

¹ Correlations between the sample means difference and standardizers of all estimators, and
² implications on biases and variances of all estimators

³ Delacre Marie¹

⁴ ¹ ULB

5 Correlations between the sample means difference and standardizers of all estimators, and
 6 implications on biases and variances of all estimators

7 **Introduction**

8 The d -family effect sizes are commonly used with “between-subject” designs where
 9 individuals are randomly assigned into one of two independent groups and groups scores
 10 means are compared. The population effect size is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \quad (1)$$

11 where both populations follow a normal distribution with mean μ_j in the j^{th}
 12 population ($j=1,2$) and common standard deviation σ . They exist different estimators of this
 13 population effect size, varying as a function of the chosen standardizer (σ). When the
 14 equality of variances assumption is met, σ is estimated by pooling both samples standard
 15 deviations (S_1 and S_2):

$$\sigma_{Cohen's\ d_s} = \sqrt{\frac{(n_1 - 1) \times S_1^2 + (n_2 - 1) \times S_2^2}{n_1 + n_2 - 2}} \quad (2)$$

16 When the equality of variances assumption is not met, we are considering three
 17 alternative estimates:

- 18 • Using the standard deviation of the control group (S_c) as standardizer:

$$S_{Glass's\ d_s} = S_c \quad (3)$$

- 19 • Using a standardizer that takes the sample sizes allocation ratio $(\frac{n_1}{n_2})$ into account:

$$S_{Shieh's\ d_s} = \sqrt{S_1^2/q_1 + S_2^2/q_2}; \quad q_j = \frac{n_j}{N} (j = 1, 2) \quad (4)$$

- 20 • Or using the square root of the non pooled average of both variance estimates (S_1^2 and
 21 S_2^2) as standardizer:

$$S_{Cohen's\ d_s} = \sqrt{\frac{(S_1^2 + S_2^2)}{2}} \quad (5)$$

22 As we previously mentioned, using these formulas implies meeting the assumption of
 23 normality. Using them when distributions are not normal will have consequences on both
 24 bias and variance of all estimators. More specifically, when samples are extracted from
 25 skewed distribution, correlations might occur between the sample means difference ($\bar{X}_1 - \bar{X}_2$)
 26 and standardizers (σ). Studying when these correlations occur is the main goal of this
 27 appendix. To this end, we will distinguish 4 situations, as a function of the sample sizes ratio
 28 ($\frac{n_1}{n_2} = 1$ vs. $\frac{n_1}{n_2} \neq 1$) and the population SD-ratio ($\frac{\sigma_1}{\sigma_2} = 1$ vs. $\frac{\sigma_1}{\sigma_2} \neq 1$), but before that, we
 29 will briefly introduce the impact of correlations on the bias.

30 **How correlations between the mean difference ($\bar{X}_1 - \bar{X}_2$) and standardizers
 31 influence the bias of estimators.**

32 **Correlations between the mean difference ($\bar{X}_1 - \bar{X}_2$) and all standardizers**
 33 **When equal population variances are estimated based on equal sample sizes
 34 (condition a)**

35 While \bar{X}_j and S_j ($j=1,2$) are uncorrelated when samples are extracted from symmetric
 36 distributions (see Figure 1), there is a non-null correlation between \bar{X}_j and S_j when
 37 distributions are skewed (Zhang, 2007).

38 More specifically, when distributions are right-skewed, there is a **positive** correlation
 39 between \bar{X}_j and S_j (see the two top plots in Figure 2), resulting in a *positive* correlation

40 between S_1 and $\bar{X}_1 - \bar{X}_2$ and in a *negative* correlation between S_2 and $\bar{X}_1 - \bar{X}_2$ (see the two
 41 bottom plots in Figure 2). This can be explained by the fact that \bar{X}_1 and $\bar{X}_1 - \bar{X}_2$ are
 42 positively correlated while \bar{X}_2 and $\bar{X}_1 - \bar{X}_2$ are negatively correlated (of course, correlations
 43 would be trivially reversed if we computed $\bar{X}_2 - \bar{X}_1$ instead of $\bar{X}_1 - \bar{X}_2$).

44 One should also notice that both correlations between S_j and $\bar{X}_1 - \bar{X}_2$ are equal, in
 45 absolute terms (possible tiny differences might be observed due to sampling error in our
 46 simulations). As a consequence, when computing a standardizer taking both S_1 and S_2 into
 47 account, it results in a standardizer that is uncorrelated with $\bar{X}_1 - \bar{X}_2$ (see Figure 3).

48 On the other hand, when distributions are left-skewed, there is a **negative** correlation
 49 between \bar{X}_j and S_j (see the two top plots in Figure 4), resulting in a *negative* correlation
 50 between S_1 and $\bar{X}_1 - \bar{X}_2$ and in a *positive* correlation between S_2 and $\bar{X}_1 - \bar{X}_2$ (see the two
 51 bottom plots in Figure 4).

52 Again, because correlations between S_j and $\bar{X}_1 - \bar{X}_2$ are similar in absolute terms, any
 53 standardizers taking both S_1 and S_2 into account will be uncorrelated with $\bar{X}_1 - \bar{X}_2$ (see
 54 Figure 5).

55 **When equal population variances are estimated based on unequal sample sizes
 56 (condition b)**

57 Even when $n_1 \neq n_2$, \bar{X}_j and S_j ($j=1,2$) remain uncorrelated as long as samples are
 58 extracted from symmetric distributions (see Figure 6). When distributions are skewed, there
 59 are again non-null correlations between \bar{X}_j and S_j , however $\text{cor}(S_1, \bar{X}_1) \neq \text{cor}(S_2, \bar{X}_2)$,
 60 because of the different sample sizes.

61 When distributions are skewed, one observes that the larger the sample size, the lower
 62 the correlation between S_j and \bar{X}_j (See Figures 7 and 7).

63 This might explain why the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ is

64 lower in the larger sample (See bottom plots in Figure ??). With no surprise, there is a
 65 positive (negative) correlation between S_1 and $\bar{X}_1 - \bar{X}_2$ and a negative (positive) correlation
 66 between S_2 and $\bar{X}_1 - \bar{X}_2$ when distribution are right-skewed (left-skewed), as illustrated in
 67 the two bottom plots of Figures 9 and 10.

68 This might also explain why the standardizers of Shieh's d_s and Cohen's d'_s are this
 69 time **correlated** with $\bar{X}_1 - \bar{X}_2$ (see Figure 5). EXPLIQUER POURQUOI SHIEH EST LE
 70 PLUS AFFECTE DES TROIS ET POURQUOI COHEN L EST PAS.

71 Reste aussi à comprendre: pourquoi la corrélation prend le signe du plus grand groupe
 72 et pas le contraire? A mon avis, simplement parce qu'en mixant le tout, on donne bcp plus
 73 de poids au 2ème groupe (mais bien écrire ça).

74 **When unequal population variances are estimated based on equal sample sizes
 75 (condition c)**

76 **When unequal population variances are estimated based on unequal sample
 77 sizes (conditions d and e)**

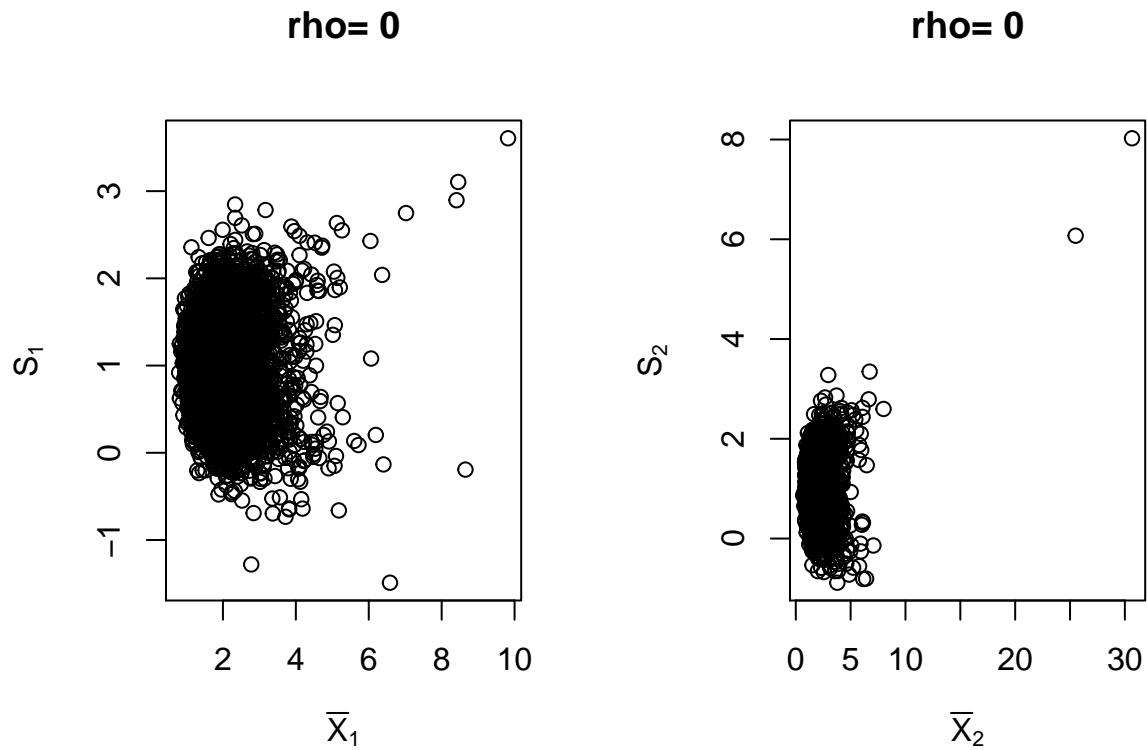


Figure 1. S_j as a function of \bar{X}_j ($j=1,2$), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

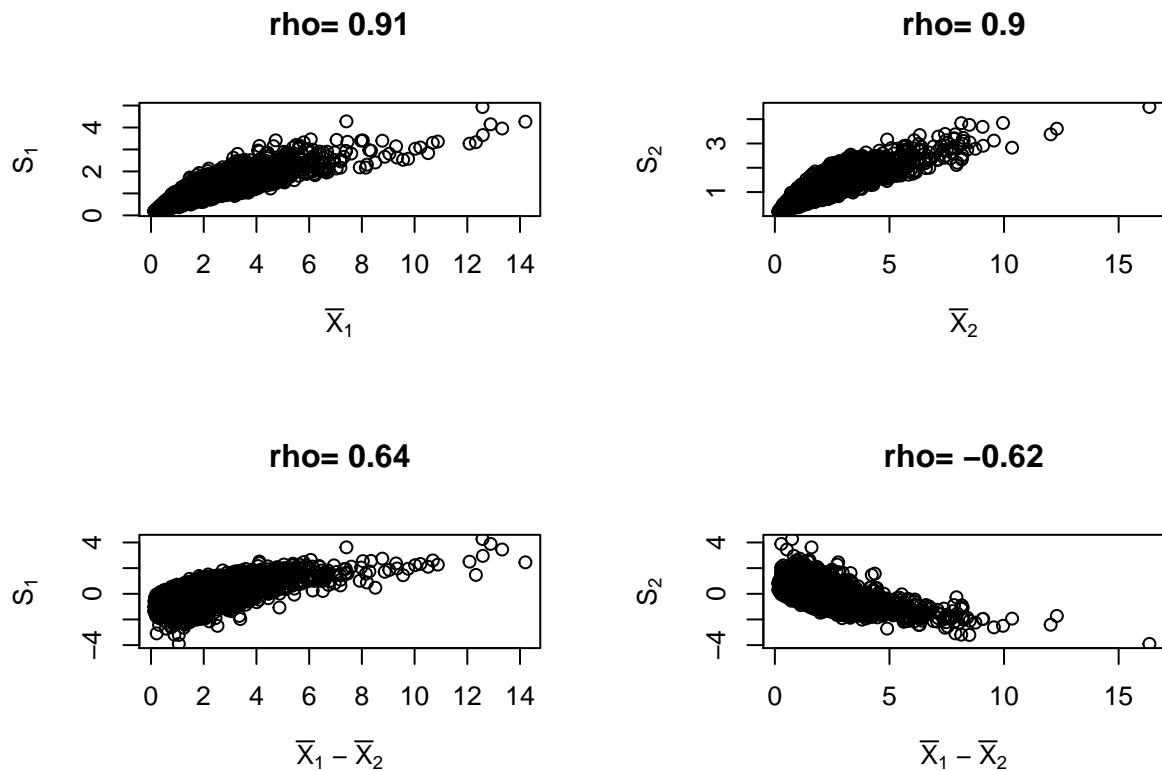


Figure 2. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$; top plots)

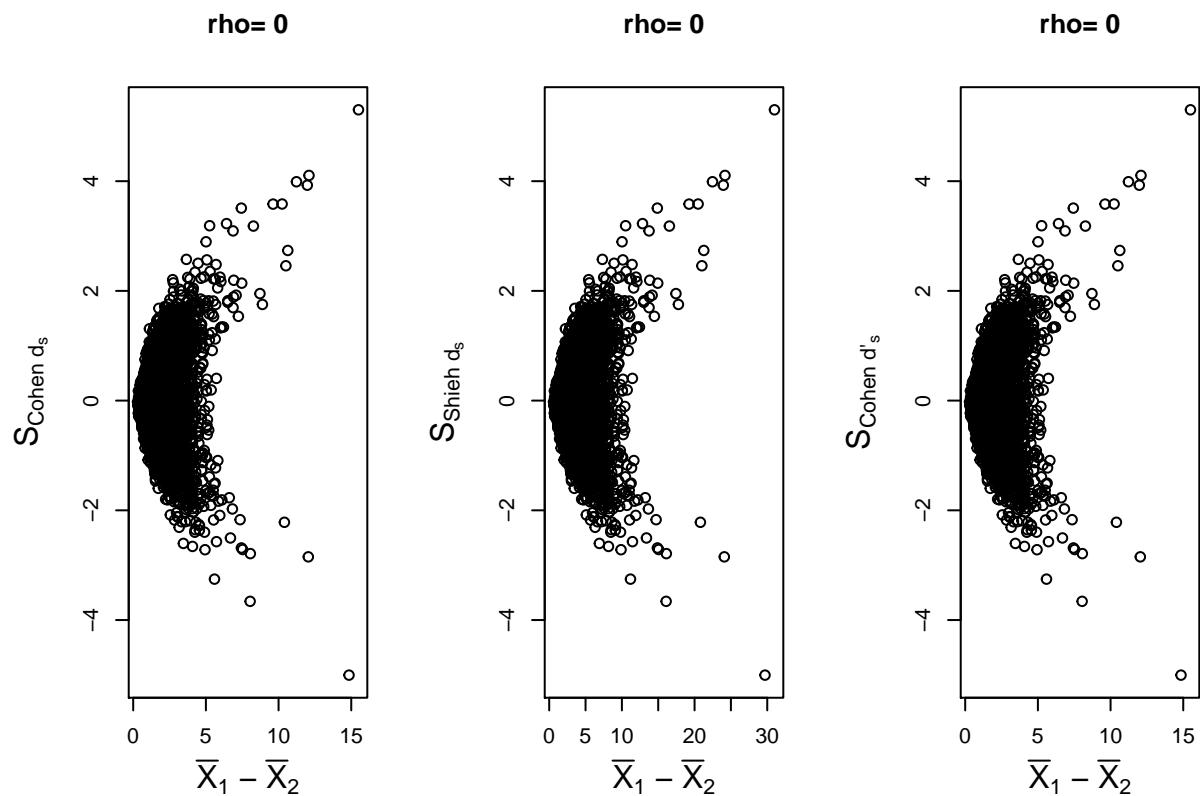


Figure 3. $S_{Glass's} d_s$, $S_{Shieh's} d_s$ and $S_{Cohen's} d_s$ as a function of the means difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

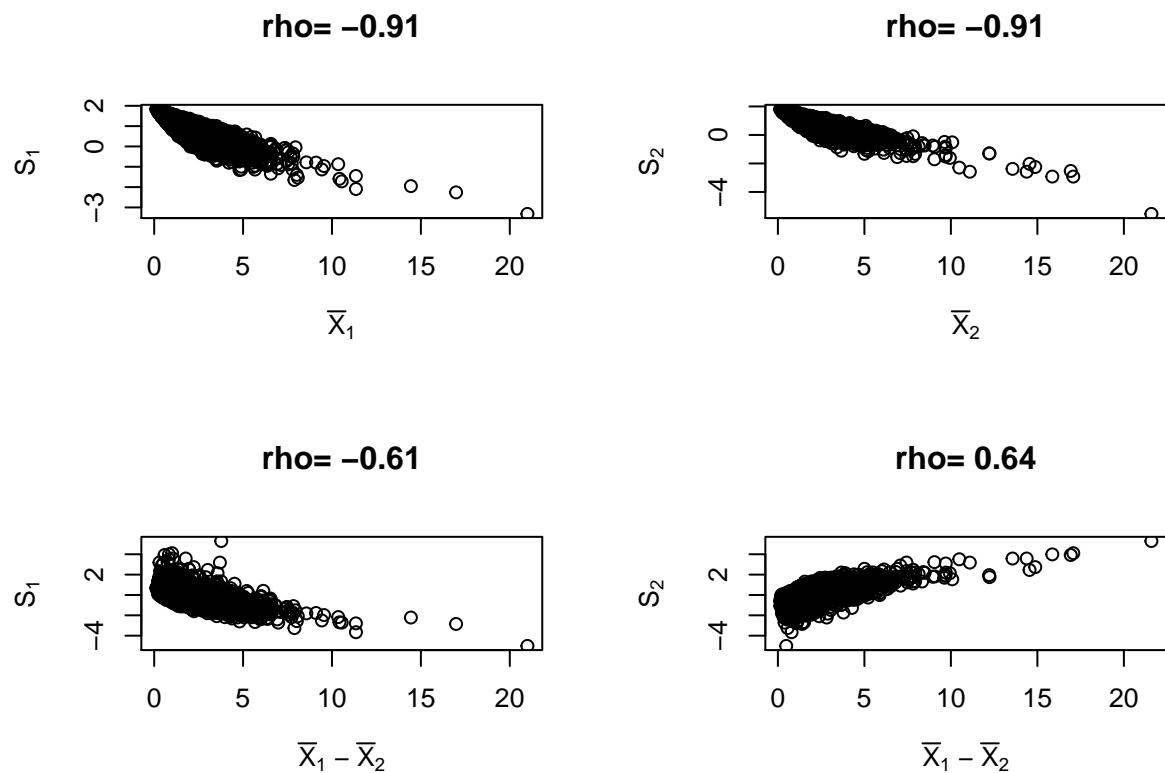


Figure 4. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$; top plots)

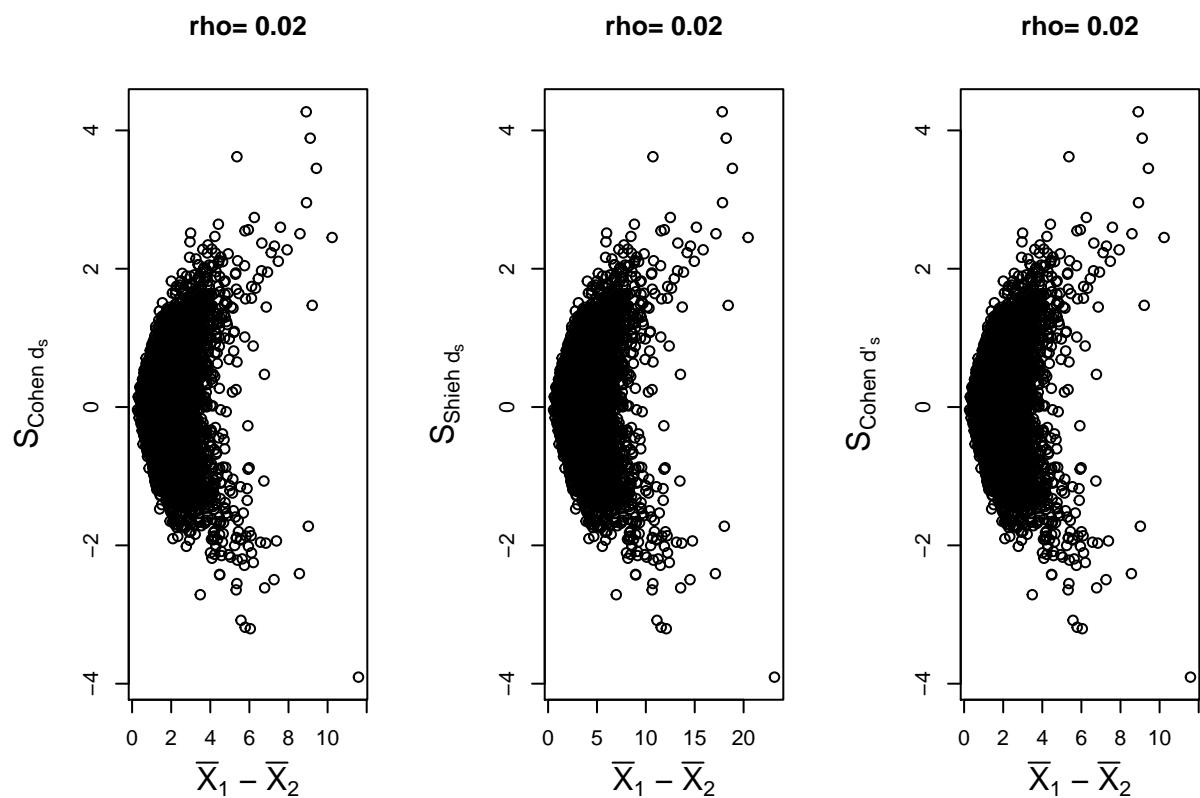


Figure 5. $S_{Glass's} d_s$, $S_{Shieh's} d_s$ and $S_{Cohen's} d'_s$ as a function of the means difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

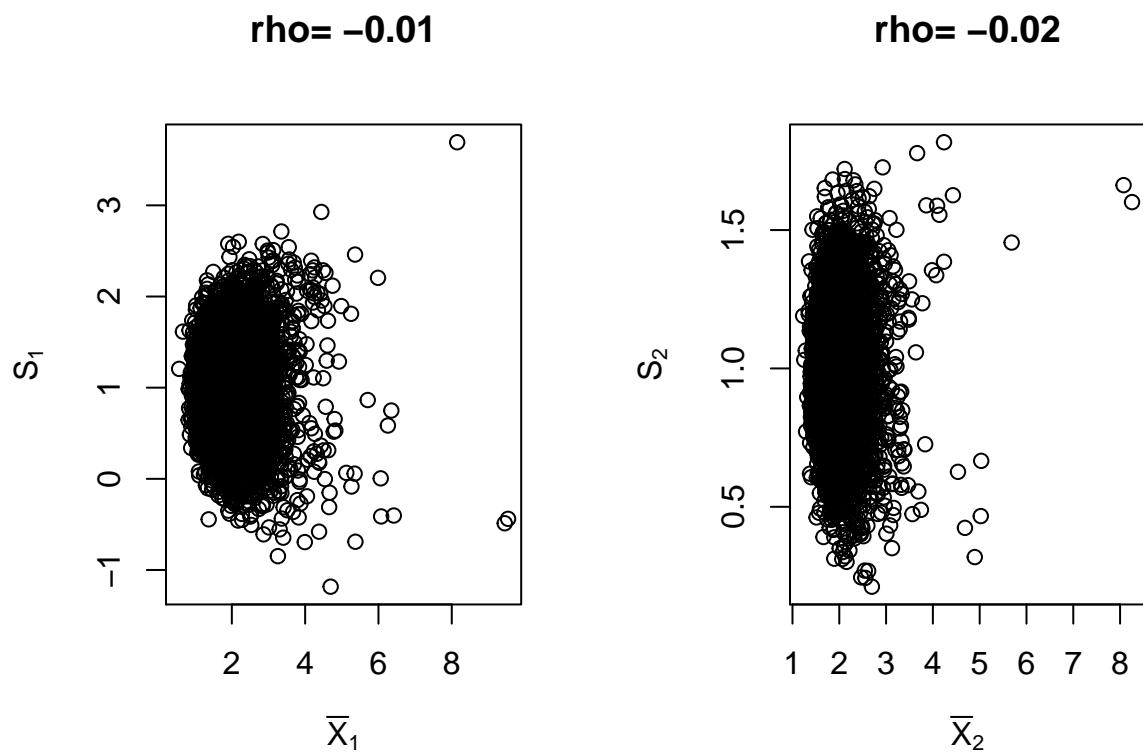


Figure 6. S_j as a function of \bar{X}_j ($j=1,2$), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

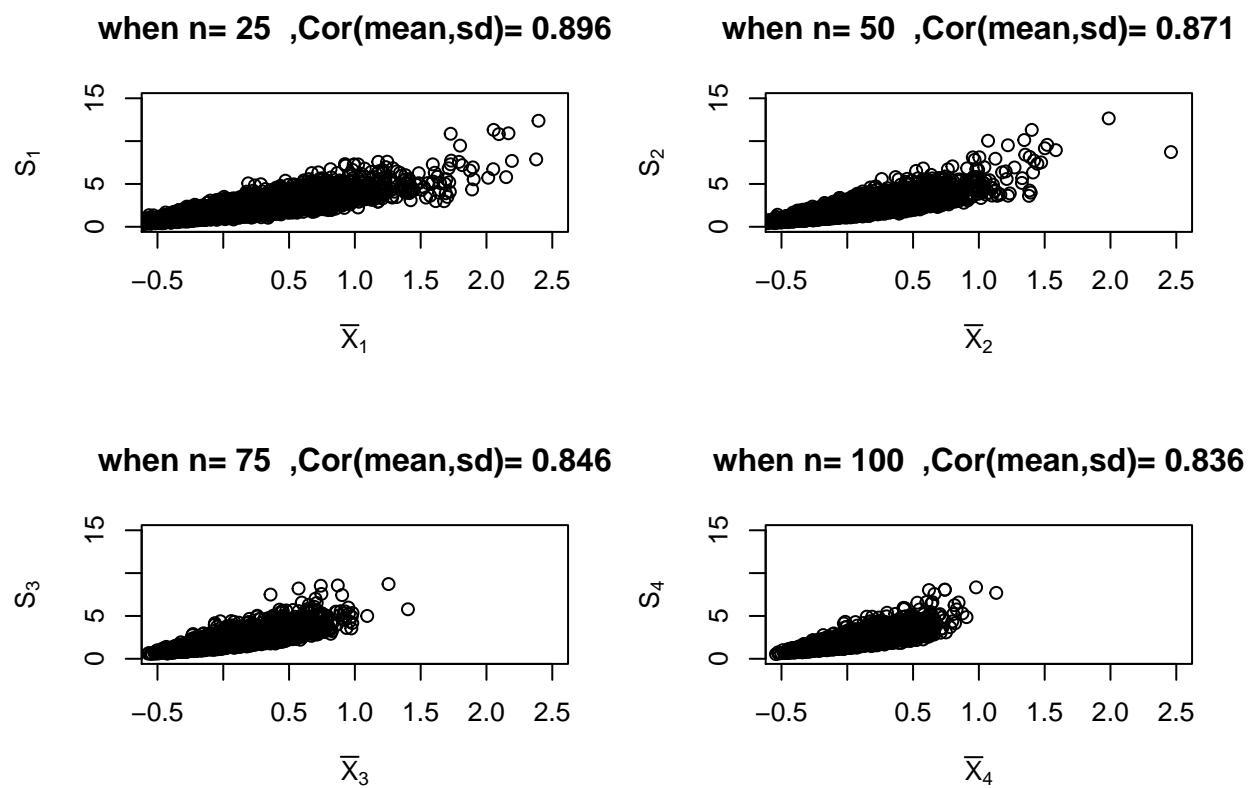


Figure 7. correlation between S_j and \bar{X}_1 when $n = 25, 50, 75$ or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

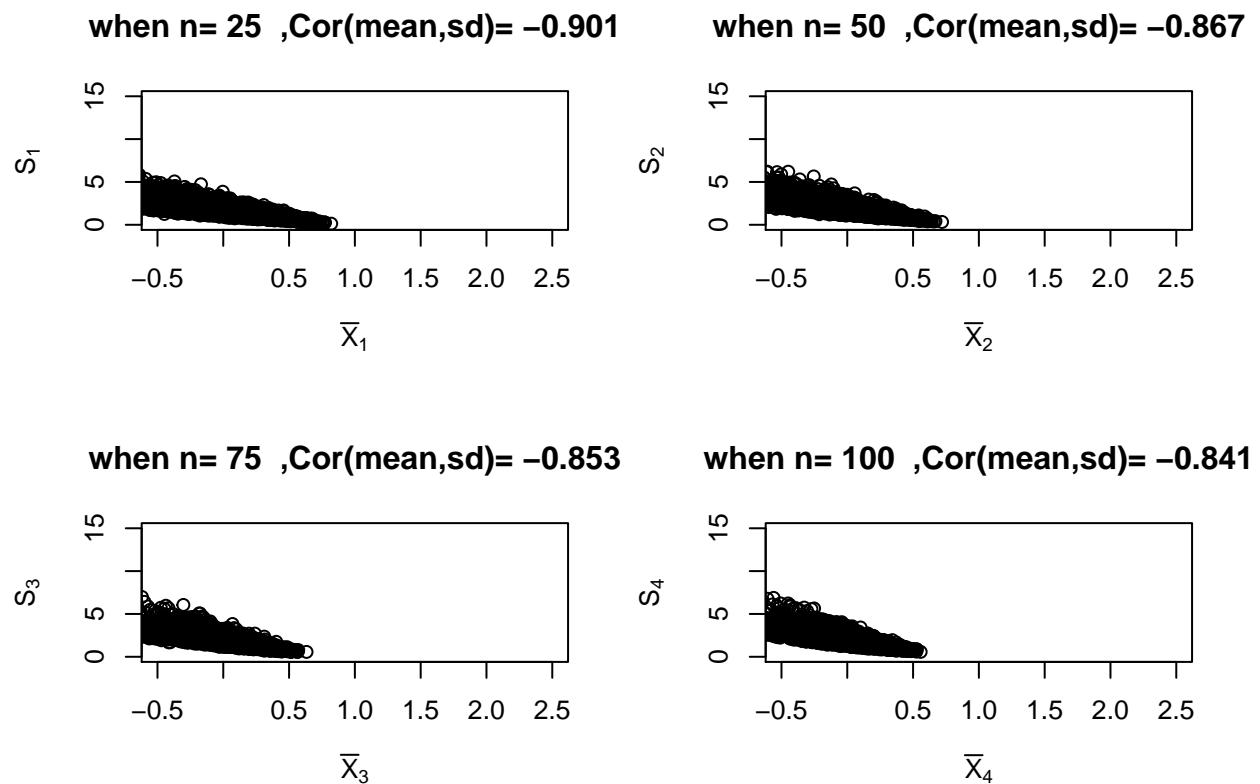


Figure 8. correlation between S_j and \bar{X}_1 when $n = 25, 50, 75$ or 100 and samples are extracted from right left distributions ($\gamma_1 = -6.32$)

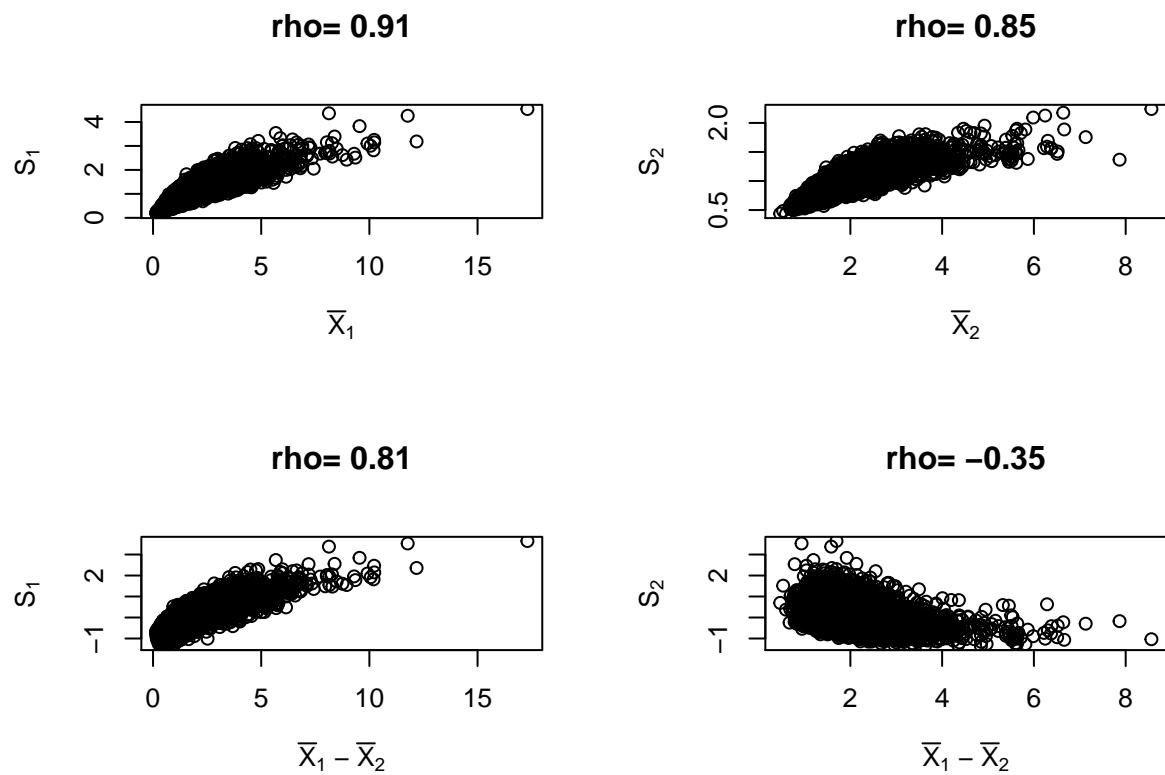


Figure 9. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$; top plots), with $n1=20$ and $n2=100$

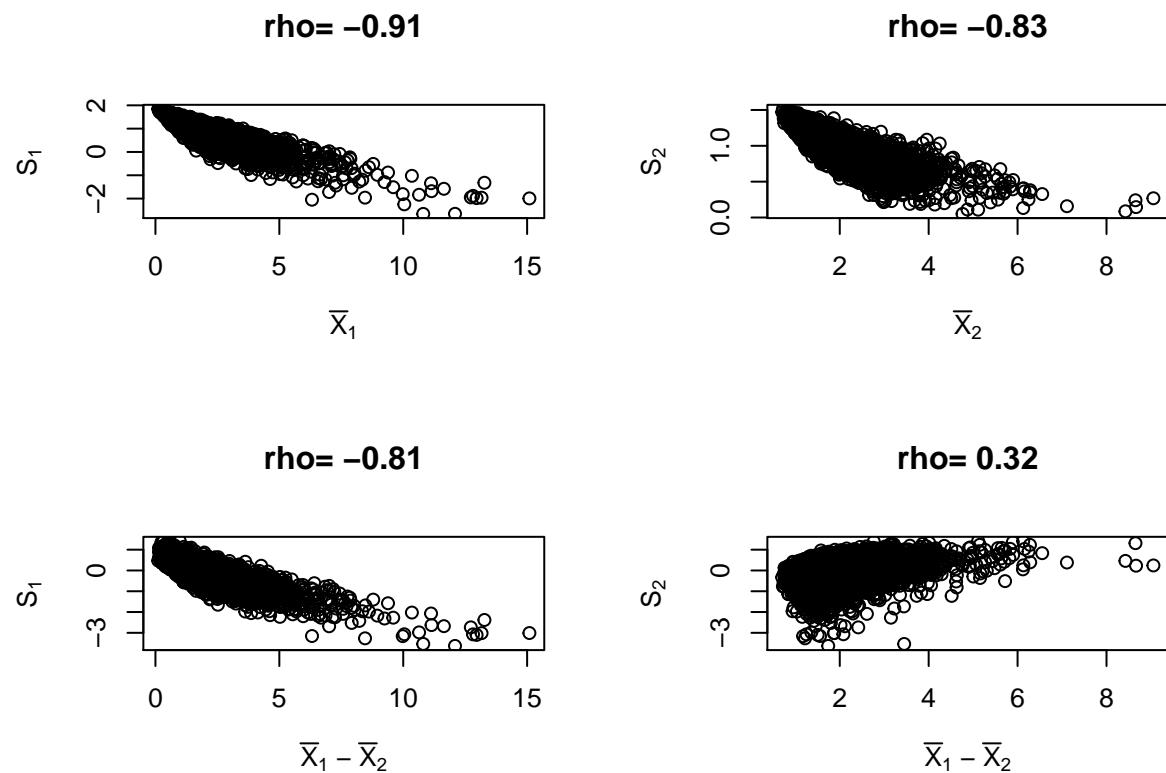


Figure 10. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$; top plots), with $n1=20$ and $n2=100$

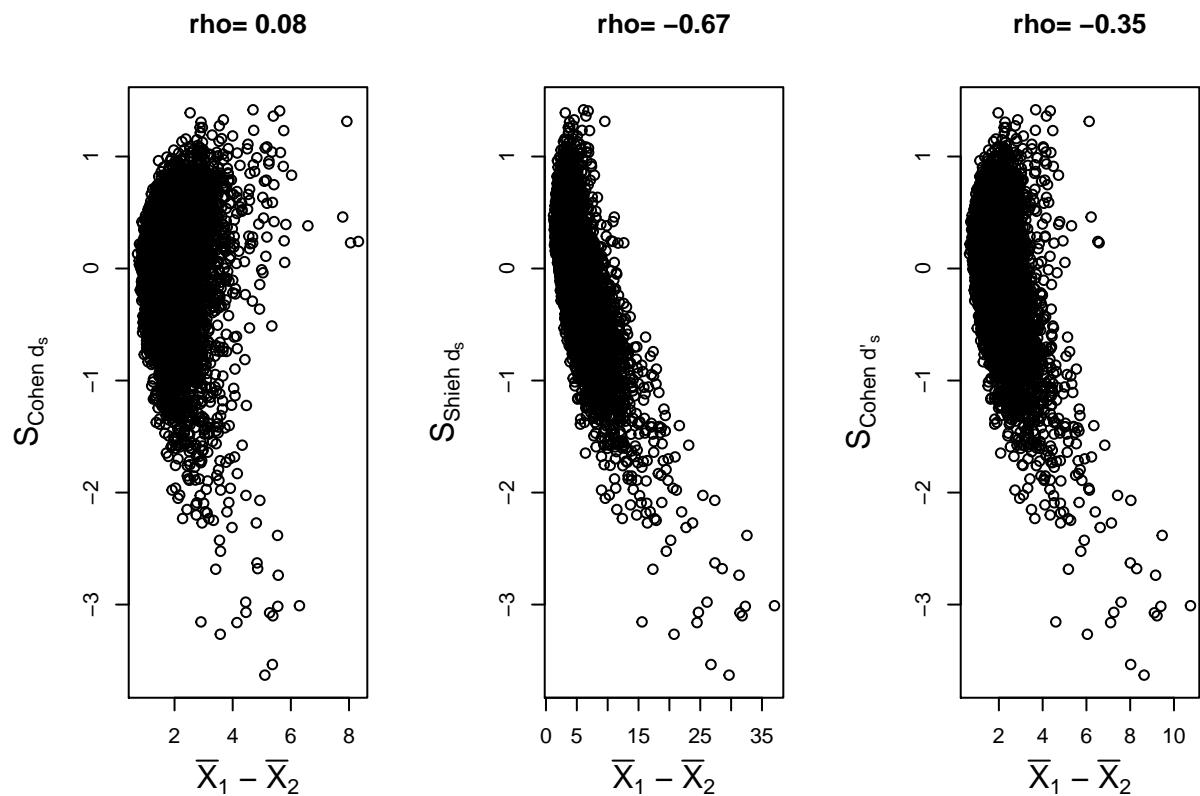


Figure 11. $S_{Glass's\,d_s}$, $S_{Shieh's\,d_s}$ and $S_{Cohen's\,d'_s}$ as a function of the means difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

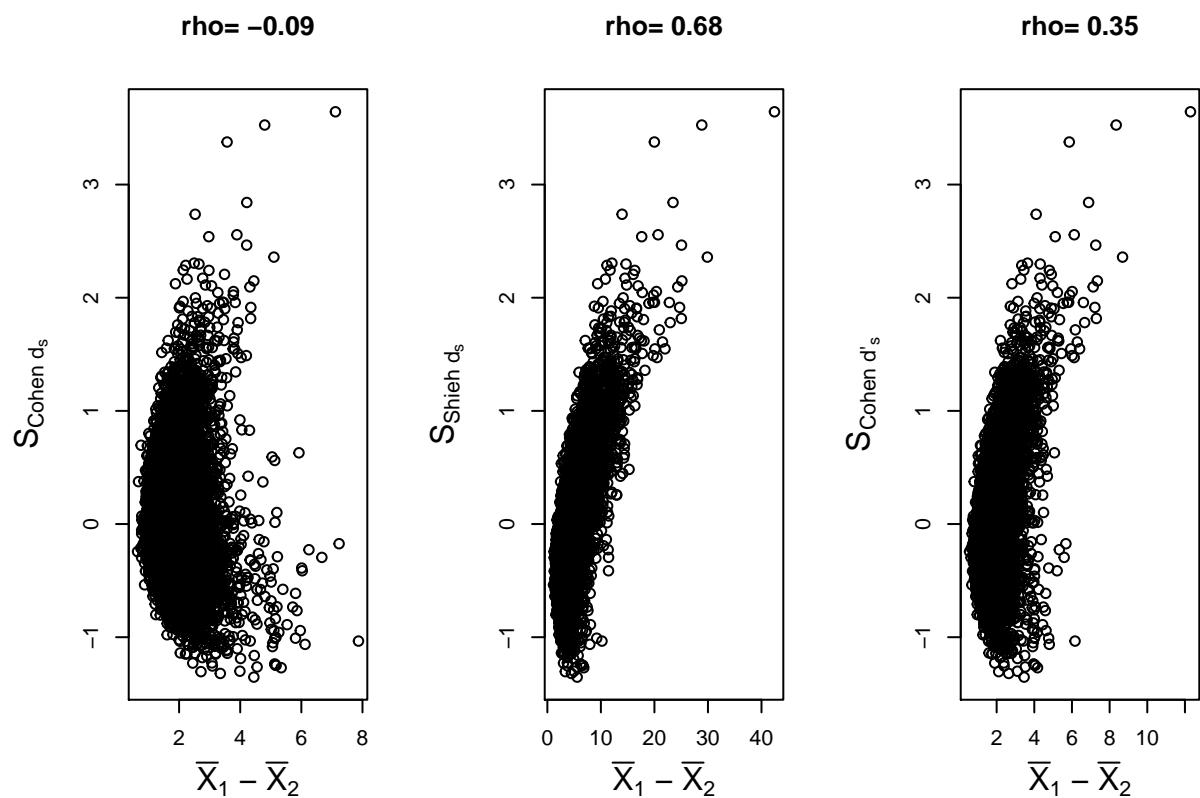


Figure 12. $S_{\text{Glass}'s \, d_s}$, $S_{\text{Shieh}'s \, d_s}$ and $S_{\text{Cohen}'s \, d'_s}$ as a function of the means difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)