equation 22.9.

Hypothesizing $Q_B$ for power analysis when there are more than two groups ($m > 2$) is difficult and requires assumptions about the expected magnitude of differences between each group (Hedges and Pigott 2001). For simplicity, I only consider the case where $Q_B$ evaluates the difference between two groups. Here $Q_B$ can be estimated as follows:

$$\bar{Q}_B = \frac{(\bar{\mu}_B)^2}{\sigma^2(\bar{\mu}_1) + \sigma^2(\bar{\mu}_2)}.$$  \hspace{1cm} (22.14)

In this equation, $\bar{Q}_B$ is based on the observed variances of the pooled effect sizes of both groups, and a hypothesis about the possible magnitude of difference between each group ($\bar{\mu}_B$). Again, small, medium, and large values for $\bar{\mu}_B$ can be hypothesized following Cohen’s (1988) rule of thumb. Using these hypotheses for $\bar{\mu}_B$, Table 22.2 describes the power tests for the

<table>
<thead>
<tr>
<th>Grouping</th>
<th>Homogeneity tests</th>
<th>Statistical power ($\rho^{\Omega_2}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Q$</td>
<td>df</td>
</tr>
<tr>
<td><strong>Within-group</strong> *</td>
<td></td>
<td></td>
</tr>
<tr>
<td>All studies</td>
<td>46.5</td>
<td>24</td>
</tr>
<tr>
<td>polyandrous</td>
<td>30.7</td>
<td>11</td>
</tr>
<tr>
<td>monandrous</td>
<td>9.2</td>
<td>12</td>
</tr>
<tr>
<td><strong>Between-group</strong> **</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fixed effects</td>
<td>6.6</td>
<td>1</td>
</tr>
<tr>
<td>random effects</td>
<td>3.4</td>
<td>1</td>
</tr>
</tbody>
</table>

* Within-group hypotheses on small $Q_B = 0.33(K - 1)$, medium $Q_B = 0.66(K - 1)$, and large $Q_B = K - 1$ amounts of heterogeneity.

** Between-group hypotheses on small $\bar{\mu}_B = 0.2$, medium $\bar{\mu}_B = 0.5$, and large $\bar{\mu}_B = 0.8$ differences between polyandrous and monandrous group effects.
between-group fixed-effects homogeneity statistics. These tests show that Torres-Vila and Jen-
nions’ (2005) meta-analysis had low power to detect a difference between polyandrous and
monandrous groups, should the “true” difference between these groups be small.

**Random-effects model: Between-study homogeneity**

Testing for within-study homogeneity is uninformative in the random-effects model because
it does not assume that studies share a common population effect size (see Hedges and Olkin
1985). However, the between-study homogeneity test under the random-effects model ($Q^*_B$) is
still useful for evaluating differences among moderator groups. The homogeneity test under the
random-effects model uses an estimate of the between-study variance ($\tau^2$) to adjust the within-
study variances of each effect size ($v_i$). Under this model, the variance of each study will equal
$v_i^* = v_i + \tau^2$, and the new random-effects variance $v^*$ replaces $v$ in Equation 22.9. This between-
study variance is estimated as follows: when $Q^*_w$ from the fixed-effects model is less than $K - 1$,
then $\tau^2$ will equal zero. Otherwise, the DerSimonian-Laird estimator is

$$\tau^2 = \frac{Q^*_w - (K - 1)}{\sum v_i^{-1}} - \frac{(\sum v_i^{-1})(\sum v_i^{-1})^{-1}}{\sum v_i^{-1}}. \tag{22.15}$$

The statistical power of $Q^*_B$ is evaluated with the following:

$$\rho^{Q^*_B} = 1 - F(c, m - 1; Q^*_B) \tag{22.16}$$

where $F(x \mid v; \lambda)$ is the same noncentral $\chi^2$ cumulative distribution function used in Equation
22.9. As in the fixed-effects model, different hypotheses about the magnitude of heterogeneity
are useful to evaluate the statistical power of $Q^*_B$. These hypotheses of $Q^*_B$ are found in Equation
22.11. Table 22.2 shows that the between-study homogeneity test assuming a random-effects
model had even lower power to detect small differences between monandrous and polyandrous
than the fixed-effects model did. This difference in power is expected given that the random-
effects model adds $\tau^2$ to the variances of each effect size. This additional variance component
results in a less sensitive, but more conservative, hypothesis test (see above section; also see

**Limitations of power statistics for meta-analysis**

An important limitation of these power analyses is that their behaviour is unknown when the
studies included in meta-analysis violate the normality assumption of statistical models. For
example, this occurs when estimates of power that are based on noncentral distributions of test
statistics (e.g., the noncentral $\chi^2$ distributions used in Equation 22.7) are tied directly to the sta-
tistical inference procedures used for data sampled from normal populations. Because of these
assumed distributions, power analyses are more sensitive to violations of statistical normality
than significance tests, which rely only on central distributions of test statistics (Hedges and
Pigott 2001).

**PRACTICAL WAYS TO IMPROVE POWER**

Given the several factors contributing to the statistical power of meta-analysis, Table 22.3 out-
lines a few approaches that can improve the statistical power of meta-analysis and homogeneity
tests. These suggestions rely mostly on improving the representation of studies among modera-
tor groups, and on improving the quality of data used to estimate effect sizes and variances.
Table 22.3. A few practical ways to improve the statistical power of meta-analysis.

<table>
<thead>
<tr>
<th>Do the following</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collate more studies.</td>
<td>Decreases variances and 95% CI of pooled effect sizes. Also allows for greater representation among moderator variables to test hypotheses, and diminishes potential for publication bias.</td>
<td>Time and resource intensive (see Chapters 4 and 5 for more details). May not improve the estimation performance of the random-effects model for pooling effect sizes.</td>
</tr>
<tr>
<td>Decide on an appropriate experimental outcome to quantify.</td>
<td>Using experimental measurements that are closely aligned with the biological effect of interest will increase the likelihood of detecting effects, and to detecting the difference between means or the strength of correlations between traits.</td>
<td>With a more explicit outcome, fewer studies will fit this definition, and thus fewer effect size data will be available for meta-analysis.</td>
</tr>
<tr>
<td>Use an appropriate effect size metric to quantify results.</td>
<td>Effect size metrics differ in the amount of information required to quantify an effect size. The more precise the effect size metric, the better the estimate of the population effect will become.</td>
<td>Using an effect size metric that is too stringent will decrease the number of studies that will fit the requirements for metric. Using effect size metrics that are too liberal will decrease the precision of the effect size to evaluate the underlying effect.</td>
</tr>
<tr>
<td>Convert effect sizes into a metric that satisfies model assumptions (e.g., transform correlations to Fisher’s z).</td>
<td>Improves variance estimates and narrows the 95% CI.</td>
<td>No disadvantage! Although, Fisher’s z are more difficult to interpret and less intuitive than raw correlations.</td>
</tr>
<tr>
<td>Use an estimate of the effect size variance that requires more information from individual studies.</td>
<td>More information is included for each study and thus increases the efficiency of the weights when pooling results.</td>
<td>Incomplete information and variable reporting of the required statistics can make this difficult (see Chapter 6). Thus variances are estimated using resampling methods or simplified surrogates of variance.</td>
</tr>
<tr>
<td>Avoid or account for the nonindependence of effect sizes (see Chapters 15 through 17).</td>
<td>Improves the estimate of pooled effect sizes and their variances, and avoids “pseudoreplication” that can inflate the conclusions drawn from the review.</td>
<td>Accounting for nonindependence requires knowledge on how effect sizes are correlated. This information is rarely available, but can be extrapolated from other sources.</td>
</tr>
<tr>
<td>Exclude outliers.</td>
<td>Decreases the deviation of the pooled effect size from the null expectation, and improves the evaluation of study heterogeneity.</td>
<td>Limits the scope of the review, and decreases the amount of potential studies parsed among moderator groups.</td>
</tr>
</tbody>
</table>

(continued)
For example, when effect size metrics that require multiple pieces of information to quantify research outcomes (e.g., Hedges’ $d$) are used over less restrictive metrics (e.g., $\ln R$), studies lacking the necessary information are often excluded from the meta-analyses. Smaller sample sizes will decrease the statistical power of meta-analysis. However, Lajeunesse and Forbes (2003) found that meta-analyses based on few, but high precision, effect sizes (relative to the same number of studies estimated with coarse effect sizes) had improved error rates because more within-study information was used to control for bias. A balance must be met between restricting analyses to studies with precise information and expanding the scope of the review by including studies with coarse effect size data. Increasing the precision of effect size data is one way to improve the statistical power of meta-analyses. The closer the effect size estimate is aligned with the biological phenomenon of interest, the stronger the hypothesis test becomes.

**CONCLUSION AND PROSPECTUS**

The scope of power analysis for meta-analysis spans a greater variety of statistical issues than those typically covered in primary research. Power statistics in primary research are more commonly used to evaluate null outcomes, but here with meta-analysis they can also serve as a tool to assess the detection of heterogeneity among effect sizes. I believe this latter application will be the real strength of power analysis for meta-analysis—given that confidence intervals are already heavily used and preferred over power statistics for evaluating null effects (Nakagawa and Cuthill 2007). Assessing heterogeneity is central to meta-analysis, because it provides the basis for hypothesis testing and exploring whether moderator variables (or covariates) explain variation in research outcomes (either experimental or ecological). Providing confidence to these tests is essential given the potential risk of overlooking or underemphasizing biologically meaningful explanations for variation in research because of low statistical power.

Finally, this chapter focuses entirely on the **retrospective** use of power analysis for meta-analysis. Another application uses power statistics **prospectively** for planning research and designing experiments where replication or funding is limited (Møller and Jennions 2001, 2002). The **prospective** approach applied this way will have limited use for conducting meta-analyses, given that meta-analysts cannot experimentally manipulate the number of published/unpublished studies available or “design” the ideal meta-analysis. Here, a meta-analysis is a

### TABLE 22.3. Continued

<table>
<thead>
<tr>
<th>Do the following</th>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impute missing data (see Chapter 13).</td>
<td>More studies are included in the overall analysis, despite not having complete information. Including studies is always better than excluding studies!</td>
<td>Requires making assumptions of the statistical distribution of missing data. These assumptions may not have a biological basis.</td>
</tr>
<tr>
<td>Use moderator groupings or covariates to explain variation among research outcomes.</td>
<td>Removes “noise” among effect sizes, decreasing the pooled variance. Can help distinguish between experimental and biological effects.</td>
<td>Including additional parameters in models decreases the performance of statistical tests (e.g., the df’s of $Q$-tests). Moderator groups will have smaller sample sizes and thus are more prone to sampling error, and have greater variances.</td>
</tr>
</tbody>
</table>

Do the following

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Impute missing data (see Chapter 13).</td>
<td>More studies are included in the overall analysis, despite not having complete information. Including studies is always better than excluding studies!</td>
</tr>
<tr>
<td>Use moderator groupings or covariates to explain variation among research outcomes.</td>
<td>Removes “noise” among effect sizes, decreasing the pooled variance. Can help distinguish between experimental and biological effects.</td>
</tr>
<tr>
<td>Requires making assumptions of the statistical distribution of missing data. These assumptions may not have a biological basis.</td>
<td>Including additional parameters in models decreases the performance of statistical tests (e.g., the df’s of $Q$-tests). Moderator groups will have smaller sample sizes and thus are more prone to sampling error, and have greater variances.</td>
</tr>
</tbody>
</table>
retrospective endeavor, limited to the diversity and abundance of published literature—these data cannot be manipulated to increase or maximize the power of meta-analyses.

However, one potential application of prospective power analysis is to determine when to stop adding studies to a meta-analysis (see Sutton et al. 2007); that is, to detect when there is minimal gain by including more research outcomes. This application could be useful when time and resources to process data for meta-analysis are limited, and exploration of moderator effects is a minor component of the review. For example, Figure 22.2A shows that reviews with more than 40 studies do not show a significant improvement for estimating the population effect of 0.5. Another application of the prospective approach would be to calculate the minimum number of studies needed to detect an existing effect (this is not the same as calculating the fail-safe number; Chapter 14). This can be applied to preliminary or exploratory meta-analyses with thousands of potential studies where effect sizes need to be extracted. This application requires a working hypothesis of the between-study variance ($\tau^2$). Here, the between-study variance can be estimated from a random subsample of the studies available for review. Hedges and Olkin (1985) provide a test to evaluate whether $\tau^2$ is nonzero. If there is little heterogeneity among effect sizes (between-study variance), then a smaller review sample size could be justified.

An important goal for performing power analyses in meta-analysis should be to inform future research. Meta-analyses may potentially identify new and interesting effects, or may provide inconclusive evidence for important hypothesis tests of ecological or evolutionary theory. However, if these review-level outcomes are identified as having low statistical power, then this serves as an important stepping point for new experiments to test the validity of these findings. Alternatively, if large effects with strong statistical power are identified, then these effects should be used for the prognostic calculation of statistical power when designing new experiments.

ACKNOWLEDGEMENTS

I thank Terri Pigott, Alex Sutton, and Michael Jennions for comments. Work on this chapter was supported by the University of South Florida and a National Science Foundation grant to the National Evolutionary Synthesis Center (EF-0423641).
REFERENCES


References


References


References

Colegrave, N. and G. D. Ruxton. 2003. Confidence intervals are a more useful complement to non-significant tests than are power calculations. *Behavioral Ecology* 14:446–450.


References


References


References


References


Festa-Bianchet, M. 1996. Offspring sex ratio studies of mammals: Does publication depend upon the quality of the research or the direction of the results? *Ecoscience* 3:42–44.


References


Gilpin, A. R. 1993 Table for conversion of Kendall’s tau to Spearman’s rho within the context of measures of magnitude of effect for meta-analysis. Educational and Psychological Measurement 53:87–92.


Greenwald, A. G., R. Gonzalez, R. J. Harris, and D. Guthrie. 1996. Effect sizes and p-values: What should be reported and what should be replicated? *Psychophysiology* 33:175–188.


References


Ioannidis, J.P.A. 1998. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. Journal of the American Medical Association 279:281–286.


Ioannidis, J.P.A. 2005c. Why most published research findings are false. PLOS Medicine 2: e124. doi:10.1371/journal.pmed.0020124


References


References


References


References


References


References


References


References


References

domized trials. Statistical Methods in Medical Research 17:279–301.
Salanti, G., F. K. Kawoura, and J. P. A. Ioannidis. 2008. Exploring the geometry of treatment net-
bias in observational studies in epidemiology: A systematic review and annotated bibliography.
Santos, E.S.A., D. Scheck, and S. Nakagawa. 2011. Dominance and plumage traits: Meta-analysis
et al. 2007. Ecological and evolutionary insights from species invasions. Trends in Ecology &
Species richness, species-area curves and Simpson’s paradox. Evolutionary Ecology Research
2:791–802.
Scherer, R. W., P. Langenberg, and E. von Elm. 2007. Full publication of results initially presented
in abstracts. Cochrane Database of Methodology Reviews Issue 2, Art. No.: MR000005. doi:
Behavioral Ecology 18:115–120.
Volume 321: Numerical Computer Methods, Part C, edited by M. L. Johnson and L. Brand,
2003. Constructing a database of individual clinical trials for longitudinal analysis. Controlled
of the control rate as a predictor of treatment efficacy in meta-analysis of clinical trials. Statistics
in Medicine 17:1923–1942.
ociations between heterogeneous treatment effects and study-level, but not patient-level, factors.
References


References


References


Sterne J.A.C., R. Harris, R. Harbord, and T. J. Steichen. 2009. Meta-Analysis in Stata: An Updated Collection from the Stata Journal. Stata Press, College Station, TX.


Walter, S. D. and X. Yao. 2007. Effect sizes can be calculated for studies reporting ranges for outcome variables in systematic reviews. *Journal of Clinical Epidemiology* 60:849–852.


References


