1   Theoretical Bias, as a function of population parameters

2   Delacre Marie[1]

3   [1] ULB

6   Theoretical Bias, as a function of population parameters

The bias

8   For all "biased" estimators, when the population effect size is null so is the bias. We

9   will therefore focus on configurations where there is a non-null population effect size. The

10   sampling distribution of Cohen's d_s (and therefore its bias) is only known under the

11   assumptions of normality and homoscedasticity. On the other hand, the biases of Glass's d_s,

12   Cohen's d_s^* and Shieh's d_s are theoretically known for all configurations where the normality

13   assumption is met. In order to simplify the analysis of their bias, it is convenient to

14   subdivide all configurations into 3 conditions:

15   - when population variances are equal across groups;

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

17   - when population variances are unequal across groups, with unequal sample sizes.

Preliminary note

19   For all previously mentioned estimators (Cohen's d_s, Glass's d_s, Cohen's d_s^* and Shieh's

20   d_s), the theoretical expectency is computed by multiplying the population effect size

21   (respectively δ_{Cohen}, δ_{Glass}, δ_{Cohen}^* and δ_{Shieh}) by the following multiplier coefficient:

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

22   where df are the degrees of freedom (see the main article). γ is *always* positive, meaning that

23   when the population effect size is not zero, all estimators will overestimate the population

24   effect size. Moreover, its limit tends to 1 when the degrees of freedom (df) tend to infinity,

25   meaning that the larger the degrees of freedom, the lower the bias.

26   While we focus on the theoretical bias of biased estimators when the normality

27   assumption is met, it is interesting to notice that our main conclusions seem to generalize to :

28   - biased estimators when samples are extracted from symmetric distributions;

29   - unbiased estimators when samples are extracted from heavy-tailed symmetric

30   distributions.

Cohen's d_s (see Table 1)

32   Under the assumptions that independant residuals are normally distributed with equal

33   variances, the **bias** of Cohen's d_s is a function of total sample size (N) and the population

34   effect size (δ_{Cohen}):

35   • The larger the population effect size, the more Cohen's d_s will overestimate δ_{Cohen};

36   • The larger the total sample size, the lower the bias (see Figure 1);

37   • Of course, considering the degrees of freedom, the sample size ratio does not matter

38   (i.e. the bias will decrease when increasing n_1, n_2 or both sample sizes).

Glass's d_s (see Table 2)

40   Because degrees of freedom depend only on the control group size (neither on σ_1 nor on

41   σ_2), there is no need to distinguish between cases where there is homoscedasticity or

42   heteroscedasticity!

43   The **bias** of Glass's d_s is a function of the control group size (n_c) and the population

44   effect size (δ_{Glass}):

45   • The larger the population effect size, the more Glass's d_s will overestimate δ_{Glass};

46   • The larger the size of the control group, the lower the bias (see the two top plots in

47   Figure 2). On the other hand, increasing the experimental group size does not impact

48   the bias (see the two bottom plots in Figure 2).

49 Cohen's d_s^* (see Table 2)

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

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

51 One can see that degrees of freedom depend only on the total sample size (N) and the

52 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

53 the population effect size (δ_{Cohen}^*), the sample size allocation ratio $\left(\frac{n_1}{n_2}\right)$ and the total sample

54 size (N).

55   • The larger the population effect size, the more Cohen's d_s^* will overestimate δ_{Cohen}^*;

56   • The further the sample size allocation ratio is from 1, the larger the bias (see Figure 3);

57   • The larger the total sample size, the lower the bias (see Figure 4).

58   **When variances are unequal across populations, with equal sample sizes.**

59 When $n_1 = n_2 = n$:

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

60 One can see that degrees of freedom depend only on the total sample size (N) and the

61 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

62 effect size (δ_{Cohen}^*), the SD-ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$ and the total sample size (N):

63   • The larger the population effect size, the more Cohen's d_s^* will overestimate δ_{Cohen}^*;

64   • The further the SD-ratio is from 1, the larger the bias (see Figure 5);

65   • The larger the total sample size, the lower the bias (see Figure 6).

66   Note: for a constant SD-ratio, σ_1 and σ_2 do not matter (see Figure 7).

[67] **When variances are unequal across populations, with unequal sample sizes.**

[68] The **bias** of Cohen's d_s^* is a function of the population effect size (δ_{Cohen}^*), the total sample

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

[70] • The larger the population effect size, the more Cohen's d_s^* will overestimate δ_{Cohen}^*;

[71] • The larger the total sample size, the lower the bias (see Figure 8);

[72] • The smallest bias always occurs when there is a positive pairing between variances and

[73]   sample sizes, because one gives more weight to the smallest variance, in the

[74]   denominator of the df computation. Moreover, the further the SD-ratio is from 1, the

[75]   further from 1 will also be the sample sizes ratio associated with the smallest bias (see

[76]   Figure 9). This can be explained by splitting the numerator and the denominator in

[77]   the df computation.

[78] As illustrated in Figure 10, for any SD-ratio, the numerator of the degrees of freedom

[79] will be maximized when sample sizes are equal across groups (and is not impacted by the

[80] SD-ratio). On the other hand, the denominator will be minimized when there is a positive

[81] pairing between variances and sample sizes. For example, when $\sigma_1 > \sigma_2$, the smallest

[82] denominator occurs when $\frac{n_1}{n_2}$ reaches its maximum value and the larger the SD-ratio, the

[83] larger the impact of the sample sizes ratio on the denominator.

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

[85] **Shieh's d_s (see Table 2)**

[86] **When variances are equal across populations.**   When $\sigma_1 = \sigma_2 = \sigma$:

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

[87]
$$\leftrightarrow df_{Shieh's\ d_s} = \frac{[\sigma^2(n_1+n_2)]^2}{n_1^2 n_2^2} \times \frac{(n_1-1)(n_2-1)}{(n_2-1)\times\frac{\sigma^4}{n_1^2}+(n_1-1)\times\frac{\sigma^4}{n_2^2}}$$

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

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

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

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 such 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 $n_1 = n_2 = n$:

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

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

107
$$\leftrightarrow df_{Shieh's\ d_s} = \frac{(\sigma_1^2 + \sigma_2^2)^2 \times (n-1)}{\sigma_1^4 + \sigma_2^4}$$

108   One can see that degrees of freedom depend on the total sample size (N) and the

109 SD-ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$. As a consequence, the bias depends on the population effect size (δ_{Shieh}), the

110 SD-ratio $\left(\frac{\sigma_1}{\sigma_2}\right)$ and the total sample size (N).

111   • The larger the population effect size, the more Shieh's d_s will overestimate δ_{Shieh};

112   • The further the SD-ratio is from 1, the larger the bias (see Figure 15);

113   • The larger the total sample size, the lower the bias (see Figure 16);

114   Note: for a constant SD-ratio, σ_1 and σ_2 do not matter (see Figure 17).

115   **When variances are unequal across populations, with unequal sample sizes.**

116 The **bias** of Shieh's d_s^* is a function of the population effect size (δ_{Shieh}), the sample sizes (n_1

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

118   • The larger the population effect size, the more Shieh's d_s will overestimate δ_{Shieh};

119   • For a constant sample sizes ratio, the larger the sample sizes, the lower the bias (See

120   Figure 18);

121   Note: When variances were equal across populations, adding subjects only in the first

122 group had the same impact on degrees of freedom (and therefore on bias) than adding

123 subjects only in the second group (see Figure 14). When variances are unequal across groups,

124 this is not true anymore (see Figure 19).

125   • The smallest bias always occurs when there is a positive pairing between variances and

126   sample sizes. Moreover, the further the SD-ratio is from 1, the further from 1 will also

127   be the sample sizes ratio associated with the smallest bias (See Figure 20);

128   Moreover, for a constant SD-ratio, the variances do not matter (See Figure 21).

In summary

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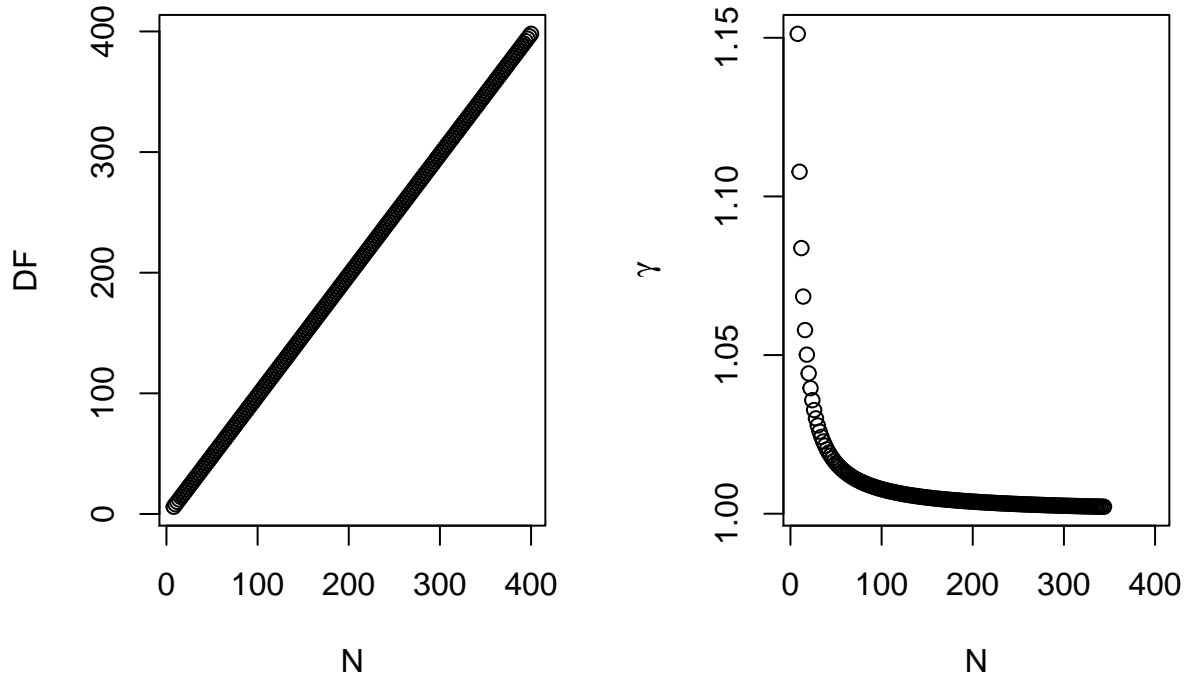
131   sample size (N):

132   • When δ_{Cohen} is null, the bias is null. In all other configurations, the larger δ_{Cohen}, the

133   more Cohen's d_s will overestimate δ_{Cohen};

134   • The bias decreases when the total sample size increases (it does not matter whether

135   one adds subjects in only one group or in both).

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

137   the control group (n_e):

138   • When δ_{Glass} is null, the bias is null. In all other configurations, the larger δ_{Glass}, the

139   more Glass's d_s will overestimate δ_{Glass};

140   • The bias decreases when the size of the control group increases. On the other hand,

141   increasing the size of the experimental group does not impact the bias.

142   The **bias** of Cohen's d_s^* is a function of the population effect size (δ_{Cohen}^*), the total

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

144   • When δ_{Cohen}^* is null, the bias is null. In all other configurations, the larger δ_{Cohen}^*, the

145   more Cohen's d_s^* will overestimate δ_{Cohen}^*;

146   • The bias decreases when the total sample size increases (it does not matter whether

147   one adds subjects in only one group or in both);

148   • The smallest bias always occurs when there is a positive pairing between $\frac{\sigma_1}{\sigma_1}$ and $\frac{n_1}{n_1}$.

149   Moreover, the larger the SD-ratio, the further from 1 is the sample sizes ratio

150   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 $\left(\frac{n_1}{n_2} \times \frac{\sigma_1}{\sigma_2}\right)$:

- 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 $\frac{\sigma_1}{\sigma_1}$ and $\frac{n_1}{n_1}$. 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").
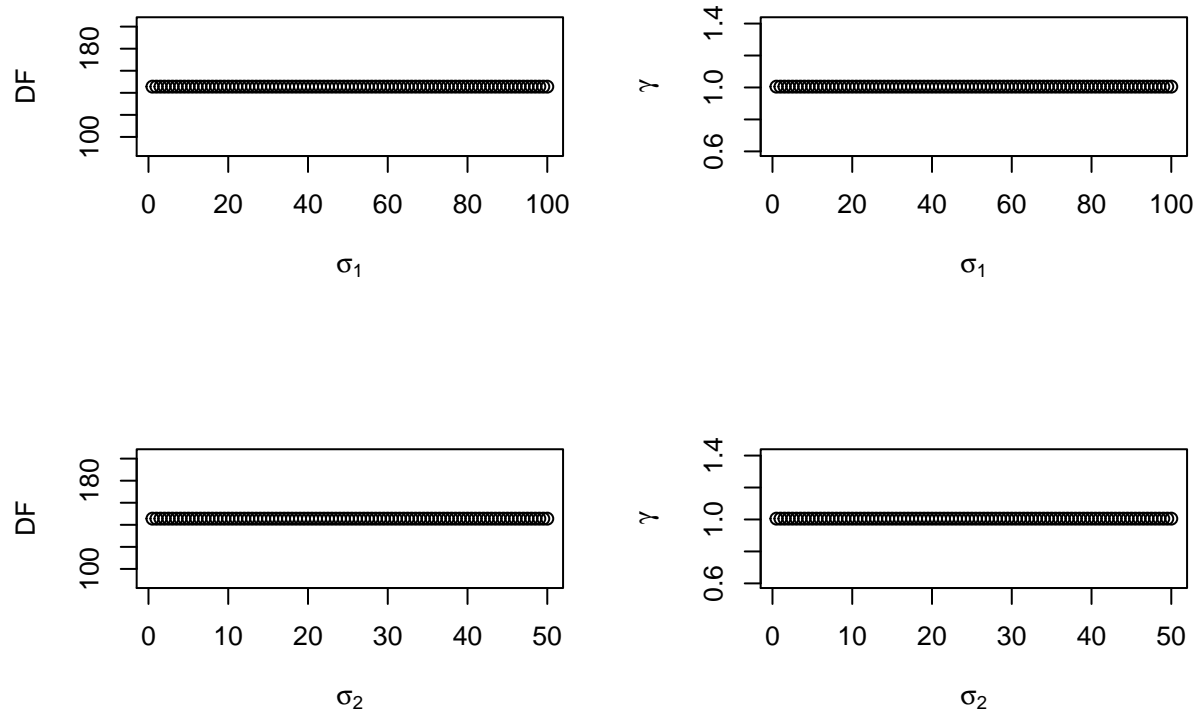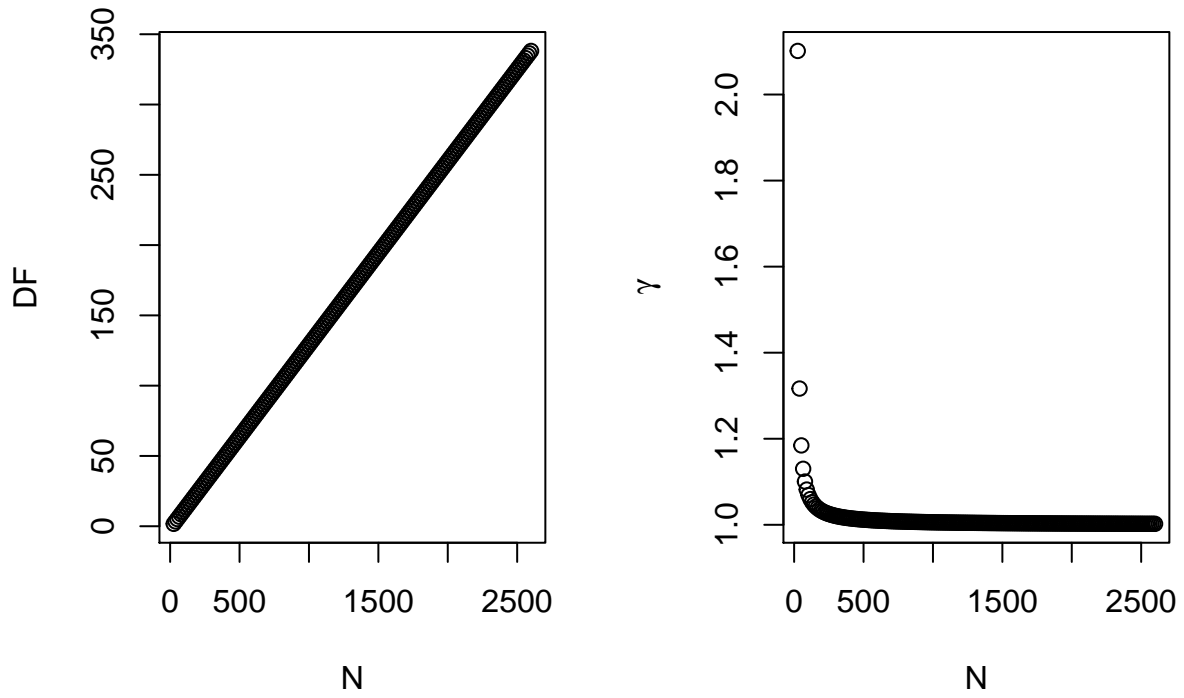
Figure 1. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s, when variances are equal across groups, as a function of the total sample size (N).
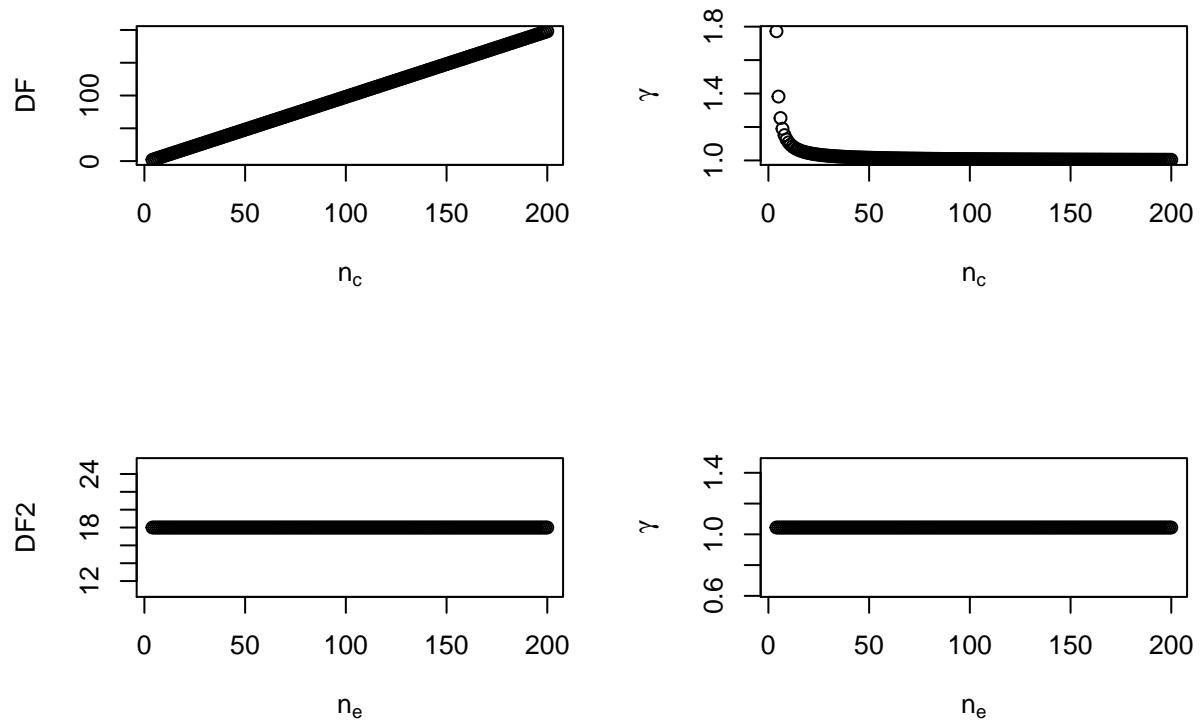
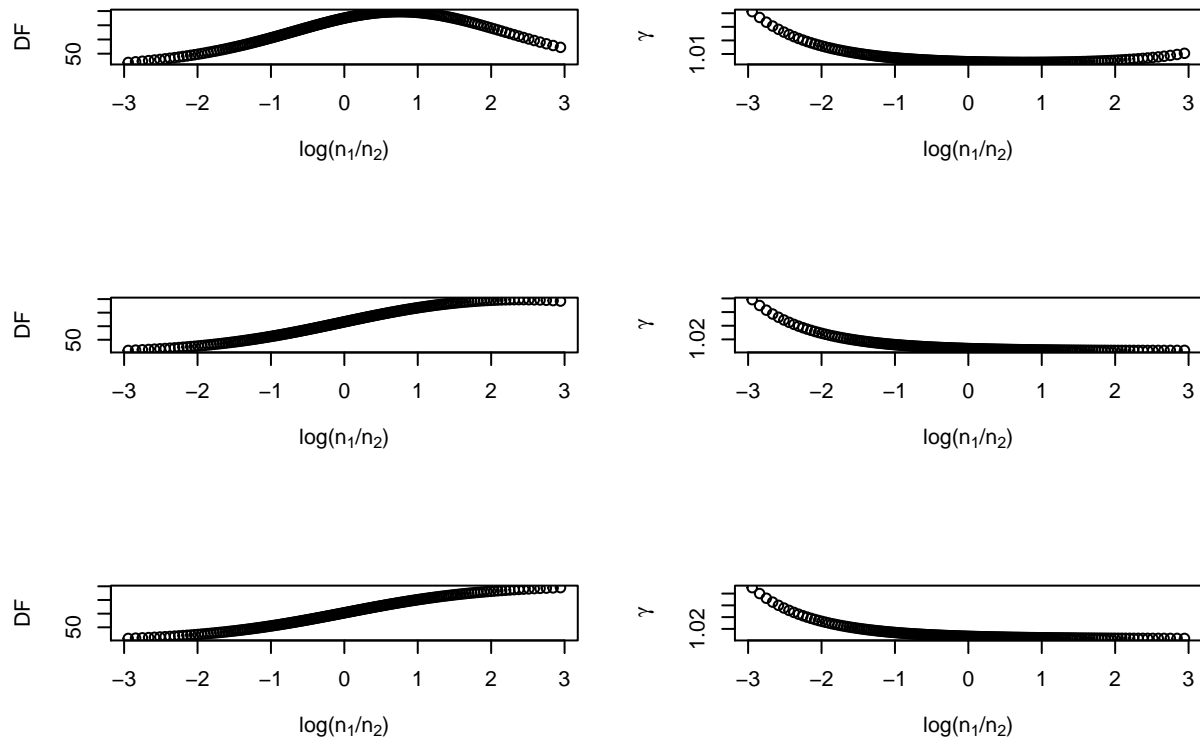*Figure 2*. Degrees of freedom (DF) and γ, when computing the bias of Glass's d_s, when variances are equal across groups, as a function of n_c (top) and n_e (bottom).
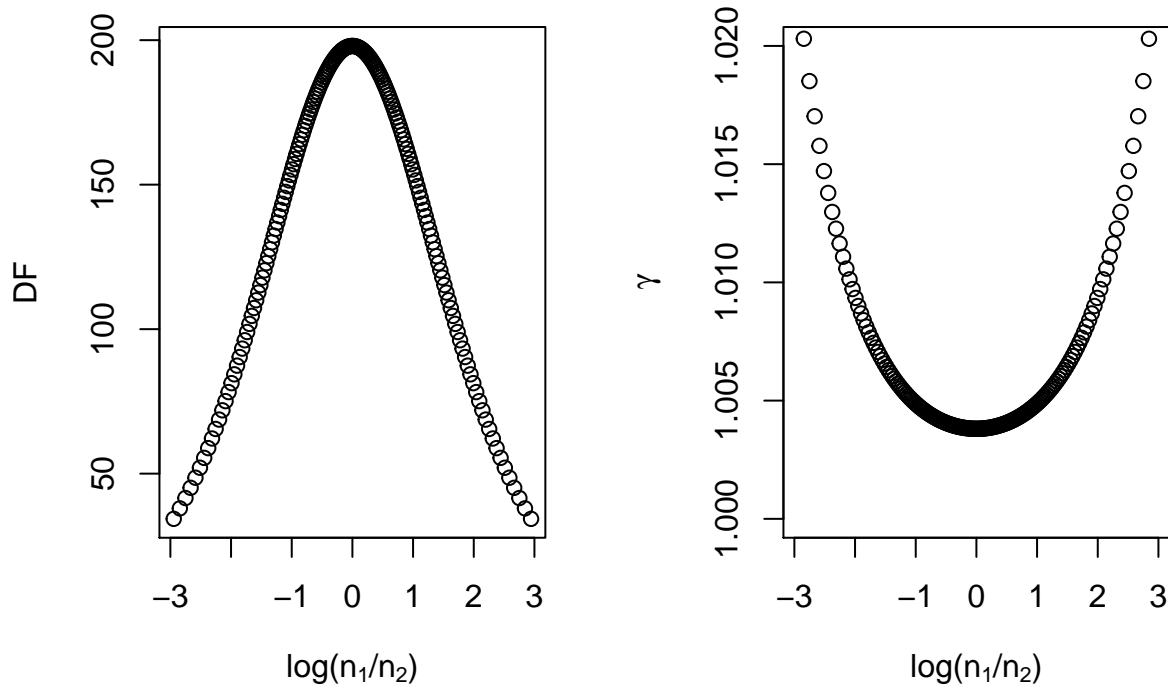
Figure 3. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s^*, when variances are equal across groups, as a function of the logarithm of the sample sizes ratio $log\left(\frac{n_1}{n_2}\right)$.
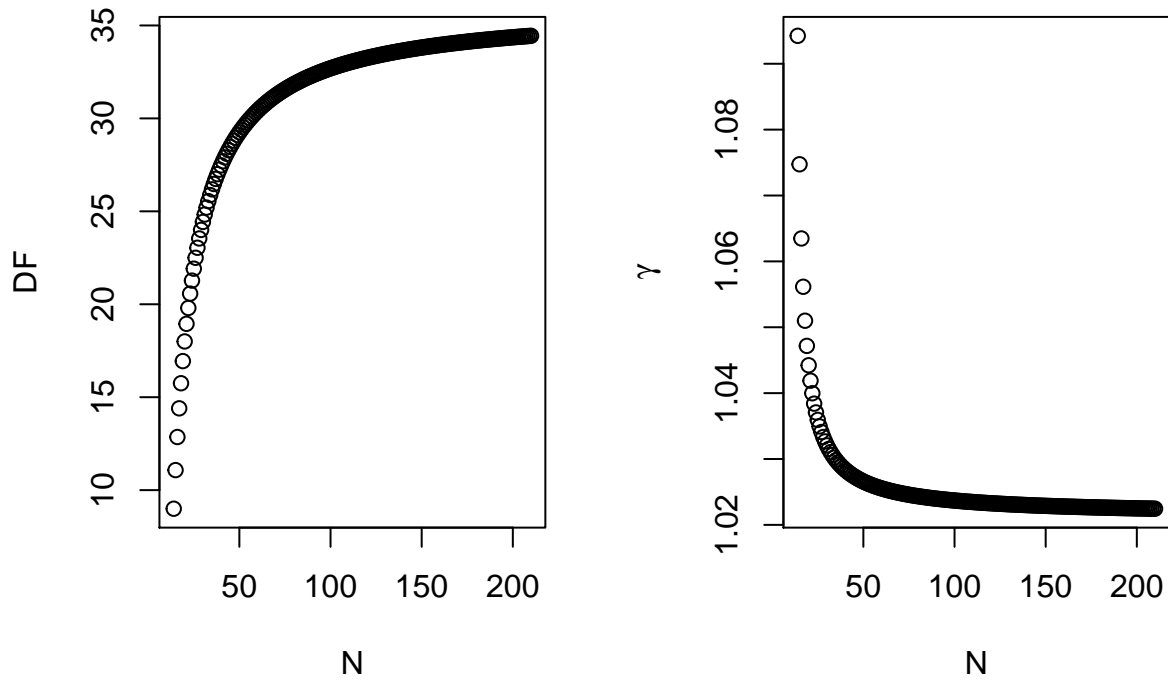
Figure 4. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s^*, when variances are equal across groups, as a function of the total sample size (N).
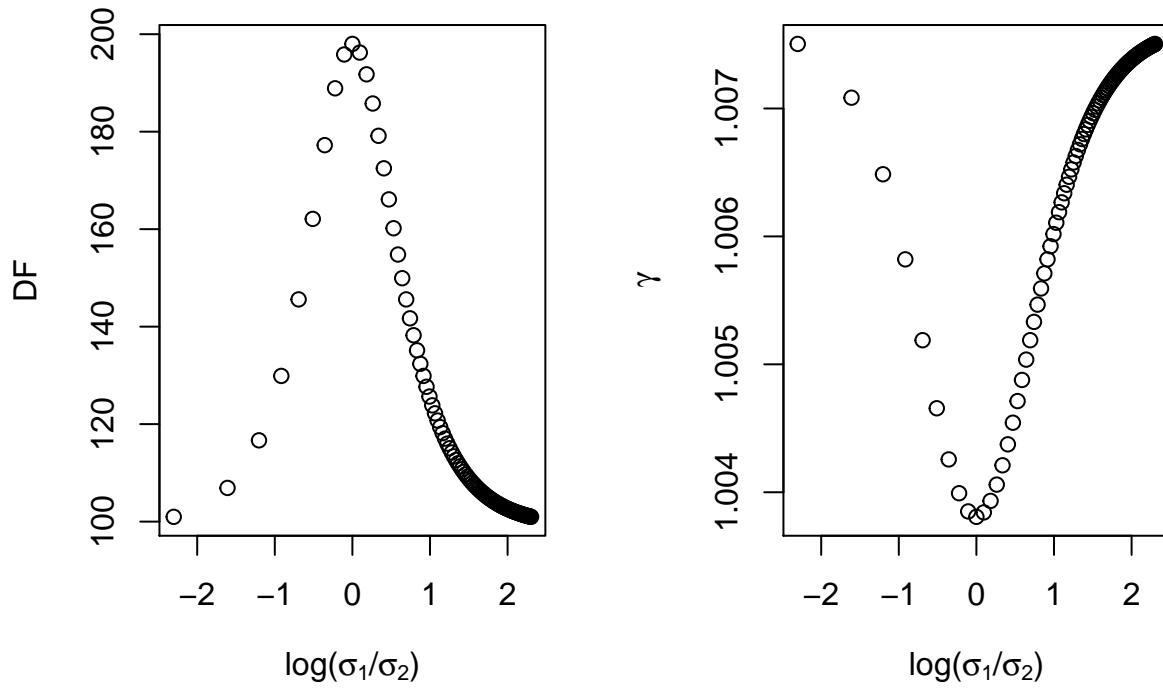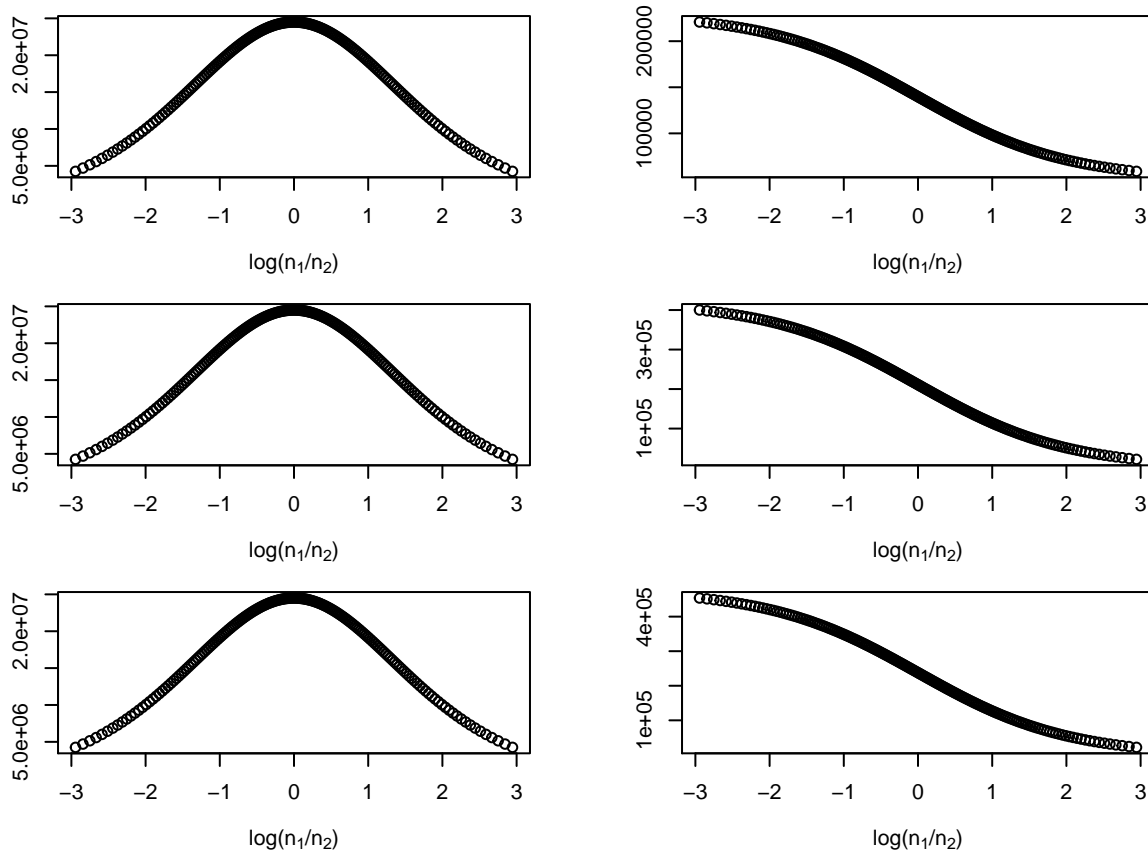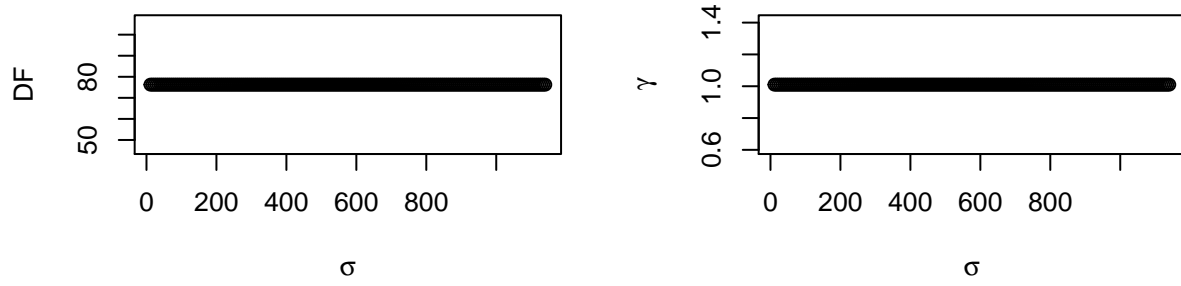
Figure 5. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s^*, when variances are unequal across groups and sample sizes are equal, as a function of the logarithm of the *SD*-ratio ($log\left(\frac{\sigma_1}{\sigma_2}\right)$).
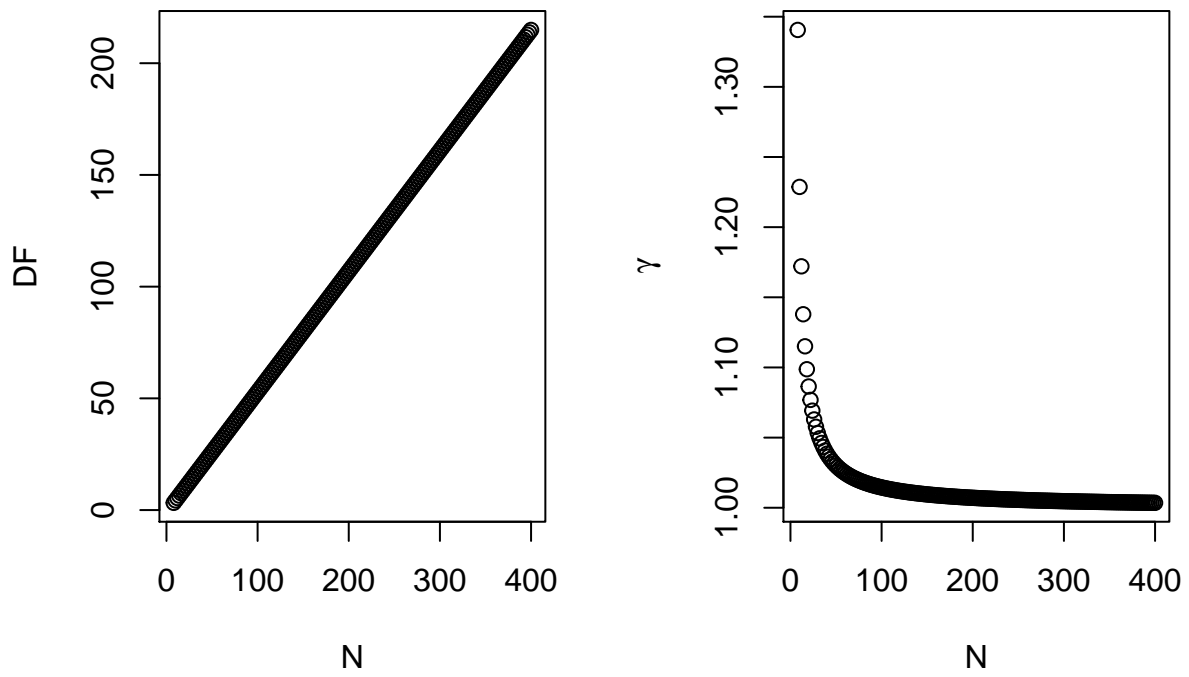
Figure 6. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s^*, when variances are unequal across groups and sample sizes are equal, as a function of the total sample size (N).
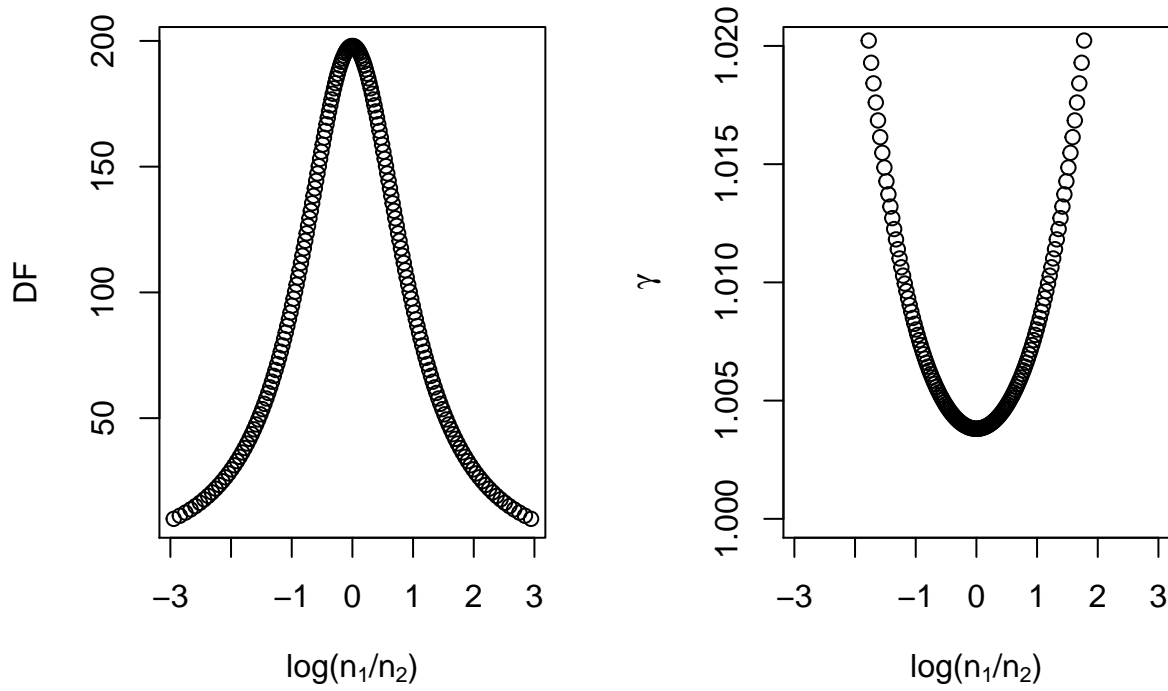
Figure 7. Degrees of freedom (DF) and γ, when computing the bias of Cohen's d_s^*, when variances are unequal across groups and sample sizes are equal, as a function of σ_1 (top plots) and σ_2 (bottom plots), for a constant SD-ratio.

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.
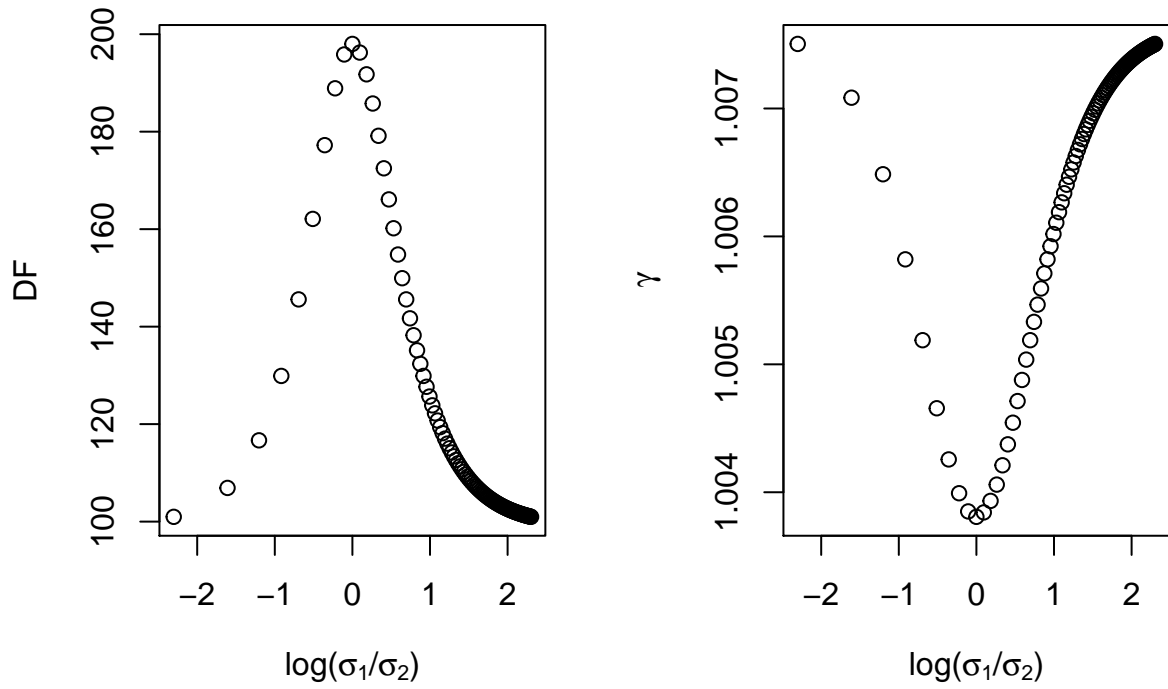
Figure 12. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are equal across groups, as a function of the logarithm of the sample sizes ratio $(log\left(\frac{n_1}{n_2}\right))$.
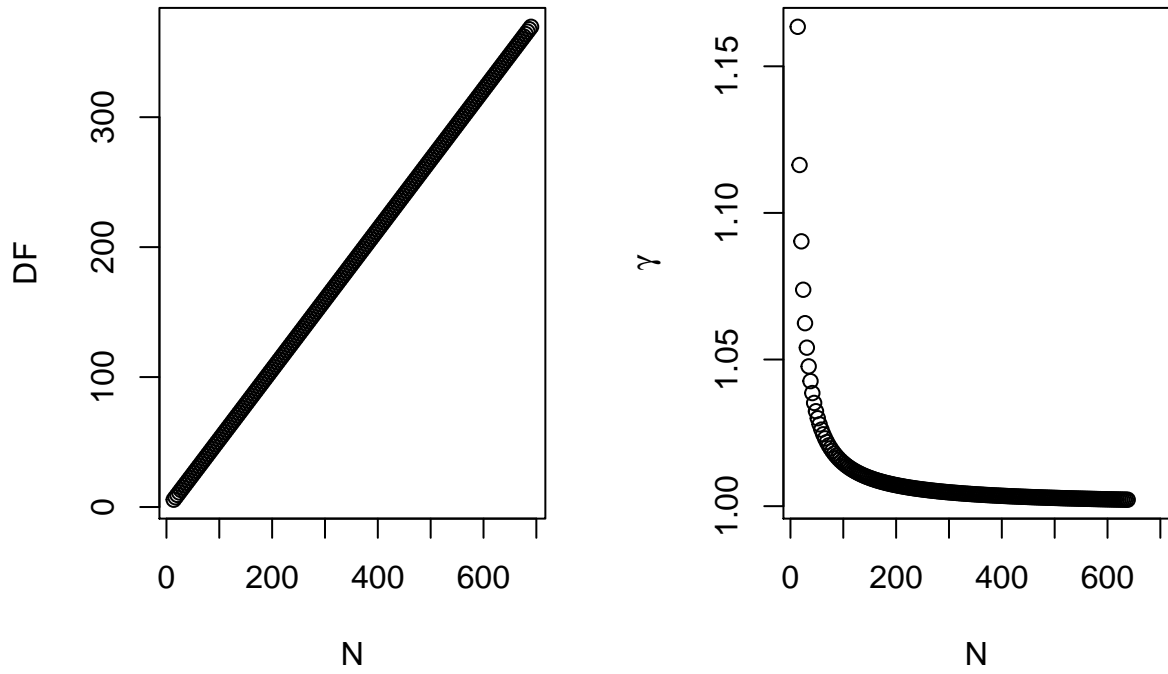
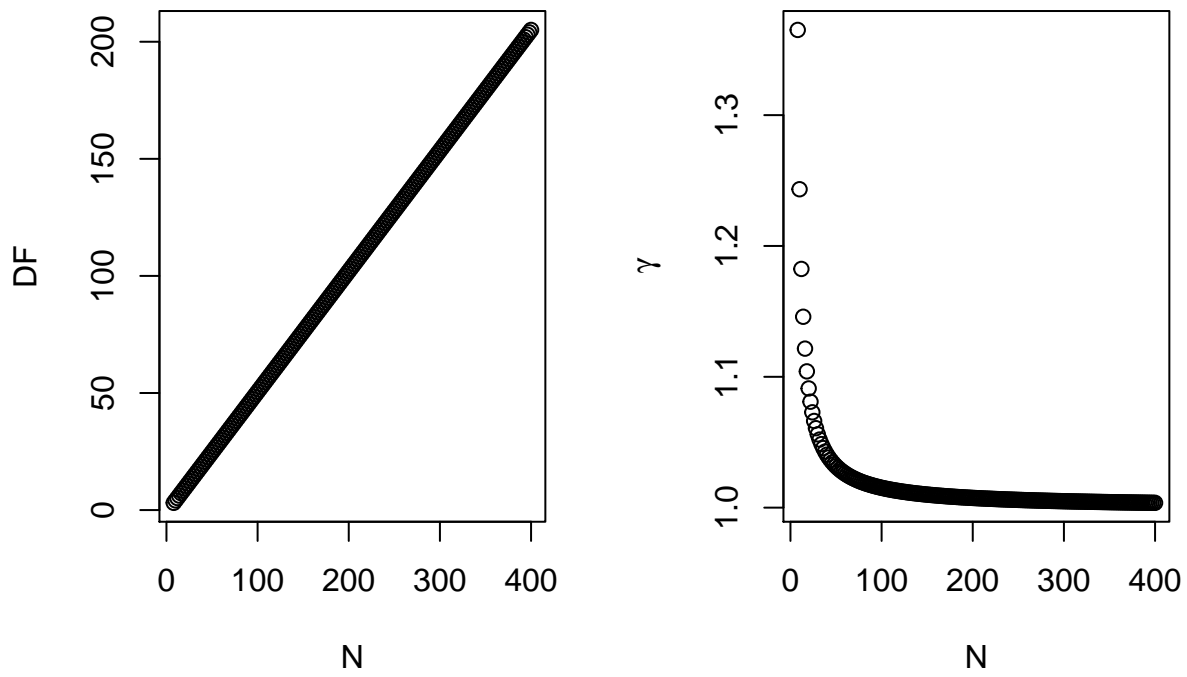*Figure 13*. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are equal across groups, as a function of the total sample size (N).
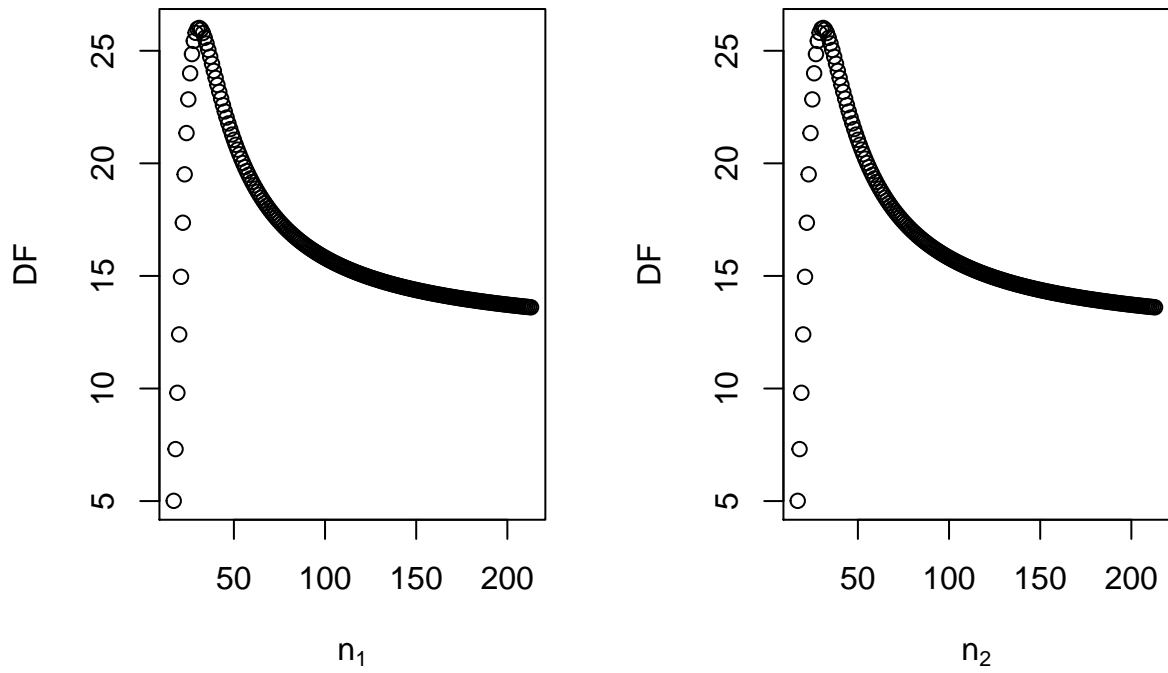
Figure 14. Degrees of freedom (DF), when computing the bias of Shieh's d_s, when variances are equal across groups, when adding subjects only in the first group (left) or in the second group (right).
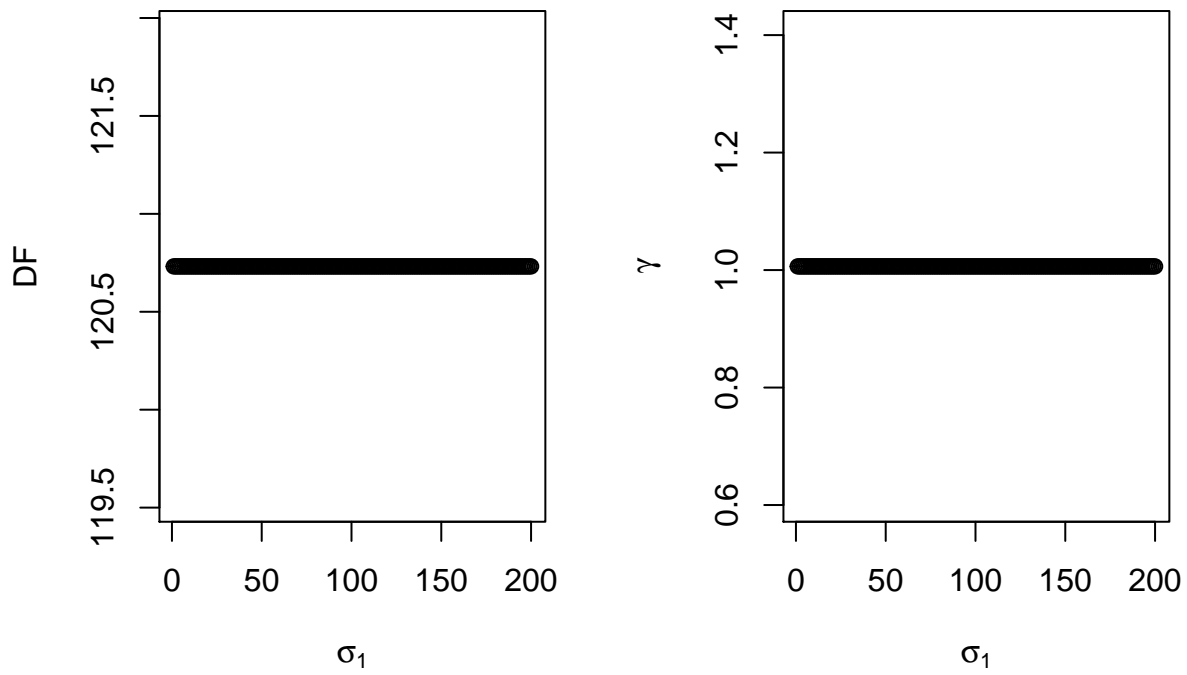
Figure 15. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are unequal across groups and sample sizes are equal, as a function of the logarithm of the SD-ratio $\left(log\left(\frac{\sigma_1}{\sigma_2}\right)\right)$.

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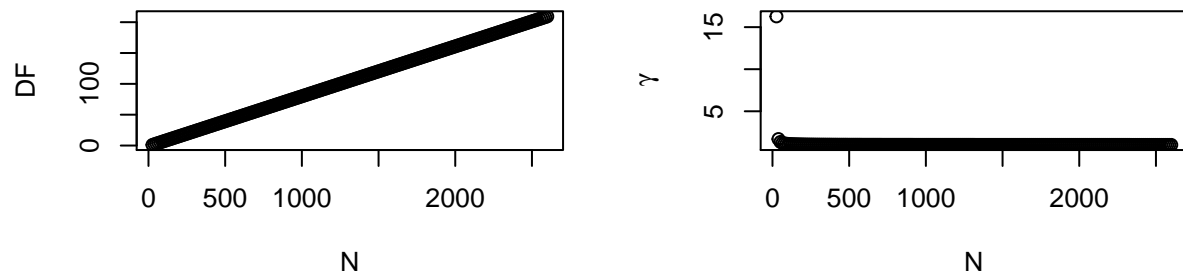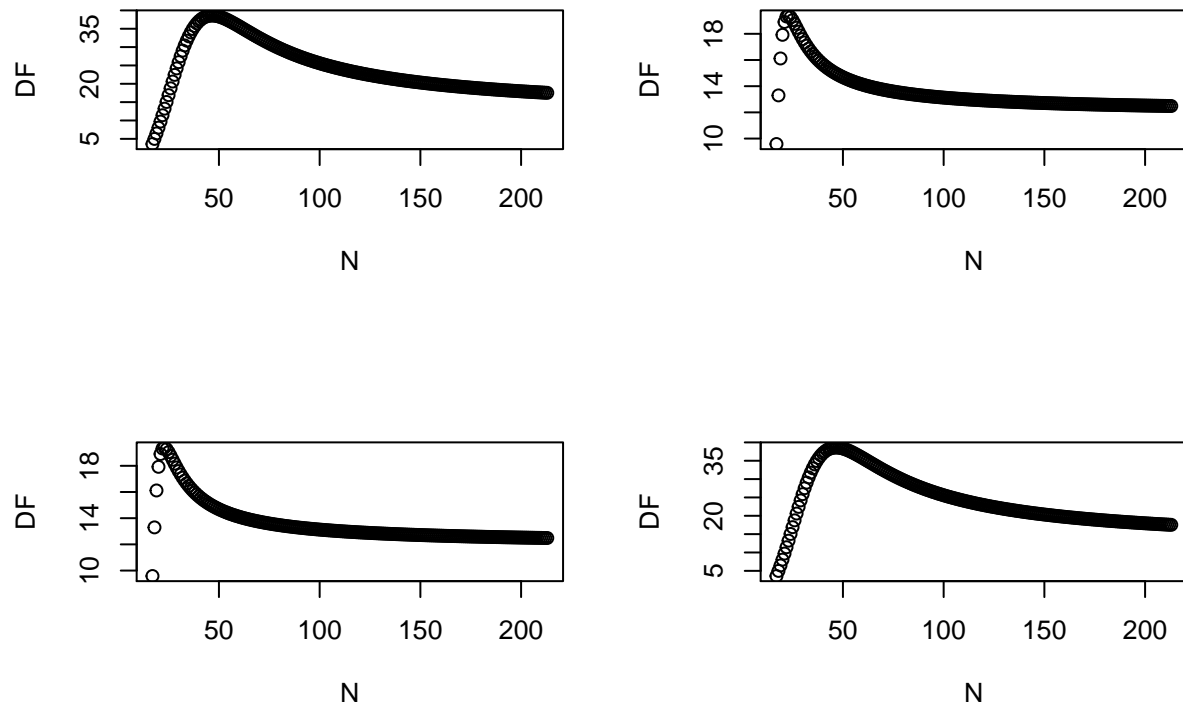
Figure 17. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances are unequal across groups and sample sizes are equal, as a function of σ_1, for a constant SD-ratio.

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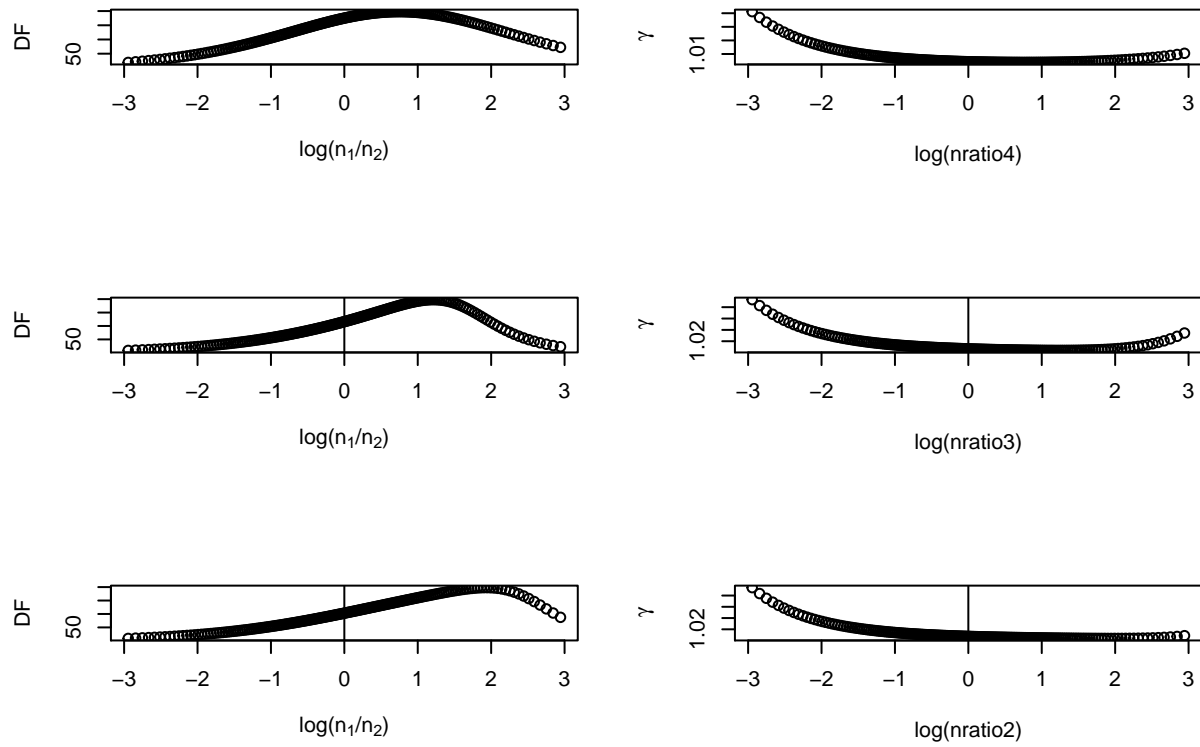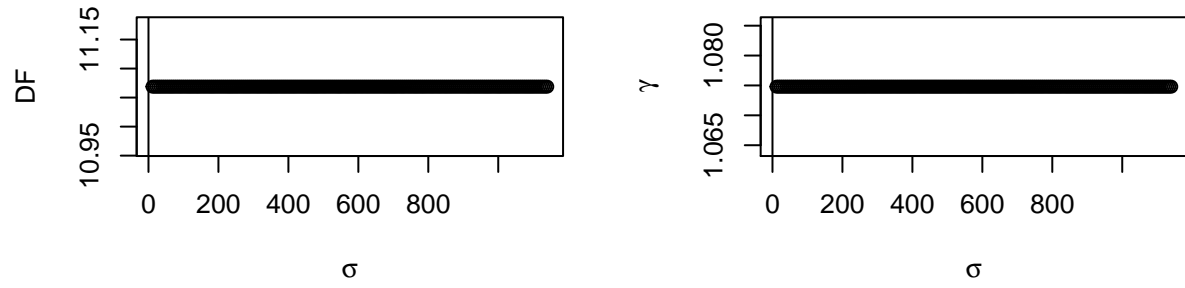
Figure 20. Degrees of freedom (DF) and γ, when computing the bias of Shieh's d_s, when variances and sample sizes are unequal across groups, as a function of the logarithm of the sample sizes ratio $\left(log\left(\frac{n_1}{n_2}\right)\right)$, when SD-ratio equals 1.46 (first row), 3.39 (second row)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