6.2-6.4 Between-Study Heterogeneity

现在，我们已经知道如何 `pool effect size` 了。

然而，原始研究间和合并后的效应量存在差异，这种差异称为 `研究间异质性`

研究间异质性也是导致真实效应量之间存在差异的原因，所以我们要使用 `tau^2` 对研究间异质性进行估计，估计真实效应量的方差（variance）。

对真实效应量的方差进行估计可以允许我们 `池化效应` ，得出真实效应量分布的平均值 μ

异质性可能来源于：

- 数据中有两个或更多的研究亚组具有不同的真实效果。

所以对异质性来源进行审查是有价值的。

> 每一个好的 meta-analysis 不仅应该报告总体效果，而且还应该说明这个估计的可信度。其中一个重要的部分是量化和分析研究之间的异质性。

学习完这一章，应该清楚：

1. **测量异质性的不同方法**以及**如何解释它们。**
2. 我们还将介绍一些工具，**这些工具使我们能够检测导致数据异质性的研究。**
3. 最后，我们讨论了在"现实世界"的 meta-analysis 中**处理大量异质性的方法**。

▼ 5.1 Measures of Heterogeneity

异质性的不同类型，包括：

1. 基线或设计相关异质性（可以通过严格限制PICOS减少）
2. 统计异质性（即使研究间几乎看上去同质，也可能会出现统计异质性）

> （Rücker and colleagues 2008）

本章主要讨论的between-study heterogeneity主要是**统计异质性**

▼ 5.1.1 Cochran's Q

基于随机效应模型的理论假设，可知真实效应量之间的差异共有两个来源：

1. `sampling error` （可用 `SE` 表示），数学表达为 ϵ_k
2. `between-study heterogeneity` 数学表达为 ζ_k

$$\hat{\theta}_k = \mu + \zeta_k + \epsilon_k$$

当我们想量化 `between-study heterogeneity` 时，我们需要知道有多少变化可归因于抽样误差（ ϵ_k ），有多少归因于真实的效应量差异（ ζ_k ） 。

- 传统的做法是使用 Q 区分**1) 抽样误差** 和 **2) 真实效应量差异**。
- Cochran's Q is defined as a `weighted sum of squares ( WSS )` ，它使用每项研究观察到的效应 $\hat{\theta}_k$与总效应 $\hat{\theta}$的偏差,用研究方差的倒数 w_k^* 加权：

$$Q = \sum_{k=1}^{K} w_k \left(\hat{\theta}_k - \hat{\theta}\right)^2$$

我们可以看到，计算cochrane' Q的公式也是使用了**逆方差加权法**

第 k 个研究观测到的效应量偏离总体效应量的值，即残差，是平方的（因此值总是正的），加权，然后求和。结果值就是 `Cochrane's` Q ，即加权的残差平方和。

- 值得注意的是，Q 的大小不仅取决于残差，还取决于该研究的权重，更大权重的研究（标准误非常低）即使与汇总效应值的差异非常小，仍会对Q 统计量产生非常大的影响（变大）。
- Q 统计量的核心价值在于检测数据中**是否存在超出抽样误差的额外变异，**如果这样说，我们可以合理的假设其余是来源于 `研究间异质性` **，让我们进行模拟以更好的描述：**

我们模拟两种情况

1. 有研究间异质性
2. 无研究间异质性

下 Q 统计量的表现：

▼ 1. 无研究间异质性

- 这意味着 $\zeta_k = 0$ ，残差 $\hat{\theta}_k - \hat{\theta}$ 全部来源于 ϵ_k ，我们可以使用rnorm进行模拟：

$$\hat{\theta}_k - \hat{\theta} \sim \mathcal{N}(0, 1)$$

```
set.seed(123) # needed to reproduce results
rnorm(n = 40, mean = 0, sd = 1) # 随机生成40个观测，均值为0，标准差为1

##  [1] -0.56048 -0.23018  1.55871  0.07051  0.12929
##  [6]  1.71506  0.46092 -1.26506 -0.68685 -0.44566
##  [...]
```

重复抽取ten thousand times：

```
set.seed(123)
error_fixed <- replicate(n = 10000, rnorm(40))
```

▼ 1. 有研究间异质性

重复抽取ten thousand times：

```
set.seed(123)
error_random <- replicate(n = 10000, rnorm(40) + rnorm(40))
```

▼ 1. 模拟 Q 统计量

```
set.seed(123)
Q_fixed <- replicate(10000, sum(rnorm(40)^2))
Q_random <- replicate(10000, sum((rnorm(40) + rnorm(40))^2))
```

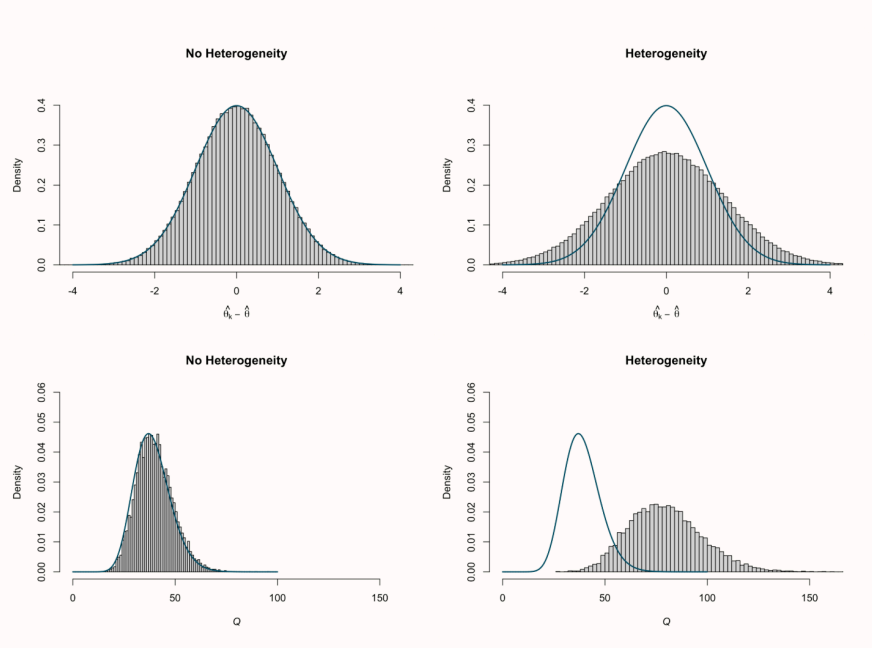
假设方差为1，因此权重可以被从公式中除去，对残差平方后求和。

解释：

概念	含义
Q 统计量	衡量研究之间效应差异的统计量
$Q \sim \chi^2_{k-1}$	在无异质性假设下，Q 的分布近似为自由度 $k-1$ 的卡方分布
卡方分布的特性	只能为正值；自由度小→右偏，大→近似正态；期望值=自由度
实践意义	如果 Q 远大于期望值，说明异质性存在

```r
# Histogram of the residuals (theta_k - theta)
# - We produce a histogram for both the simulated values in
#   error_fixed and error_random
# - `lines` is used to add a normal distribution in blue.

hist(error_fixed,
     xlab = expression(hat(theta[k])~-~hat(theta)), prob = TRUE,
     breaks = 100, ylim = c(0, .45), xlim = c(-4,4),
     main = "No Heterogeneity")
lines(seq(-4, 4, 0.01), dnorm(seq(-4, 4, 0.01)),
      col = "blue", lwd = 2)

hist(error_random,
     xlab = expression(hat(theta[k])~-~hat(theta)), prob = TRUE,
     breaks = 100,ylim = c(0, .45), xlim = c(-4,4),
     main = "Heterogeneity")
lines(seq(-4, 4, 0.01), dnorm(seq(-4, 4, 0.01)),
      col = "blue", lwd = 2)


# Histogram of simulated Q-values
# - We produce a histogram for both the simulated values in
#   Q_fixed and Q_random
# - `lines` is used to add a chi-squared distribution in blue.

# First, we calculate the degrees of freedom (k-1)
# remember: k=40 studies were used for each simulation
df <- 40-1

hist(Q_fixed, xlab = expression(italic("Q")), prob = TRUE,
     breaks = 100, ylim = c(0, .06),xlim = c(0,160),
     main = "No Heterogeneity")
lines(seq(0, 100, 0.01), dchisq(seq(0, 100, 0.01), df = df),
      col = "blue", lwd = 2)

hist(Q_random,  xlab = expression(italic("Q")), prob = TRUE,
     breaks = 100, ylim = c(0, .06), xlim = c(0,160),
     main = "Heterogeneity")
lines(seq(0, 100, 0.01), dchisq(seq(0, 100, 0.01), df = df),
      col = "blue", lwd = 2)
```

输出如下：



对于异质性的例子，分布看起来完全不同。模拟的数据似乎根本不符合预期的分布。 值明显地向右移动；分布的平均值大约是它的两倍。我们可以得出结论，当研究间存在大量异质性时，Q 的值大大高于我们在假设无异质性时所期望的 $K-1$的值。这并不奇怪，因为我们在数据中添加了额外的变量来模拟研究间异质性的存在。

- Cochran's Q 是一个非常重要的统计量，主要是因为其他常用的量化异质性的方法，如 $Higgins$和 $Thompson$的 I^2 统计量和 H^2 都是基于它。

> 注意：我们不应该只依赖于 Q 统计量，要结合使用。

▼ 5.1.2 Higgins & Thompson's I^2 Statistic

- 基于 Q 衡量研究间异质性，它被定义为由非抽样误差引起的效应量中变异的百分比。

数学表达式如下：

$$I^2 = \frac{Q-(K-1)}{Q}$$

R模拟一下：

```
# Display the value of the 10th simulation of Q
Q_fixed[10]
## [1] 35.85787

# Define k
k <- 40

# Calculate I^2
(Q_fixed[10] - (k-1))/Q_fixed[10]
## [1] -0.08762746
```

≈ 0，研究间异质性占总变异比为零

```
(Q_random[10] - (k-1))/Q_random[10]

> (Q_random[10] - (k-1))/Q_random[10]
[1] 0.5692061
```

存在异质性了，正如前所述，基于cochrane's Q 计算，当Q统计量远高于χ^2_{k-1}的期望值时，表示异质性存在，具体结合 p 值和 I^2 的经验性判断：

- I^2 `= 25%` : low heterogeneity
- I^2 `= 50%` : moderate heterogeneity
- I^2 `= 75%` : substantial heterogeneity

▼ 5.1.4 Heterogeneity Variance τ^2 & Standard Deviation τ

请报告：which method has been used to calculate the confidence intervals around τ^2 in your meta-analysis.

> 在{meta包}中，计算 `tau^2` 周围置信区间的方法主要包括 `Q -Profile method` 和 `Jackson method` ，值得注意的是，除了使用 `DL method` 时 {meta}包默认使用 `J` 方法，其余均以 `QP` 法为默认方法。仅需特殊制定：`method.tau.ci = "QP" / method.tau.ci = "J"`

▼ 5.2 Which Measure Should I Use?

当我们评估和报告异质性时，需要稳健的估计，而且不太受统计功效的影响。

> When we assess and report heterogeneity in a meta-analysis, we need a measure which is robust, and not too heavily influenced by statistical power.

例如cochrane Q 会随着样本量和精确度的增加而增加（更容易出现显著的异质性，而在样本量不足时，其统计功效又不足，所以将显著性水平设置为 Q < 0.10）

> Cochran's *Q* increases both when the number of studies increases, and when the precision (i.e. the sample size of a study) increases.

所以当我们进行研究间异质性估计时，不能仅仅依赖于 Q 统计量（因为其很大程度上依赖于统计效力，并且随着样本量的变化而变化 {不稳健} ）。

▼ I^2

比较稳健，因为不受到样本量变化的影响（因为被转换成了比例），并且易于解释（很多研究中对其进行汇报），如果对 I^2 的不确定性进行汇报再好不过：`I^2 = 62.6% [37.9%; 77.5%]`

- 但是仅汇报 I^2 不够，因为随着样本量的增加，抽样误差趋于零，此时 I^2 的增加仅因为样本量的增加，因此：

✅ 1. 报告 τ^2（Tau-squared）

- 这是真实异质性方差的估计值，**有单位、反映真实变异量大小**。
- τ² 不直接依赖样本量，**更稳健、更实用**，尤其是在模型构建或预测新研究时。

✅ 2. 报告 I^2 的置信区间

- 提供 I^2 的 95% CI 可以告诉读者估计是否稳定。
- 如果 CI很宽（如 10% – 90%），那就说明不确定性很高。

✅ 3. 报告 预测区间 （prediction interval）

- 提供将来某项研究可能落在的效应范围，比 I^2更直观有意义。

▼ τ^2

- 更稳健，不受到样本量和精确度变化的影响，但不好解释

> ***Prediction intervals (PIs) are a good way to overcome this limitation.*** (IntHout et al. 2016)

- PIs为我们提供了一个范围，我们可以根据现有证据预期未来研究的效应量可能的落点范围（就像预判网球的落点）

- 如果预测区间完全大于零（positive，尽管真实效应量之间的变异存在），则干预可能在各种情况下都是有益的；如果它包含零，则效果不太确定（尽管一个宽泛的预测区间很常见）

报告内容	是否推荐	原因
I^2 点估计	✅	描述变异的比例，直观
I^2 置信区间	✅✅	显示估计不确定性，避免误解
τ^2	✅✅✅	更稳健，不受样本量影响
Q 及其 p 值	⚠️ 有限作用	太依赖样本量，容易误导
预测区间	✅✅	表达未来研究可能效应范围

计算预测 `PIs` around the $\hat{\mu}$，我们使用`tau^2` 和观测真实效应量均值的 `SE`：

The formula for 95% prediction intervals looks like this:

$$\hat{\mu} \pm t_{K-1,0.975}\sqrt{SE^2_{\hat{\mu}} + \hat{\tau}^2}$$

```
qt(0.975, df = 5)
```

幸运的是不理解计算公式也没关系，{meta} 提供了计算预测区间的功能`prediction = TRUE`

建议至少总是报告I^2（带置信区间），以及预测区间，并相应地解释结果。

▼ 5.3 Assessing Heterogeneity in *R*

```
m.gen <- update(m.gen, prediction = TRUE)

summary(m.gen)

## Review:     Third Wave Psychotherapies
##
## [...]
```

```
##
## Number of studies combined: k = 18
##
##                               SMD            95%-CI    t  p-value
## Random effects model (HK) 0.5771 [ 0.3782; 0.7760] 6.12 < 0.0001
## Prediction interval              [-0.0572; 1.2115]
##
## Quantifying heterogeneity:
##  tau^2 = 0.0820 [0.0295; 0.3533]; tau = 0.2863 [0.1717; 0.5944];
##  I^2 = 62.6% [37.9%; 77.5%]; H = 1.64 [1.27; 2.11]
##
## Test of heterogeneity:
##      Q d.f. p-value
##  45.50   17  0.0002
##
## Details on meta-analytical method:
## - Inverse variance method
## - Restricted maximum-likelihood estimator for tau^2
## - Q-profile method for confidence interval of tau^2 and tau
## - Hartung-Knapp adjustment for random effects model (df = 17)
## - Prediction interval based on t-distribution (df = 16)
```

所以我们关注：

1. Q统计量 和 显著性
2. I^2 统计量和 95%置信区间
3. `tau^2` 和95%置信区间
4. PI 和95%置信区间

Reporting the Amount of Heterogeneity In Your Meta-Analysis
Here is how we could report the amount of heterogeneity we found in our example:

> "The between-study heterogeneity variance was estimated at ˆτ2 = 0.08 (95%CI: 0.03-0.35), with an I2 value of 63% (95%CI: 38-78%). The prediction interval ranged from g = -0.06 to 1.21, indicating that negative intervention effects cannot be ruled out for future studies."

我们可以看到，研究的真实效应量之间存在一定的差异，有必要进一步探索这些差异的来源：

- 也许是因为几项研究异常值，导致我们对合并效应量的低估或高估
- 为了解决这些问题，我们现在将了解使我们能够评估汇总结果的稳健性的程序：
- `Outlier analysis（异常值分析）`
- `Influence analysis（影响力分析）`

排查那些研究对结果的影响较大。

当观测到中等到大的异质性时，应该怎么办?

- 当 I^2 > 50% 时，即表明中等到大的异质性来源于研究间观测到的真实效应量的变异，但这不是iron-clad的，而是"rule of thumb"，所以要事先指定分析计划，而不是根据喜好和无理由的情况下随意的删除研究，这种做法是不严谨的"后验操作（p-hacking）"风险，这会让分析结果受到研究者主观意图（researcher agenda）污染。

▼ 5.4 Outliers & Influential Cases

- 前面提到，一两项研究可能为极端值（过大或过小），影响我们最终的效应量池化，一个好的方法是**移除这些研究后重新池化效应量并观察**。
- 另一方面，我们想看池化后的效应量是否稳健（是否会受到单个研究的影响，是否会有单个研究将池化效应量拉到某个方向），这种研究我们称之为 `influential cases`

▼ 5.4.1 Basic Outlier Removal

有几种方法可以识别 `influential cases`：

1. `"Brute force" approach`：如果某项研究的置信区间和池化后观测到的真实效应量的置信区间不重叠，请检查纳入观测效应量是否出现以下两种情况：
   - 效应极小：for which the **upper bound** of the 95% confidence interval is **lower** than the **lower bound** of the pooled effect confidence interval (i.e. extremely **small** effects)
   - 效应极大：for which the **lower bound** of the 95% confidence interval is **higher** than the **upper bound** of the pooled effect confidence interval (i.e. extremely **large** effects).

   `dmetar` 可以完成这个算法，其以meta的输出结果为对象。确保你已经加载了它：

```
find.outliers(m.gen)
```

```
## Identified outliers (random-effects model)
## ----------------------------------------
## "DanitzOrsillo", "Shapiro et al."
##
## Results with outliers removed
## ----------------------------
## Number of studies combined: k = 16
##
##                         SMD           95%-CI    t  p-value
## Random effects model 0.4528 [0.3257; 0.5800] 7.59 < 0.0001
## Prediction interval         [0.1693; 0.7363]
##
## Quantifying heterogeneity:
##  tau^2 = 0.0139 [0.0000; 0.1032]; tau = 0.1180 [0.0000; 0.3213];
##  I^2 = 24.8% [0.0%; 58.7%]; H = 1.15 [1.00; 1.56]
##
## Test of heterogeneity:
##      Q d.f. p-value
##  19.95   15  0.1739
##
## [...]
```

We see that the `find.outliers` function has detected two outliers, "`DanitzOrsillo`" and "`Shapiro et al.`".

- 当排除了这些研究后，`find.outliers` function rerun 了结果。

现在观测一下排除outliers后的结果：

```
summary(find.outliers(m.gen)$m.random)

> summary(find.outliers(m.gen)$m.random)
Review:     Third Wave Psychotherapies

                          SMD            95%-CI %W(random) exclude
Call et al.            0.7091 [ 0.1979; 1.2203]        4.4
Cavanagh et al.        0.3549 [-0.0300; 0.7397]        6.9
DanitzOrsillo          1.7912 [ 1.1139; 2.4685]        0.0       *
de Vibe et al.         0.1825 [-0.0484; 0.4133]       13.1
Frazier et al.         0.4219 [ 0.1380; 0.7057]       10.4
Frogeli et al.         0.6300 [ 0.2458; 1.0142]        6.9
```

```
Gallego et al.          0.7249 [ 0.2846; 1.1652]        5.6
Hazlett-Stevens & Oren 0.5287 [ 0.1162; 0.9412]        6.2
Hintz et al.            0.2840 [-0.0453; 0.6133]        8.6
Kang et al.             1.2751 [ 0.6142; 1.9360]        2.8
Kuhlmann et al.         0.1036 [-0.2781; 0.4853]        7.0
Lever Taylor et al.     0.3884 [-0.0639; 0.8407]        5.4
Phang et al.            0.5407 [ 0.0619; 1.0196]        4.9
Rasanen et al.          0.4262 [-0.0794; 0.9317]        4.5
Ratanasiripong          0.5154 [-0.1731; 1.2039]        2.6
Shapiro et al.          1.4797 [ 0.8618; 2.0977]        0.0       *
Song & Lindquist        0.6126 [ 0.1683; 1.0569]        5.6
Warnecke et al.         0.6000 [ 0.1120; 1.0880]        4.8

Number of studies: k = 16

                              SMD           95%-CI    t  p-value
Random effects model (HK) 0.4528 [0.3257; 0.5800] 7.59 < 0.0001
Prediction interval               [0.1705; 0.7352]

Quantifying heterogeneity (with 95%-CIs):
 tau^2 = 0.0139 [0.0000; 0.1032]; tau = 0.1180 [0.0000; 0.3213]
 I^2 = 24.8% [0.0%; 58.7%]; H = 1.15 [1.00; 1.56]

Test of heterogeneity:
     Q d.f. p-value
 19.95   15  0.1739

Details of meta-analysis methods:
- Inverse variance method
- Restricted maximum-likelihood estimator for tau^2
- Q-Profile method for confidence interval of tau^2 and tau
- Calculation of I^2 based on Q
- Hartung-Knapp adjustment for random effects model (df = 15)
- Prediction interval based on t-distribution (df = 15)
```

我们可以看到，两项被认为是异常值的研究的权重被设置为零（排除在分析之外了），效应量被重新计算（相应的置信区间、异质性指标和预测区间均重新计算），结果表明：

1. Q统计量无统计学意义（不显著）
2. I^2 统计量的减少：from 63% to 25%
3. `tau^2` 的置信区间现在也包括0了（异质性不显著）
4. 预测区间变得更加 narrowed（仅包含positive方向）

▼ 5.4.2 Influence Analysis

> 除了极端值（outliers）会影响结果的robust，一些研究观测到的效应量也会极大程度的影响池化后的总体效应量

当**研究的显著结果**依赖于**单一研究时**，也就意味着如果**将该研究removed**，结果可能**没有统计学意义**。因此，为了验证我们结果的稳健性，应该对此进行**探索**：

- 首先区分 `outliers cases` 和 `influence cases` ，outliers cases 只是数据极端，remove之后可能不会影响结果，而 influence cases 恰恰相反，即使数据并不极端，将其移除后可能会影响重分析的结果（即影响效应量池化的稳健性）

有许多工具可以实现这个探索：

1. `leave-one-out` method
   - In this approach, we recalculate the results of our meta-analysis K `times`, each time `leaving out one study`
2. **根据这些数据，我们可以计算出不同的影响力诊断：**
   - 影响诊断使我们能够检测对我们的元分析的总体估计**影响最大的研究**，并让我们评估这种巨大的影响是否影响了我们的汇总效应。

The {dmetar} package contains a function called `InfluenceAnalysis`：

注意：

1. **但是该函数只能以{meta}包的输出作为对象进行输入
2. default是FE，请设置 `random = TRUE`

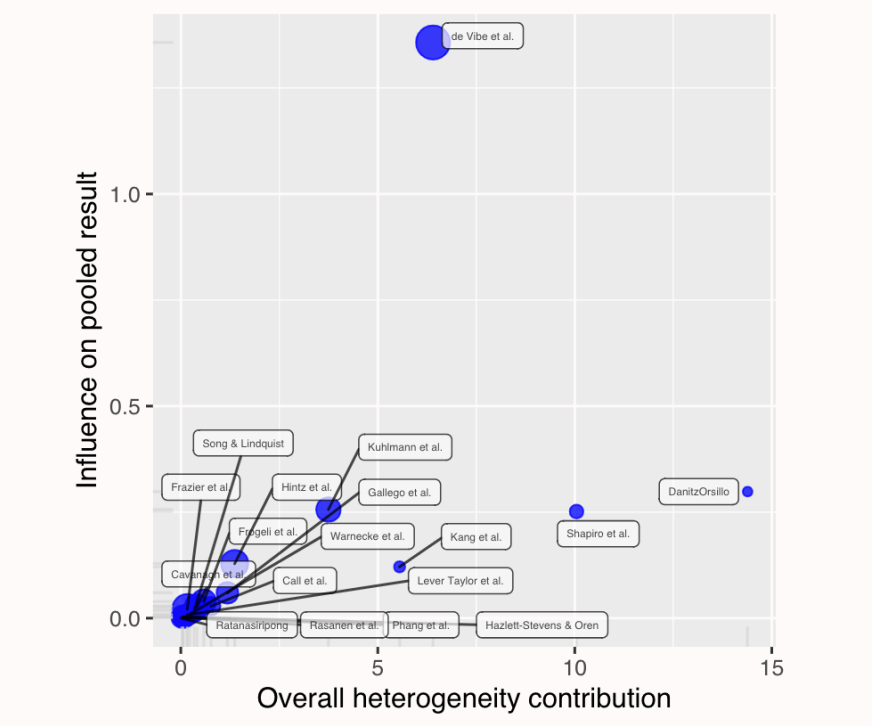
```
m.gen.inf <- InfluenceAnalysis(m.gen, random = TRUE)
```

The `InfluenceAnalysis` function creates four influence diagnostic plots:

	功能	排序依据
Baujat plot	找出对异质性和效应影响最大的研究	Q统计量 + 效应变化
Influence diagnostics	用回归影响指标识别关键研究	Cook's D、DFBETAS 等
Leave-one-out (效应量排序)	看删除后效应如何变	效应量变化
Leave-one-out (I^2排序)	看删除后异质性如何变	I^2 变化

▼ 5.4.2.1 Baujat Plot

```
plot(m.gen.inf, "baujat")
```



- X axis表示异质性的程度（as measured by the Cochrane's Q）
- Y axis 表示 对于池化效应量的影响

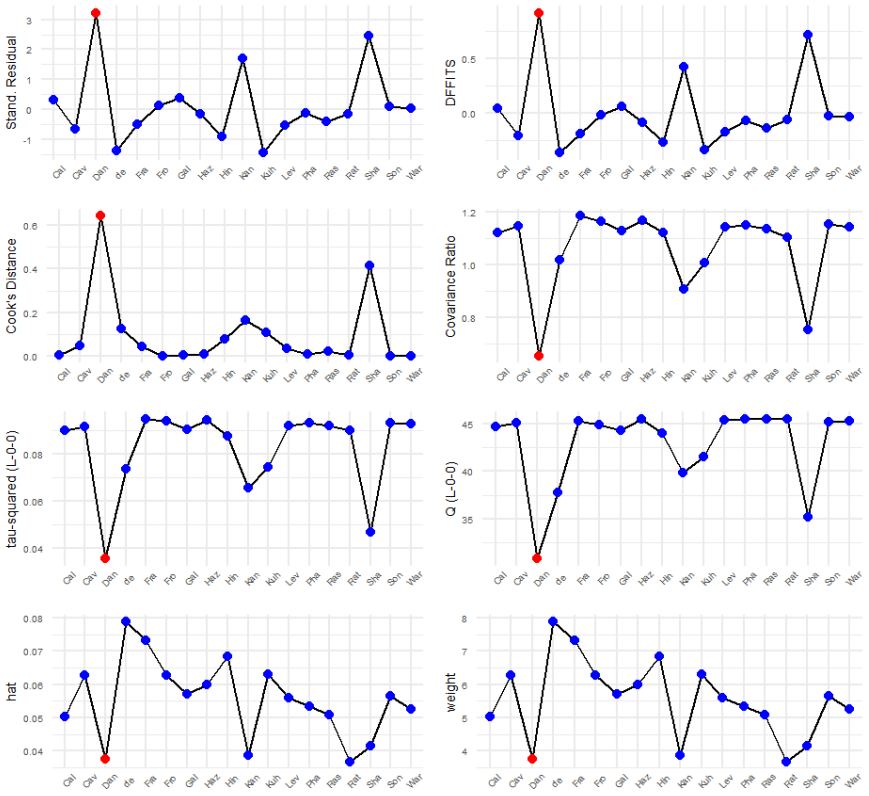
`Baujat图` 用于诊断荟萃分析中过度影响异质性的研究。靠近右侧的研究，对整体异质性有很大贡献，可能为异常值的相关研究。靠近右上角的研究，可能会对整个分析有特别的影响，因为对异质性和合并效应都有很大的影响，如"de Vibe et al."。

▼ 5.4.2.2 Influence Diagnostics

可视化 `influence diagnostics` 的结果：

```
plot(m.gen.inf, "influence")
```

Studies determined to be influential are displayed in red in the plot generated by the `InfluenceAnalysis` function:



- 诊断图提供了单个研究与我们拟合模型之间的拟合程度，哪些拟合的好，哪些拟合的不好

▼ 5.4.2.2.1 Externally Standardized Residuals

$$t_k = \frac{\hat{\theta}k - \hat{\mu}_{\setminus k}}{\sqrt{\mathrm{Var}(\hat{\mu}_{\setminus k}) + \hat{\tau}^2_{\setminus k} + s^2_k}}$$

理解一下：

对第 k 项研究来说，它衡量的是“该研究的观测效应量 $\hat{\theta}_k$ 与**不含它时**重新计算得到的池化效应 $\hat{\mu}_{\setminus k}$ 之间的差距”，并且把这条差距除以一个综合的标准误来做**标准化**——这样就能把**不同研究**放到**同一量纲**，直接比较谁偏离得过大。

综合的标准误：

(1) the **variance** of the external effect (i.e. the squared standard error of $\mu^2_{\setminus k}$

(2) the $\tau 2$ estimate of the external pooled effect,

(3) the variance of k

- 所以综合的标准误就是除以（1）剔除第K项研究后合并效应量的方差 + （2）剔除第K项研究后合并效应量的tau^2 + （3）第K个研究标准误的平方。三者之和的平方根以标准化。

`influence cases` **de 判断标准：**

- $|t_k| > 1.96$: ：统计上显著偏离
- $|t_k| > 2.5$: ：建议进一步审查
- $|t_k| > 3$ ：高度偏离，疑似异常值

▼ 5.4.2.2.2 DFFITS Value

$$\mathrm{DFFITS}_k = \frac{\hat{\mu} - \hat{\mu}_{\backslash k}}{\sqrt{\frac{w_k^{()}}{\sum_{k=1}^{K} w_k^{()}} \cdot \left(s_k^2 + \hat{\tau}_{\backslash k}^2\right)}}$$

> **用来判断：某一研究对合并效应（pooled effect）的影响有多大。**

它和外部标准化残差 t_k 可比，形式接近，但关注点是 第K个研究**"是否对模型结果有显著影响"**，而不仅是偏离。

▼ 5.4.2.2.3 Cook's Distance

与DFFIT基本一致，唯一的区别是差异被^2了，所以只能取正值：

$$D_k = \frac{\left(\hat{\mu} - \hat{\mu}_{\backslash k}\right)^2}{\sqrt{s_k^2 + \hat{\tau}^2}} \cdot$$

= D_k 值越大，数据越有可能是 `influence cases`

▼ 5.4.2.2.4 Covariance Ratio

$$\mathrm{CovRatio}_k = \frac{Var(\hat{\mu}_{\backslash k})}{Var(\hat{\mu})}$$

`CovRatio` **衡量的是某项研究对 meta 分析结果"** `精度` **"的影响。**

若 CovRatio < 1，说明删掉该研究后方差更小、结果更稳定，它可能是"influence cases"或不太可靠的效应量。

相反，CovRatio > 1 的研究是"稳定器"——提高了 meta 结果的可靠性。

意思就是：

当删除第k个研究后，重新合并效应量的方差和原始效应量的方差的比例小于1，代表剔除了第k个研究后，估计变得更精确了，也就代表第k个研究可能是对合并结果**增加了整体的不确定性**的研究。

▼ 5.4.2.2.5 Leave-One-Out $\tau 2$ and Q Values

我们 `逐个删除每一项研究`，观察模型中异质性指标 `tau^2` 和 `Cochran's` Q 是否发生明显变化。意思就是，删除第k项研究后，tau^2和Q变小了，所以这个研究可能是产生异质性的来源。

然而，有一些有用的"经验法则"可以指导我们的决定。如果**满足以下条件之一**，则影响分析函数**将研究视为有影响的案例**：

$$\mathrm{DFFITS}_k > 3\sqrt{\frac{1}{k-1}}$$

$$D_k > 0.45$$

$$\mathrm{hat}_k > \frac{3}{k}$$

模拟一下帮助我更好的理解这个内容：

假设共有20项研究：

指标	数值
DFFITS	1.1
Cook's D	0.52
Hat 值	0.18

我们现在计算每项的阈值：

指标	判断标准	结果
DFFITS	$> 3 \cdot 1/(k-1) = 3 \cdot 1/19 \approx 0.688$	✅ 1.1 > 0.688 → 超标

Cook's D	> 0.45	✅ 0.52 > 0.45 → 超标
Hat 值	$> 3/k = 3/20 = 0.15$	✅ 0.18 > 0.15 → 超标

	含义
DFFITS 超阈值	删除该研究会显著改变合并效应估计
Cook's D 超阈值	该研究对模型拟合整体影响较大
Hat 值超阈值	在模型结构中权重较高（高杠杆）

这进一步证实了这两项研究可能**扭曲了我们的综合效应估计**，并导致了我们在最初的荟萃分析中发现的**部分研究间异质性**。

> **This further corroborates that the two studies could have distorted our pooled effect estimate, and cause parts of the between-study heterogeneity we found in our initial meta-analysis.**

影响力分析的目的：

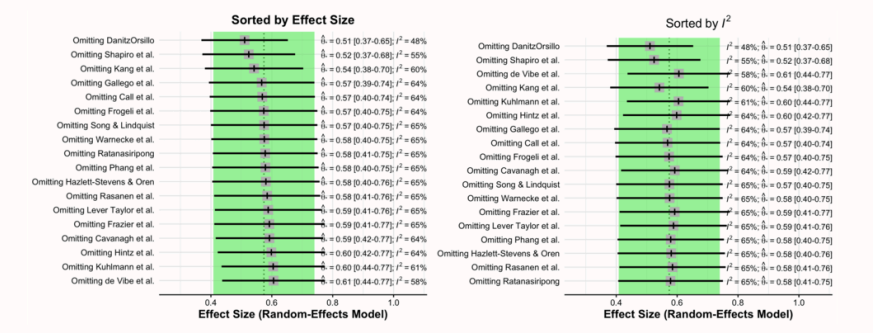
目的	说明
是否存在 **离群值（outliers）**	哪些研究偏离了合并效应的整体趋势（比如 t_k）
是否存在 **高杠杆研究（leverage）**	哪些研究由于样本大、方差小，在模型中占有"主导权重（`hat` 和 `weight` 分析）
哪些研究 **影响了合并效应的估计值（effect size）**	例如 DFFITS、Leave-one-out 检查研究删除后的效应变化
哪些研究 **影响了模型的异质性**（τ^2、I^2、Q）	是否某些研究导致总体不一致性明显变高
哪些研究 **改变了模型的拟合结构或稳定性**	如 Cook's D 指出研究对整体回归的拟合影响

总的来说，`Influence analysis` 是探索是否有 *异常研究* 影响了 *合并效应量* 和 *异质性*

▼ 5.4.2.3 Leave-One-Out Meta-Analysis Results

```
plot(m.gen.inf, "es")  # 按照池化效应量（lower to higher）排序
```

```
plot(m.gen.inf, "i2")  # 按照I^2值排序
```



当每次排除一项研究时，重新计算池化效应（with 95%CI）和I^2，绿色的阴影和中心的虚线代表的是原始池化效应量的点估计和95置信区间，

可以和 `Baujat` 图，`influence analysis` 连用，找到影响池化效应观测、观测精确度 和 异质性的来源：最后，因此，我们也应该**进行并报告排除这两项研究的敏感性分析结果（敏感性分析）**：也就是我们要进行outliers （观测置信区间是否overlap）和 influence analysis （结合baujat图、influence diagnostics、）以及 leave one-out （effect排序和I^2排序）分析，找到对池化效应量估计、对估计的精确性以及异质性贡献**最大的原始研究**，并在同一表格内汇报**剔除这些研究前后观测到**的池化效应量（with 置信区间）、异质性的变化，并做出相应解释（为什么移除这些研究）。

▼ 5.4.3 GOSH Plot Analysis

Let us see what happens if we `rerun` the meta-analysis while removing the **three studies** that the `gosh.diagnostics` function has identified.

```
update(m.gen, exclude = c(3, 4, 16)) %>%
  summary()

  ## Review:     Third Wave Psychotherapies
##                            SMD            95%-CI %W(random) exclude
## Call et al.            0.7091 [ 0.1979; 1.2203]        4.6
## Cavanagh et al.        0.3549 [-0.0300; 0.7397]        8.1
## DanitzOrsillo          1.7912 [ 1.1139; 2.4685]        0.0       *
## de Vibe et al.         0.1825 [-0.0484; 0.4133]        0.0       *
## Frazier et al.         0.4219 [ 0.1380; 0.7057]       14.8
## Frogeli et al.         0.6300 [ 0.2458; 1.0142]        8.1
## Gallego et al.         0.7249 [ 0.2846; 1.1652]        6.2
## Hazlett-Stevens & Oren 0.5287 [ 0.1162; 0.9412]        7.0
## Hintz et al.           0.2840 [-0.0453; 0.6133]       11.0
## Kang et al.            1.2751 [ 0.6142; 1.9360]        2.7
## Kuhlmann et al.        0.1036 [-0.2781; 0.4853]        8.2
## Lever Taylor et al.    0.3884 [-0.0639; 0.8407]        5.8
## Phang et al.           0.5407 [ 0.0619; 1.0196]        5.2
## Rasanen et al.         0.4262 [-0.0794; 0.9317]        4.7
## Ratanasiripong         0.5154 [-0.1731; 1.2039]        2.5
## Shapiro et al.         1.4797 [ 0.8618; 2.0977]        0.0       *
## Song & Lindquist       0.6126 [ 0.1683; 1.0569]        6.1
## Warnecke et al.        0.6000 [ 0.1120; 1.0880]        5.0
##
## Number of studies combined: k = 15
##
##                         SMD           95%-CI    t  p-value
## Random effects model 0.4819 [0.3595; 0.6043] 8.44 < 0.0001
## Prediction interval         [0.3586; 0.6053]
##
## Quantifying heterogeneity:
##  tau^2 < 0.0001 [0.0000; 0.0955]; tau = 0.0012 [0.0000; 0.3091];
##  I^2 = 4.6% [0.0%; 55.7%]; H = 1.02 [1.00; 1.50]
##
## Test of heterogeneity:
##      Q d.f. p-value
##  14.67   14  0.4011
## [...]
```

我们可以发现：

- studies number 3 and 16 are "`DanitzOrsillo`" and "`Shapiro et al.`"。这两项研究在之前的分析中（outliers 和 influence analysis）也被发现具有影响力。
- Study 4 is the one by "de Vibe"。虽然不是离群值和强影响值，但是这个研究具有极高的权重，即使其效应量与总体估计的偏差不大，但仍可能拉动总体效应量，对整体合并效应产生了实质影响，**所以也必须被当作一个影响性研究看待**。

删除三项具有强影响力的研究后，模型的异质性（$\tau^2 I^2$）几乎消失，说明它们是**异质性的关键来源**，而合并效应量虽然略有下降，但仍在同一数量级，**表明整体结果稳健（最初计算的平均效应不太受异常值和有影响力的研究的影响）**。

我们可以使用以下格式进行汇报：

1. 确定在我们的元分析中哪些研究是 `influence cases`
2. 报告sensitivity analysis的结果（剔除 `influence cases` 后的重分析）
3. 创建一个table展示**剔除前后**的结果，内容至少包括：
   - pooled effect, its confidence interval and p value
   - few measures of heterogeneity（I^2, τ^2 et al. as well as the confidence interval thereof）
   - prediction intervals
4. 汇报我们剔除了哪些原始研究（汇报新结果基于哪些研究）

Below is an example of how such a table looks like for our `m.gen` meta-analysis from before:

Analysis	*g*	95% CI	*p*	95% PI	I^2	95% CI (I^2)
Main Analysis	0.58	0.38 – 0.78	<0.001	-0.06 – 1.22	63%	39 – 78
Infl. Cases Removed[1] [b]	0.48	0.36 – 0.60	<0.001	0.36 – 0.61	5%	0 – 56

[1] Removed as outliers: DanitzOrsillo, de Vibe, Shapiro.

[b] Influential studies identified via influence diagnostics, outliers analysis and GOSH (e.g., DFFITS, Cook's D, standardized residuals) were removed in this sensitivity analysis.

这种类型的表非常方便，因为我们还可以添加其他敏感性分析结果的进一步行。例如，如果我们进行的分析只考虑了低偏倚风险的研究（第1.4.5章），我们可以在第三行报告结果。