Statistical Methods for Meta-Analysis

September 16, 2020

Alessio Crippa

Department of Medical Epidemiology and Biostatistics
Karolinska Institutet

Systematic review

Fixed-effect model

Random-effects model

Heterogeneity

Model choice

Sensitivity analysis

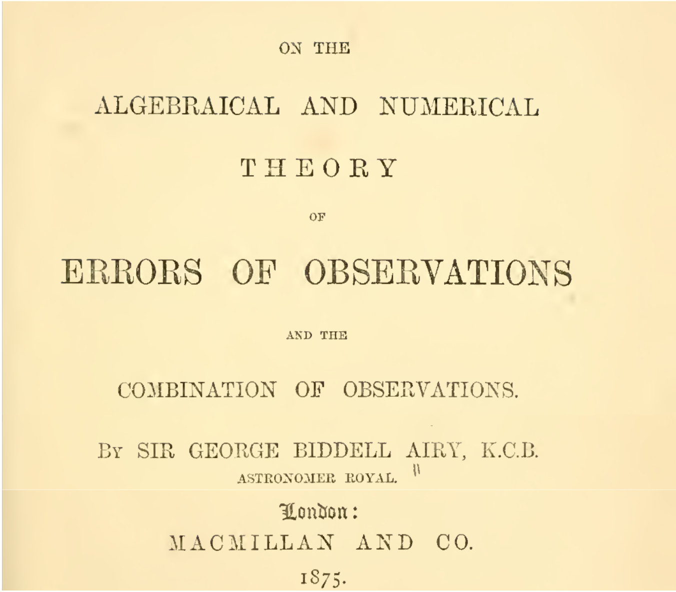
Systematic review

Systematic review

The process of systematically reviewing and integrating research evidence is described with different terms (systematic review, meta-analysis, research synthesis, overview, pooling).

The aim of a systematic review is to evaluate and syntethise the results published over time in the literature on a specific field.

Meta-analysis is the statistical method to combine or pool results obtained from distinguished, but comparable studies.



ON THE

ALGEBRAICAL AND NUMERICAL

THEORY

OF

ERRORS OF OBSERVATIONS

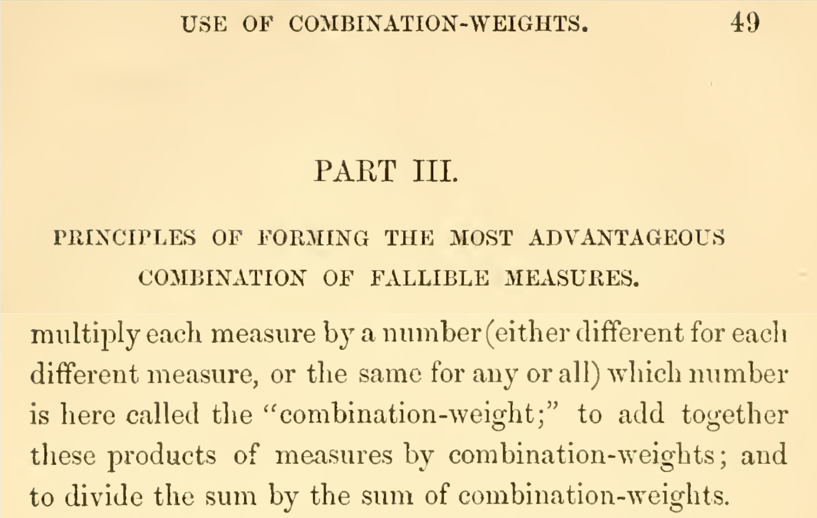
AND THE

COMBINATION OF OBSERVATIONS.

BY SIR GEORGE BIDDELL AIRY, K.C.B.

ASTRONOMER ROYAL.

London:
MACMILLAN AND CO.
1875.



USE OF COMBINATION-WEIGHTS. 49

PART III.

PRINCIPLES OF FORMING THE MOST ADVANTAGEOUS COMBINATION OF FALLIBLE MEASURES.

multiply each measure by a number (either different for each different measure, or the same for any or all) which number is here called the "combination-weight;" to add together these products of measures by combination-weights; and to divide the sum by the sum of combination-weights.

Models for meta-analysis

Two common models:

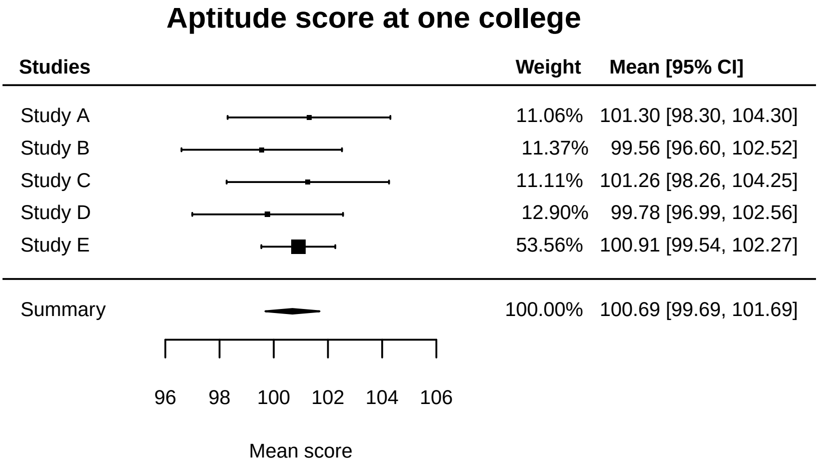
- Common or fixed-effect model: one true effect size. Differences in the observed effect sizes are only due to sampling error.
- Random-effects model: the true effect size might varies across studies. Differences in the observed may also be due to differences in the studies (mixes of participants, implementations of interventions, etc.)

The choice between the two models is critical. It doesn't affect only the computations, but also the goal of the analyses and the interpretations of the results.

As formulas are not always intuitive, it helps to keep in mind that the summary effect is a weighted mean of the observed effect sizes.

Fixed-effect model

Example of a fixed-effect analysis



Studies	Weight	Mean [95% CI]
Study A	11.06%	101.30 [98.30, 104.30]
Study B	11.37%	99.56 [96.60, 102.52]
Study C	11.11%	101.26 [98.26, 104.25]
Study D	12.90%	99.78 [96.99, 102.56]
Study E	53.56%	100.91 [99.54, 102.27]
Summary	100.00%	100.69 [99.69, 101.69]

Fixed-effect model

All the studies share a common (true and unknown) effect size, denoted by θ. Effect sizes might differ because of random variability.

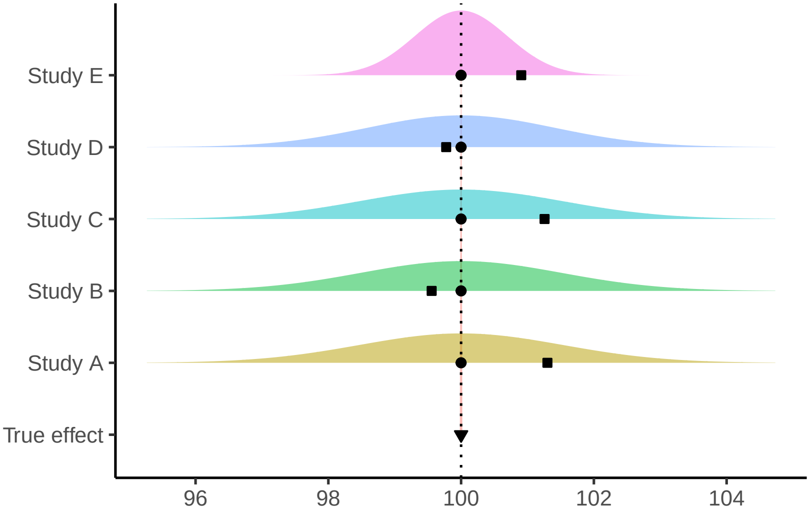
$$Y_i = \theta + \epsilon_i$$

i is an index for the study $i = 1, \ldots, k$

Y_i is the observed effect size (mean) in the i-th study

θ is the unknown true parameter

ϵ_i is residual deviation from the common effect in the i-th study


Study E
Study D
Study C
Study B
Study A
True effect
96
98
100
102
104

Inverse variance weighted mean

We aim to estimate the unknown true effects starting from the observed effect sizes.

The most precise (which minimizes the variance) estimate is a weighted mean with weights equal to the inverse of the study's variance:

$$W_i = \frac{1}{V_{Y_i}} = \frac{1}{\mathrm{SE}(Y_i)^2}$$

V_{Y_i} is the within-study variance for study, e.g. an estimate of the (square) standard error of the observed mean score in the *i*-th study.

Point estimation

The weighted mean (M) is computed as

$$M = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}$$

the sum of the products $W_i Y_i$ (effect size multiplied by weight) divided by the sum of the weights.

Variance and interval estimation

The variance and standard error of the summary effect are given by:

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i}$$

$$SE_M = \sqrt{V_M}$$

95% confidence interval for the summary effect can be calculated using the normal approximation

$$M \pm 1.96 \times SE_M$$

Hand calculations

```
    study     y    se     v     w    wy std_w
1 Study A 101.3 1.532 2.347 0.426  43.2 0.111
2 Study B  99.6 1.511 2.283 0.438  43.6 0.114
3 Study C 101.3 1.528 2.336 0.428  43.3 0.111
4 Study D  99.8 1.419 2.013 0.497  49.6 0.129
5 Study E 100.9 0.696 0.485 2.063 208.2 0.536
```

$$M = \frac{101.301 \cdot 0.426 + 99.556 \cdot 0.438 + 101.257 \cdot 0.428 + 99.775 \cdot 0.497 + 100.906 \cdot 2.063}{0.426 + 0.438 + 0.428 + 0.497 + 2.063} =$$

$$= \frac{43.169 + 43.611 + 43.345 + 49.571 + 208.182}{3.852} = \frac{387.878}{3.852} = 100.7$$

$$V_M = \frac{1}{0.426 + 0.438 + 0.428 + 0.497 + 2.063} = \frac{1}{3.852} = 0.26$$

$$M \pm 1.96 \times SE_M = 100.7 \pm 1.96 \cdot \sqrt{0.26} = (99.7, 101.7)$$

Hypothesis testing

A Z-value test can be used for testing the null hypothesis that the true efftc is equal to zero $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$

$$Z = \frac{M}{SE_M}$$

A two-sided p-value is obtained from the cumulative standard normal function

$$p = 2[1 - \Phi(|Z|)]$$

Studies on Prophylactic Use of Lidocaine After a Heart Attack (Hine et al. (1989))

Code available at https://rpubs.com/alecri/code_meta

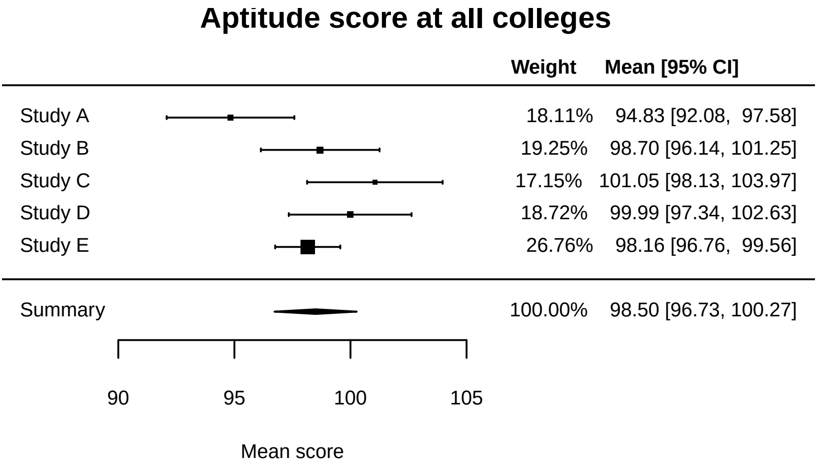
Meta-analysis of death rates in 6 randomized controlled trials evaluating mortality from prophylactic use of lidocaine in acute myocardial infarction.

Research question: Is there a detrimental effect of lidocaine on mortality?

The studies, taken individually, are too small to detect important differences in mortality rates.

Random-effects model

Example of a random-effects analysis



Aptitude score at all colleges

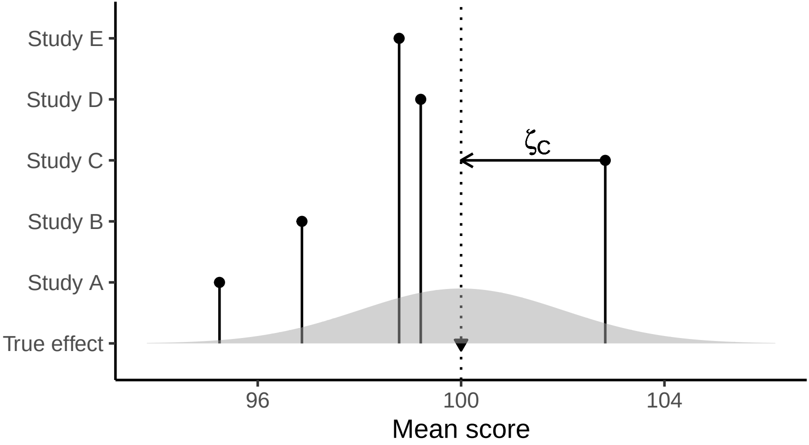
	Weight	Mean [95% CI]
Study A	18.11%	94.83 [92.08, 97.58]
Study B	19.25%	98.70 [96.14, 101.25]
Study C	17.15%	101.05 [98.13, 103.97]
Study D	18.72%	99.99 [97.34, 102.63]
Study E	26.76%	98.16 [96.76, 99.56]
Summary	100.00%	98.50 [96.73, 100.27]

The assumption of a common effect size for all the studies is often implausible.

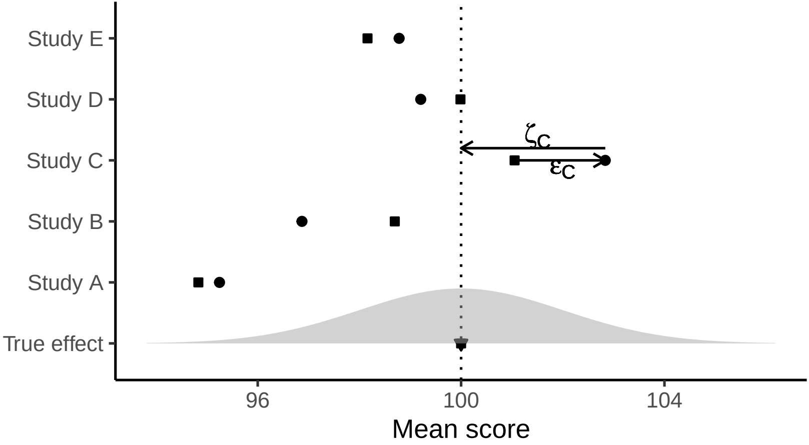
We assume instead that there is a distribution of true effect sizes and that those studies are somehow similar, so that it makes sense to combine this information.

Our aim is to estimate the mean or summary effect of this distribution.

Between-studies variation



Within-study variation



Random-effects model

$$Y_i = \mu + \zeta_i + \epsilon_i$$

i is an index for the study $i = 1, \ldots, k$

Y_i is the observed effect size (mean) in the i-th study

μ is the overall mean of the (true) study-specific effects

ζ_i is distance between the overall mean and the study-specific (true) effect

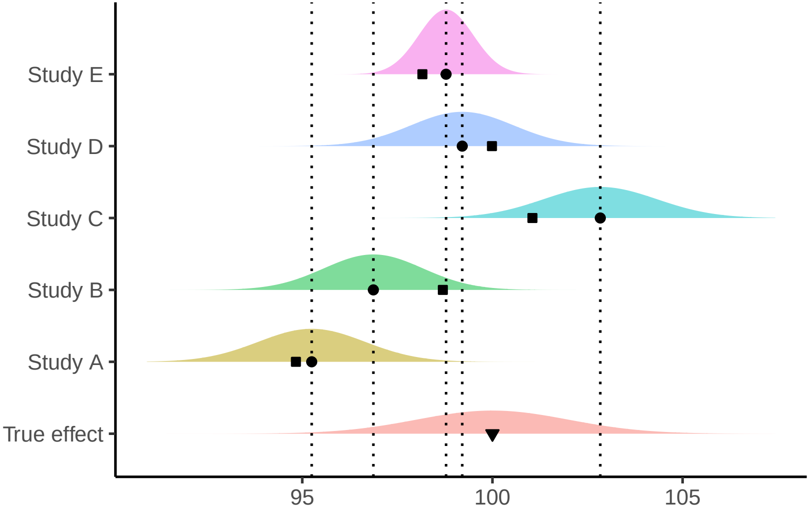
$\theta_i = \mu + \zeta_i$ is the (true) effect in the i-th study

ϵ_i is residual deviation from the true effect in the i-th study

Key points

The distance from μ and θs depends on τ the standard deviation of the distribution of the true effects (τ^2 is the variance)

The distance from θ_i and Y_i depends on V_{Y_i}, the sampling distribution. It varies across studies.


Study E
Study D
Study C
Study B
Study A
True effect
95
100
105

Inverse variance weighted mean

We aim to estimate the overall mean of the distribution of unknown effect sizes.

As before, the weights are proportion to the inverse of that study's variance. However, under random-effects model, the study's variances are given by $\tau^2 + V_{Y_i}$.

We need an estimate for the between-studies variance.

Der Simonian and Laird estimator of τ^2

Der Siminonian & Laird proposed a moment based estimate of the between-studies variance.

$$T^2 = \max\left\{ \frac{Q-(k-1)}{\sum_{i=1}^k W_i - \frac{\sum_{i=1}^k W_i^2}{\sum_{i=1}^k W_i}}, 0 \right\}$$

where Q is the heterogeneity test statistic

$$Q = \sum_{i=1}^{k} W_i Y_i^2 - \frac{\left(\sum_{i=1}^k W_i Y_i\right)^2}{\sum_{i=1}^k W_i}$$

Point estimation

The random effects estimate is the weighted average

$$M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}$$

where the weights W_i^* incorporate the between-studies variance:

$$W_i^* = \frac{1}{V_{Y_i}^*} = \frac{1}{V_{Y_i} + T^2}$$

Interval and hypothesis testing

$$V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*}$$

95% confidence interval for the summary effect can be calculated using the normal approximation

$$M \pm 1.96 \times SE_M^*$$

A test for the the null hypothesis that the mean effect μ is equal to 0 can be obtained as in the fixed-effect meta-analysis.

Hand calculations

```
    study     y    se     v     w    wy    y2   wy2    w*
1 Study A  94.8 1.403 1.967 0.508  48.2  8993  4572 0.221
2 Study B  98.7 1.304 1.701 0.588  58.0  9741  5727 0.234
3 Study C 101.1 1.489 2.218 0.451  45.6 10212  4603 0.209
4 Study D 100.0 1.349 1.820 0.550  55.0  9998  5494 0.228
5 Study E  98.2 0.715 0.511 1.955 191.9  9635 18841 0.325
```

$$\sum_{i=1}^{k} W_i Y_i^2 = 8993.249 \cdot 0.508 + 9740.95 \cdot 0.588 + 10211.553 \cdot 0.451 +$$

$$9997.518 \cdot 0.55 + 9635.232 \cdot 1.955 = 39237.17$$

$$\sum_{i=1}^{k} W_i Y_i = 94.833 \cdot 0.508 + 98.696 \cdot 0.588 + 101.052 \cdot 0.451 + 99.988 \cdot 0.55 + 98.159 \cdot 1.955 = 398.68$$

$\sum_{i=1}^{k} W_i = 4.05$ and $\sum_{i=1}^{k} W_i^2 = 4.93$

$$Q = 39237.17 - \frac{398.68^2}{4.05}$$

$$T^2 = \frac{11.272 - (5-1)}{4.05 - \frac{4.93}{4.05}} = 2.566$$

Distributional assumptions

$\epsilon_i \sim N(0, V_{Y_i})$ or $\theta_i \sim N(\theta, V_{Y_i})$

- the moment-based estimator of τ^2 does not require additional assumption about the distribution of ζ_i;
- the pooled effect M^* and its SE are also independent on a distributional assumption for the random-effects;
- confidence intervals relies on approximating distributions or central limit theorem ($k >> 0$).

$\zeta_i \sim N(0, \tau^2)$

- alternative estimator of τ^2;
- facilitates computation of confidence intervals;
- allows to give predictions for new studies.

Studies on the Effectiveness of the BCG Vaccine Against Tuberculosis (Colditz et al. (1994))

Meta-analysis of 13 studies examining the effectiveness of the Bacillus Calmette-Guerin (BCG) vaccine against tuberculosis.

Research question: What is the overall effectiveness of the BCG vaccine for preventing tuberculosis?

Additional question: Are there any moderator variables that may potentially influence the size of the effect?

Heterogeneity

The aim of a meta-analysis is not only to provide a summary measure, but also to make sense of the pattern of the observed results.

Additional important questions to address:

- Is there evidence of heterogeneity in true effect sizes?
- What is the variance of the true effects?
- What are the implications of the observed heterogeneity?
- What proportion of the observed dispersion can be attributed to differences among studies?
- Can we explain (part of) the observed heterogeneity?

Heterogeneity across studies

Differences between studies might be due to:

- design, and follow-up
- populations, participants, patients
- treatment or exposure, intervention
- outcome definition

Heterogeneity usually refers to variation in the *true* effect sizes.

We want to describe and quantify this variation.

Problem: the variation in the *observed* effect sizes is partly spurious.

It consists of both (true) heterogeneity and random error.

A set of alternative measures and statistics are available which tackle different aspect of the heterogeneity:

- Q statistic;
- results from the test based on Q;
- between-studies variance (T^2);
- between-studies standard deviation (T);
- ratio of true heterogeneity to total observed variation (I^2)

Disentangle between-studies variation from observed variation

The idea behind the Der Simonian and Laird estimator

1. Compute the total amount of variation observed across studies.
2. Estimate the expected variation under the fixed-effect model.
3. The excess variation (if any) is assumed to reflect differences in effect sizes (statistical heterogeneity)

The Q statistic

Q is a weighted sum of squares of the observed effect sizes, which quantifies the total amount of variation observed across studies.

$$Q = \sum_{i=1}^{k} W_i (Y_i - M)^2 = \sum_{i=1}^{k} \left(\frac{Y_i - M}{S_{Y_i}} \right)^2$$

where W_i and M are the weights and summary measure from a fixed-effect model.

Q is a standardized measure which is independent from the metric of the effect sizes.

Under the fixed-effect assumption, the expected value of Q is its degrees of freedom $E[Q] = \mathrm{df} = k - 1$

The excess variation is given by $Q - \mathrm{df}$ which will be attributed to differences in the true effects from study to study.

Note: the excess variation might be negative if $Q < \mathrm{df}$.

Q reflects the total dispersion while $Q - \mathrm{df}$ the excess variation. They are not easy to interpret: they are sums, and so they depends strongly on the number of studies.
Q is on a standardized scale. Some measures expressed as ratio or on the same scale as the effect size might be preferable.

Testing the assumption of homogeneity in effects

Researchers are often interested in testing if heterogenity is statistically significant.
The null hypothesis is that the underlying true effect size is the same for all studies.

The *p*-value of the heterogeneity test is calculated comparing the statistic Q with a Chi-Square distribution with degrees of freedom equal to $k-1$.

Usually, an $\alpha = 0.1$ is chosen as cut off for the *p*-value.

A statistically significant result means that there is evidence against there being one common effect size.
Failing to reject the null hypothesis does not imply that studies are homogeneous.

It is known that the test for heterogeneity

- has low power when only a few studies
- is possibly oversensitive when many studies
- is difficult to compare across meta-analyses

Note: both Q and the corresponding p-value do not estimate the magnitude of the true dispersion.

Estimating τ^2

τ^2 is the variance of the true effect sizes. It is in the same metric as the effect and reflects the absolute amount of variation.

Der Simonian and Laird proposed the estimator:
$T^2 = \max\left(0; \frac{Q-\mathrm{df}}{C}\right)$

As it might generate negative estimates, it is truncated to 0 in case $Q < \mathrm{df}$.
It doesn't require any assumption about the distribution of random-effects, it easy to compute and explain.

Alternative estimators exist. Many statisticians favor a restricted maximum likelihood method.

Tau

τ refers to the actual standard deviation of the true effect sizes. An estimator for τ is simply $T = \sqrt{T^2}$.

T in on the same scale of the effect size. It can be used to describe the distribution of effect sizes about the mean effect.

e.g. (Under the normality assumption), we expect 95% of the true effects will fall in the range $M^* \pm 1.96\,T$

T^2 and T are absolute measures. Sometimes, it can be easier to think about heterogeneity independent of the scale.

A measure of heterogeneity

Higgins et al (2002) proposed a statistic I^2 which quantifies the amount of total variability attributed to between study heterogeneity.

$$I^2 = \frac{Q - \mathrm{df}}{Q} \times 100\% = \frac{\mathrm{Variance}_{\mathrm{between}}}{\mathrm{Variance}_{\mathrm{total}}} \times 100\%$$

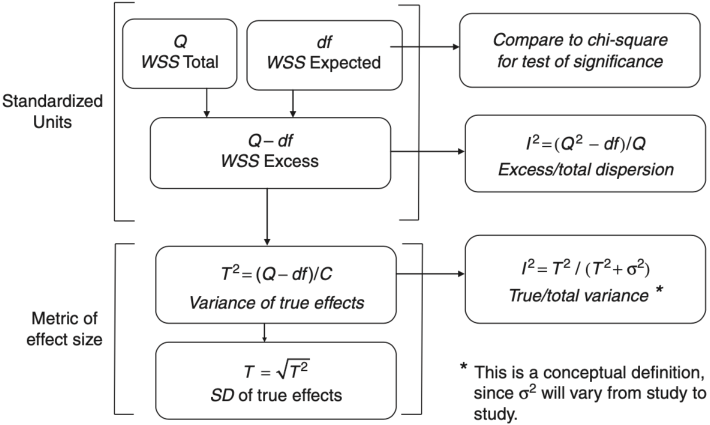
Advantages of I^2 are:

- intuitive interpretation, it is a percentage
- it is easy to calculate from previously published meta-analysis
- the interpretation is not dependent on the choice of measure of association

I^2 reflects the extent of overlap of confidence intervals. It is convenient to interpret as a measure of inconsistency, and not as a measure of the real variation.

Note: The I^2 statistic is a descriptive statistic and not an estimate of any underlying quantity.

Recap



Standardized Units

Q
WSS Total

df
WSS Expected

Compare to chi-square for test of significance

Q – *df*
WSS Excess

$I^2 = (Q^2 - df)/Q$
Excess/total dispersion

Metric of effect size

$T^2 = (Q - df)/C$
Variance of true effects

$I^2 = T^2 / (T^2 + \sigma^2)$
True/total variance *

$T = \sqrt{T^2}$
SD of true effects

* This is a conceptual definition, since σ^2 will vary from study to study.

Figure 16.3 Flowchart showing how T^2 and I^2 are derived from *Q* and *df*.

What to do in case of substantial heterogeneity?

If the studies are highly heterogeneous you need to determine how to proceed.

- Don't combine results
  - Do a qualitative systematic review (describe trials and evidence narratively)
- Combine with a random effects model
- Combine only a subset of initial studies (those that are similar)
- Try to explain heterogeneity:
  - subgroup analyses
  - meta-regression

Exploring Heterogeneity

- Exploring Heterogeneity
  - do separate meta-analyses on subgroups of studies according to study characteristics that may influence the pooled association (gender, study design, geographical area, year of publication);
  - compare the combined results with analogue to an ANOVA model.
- Meta-regression
  - similar to standard regression;
  - the pooled effect size is modeled as a linear function of one or more explanatory variables

Note: Both approaches may have low power, they require several studies.

Studies on the Effects of Diuretics in Pregnancy (Viechtbauer et al. (2007))

Meta-analysis on 9 studies examining the effects of diuretics in pregnancy on various outcomes, including the presence of any form of pre-eclampsia, perinatal death, stillbirth, and neonatal death.

Research questions: Are there any health-related conditions in the use of diuretics during pregnancy? How heterogeneous are the results?

Model choice

Fixed vs Random-effects model

The differences between the fixed and random-effects model are critical with rispects to

- assumption: one true effect vs. a distribution of true effects;
- goals of the analysis: combining estimates vs. summarizing a distribution;
- interpretation of the statistics: estimate of the mean vs mean of the mean estimates;
- computation (estimate of τ^2).

Differences in the computation

The variance of the combined effect is larger in a random-effects model.

It also includes τ^2, so the weights will be smaller, and similarly their sum $\left(\mathrm{Var}(M) = \frac{1}{\sum_{i=1}^{k} W_i}\right)$.

This is also true because it's an estimate of a different parameter (one mean vs mean of populations)

Differences in precision

Let us assume $V_{Y_i} = \sigma^2$ and $n_i = n \, \forall i$

$$V_M = \frac{1}{\sum_{i=1}^{k} W_i} = \frac{1}{\sum_{i=1}^{k} \frac{n_i}{\sigma^2}} = \frac{\sigma^2}{\sum_{i=1}^{k} n_i} = \frac{\sigma^2}{kn}$$

- standard deviation of the outcome (higher sd, less precise);
- sample size in the individual studies (n increases, more precise).

$$V_{M^*} = \frac{\sigma^2}{kn} + \frac{T^2}{k}$$

- standard deviation of the true effect sizes (higher T^2, less precise);
- number of studies (higher k, more precise).

Difference in weights' definition

The estimate of the combined effect is almost always different under the two models (unless $T^2 = 0$), because the weights are different.

$W_i = \frac{1}{V_{Y_i}}$: the information from smaller studies is minimal

$W_i^* = \frac{1}{V_{Y_i} + T^2}$: T^2 reduces the relative differences among the weights

In a random-effects model, large studies lose influence and small studies gain influence.

Which model should we choose?

Fixed-effect model

- there is good reason to believe that all the studies are functionally identical;

- the goal is to compute the common effect size (not be generalized beyond the observed population).

Random-effects model

- the studies have enough in common but are not identical;

- the goal of the analysis is usually to generalize to a range of scenarios.

Sensitivity analysis

Studies on Lung Cancer Risk from ETS Exposure (Hackshaw et al. (1997))

Meta-analysis on 37 studies (4 cohort, 33 case-control) reporting results on the risk of lung cancer from environmental tobacco smoke (ETS) exposure from the spouse in women who are lifelong nonsmoker.

Research questions: What is the accumulated evidence of evidence on lung cancer and environmental tobacco smoke? How robust are the combined results?

Publication Bias

Precision increases as study size increases.

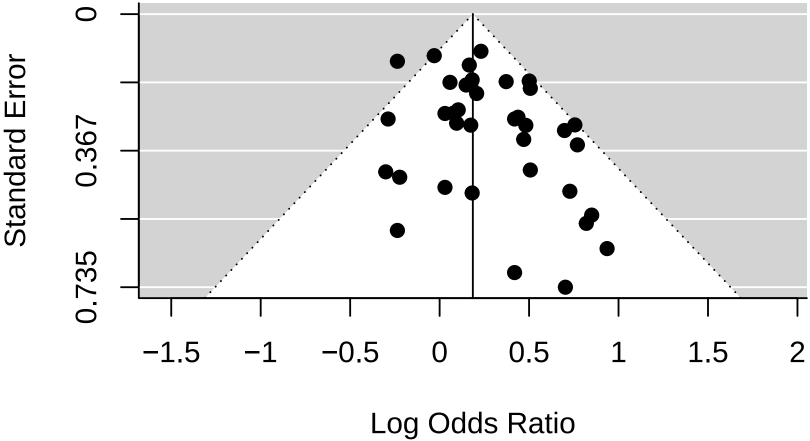
Results from small studies will be more scattered than larger studies.

In the absence of bias, a plot of precision vs. the estimate (funnel plot) will resemble a symmetrical inverted funnel.

If there is publication bias, because smaller studies showing no statistically significant effects remain unpublished, then this will lead to asymmetry in the funnel plot.

A pseudo-confidence region is usually represented within bounds $M \pm 1.96\mathrm{SE}$.

Funnel plot



Egger's test

Inspection of the funnel plot might be subjective. Egger (BMJ, 1997) and colleagues proposed a simple test to detect asymmetry in the funnel plot:

$$E[\text{standardized effect}] = a + b \times \text{precision}$$

First, standardizes the study-specific effect sizes by subtracting the pooled effect and dividing by standard error.

Second, fit a regression of standardized effect size against their precision (inverse of the standard error).

Third, tests whether intercept is significantly different from zero.

Funnel plot asymmetry may reflect something other than publication bias:

- there may be true heterogeneity leading to a different effect depending on study size;
- the intervention effect may differ between small and large studies.

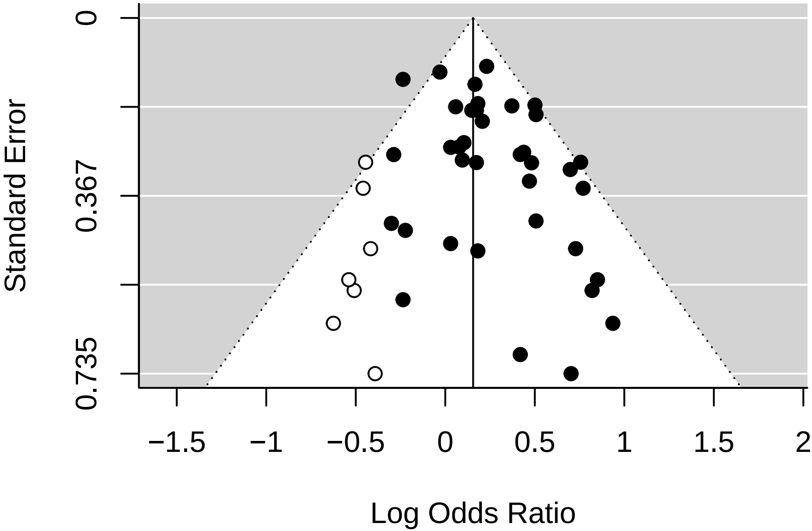
One can view funnel plots as displaying evidence for "small study effects" in general rather than publication bias in particular.

Trim-and-fill analysis

The trim and fill method is a nonparametric (rank-based) data augmentation technique proposed by Duval and Tweedie (2005).

It estimates the number of studies missing due to the suppression of the most extreme results on one side of the funnel plot.
The method augments the observed data so that the funnel plot is more symmetric.

Note: the method examines the sensitivity of the results to a particular form of publication bias. It does not provide a more "valid" summary estimate.


Standard Error
0
0.367
0.735
−1.5
−1
−0.5
0
0.5
1
1.5
2
Log Odds Ratio

Cumulative meta-analysis

A cumulative meta-analysis describes the accumulation of evidence.

It consists of performing a new meta-analysis as the available estimates are added to the analysis in (typically) chronological order.

It allows

- the study of trends in efficacy;
- to determine when a new treatment appears to be significantly effective or deleterious.