Análise e manejo de outliers

Clarissa F. D. Carneiro

2023-06-05

Objetivo: Comparar valores de médias e desvios entre a extracao original e re-extracao. A amostra re-extraída foi selecionada por serem outliers, ou no tamanho de efeito ou na variancia dele.

Ao tentar fazer a meta-análise inicial, houve um erro que aponta para uma diferenca muito grande entre a menor e maior variancia (ou n). Os autores da funcao de análise indicam que isso pode ocorrer por erros de extracao de dados.

Comecamos conferindo a distribuicao de sample sizes e de vi para ver onde pode estar o problema:

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.000 3.000 3.000 4.216 4.000 24.000

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.00 3.00 3.00 4.22 4.00 24.00

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.1 0.8 2.3 1489.0 6.9 1494773.7 6

| rayyan.key | yi | vi |
| --- | --- | --- |
| rayyan-115935411 | -3458.06035 | 1494773.6746 |
| rayyan-115935411 | -900.85105 | 101442.5762 |
| rayyan-115922365 | 1158.08205 | 83822.6278 |
| rayyan-115923162 | -920.56888 | 70621.2546 |
| rayyan-115935519 | -156.71039 | 2047.1788 |
| rayyan-115934663 | -167.47123 | 1753.4132 |
| rayyan-115922941 | -114.15348 | 1086.5848 |
| rayyan-115934780 | -91.29182 | 695.1830 |
| rayyan-115922941 | -82.20311 | 563.7792 |
| rayyan-115926424 | -141.59031 | 501.3954 |

| rayyan.key | yi | vi |
| --- | --- | --- |
| rayyan-115928295 | -0.3530388 | 0.1015580 |
| rayyan-115928431 | -0.7329931 | 0.1016343 |
| rayyan-115928431 | -0.8843718 | 0.1045490 |
| rayyan-115928431 | -0.9461367 | 0.1058949 |
| rayyan-115928431 | -1.0620553 | 0.1086662 |
| rayyan-115928431 | -1.1683500 | 0.1114886 |
| rayyan-115928431 | -1.2023444 | 0.1124480 |
| rayyan-115921714 | -0.3598424 | 0.1129095 |
| rayyan-115928431 | -1.2202264 | 0.1129637 |
| rayyan-115928431 | -1.2354837 | 0.1134098 |

Além desses casos, observamos também outras comparacoes com tamanhos de efeito muito extremos. Decidimos conferir esses outliers, comecando pelos mais extremos. Dependendo de quantos erros encontrarmos, devemos decidir até que threshold deveríamos conferir todas as extracoes.

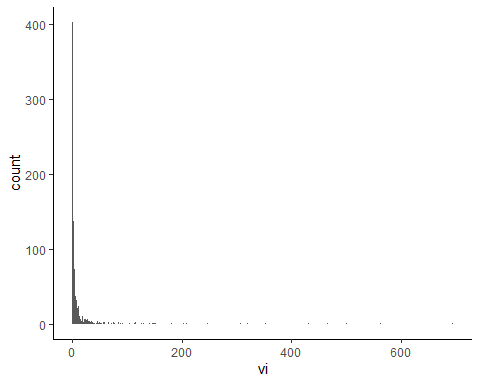
Comparacoes com g menor ou igual a -100: 7

Comparacoes com g menor ou igual a -50 (e maior que -100): 10

Comparacoes com g menor ou igual a -20 (e maior que -50): 62

Comparacoes com g menor ou igual a -10 (e maior que -20): 179

Para decidir os bins de variancia, podemos olhar pra distribuicao dos valores:



Comparacoes com vi maior ou igual a 1000: 7

Comparacoes com vi maior ou igual a 500 (e menor que 1000): 3

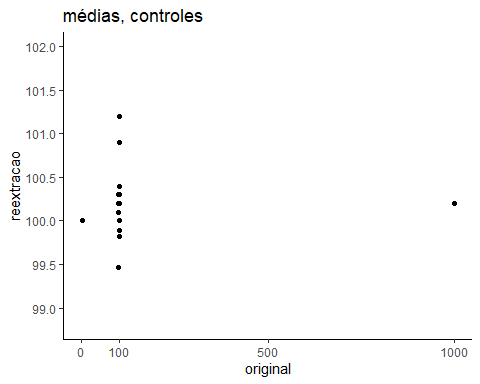
Comparacoes com vi maior ou igual a 100 (e menor que 500): 21

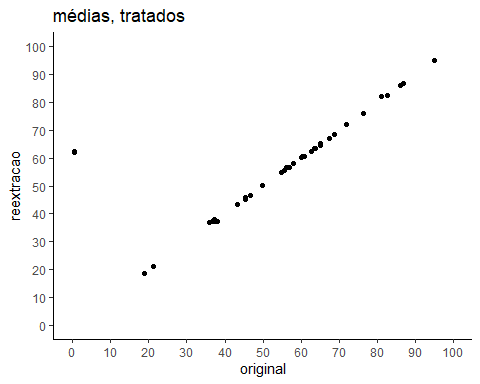
Comparacoes com vi maior ou igual a 50 (e menor que 100): 19

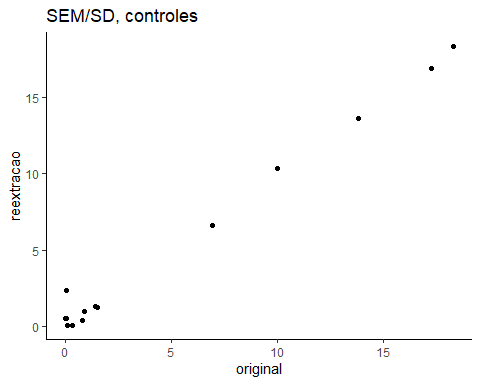
Comparacoes com vi maior ou igual a 10 (e menor que 50): 171

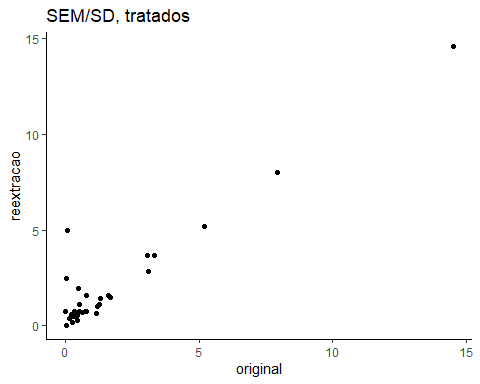
No primeiro round de verificacoes, decidimos re-extrair médias e variacao das comparacoes com tamanho de efeito menor ou igual a -50, e variancia maior ou igual a 100.

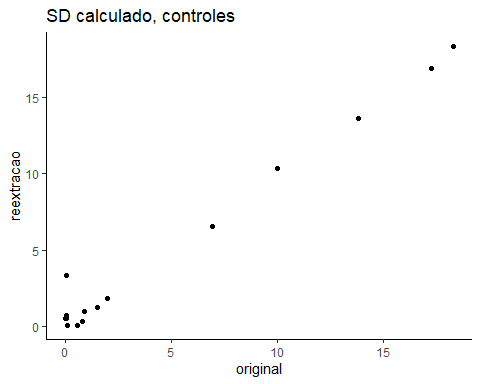
Correlacoes:

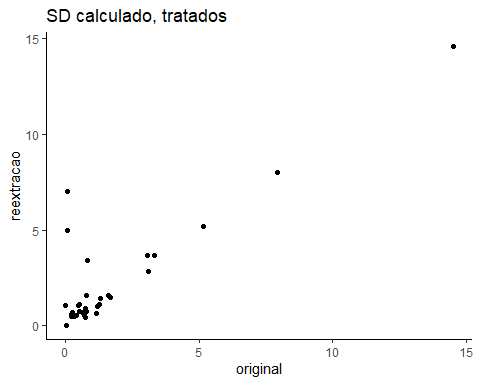


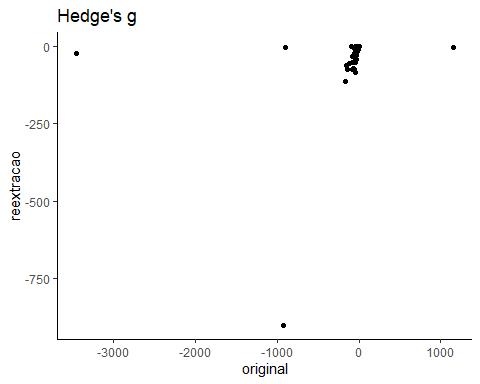


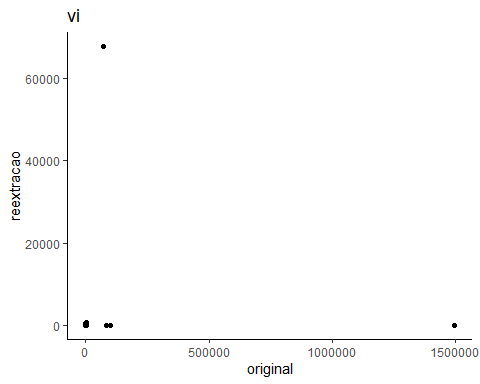












*OBS:* Uma mesma comparacao segue como outlier em ES e vi.

Além disso, durante a limpeza de n\_definition, notei erros na extracao dos valores de n. De 11 artigos revisados inicialmente, 5 estavam sem o valor de n extraído sendo que o artigo reporta médias de replicatas e informa o número de replicatas.

Devemos decidir se vale revisar todos os casos sem n:

Total de comparacoes a revisar: 204

Total de artigos: 63