Introduction to Meta-Analysis

Every single study is just a data-point in a future meta-analysis. If you draw small samples from a population, the mean and standard deviation in the sample can differ considerably from the mean and standard deviation in the population. There is great variability in small samples. Parameter estimates from small samples are very imprecise, and therefore the 95% confidence intervals around effect sizes are very wide. Indeed, this led Cohen (1994) to write "I suspect that the main reason [confidence intervals] are not reported is that they are so embarrassingly large!" If we want a more precise estimate of our parameter of interest, such as the mean difference or correlation in the population, we need either run extremely large single studies, or alternatively, combine data from several studies by performing a **meta-analysis**. The most common approach to combine studies is to perform a meta-analysis of effect size estimates.

You can perform a meta-analysis for a set of studies in a single article you plan to publish (often called an **internal meta-analysis**), or you can search the literature for multiple studies reported in as many different articles as possible, and perform a meta-analysis on all studies others have published. An excellent introduction to meta-analyses is provided in the book by Borenstein, Hedges, Higgins, and Rothstein (2009). There is commercial software you can use to perform meta-analyses, but I highly recommend *against* using such software. Almost all commercial software packages lack transparency, and do not allow you to share your analysis code and data with other researchers. In this assignment, we will be using R to perform a meta-analysis of effect sizes, using the **metafor** package by Viechtbauer (2010). An important benefit of using metafor is that your meta-analysis can be made completely reproducible.

Single study meta-analysis

Open the file introduction_to_meta-analysis.R which contains the code needed to follow along and answer the questions. **Run line 1-4 and make sure the MBESS and metafor packages are installed.** Let's first begin with something you will hardly ever do in real life: a meta-analysis of a single study. This is a little silly, because a simple *t*-test or correlation will tell you the same thing – but it is educational to compare a *t*-test with a meta-analysis of a single study, before we look at how to combine multiple studies into a meta-analysis.

A difference between an independent *t*-test and a meta-analysis is that a *t*-test is performed on the raw data, while a meta-analysis is performed on the effect size(s) of individual studies. The `metafor` R package contains a very useful function called 'escalc' that can be used to calculate effect sizes, their variances, and confidence intervals around effect size estimates. So let's start by calculating the effect size the enter into our meta-analysis. The code below (and in the R file under **Part 1**) can be used to calculate the **standardized mean difference** (SMD) from two independent groups from **means** (specified by m1i and m2i), **standard deviations** (sd1i and sd2i) , and the number of observations in each group (n1i and n2i). By default, metafor calculates the effect size '**Hedges' g**' which is the unbiased version of Cohen's d (see week four of my previous MOOC, or Lakens, 2013, for a primer on effect sizes).

```
# We calculate the standardized mean difference
# We store it as the variable g (because by default, Hedges' g is computed)
g <- escalc(measure = "SMD",
            n1i = 50, #sample size in group 1 is 50
            m1i = 5.6, #observed mean in group 1 is 5.6
            sd1i = 1.2, #observed standard deviation in group 1 is 1.2
            n2i = 53, #sample size in group 2 is 50
            m2i = 4.9, #observed mean in group 1 is 4.9
            sd2i = 1.3) #observed standard deviation in group 2 is 1.3
```

The output gives you Hedge's g (under the yi column, which always returns the effect size, in this case the standardized mean difference) and the variance of the effect size estimate (under vi).

```
      yi          vi
1     0.5547      0.0404
```

As explained in Borenstein, Hedges, Higgins, and Rothstein (2009, formula 4.18 to 4.24) the standardized mean difference Hedges' g is calculated by dividing the difference between means by the pooled standard deviation, multiplied by a correction factor, J:

$$J = (1 - \ 3/(4df - 1))$$

$$g = J \times \left(\frac{\bar{X}_1 - \bar{X}_2}{S_{within}}\right)$$

and a very good approximation of the variance of the standardized mean difference (SMD)Hedges' g is provided by:

$$Vg = J^2 \times \left(\frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2)}\right)$$

The variance of the standardized mean difference depends only on the sample size (n_1 and n_2) and the value of the standardized mean difference itself. **To perform the**

required calculations for a meta-analysis, you need the effect sizes and their variance. This means that if you have coded the effect sizes and the sample sizes (per group) from studies in the literature, you have the information you need to perform a meta-analysis.

You do not need to manually calculate the effect size and its variance using the two formula above – the `escalc` function does this for you. We can now easily perform a single study meta-analysis using the `rma` function in the `metafor` package:

```
rma(yi, vi, data = g)
```

Which gives the output:

```
Fixed-Effects Model (k = 1)

Test for Heterogeneity:
Q(df = 0) = 0.0000, p-val = 1.0000

Model Results:
estimate      se    zval    pval   ci.lb   ci.ub
  0.5547  0.2009  2.7612  0.0058  0.1610  0.9485  **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here, we see under Model Results the effect size Hedges' g (0.55) and the standard error (which is just the square root of the variance we calculated above, √0.04), the Z-test statistic testing the mean difference against the null-hypothesis (2.76), and the 95% confidence interval [ci.lb = 0.16; ci.ub = 0.95] around the effect size (the interval width can be specified using the 'level =' option). We also get the *p*-value for the test of the meta-analytic effect size against 0. In this case we can reject the null-hypothesis ($p = 0.0058$).

In a meta-analysis, a Z-test is used to examine whether the null-hypothesis can be rejected. This assumes a normally distributed random effect size model. Normally, you would analyze data from a single study with two groups using a *t*-test, which not surprisingly uses a *t*-distribution. If we directly compare a single-study meta-analysis, based on a Z-test, with a normal *t*-test, we will see some tiny differences in the results. The *t*-value would be 2.83, and the *p*-value would be 0.0055. With large enough sample sizes (which is commonly true in a meta-analysis) the difference between a Z-test and *t*-test is not noticeable, and for this reason the Z-test is used in meta-analyses.

Simulating a single study meta-analysis

Now we are ready to simulate single studies. This code simulates two groups of 50 participants. Both groups perform an IQ test. By default, and IQ test yields a mean IQ score of 100, and a standard deviation of 15, in the general population. However, one group has followed an expensive year-long IQ training program. This IQ training program

has a proven track record of boosting the IQ score of participants by 6 IQ points, to a mean of 106. We simulate data from a control group and an intervention group using the `rnorm` function (lines 42 to 47) of 50 people each with the specified population parameters.

Note that in line 34 the `set.seed(1000)` command is used to make the simulation reproducible. As long as this command is executed before the lines below it are run, the output of the simulation will be the same as the result in this assignment. If you comment this line of code out (by placing a # in front of it, as in all the comments throughout the code explaining what the code does) the simulation results will vary each time you run the code.

Let's first analyze the simulated data using a *t*-test (line 50). We will calculate the effect size Hedges' g (line 53 to 59), and the 95% confidence interval around the effect size (line 62 to 65), using the widely used MBESS package (Kelley, 2007). The MBESS packages uses the *t*-distribution when calculating confidence intervals (so we can confirm this really makes only a tiny difference compared to using the Z-distribution in a meta-analysis with a sample size of 50 observations per group). Select and run the code up to line 65, and scroll up in the console window to see all the output, which should look like:

```
        Two Sample t-test

data:  x and y
t = 0.2508718, df = 98, p-value = 0.8024385
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -5.180770221  6.680207710
sample estimates:
  mean of x   mean of y
103.6205476 102.8708289

[1] 0.04978922624

$Lower.Conf.Limit.smd
[1] -0.3420089986

$Upper.Conf.Limit.smd
[1] 0.4421021361
```

The *t*-test results indicate that in the current sample the means of both groups are very similar (M_x = 103.6 and M_y = 102.9), with a 95% confidence around the mean difference from -5.18 to 6.68, and this difference is not statistically different from 0 (p = 0.802). The output from the SMD function (rounded to 3 digits) indicates the standardized mean difference is close to zero (g = 0.0498), and the 95% confidence intervals around the

standardized mean difference include zero [-0.342; 0.442]. Because we know the population parameters in the simulation, we are certain there is a true effect in the population, but in the current sample, we were not able to statistically detect it.

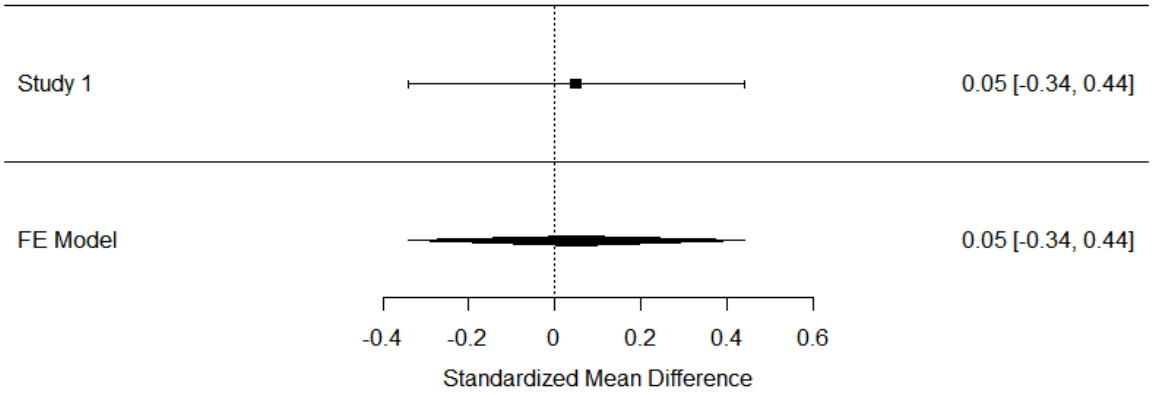
We can also analyze this data as a single study meta-analysis. In line 68-74 the `escalc` function is used to calculate Hedges' g. In line 76 a meta-analysis is performed with the `rma` function, which is stored as the variable 'result'. The results are printed in line 77. Line 78 plots a graphical summary of the meta-analysis, known as a **forest plot**. Run these lines (make sure the plot window in RStudio is large enough to see the forest plot)..

The Z-value, *p*-value, and 95% confidence interval in the output of the meta-analysis is very similar to the results we've seen above for the *t*-test:

```
Fixed-Effects Model (k = 1)
Test for Heterogeneity:
Q(df = 0) = 0.0000, p-val = 1.0000

Model Results:
Estimate    se          zval        pval        ci.lb       ci.ub
0.0498      0.2000      0.2489      0.8034      -0.3423     0.4418
```

Under 'Model Results' we find Hedges' g is 0.0498, and the confidence interval around the effect size (ci.lb and ci.ub for the lower bound and upper bound) is similar (although not identical, to that calculated for the *t*-test, [-0.34; 0.44], and the *p*-value is quite comparable as well ($p = 0.803$). We can't reject the null hypothesis based on this single study.

The forest plot is presented below. We see the effect size for Study 1 marked by the black square at 0.05, and the confidence interval is visualized by lines extending to – 0.34 on the left and 0.44 on the right. The numbers are printed on the right-hand side of the forest plot. On the lower half of the forest plot, we see a stretched-out diamond ♦. The diamond summarizes the meta-analytic effect size estimate, with the center being at the meta-analytic effect size estimate, and the left and right endpoints at the 95% confidence interval of the meta-analytic effect size estimate. Because we only have a single study, the meta-analytic effect size estimate is the same as the effect size estimate for our single study.



Meta-analyses get a bit more exciting when we are using them to analyze results from multiple studies. When multiple studies are combined in a meta-analysis, effect size estimates are not simply averaged, but they are **weighted** by the **precision** of the effect size estimate, which is determined by the sample size of the study. Thus, the larger the sample size of an individual study, the more weight it gets in the meta-analysis, meaning that it has more influence on the meta-analytic effect size estimate.

Simulating meta-analyses of mean standardized differences

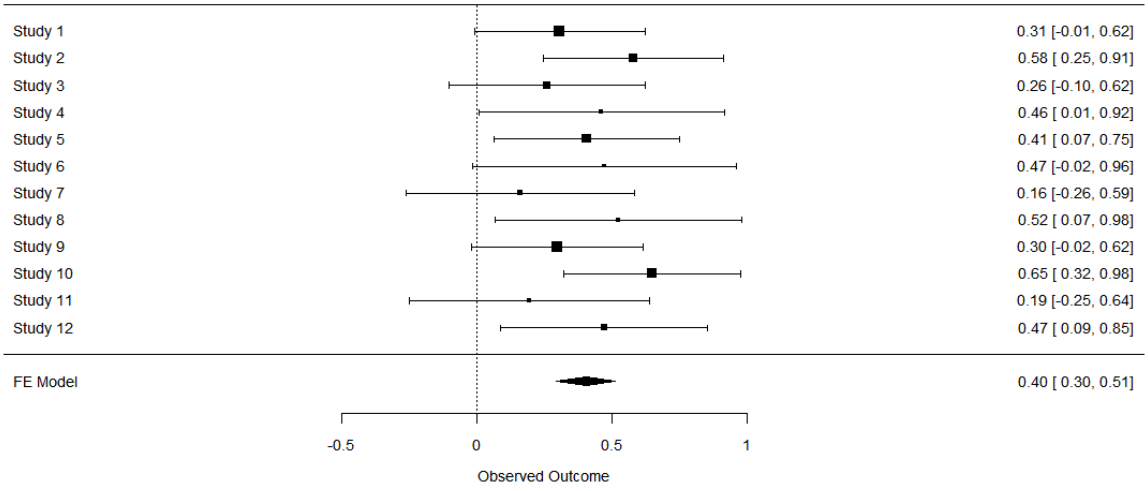
Open the file simulating_meta_analyses_smd.R. In this file we again simulate data for two groups, one with a mean of 100, and one with a mean of 106, both groups with a standard deviation of 15 (lines 7 to 10). To perform a meta-analysis, we need to calculate the effect size for each study, and the variance of the effect size. An empty dataframe with the column names yi and vi is created in line 11. We then simulate studies (the number of simulated studies is 12, which you can change in line 5).

Both groups of participants have the same size in the simulation. To simulate studies with some variation in the sample sizes, the sample size in each group (n) is determined by randomly drawing a value between 30 and 80 (line 14). If in your field sample sizes are much larger of smaller, feel free to change these numbers to something more realistic in your field. Two normally distributed samples are simulated (lines 15 and 16), and the `escalc` function is used to calculate and store the effect size (yi) and variance (vi) for each simulated study in a dataframe called `metadata`. We can then perform the meta-analysis, print the results, and plot the forest plot (lines 21 to 23).

Run all lines of code. Note that we are simulating studies, without setting a seed – so your results will differ from the results described below. but should be somewhat similar.

```
Fixed-Effects Model (k = 12)

Test for Heterogeneity:
Q(df = 11) = 7.2803, p-val = 0.7759

Model Results:

estimate      se    zval    pval   ci.lb   ci.ub
  0.4044  0.0551  7.3383  <.0001  0.2964  0.5124  ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We see a test for **heterogeneity**. We will return to heterogeneity tests later. We see the model results, which in this specific simulation yielded a meta-analytic effect size estimate of 0.40 (your result could be smaller or larger, but should be close – feel free to run all lines of code multiple times to explore the amount of variation you can expect). The confidence interval around the effect size estimate [0.30; 0.51] is much narrower than we saw before for a single study. This is because the 12 studies we simulated together have quite a large sample size, and the larger the sample size, the smaller the standard error, and thus the narrower the confidence interval is. The meta-analytic effect size estimate is statistically different from 0 ($p < 0.0001$) so we can reject the null hypothesis if we use an alpha level of 0.05. The forest plot provides a more detailed overview of the individual studies.



We see 12 rows, one for each study, each with their own effect size and confidence interval. If you look closely, you can see the squares that indicate the effect size estimate for each study differ in size. The larger the sample size, the bigger the square. Study 4 had a relatively small sample size, which can be seen both by the small square, and the relatively wide confidence interval. Study 1 had a larger sample size, and thus a slightly larger square and narrower confidence interval. At the bottom of the graph we find the meta-analytic effect size and its confidence interval, both visualized by a diamond and numerically. The model is referred to as a FE Model, or **Fixed Effect (FE) model**. The alternative approach is a RE Model, or **Random Effects (RE) model**.

Fixed Effect vs Random Effects

There are two possible models when performing a meta-analysis. One model, known as a fixed effect model, assumes there is one effect size that generates the data in all studies in the meta-analysis. This model assumes there is no variation between individual studies – all have exactly the same true effect size. The perfect example of this is the simulations we have done so far. We specified a single true effect in the population, and generated random samples from this population effect.

Alternatively, one can use a model where the true effect differs in some way in each individual study. We don't have a single true effect in the population, but a range of **randomly distributed** true effect sizes (hence the 'random effects' model). Studies differs in some way from each other (or some sets of studies differ from other sets), and their true effect sizes differ as well. Note the difference between a fixed effect model, and a random effect**s** model, in that the plural 'effects' is used only in the latter. Borenstein et al (2009) state there are two reasons to use a fixed effect model: When all studies are functionally equivalent, and when your goal is *not* to generalize to other populations. This makes the random effects model generally the better choice, although some people have raised the concern that random-effects models give more weight to smaller studies, which can be more biased. By default, metafor will use a random effects model. We used the `method="FE"` command to explicitly ask for a fixed effect model. In the meta-analyses we will simulate in the rest of this assignment we will leave out this command and simulate random effects meta-analyses.

Q1: Run the code 20 times. Compare the effect size estimate for Study 1 (the first simulated study in the meta-analysis) with the meta-analytic effect size estimate. Which statement is true?

A) The effect size estimate for Study 1 is **less** variable, and **less** often statistically significant, than the meta-analytic effect size estimate.

B) The effect size estimate for Study 1 is **less** variable, and **more** often statistically significant, than the meta-analytic effect size estimate.

C) The effect size estimate for Study 1 is **more** variable, and **less** often statistically significant, than the meta-analytic effect size estimate.

D) The effect size estimate for Study 1 is **more** variable, and **more** often statistically significant, than the meta-analytic effect size estimate.

Simulating meta-analyses for dichotomous outcomes

Although meta-analyses on mean differences are very common, a meta-analysis can be performed on many different effect sizes. To show a slightly less common example, let's simulate a meta-analysis based on odds ratios. Sometimes the main outcome in an experiment is a dichotomous variable, such as the success or failure on a task. In such study designs we can calculate risk ratios, odds ratios, or risk differences as the effect size measure. Risk differences are sometimes judged easiest to interpret, but odds ratios are most often used for a meta-analysis because they have attractive statistical properties. An **odds ratio** is a ratio of two odds. To illustrate how an odds ratio is calculated, it is useful to consider the four possible outcomes in a 2 x 2 table of outcomes:

	Success	Failure	N
Experimental	*A*	*B*	*n1*
Control	*C*	*D*	*n2*

The odds ratio is calculated as: $OR = \frac{AD}{BC}$.

The meta-analysis is performed on log transformed odds ratios (because log transformed odds ratios are symmetric around 1, see Borenstein et al., 2009), and thus the log of the odds ratio is used, which has a variance which is approximated by: $\mathrm{Var}(\log OR) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$.

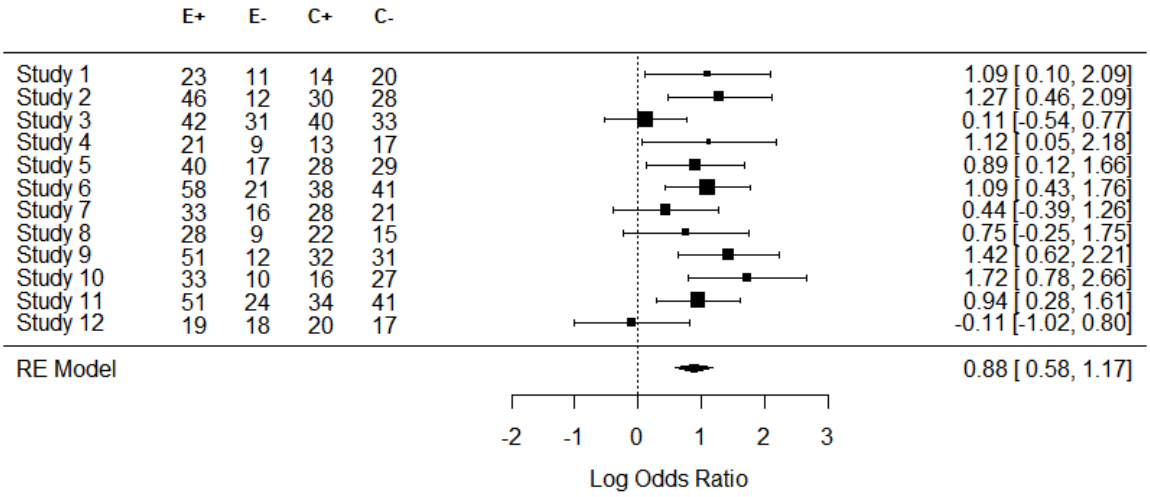
Let's assume that we train students in using a spaced learning strategy (they work through a textbook every week instead of cramming the week before the exam). Without such training, 70 out of 100 students succeed in passing the course after the first exam, but with this training, 80 out of 100 students pass.

	Success	Failure	N
Experimental	80	20	100
Control	70	30	100

The odds of passing in the experimental group is 80/20, or 4, while odds in the control condition are 70/30, or 2.333. The ratio of these two odds is then: 4/2.333 = 1.714, or:

$$OR = \frac{80 \times 30}{20 \times 70} = 1.714$$

Open the file simulating_meta_analyses_or.R. This script simulates studies with dichotomous outcomes, where you can set the percentage of successes and failures in the experimental and control condition (lines 7 and 8). In the script, by default the percentage of success in the experimental condition is 70%, and in the control condition it is 50%. Run the script.



The forest plot presents the studies and four columns of data after the study label, which contain the number of successes and failures in the experimental groups (E+ and E-), and the number of successes and failures in the control group (C+ and C-). Imagine we study the percentage of people who get a job within 6 months after a job training program, compared to a control condition. In Study 1, which had 34 participants in each condition, 23 people in the job training condition got a job within 6 months, and 11 did not get a job. In the control condition, 14 people got a job, but 20 did not. The effect size estimate for the random effects model is 0.88. Feel free to adjust the number of studies, or the sample sizes in each study, to examine the effect it has on the meta-analytic effect size estimate (in general, the more data, the more accurate the estimate).

Heterogeneity

Although researchers often primarily use meta-analysis to compute a meta-analytic effect size estimate, and test whether this effect is statistically different from zero, **an arguably much more important use of meta-analyses is to explain variation between (sets of) studies**. This variation among (sets of) studies is referred to as **heterogeneity**. Tests have been developed to examine whether the studies included in a meta-analysis vary more than would be expected if the underlying true effect size in all studies was the same, and measures have been developed to quantify this variation.

If all studies have the same true population effect size, the only source of variation is random error. If there are real differences between (sets of) studies, there are two sources of variation, namely random variation from study to study, *and* real differences in effect sizes in (sets of) studies.

A classical measure of heterogeneity is Cochran's Q statistic, which is the weighted sum of the squared differences between effect size estimates in each study, and the meta-analytic effect size estimate. The Q statistic can be used to test whether the absence of heterogeneity can be statistically rejected (by comparing it to the expected amount of variation, which is the degrees of freedom, *df*, or the number of studies -1, see Borenstein et al., 2009), but it can have low power if the number of studies in the meta-analysis is small (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

On theoretical grounds one might argue that some heterogeneity will always happen in a meta-analysis, and therefore it is more interesting to quantify the extent to which there is heterogeneity. The I^2 index measures the extent of true heterogeneity. It is calculated as follows: $I^2 = \frac{(Q-k-1)}{Q} \times 100\%$, where the k is the number of studies (and k-1 is the degrees of freedom). I^2 ranges from 0 to 100 and can be interpreted as the percentage of the total variability in a set of effect sizes that is due to heterogeneity. When $I^2 = 0$ all variability in the effect size estimates can be explained by within-study error, and when $I^2 = 50$ half of the total variability can be explained by true heterogeneity. I^2 values of 25%, 50%, and 75% can be interpreted as low, medium, and high heterogeneity.

Open the file heterogeneity.R. This script simulates a similar meta-analysis to the example for dichotomous outcomes above, but with a small variation. The first half of the simulated experiments are based on the population success rates specified in lines 7 and 8, but the second half of the simulated experiments are based on the population success rates specified in lines 25 and 26. Thus, in this set of studies, the odds ratio differs for the first half of the study , compared to the second half (successes in Group 1

and 2 are set to 0.2 and 0.7 for the first half, but to 0.7 and 0.9 in the second half). There is true heterogeneity. In line 46 we use the 'confint' function in the metafor package to report both the I^2 statistic, and its confidence interval.

Run the script 20 times, each time noting the I^2 statistic and the 95% confidence interval around it. You will see widely varying estimates for I^2. Because we are simulating only 12 studies, estimates of heterogeneity have large variability. Increase the number of simulated experiments in line 5 from 12 to 200.

Q2: Re-run the simulation 20 times. Compare the I^2 value and its 95% CI when simulating meta-analyses of 12 and 200 studies. Which statement is true?

A) The I^2 values are **more** variable when simulating 200 studies, and the 95% CI is **wider**.
B) The I^2 values are **more** variable when simulating 200 studies, and the 95% CI is **narrower**.
C) The I^2 values are **less** variable when simulating 200 studies, and the 95% CI is **wider**.
D) The I^2 values are **less** variable when simulating 200 studies, and the 95% CI is **narrower**.

In this assignment, you've learned the basics of combining studies in a meta-analysis.

You can use meta-analyses to combine results in a single paper, or to evaluate effects in the literature. When combining effects in the literature, one of the main goals of a meta-analysis is to test or develop theories that can explain the variability in effect size estimates. Performing meta-analyses is relatively easy whenever you have access to the effect sizes (or the necessary information to compute effect sizes) and the sample sizes for each study. By combining information, you can achieve a much higher precision in effect size estimates. Excellent introductions to performing and interpreting meta-analyses can be found in Borenstein et al (2009) and Cummings (2013).

References

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). Introduction to meta-analysis. Chichester, United Kingdom: John Wiley & Sons, Ltd.

Cumming, G. (2013). Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. New York, NY: Routledge.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? Psychological Methods, 11(2), 193.

Kelley, K. (2007). Methods for the behavioral, educational, and social sciences: An R package. Behavior Research Methods, 39(4), 979–984.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. Journal of Statistical Software, 36(3), 1–48.