

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

- 3 Correlations between the sample mean difference and standardizers of all estimators, and
 4 implications on biases and variances of all estimators

5 **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 . There 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):

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

- 6 When the equality of variances assumption is not met, we are considering three
 7 alternative estimates:

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

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

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

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

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

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

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

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

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

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

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

³¹ “All other parameters being equal” is mentioned because it is always possible that
³² other factors in action have an opposite effect on bias and variance in order that increasing
³³ the magnitude of the correlation between S_i and $\bar{X}_1 - \bar{X}_2$ does not necessarily reduce the

³⁴ bias and the variance. For example, when population variances are equal across groups and
³⁵ sample sizes are unequal, we will see below that the lower n_j , the larger the magnitude of
³⁶ the 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
³⁷ positive, the smaller the sample size, the larger the positive correlation. At the same time,
³⁸ we know that increasing the sample size decreases the bias. This is a nice example of
³⁹ situations where two factors might have an opposite action on bias.

⁴⁰ Correlations between the mean difference ($\bar{X}_1 - \bar{X}_2$) and all standardizers

⁴¹ When equal population variances are estimated based on equal sample sizes ⁴² (condition a)

⁴³ While \bar{X}_j and S_j ($j=1,2$) are uncorrelated when samples are extracted from
⁴⁴ symmetric distributions (see Figure 1), there is a non-null correlation between \bar{X}_j and S_j
⁴⁵ when distributions are skewed (Zhang, 2007).

⁴⁶ More specifically, when distributions are right-skewed, there is a **positive** correlation
⁴⁷ between \bar{X}_j and S_j (see the two top plots in Figure 2), resulting in a *positive* correlation
⁴⁸ 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
⁴⁹ two bottom plots in Figure 2). This can be explained by the fact that \bar{X}_1 and $\bar{X}_1 - \bar{X}_2$ are
⁵⁰ positively correlated while \bar{X}_2 and $\bar{X}_1 - \bar{X}_2$ are negatively correlated (of course,
⁵¹ correlations would be trivially reversed if we computed $\bar{X}_2 - \bar{X}_1$ instead of $\bar{X}_1 - \bar{X}_2$).

⁵² One should also notice that both correlations between S_j and $\bar{X}_1 - \bar{X}_2$ are equal, in
⁵³ absolute terms (possible tiny differences might be observed due to sampling error in our
⁵⁴ simulations). As a consequence, when computing a standardizer taking both S_1 and S_2 into
⁵⁵ account, it results in a standardizer that is uncorrelated with $\bar{X}_1 - \bar{X}_2$ (see Figure 3).

⁵⁶ On the other hand, when distributions are left-skewed, there is a **negative**
⁵⁷ correlation between \bar{X}_j and S_j (see the two top plots in Figure 4), resulting in a *negative*

58 correlation between S_1 and $\bar{X}_1 - \bar{X}_2$ and in a *positive* correlation between S_2 and $\bar{X}_1 - \bar{X}_2$
 59 (see the two bottom plots in Figure 4).

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

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

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

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

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

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

76 - When computing $S_{\text{Cohen's } d_s^*}$, the same weight is given to both S_1 and S_2 . Therefore,
 77 it does not seem surprising that the sign of the correlation between $S_{\text{Cohen's } d_s^*}$ and $\bar{X}_1 - \bar{X}_2$
 78 is the same as the size of the correlation between $\bar{X}_1 - \bar{X}_2$ and the SD of the smallest
 79 sample;

80 - When computing $S_{\text{Shieh's } d_s}$, more weight is given to the SD of the smallest sample,
 81 it is therefore not really surprising to observe that the correlation between $S_{\text{Shieh's } d_s}$ and
 82 $\bar{X}_1 - \bar{X}_2$ is closer of the correlation between the SD of the smallest group and $\bar{X}_1 - \bar{X}_2$

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

84 - When computing S_{Cohen} , more weight is given to the SD of the largest sample,

85 which by compensation effect brings the correlation very close to 0.

86 The correlation between $\bar{X}_1 - \bar{X}_2$ and respectively S_1, S_2 , the standardizer of Cohen's

87 d_s^* , the standardizer of Shieh's d_s and the standardizer of Cohen's d_s are summarized in

88 Table 1.

89 **When unequal population variances are estimated based on equal sample sizes**

90 **(condition c)**

91 When distributions are skewed, there are again non-null correlations between \bar{X}_j and

92 S_j . As illustrated in Figures 12 and 13, the correlation remains the same for any

93 population SD (σ). However, the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$

94 differs: it is stronger in the sample extracted from the larger population variance (see

95 Figures 14 and 15).

96 This also explains that when computing a standardizer that takes both S_1 and S_2

97 into account, it results in a standardizer that is correlated with $\bar{X}_1 - \bar{X}_2$ (see Figures 16

98 and 17). The correlation between the mean difference ($\bar{X}_1 - \bar{X}_2$) and respectively the

99 standardizer of Shieh's d_s , Cohen's d_s^* and Cohen's d_s will have the same sign as the

100 correlation between ($\bar{X}_1 - \bar{X}_2$) and the larger SD . Table 2 summarizes the sign of the

101 correlation between $\bar{X}_1 - \bar{X}_2$ and respectively S_1, S_2 and the three standardizers taking

102 both S_1 and S_2 into 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		
	right-skewed	left-skewed
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
	$S_{Cohen's\ d_s^*}$: 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>

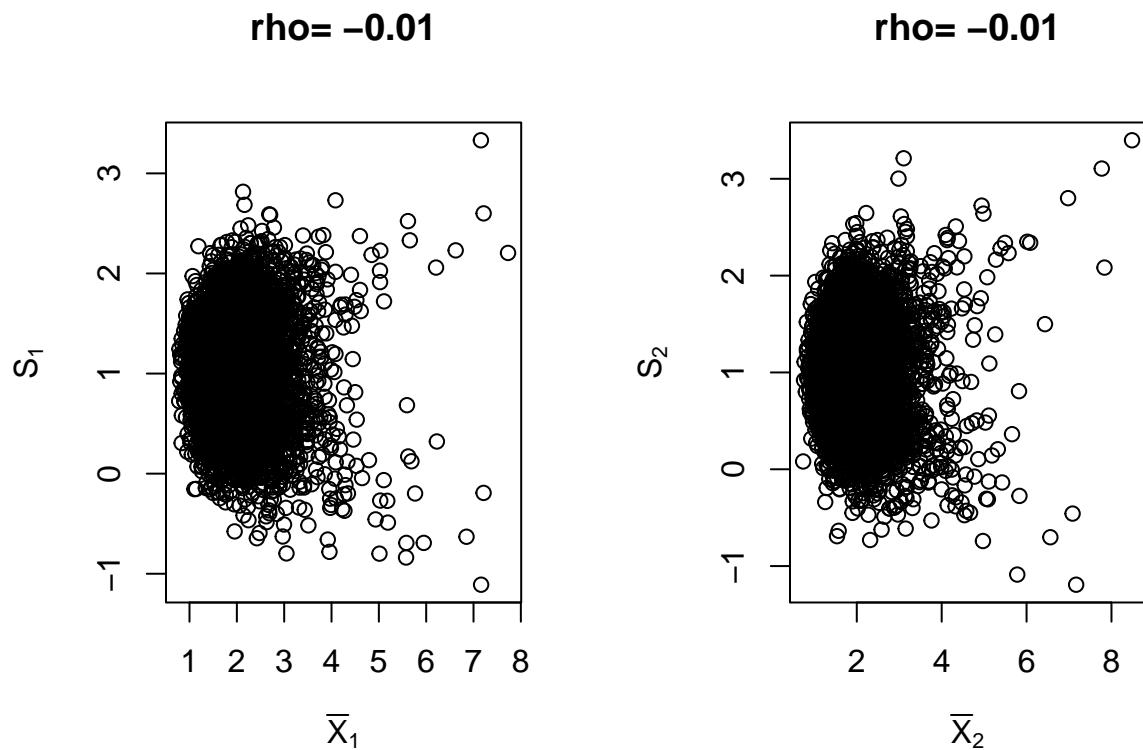


Figure 1. S_j as a function of \bar{X}_j ($j=1,2$), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

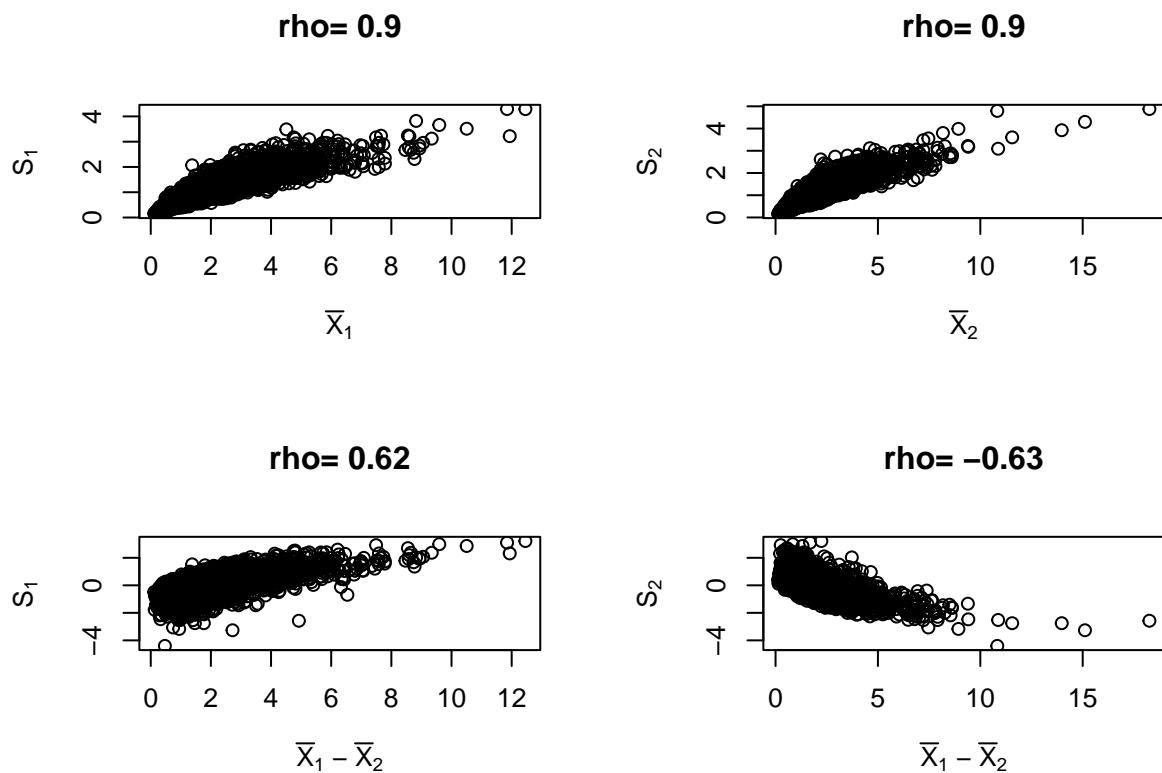


Figure 2. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

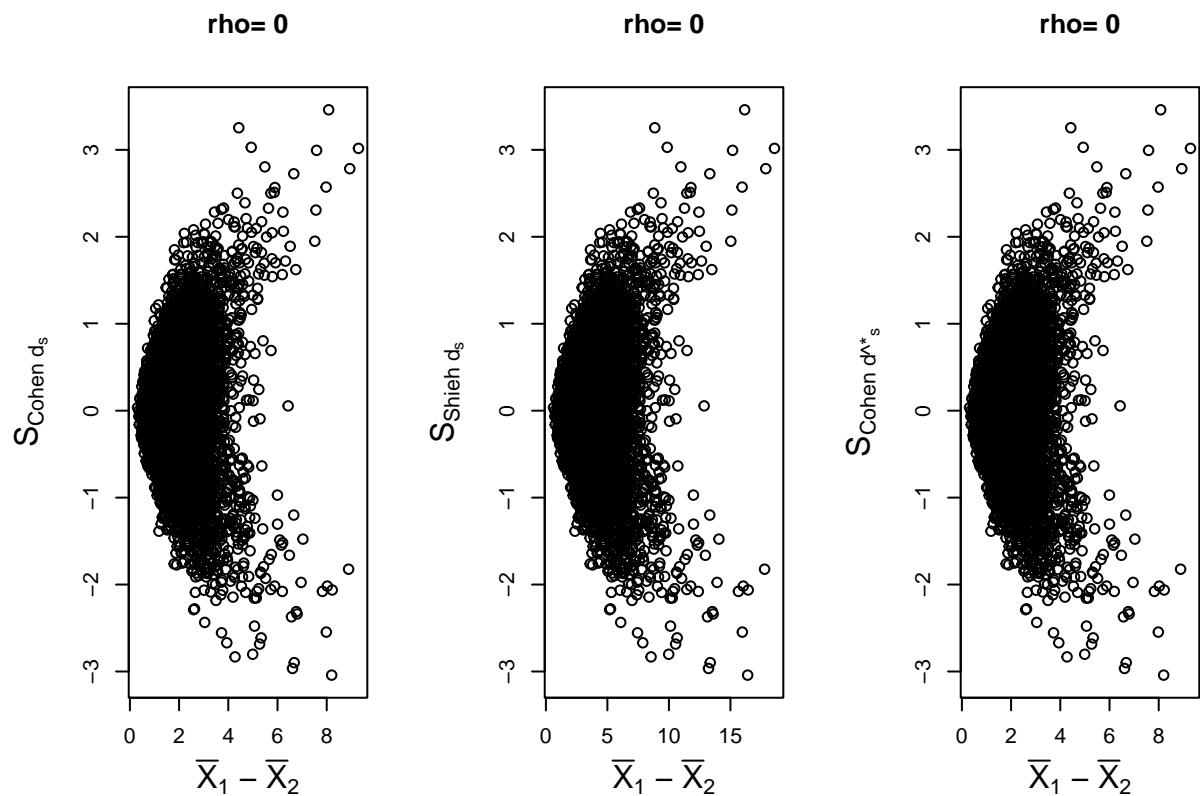


Figure 3. $S_{Cohen's} d_s$, $S_{Shieh} d_s$ and $S_{Cohen's} d_s^*$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

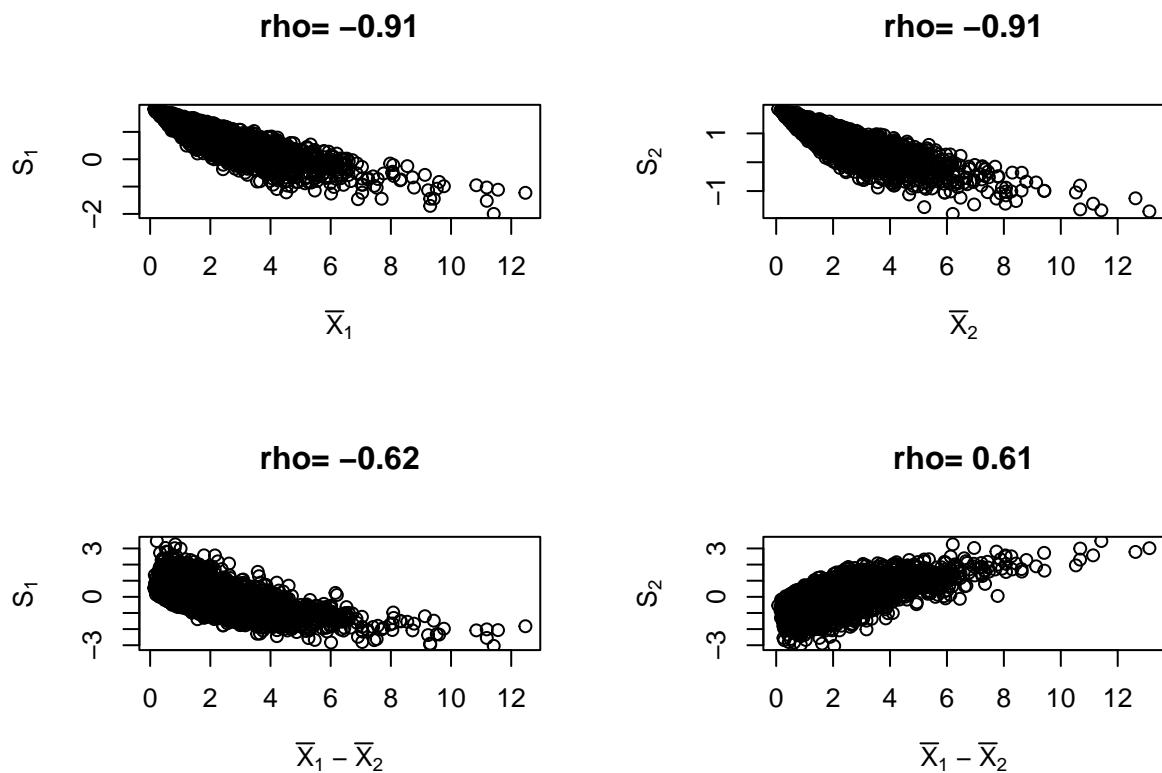


Figure 4. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

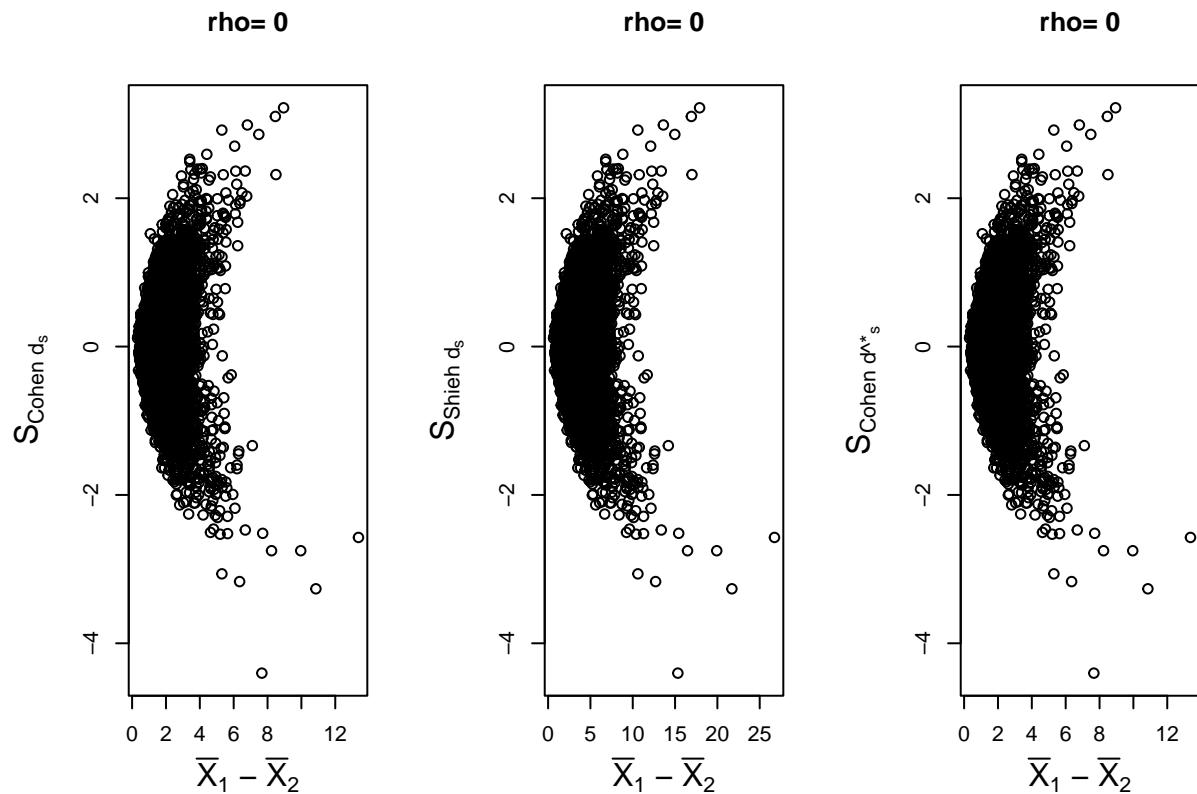


Figure 5. $S_{Cohen's} d_s$, $S_{Shieh's} d_s$ and $S_{Cohen's} d_s^*$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

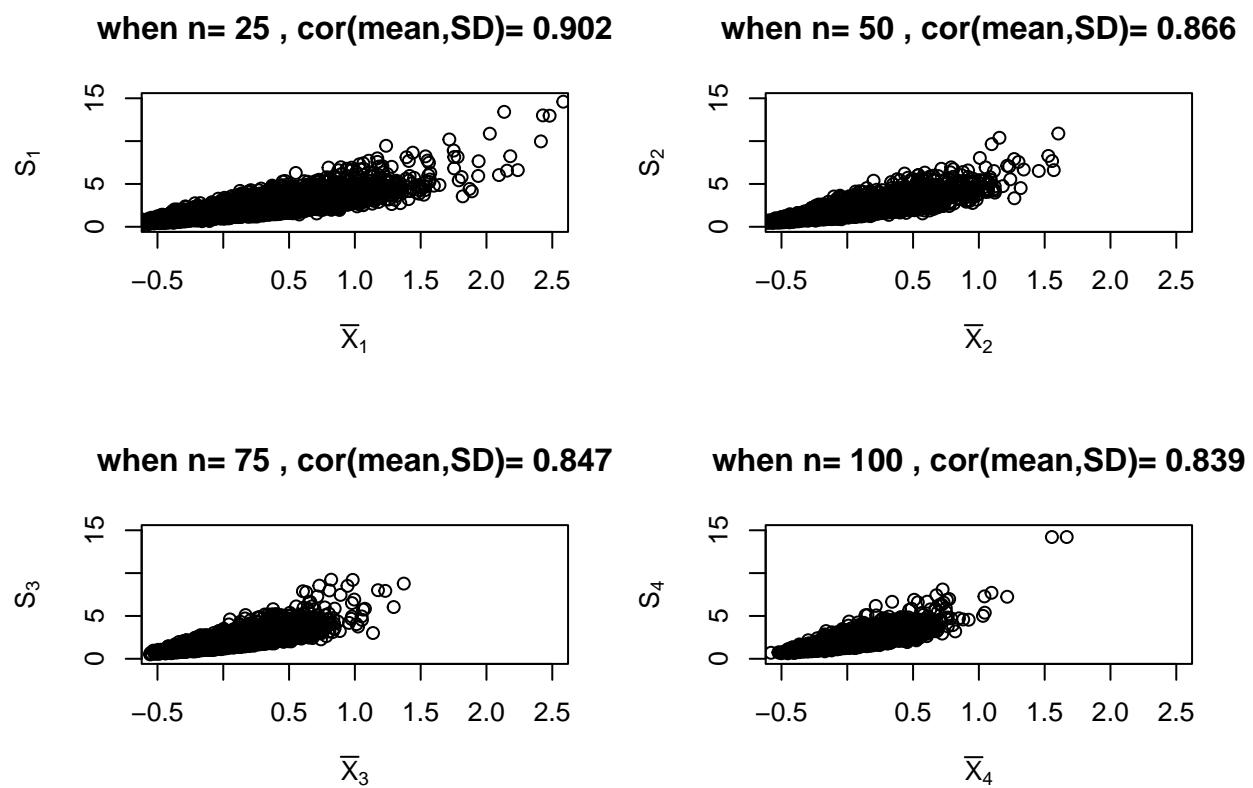


Figure 6. Correlation between S_j and \bar{X}_j when $n = 25, 50, 75$ or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

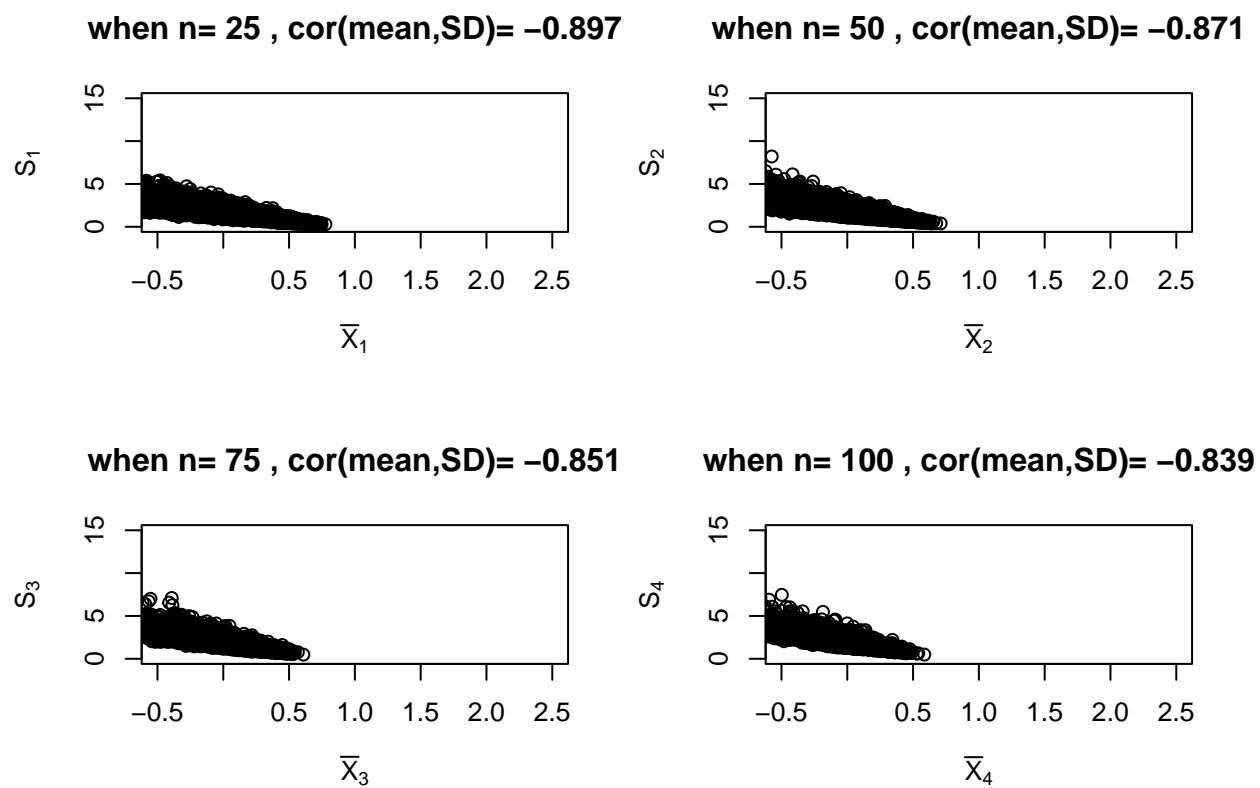


Figure 7. Correlation between S_j and \bar{X}_j when $n = 25, 50, 75$ or 100 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

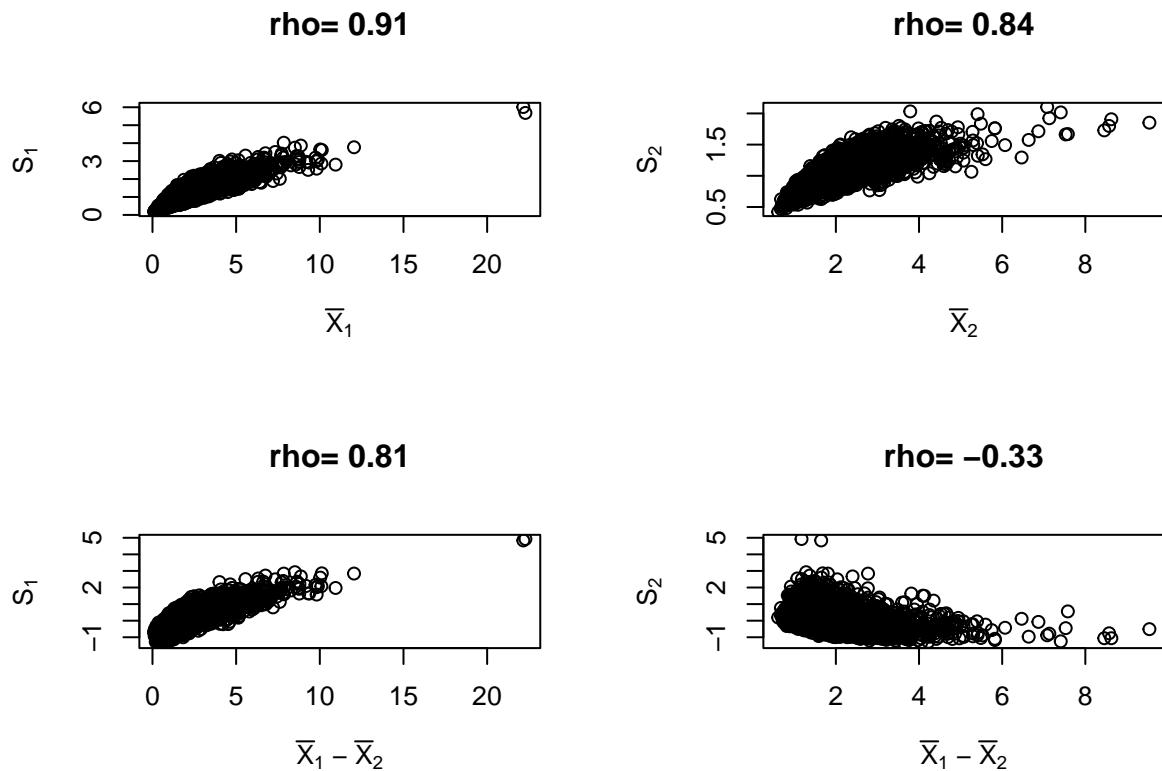


Figure 8. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $n_1=20$ and $n_2=100$

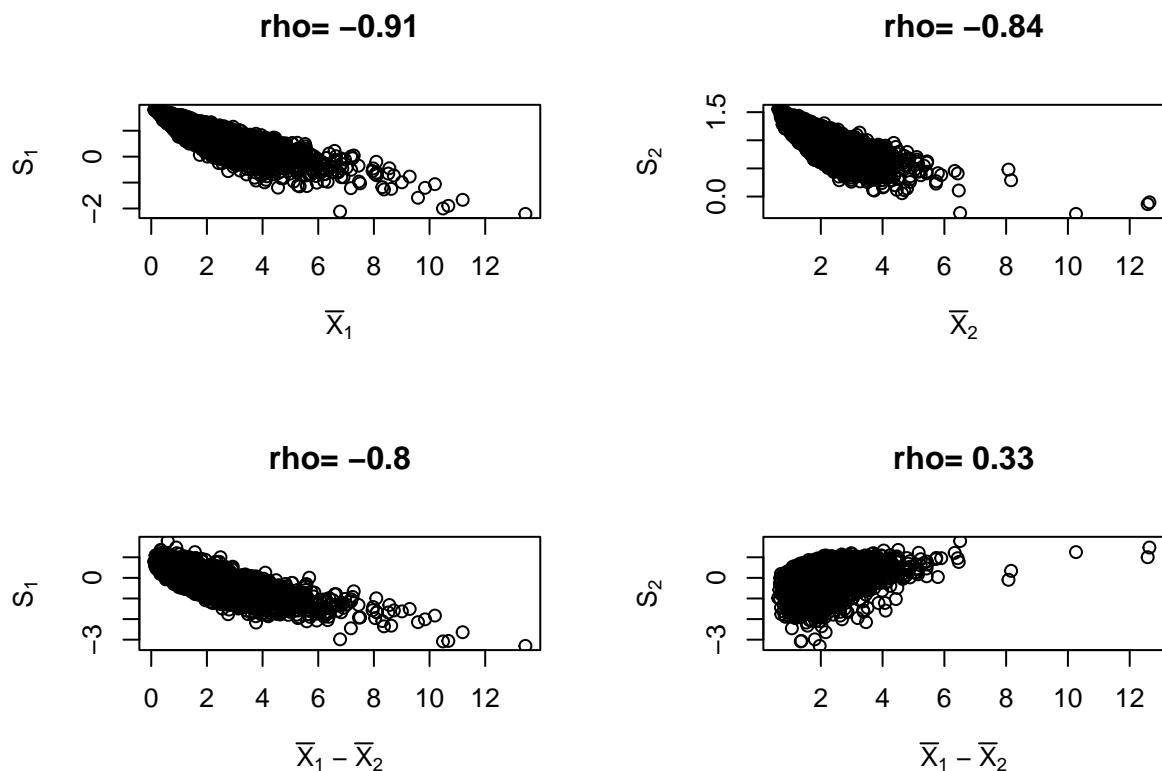


Figure 9. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $n_1=20$ and $n_2=100$

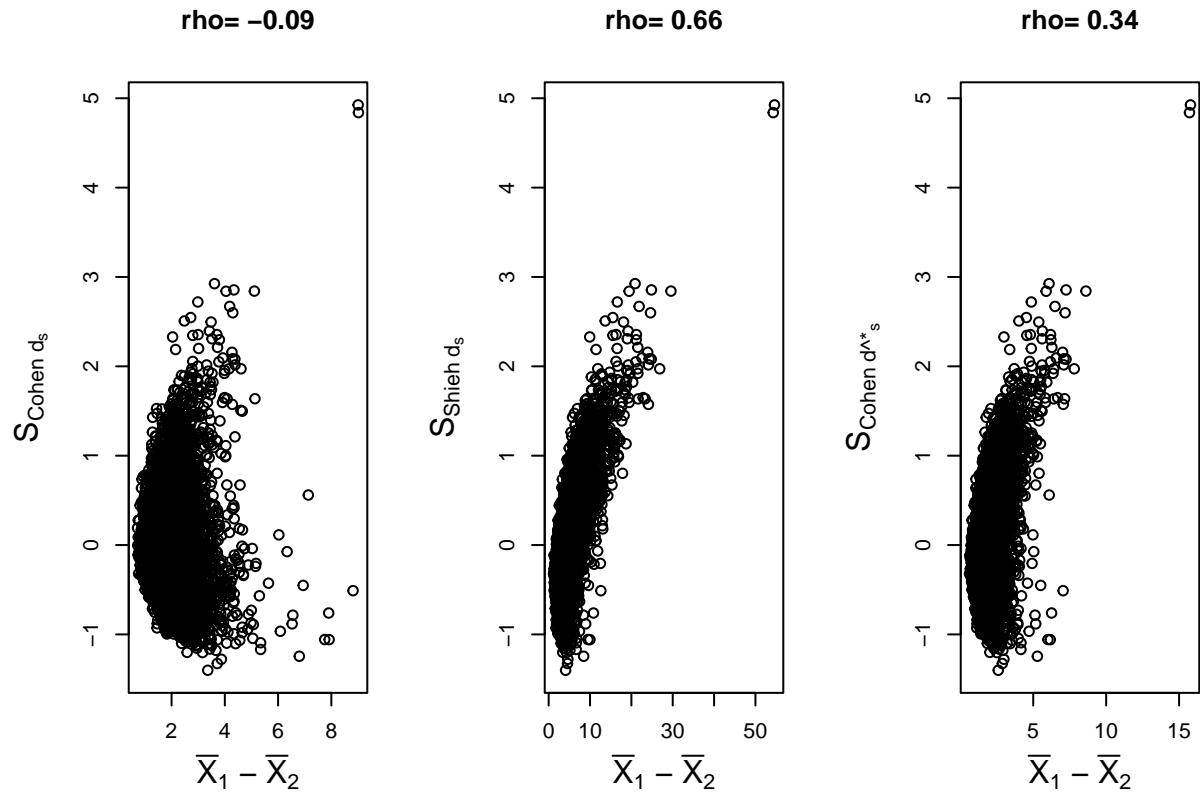


Figure 10. $S_{Cohen's\ ds}$, $S_{Shieh's\ ds}$ and $S_{Cohen's\ d^*s}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$, with $n_1=20$ and $n_2=100$)

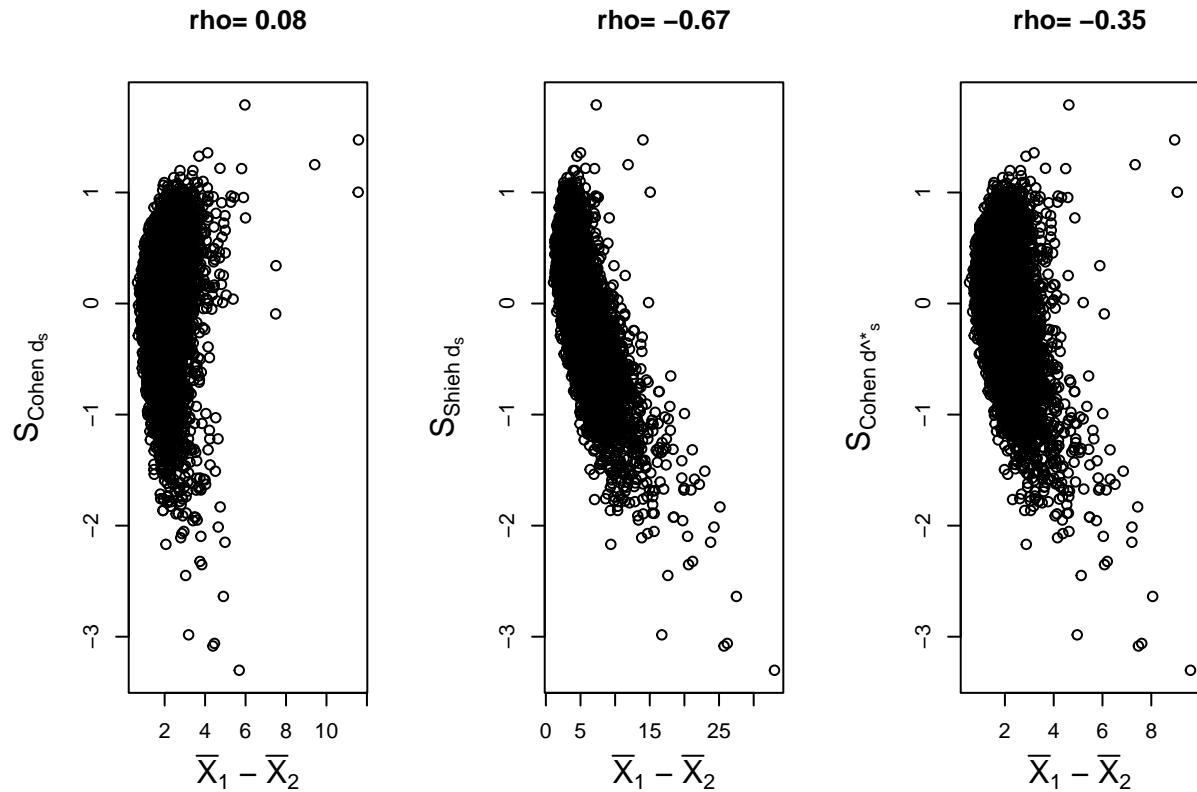


Figure 11. $S_{Cohen's\ ds}$, $S_{Shieh's\ ds}$ and $S_{Cohen's\ d^*s}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $n_1=20$ and $n_2=100$

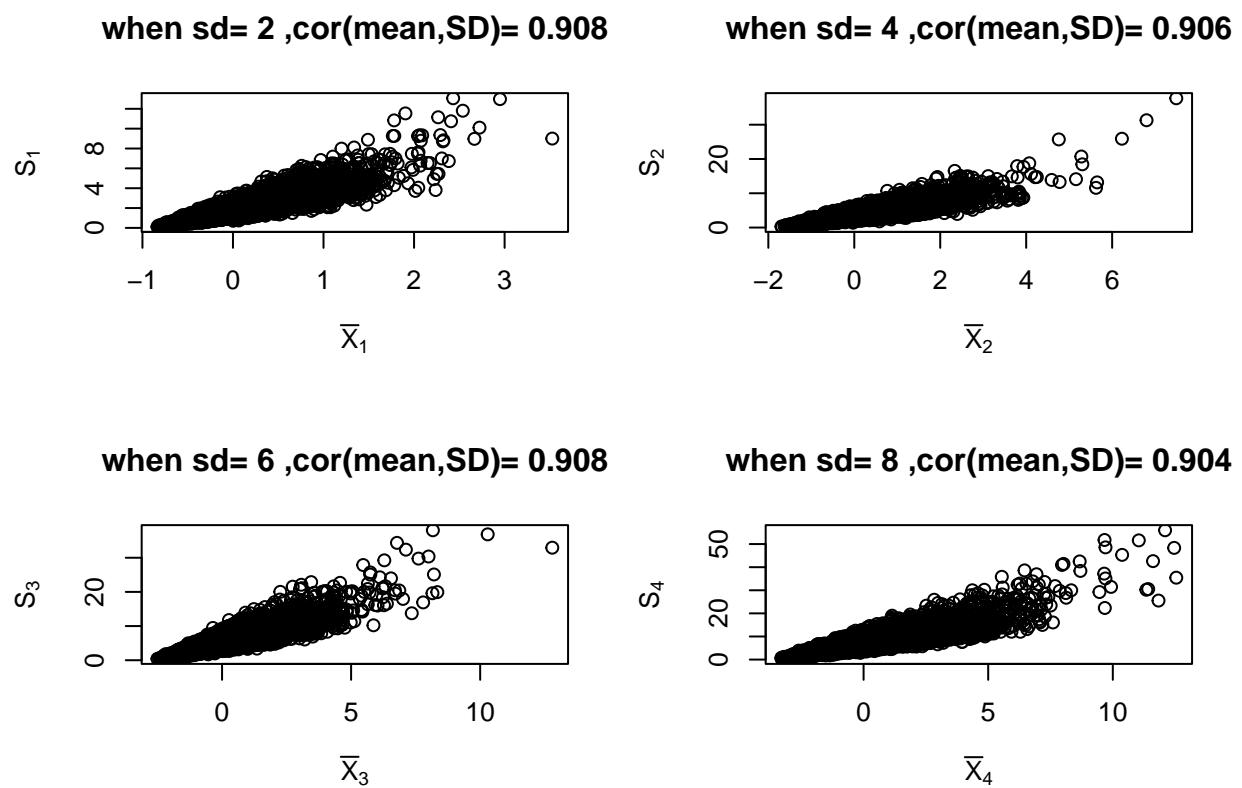


Figure 12. Correlation between S_j and \bar{X}_j when $SD = 2, 4, 6$ or 8 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

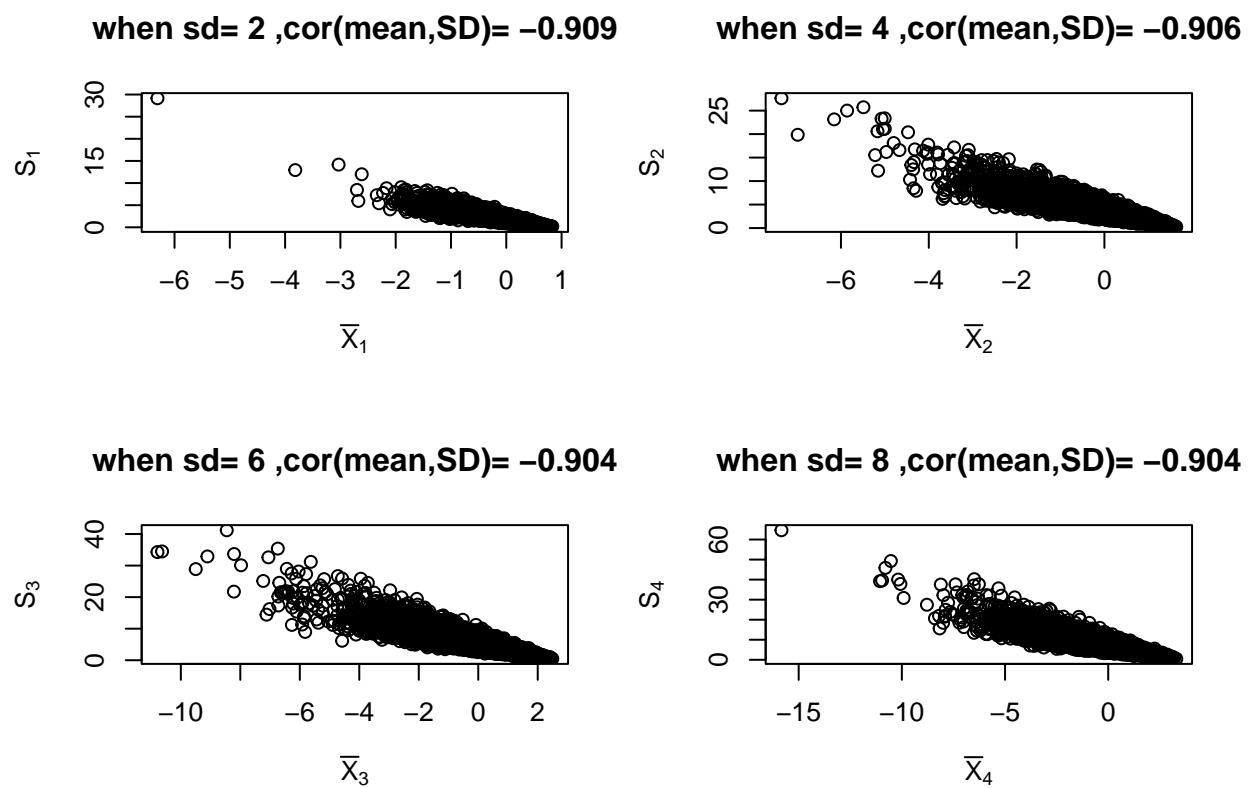


Figure 13. Correlation between S_j and \bar{X}_j when $SD = 2, 4, 6$ or 8 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

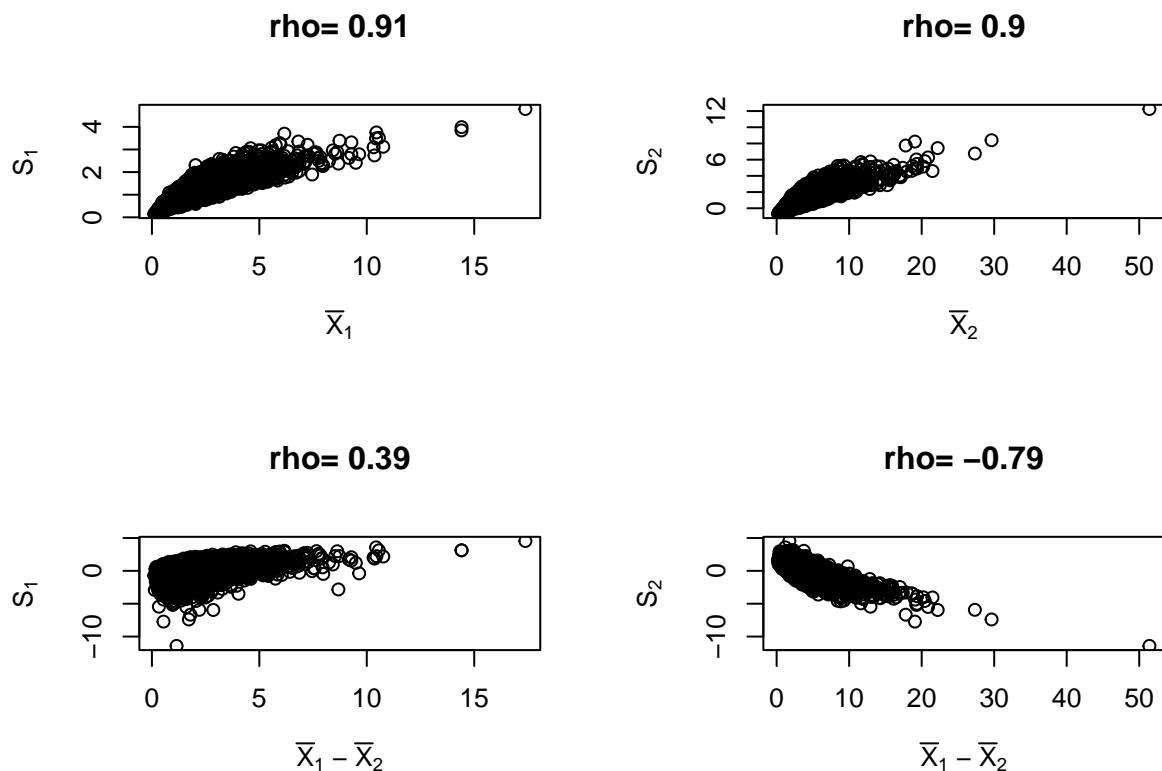


Figure 14. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $S_1=2$ and $S_2=4$

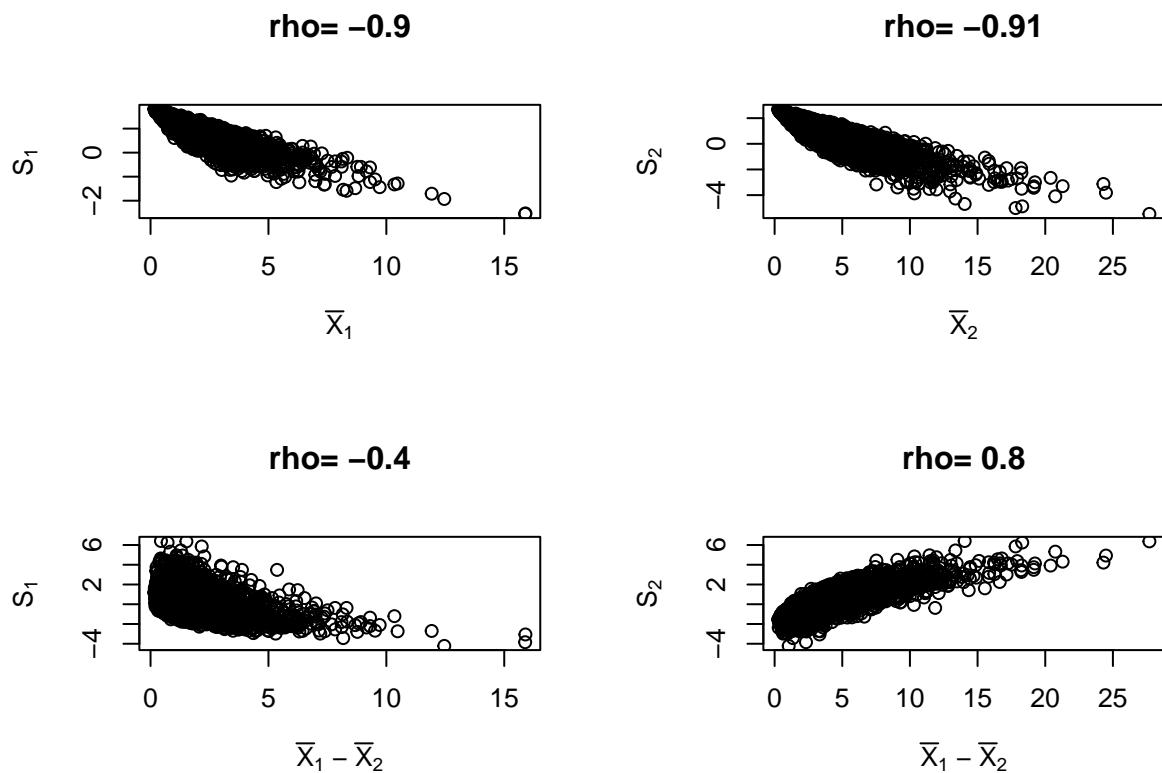


Figure 15. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $S_1=2$ and $S_2=4$

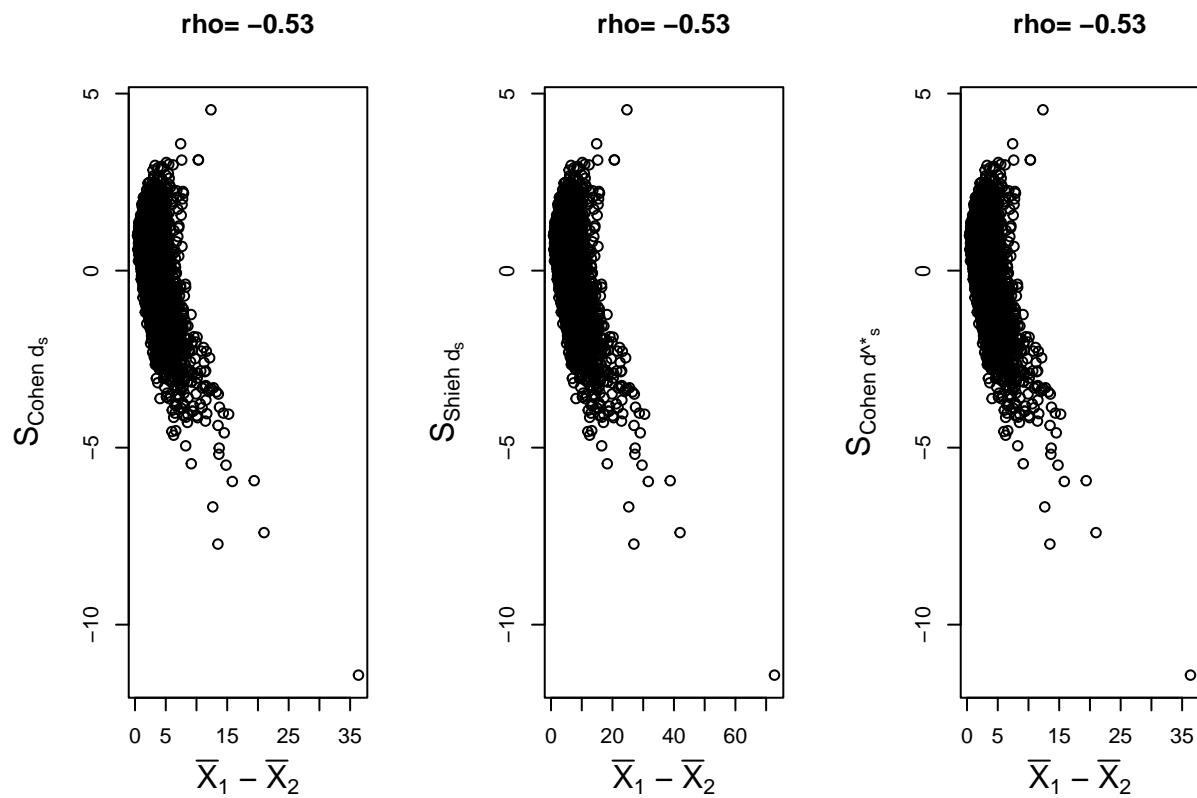


Figure 16. $S_{Cohen's\,d_s}$, $S_{Shieh\,d_s}$ and $S_{Cohen\,d_s^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $S_1=2$ and $S_2=4$

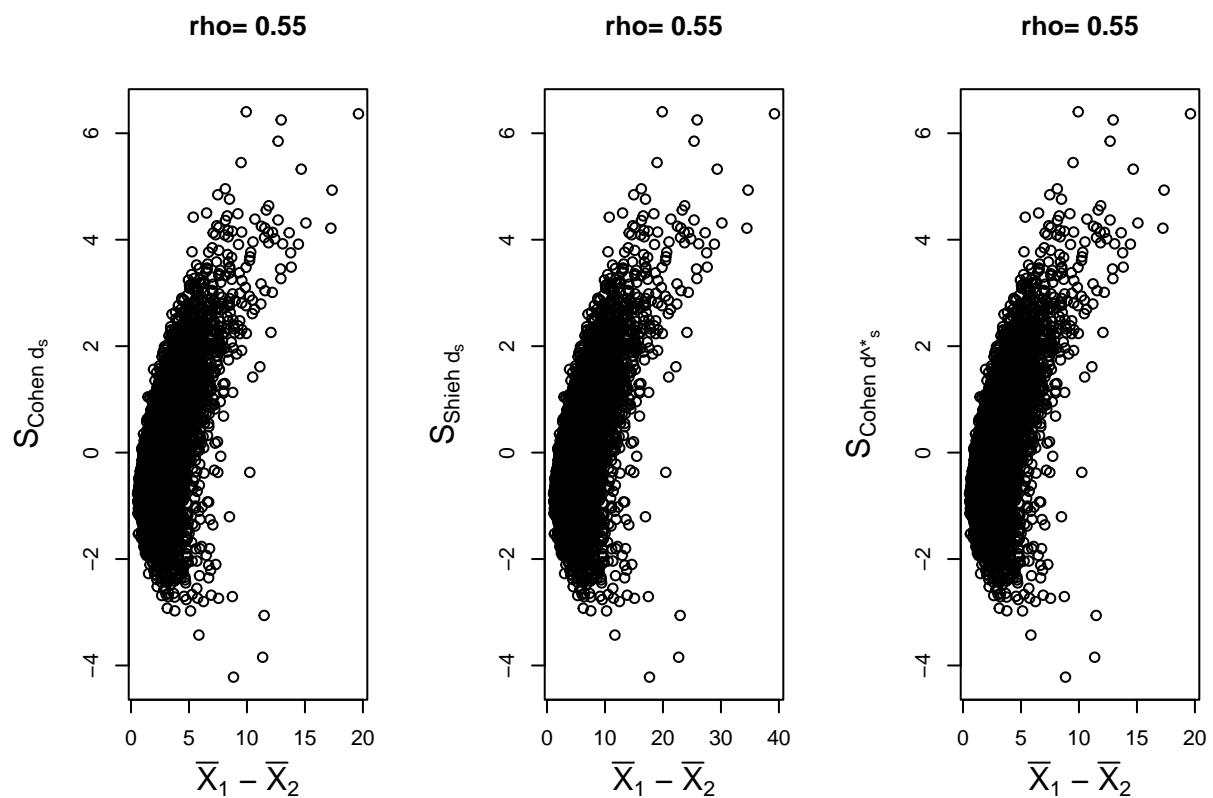


Figure 17. $S_{Cohen's\,d_s}$, $S_{Shieh's\,d_s}$ and $S_{Cohen's\,d_s^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $S_1=2$ and $S_2=4$