

¹ Theoretical Bias, as a function of population parameters

² Delacre Marie¹

³ ¹ ULB

Theoretical Bias, as a function of population parameters

The bias

For all “biased” estimators, when the population effect size is null so is the bias. We will therefore focus on configurations where there is a non-null population effect size. The sampling distribution of Cohen’s d_s (and therefore its bias) is only known under the assumptions of normality and homoscedasticity. On the other side, the biases of Glass’s d_s , Cohen’s d'_s and Shieh’s d_s are theoretically known for all configurations where the normality assumption is met, whatever variances are equal across groups or not. In order to simplify the analysis of their bias, it is convenient to subdivide all configurations 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.

Preliminary note

For all previously mentioned estimators (Cohen’s d_s , Glass’s d_s , Cohen’s d'_s and Shieh’s d_s), the theoretical expectancy is computed by multiplying the population effect size by the following multiplier coefficient:

$$\gamma = \frac{\sqrt{\frac{df}{2}} \times \Gamma \frac{df-1}{2}}{\Gamma \frac{df}{2}} \quad (1)$$

Where df are the degrees of freedom. γ is *always* positive, meaning that when the population effect size is not zero, all estimators will overestimate the real population effect size. Moreover, its limit tends to 1 when the degrees of freedom (df) tend to infinity, meaning that the larger the degrees of freedom, the lower the bias.

While we focus on the theoretical bias of biased estimators (Cohen’s d_s , Glass’s d_s , Shieh’s d_s and Cohen’s d'_s) when the normality assumption is met, it is interesting to notice that our main conclusions seem to generalize to :

- biased estimators when samples are extracted from symmetric distributions;
- unbiased estimators when samples are extracted from heavy-tailed symmetric distributions.

Cohen's d_s (see Table 1)

Under the assumptions that independant residuals are normally distributed with equal variances, 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 (see Figure 1);
- Of course, considering the degrees of freedom, the sample size ratio does not matter (i.e. the bias will decrease, whatever one increases n_1 , n_2 or both sample sizes).

Glass's d_s (see Table 2)

Because degrees of freedom depend only on the control group size (neither on σ_1 nor on σ_2), 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 control group size (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 (see the two top plots in Figure 2). On the other side, increasing the experimental group size does not impact the bias (see the two bottom plots in Figure 2).

Cohen's d'_s (see Table 2)

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 $\left(\frac{n_1}{n_2}\right)$. 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 $\left(\frac{n_1}{n_2}\right)$ 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 3);
- The larger the total sample size, the lower the bias (see Figure 4).

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 $\left(\frac{\sigma_1}{\sigma_2}\right)$. As a consequence, the **bias** of Cohen's d'_s is a function of the population effect size (δ'_{Cohen}), the *SD*-ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$ 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 5);
- The larger the total sample size, the lower the bias (see Figure 6).

Note: for a constant *SD*-ratio, σ_1 and σ_2 don't matter (see Figure 7).

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 (N), and the interaction between the sample sizes ratio and the SD -ratio ($\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}$):

- 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 (see Figure 8);
- The smallest bias always occurs when there is a positive pairing between variances and sample sizes, because one gives more weight to the smallest variance, in the denominator of the df computation. Moreover, the further the SD -ratio is from 1, the further from 1 will also be the sample sizes ratio associated with the smallest bias (see Figure 9). This can be explained by splitting the numerator and the denominator in the df computation.

As illustrated in Figure 10, for any SD -ratio, the numerator of the degrees of freedom will be maximized when sample sizes are equal across groups (and is not impacted by the SD -ratio). On the other side, the denominator will be minimized when there is a positive pairing between variances and sample sizes. For example, when $\sigma_1 > \sigma_2$, the smallest denominator occurs when $\frac{n_1}{n_2} = \max\left(\frac{n_1}{n_2}\right)$ and the larger the SD -ratio, the larger the impact of the sample sizes ratio on the denominator.

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

Shieh's d_s (see Table 2)

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

$$df_{Shieh's d_s} = \frac{\left(\frac{n_2\sigma^2 + n_1\sigma^2}{n_1n_2}\right)^2}{\frac{(n_2-1)\left(\frac{\sigma^2}{n_1}\right)^2 + (n_1-1)\left(\frac{\sigma^2}{n_2}\right)^2}{(n_1-1)(n_2-1)}}$$

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

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

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

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

One can see that degrees of freedom depend only on the total sample size (N) and the sample size allocation ratio $\left(\frac{n_1}{n_2}\right)$. 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 $\left(\frac{n_1}{n_2}\right)$ 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);
- For a constant sample sizes ratio, the larger the total sample size, the lower the bias (see Figure 13).

Note: when computing Cohen's d'_s , degrees of freedom increased when adding subjects in either one or both groups, even when the sample size ratio increased. When computing Shieh's d_s , this is not true anymore: there is a larger impact of the sample sizes ratio, in order that increasing the sample sizes ratio when adding subjects in only one group can decrease the degrees of freedom and therefore, increase the bias (See Figure 14).

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) and the SD -ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$. As a consequence, the bias depends on the population effect size (δ_{Shieh}), the SD -ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$ 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 15);
- The larger the total sample size, the lower the bias (see Figure 16);

Note: for a constant SD -ratio, σ_1 and σ_2 don't matter (see Figure 17).

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 interaction between the sample sizes ratio and the SD -ratio $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$:

- The larger the population effect size, the more *Shieh's* d_s will overestimate δ_{Shieh} ;
- For a constant sample sizes ratio, the larger the sample sizes, the lower the bias (See Figure 18);

Note: When variances were equal across populations, adding subjects only in the first group had the same impact on degrees of freedom (and therefore on bias) than adding subjects only in the second group (see Figure 14). When variances are unequal across groups, this is not true anymore (see Figure 19).

- The smallest bias always occurs when there is a positive pairing between variances and sample sizes. Moreover, the further the SD -ratio is from 1, the further from 1 will also be the sample sizes ratio associated with the smallest bias (See Figure 20);

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

In summary

The **bias** of Cohen's d_s is a function of the population effect size δ_{Cohen} and the total sample size (N):

- When δ_{Cohen} is null, the bias is null. In all other configurations, the larger δ_{Cohen} , the more Cohen's d_s will overestimate δ_{Cohen} ;
- The bias decreases when the total sample size increases (whatever one adds subjects in only one group or in both).

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

- When δ_{Glass} is null, the bias is null. In all other configurations, the larger δ_{Glass} , the more Glass's d_s will overestimate δ_{Glass} ;
- The bias decreases when the size of the control group increases. On the other side, increasing the size of the experimental group does not impact the bias.

The **bias** of Cohen's d'_s is a function of the population effect size (δ'_{Cohen}), the total sample size, and the interaction between the sample sizes ratio and the SD -ratio ($\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}$):

- When δ'_{Cohen} is null, the bias is null. In all other configurations, the larger δ'_{Cohen} , the more Cohen's d'_s will overestimate δ'_{Cohen} ;
- The bias decreases when the total sample size increases (whatever one adds subjects in only one group or in both);
- The smallest bias always occurs when there is a positive pairing between $\frac{\sigma_1}{\sigma_2}$ and $\frac{n_1}{n_2}$.

Moreover, the larger the SD -ratio, the further from 1 is the sample sizes ratio associated with the smallest bias.

The **bias** of Shieh's d'_s is a function of the population effect size (δ_{Shieh}), the total sample size, and the interaction between the sample sizes ratio and the SD -ratio ($\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}$):

- When δ_{Shieh} is null, the bias is null. In all other configurations, the larger δ_{Shieh} , the more Shieh's d_s will overestimate δ_{Shieh} ;
- For a constant sample sizes ratio, the bias decreases when the total sample size increases;
- The smallest bias always occurs when there is a positive pairing between $\frac{\sigma_1}{\sigma_2}$ and $\frac{n_1}{n_2}$. Moreover, the larger the SD -ratio, the further from 1 is the sample sizes ratio associated with the smallest bias (for more details, see "Theoretical Bias, as a function of population parameters").

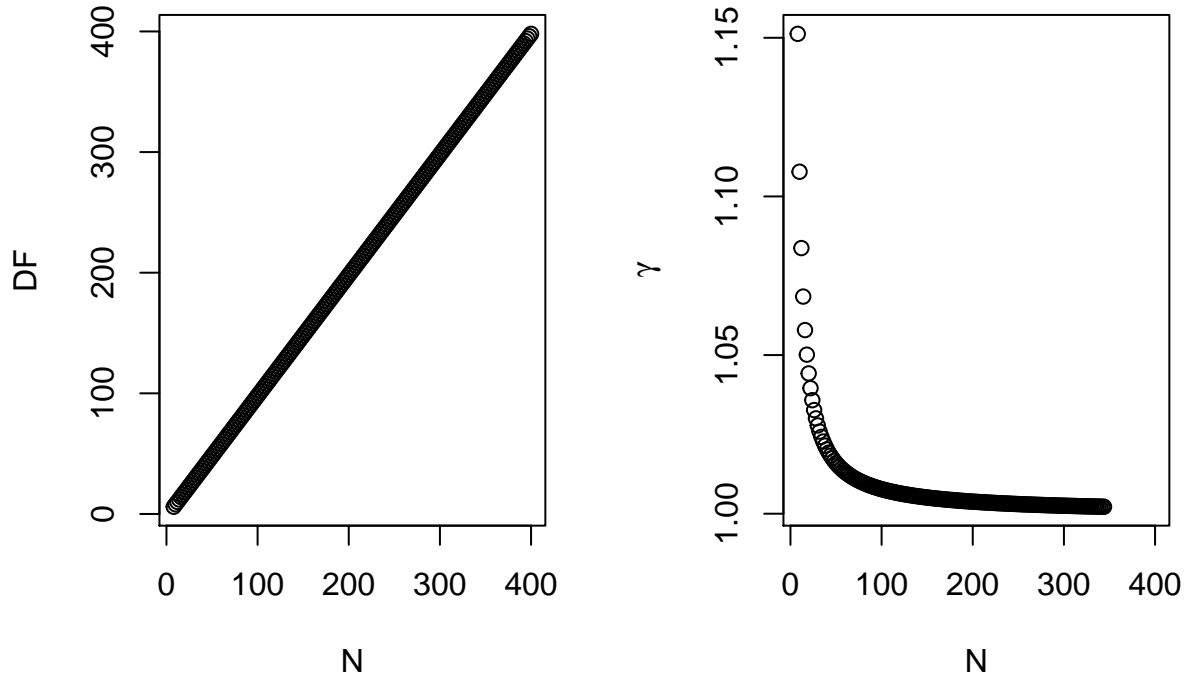


Figure 1. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d_s , when variances are equal across groups, as a function of the total sample size (N).

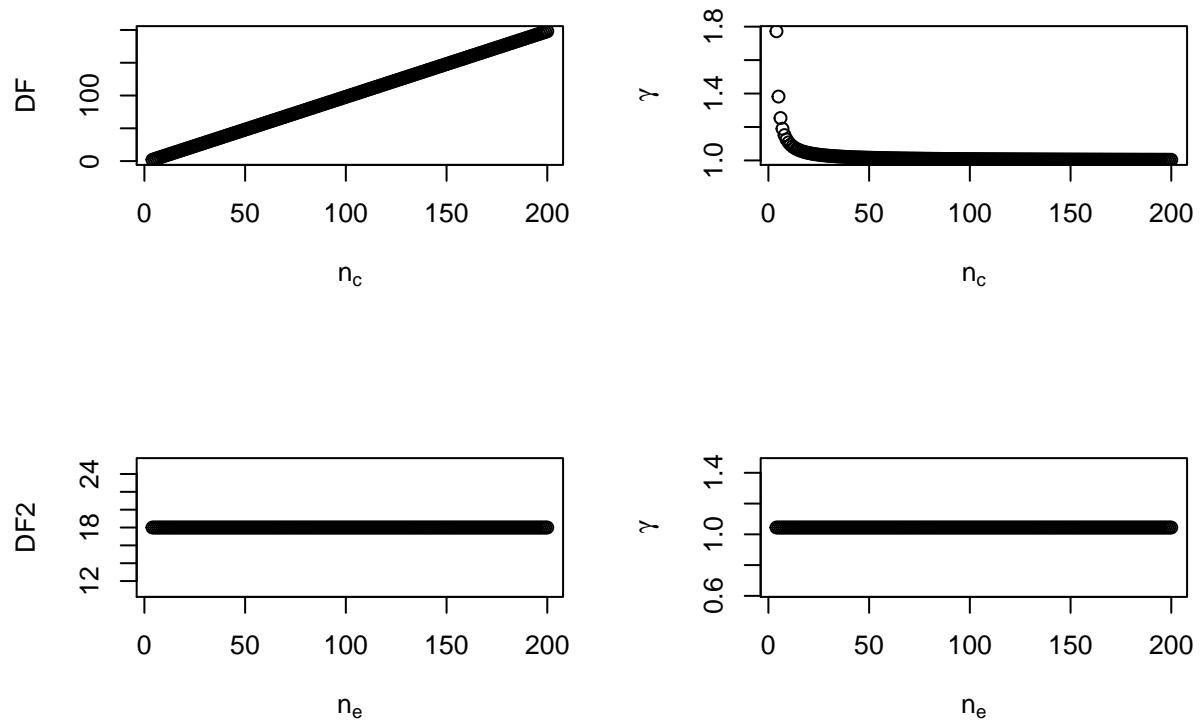


Figure 2. Degrees of freedom (DF) and γ , when computing the bias of Glass's d_s , when variances are equal across groups, as a function of n_c (top) and n_e (bottom).

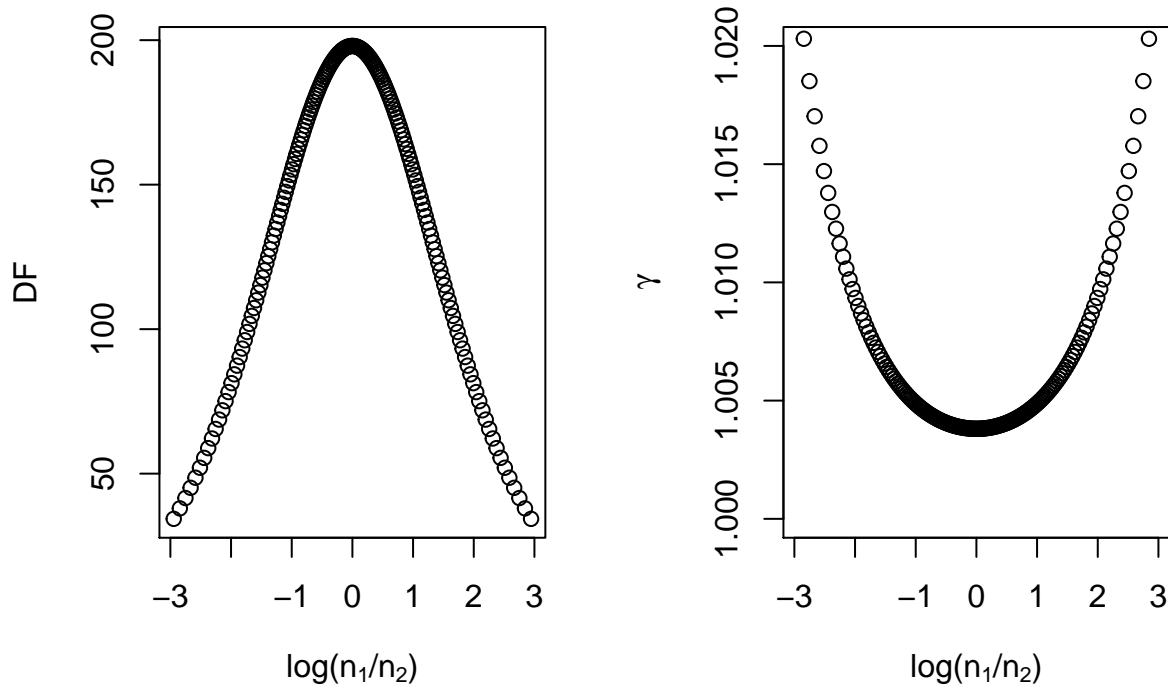


Figure 3. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances are equal across groups, as a function of the logarithm of the sample sizes ratio $\log\left(\frac{n_1}{n_2}\right)$.

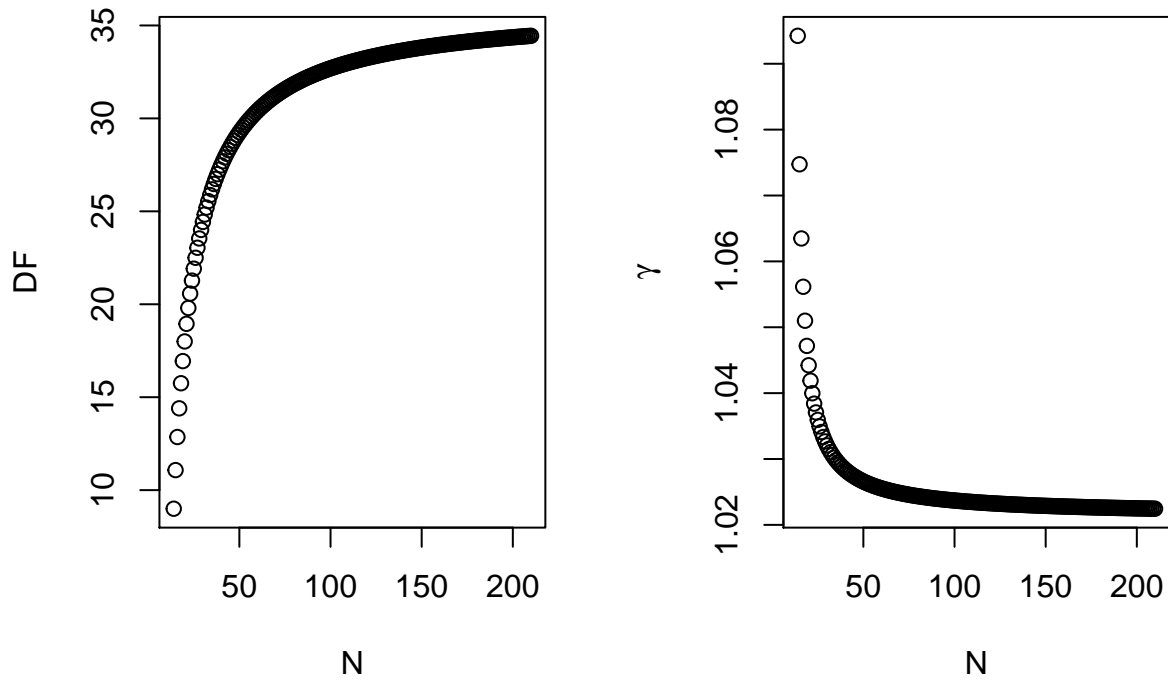


Figure 4. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances are equal across groups, as a function of the total sample size (N).

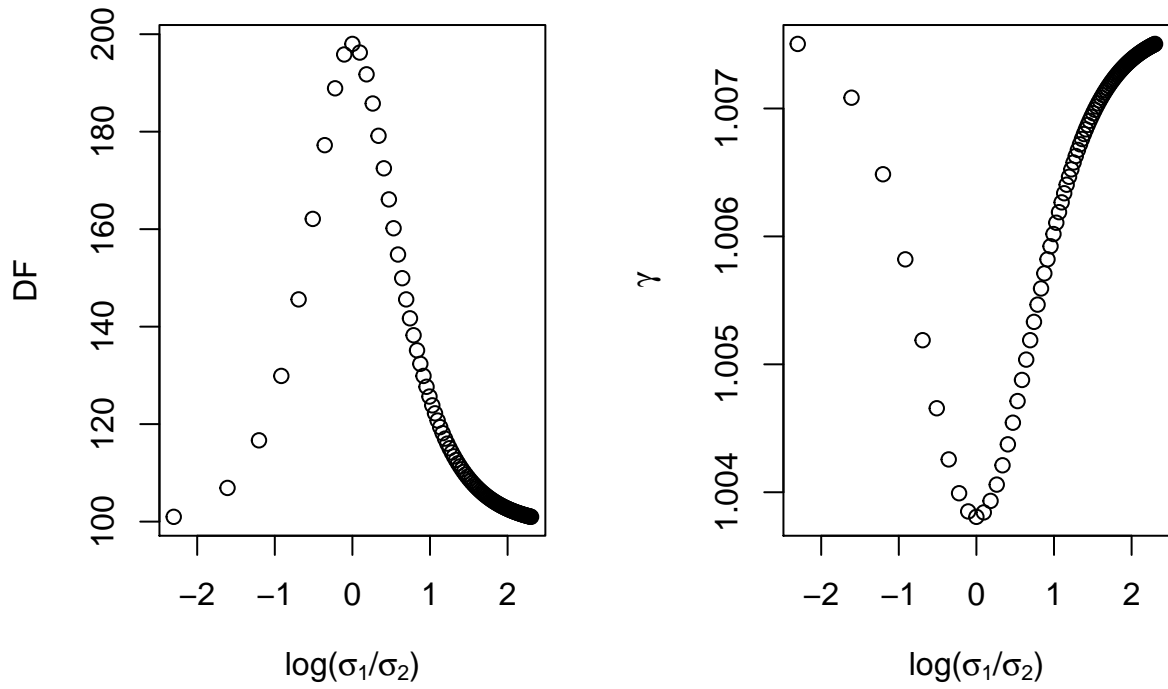


Figure 5. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances are unequal across groups and sample sizes are equal, as a function of the logarithm of the SD -ratio ($\log\left(\frac{\sigma_1}{\sigma_2}\right)$).

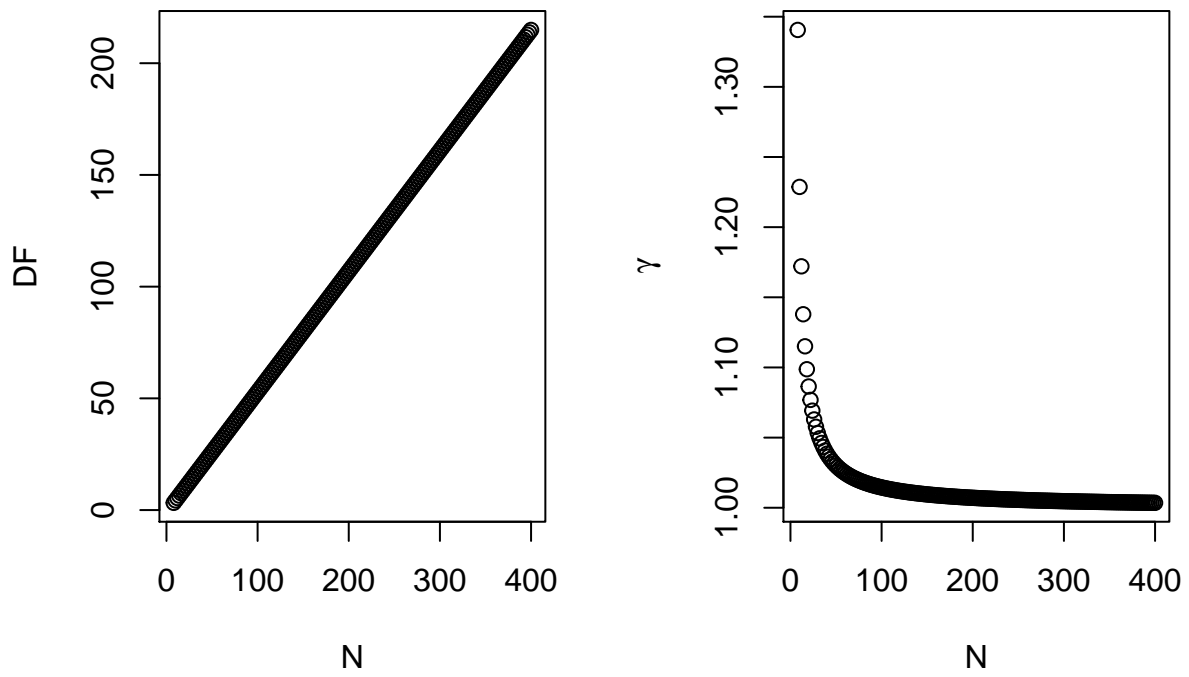


Figure 6. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances are unequal across groups and sample sizes are equal, as a function of the total sample size (N).

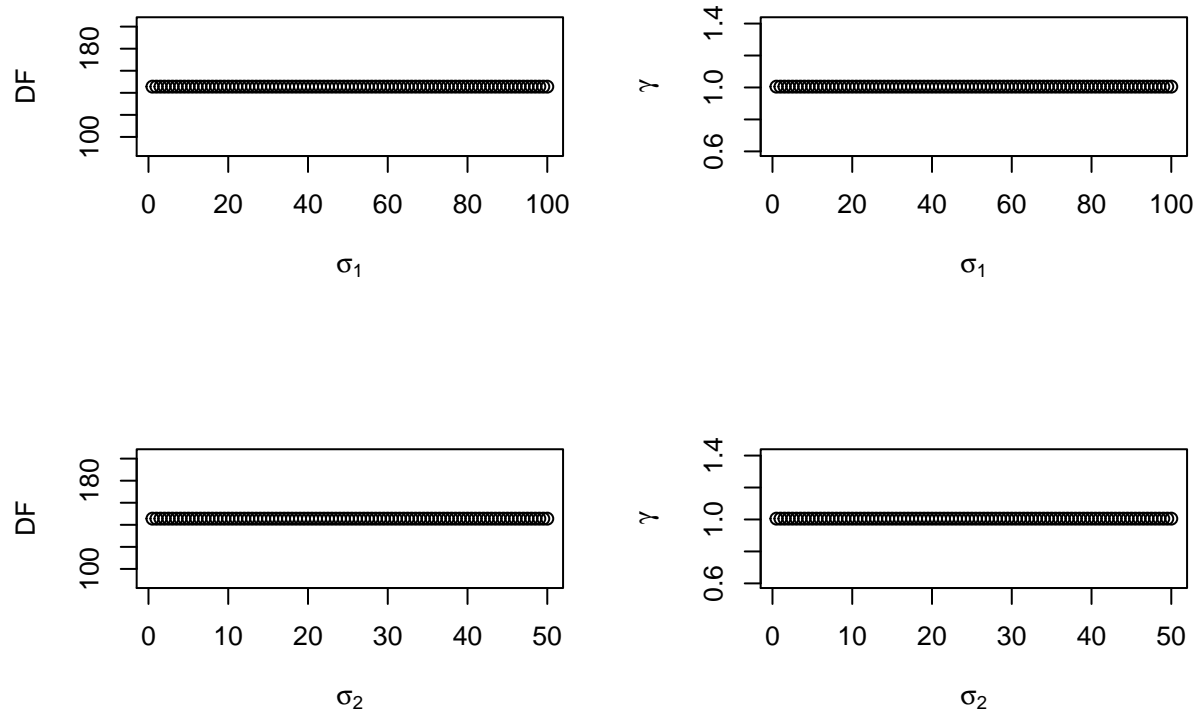


Figure 7. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances are unequal across groups and sample sizes are equal, as a function of σ_1 (top plots) and σ_2 (bottom plots), for a constant SD -ratio.

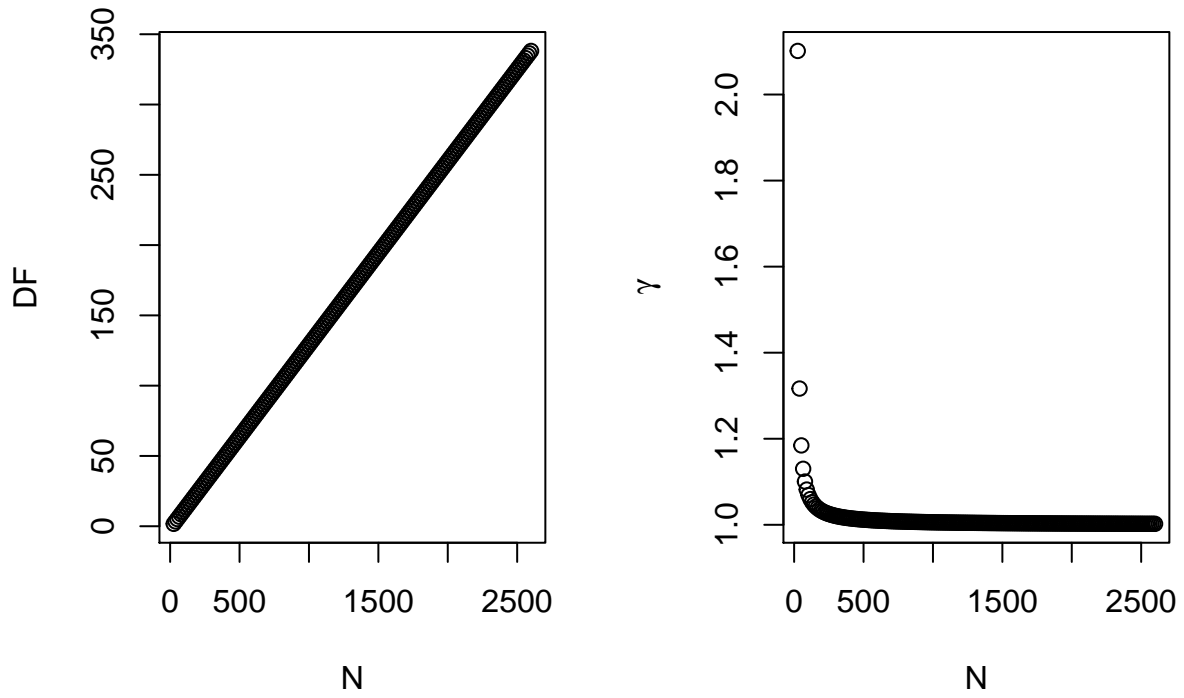


Figure 8. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances and sample sizes are unequal across groups, as a function of the total sample size (N).

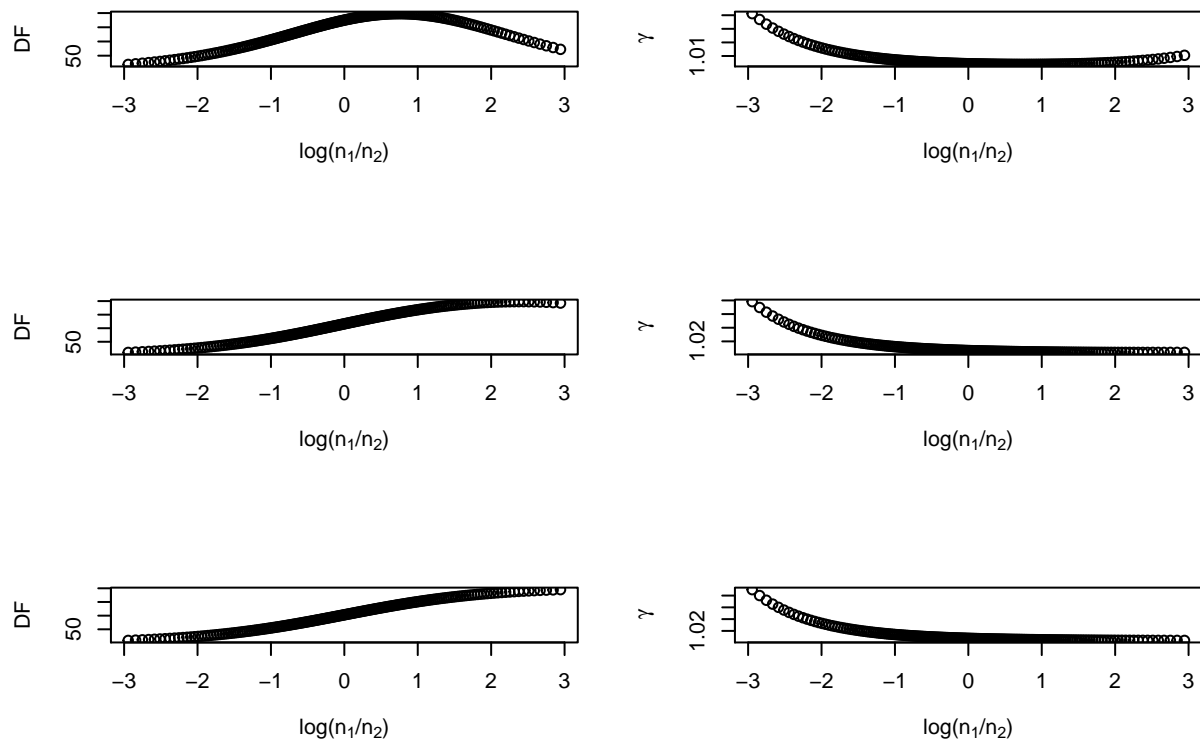


Figure 9. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s when variances and sample sizes are unequal across groups, as a function of the logarithm of the sample sizes ratio ($\log\left(\frac{n_1}{n_2}\right)$), when SD -ratio equals 1.46 (first row), 3.39 (second row) or 7 (third row).

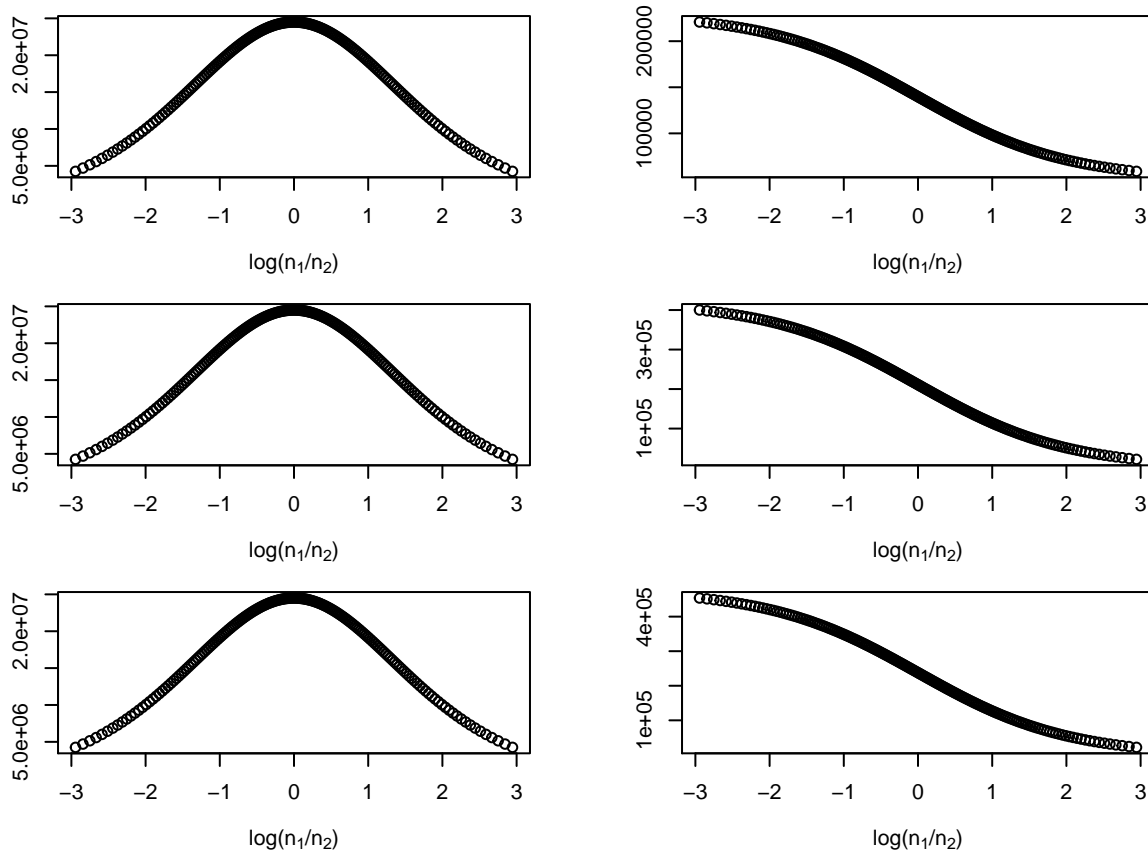


Figure 10. Numerator and denominator of the degrees of freedom (DF) computation, when computing the bias of Cohen's d'_s when variances and sample sizes are unequal across groups, as a function of the logarithm of the sample sizes ratio ($\log\left(\frac{n_1}{n_2}\right)$), when SD -ratio equals 1.46 (first row), 3.39 (second row) or 7 (third row).

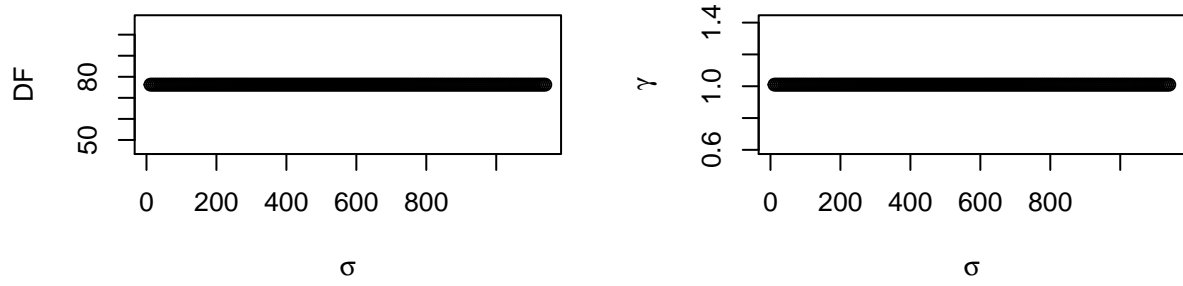


Figure 11. Degrees of freedom (DF) and γ , when computing the bias of Cohen's d'_s , when variances and sample sizes are unequal across groups, as a function of $\sigma = \frac{(\sigma_1^2 + \sigma_2^2)}{2}$, for a constant SD -ratio.

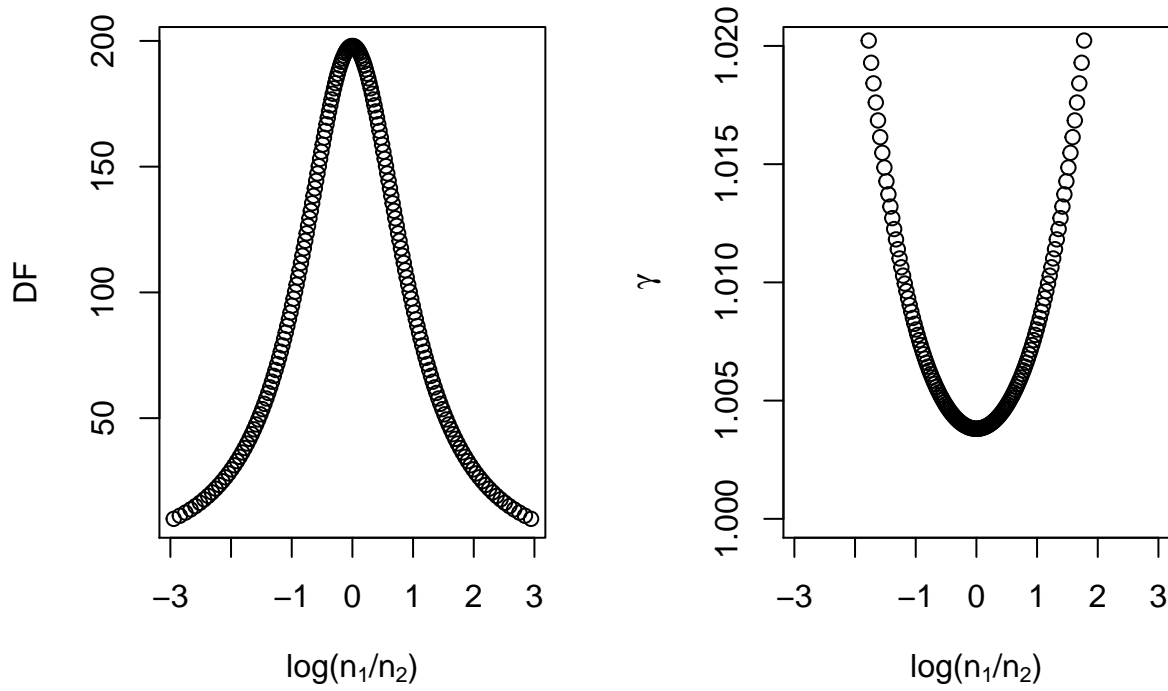


Figure 12. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances are equal across groups, as a function of the logarithm of the sample sizes ratio $(\log(\frac{n_1}{n_2}))$.

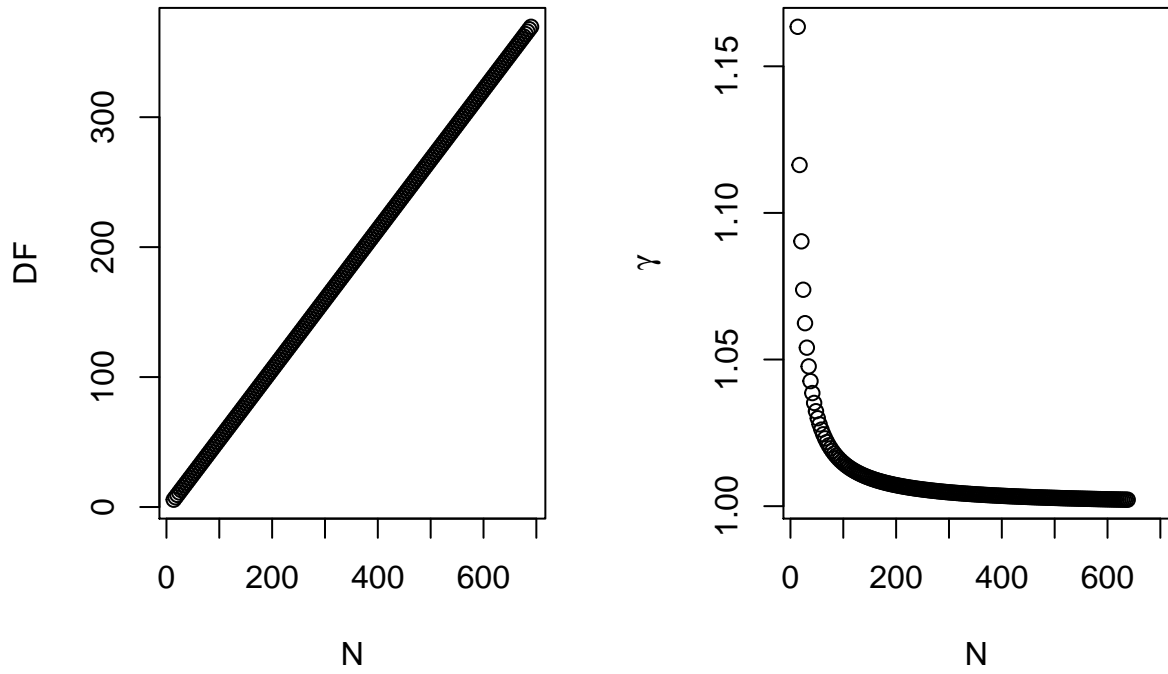


Figure 13. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances are equal across groups, as a function of the total sample size (N).

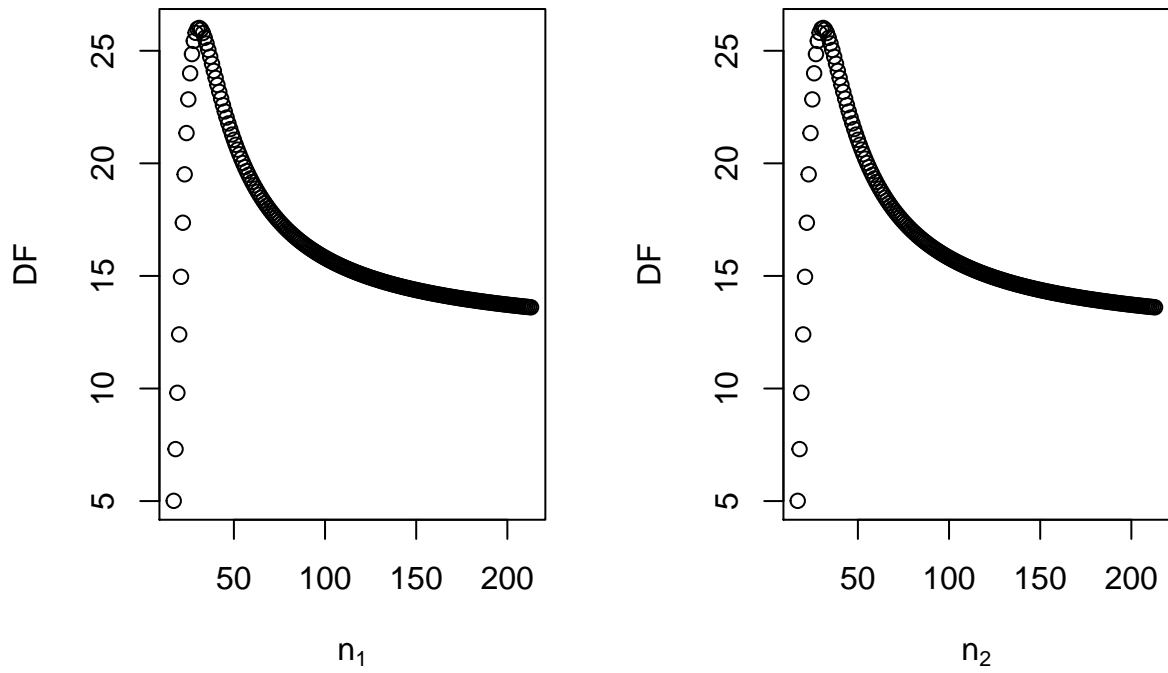


Figure 14. Degrees of freedom (DF), when computing the bias of Shieh's d_s , when variances are equal across groups, , when adding subjects only in the first group (left) or in the second group (right).

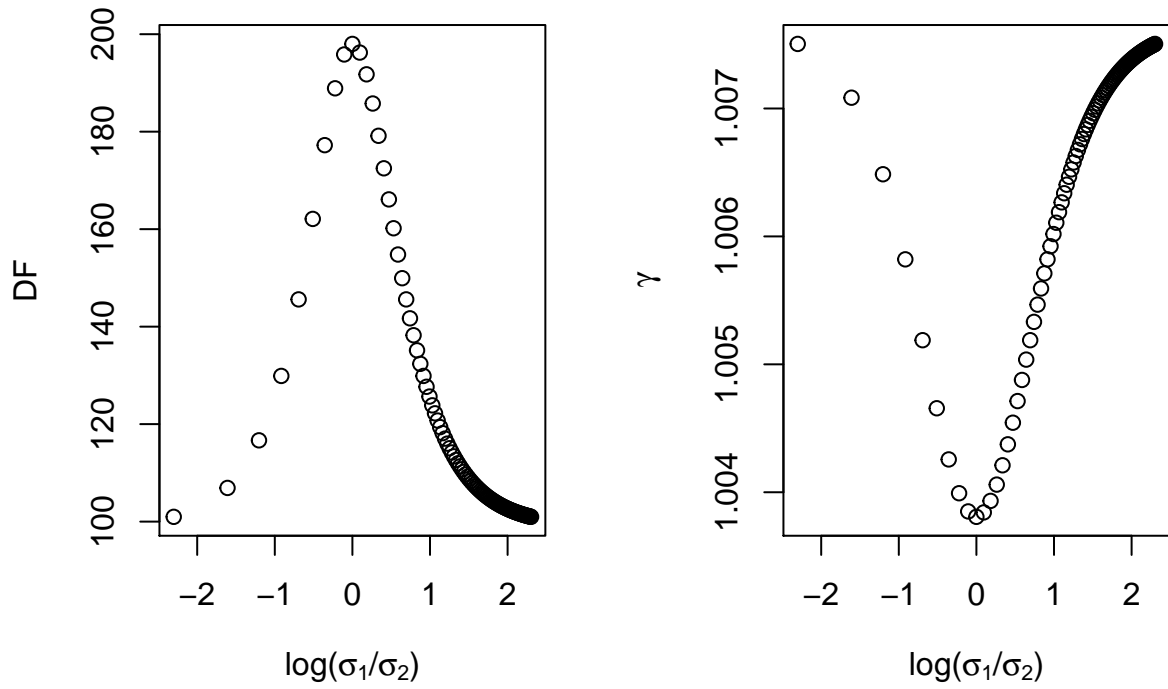


Figure 15. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances are unequal across groups and sample sizes are equal, as a function of the logarithm of the SD -ratio ($\log\left(\frac{\sigma_1}{\sigma_2}\right)$).

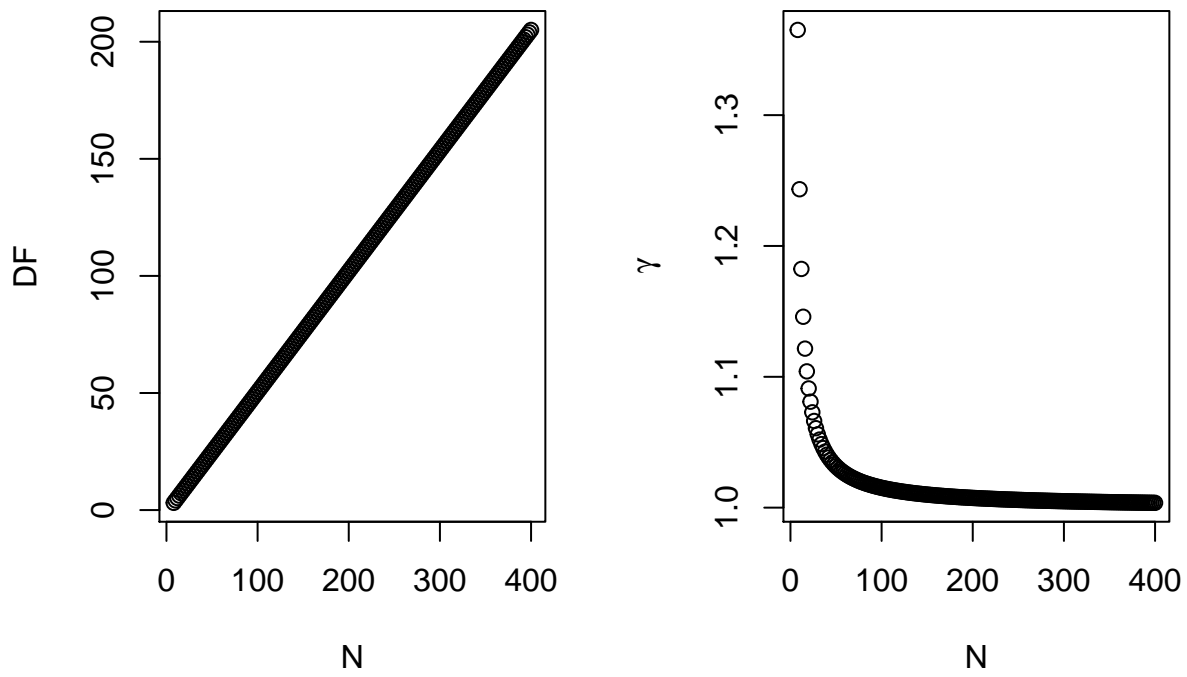


Figure 16. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances are unequal across groups and sample sizes are equal, as a function of the total sample size (N).

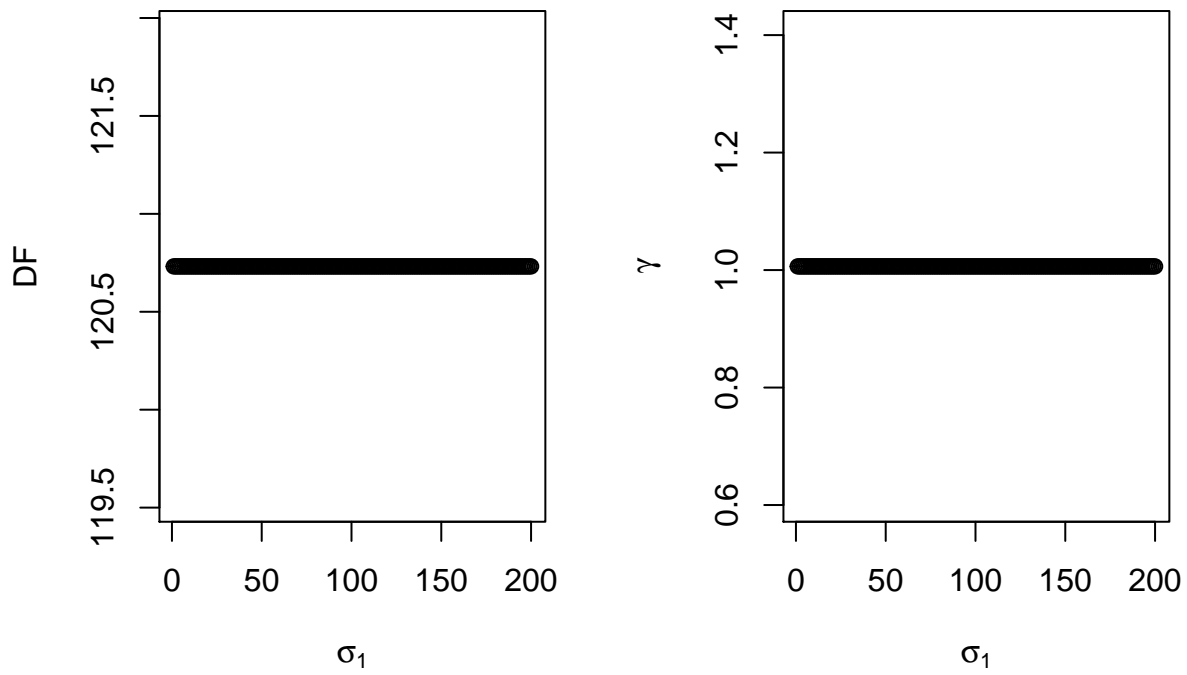


Figure 17. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances are unequal across groups and sample sizes are equal, as a function of σ_1 , for a constant SD -ratio.

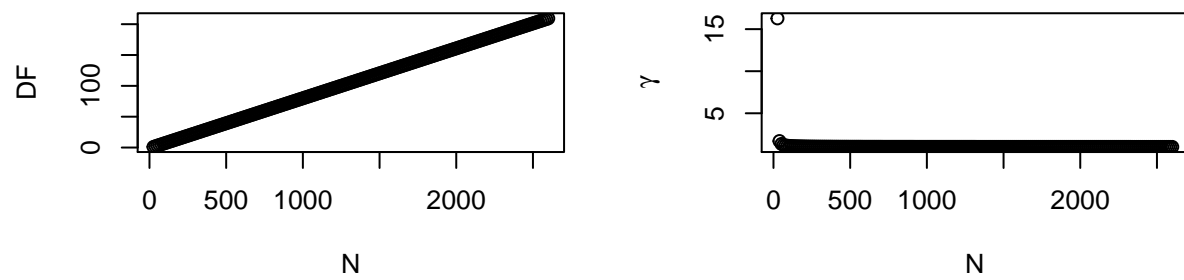


Figure 18. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances and sample sizes are unequal across groups, as a function of the total sample size (N).

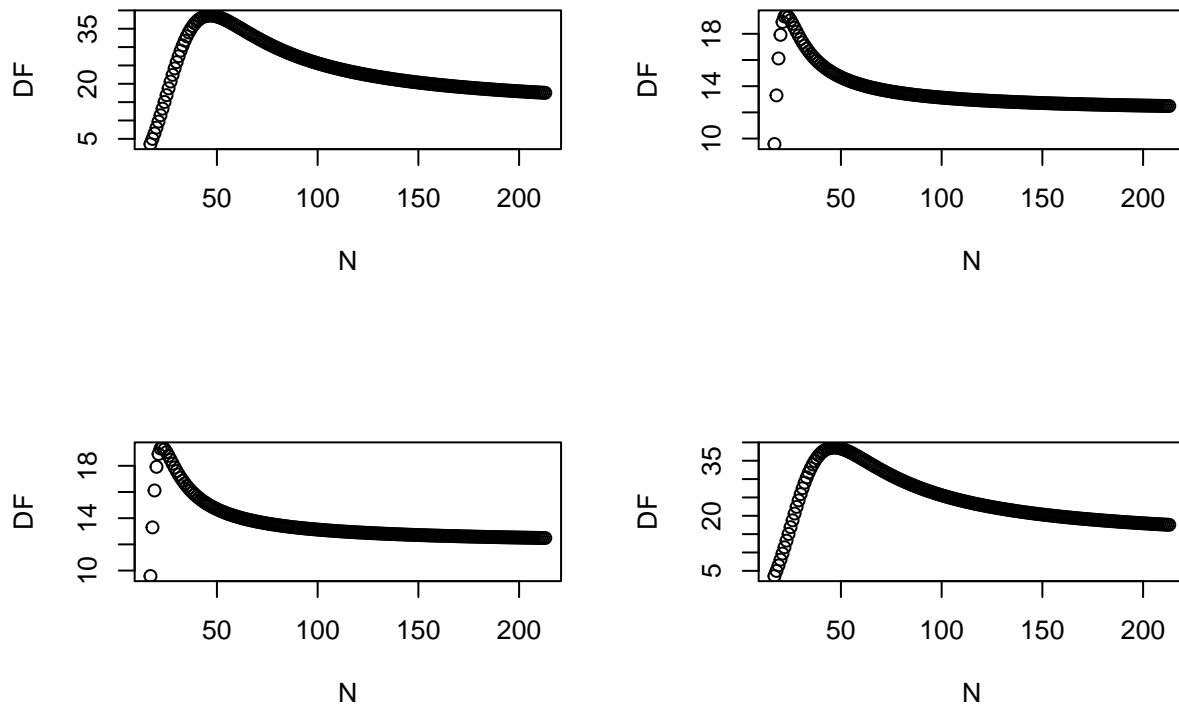


Figure 19. Degrees of freedom (DF), when computing the bias of Shieh's d_s , when variances and sample sizes are unequal across groups, as a function of the total sample size, when adding subjects only in one group (either in the first group; see top plots; or in the second group; see bottom plots), and $\sigma_1 > \sigma_2$ (left plots) or $\sigma_1 < \sigma_2$ (right plots).

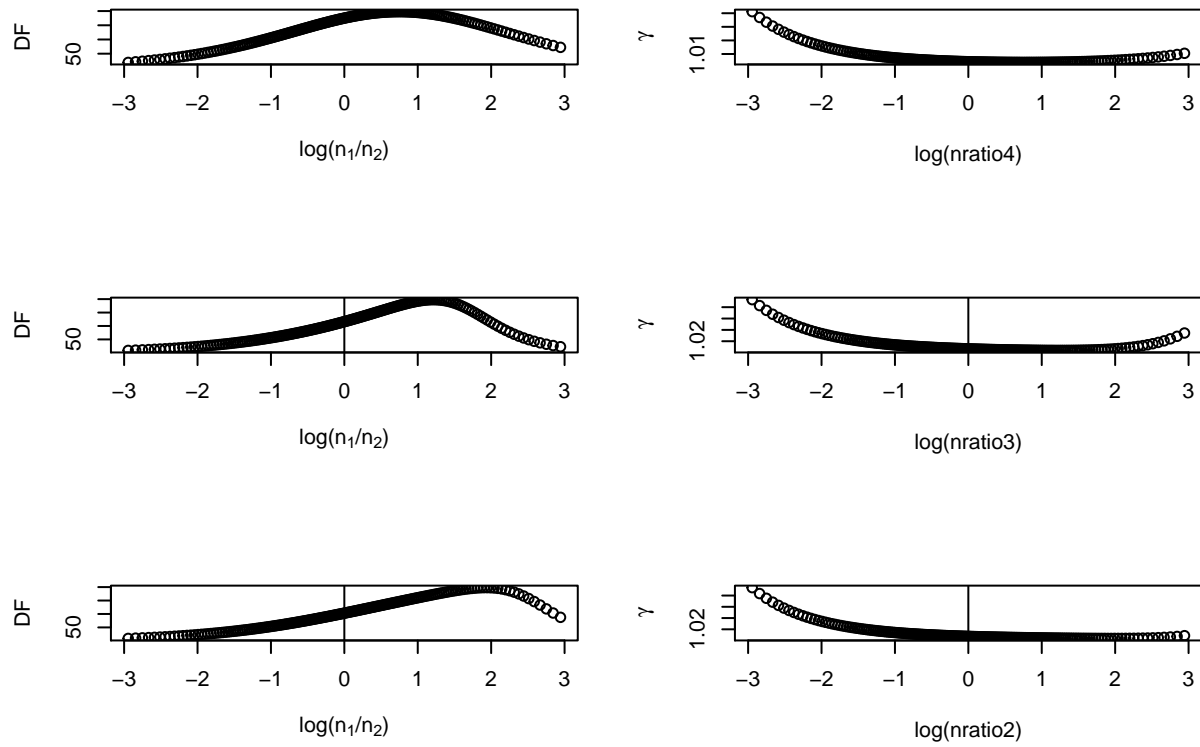


Figure 20. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances and sample sizes are unequal across groups, as a function of the logarithm of the sample sizes ratio ($\log\left(\frac{n_1}{n_2}\right)$), when SD -ratio equals 1.46 (first row), 3.39 (second row) or 7 (third row).

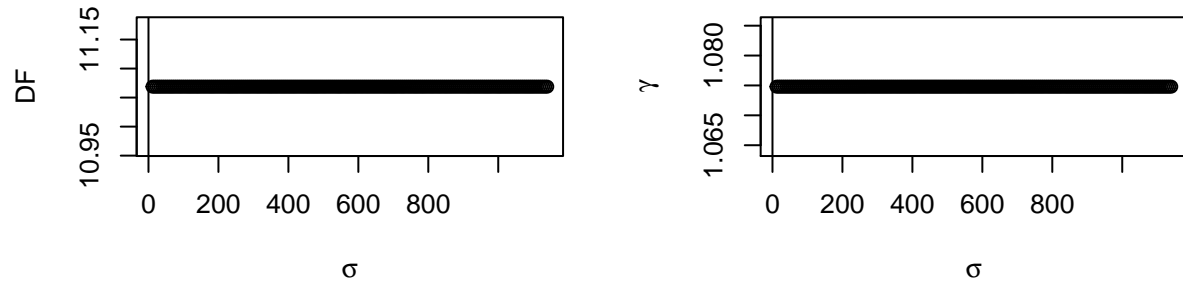


Figure 21. Degrees of freedom (DF) and γ , when computing the bias of Shieh's d_s , when variances and sample sizes are unequal across groups, as a function of σ_1 and σ_2 , for a constant SD -ratio.