Theoretical Bias, as a function of population parameters

Delacre Marie[1]

[1] ULB

Theoretical Bias, as a function of population parameters

The bias

For all 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 subdivise 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:

$$\gamma = \frac{\sqrt{\frac{df}{2}} \times \Gamma \frac{df-1}{2}}{\Gamma \frac{df}{2}} \tag{1}$$

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.

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 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 (see the two top plots in Figure 2). On the other side, increasing the size of the experimental group does not impact the bias (see the two bottom plots in Figure 2).

Cohen's d'_s (see Table 3)

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

- 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 in Figure 8)
- The smallest bias always occure when there is a positive pairing between variances and sample size, because one gives more weight to the smallest variance in the denominator

of the df computation. Moreover, the larger the SD-ratio, the further from 1 is the sample sizes ratio associated with the smallest bias. This can be explained by splitting the numerator and the denominator in the DF computation (see Figure 9).

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 $\frac{n_1}{n_2} = max(\frac{n_1}{n_2}))$ and the larger the SD-ratio, the larger the impact of the sample sizes ratio on the denominator. In the end, the larger the SD-ratio, the further from 1 is the sample sizes ratio associated with the larger degrees of freedom.

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

Shieh's d_s (see Table 4)

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 ($\frac{n_1}{n_2}$). 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 ($\frac{n_1}{n_2}$) 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)

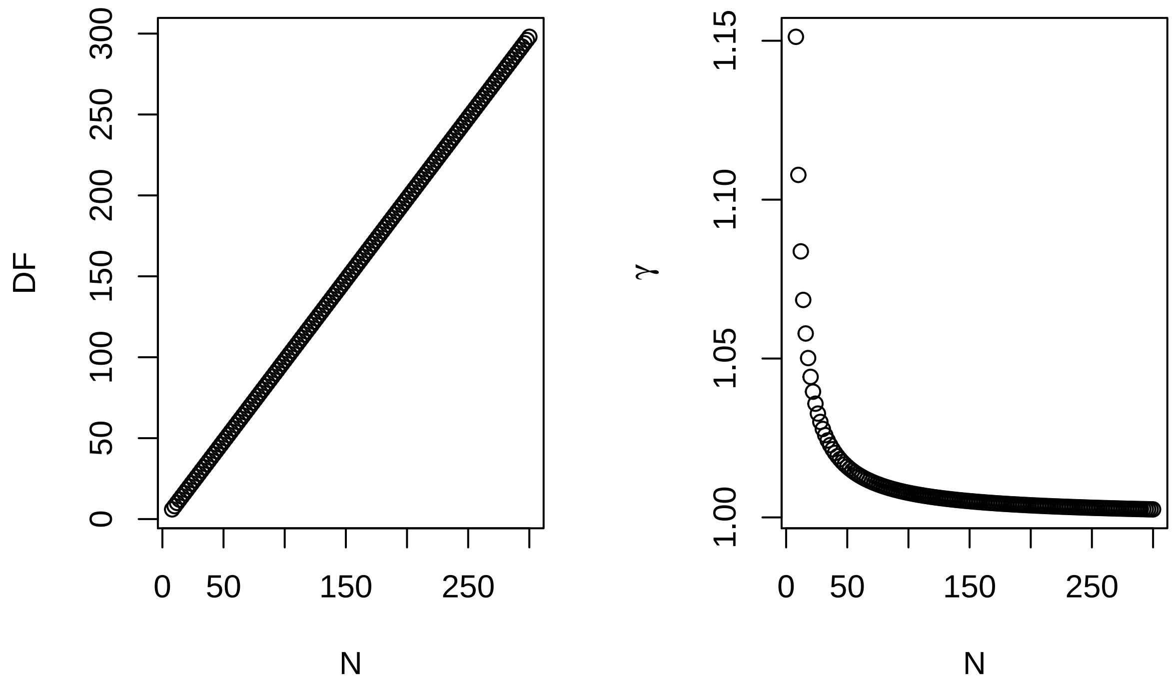

Figure 1. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s, when variances are equal across groups, as a function of the total sample size (N)

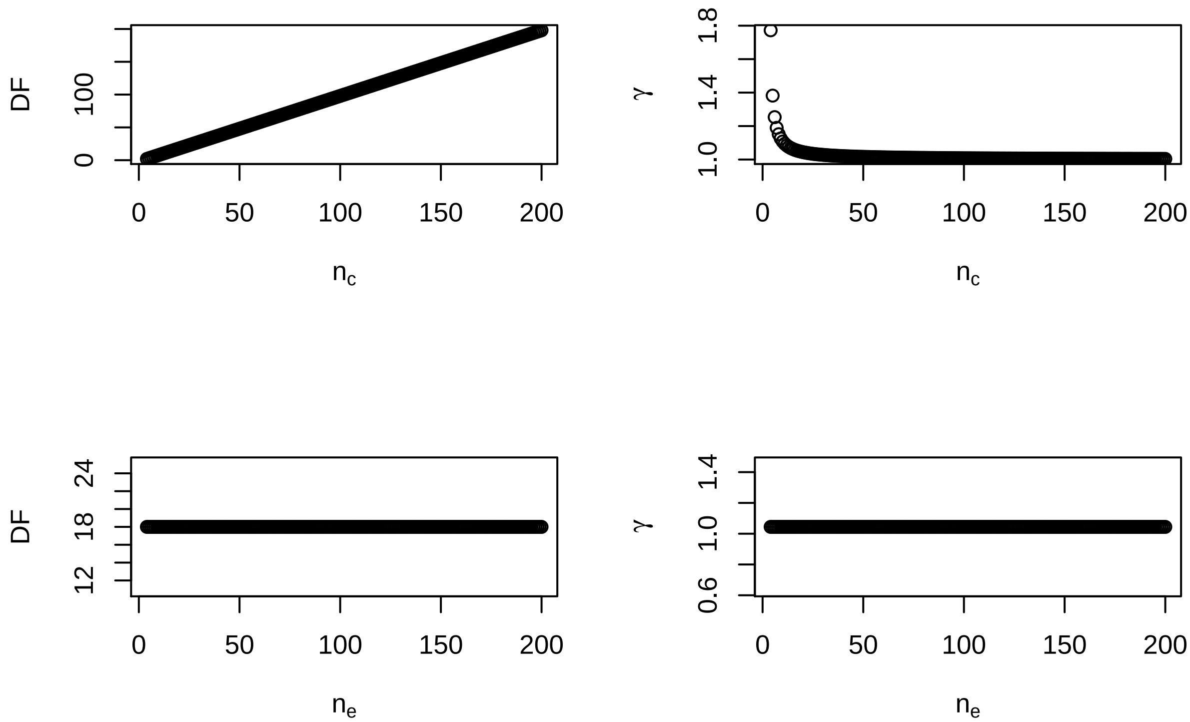

Figure 2. Degrees of freedom (DF) and γ, when computing the bias of Glass's d_s, as a function of n_c (top) and n_e (bottom)

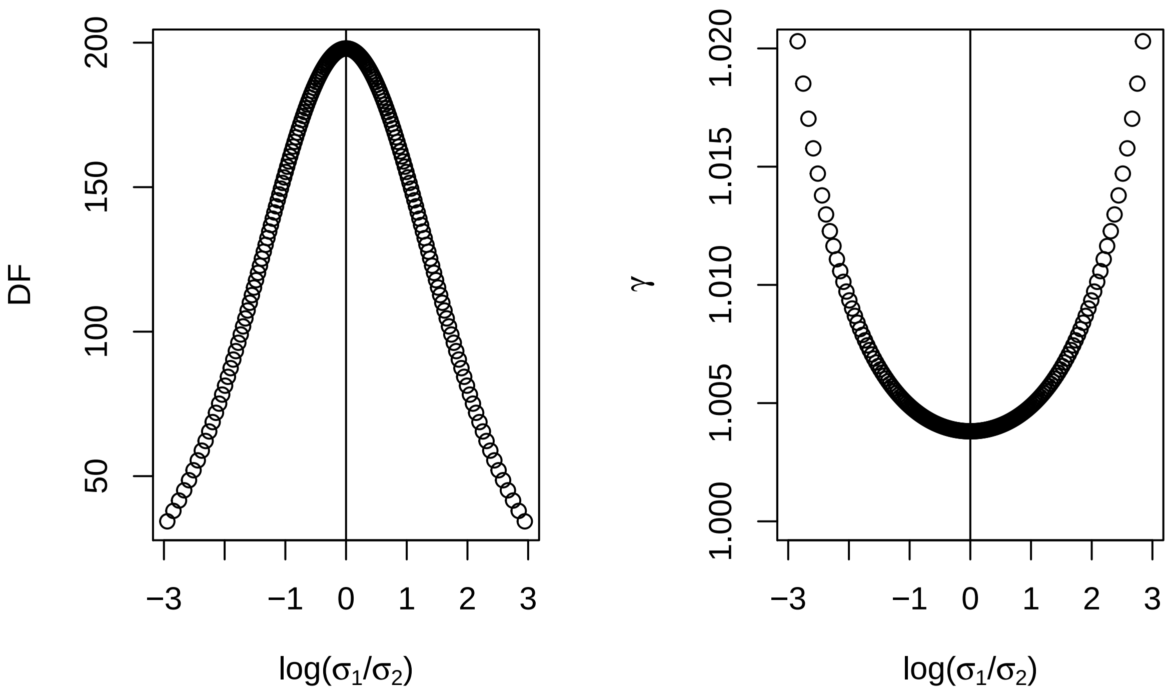

Figure 3. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d'_s, when variances are equal across groups, as a function of the logarithm of the sample sizes ratio $log\left(\frac{n_1}{n_2}\right)$

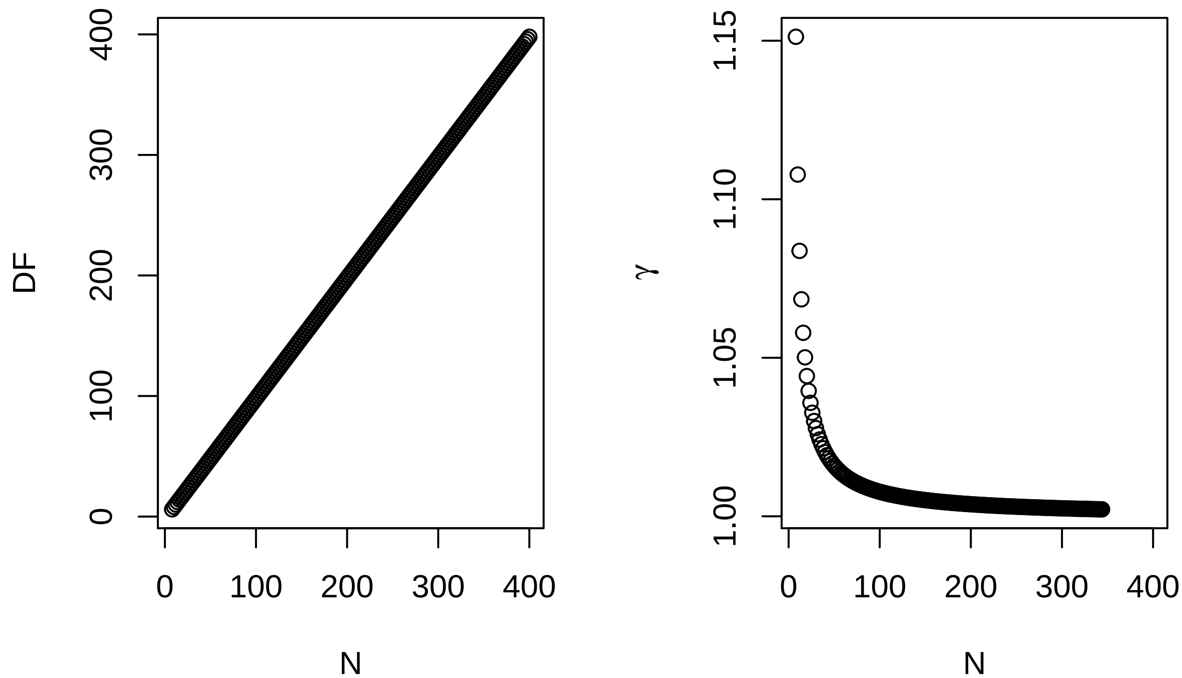

Figure 4. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d'_s, when variances are equal across groups, as a function of the total sample size (N)

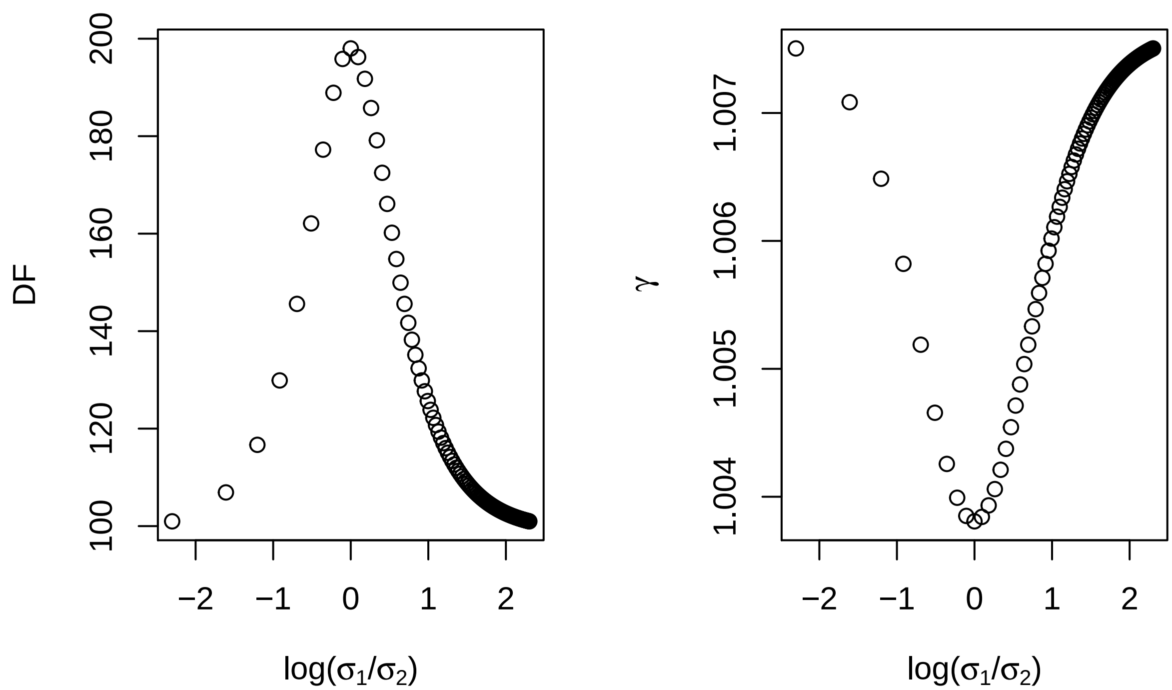

Figure 5. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s', when variances are unequal across groups and sample sizes are equal, as a function of the logarithm of the *SD*-ratio ($log\left(\frac{\sigma_1}{\sigma_2}\right)$)

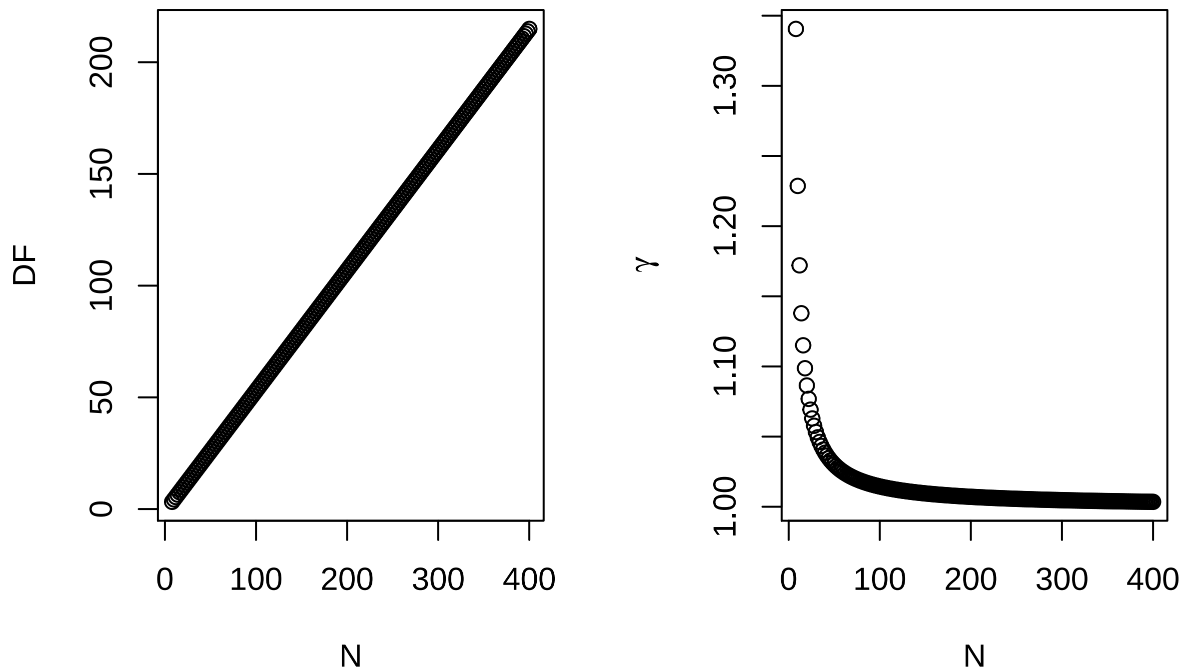

Figure 6. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s', when variances are unequal across groups and sample sizes are equal, as a function of the total sample size (N)

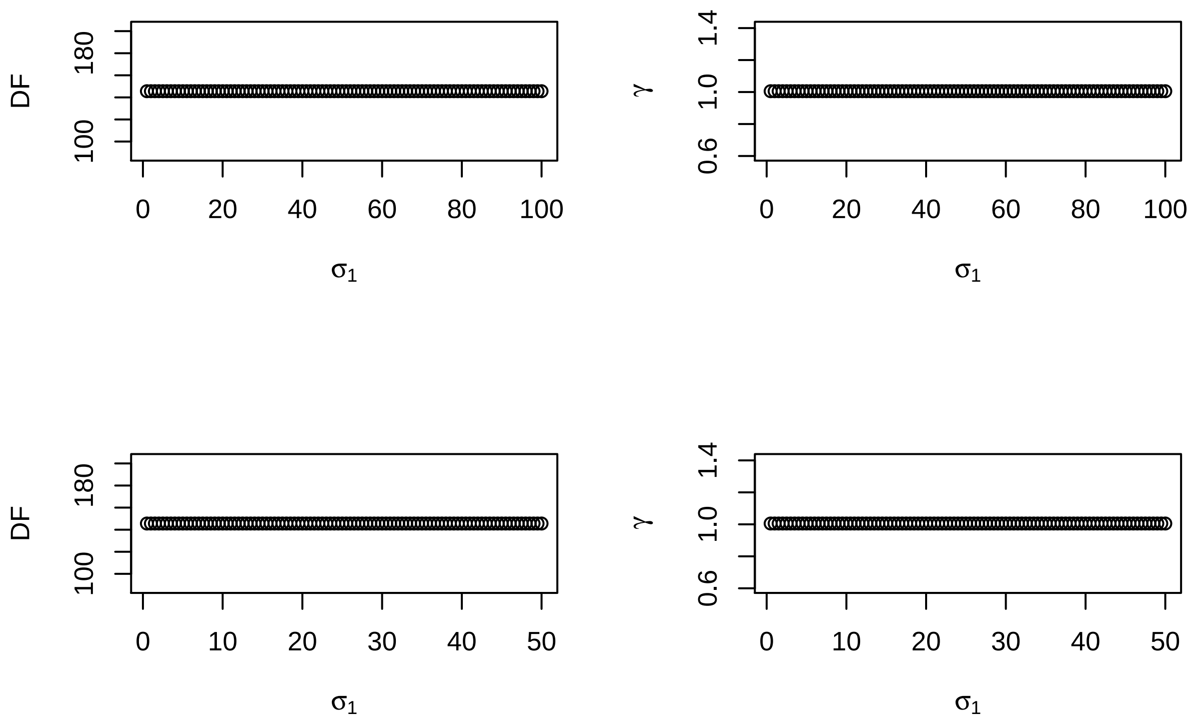

Figure 7. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d'_s, when variances are unequal across groups and sample sizes are equal, as a function of σ_1 and σ_2, for a constant SD-ratio

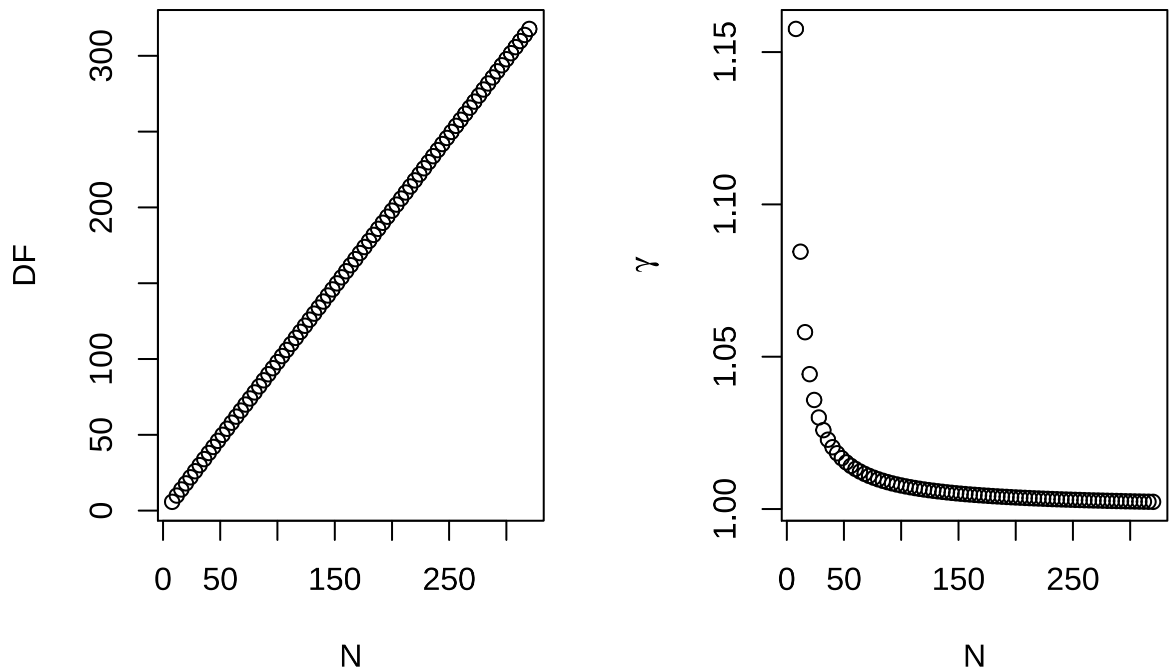

Figure 8. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d'_s, when variances and sample sizes are unequal across groups, as a function of the total sample size (N)

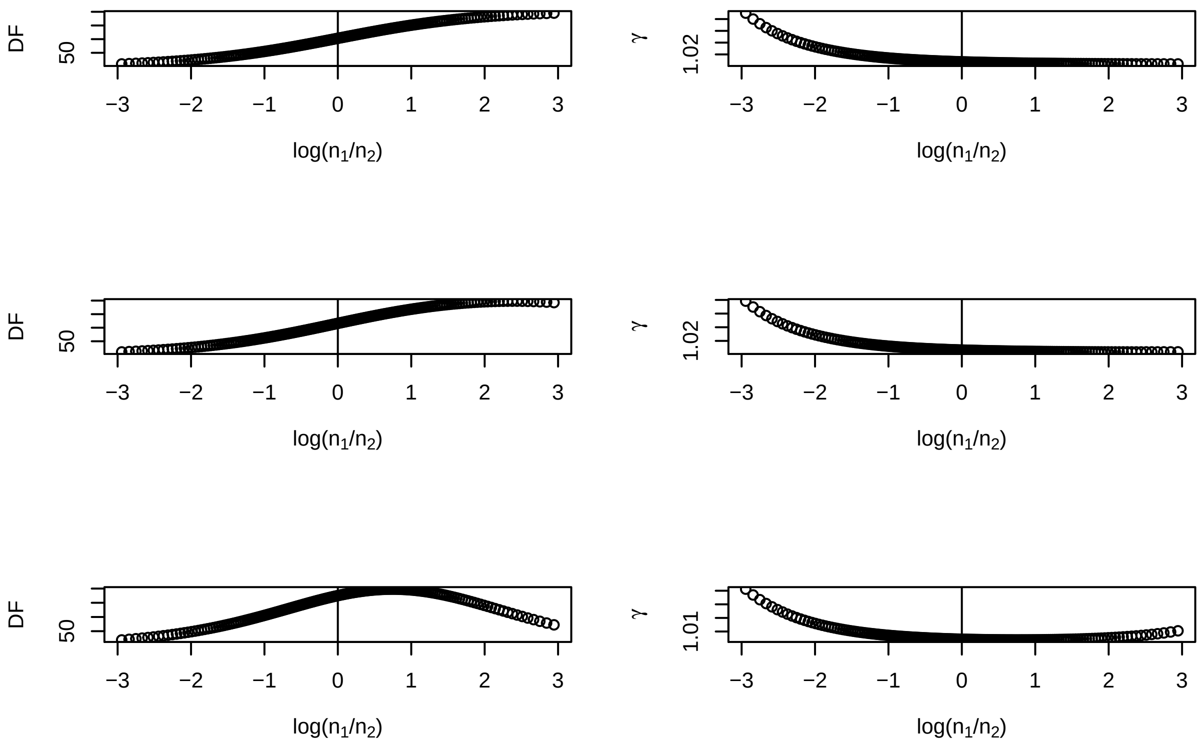

Figure 9. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d'_s, when variances and sample sizes are unequal across groups, as a function of the logarithm of the sample sizes ratio ($log\left(\frac{n_1}{n_2}\right)$), when SD-ratio equals 1.46 (first row), 3.39 (second row) or 7 (third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

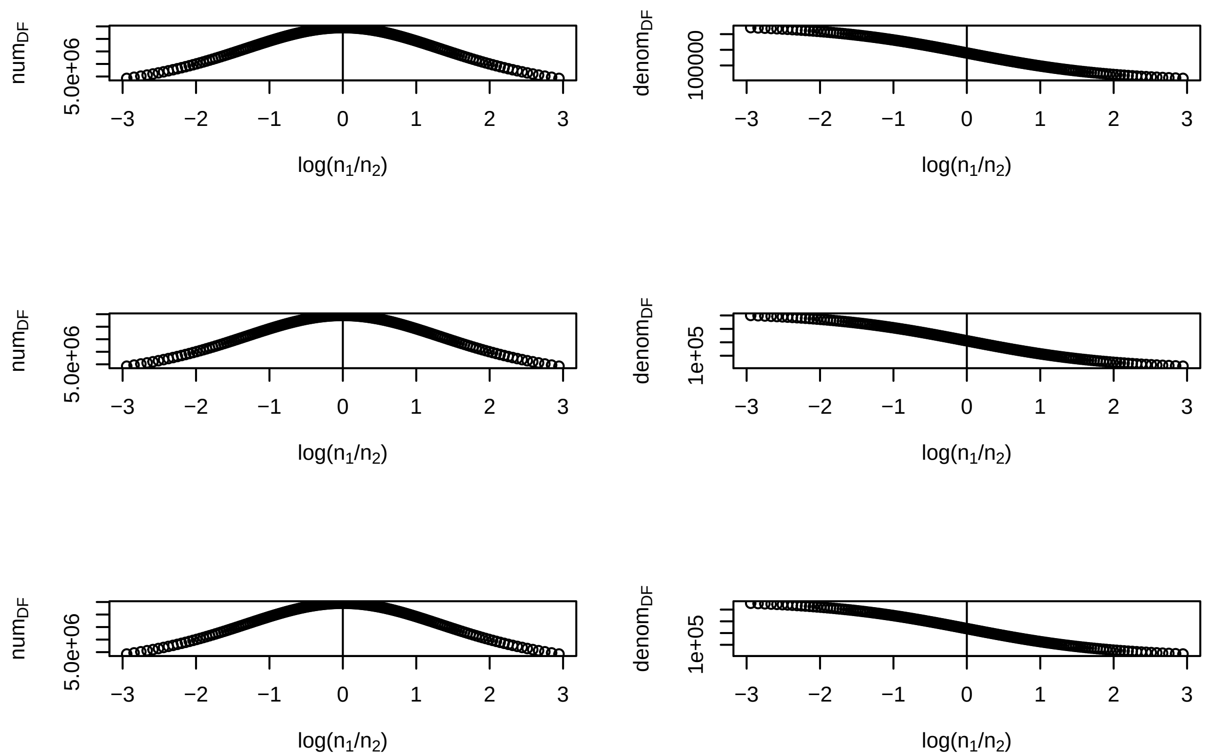

Figure 10. numerator and denominator of the degrees of freedom (DF) computation, when computing the bias of Cohen's d'_s, when variances and sample sizes are unequal across groups, as a function of the logarithm of the sample sizes ratio ($log\left(\frac{n_1}{n_2}\right)$), when SD-ratio equals 1.46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

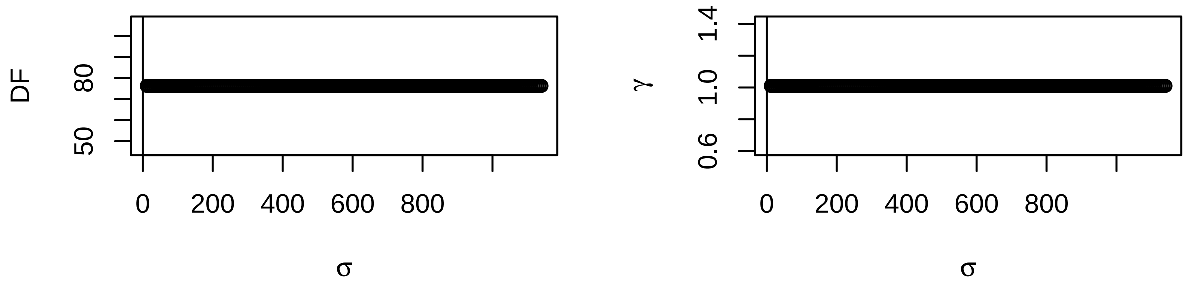

Figure 11. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d'_s, when variances and sample sizes are unequal across groups, as a function of $\sigma = \frac{(\sigma_1^2+\sigma_2^2)}{2}$, for a constant SD-ratio

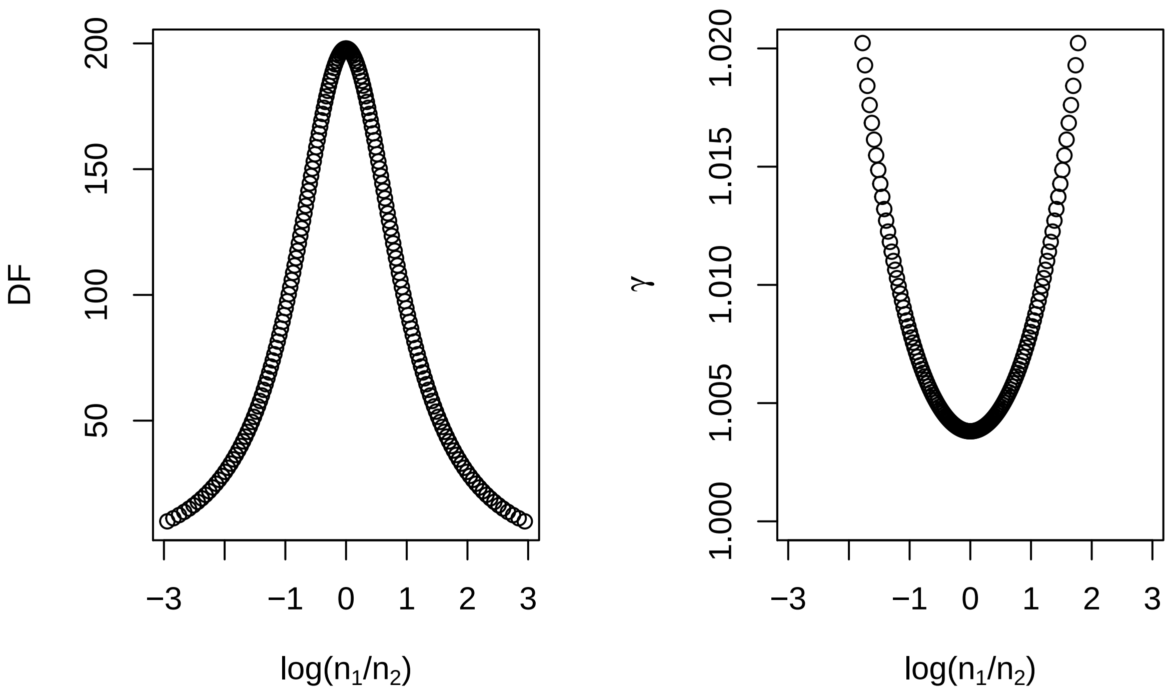

Figure 12. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are equal across groups, as a function of the logarithm of the sample sizes ratio $\left(log\left(\frac{n_1}{n_2}\right)\right)$

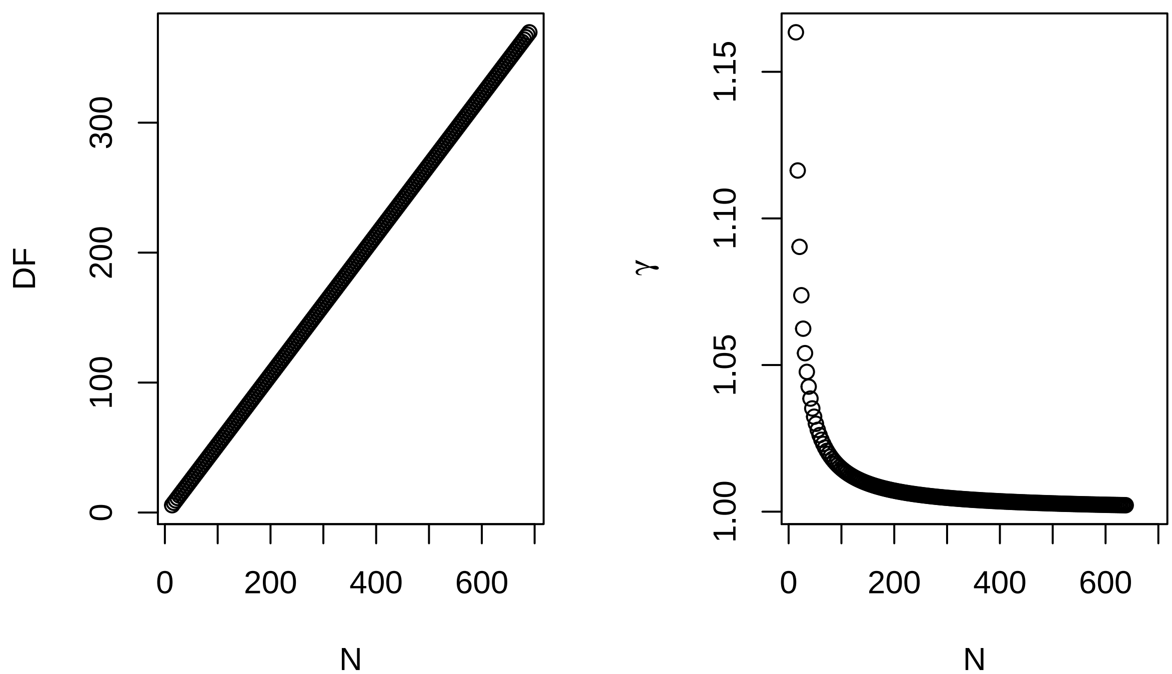

Figure 13. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are equal across groups, as a function of the total sample size (N)

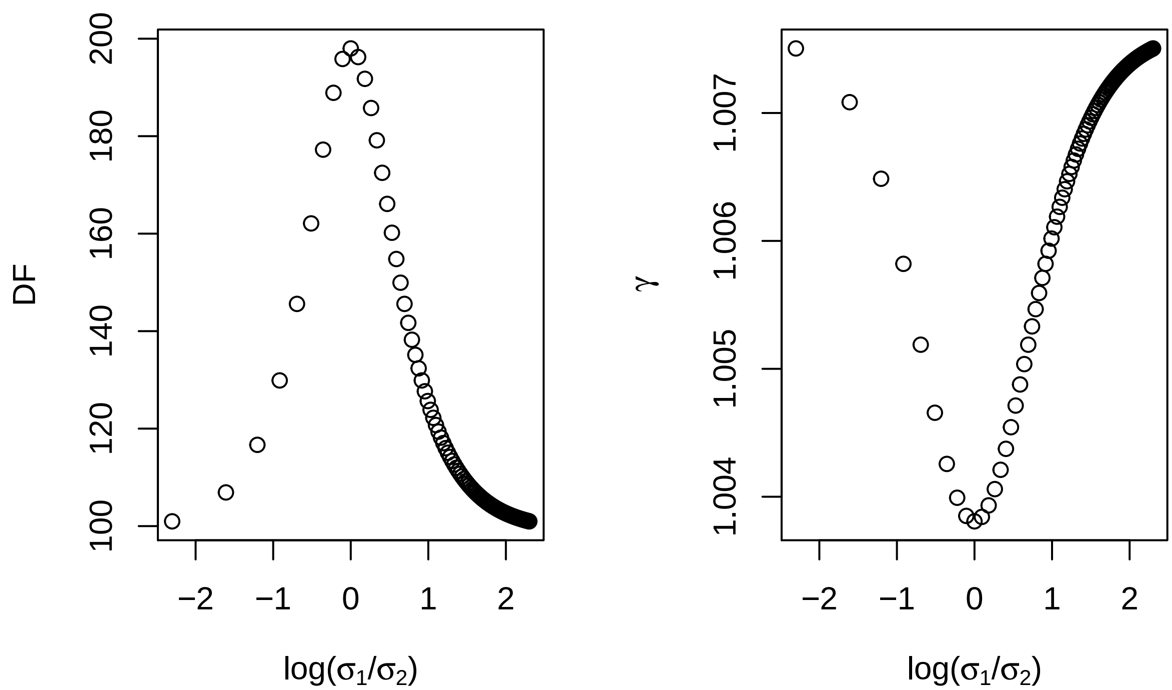

Figure 14. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are unequal across groups and sample sizes are equal, as a function of the logarithm of the SD-ratio ($log\left(\frac{\sigma_1}{\sigma_2}\right)$)

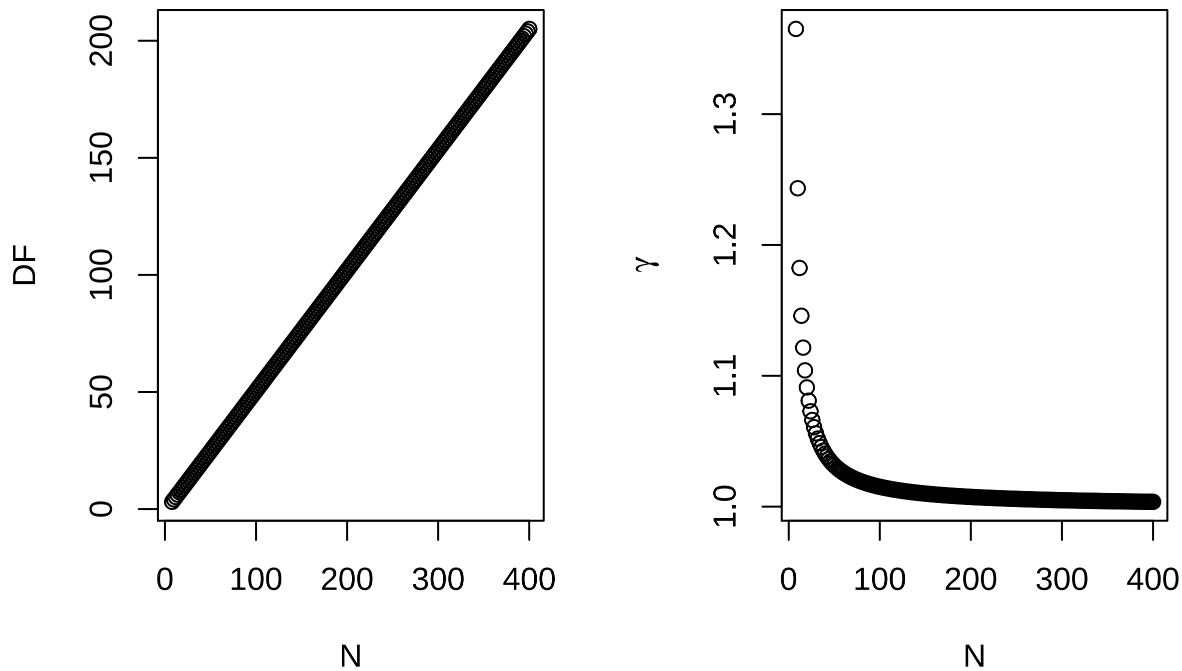

Figure 15. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are unequal across groups and sample sizes are equal, as a function of the total sample size (N)

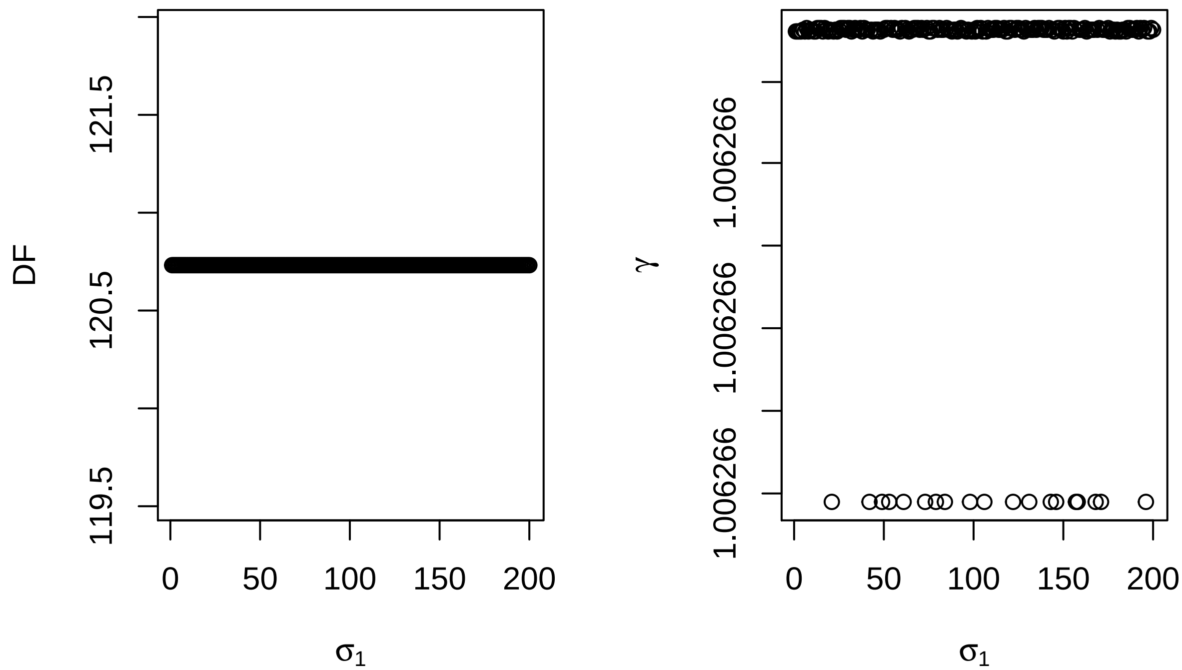

Figure 16. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are unequal across groups and sample sizes are equal, as a function of σ_1, for a constant *SD*-ratio

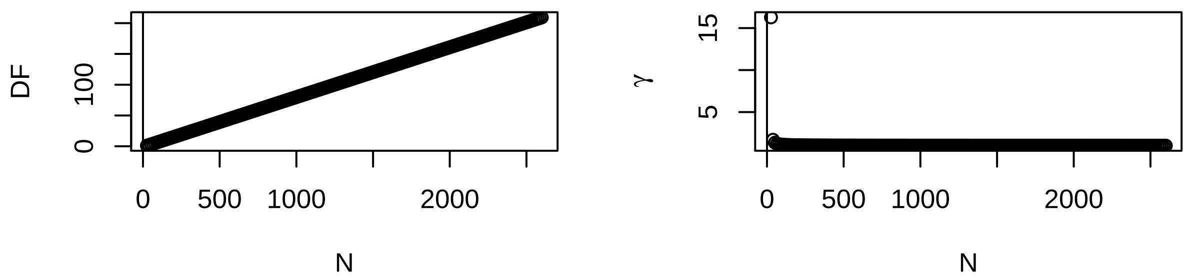

Figure 17. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances and sample sizes are unequal across groups, as a function of the total sample size (N)

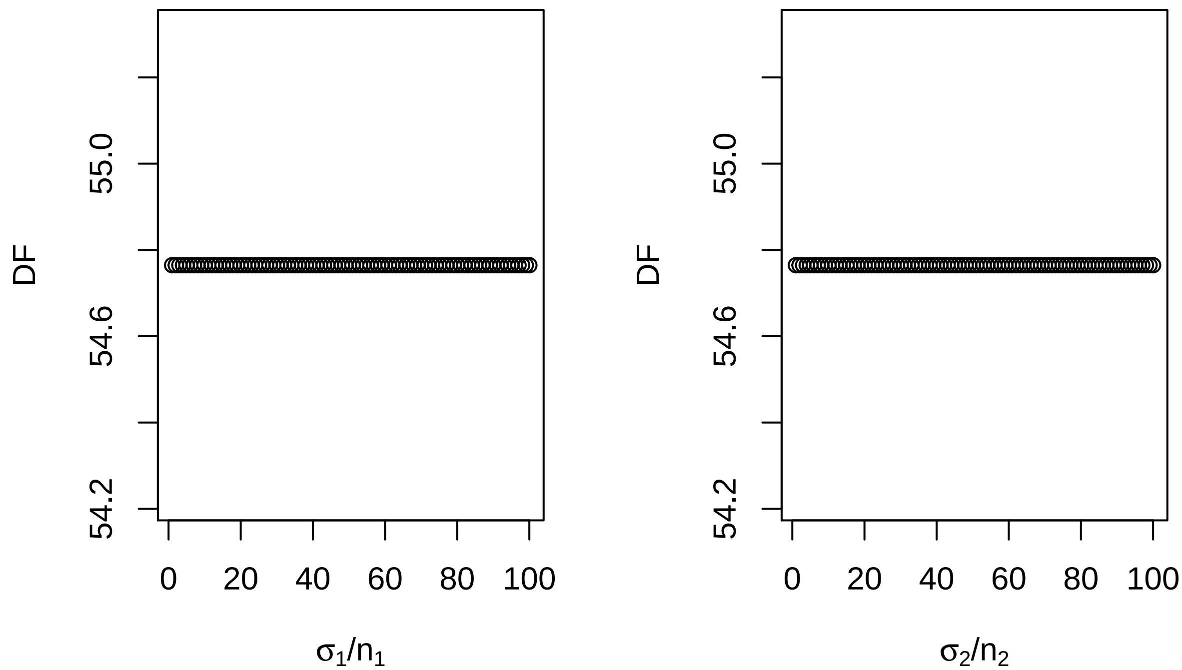

Figure 18. Degrees of freedom (DF) when computing the bias of Shieh's d_s, when variances and sample sizes are unequal across groups, as a function of the variances and sample sizes ratios ($\frac{\sigma_j}{n_j}$)

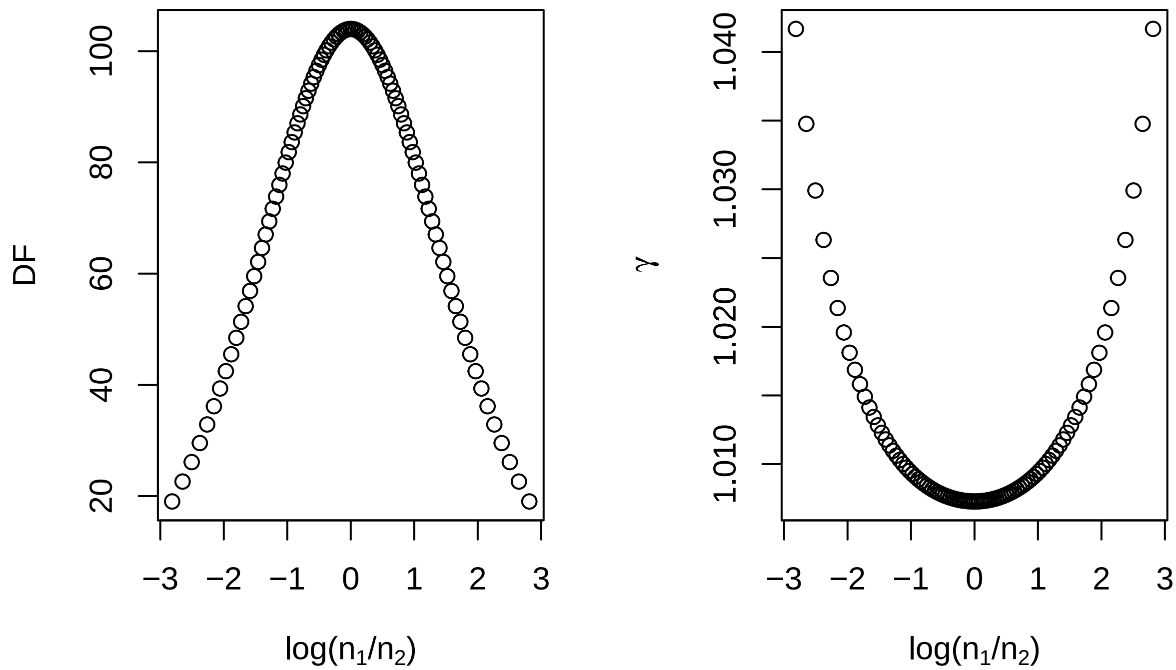

Figure 19. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances and sample sizes are unequal across groups and $\frac{\sigma_1^2}{n_1} = \frac{\sigma_2^2}{n_2}$, as a function of the logarithm of the sample sizes ratio ($log\left(\frac{n_1}{n_2}\right)$)

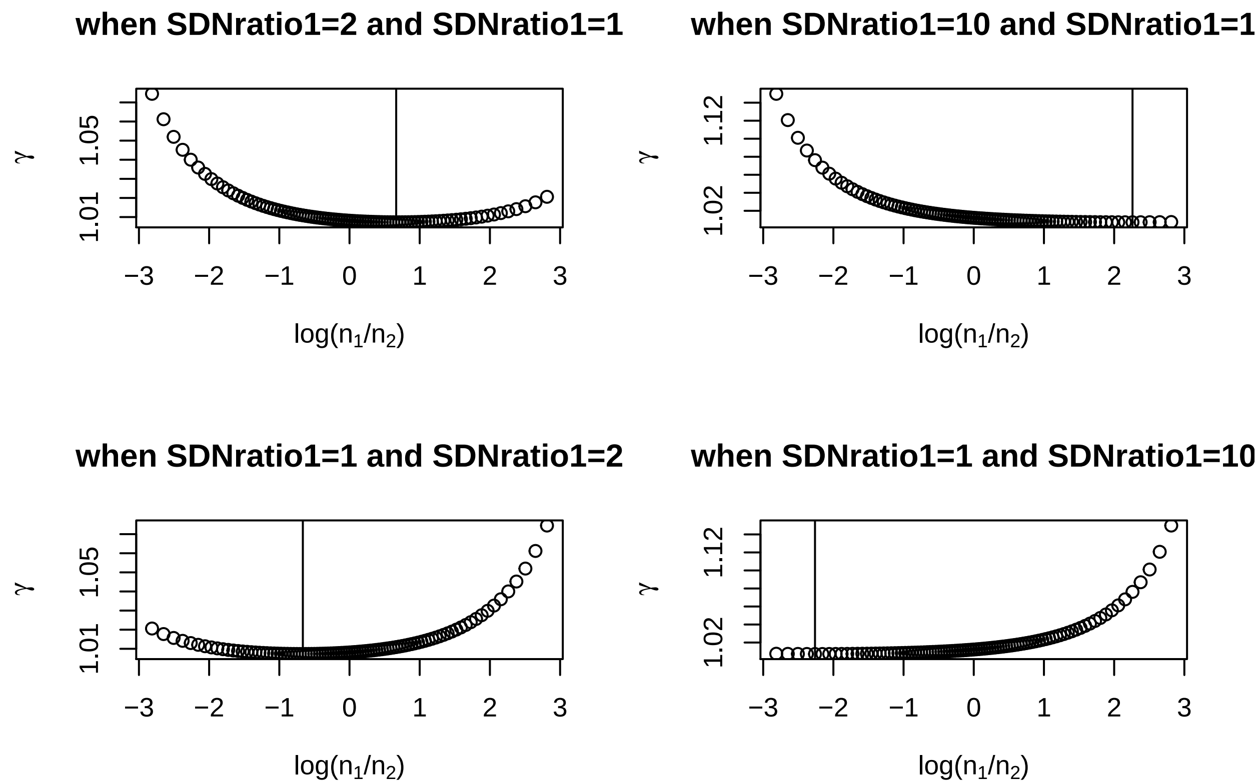

Figure 20. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances and sample sizes are unequal across groups and either $\frac{\sigma_1^2}{n_1} > \frac{\sigma_2^2}{n_2}$ (top) or $\frac{\sigma_1^2}{n_1} < \frac{\sigma_2^2}{n_2}$ (bottom), as a function of the logarithm of the sample sizes ratio ($log\left(\frac{n_1}{n_2}\right)$)

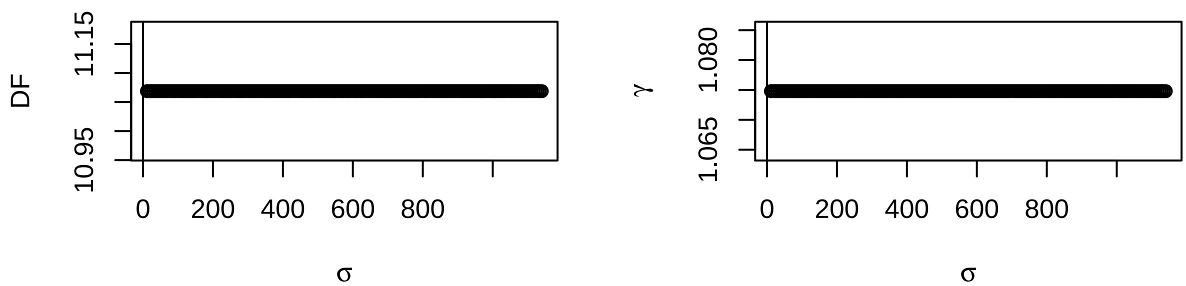

Figure 21. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances and sample sizes are unequal across groups, as a function of σ_1 and σ_2, for a constant SD-ratio