

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

3 Delacre Marie¹

4 ¹ ULB

5 Author Note

6 Correspondence concerning this article should be addressed to Delacre Marie, . E-mail:

7 Correlations between the sample mean difference and standardizers of all estimators, and
 8 implications on biases and variances of all estimators

9 **Introduction**

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

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

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

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

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

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

$$S_{Glass's\ d_s} = S_c$$

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

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

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

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

16 As we previously mentioned, the use of these formulas requires to meet the assumption
 17 of normality. Using them when distributions are not normal will have consequences on both
 18 bias and variance of all estimators. More specifically, when samples are extracted from
 19 skewed distributions, correlations might occur between the sample mean difference ($\bar{X}_1 - \bar{X}_2$)
 20 and standardizers (S). Throughout this Supplemental Material, we will study when these
 21 correlations occur. To this end, we will distinguish 3 situations:

- 22 - when $\sigma_1 = \sigma_2$ and $n_1 = n_2$ (condition a);
 23 - when $\sigma_1 = \sigma_2$ and $n_1 \neq n_2$ (condition b);
 24 - when $\sigma_1 \neq \sigma_2$ and $n_1 = n_2$ (condition c).

25 Before studying conditions a, b and c, we will briefly introduce the impact of these
 26 correlations on the bias. Note that we will compute correlations using the coefficient of
 27 Spearman's ρ . We decided to use Spearman's ρ instead of Pearson's ρ because some plots
 28 revealed non-perfectly linear relations.

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

31 When distributions are right-skewed, there is a positive (negative) correlation between
 32 S_1 (S_2) and $(\bar{X}_1 - \bar{X}_2)$. When distributions are left-skewed, there is a negative (positive)
 33 correlation between S_1 (S_2) and $(\bar{X}_1 - \bar{X}_2)$. When the population mean difference ($\mu_1 - \mu_2$)
 34 is positive (like in our simulations), all other parameters being equal, an estimator is always
 35 less biased and variable when choosing a standardizer that is positively correlated with
 36 $\bar{X}_1 - \bar{X}_2$ than when choosing an estimator that is negatively correlated with $\bar{X}_1 - \bar{X}_2$. When
 37 the population mean difference is negative, the reverse is true.

38 "All other parameters being equal" is mentioned because it is always possible that
 39 other factors in action have an opposite effect on bias and variance in order that increasing
 40 the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ does not necessarily reduce the

41 bias and the variance. For example, when population variances are equal across groups and
 42 sample sizes are unequal, we will see below that the lower n_j , the larger the magnitude of the
 43 correlation between S_j and $\bar{X}_1 - \bar{X}_2$. When the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ is
 44 positive, the smaller the sample size, the larger the positive correlation. At the same time,
 45 we know that increasing the sample size decreases the bias. This is a nice example of
 46 situations where two factors might have an opposite action on bias.

47 **Correlations between the mean difference ($\bar{X}_1 - \bar{X}_2$) and all standardizers**

48 **When equal population variances are estimated based on equal sample sizes**
 49 **(condition a)**

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

53 More specifically, when distributions are right-skewed, there is a **positive** correlation
 54 between \bar{X}_j and S_j (see the two top plots in Figure 2), resulting in a *positive* correlation
 55 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
 56 bottom plots in Figure 2). This can be explained by the fact that \bar{X}_1 and $\bar{X}_1 - \bar{X}_2$ are
 57 positively correlated while \bar{X}_2 and $\bar{X}_1 - \bar{X}_2$ are negatively correlated (of course, correlations
 58 would be trivially reversed if we computed $\bar{X}_2 - \bar{X}_1$ instead of $\bar{X}_1 - \bar{X}_2$).

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

63 On the other hand, when distributions are left-skewed, there is a **negative** correlation
 64 between \bar{X}_j and S_j (see the two top plots in Figure 4), resulting in a *negative* correlation

65 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
 66 bottom plots in Figure 4).

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

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

72 When distributions are skewed, there are again non-null correlations between \bar{X}_j and
 73 S_j , however $\text{cor}(S_1, \bar{X}_1) \neq \text{cor}(S_2, \bar{X}_2)$, because of the different sample sizes.

74 When distributions are skewed, one observes that the larger the sample size, the lower
 75 the correlation between S_j and \bar{X}_j (See Figures 6 and 7).

76 This might explain that the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ is
 77 lower in the larger sample (see bottom plots in Figures 8 and 9). With no surprise, there is a
 78 positive (negative) correlation between S_1 and $\bar{X}_1 - \bar{X}_2$ and a negative (positive) correlation
 79 between S_2 and $\bar{X}_1 - \bar{X}_2$ when distribution are right-skewed (left-skewed), as illustrated in
 80 the two bottom plots of Figures 8 and 9.

81 This might also explain that standardizers of Shieh's d_s and Cohen's d_s^* are
 82 **correlated** with $\bar{X}_1 - \bar{X}_2$ (see Figures 10 and 11):

83 - When computing $S_{\text{Cohen's } d_s^*}$, the same weight is given to both S_1 and S_2 . Therefore,
 84 it does not seem surprising that the sign of the correlation between $S_{\text{Cohen's } d_s^*}$ and $\bar{X}_1 - \bar{X}_2$ is
 85 the same as the size of the correlation between $\bar{X}_1 - \bar{X}_2$ and the SD of the smallest sample;
 86 - When computing $S_{\text{Shieh's } d_s}$, more weight is given to the SD of the smallest sample, it
 87 is therefore not really surprising to observe that the correlation between $S_{\text{Shieh's } d_s}$ and
 88 $\bar{X}_1 - \bar{X}_2$ is closer of the correlation between the SD of the smallest group and $\bar{X}_1 - \bar{X}_2$

89 (i.e. $|cor(S_{Shieh's\ d_s}, \bar{X}_1 - \bar{X}_2)| > |cor(S_{Cohen's\ d_s^*}, \bar{X}_1 - \bar{X}_2)|$);

90 - When computing S_{Cohen} , more weight is given to the SD of the largest sample, which
91 by compensation effect, brings the correlation very close to 0.

92 The correlation between $\bar{X}_1 - \bar{X}_2$ and respectively S_1, S_2 , the standardizer of Cohen's
93 d_s^* , the standardizer of Shieh's d_s and the standardizer of Cohen's d_s are summarized in
94 Table 1:

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

97 When distributions are skewed, there are again non-null correlations between \bar{X}_j and
98 S_j . As illustrated in Figures 12 and 13, the correlation remains the same for any population
99 $SD (\sigma)$. However, the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ differs: it is
100 stronger in the sample extracted from the larger population variance (see Figures 14 and 15).

101 This also explain that when computing a standardizer that takes both S_1 and S_2 into
102 account, it results in a standardizer that is correlated with $\bar{X}_1 - \bar{X}_2$ (see Figures 16 and 17).
103 The correlation between the mean difference ($\bar{X}_1 - \bar{X}_2$) and respectively the standardizer of
104 Shieh's d_s , Cohen's d_s^* and Cohen's d_s will have the same sign as the correlation between
105 ($\bar{X}_1 - \bar{X}_2$) and the larger SD . Table 2 summarizes the sign of the correlation between
106 $\bar{X}_1 - \bar{X}_2$ and respectively S_1, S_2 and the three standardizers taking both S_1 and S_2 into
107 account (see "Others" in the Table).

Table 1

Correlation between standardizers (S_1 , S_2 , $S_{Cohen's\ d_s}$ and others) and $\bar{X}_1 - \bar{X}_2$, when samples are extracted from skewed distributions with equal variances, and $n_1 = n_2$ (condition a) or $n_1 \neq n_2$ (condition b)

		population		
		distribution		
		<i>right-skewed</i>	<i>left-skewed</i>	
When $n_1 = n_2$				
	S_1 : positive			S_1 : negative
	S_2 : negative			S_2 : positive
	$S_{Cohen's\ d_s}$: null			$S_{Cohen's\ d_s}$: null
	$S_{Shieh's\ d_s}$: null			$S_{Shieh's\ d_s}$: null
	$S_{Cohen's\ d_s^*}$: null			$S_{Cohen's\ d_s^*}$: null
When $n_1 > n_2$				
	S_1 : positive			S_1 : negative
	S_2 : negative			S_2 : positive
	$S_{Cohen's\ d_s}$: null			$S_{Cohen's\ d_s}$: null
	$S_{Shieh's\ d_s}$: negative			$S_{Shieh's\ d_s}$: positive
	$S_{Cohen's\ d_s^*}$: positive (but very small)			$S_{Cohen's\ d_s^*}$: negative (but very small)
When $n_1 < n_2$				
	S_1 : positive			S_1 : negative
	S_2 : negative			S_2 : positive
	$S_{Cohen's\ d_s}$: negative (but very small)			$S_{Cohen's\ d_s}$: positive (but very small)
	$S_{Shieh's\ d_s}$: positive			$S_{Shieh's\ d_s}$: negative

population	distribution
$S_{Cohen's\ d_s^*}$: positive	$S_{Cohen's\ d_s^*}$: negative

Table 2

Correlation between standardizers (S_1, S_2 and others) and $\bar{X}_1 - \bar{X}_2$, when samples are extracted from skewed distributions with equal sample sizes, as a function of the SD-ratio.

population distribution		
	<i>right-skewed</i>	<i>left-skewed</i>
When $\sigma_1 = \sigma_2$	S_1 : <i>positive</i> S_2 : <i>negative</i> Others: <i>null</i>	S_1 : <i>negative</i> S_2 : <i>positive</i> Others: <i>null</i>
When $\sigma_1 > \sigma_2$	S_1 : <i>positive</i> S_2 : <i>negative</i> Others: <i>positive</i>	S_1 : <i>negative</i> S_2 : <i>positive</i> Others: <i>negative</i>
When $\sigma_1 < \sigma_2$	S_1 : <i>positive</i> S_2 : <i>negative</i> Others: <i>negative</i>	S_1 : <i>negative</i> S_2 : <i>positive</i> Others: <i>positive</i>

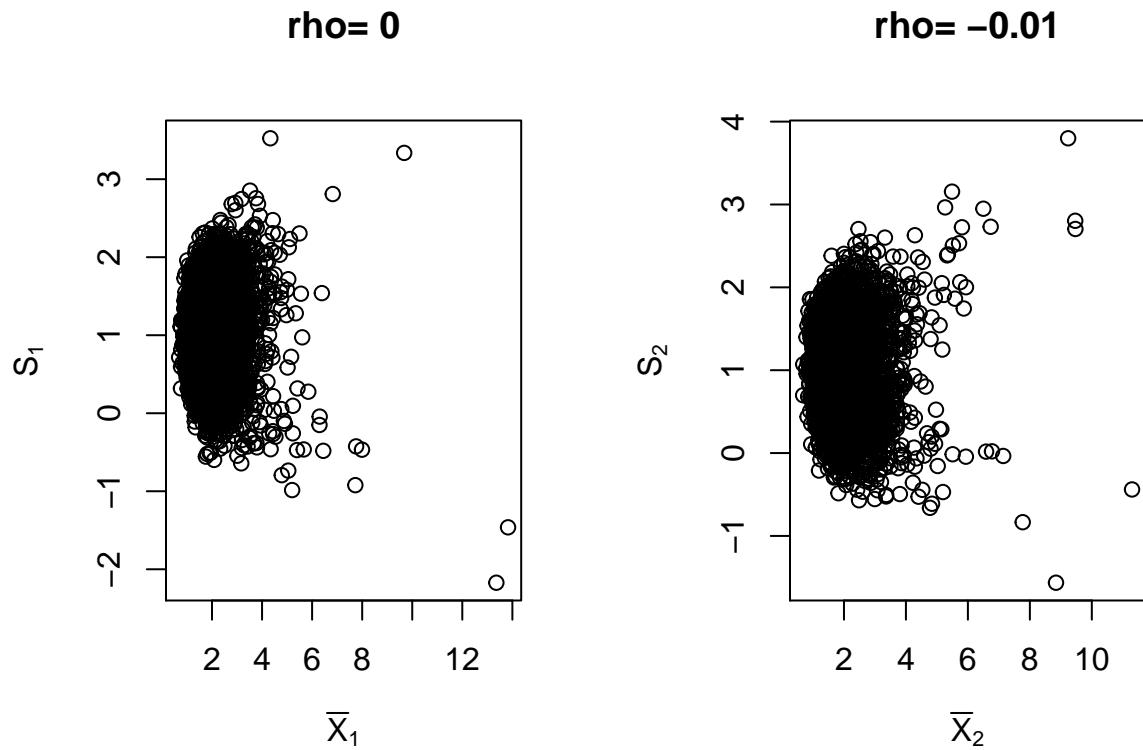


Figure 1. S_j as a function of \bar{X}_j ($j=1,2$), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

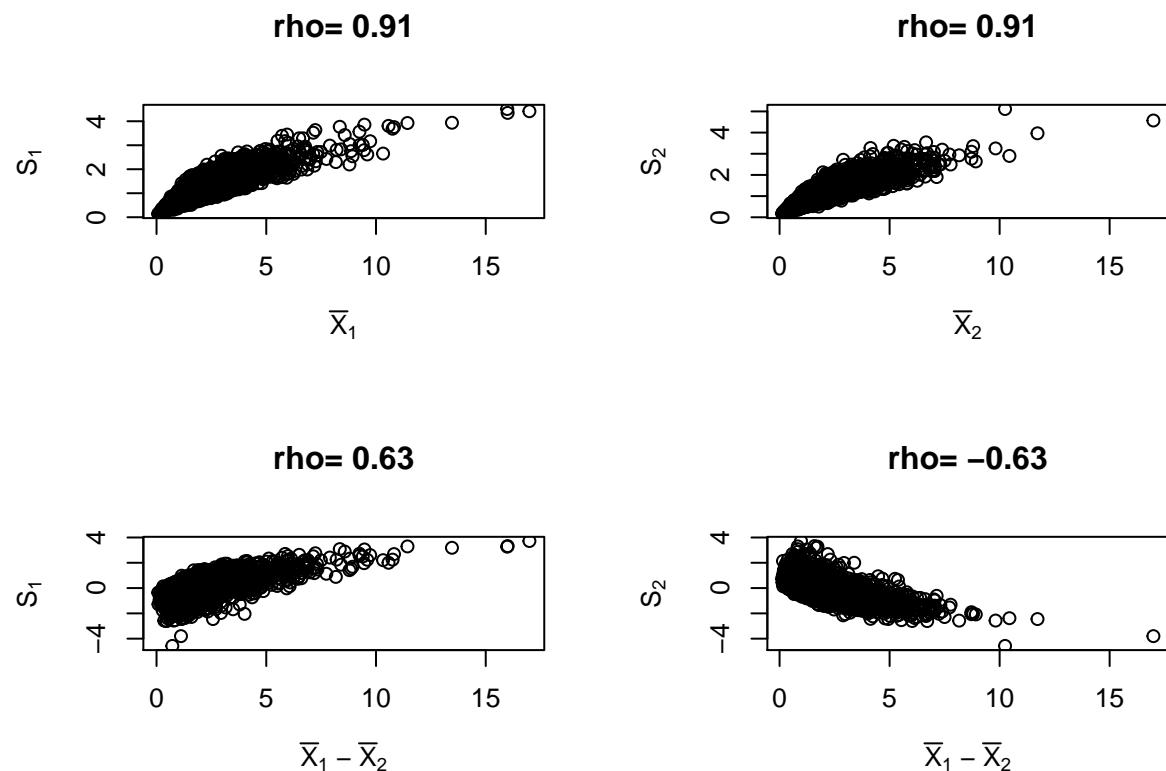


Figure 2. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

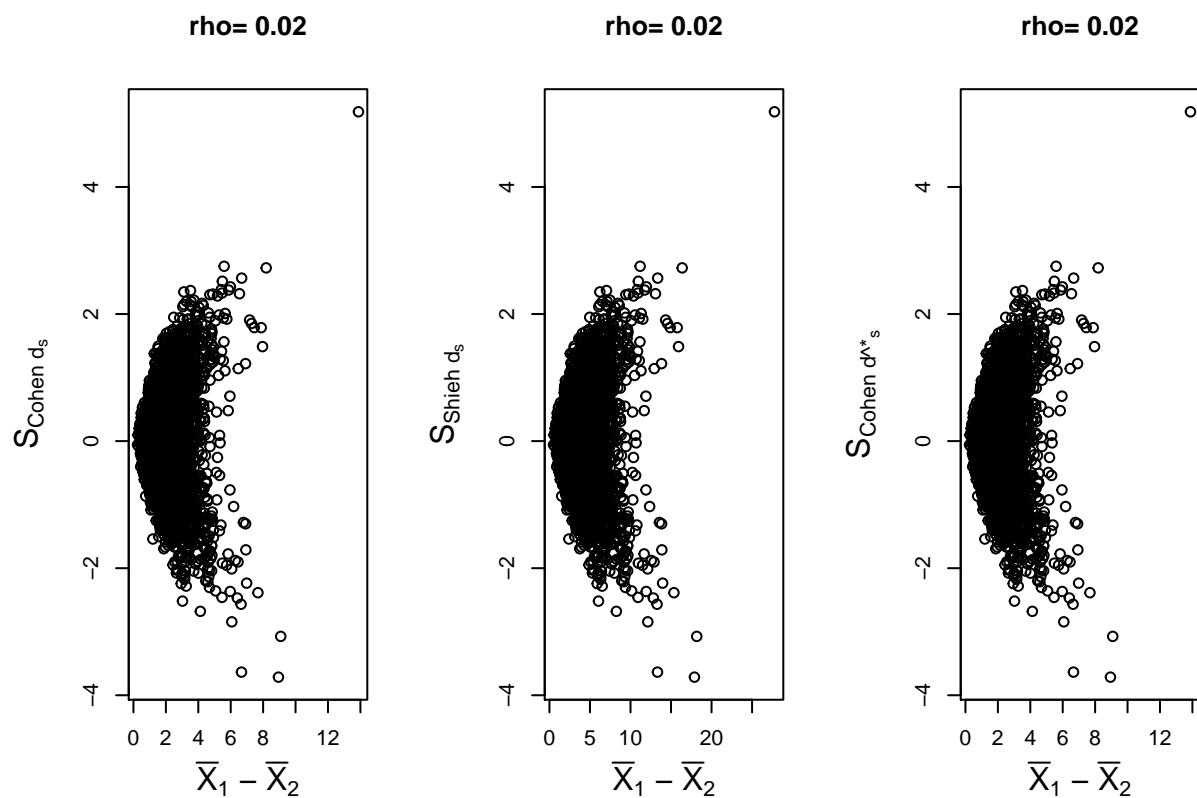


Figure 3. $S_{Glass's} d_s$, $S_{Shieh's} d_s$ and $S_{Cohen's} d_s^*$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

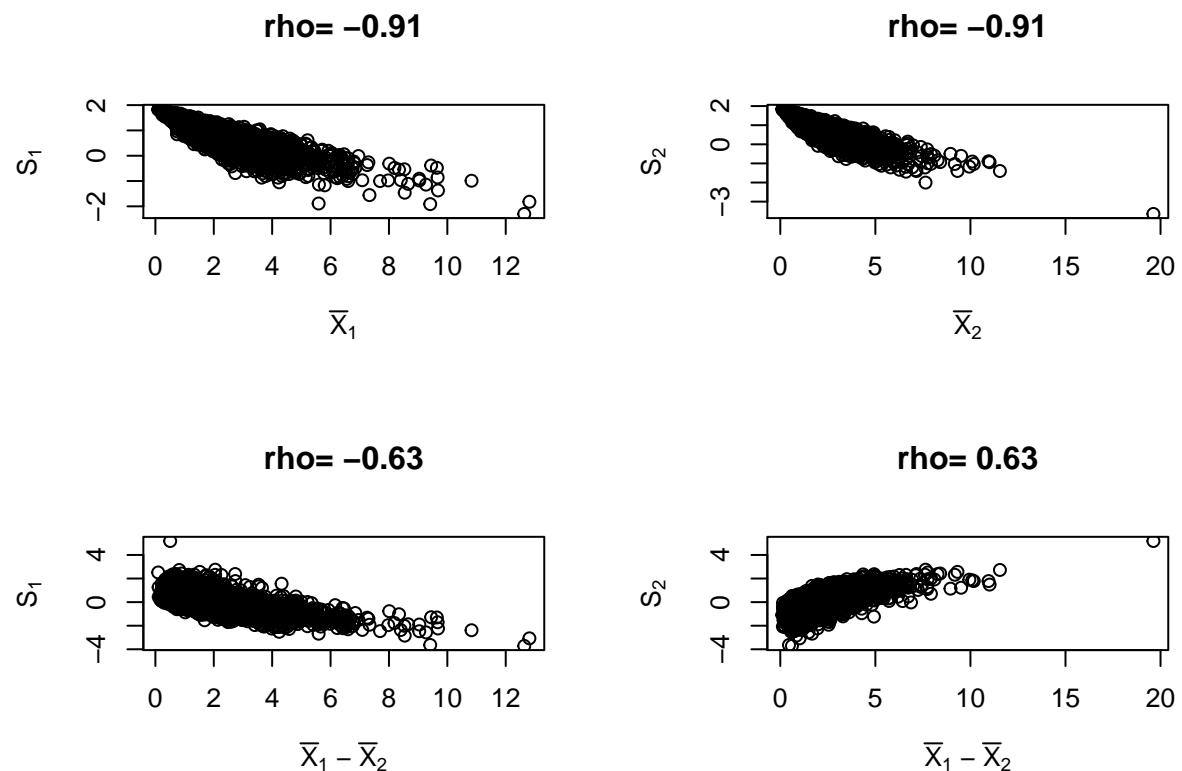


Figure 4. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

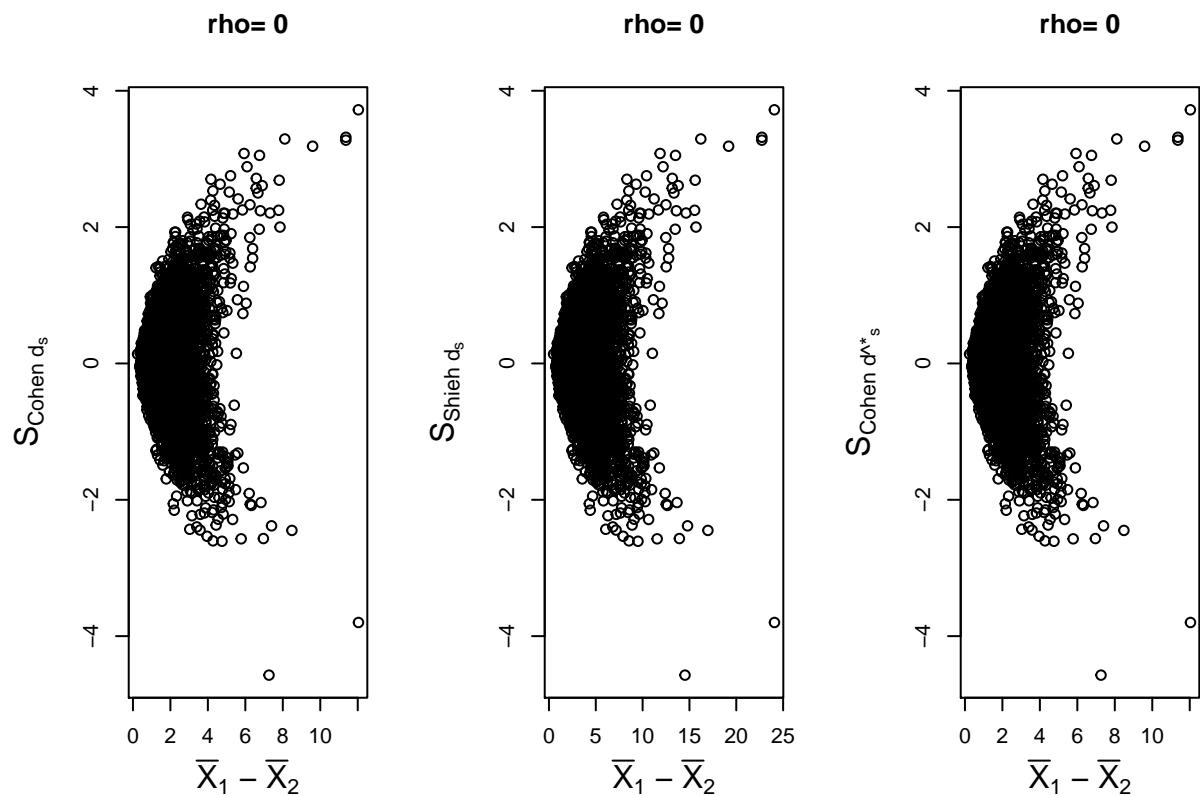


Figure 5. $S_{\text{Glass}'s} d_s$, $S_{\text{Shieh}'s} d_s$ and $S_{\text{Cohen}'s} d_s^*$ as a function of the mean difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

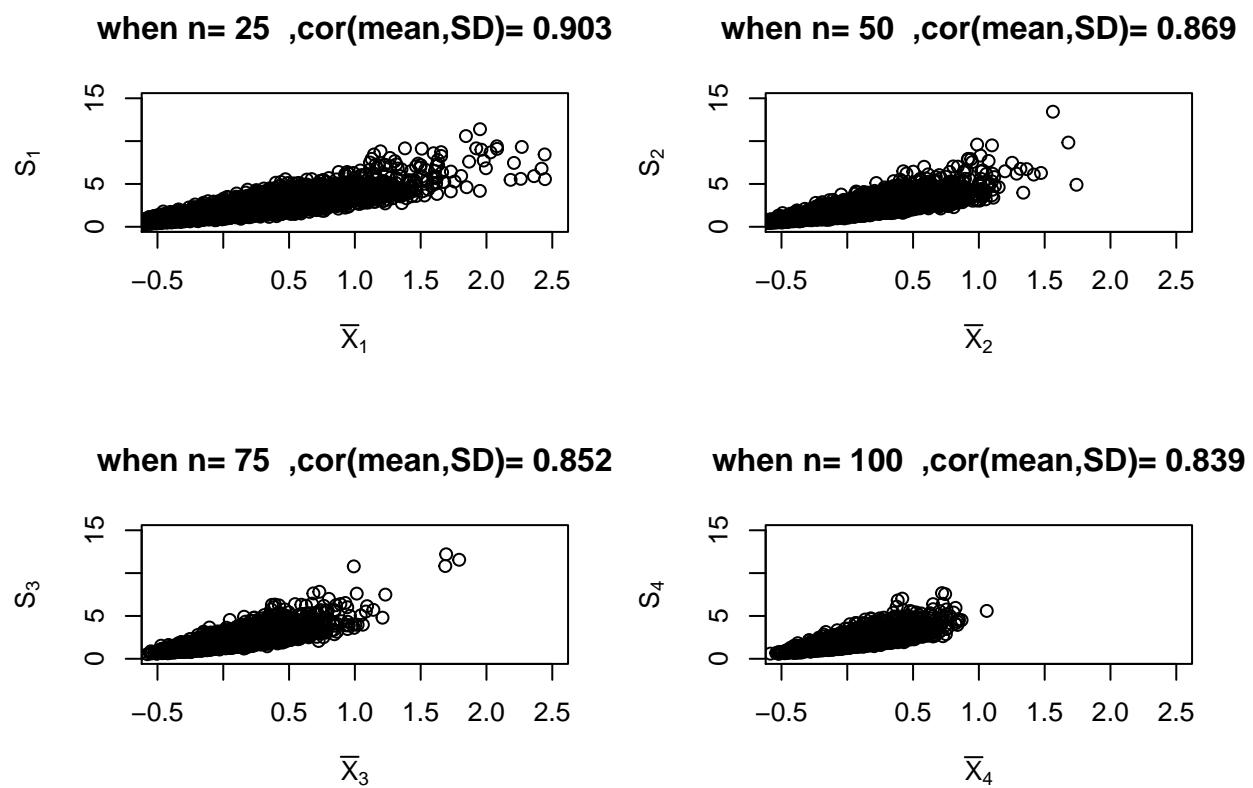


Figure 6. correlation between S_j and \bar{X}_j when $n = 25, 50, 75$ or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

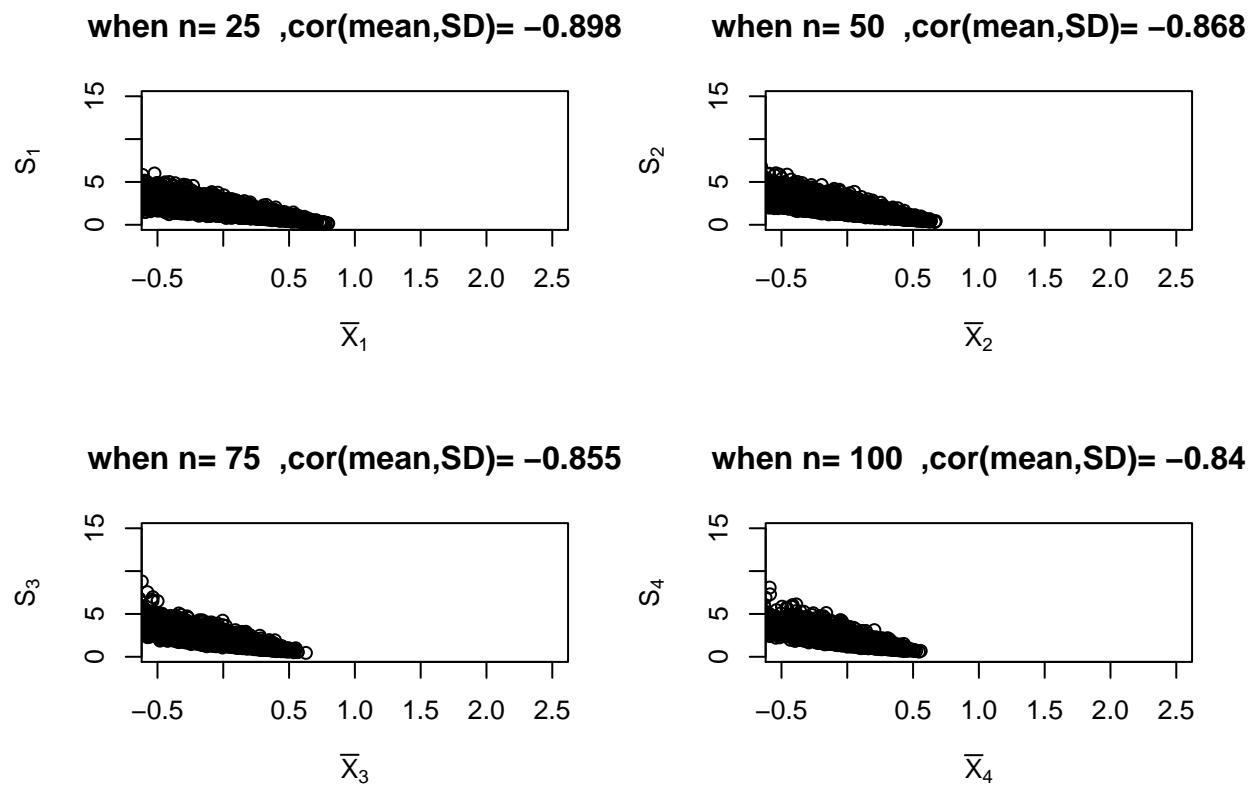


Figure 7. correlation between S_j and \bar{X}_j when $n = 25, 50, 75$ or 100 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

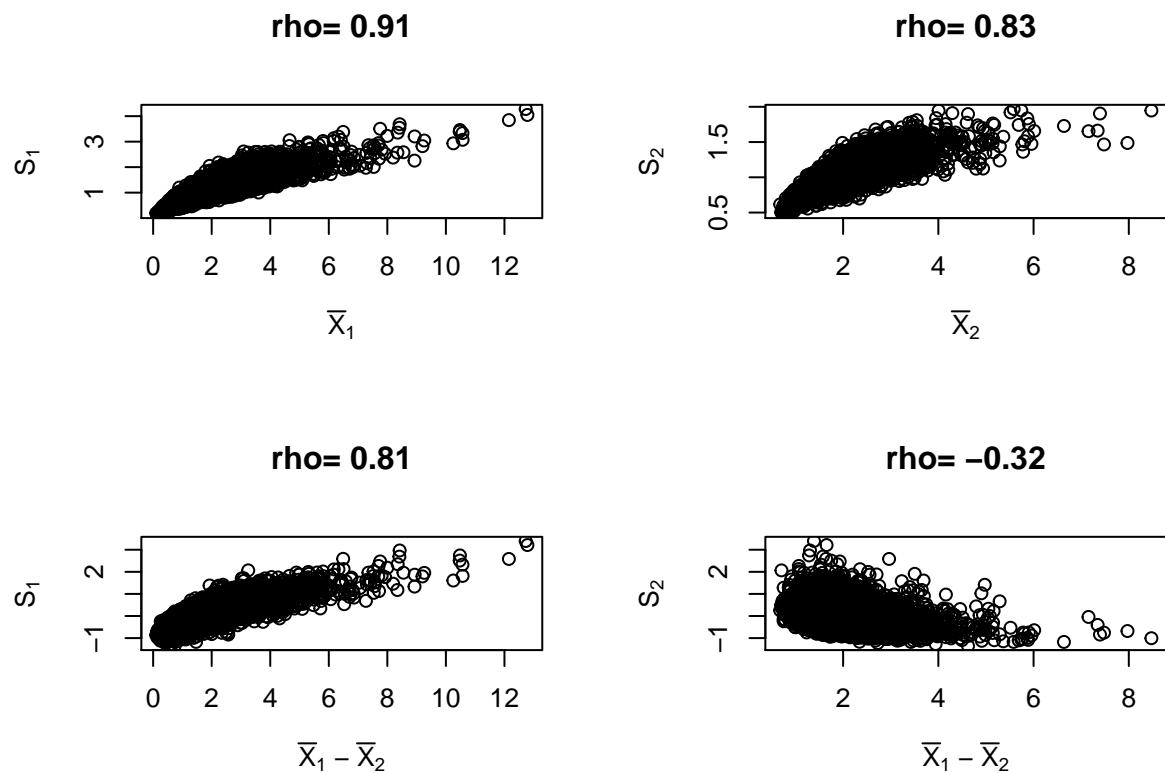


Figure 8. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $n_1=20$ and $n_2=100$

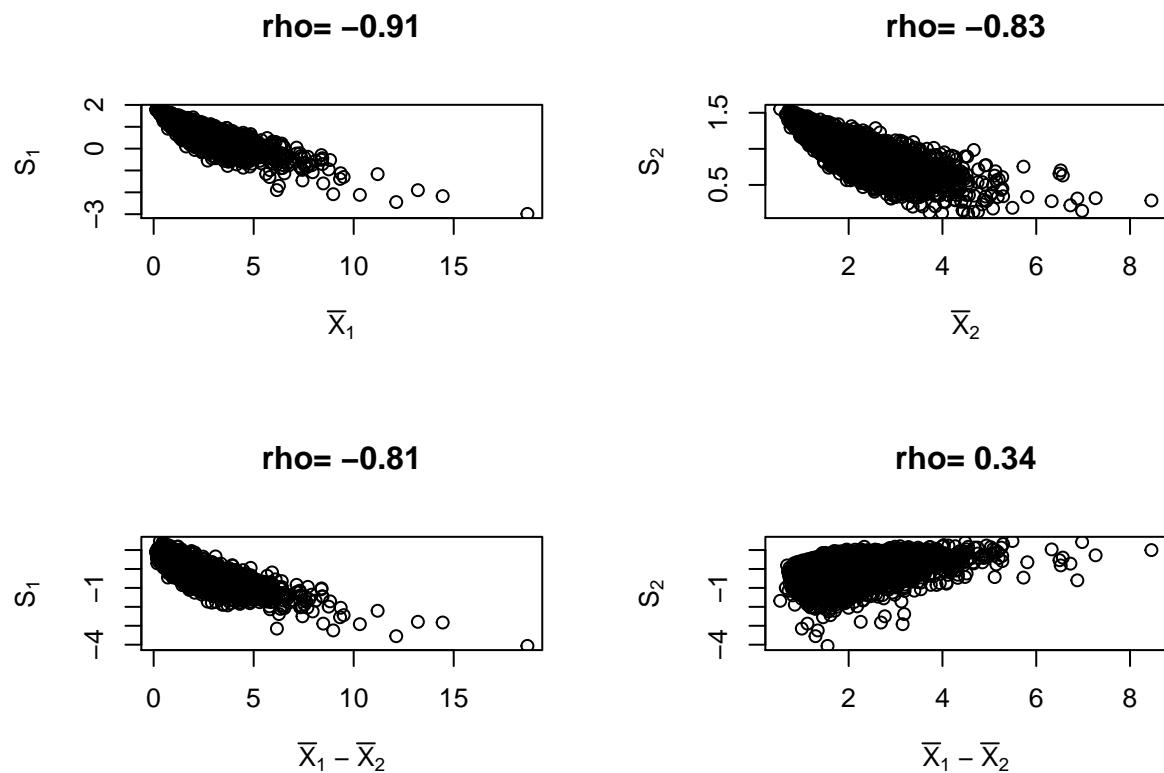


Figure 9. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $n_1=20$ and $n_2=100$

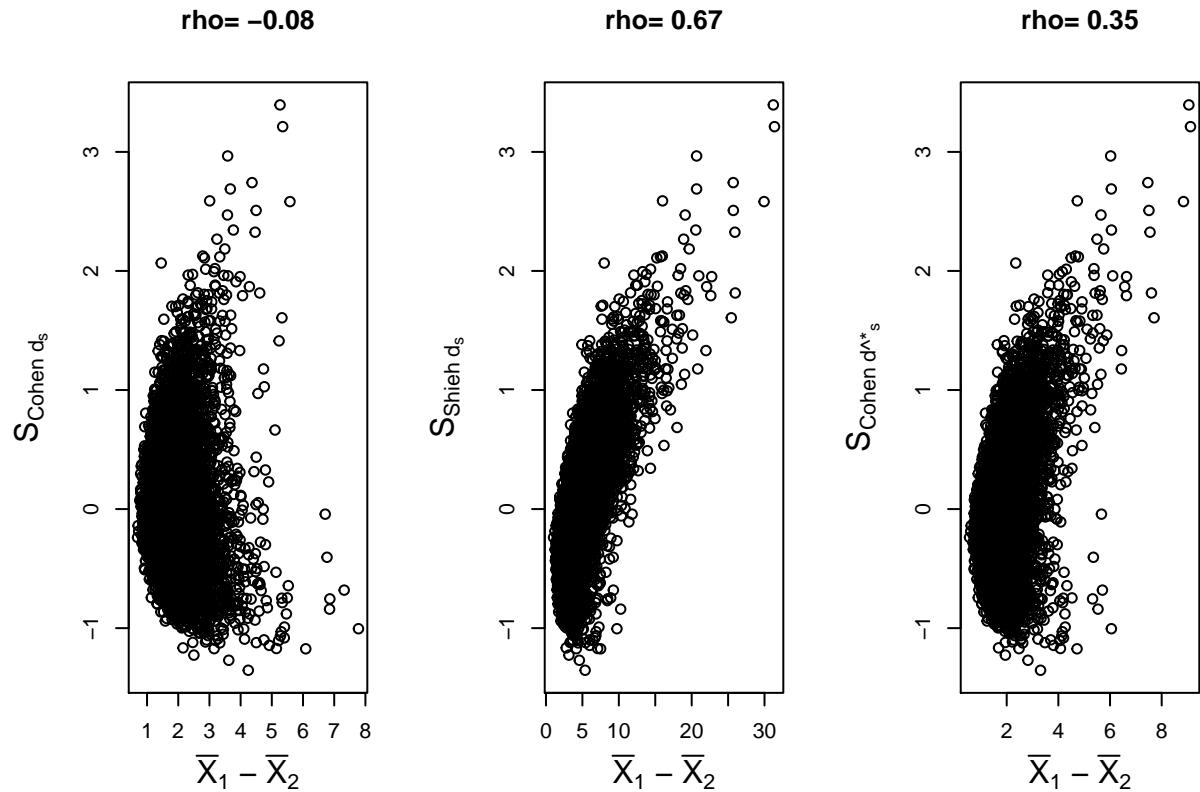


Figure 10. $S_{Cohen's\ ds}$, $S_{Shieh's\ ds}$ and $S_{Cohen's\ d^*_s}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$, with $n_1=20$ and $n_2=100$)

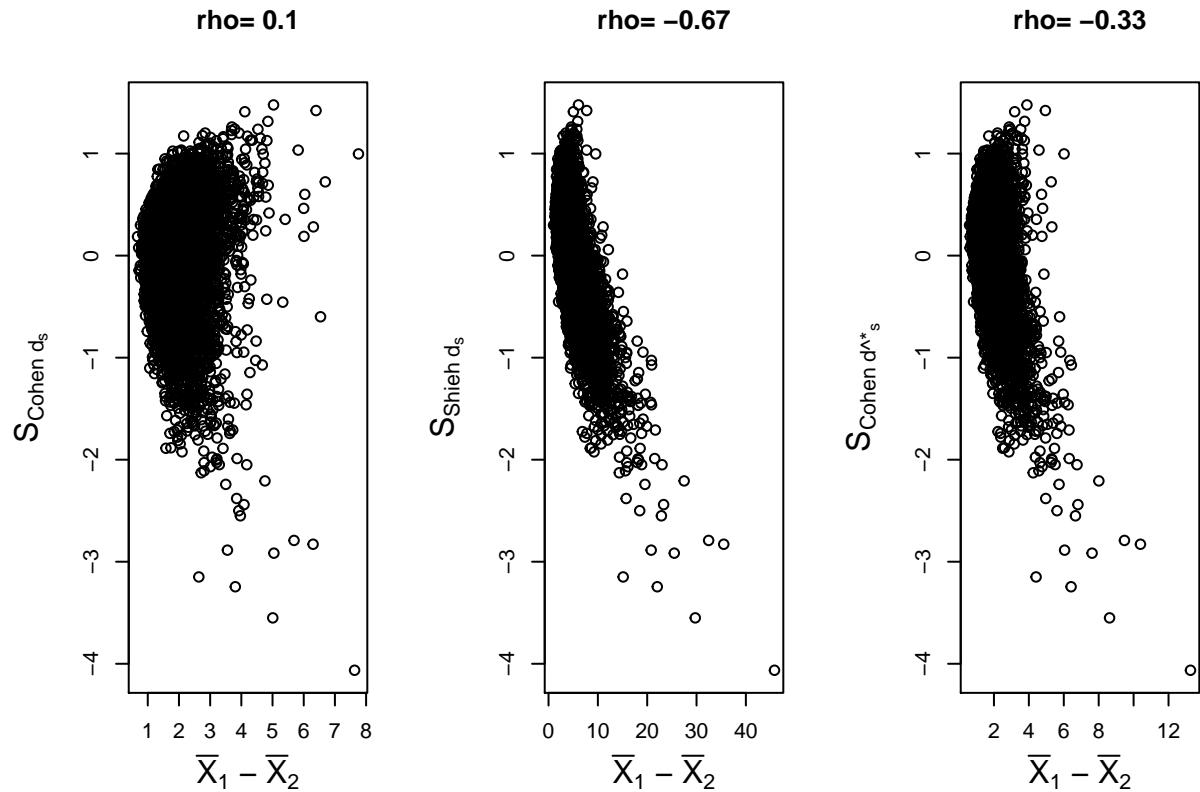


Figure 11. $S_{Cohen's\ ds}$, $S_{Shieh's\ ds_s}$ and $S_{Cohen's\ d^*s}$ as a function of the mean difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $n_1=20$ and $n_2=100$

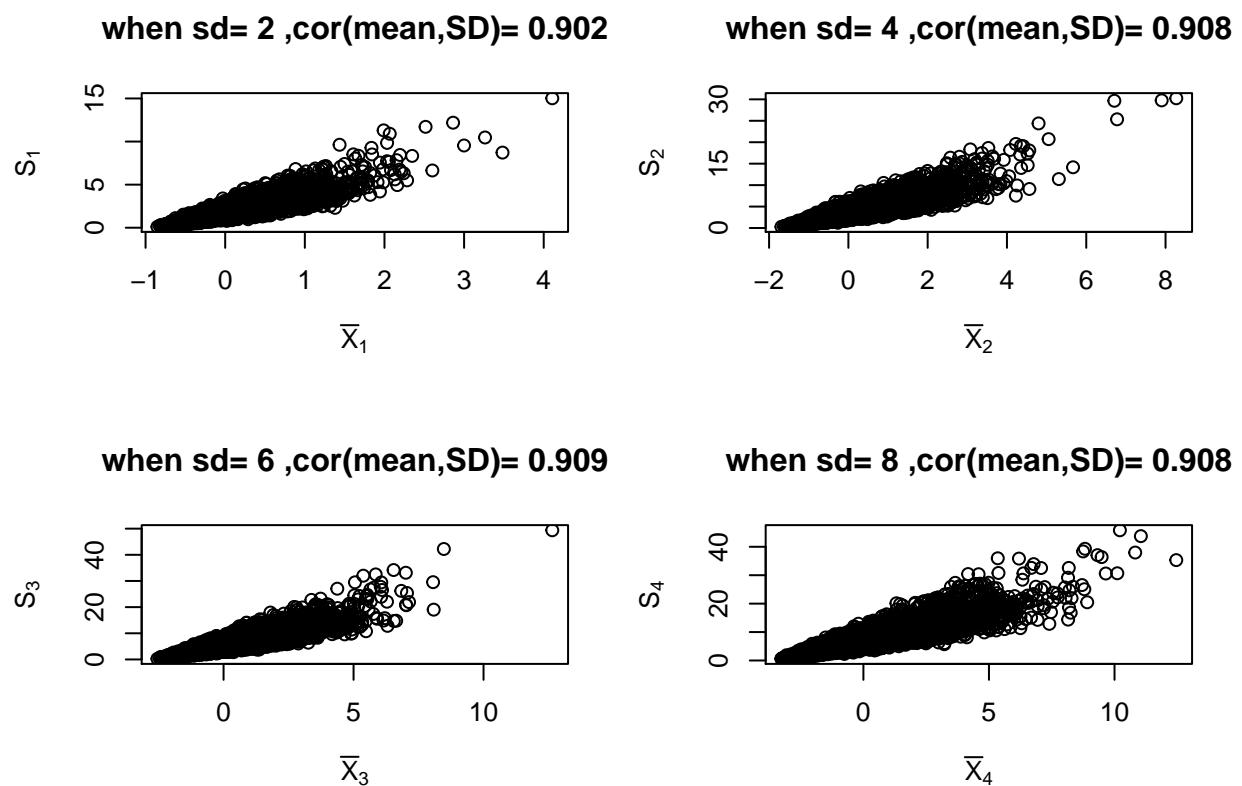


Figure 12. correlation between S_j and \bar{X}_j when $SD = 2, 4, 6$ or 8 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

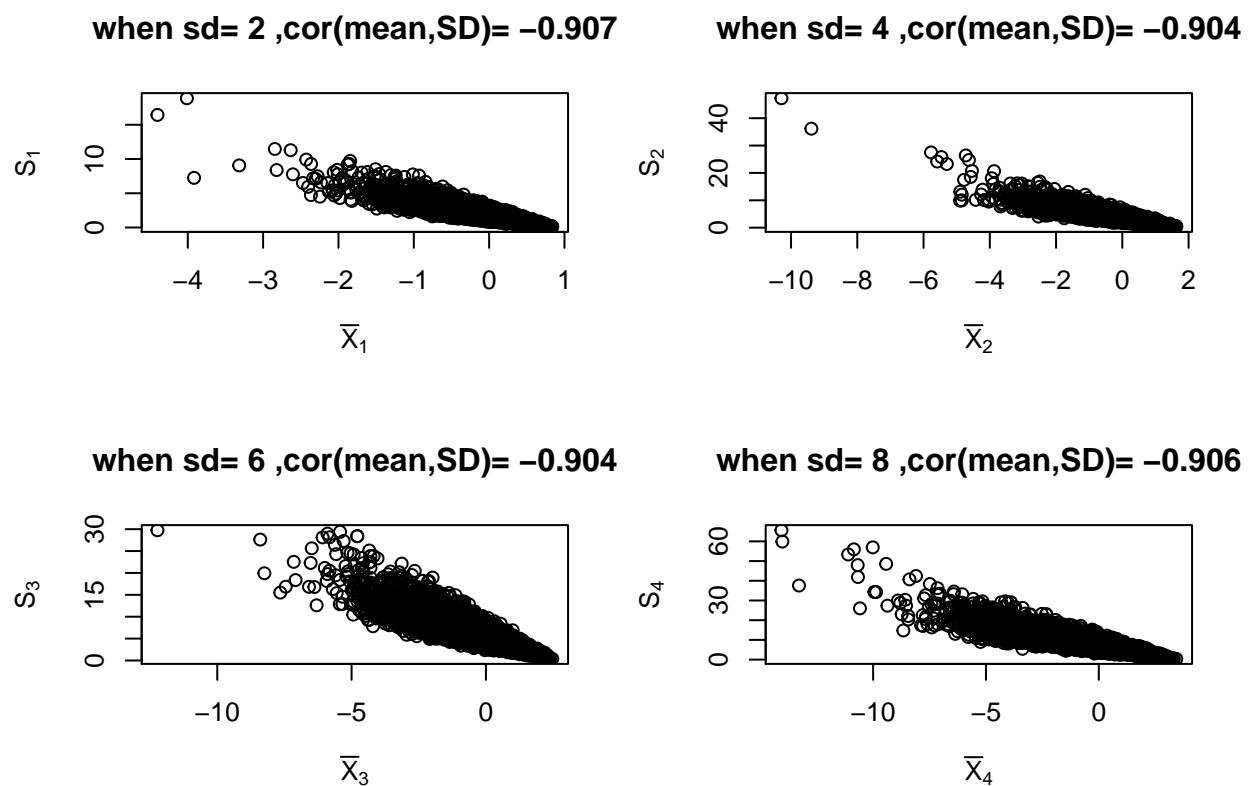


Figure 13. correlation between S_j and \bar{X}_j when $SD = 2, 4, 6$ or 8 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

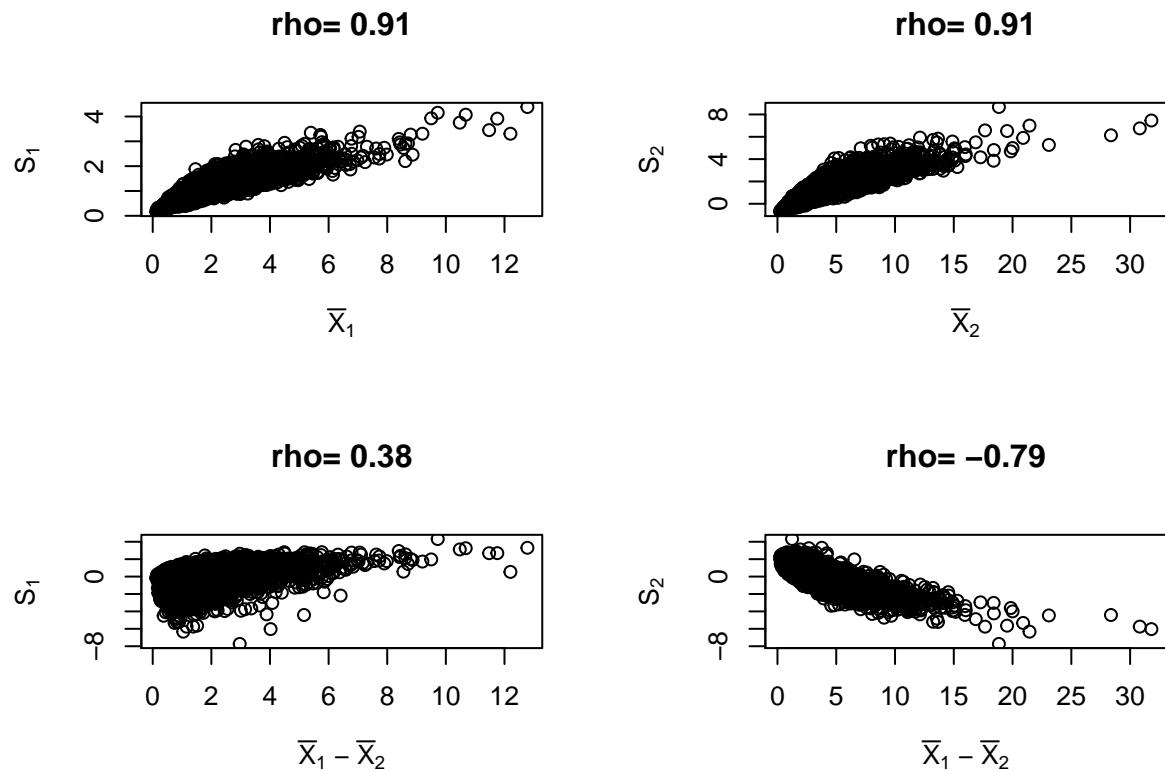


Figure 14. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $S_1=2$ and $S_2=4$

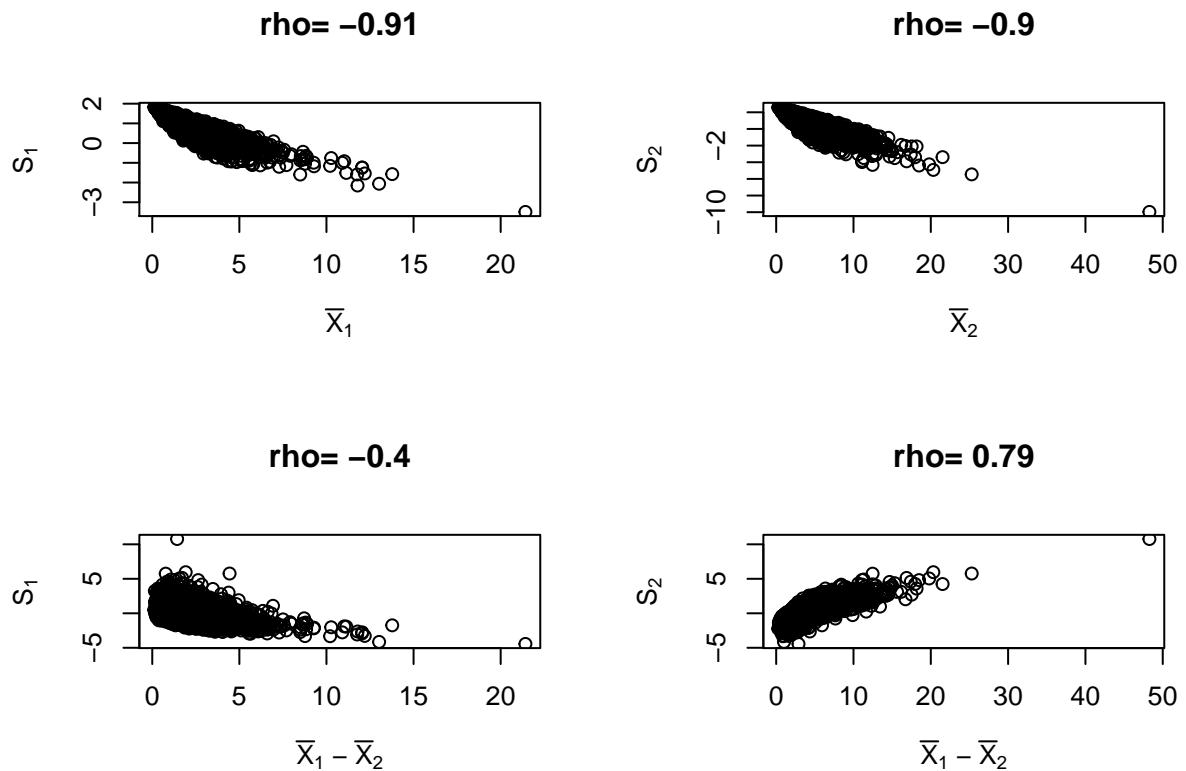


Figure 15. S_j ($j=1,2$) as a function \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $S_1=2$ and $S_2=4$

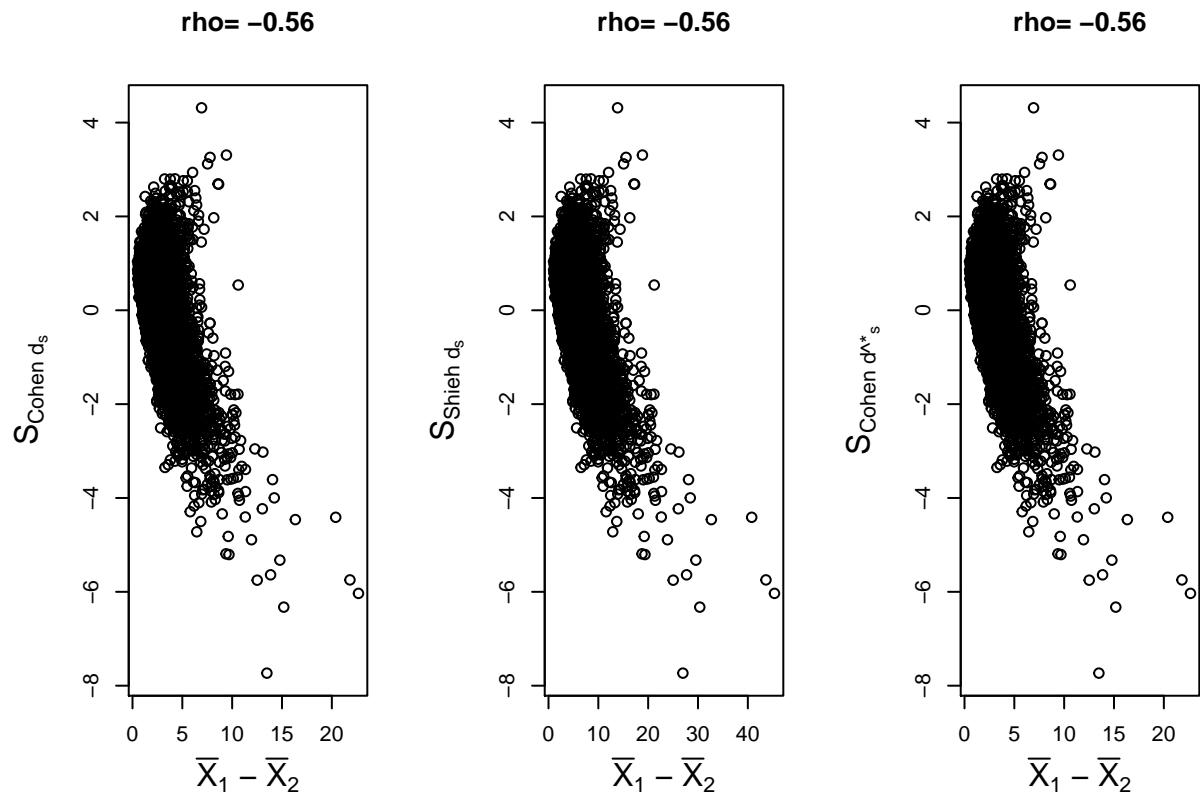


Figure 16. $S_{Cohen's} d_s$, $S_{Shieh} d_s$ and $S_{Cohen's} d_s^*$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $S_1=2$ and $S_2=4$

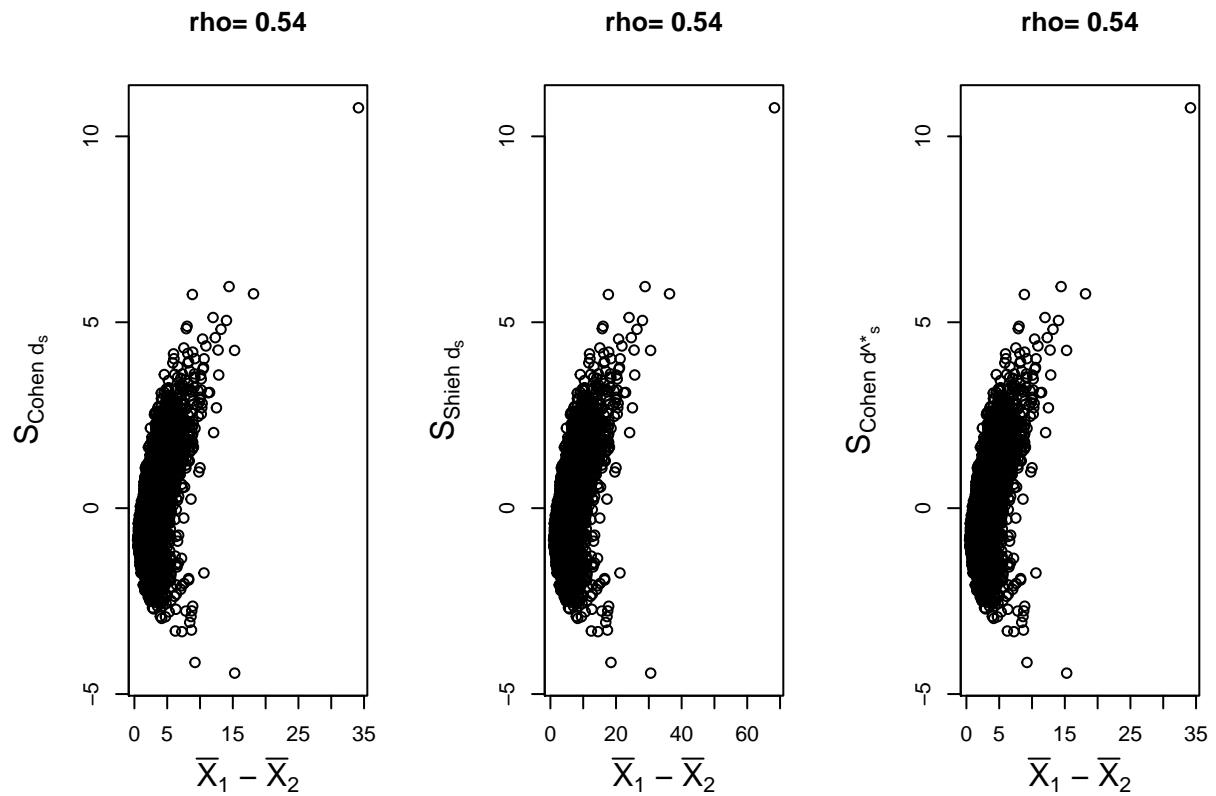


Figure 17. $S_{Cohen's\,d_s}$, $S_{Shieh's\,d_s}$ and $S_{Cohen's\,d_s^*}$ as a function of the mean difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $S_1=2$ and $S_2=4$