Theoretical Bias and variance, as a function of population parameters

Delacre Marie[1]

[1] ULB

Theoretical Bias and variance, as a function of population parameters

The bias

For all estimators, when the population effect size is null so is the bias. We will subdivise all configurations when there is a non-null population effect into 3 conditions:

- when variances are equal across groups,

- when variances are unequal across groups, with equal sample sizes

- when variances are unequal across groups, with unequal sample sizes

Cohen's d_s

When variances are equal across populations. The **bias** of Cohen's d_s is a function of total sample size (N) and the population effect size (δ_{Cohen}):

- The larger the population effect size, the more Cohen's d_s will overestimate δ_{Cohen}.
- The larger the total sample size, the lower the bias. The bias tends to zero when the total sample size tends to infinity (see Figure 1)

Of course, considering the degrees of freedom, the sample size ratio does not matter… (see Figure 2)

Glass's d_s

Because degrees of freedom depend only on the sample size of the control group, there is no need to distinguish between cases where there is homoscedasticity or heteroscedasticity!

The **bias** of Glass's d_s is a function of the sample size of the control group (n_c) and the population effect size (δ_{glass}):

- The larger the population effect size, the more Glass's d_s will overestimate δ_{Glass}.

- The larger the size of the control group, the lower the bias. The bias tends to zero when the sample size of the control group tends to infinity (see Figure 3)

Cohen's d'_s

When variances are equal across populations. When $\sigma_1 = \sigma_2 = \sigma$:

$$df_{Cohen's\ d'_s} = \frac{(n_1-1)(n_2-1)(2\sigma^2)^2}{(n_2-1)\sigma^4 + (n_1-1)\sigma^4} = \frac{(n_1-1)(n_2-1)\times 4\sigma^4}{\sigma^4(n_1+n_2-2)} = \frac{4(n_1-1)(n_2-1)}{n_1+n_2-2}$$

One can see that degrees of freedom depend only on the total sample size (N) and the sample size allocation ratio. As a consequence, the **bias** of Cohen's d'_s is a function of the population effect size (δ'_{Cohen}), the sample size allocation ratio and the total sample size (N).

- The larger the population effect size, the more $Cohen's\ d'_s$ will overestimate δ'_{Cohen}
- The further the sample size allocation ratio is from 1, the larger the bias (see Figure 4)
- The larger the total sample size, the lower the bias (see Figure 5)

When variances are unequal across populations, with equal sample sizes. When $n_1 = n_2 = n$:

$$df_{Cohen's\ d'_s} = \frac{(n-1)^2(\sigma_1^2+\sigma_2^2)^2}{(n-1)(\sigma_1^4+\sigma_2^4)} = \frac{(n-1)(\sigma_1^4+\sigma_2^4+2\sigma_1^2\sigma_2^2)}{\sigma_1^4+\sigma_2^4}$$

One can see that degrees of freedom depend only on the total sample size (N) and the SD-ratio. As a consequence, the **bias** of Cohen's d'_s is a function of the population effect size (δ'_{Cohen}), the SD-ratio and the total sample size (N):

- The larger the population effect size, the more $Cohen's\ d'_s$ will overestimate δ'_{Cohen}
- The further the SD-ratio is from 1, the larger the bias (see Figure 6)
- The larger the total sample size, the lower the bias (see Figure 7)

Note: for a constant SD-ratio, the size of the variance does not matter (see Figure 8)

When variances are unequal across populations, with unequal sample sizes. The **bias** of Cohen's d'_s is a function of the population effect size (δ'_{Cohen}), the total sample size, and the sample sizes and variances pairing :

- The larger the population effect size, the more $Cohen's\ d'_s$ will overestimate δ'_{Cohen}
- When there is a positive pairing between sample sizes and variances, one gives more weight to the smallest variance. As a consequence, the denominator in the df computation decreases, the degrees of freedom increase and therefore, the bias decreases (see the two plots on the top in Figure 9). On the other size, where there is a negative pairing between sample sizes and variances, one gives more weight to the largest variance. As a consequence, the denominator in the df computatin increases, the degrees of freedom decrease and therefore, the bias increase (see the two plots in the bottom of Figure 9).
- The larger the total sample size, the lower the bias (illustration in Figure 10)

Note: for a constant SD-ratio, the variance does not matter. (See Figure 11)

Shieh's d_s

When variances are equal across populations. When $\sigma_1 = \sigma_2 = \sigma$:

$$df_{Shieh's\ d_s} = \frac{\left(\frac{n_2\sigma^2+n_1\sigma^2}{n_1 n_2}\right)^2}{\frac{(n_2-1)\left(\frac{\sigma^2}{n_1}\right)^2+(n_1-1)\left(\frac{\sigma^2}{n_2}\right)^2}{(n_1-1)(n_2-1)}}$$

$$\leftrightarrow df_{Shieh's\ d_s} = \frac{[\sigma^2(n_1+n_2)]^2}{n_1^2 n_2^2} \times \frac{(n_1-1)(n_2-1)}{(n_2-1)\times\frac{\sigma^4}{n_1^2}+(n_1-1)\times\frac{\sigma^4}{n_2^2}}$$

$$\leftrightarrow df_{Shieh's\ d_s} = \frac{\sigma^4 N^2}{n_1^2 n_2^2} \times \frac{(n_1-1)(n_2-1)}{\sigma^4\left(\frac{n_2-1}{n_1^2}+\frac{n_1-1}{n_2^2}\right)}$$

$$\leftrightarrow df_{Shieh's\ d_s} = \frac{N^2(n_1-1)(n_2-1)}{n_1^2 n_2^2\left(\frac{n_2^2(n_2-1)+n_1^2(n_1-1)}{n_1^2 n_2^2}\right)}$$

$$\leftrightarrow df_{Shieh's\ d_s} = \frac{N^2(n_1-1)(n_2-1)}{n_2^2(n_2-1)+n_1^2(n_1-1)}$$

One can see that degrees of freedom depend only on the total sample size (N) and the sample size allocation ratio. As a consequence, the **bias** of Shieh's d'_s is a function of the population effect size (δ_{Shieh}), the sample size allocation ratio and the total sample size (N).

- The larger the population effect size, the more $Shieh's\ d_s$ will overestimate δ_{Shieh}
- The further the sample size allocation ratio is from 1, the larger the bias (see Figure 12)
- The larger the total sample size, the lower the bias (see Figure 13)

When variances are unequal across populations, with equal sample sizes. When $n_1 = n_2 = n$:

$$df_{Shieh's\ d_s} = \frac{\left(\frac{\sigma_1^2+\sigma_2^2}{n}\right)^2}{\frac{(\sigma_1^2/n)^2+(\sigma_2^2/n)^2}{n-1}}$$

$$df_{Shieh's\ d_s} = \frac{(\sigma_1^2+\sigma_2^2)^2}{n^2} \times \frac{n-1}{\frac{\sigma_1^4+\sigma_2^4}{n^2}}$$

$$df_{Shieh's\ d_s} = \frac{(\sigma_1^2+\sigma_2^2)^2 \times (n-1)}{\sigma_1^4+\sigma_2^4}$$

One can see that degrees of freedom depend on the total sample size (N), the SD-ratio. As a consequence, the bias depends on the population effect size (δ_{Shieh}), the SD-ratio and the total sample size (N).

- The larger the population effect size, the more $Shieh's\ d_s$ will overestimate δ_{Shieh}
- The further the SD-ratio is from 1, the larger the bias (see Figure 14)
- The larger the total sample size, the lower the bias (see Figure 15)

Note: for a constant SD-ratio, the size of the variance does not matter (see Figure 16)

When variances are unequal across populations, with unequal sample sizes. The **bias** of Shieh's d'_s is a function of the population effect size (δ_{Shieh}), the sample sizes (n_1 and n_2), and the pairing between sample sizes and variances and sample sizes ratios.

- The larger the population effect size, the more $Shieh's\ d_s$ will overestimate δ_{Shieh}
- The larger the sample sizes, the lower the bias (illustration in Figure 17)
- The variances and sample sizes ratios don't matter per se (see Figure 18). However, the pairing between these ratios and sample sizes has an effect on the bias:
  - When $\frac{\sigma_1^2}{n_1} = \frac{\sigma_2^2}{n_2}$, the smallest bias occurs when sample sizes are equal across groups. The further the sample sizes ratio is from 1, the larger the bias (see Figure 19).
  - When $\frac{\sigma_1^2}{n_1} \neq \frac{\sigma_2^2}{n_2}$, the minimum bias will always occure when $min(\frac{\sigma_j^2}{n_j})$ will be associated with $min(n_j)$. In other word, when $\frac{\sigma_1^2}{n_1} > \frac{\sigma_2^2}{n_2}$, the sample sizes ratio associated with the minimum bias will be positive, meaning that $n_1 > n_2$ (and the larger the difference between $\frac{\sigma_1^2}{n_1}$ and $\frac{\sigma_2^2}{n_2}$, the further from 1 will be this sample sizes ratio; see the two top plots in Figure 20). On the other side, when $\frac{\sigma_1^2}{n_1} < \frac{\sigma_2^2}{n_2}$, the sample sizes ratio associated with the minimum bias will be negative, meaning that $n_1 < n_2$ (and the larger the difference between $\frac{\sigma_1^2}{n_1}$ and $\frac{\sigma_2^2}{n_2}$, the further from 1 will be this sample sizes ratio; see the two bottom plots in Figure 20).

Moreover, for a constant SD-ratio, the variances don't matter either. (See Figure 21)

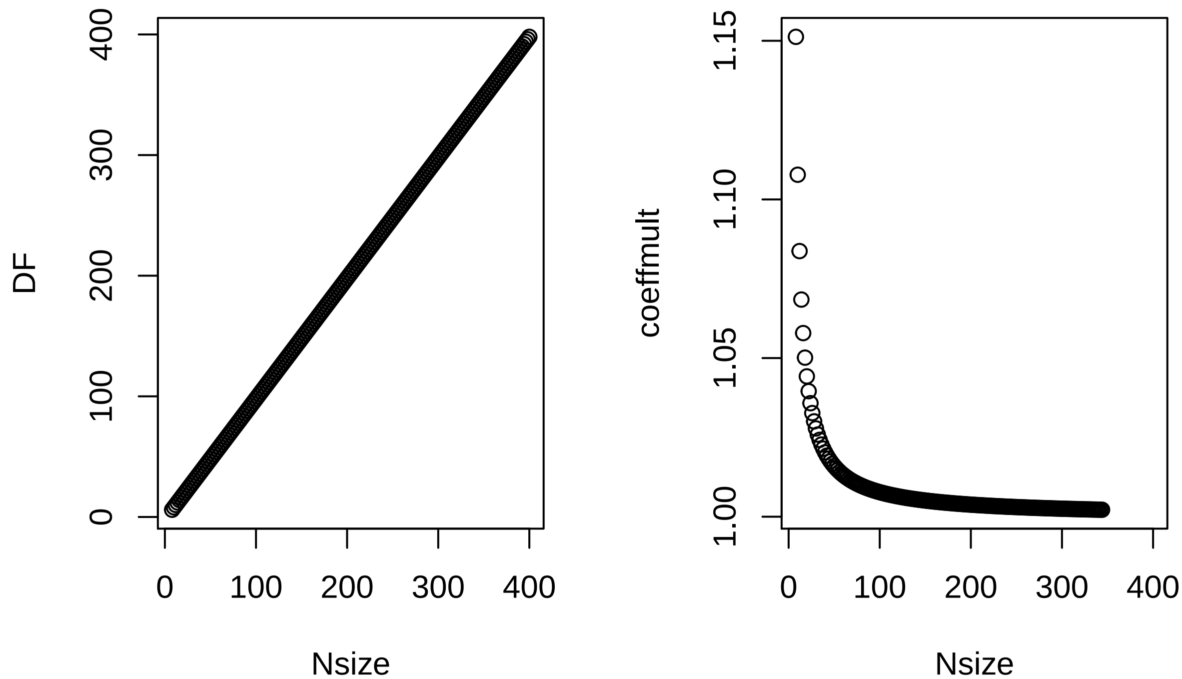
DF
0
100
200
300
400
Nsize
0
100
200
300
400
coeffmult
1.00
1.05
1.10
1.15
Nsize
0
100
200
300
400

Figure 1. . . .

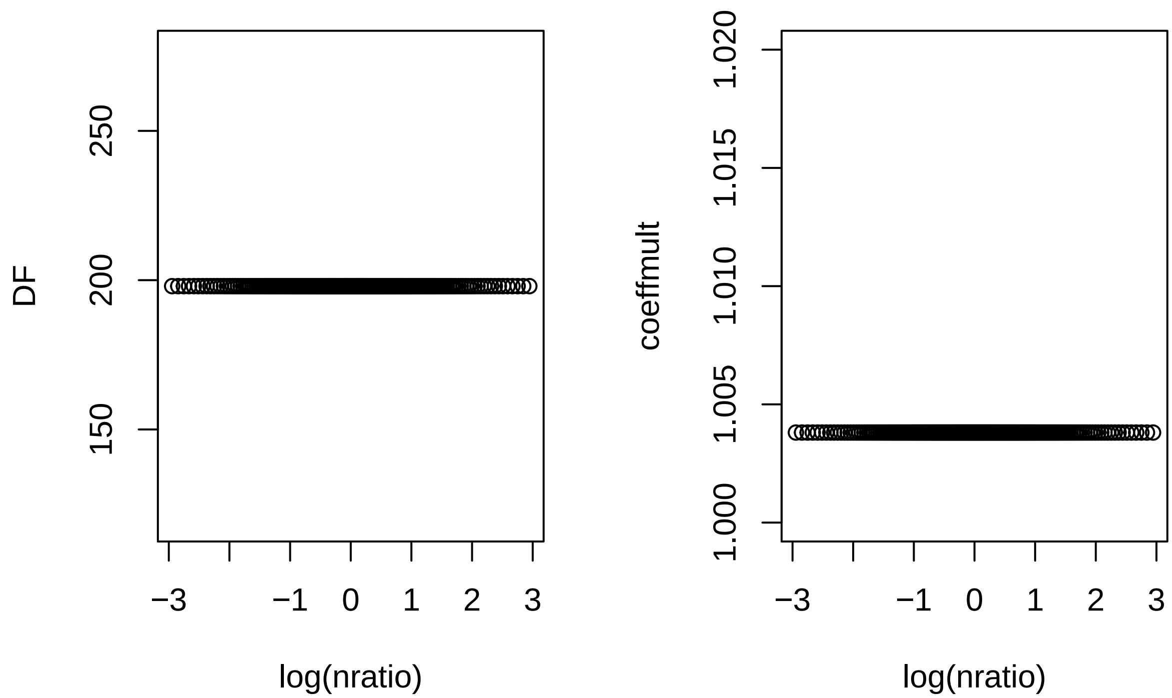
DF
250
200
150
−3
−1
0
1
2
3
log(nratio)
coeffmult
1.020
1.015
1.010
1.005
1.000
−3
−1
0
1
2
3
log(nratio)

Figure 2. ...

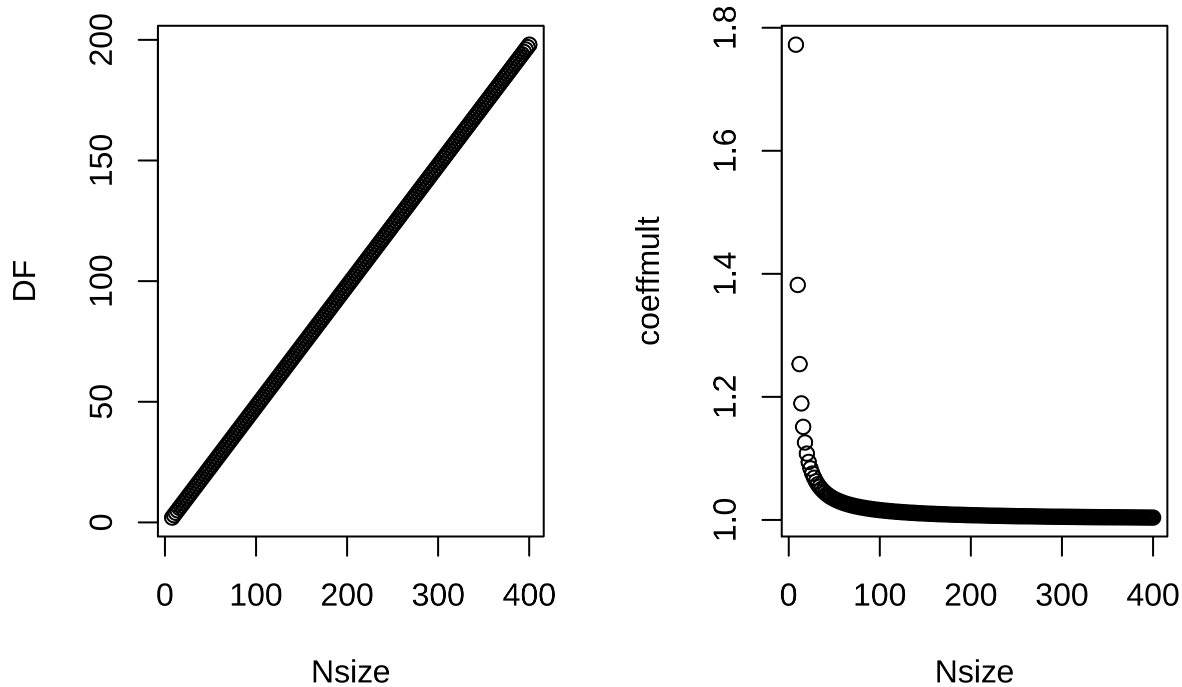
DF
0
50
100
150
200
0
100
200
300
400
Nsize
coeffmult
1.0
1.2
1.4
1.6
1.8
0
100
200
300
400
Nsize

Figure 3. . . .

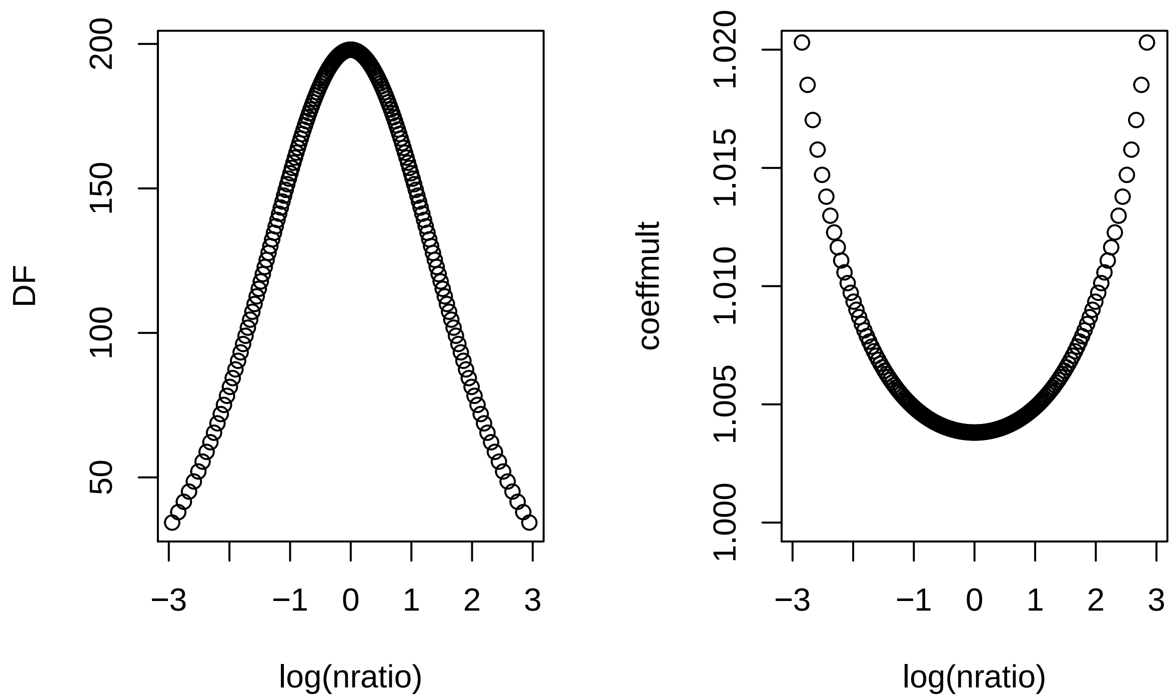
DF
200
150
100
50
−3
−1
0
1
2
3
log(nratio)
coeffmult
1.020
1.015
1.010
1.005
1.000
−3
−1
0
1
2
3
log(nratio)

Figure 4. . . .

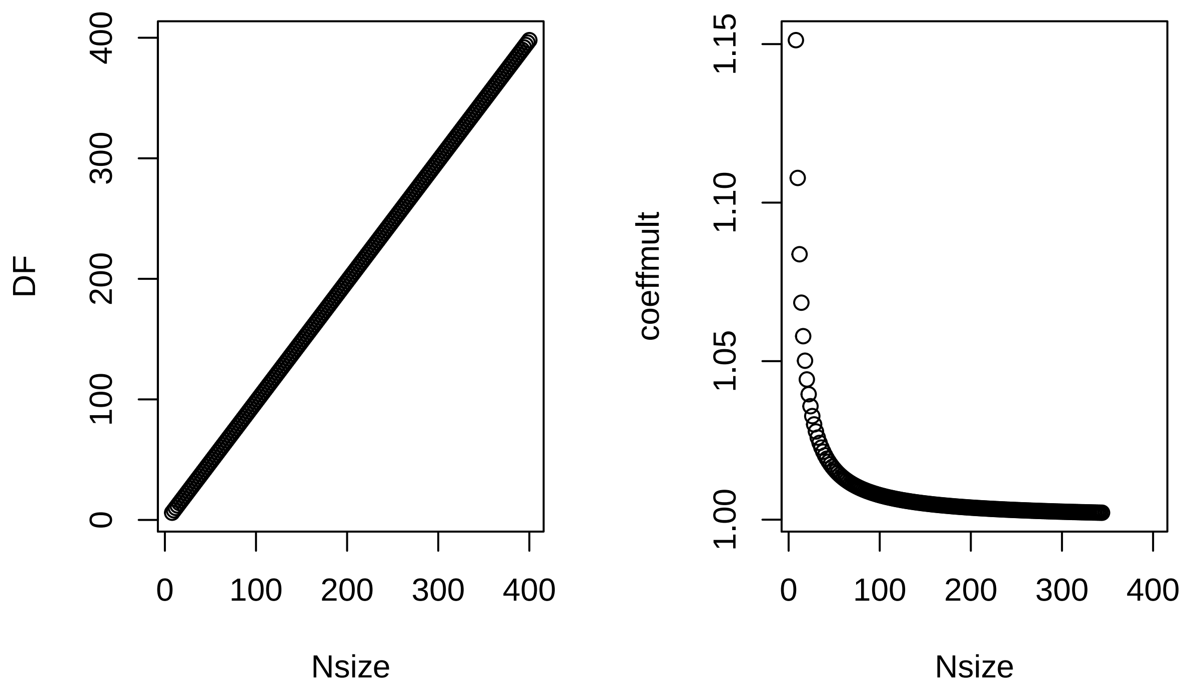
DF
0
100
200
300
400
Nsize
0
100
200
300
400
coeffmult
1.00
1.05
1.10
1.15
Nsize
0
100
200
300
400

Figure 5. ...

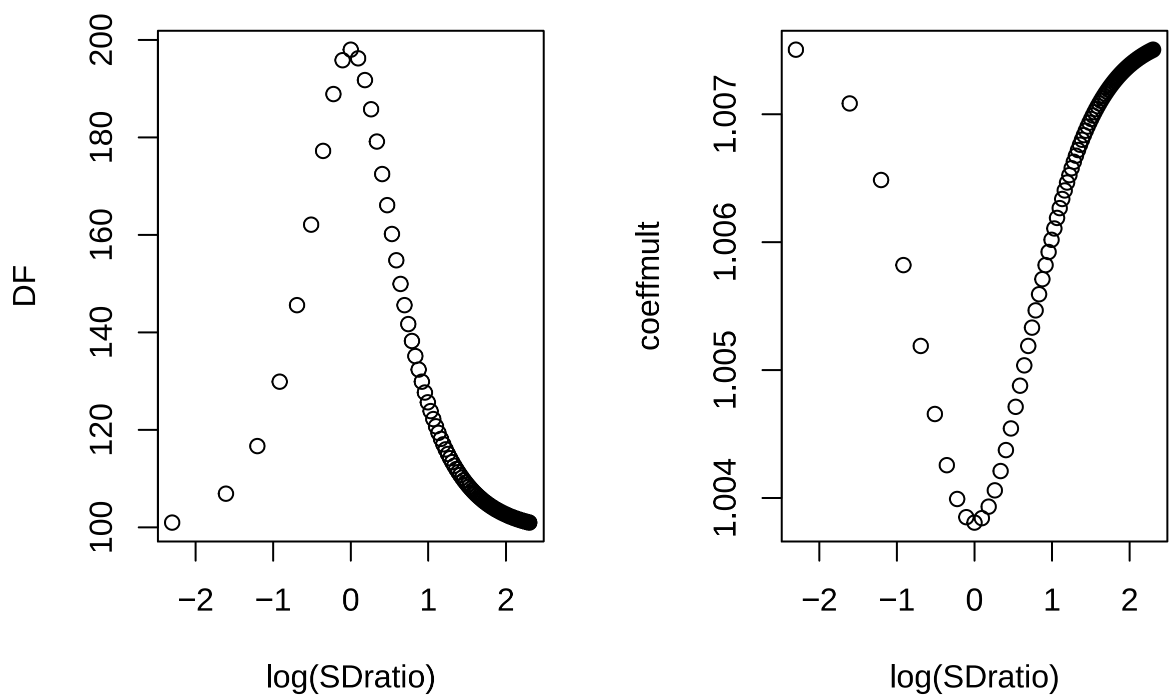
DF
200
180
160
140
120
100
−2
−1
0
1
2
log(SDratio)
coeffmult
1.007
1.006
1.005
1.004
−2
−1
0
1
2
log(SDratio)

Figure 6. ...

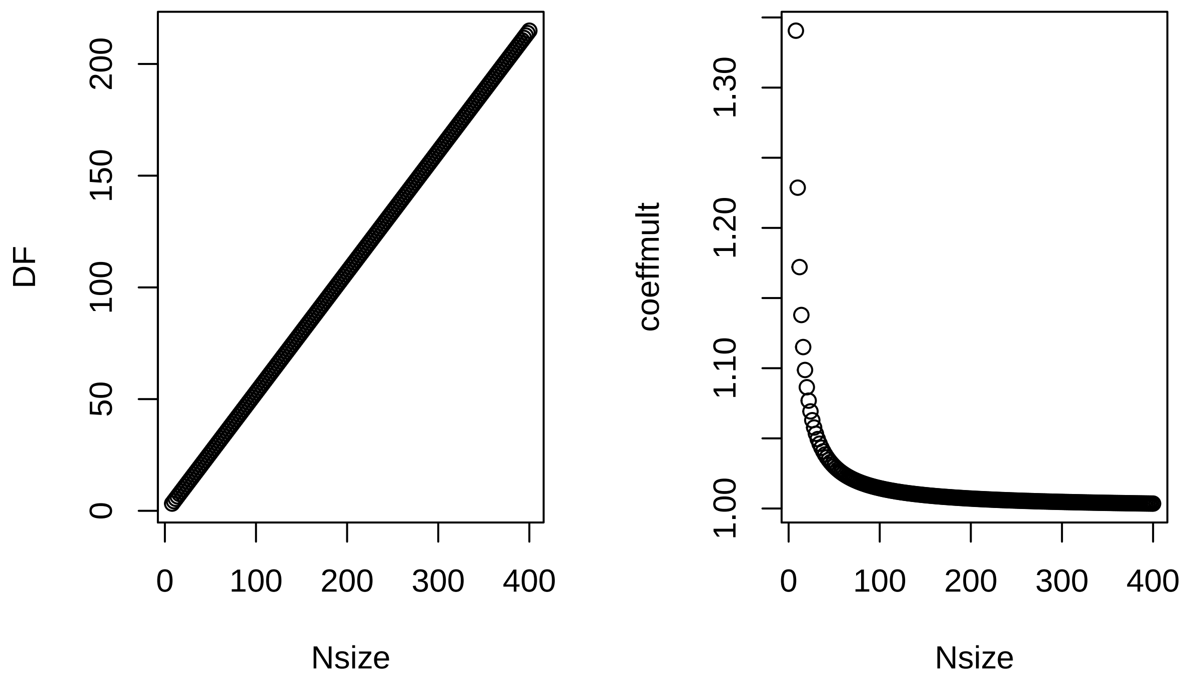
DF
0
50
100
150
200
0
100
200
300
400
Nsize
coeffmult
1.00
1.10
1.20
1.30
0
100
200
300
400
Nsize

Figure 7. ...

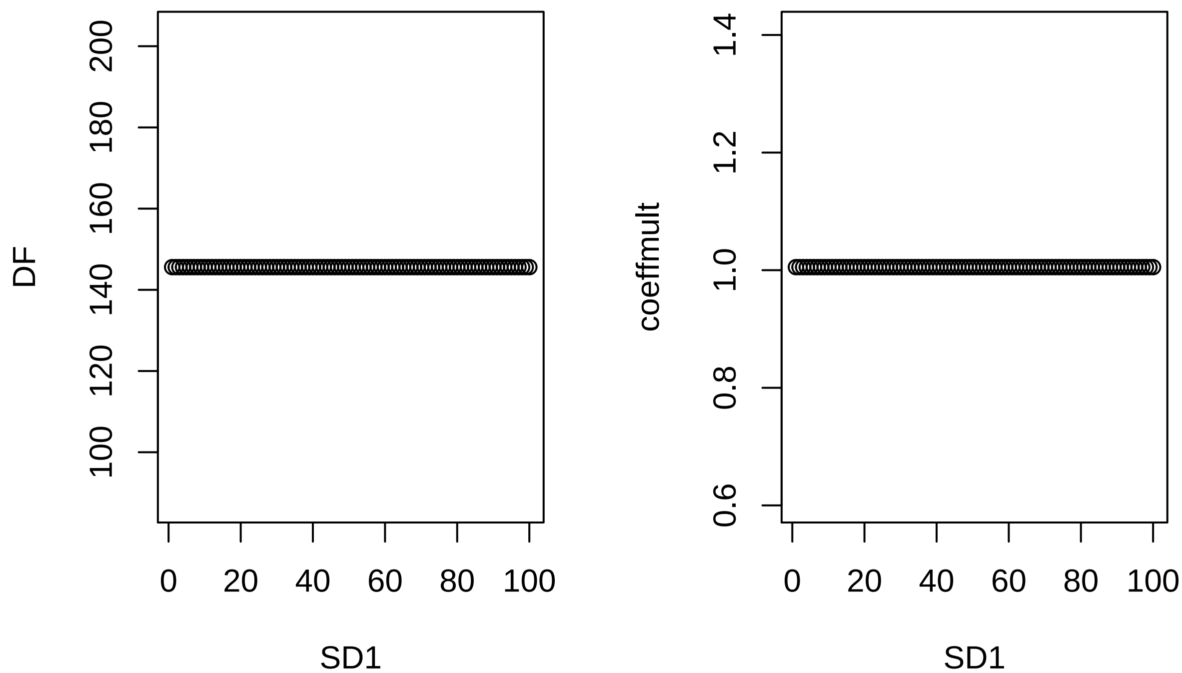
DF
200
180
160
140
120
100
0
20
40
60
80
100
SD1
coeffmult
1.4
1.2
1.0
0.8
0.6
0
20
40
60
80
100
SD1

Figure 8. ...

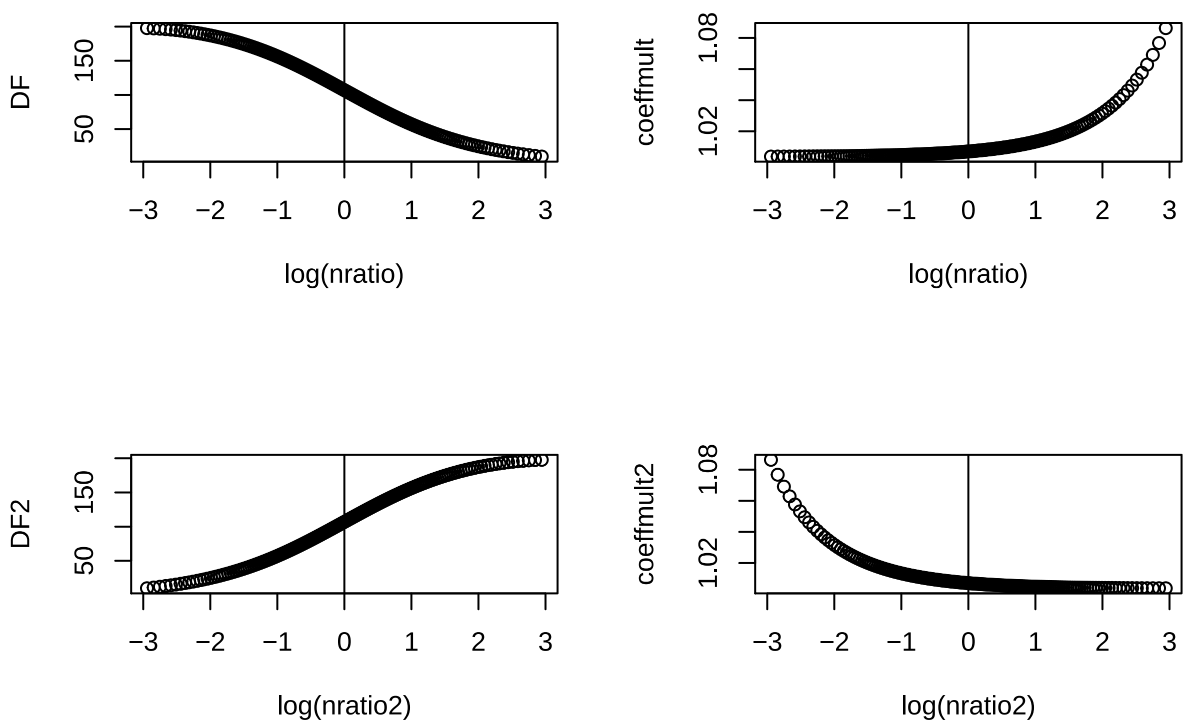
DF
150
50
−3
−2
−1
0
1
2
3
log(nratio)
coeffmult
1.08
1.02
log(nratio)
DF2
log(nratio2)
coeffmult2
log(nratio2)

Figure 9. ...

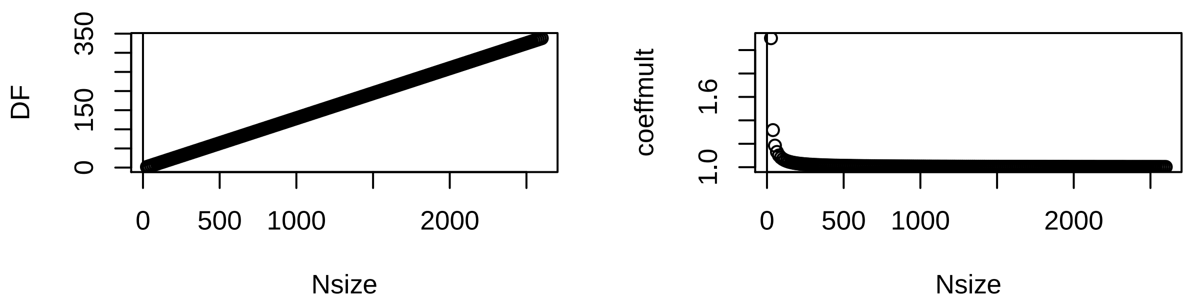
DF
0
150
350
0
500
1000
2000
Nsize
coeffmult
1.0
1.6
0
500
1000
2000
Nsize

Figure 10. ...

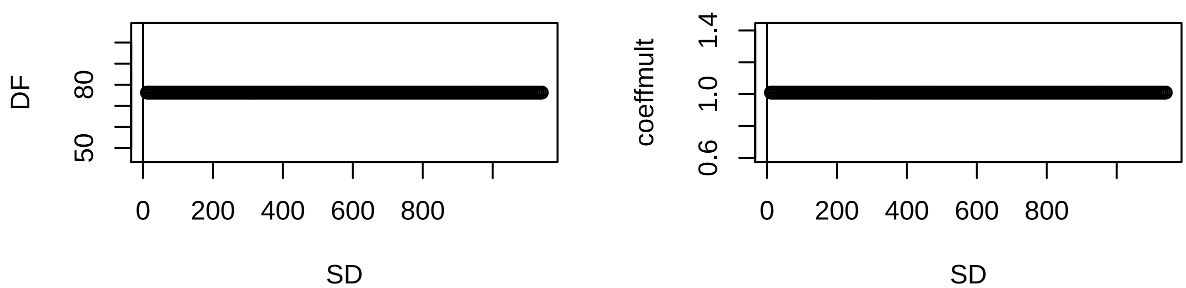
DF
50
80
0
200
400
600
800
SD
coeffmult
0.6
1.0
1.4
0
200
400
600
800
SD

Figure 11. . . .

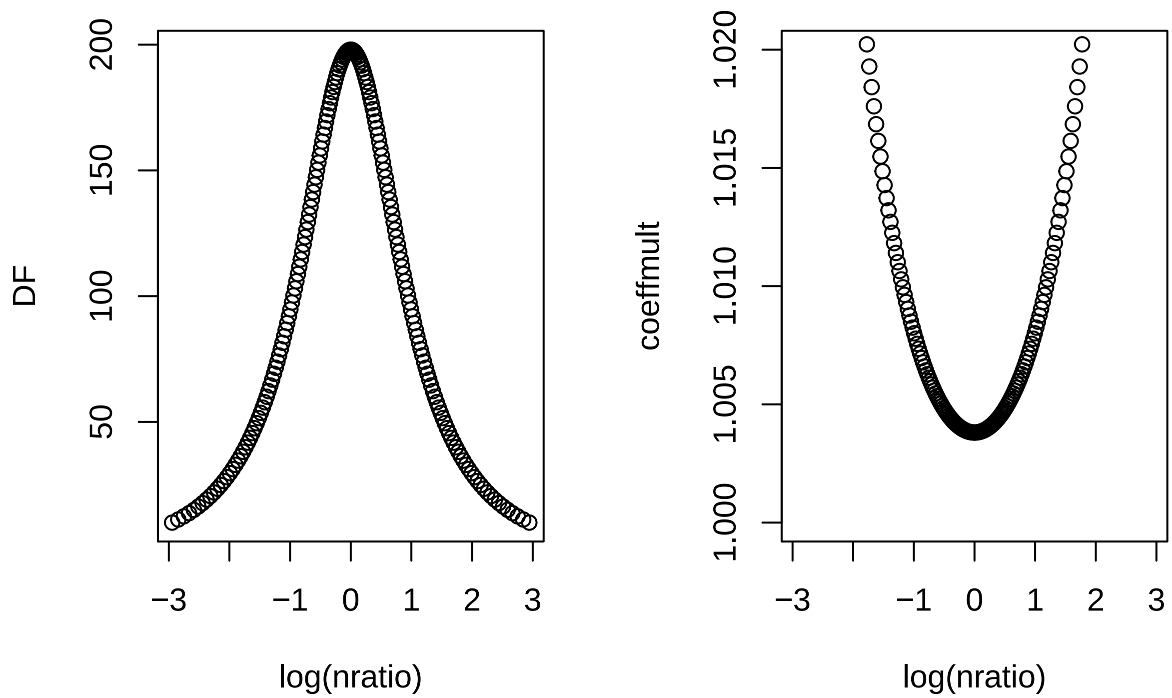
DF
200
150
100
50
−3
−1
0
1
2
3
log(nratio)
coeffmult
1.020
1.015
1.010
1.005
1.000
−3
−1
0
1
2
3
log(nratio)

Figure 12. . . .

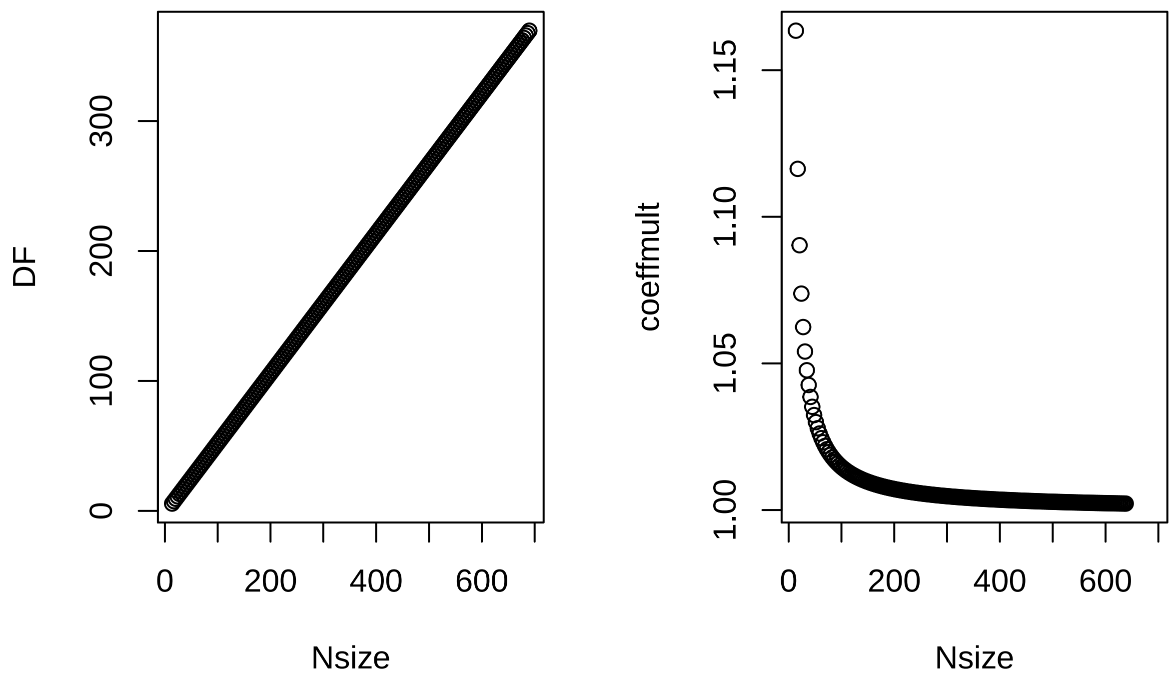
DF
0
100
200
300
0
200
400
600
Nsize
coeffmult
1.00
1.05
1.10
1.15
0
200
400
600
Nsize

Figure 13. ...

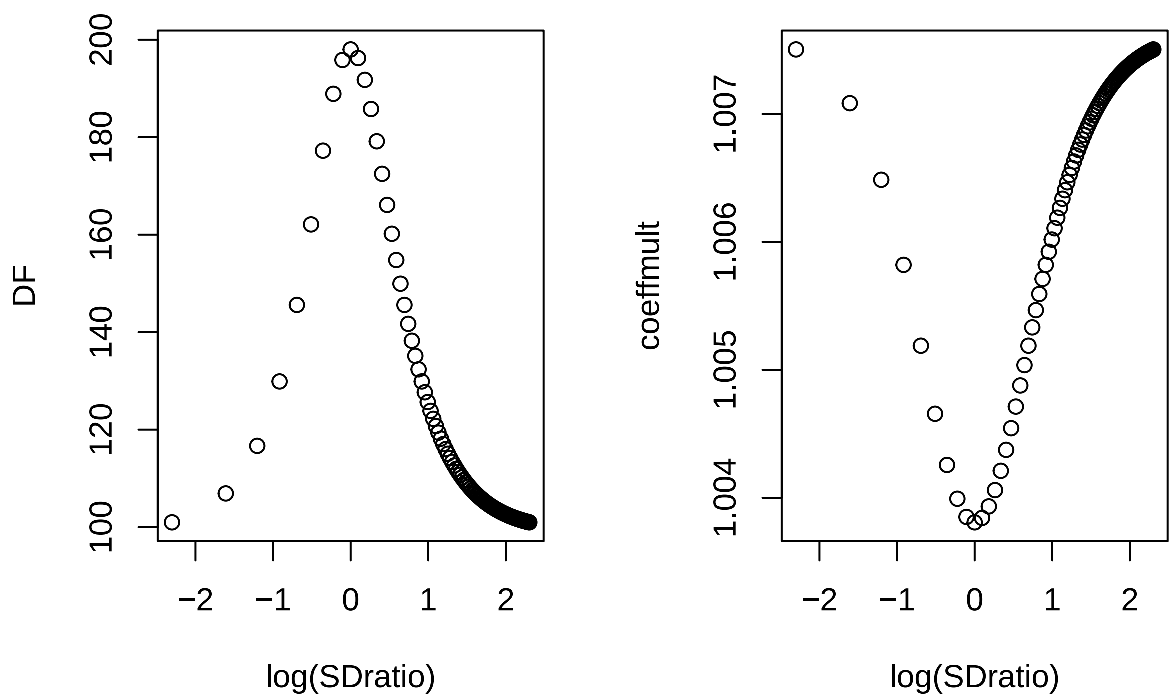
DF
200
180
160
140
120
100
−2
−1
0
1
2
log(SDratio)
coeffmult
1.007
1.006
1.005
1.004
−2
−1
0
1
2
log(SDratio)

Figure 14. . . .

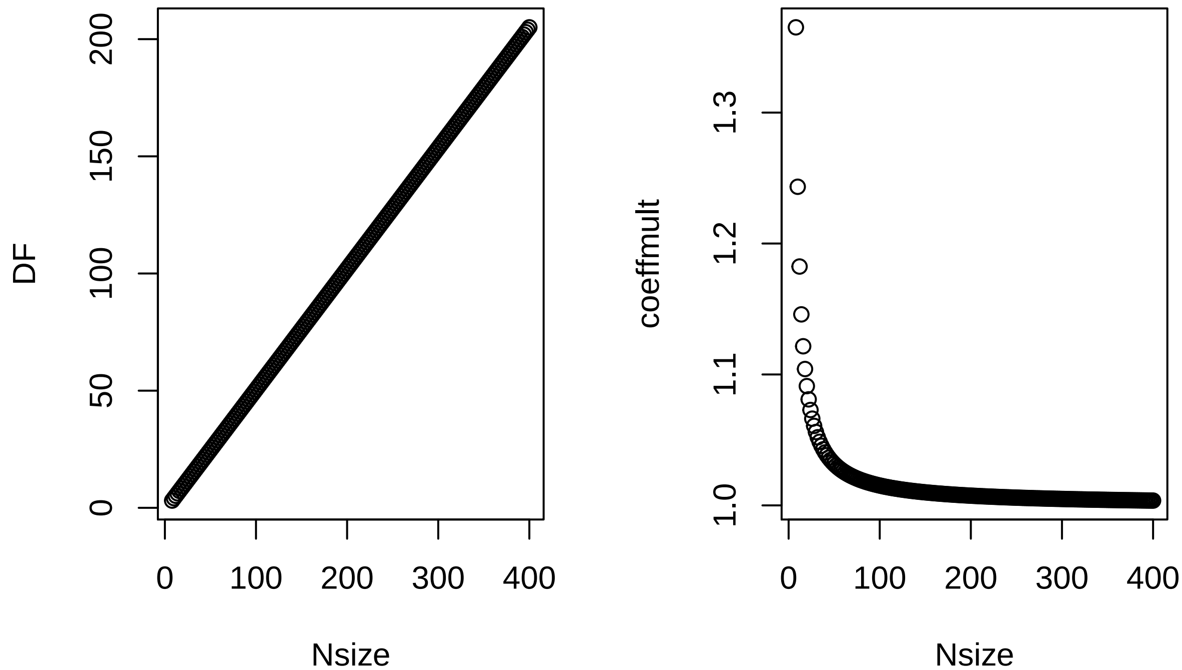

Figure 15. bias of Cohen's d'_s as a function of the total sample size, when variances are equal across groups

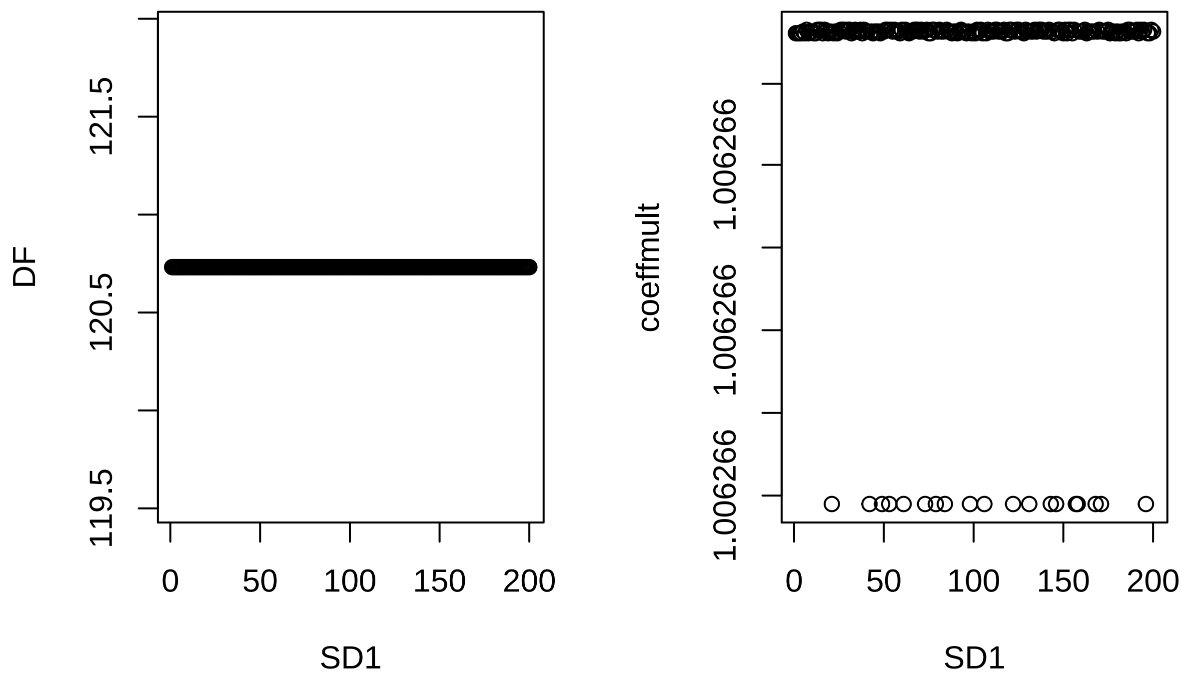
DF
121.5
120.5
119.5
0
50
100
150
200
SD1
coeffmult
1.006266
1.006266
1.006266
0
50
100
150
200
SD1

Figure 16. ;;;

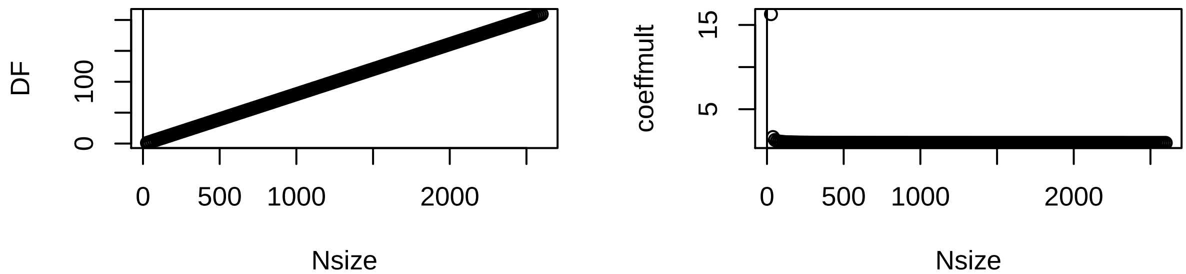
DF
0
100
0
500
1000
2000
Nsize
coeffmult
5
15
0
500
1000
2000
Nsize

Figure 17. ...

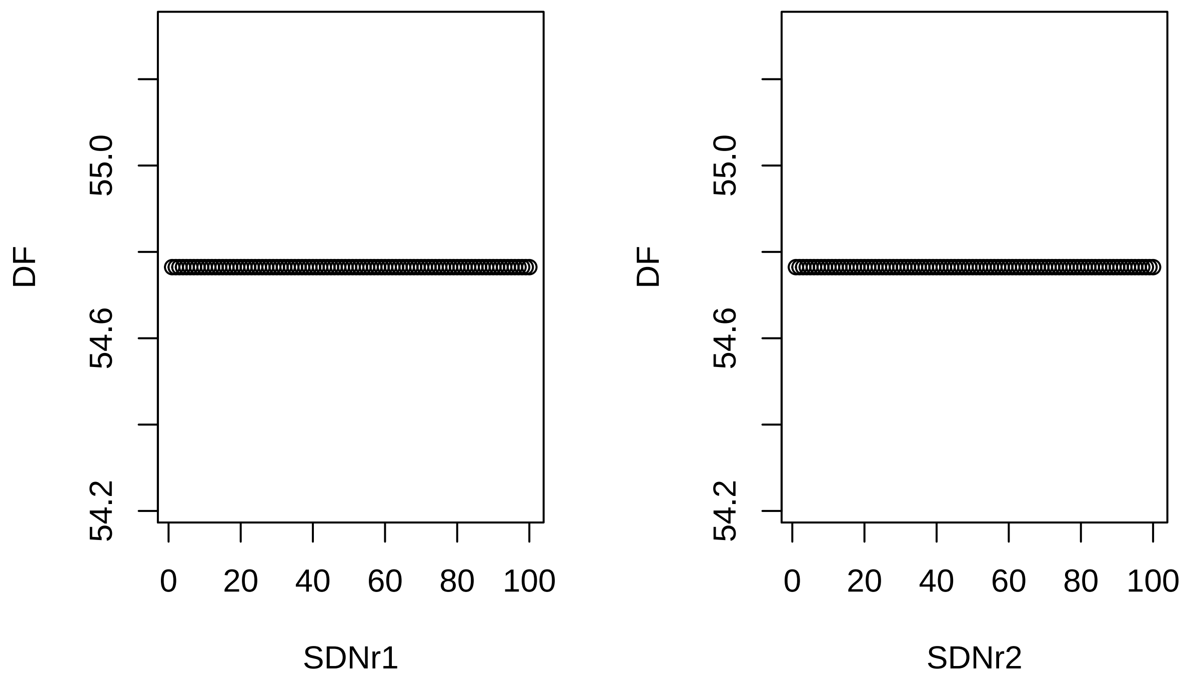
DF
55.0
54.6
54.2
0
20
40
60
80
100
SDNr1
DF
55.0
54.6
54.2
0
20
40
60
80
100
SDNr2

Figure 18. ...

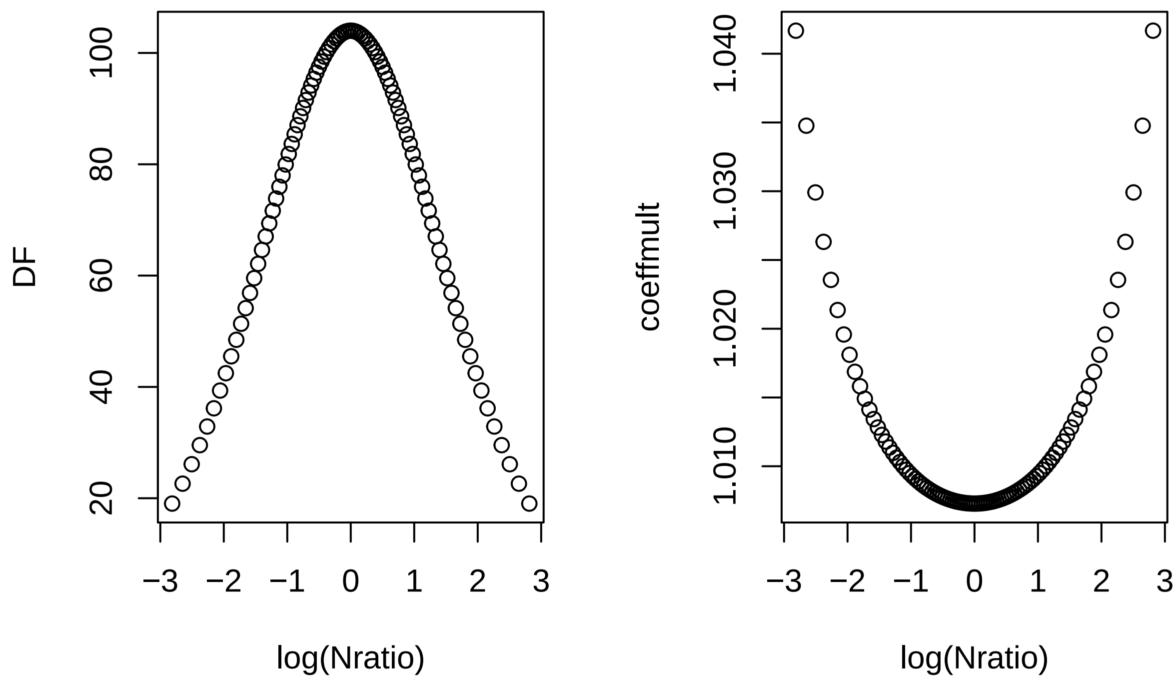
DF
100
80
60
40
20
−3 −2 −1 0 1 2 3
log(Nratio)
coeffmult
1.040
1.030
1.020
1.010
−3 −2 −1 0 1 2 3
log(Nratio)

Figure 19. ...

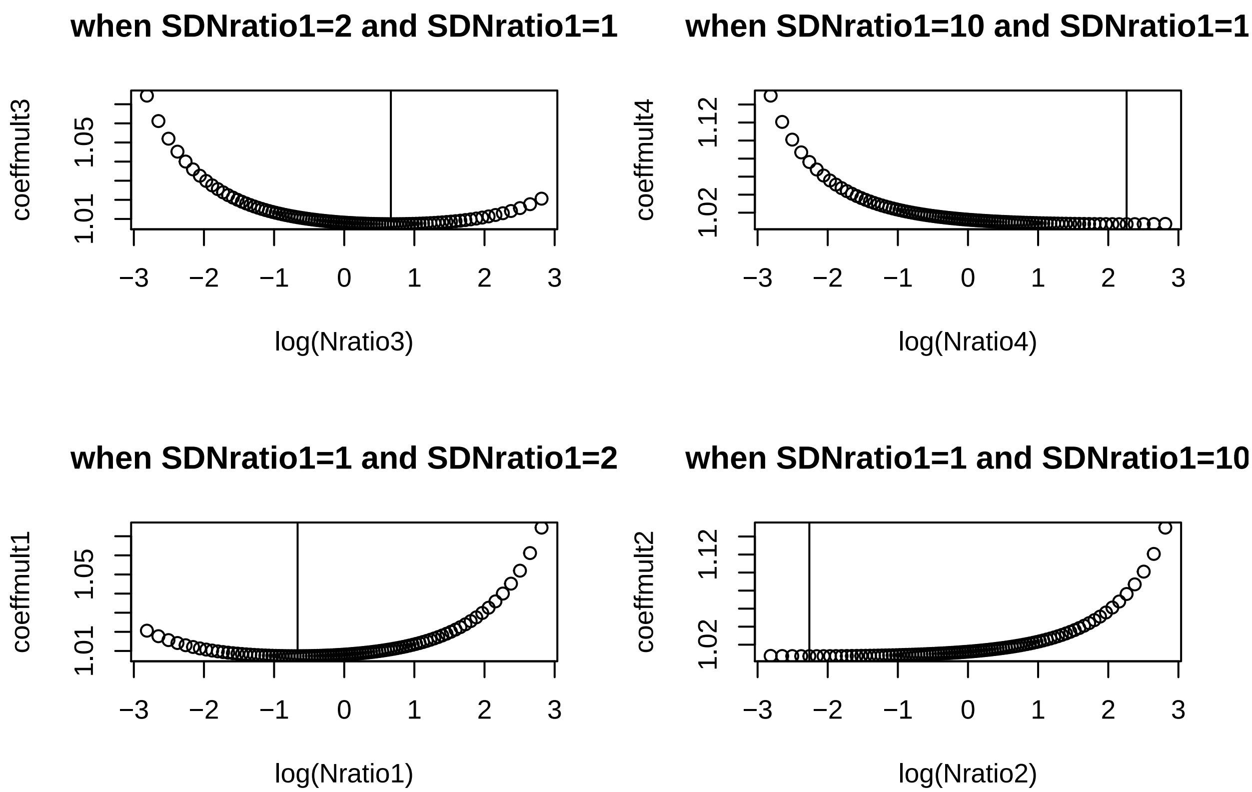
when SDNratio1=2 and SDNratio1=1
coeffmult3
1.01
1.05
−3 −2 −1 0 1 2 3
log(Nratio3)
when SDNratio1=10 and SDNratio1=1
coeffmult4
1.02
1.12
−3 −2 −1 0 1 2 3
log(Nratio4)
when SDNratio1=1 and SDNratio1=2
coeffmult1
1.01
1.05
−3 −2 −1 0 1 2 3
log(Nratio1)
when SDNratio1=1 and SDNratio1=10
coeffmult2
1.02
1.12
−3 −2 −1 0 1 2 3
log(Nratio2)

Figure 20. ...

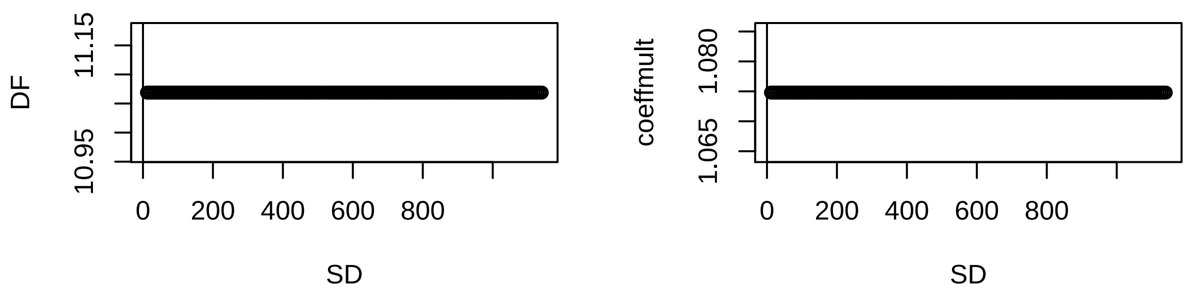
DF
11.15
10.95
0
200
400
600
800
SD
coeffmult
1.080
1.065
0
200
400
600
800
SD

Figure 21. . . .