Methodology of Meta-Analysis

Elena Kulinskaya

School of Computing Sciences,
University of East Anglia

November 2020

Software and Resources

Review Manager (RevMan is the software used for preparing and maintaining Cochrane reviews https://training.cochrane.org/online-learning/core-software-cochrane-reviews/revman

OpenMEE is open-source, cross-platform software for ecological and evolutionary meta-analysis http://www.cebm.brown.edu/openmee/

R libraries

meta is a meta analysis package by Guido Schwarzer.

metafor is a meta analysis package by Wolfgang Viechtbauer, includes meta regression.

mvmeta is a package by Antonio Gasparrini, includes multivariate and univariate meta-analysis and meta-regression.

Weighting the evidence

In Session 1 we have dealt with various effect measures. They are a "common currency" to compare results of different studies using different measures. At the start of an meta-analysis, we convert effects reported in different studies into a single comparable effect measure.

An important feature of a meta-data is that every study has a different sampling error, additionally there may be between-study variation which needs taking into account. Standard statistical methods such as ANOVA or regression, cannot be used because the homogeneity of variances assumption is violated.

Therefore weighted statistical methods are traditionally used in meta-analysis. These methods weigh the studies according to their precision.

Notation and assumptions

Consider an experiment with two arms: Experiment and Control, of a total sample size N. Meta Analysis (MA) uses the following info:

Sample sizes:	$n_C = qN, \quad n_E = (1-q)N, \quad$ *where* q=n_C/N
Effect of interest:	$\theta, \quad$ say $\theta = (\mu_E - \mu_C)/\sigma$
Estimated by:	$\hat{\theta}, \quad$ say $\hat{\theta} = (\bar{X}_E - \bar{X}_C)/S$
Assumptions:	$\hat{\theta}_i$ is normally distributed $N(\theta_i, v_i)$, Studies are independent!
Variance of $\hat{\theta}$	v estimated by $\hat{v}, \quad$ typically $v = v_0/N$
Weights for θs:	$w = 1/v$
Weights used in MA:	$\hat{w} = 1/\hat{v}$
Number of studies in MA:	K

Fixed effects model

Given estimated effects from K studies $\hat{\theta}_1, \cdots, \hat{\theta}_K$, with $\hat{\theta}_i \sim N(\theta_i, v_i)$,

Given the homogeneity of effects: $\theta_1 = \theta_2 = \cdots = \theta_K = \theta$

the combined effect θ is estimated as the weighted mean
$\hat{\theta} = (w_1\hat{\theta}_1 + \cdots + w_K\hat{\theta}_K)/W$, where weights $w_i = 1/v_i$ and $W = \sum w_i$.

Thus, the studies are weighted according to their precision. The smaller the variance, the larger the weight.

Estimating the weights $\hat{w}_i = 1/\hat{v}_i^2$

The combined effect θ is estimated as the weighted mean $\hat{\theta} = (w_1\hat{\theta}_1 + \cdots + w_K\hat{\theta}_K)/W$, where weights $w_i = 1/v_i$ and $W = \sum w_i$.

Difficulty: Need to use estimated weights $\hat{w}_i$ instead of constants $w_i = 1/v_i$.

Generally, we need to estimate the variance of an effect measure, $\hat{v}$, and then calculate weights $\hat{w} = 1/\hat{v}$.

Standard MA treats these estimated weights as if they were known. This may result in spurious conclusions. These concerns will be addressed later.

Standard inference in fixed effects model

Combined effect $\hat{\theta} \sim N(\theta, 1/W)$, so the nominal confidence limits are $\hat{\theta} \pm W^{-1/2} z_{1-\alpha/2}$. Estimated $\hat{W}$ is substituted in practice.

Wald test: To test $H_0 : \hat{\theta} = 0$ vs alternative $H_1 : \hat{\theta} \neq 0$, compare $|W^{-1/2}\hat{\theta}|$ to critical value $z_{\alpha/2}$ from $N(0, 1)$ distribution

Cochran's Q test for heterogeneity:

$H_0 : \theta_1 = \theta_2 = \cdots = \theta_K$ vs alternative H_1: some θ's differ,

$$Q = \sum w_i(\hat{\theta}_i - \hat{\theta})^2 \sim \chi^2_{K-1}, \text{ where } K \text{ is the number of studies.}$$

If rejected, **random effects model** can be used

Software

metafor uses the **escalc** function to calculate estimated effect measures commonly used in meta-analyses and their variances, and then the **rma** function is used for meta-analysis and meta-regression.

There are two different interfaces for using the **escalc**, a default and a formula interface. For the default interface, the arguments of the function are

escalc(measure, ai, bi, ci, di, n1i, n2i, m1i, m2i, sd1i, sd2i, xi, mi, ri, ni, data = NULL, add = 1/2, to = "only0", vtype = "LS", append = FALSE) where measure is a character string specifying which outcome measure should be calculated.

Effect measures calculated by the **escalc** function

"RR": The log relative risk
"OR": The log odds ratio
"RD": The risk difference
"MD": The raw mean difference
"SMD": The standardized mean difference
"ROM": The log ratio of means
"ZCOR" z-transformed correlation coefficient

and others.

Dataset 4: The impact of reed management on wildlife, Valkama et al. (2008)

The authors found that reed management modifies the structure of re-growing reed stands: reed stems were shorter and denser in managed sites than in unmanaged sites.

1.4.1 Enter the data and print the names of the variables. Use procedure escalc in metafor to calculate the log Response Ratio and its variance.

Syntax:

```
valkama<-read.csv("C:/meta course/Valkama2008.csv")
names(valkama)
dat <- escalc(measure="ROM", m1i=Mean_treat, sd1i=SD_treat,
 n1i=N_treat, m2i=Mean_control, sd2i=SD_control, n2i=N_control,
 data=valkama, append=TRUE)
names(dat)
```

names(dat)

```
 [1] "study"                          "Author"
 [3] "Year"                           "Journal"
 [5] "Country"                        "Vegetation"
 [7] "Phragmites.dominated"           "habitat"
 [9] "salinity"                       "Management.1"
[11] "Management.2"                   "M..frequency.year"
[13] "M..freuency.period"             "Duration.of.the.follow.up.period..years"
[15] "Time.of.managemant"             "Table.Figure"
[17] "Species.1"                      "common.name"
[19] "Order"                          "Family"
[21] "Species.2"                      "Variable.1"
[23] "Variable.2"                     "Mean_control"
[25] "SD_control"                     "N_control"
[27] "Mean_treat"                     "SD_treat"
[29] "N_treat"                        "yi"
[31] "vi"
```

Meta analysis of the reed data

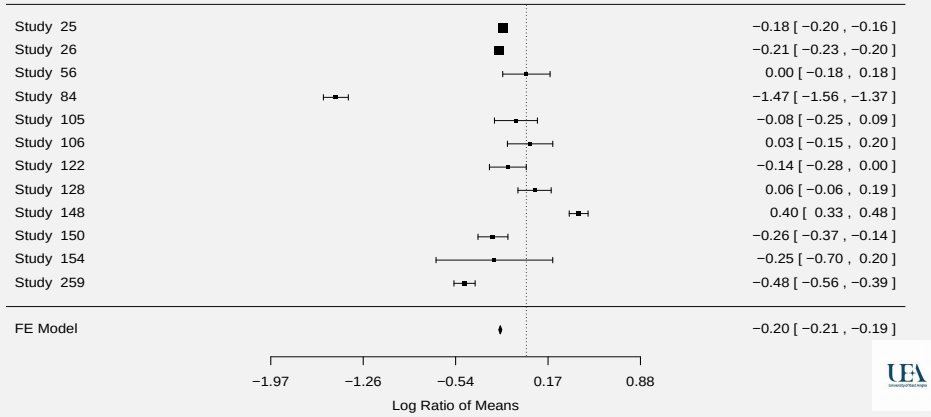
1.4.2 Perform fixed effects meta-analysis of the stem height of re-growing reed. Use variables Species.2 and Variable.2 to select reed as the species and stem height as the response. Which model (random or fixed effects) should be used? Produce the forest plot. What are your conclusions?

```
valkama_r1<-rma(yi,vi, data=dat,
subset=(Variable.2=="stem height")
         &(Species.2=="reed"),method="FE")
forest(valkama_r1)
summary(valkama_r1)
```

Output for the reed data, FEM

```
> summary(valkama_r1)

 Fixed-Effects Model (k = 12)

   logLik   Deviance        AIC        BIC
-483.3958   966.7916   968.7916   969.2765

 Test for Heterogeneity:
Q(df = 11) = 1016.4447, p-val < .0001

 Model Results:
estimate       se     zval     pval    ci.lb    ci.ub
 -0.2006   0.0060 -33.6250   <.0001  -0.2123  -0.1889      ***
```

Forest plot for the reed data



Some comments on the reed data analysis

- Forest plot depicts effects and their confidence intervals for $\hat{\theta}_i$.
- Heterogeneity is significant. Thus we should repeat the analysis using the random effects model.
- There is one far outlier: study 84. This alone may explain heterogeneity.
- As a sensitivity analysis, try to take this study out and redo the analysis. Also investigate the reasons for this extremely negative effect.
- The overall effect is significantly negative, confirming the conclusion that the reed height is decreased in managed sites.

More on heterogeneity

Cochran's Q test is not very powerful (works well with study sizes from 80 or K from 40). It would be good to quantify heterogeneity.

"...since clinical and methodological diversity always occur in a meta-analysis, statistical heterogeneity is inevitable (Higgins 2003). Thus the test for heterogeneity is irrelevant to the choice of analysis... move the focus away from testing whether heterogeneity is present to assessing its impact on the meta-analysis (Cochrane handbook)

Quantifying heterogeneity

Inconsistency index $I^2 = \frac{Q-(K-1)}{Q} \times 100\%$ (Higgins&Thompson, 2002)

"measures the extent of inconsistency among the studies' results, and is interpreted as approximately the proportion of total variation in study estimates that is due to heterogeneity rather than sampling error." (RevMan User Guide)

A rough guide to interpretation of I^2 is as follows:

0% to 40%: might not be important;

30% to 60%: may represent moderate heterogeneity;

50% to 90%: may represent substantial heterogeneity;

75% to 100%: considerable heterogeneity.

(Cochrane Handbook)

Drawback

When the average study size $\bar{n}$ increases, $I^2 \to 100\%$ regardless of true heterogeneity (Rücker et al, 2008)

Output for the reed data, REM

```
> valkama_r1<-rma(yi,vi, data=dat, subset=(Variable.2=="stem height")
> &(Species.2=="reed"),method="REML")
> summary(valkama_r1)
Random-Effects Model (k = 12; tau^2 estimator: REML)
  logLik  Deviance       AIC       BIC
 -8.2109   16.4217   20.4217   21.2175

tau^2 (estimate of total amount of heterogeneity): 0.2049 (SE = 0.0903)
tau (sqrt of the estimate of total heterogeneity): 0.4527
I^2 (% of total variability due to heterogeneity): 99.68%
H^2 (total variability / sampling variability):    311.90
 Test for Heterogeneity:
Q(df = 11) = 1016.4447, p-val < .0001
Model Results:
estimate       se     zval     pval    ci.lb    ci.ub
 -0.2146   0.1329  -1.6145   0.1064  -0.4752   0.0459
```

Inference on $\hat{\theta}$ in FEM

z-transformed correlation coefficient is the 'safest' setting for the standard meta-analysis. All assumptions are satisfied, and therefore inference works.

The first violation of assumptions for all other measures is treating the weights $\hat{w}_i$ as if they are constant. Additionally, for Hedges' d, for response ratio and for all measures based on binary data, the effects and the weights are not independent.

Li, Shi&Ross (1994) demonstrated that $\hat{W}^{-1}$ underestimates the variance of the combined effect in a general setting.

This should lead to considerable inflation of type I error of the Wald test, and too narrow CIs for the combined effect when n is small. For large enough sample sizes, everything works.

Permutation test

Permutation test is a good way not to make wrong assumptions. For models without moderators, the permutation test is carried out by permuting the signs of the observed effect sizes or outcomes.

The (two-sided) p-value of the permutation test is then equal to two times the proportion of times that the test statistic under the permuted data is as extreme or more extreme than under the actually observed data.

There are 2^K permutations for K studies, so the smallest possible (two-sided) p-value is $1/2^{K-1}$. This is .0625 when k=5 and .03125 when k=6. Therefore, k must be at least equal to 6 to reject the null hypothesis at $\alpha = 0.05$.

Syntax for permutation test

The function permutest carries out permutation tests for objects of class "rma.uni".

Usage

permutest(x, exact=FALSE, iter=1000, progbar=TRUE, retpermdist=FALSE, digits=x*digits*, *permci* = *FALSE*, ...)

Permutation test for reed height analysis

```
> permutest(valkama_r1)
Running 1000 iterations for approximate permutation test.
Model Results:
         estimate      se      zval   pval*    ci.lb    ci.ub
intrcpt   -0.2006  0.0060  -33.6250  0.0780  -0.2123  -0.1889   .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> permutest(valkama_r1, iter=10000)
Running 4096 iterations for exact permutation test.
Model Results:
         estimate      se      zval   pval*    ci.lb    ci.ub
intrcpt   -0.2006  0.0060  -33.6250  0.0908  -0.2123  -0.1889   .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Inference on heterogeneity

Cochran's Q test for heterogeneity behaves differently for different effect measures.

It is very liberal for the MD and very conservative for SMD in the case of small sample sizes, especially for large K.

In the balanced case, type I error is close to nominal level for $n \geq 80$ (Viechtbauer, 2007).

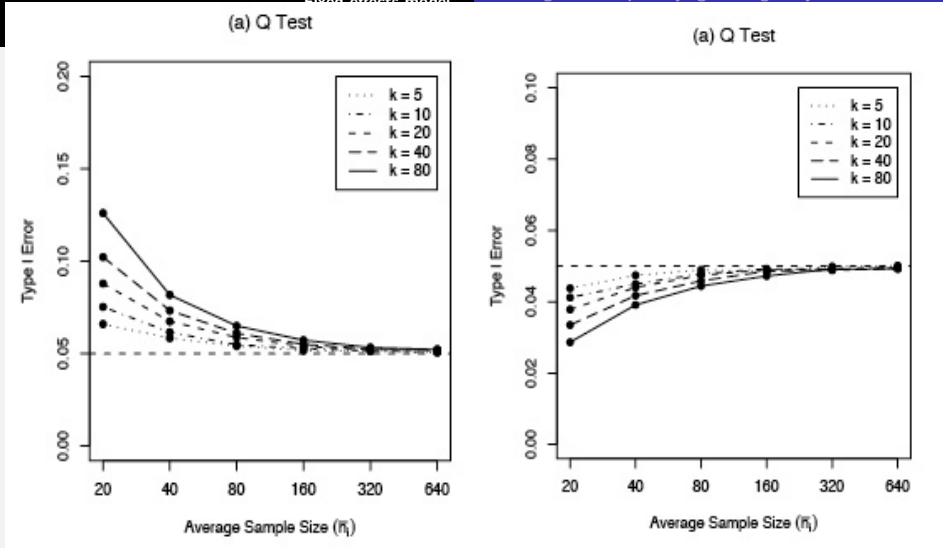


Figure: *Type I error of Cochran's Q for MD and SMD (Viechtbauer (2007), BJMSP, 60, 29-60)*

Random effects model (REM)

In the REM, estimated effects $\hat{\theta}_i \sim N(\theta_i, v_i)$ *conditionally on* θ_i, and the parameters $\theta_i \sim N(\theta, \tau^2)$.

Then, *unconditionally*, $\hat{\theta}_i \sim N(\theta_i, v_i + \tau^2)$.

Parameter τ^2 is the between-study variance component

The REM weights are $w_i^* = (v_i + \tau^2)^{-1}$, more homogeneous than the FEM weights w_i.
The combined effect θ is estimated as the weighted mean $\hat{\theta} = (w_1^*\hat{\theta}_1 + + w_K^*\hat{\theta}_K)/W^*$, where $W^* = \sum w_i^*$.

Difficulty: Need to estimate τ^2!

The conceptual difference between fixed and random effects models

In FEM, analysis is concerned with the studies already done

In REM, the studies in MA are a random sample from a hypothetical population of all studies; analysis aims to generalise the conclusions to this population.

FEM: was there an effect on average in the studies at hand?

REM: will there be an effect 'on average"? (Baily, 1987, *Stat Med*, **6**, 351-358)

The range of views by statisticians and different Cochrane groups varies from using exclusively FEM to choice based on heterogeneity to using exclusively REM.

Standard inference on $\hat{\theta}$ in random effects model

Combined effect $\hat{\theta} \sim N(\theta, 1/W^*)$, where W^* is the sum of weights w_i^*

The confidence limits are $\hat{\theta} \pm (W^*)^{-1/2} z_{1-\alpha/2}$

Wald test: To test $H_0 : \hat{\theta} = 0$ vs alternative $H_1 : \hat{\theta} \neq 0$, compare $|(W^*)^{-1/2}\hat{\theta}|$ to critical value $z_{\alpha/2}$ from $N(0,1)$ distribution

Since $w_i^* < w_i$, the variance of $\hat{\theta}$ under the REM is larger: $1/W^* < 1/W$. Consequently, REM confidence intervals are usually wider, and the power of the Wald test is lower than under FEM.

Cochran's Q test for heterogeneity: $H_0 : \tau^2 = 0$ vs alternative $H_1 : \tau^2 > 0$

The Q statistic and its null distribution are the same as in FEM! Power differs.

Estimating the variance component τ^2

The standard approach of DerSimonian and Laird (1986):

under heterogeneity ($\tau^2 > 0$),

$\mathrm{E}(Q) = (K-1) + (S_1 - \frac{S_1}{S_2})\tau^2$, where $S_r = \sum w_i^r$, so

$$\hat{\tau}^2 = \max\Big(0, \frac{Q-(K-1)}{S_1 - \frac{S_1}{S_2}}\Big)$$

When $\hat{\tau}^2 = 0$, inference about the combined effect $\hat{\theta}$ is the same as under FEM

Other approaches to estimation of $\hat{\tau}^2$ include ML, REML and other estimators, see Viechtbauer(2005),(2007) for comparison. REML is considered a gold standard, but for SMD and log R, PM may be better.

Output for the reed data, REM

```
> valkama_r1<-rma(yi,vi, data=dat, subset=(Variable.2=="stem height") &
    (Species.2=="reed"),method="REML")
> summary(valkama_r1)
Random-Effects Model (k = 12; tau^2 estimator: REML)
  logLik  Deviance       AIC       BIC
 -8.2109   16.4217   20.4217   21.2175

tau^2 (estimate of total amount of heterogeneity): 0.2049 (SE = 0.0903)
tau (sqrt of the estimate of total heterogeneity): 0.4527
I^2 (% of total variability due to heterogeneity): 99.68%
H^2 (total variability / sampling variability):    311.90
 Test for Heterogeneity:
Q(df = 11) = 1016.4447, p-val < .0001
Model Results:
estimate       se     zval     pval    ci.lb    ci.ub
 -0.2146   0.1329  -1.6145   0.1064  -0.4752   0.0459
---
```


means, variances and sample sizes for the reed data

```
> cbind(round(dat1$yi,3),round(dat1$vi,5),
    dat1$N_control,dat1$N_treat)
       [,1]    [,2] [,3] [,4]
 [1,] -0.179 0.00008  100  100
 [2,] -0.213 0.00008  100  100
 [3,]  0.001 0.00867    7    2
 [4,] -1.468 0.00241  200  200
 [5,] -0.080 0.00713    6    6
 [6,]  0.029 0.00795    6    6
 [7,] -0.142 0.00525   13   13
 [8,]  0.064 0.00427   25   25
 [9,]  0.404 0.00138   25   25
[10,] -0.257 0.00346   25   25
[11,] -0.246 0.05267    4    4
[12,] -0.476 0.00173  190  190
```

Weights for the reed data, FEM vs REM

Compare the weights between the 2 methods (FEM vs. REM): REM makes all weights relatively equal (homogeneous).

```
> (1/(valkama_r1$vi+valkama_r1$tau^2))/
          sum(1/(valkama_r1$vi+valkama_r1$tau^2))

 [1] 0.09453007 0.09452277 0.07849676 0.08957108 0.08095242 0.07962766
 [7] 0.08417789 0.08596715 0.09170094 0.08749199 0.04200741 0.09095385

> (1/(valkama_r1$vi))/sum(1/(valkama_r1$vi))

 [1] 0.4587807549 0.4403413876 0.0041056146 0.0147909514 0.0049900976
 [6] 0.0044771522 0.0067788326 0.0083405725 0.0258796417 0.0102830650
[11] 0.0006757601 0.0205561697
```

Standard inference on τ^2 in REM

Malzahn, Bohning and Holling (2000) show that when using Hedges' d :

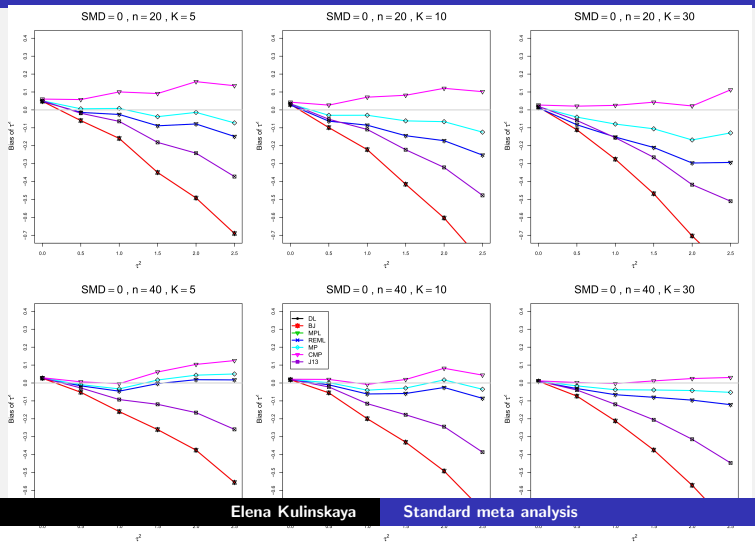
$\hat{\tau}^2_{DL}$ has considerable negative bias which increases strongly in magnitude with the actual τ^2. This is the result of ignoring the randomness of $\hat{w}_i$.

The PM estimator is better than REML for SMD and about the same for log RR, see Bakbergenuly et al.(2019) Statistics in Medicine, 1–21 for MA of MD and SMD.

For Valkama reed data, $\tau^2_{DL} = 0.0602$, $\tau^2_{REML} = 0.2049$ and $\tau^2_{PM} = 0.2003$ resulting in significant effect of reed management when using DL and ns effect for REML and PM.

All estimators are not reliable for small n and/or small K.

Bias of between-study variance for $SMD = 0$ and $n_T = n_C$ with equal sample sizes across studies



Estimation of the combined effect θ_{comb}

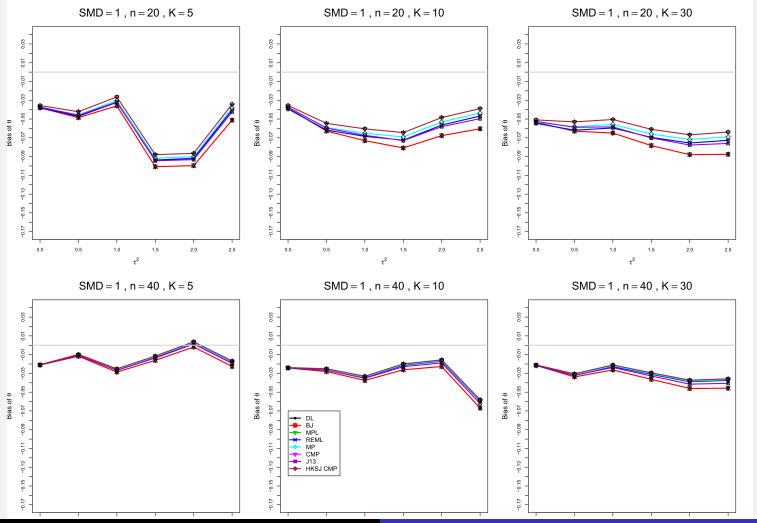
Because the estimated effects $\hat{\theta}_j$ and their weights $\hat{w}_j$ based on the inverse variances are not independent, the combined estimator $\hat{\theta}_{comb}$ is biased when the sample sizes are small.

This is true both for the SMD, and for the log response ratio.

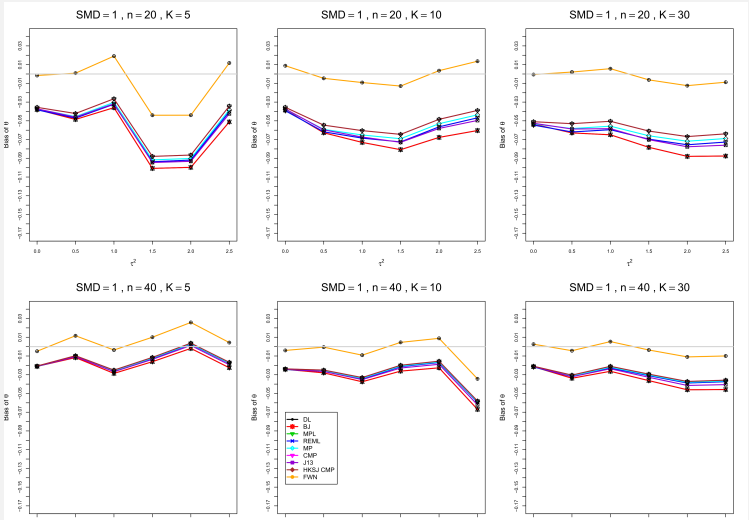
The only remedy to these biases is to use fixed weights (equal or inverse sample sizes based), but this is not yet part of mainstream packages.

Log response ratio is additionally biased due to boundary problems at zero. See Bakbergenuly et al. (2020) BMC Medical Research Methodology, 20:263 on MA for LRR.

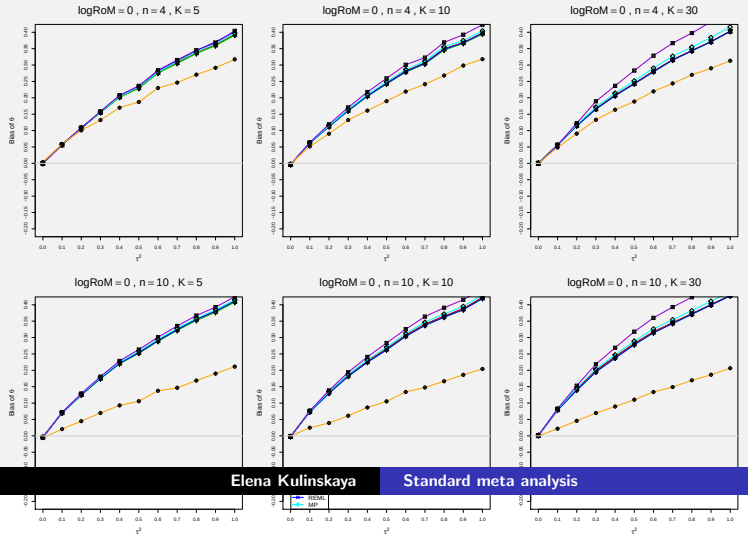
Bias of the combined effect for $SMD = 1$ and $n_T = n_C$



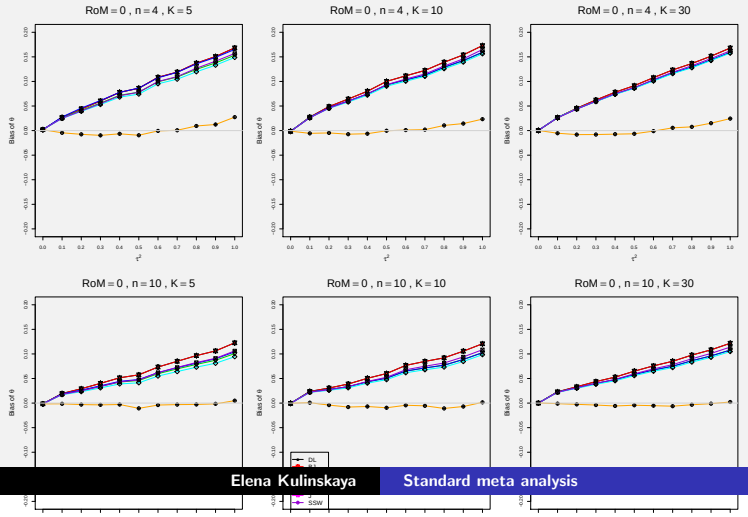
Bias of the combined effect for $SMD = 1$ and $n_T = n_C$



Bias of the combined log response ratio when $\bar{X}_C = \bar{X}_T = 1$ and $n_T = n_C$



Bias of the combined log response ratio when $\bar{X}_C = \bar{X}_T = 4$ and $n_T = n_C$



Standard inference on θ_{comb} in REM

Since $\text{Var}(\hat{\theta}) = (W^*)^{-1}$ is estimated as $(\sum \hat{w}_i^*)^{-1}$, it is underestimated. This should lead to considerable inflation of type I error and too narrow CIs for the combined effect.

Simulations in Viechtbauer (2005) show:

Type 1 error and confidence levels are only controlled adequately when τ^2 is was close to zero and/or K is large.

For small K, the probability of falsely rejecting H_0 became increasingly inflated for all methods as the average sample size $\bar{n}$ and τ^2 increased.

For example, the type I error for $K = 5$ and $\bar{n} = 80$ is 12%, and for $K = 10$ it is 8%, reducing to 6% at $K = 40$.

Hartung–Knapp–Sidik–Jonkman method (HKSJ)

Hartung and Knapp (2001), Sidik and Jonkman (2002) have proposed to take into account the randomness of weights by using the t-distribution for inference.

Their method can be used with any good estimator of τ^2, such as REML or PM.

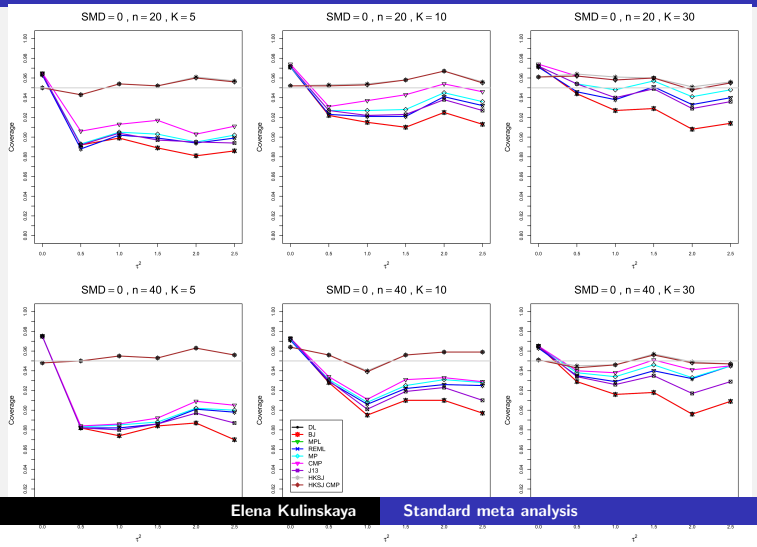
IntHout et.al (2014) have shown that the t confidence intervals based on HKSJ method result in improved coverage compared to the standard methods.

Use test="knha".

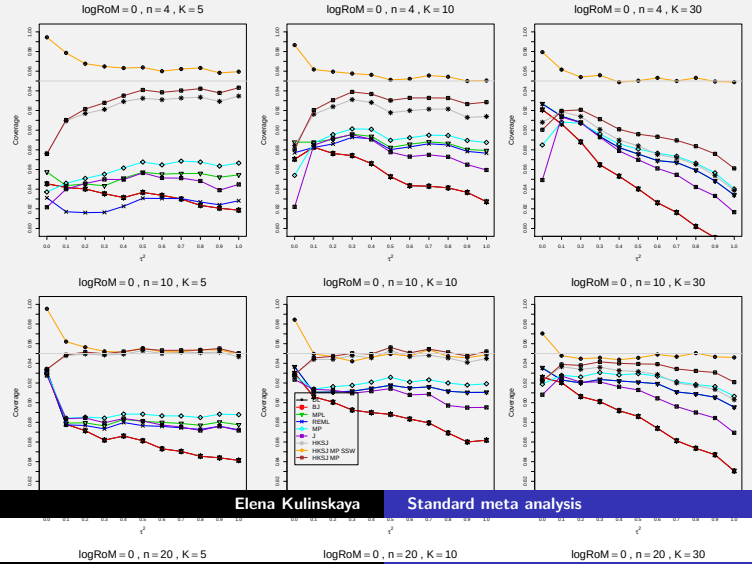
MA of the reed data, standard vs HKSJ t-based inference

```
> valkama_r1<-rma(yi,vi, data=dat, subset=(Variable.2=="stem height")
&(Species.2=="reed"),method="REML")
> summary(valkama_r1)

Random-Effects Model (k = 12; tau^2 estimator: REML)
tau^2 (estimated amount of total heterogeneity): 0.2049 (SE = 0.0903)
Model Results:
estimate      se     zval    pval    ci.lb   ci.ub
 -0.2146  0.1329  -1.6145  0.1064  -0.4752  0.0459

> valkama_r2<-rma(yi,vi, data=dat, subset=(Variable.2=="stem height")
 &(Species.2=="reed"),method="REML",test="knha")
> summary(valkama_r2)

Model Results:
estimate      se     tval    pval    ci.lb   ci.ub
 -0.2146  0.1315  -1.6325  0.1308  -0.5040  0.0747
```


Coverage of the combined effect for $SMD = 0$ and $n_T = n_C$



Coverage of the combined log response ratio when $\bar{X}_T = \bar{X}_C = 4$ and $n_T = n_C$



Summary

In this session we dealt with the standard meta analysis based on inverse variances.

- The assumptions are violated by treating the weights $\hat{w}_i$ as if they are constant. For many measures there is additionally dependence of the effects and their variances/weights.
- Standard inference in FEM works for large enough sample sizes n. Permutation tests can be used for small n.
- Standard inference in REM is very dubious for small K and n, but works OK for large K and large n. Always use the HKSJ method.
- Better not to use the response ratio!