Annexes

Annexes du Chapitre 3

Appendix 1 : The Mathematical Development of the F-test, W-test, and F^*-test: Numerical Example

Descriptive statistics are presented in Table A1. The raw data are available here : https://github.com/mdelacre/W-ANOVA/tree/master/Functions (see "practical example.R"). The dependent variable is a score that can vary from 0 to 40. The independent variable is a three-level factor A (levels = A_1, A_2 and A_3).

Table A1. *Summary of the data of the fictive case*

	A1	A2	A3
n_j	41	21	31
$\bar{X}$	24	23	27
S^2	81.75	10.075	38.40

The overall mean (i.e. the mean of the global dataset) is a weighted mean of the sample means:

$$\bar{X}_{..} = \frac{(41 \times 24) + (21 \times 23) + (31 \times 27)}{41 + 21 + 31} = \frac{2304}{93} \approx 24.77$$

The F-test statistic and degrees of freedom are computed by applying equation 1 and related degrees of freedom :

$$F = \frac{\frac{1}{3-1}[41 \times (24 - \frac{2304}{93})^2 + 21 \times (23 - \frac{2304}{93})^2 + 31 \times (27 - \frac{2304}{93})^2]}{\frac{1}{93-3}[(41-1) \times 81.75 + (21-1) \times 10.075 + (31-1) \times 38.4]} \approx 2.377$$

$$df_n = 3 - 1 = 2$$

$$df_d = 93 - 3 = 90$$

The F^*-test statistic and his degrees of freedom are computed by applying equation 2 and related degrees of freedom :

$$F^* = \frac{41 \times (24 - \frac{2304}{93})^2 + 21 \times (23 - \frac{2304}{93})^2 + 31 \times (27 - \frac{2304}{93})^2}{(1 - \frac{41}{93}) \times 81.75 + (1 - \frac{21}{93}) \times 10.075 + (1 - \frac{31}{93}) \times 38.4} \approx 3.088$$

$$df_n = 3 - 1 = 2$$

$$df_d = \frac{1}{\frac{(\frac{(1-\frac{41}{93})\times 81.75}{\sum_{j=1}^{k}(1-\frac{n_j}{N})S_j^2})^2}{41-1} + \frac{(\frac{(1-\frac{21}{93})\times 10.075}{\sum_{j=1}^{k}(1-\frac{n_j}{N})S_j^2})^2}{21-1} + \frac{(\frac{(1-\frac{31}{93})\times 38.4}{\sum_{j=1}^{k}(1-\frac{n_j}{N})S_j^2})^2}{31-1}} \approx 81.149$$

where $\sum_{j=1}^{k}(1 - \frac{n_j}{N}) \times S_j^2 \approx 79.11$

Finally, the W-test and his degrees of freedom are computed by applying equation 3 and related degrees of freedom :

$$W = \frac{\frac{1}{3-1}[\frac{41}{81.75}(24 - \bar{X}')^2 + \frac{21}{10.075}(23 - \bar{X}')^2 + \frac{31}{38.4}(27 - \bar{X}')^2]}{\frac{2(3-2)}{3^2-1}[(\frac{1}{41-1})(1 - \frac{\frac{41}{81.75}}{w})^2 + (\frac{1}{21-1})(1 - \frac{\frac{21}{10.075}}{w})^2 + (\frac{1}{31-1})(1 - \frac{\frac{31}{38.4}}{w})^2] + 1} \approx 4.606$$

where: $w = \sum_{j=1}^{k} w_j \approx 3.39$ and $\bar{X}' = \frac{\sum_{j=1}^{k}(w_j \bar{X}_j)}{w} \approx 24.1$

$$df_n = 3 - 1$$

$$df_d = \frac{3^2 - 1}{3[\frac{(1-\frac{w_j}{w})^2}{41-1} + \frac{(1-\frac{w_j}{w})^2}{21-1} + \frac{(1-\frac{w_j}{w})^2}{31-1}]} \approx 59.32$$

One should notice that in this example, the biggest sample size has the biggest variance. As previously mentioned, it means that the F-test will be too conservative, because the F value decreases. The F^*-test will also be a little too conservative, even if the test is less affected than the F-test. As a consequence: $W > F^* > F$.

Appendix 2 : Justification for the choice of distributions in simulations

The set of simulations described in the article was repeated for 7 distributions. We used R commands to generate data from different distributions:

— k normal distributions (Figure A2.1) : in order to assess the Type I error rate and power of the different tests under the assumption of normality, data were generated by means of the function "rnorm" (from the package "stats"; "R: The Normal Distribution," 2016).

— k double exponential distributions (Figure A2.2) : In order to assess the impact of high kurtosis on the Type I error rate and power of all tests, data were generated by means of the function "rdoublex" (from the package "smoothmest"; "R: The double exponential (Laplace) distribution," 2012).

— k mixed normal distributions (Figure A2.3) : In order to assess the impact of extremely high kurtosis on the Type I error rate and power of all tests, regardless of variance, data were generated by means of the function "rmixnorm" (from the package "bda"; Wang & Wang, 2015).

— k normal right skewed distributions (Figure A2.4) : In order to assess the impact of moderate skewness on the Type I error rate and power, data were generated by means of the function "rsnorm" (from the package "fGarch"; "R: Skew Normal Distribution," 2017). The normal skewed distribution was chosen because it is the only skewed distribution where the standard deviation ratio can vary without having an impact on skewness.

— k-1 normal left skewed distributions (Figure A2.5) and 1 normal right skewed distribution (Figure A2.4) : In order to assess the impact of unequal shapes, in terms of skewness, on the Type I error rate and power, when data have moderate skewness, data were generated by means of the functions "rsnorm" (from the package "fGarch"; "R: Skew Normal Distribution," 2017).

— k-1 chi-squared distributions with two degrees of freedom (See Figure A2.6), and one normal rigt skewed distribution (Figure A2.4) : In order to assess the impact of high asymetry on the Type I error rate an power, k-1 distributions were generated by means of the functions "rchisq" ("R: The (non-central) Chi-squared Distribution," 2016). The last distribution was generated by means of "rsnorm" in order to follow a normal right skewed distribution with a mean of 2 (from the package "fGarch"; "R: Skew Normal Distribution," 2017). Because the chi-squared is non-negative, it is not possible to generate chi-squared where population SD= 1, 4 or 8 and population mean is the same than the chi-squared with two degrees of freedom. However, we wanted to assess the impact of different SD-ratio on Type I error rate. For these reasons, the last distribution was generated by means of "rsnorm" in order to follow a normal skewed distribution with positive skewness of +0.99 and mean = 2 (from the package "fGarch"; "R: Skew Normal Distribution," 2017).

— k-1 chi-squared distributions with two degrees of freedom (See Figure A2.6), and one normal left skewed distribution (Figure A2.5) : In order to assess the impact of unequal shapes, in terms of skewness, on Type I error rate and power when distributions have extreme skewness, k-1 distributions were generated by means of the functions "rchisq" ("R: The (non-central) Chi-squared Distribution," 2016). The last distribution was generated by means of "rsnorm" in order to follow a normal right skewed distribution with a mean of 2 (from the package "fGarch"; "R: Skew Normal Distribution," 2017)

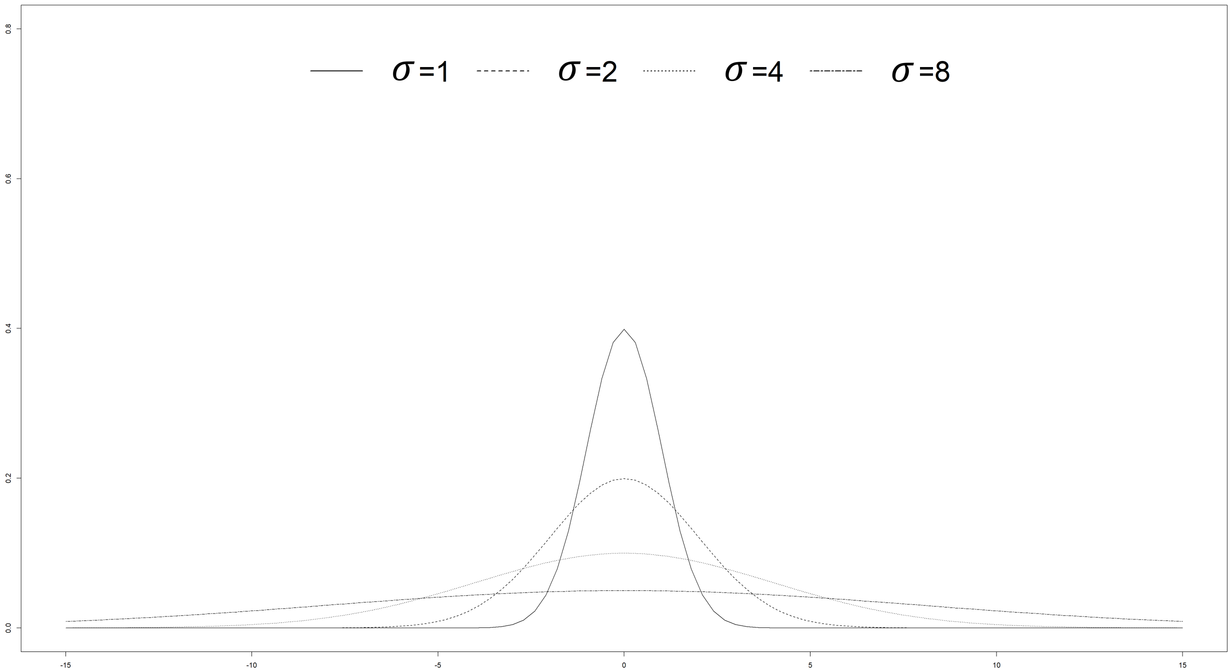

Figure A2.1 : centered normal probability density function, as a function of the population SD

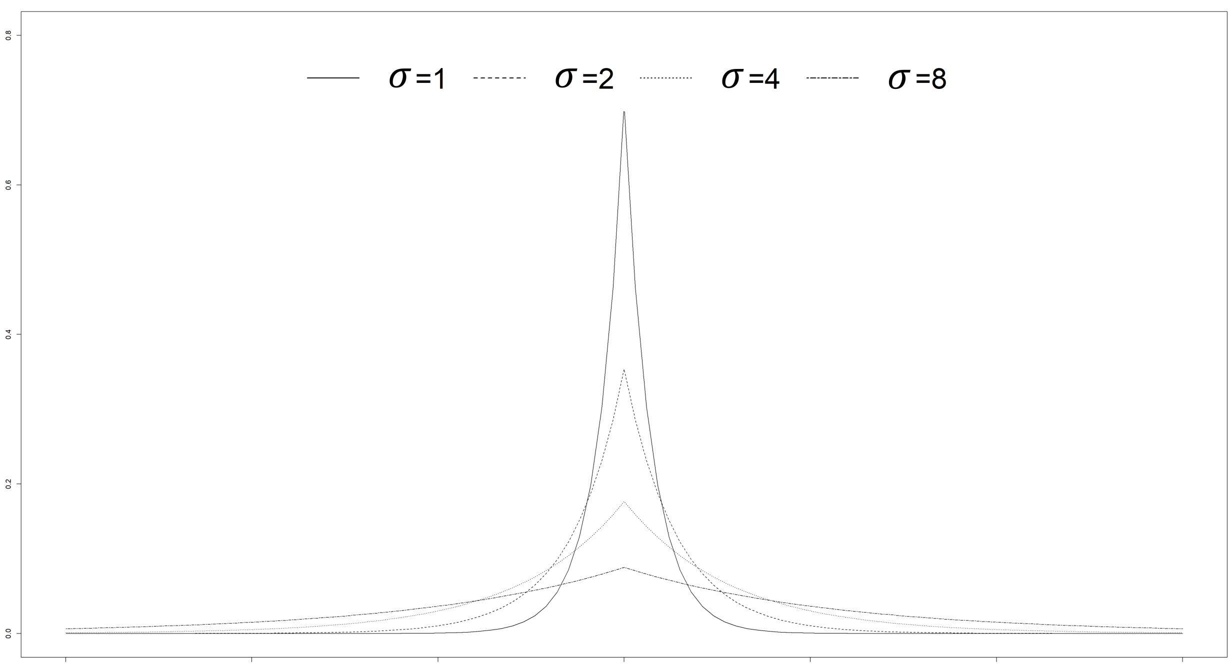

Figure A2.2 : centered double exponential probability density function, as a function of the population SD

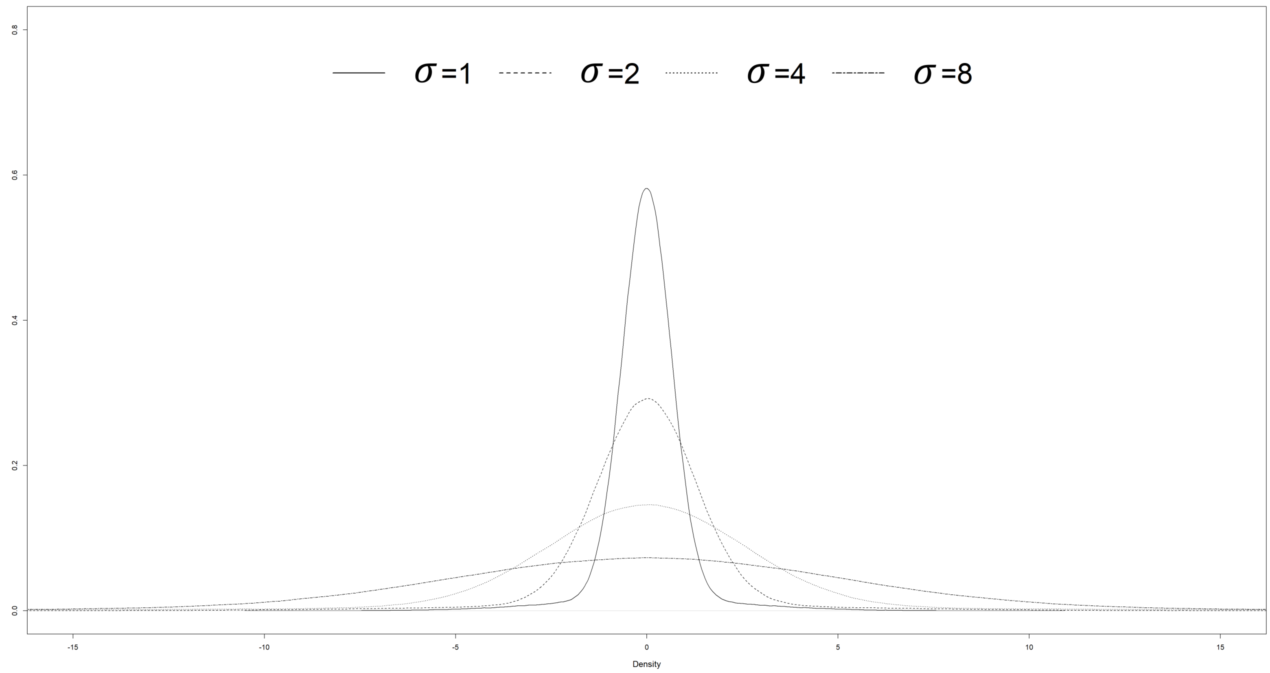

Figure A2.3 : centered mixed normal probability density function, as a function of the population SD

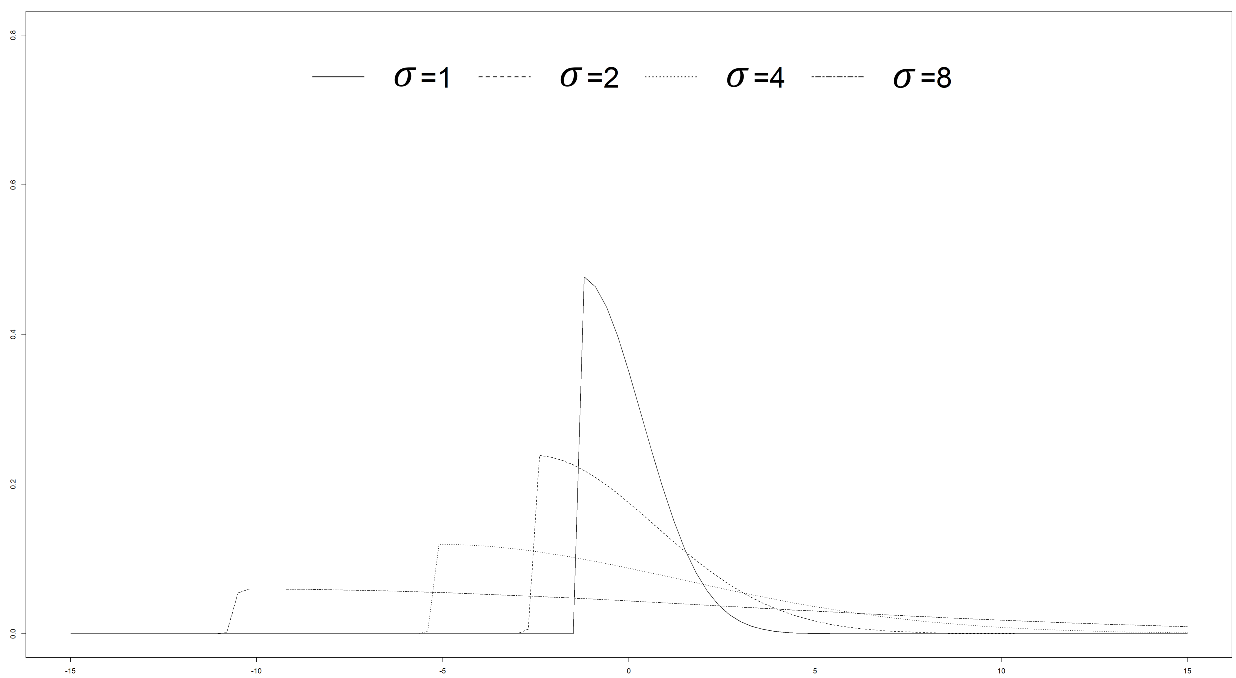

Figure A2.4 : centered normal right skewed probability density function, as a function of the population SD

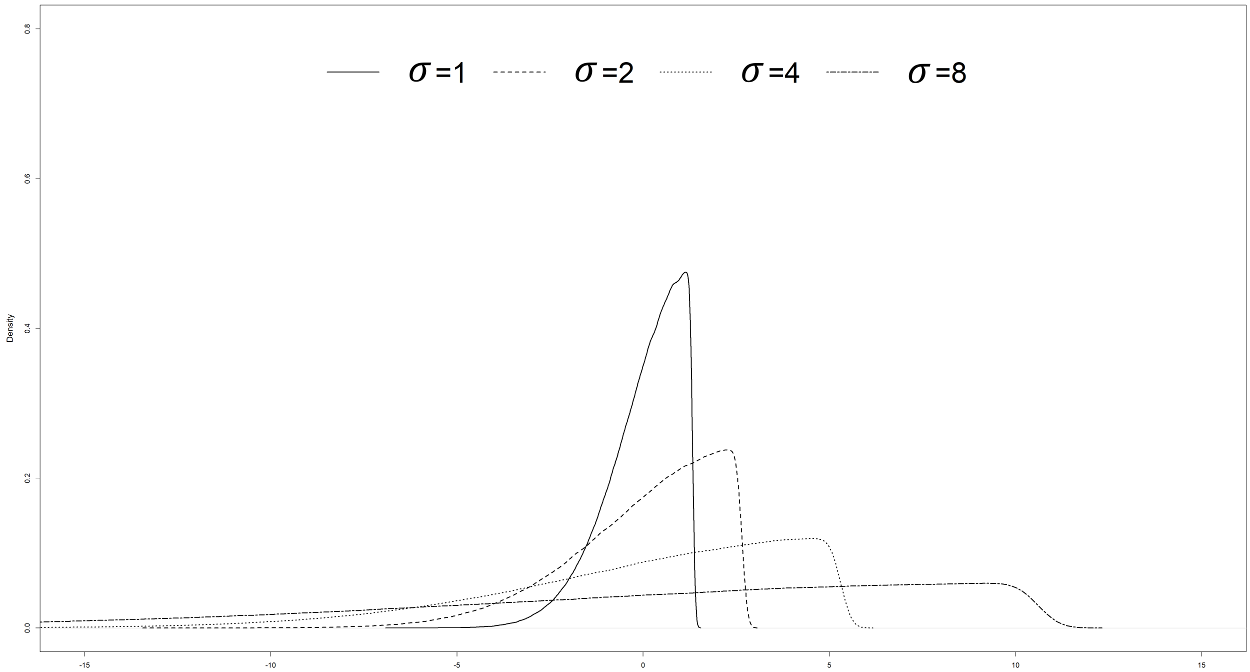

Figure A2.5 : centered normal left skewed probability density function, as a function of the population SD

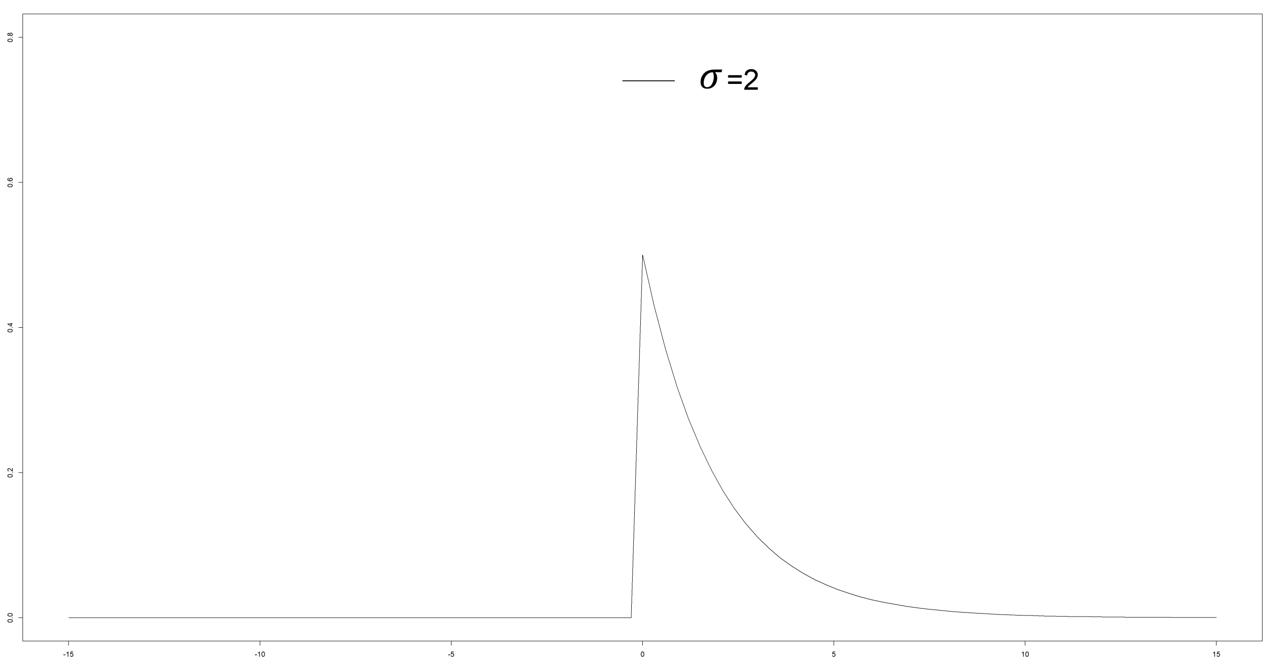

Figure A2.6 : chi-squared with 2 degrees of freedom probability density function, when the population SD equals 2

Annexes du Chapitre 4

Appendix 1 : The bias of Cohen's *d* is twice as large as the bias of Shieh's *d* when population variances and sample sizes are equal across groups: mathematical demonstration.

As mentioned in Table 2, the bias of Cohen's d is defined as

$$Bias_{Cohen's\ d} = \delta_{Cohen} \times \left(\frac{\sqrt{\frac{df_{Student}}{2}} \times \Gamma\left(\frac{df_{Student}-1}{2}\right)}{\Gamma\left(\frac{df_{Student}}{2}\right)} - 1 \right) \tag{20}$$

with

$$\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{(n_1-1)\times\sigma_1^2+(n_2-1)\times\sigma_2^2}{n_1+n_2-2}}}$$

and

$$df_{Student} = n_1 + n_2 - 2$$

As mentioned in Table 3, the bias of Shieh's d is defined as

$$Bias_{Shieh's\ d} = \delta_{Shieh} \times \left(\frac{\sqrt{\frac{df_{Welch}}{2}} \times \Gamma\left(\frac{df_{Welch}-1}{2}\right)}{\Gamma\left(\frac{df_{Welch}}{2}\right)} - 1 \right) \tag{21}$$

with

$$\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{\sigma_1^2}{n_1/N} + \frac{\sigma_2^2}{n_2/N}}} \quad (N = n_1 + n_2)$$

and

$$df_{Welch} = \frac{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)^2}{\frac{(\sigma_1^2/n_1)^2}{n_1-1} + \frac{(\sigma_2^2/n_2)^2}{n_2-1}}$$

When $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$, δ_{Cohen} is twice larger than δ_{Shieh}, as shown below in equations 22 and 23:

$$\delta_{Cohen} = \frac{\mu_1 - \mu_2}{\sqrt{\frac{2(n-1)\sigma^2}{2(n-1)}}} = \frac{\boldsymbol{\mu_1 - \mu_2}}{\boldsymbol{\sigma}} \tag{22}$$

$$\delta_{Shieh} = \frac{\mu_1 - \mu_2}{\sqrt{2\left(\frac{\sigma^2}{n/(2n)}\right)}} = \frac{\boldsymbol{\mu_1 - \mu_2}}{\boldsymbol{2\sigma}} \tag{23}$$

Moreover, degrees of freedom associated with Student's t-test and Welch's t-test are identical, as shown below in equations 24 and 25:

$$df_{Student} = \boldsymbol{2(n-1)} \tag{24}$$

$$df_{Welch} = \frac{\left[2(\sigma^2/n)\right]^2}{\frac{2(\sigma^2/n)^2}{n-1}} = \boldsymbol{2(n-1)} \tag{25}$$

Equations 20 and 21 can therefore be redefined as follows:

$$Bias_{Cohen's\ d} = \frac{\mu_1 - \mu_2}{\sigma} \times \left(\frac{\sqrt{n-1} \times \Gamma\left(\frac{2n-3}{2}\right)}{\Gamma(n-1)} - 1\right)$$

$$Bias_{Shieh's\ d} = \frac{\mu_1 - \mu_2}{\boldsymbol{2}\sigma} \times \left(\frac{\sqrt{n-1} \times \Gamma\left(\frac{2n-3}{2}\right)}{\Gamma(n-1)} - 1\right)$$

We can therefore conclude that the bias of Cohen's d is twice larger than the bias of Shieh's d.

Appendix 2 : The variance of Cohen's *d* is four times larger than the bias of Shieh's *d* when population variances and sample sizes are equal across groups: mathematical demonstration.

The variance of Cohen's d is defined in Table 2 as

$$Var_{Cohen's\ d} = \frac{N \times df_{Student}}{n_1 n_2 \times (df_{Student} - 2)} + \delta^2_{Cohen} \left[\frac{df_{Student}}{df_{Student} - 2} - \left(\frac{\sqrt{\frac{df_{Student}}{2}} \times \Gamma\left(\frac{df_{Student}-1}{2}\right)}{\Gamma\left(\frac{df_{Student}}{2}\right)} \right)^2 \right] \tag{26}$$

and the variance of Shieh's d is defined in Table 3 as

$$Var_{Shieh's\ d} = \frac{df_{Welch}}{(df_{Welch} - 2)N} + \delta^2_{Shieh} \left[\frac{df_{Welch}}{df_{Welch} - 2} - \left(\frac{\sqrt{\frac{df_{Welch}}{2}} \times \Gamma\left(\frac{df_{Welch}-1}{2}\right)}{\Gamma\left(\frac{df_{Welch}}{2}\right)} \right)^2 \right] \tag{27}$$

We have previously shown in equations 24 and 25 that degrees of freedom associated with Student's t-test and Welch's t-test equal $2(n-1)$, when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$. As a consequence, the first term of the addition in equation 26 is 4 times larger than the first term of the addition in equation 27:

$$\frac{N \times df_{Student}}{n_1 n_2 \times (df_{Student} - 2)} = \frac{2n \times 2(n-1)}{n^2 \times (2n-4)} = \boldsymbol{\frac{4(n-1)}{n(2n-4)}}$$

$$\frac{df_{Welch}}{(df_{Welch} - 2)N} = \frac{2(n-1)}{2n(2n-4)} = \boldsymbol{\frac{n-1}{n(2n-4)}}$$

We have also previously shown in equations 22 and 23 that δ_{Cohen} is twice larger than δ_{Shieh} when $n_1 = n_2 = n$ and $\sigma_1 = \sigma_2 = \sigma$ and, therefore, δ^2_{Cohen} is four times larger than δ^2_{Shieh}. As a consequence, the second term of the addition in equation 26 is also 4 times larger than the second term of the addition in equation 27. Because both terms of the addition in equation 26 are four times larger than those in equation 27, we can conclude that the variance of Cohen's d is four times larger than the variance of Shieh's d.