

- 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**

- 6 The d -family effect sizes are commonly used for mean differences between groups or
 7 conditions. In all generality, the population effect size is defined as

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

where μ_1 and μ_2 respectively stand for the population means of population 1 and 2, and σ represents the “variability” of the combined observations from both population. There exist different estimators of the effect size measure in expression 1. For all, the mean difference is estimated by the difference $\bar{X}_1 - \bar{X}_2$ of both sample means. When the equality of variances assumption is met, σ is estimated by pooling both sample standard deviations (S_1 and S_2):

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

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

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

$$S_{Glass'\ d} = S_c$$

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

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

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

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

- 11 As we previously mentioned, the use of these formulas requires to meet the
 12 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).

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

How correlations between the mean difference ($\bar{X}_1 - \bar{X}_2$) and standardizers 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.

“All other parameters being equal” is mentioned because it is always possible that other factors in action have an opposite effect on bias and variance in order that increasing the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ does not necessarily reduce the bias and the variance. For example, when population variances are equal across groups and sample sizes are unequal, we will see below that the lower n_j , the larger the magnitude of

38 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
 39 positive, the smaller the sample size, the larger the positive correlation. At the same time,
 40 we know that increasing the sample size decreases the bias. This is a nice example of
 41 situations where two factors might have an opposite action on bias.

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

43 **When equal population variances are estimated based on equal sample sizes
 44 (condition a)**

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

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

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

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

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

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

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

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

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

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

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

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

- When computing $S_{\text{Cohen's } d}$, more weight is given to the SD of the largest sample, which by compensation effect brings the correlation very close to 0.

87 The correlation between $\bar{X}_1 - \bar{X}_2$ and respectively S_1, S_2 , the standardizer of Cohen's
88 d^* , the standardizer of Shieh's d and the standardizer of Cohen's d are summarized in
89 Table 1.

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

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

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

Table 1

Correlation between standardizers (S_1 , S_2 , $S_{Cohen's\ d}$, $S_{Shieh's\ d}$ and $S_{Cohen's\ d^}$) 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
	$S_{Cohen's\ d}$: null	$S_{Cohen's\ d}$: null
	$S_{Shieh's\ d}$: null	$S_{Shieh's\ d}$: null
	$S_{Cohen's\ d^*}$: null	$S_{Cohen's\ d^*}$: null
When $n_1 > n_2$		
	S_1 : positive	S_1 : negative
	S_2 : negative	S_2 : positive
	$S_{Cohen's\ d}$: null	$S_{Cohen's\ d}$: null
	$S_{Shieh's\ d}$: negative	$S_{Shieh's\ d}$: positive
	$S_{Cohen's\ d^*}$: positive (but very small)	$S_{Cohen's\ d^*}$: negative (but very small)
When $n_1 < n_2$		
	S_1 : positive	S_1 : negative
	S_2 : negative	S_2 : positive
	$S_{Cohen's\ d}$: negative (but very small)	$S_{Cohen's\ d}$: positive (but very small)
	$S_{Shieh's\ d}$: positive	$S_{Shieh's\ d}$: negative
	$S_{Cohen's\ d^*}$: positive	$S_{Cohen's\ d^*}$: 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>

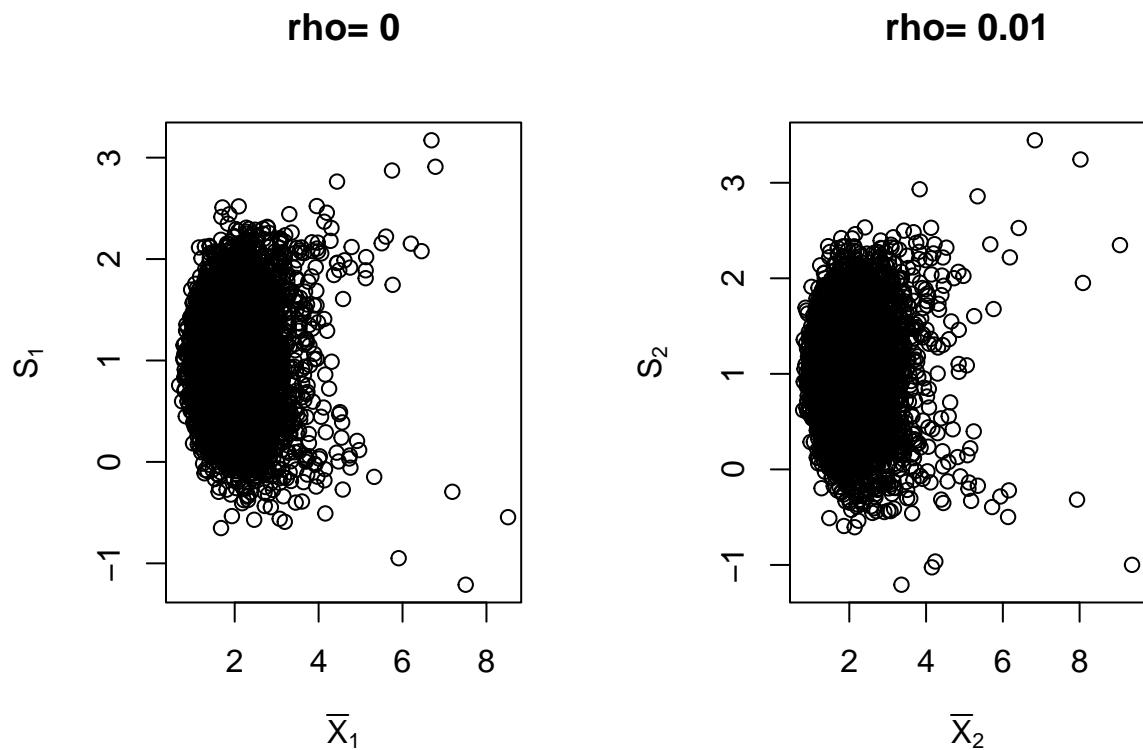


Figure 1. S_j as a function of \bar{X}_j ($j=1,2$), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

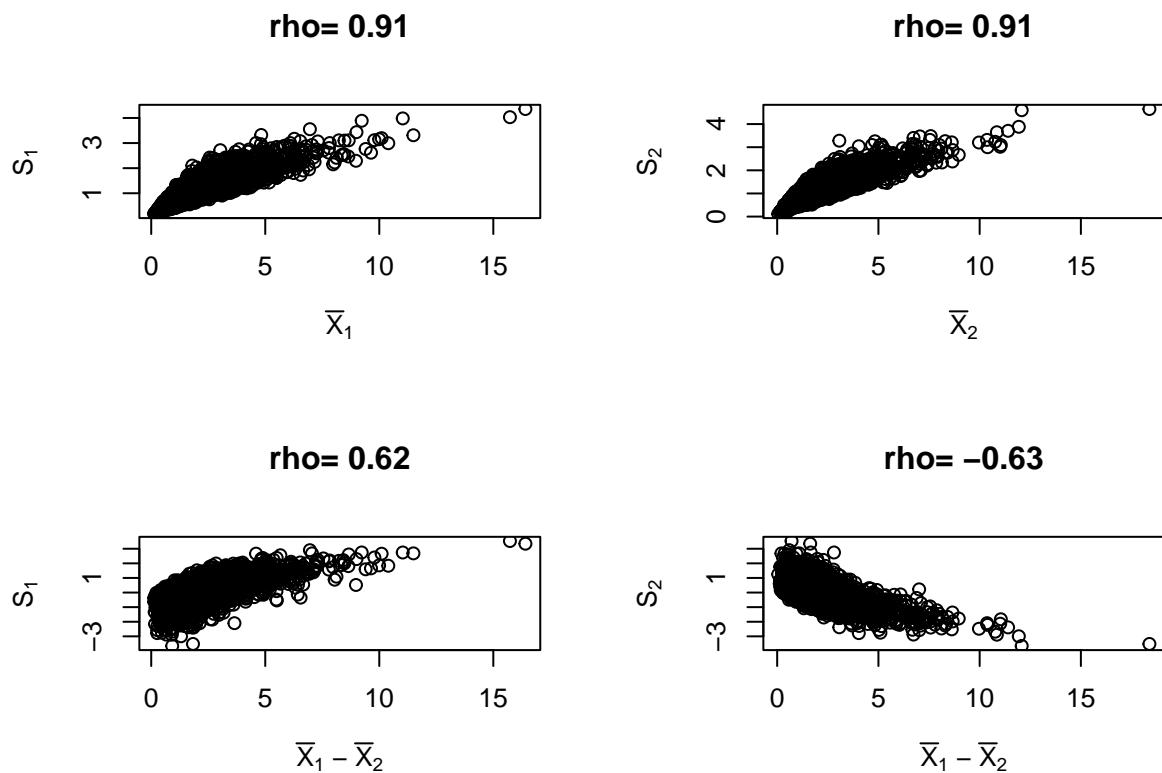


Figure 2. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

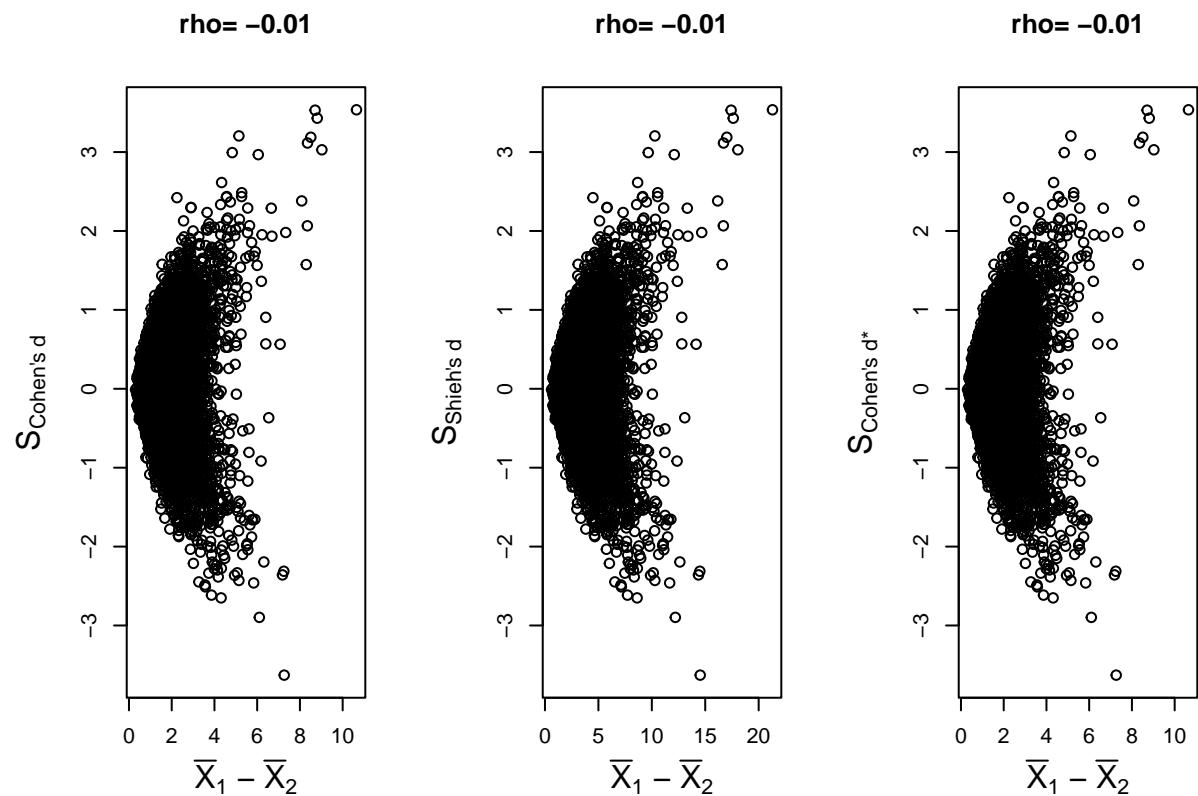


Figure 3. $S_{Cohen's\ d}$, $S_{Shieh's\ d}$ and $S_{Cohen's\ d^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

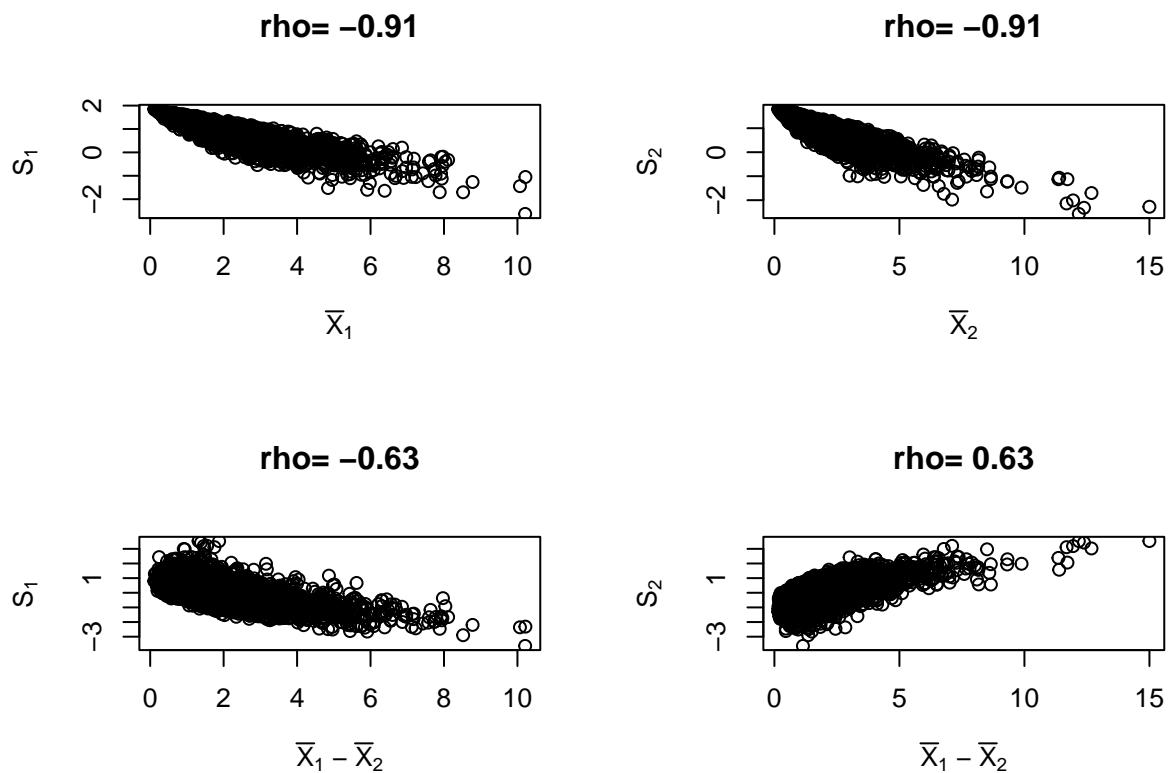


Figure 4. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

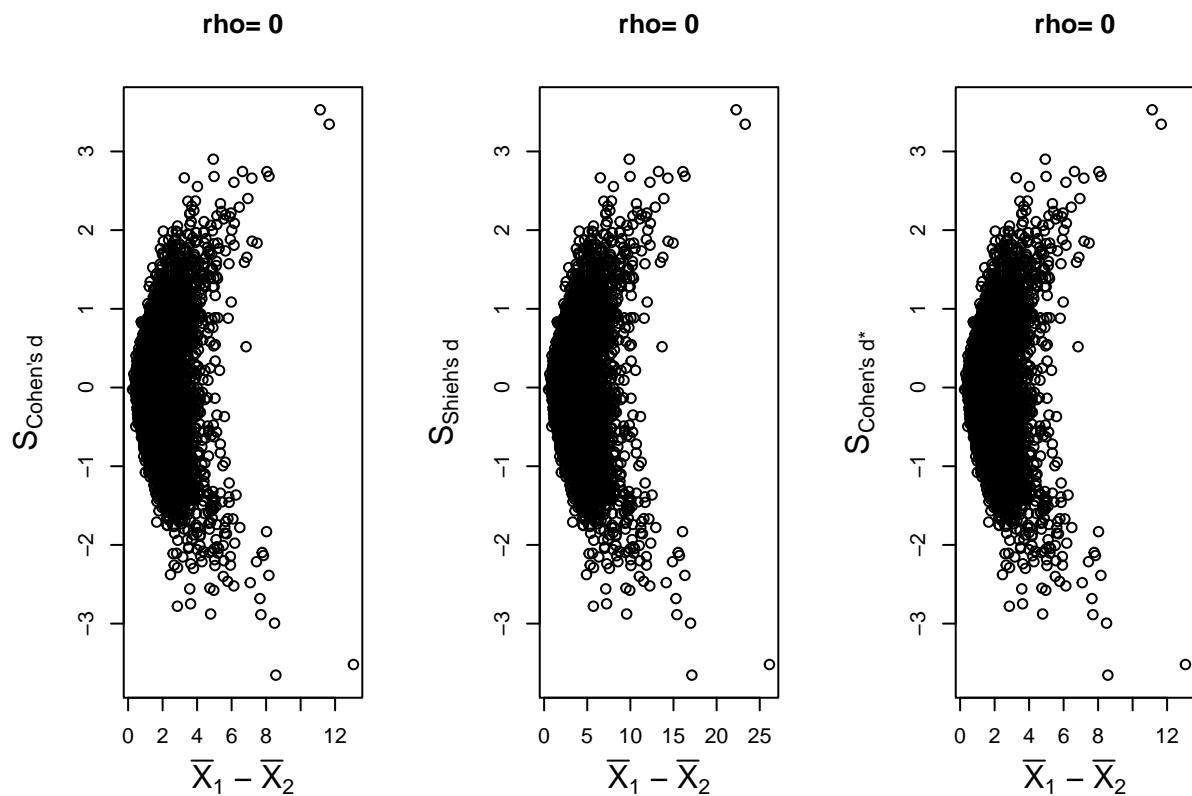


Figure 5. $S_{Cohen's\ d}$, $S_{Shieh's\ d}$ and $S_{Cohen's\ d^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

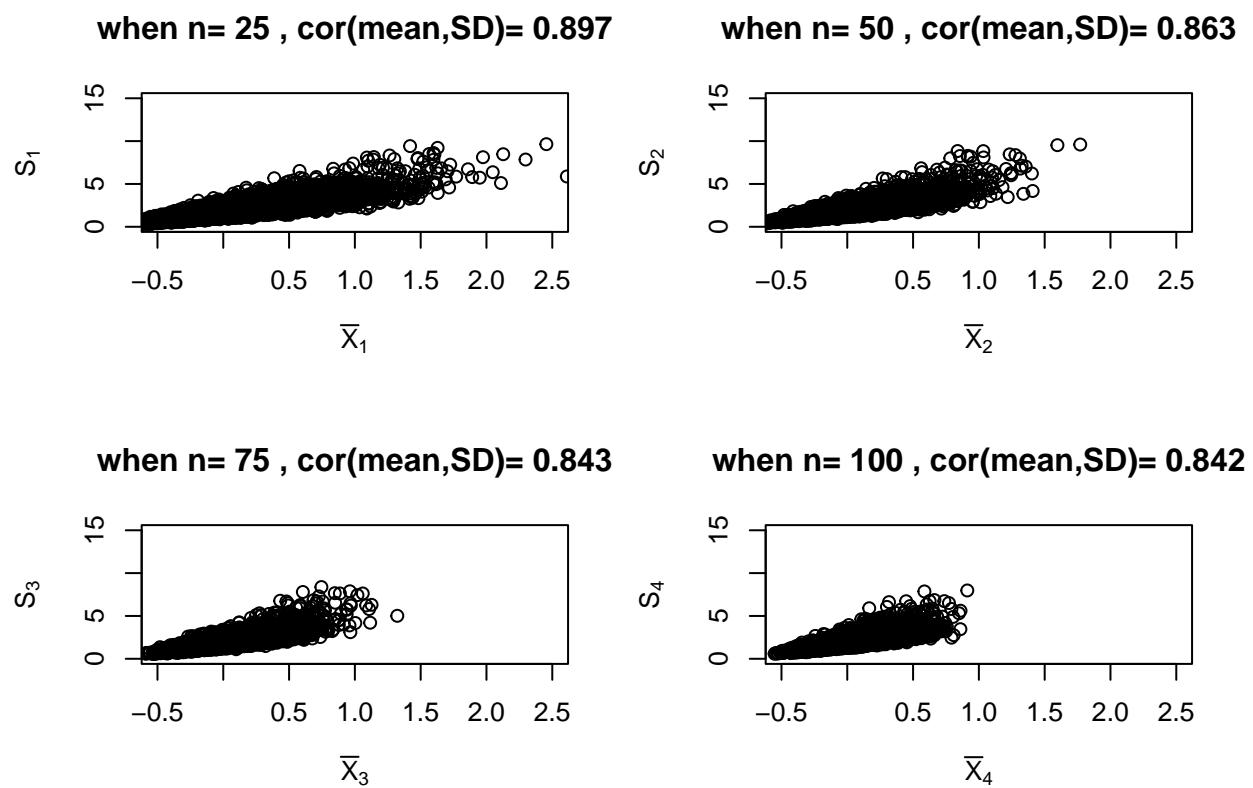


Figure 6. Correlation between S_j and \bar{X}_j when $n = 25, 50, 75$ or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

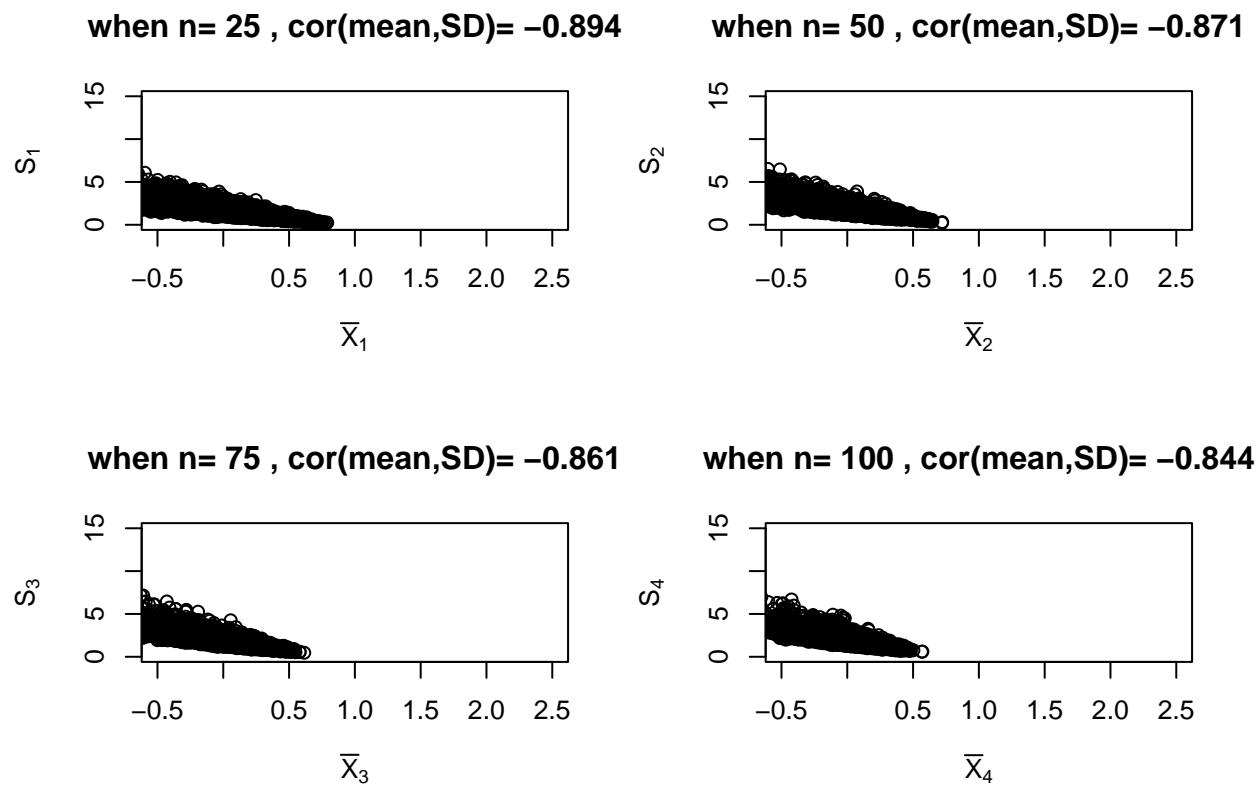


Figure 7. Correlation between S_j and \bar{X}_j when $n = 25, 50, 75$ or 100 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

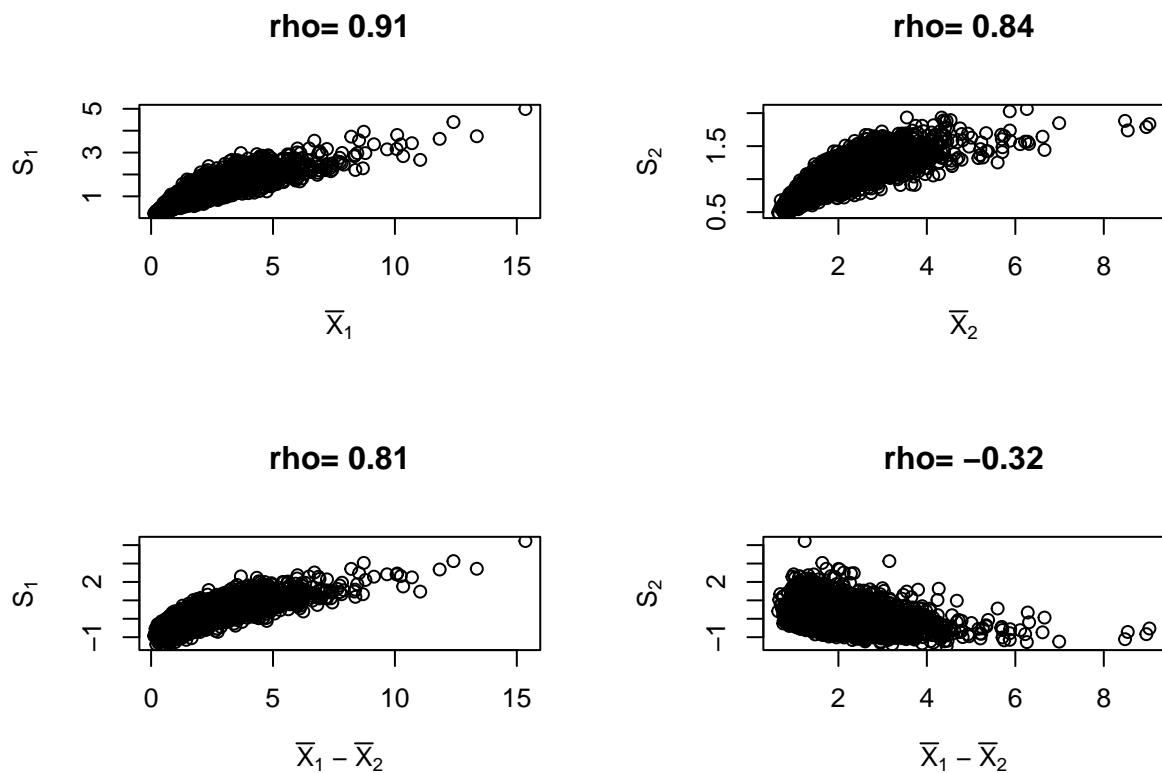


Figure 8. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $n_1=20$ and $n_2=100$

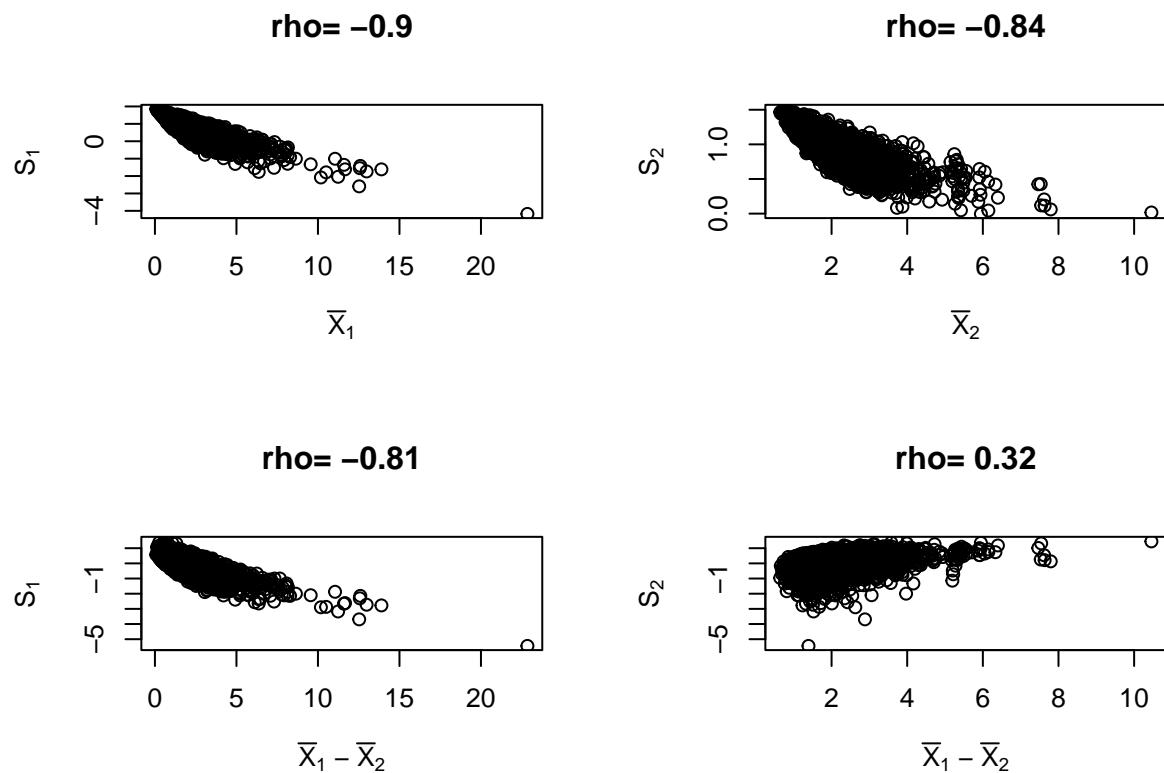


Figure 9. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $n_1=20$ and $n_2=100$

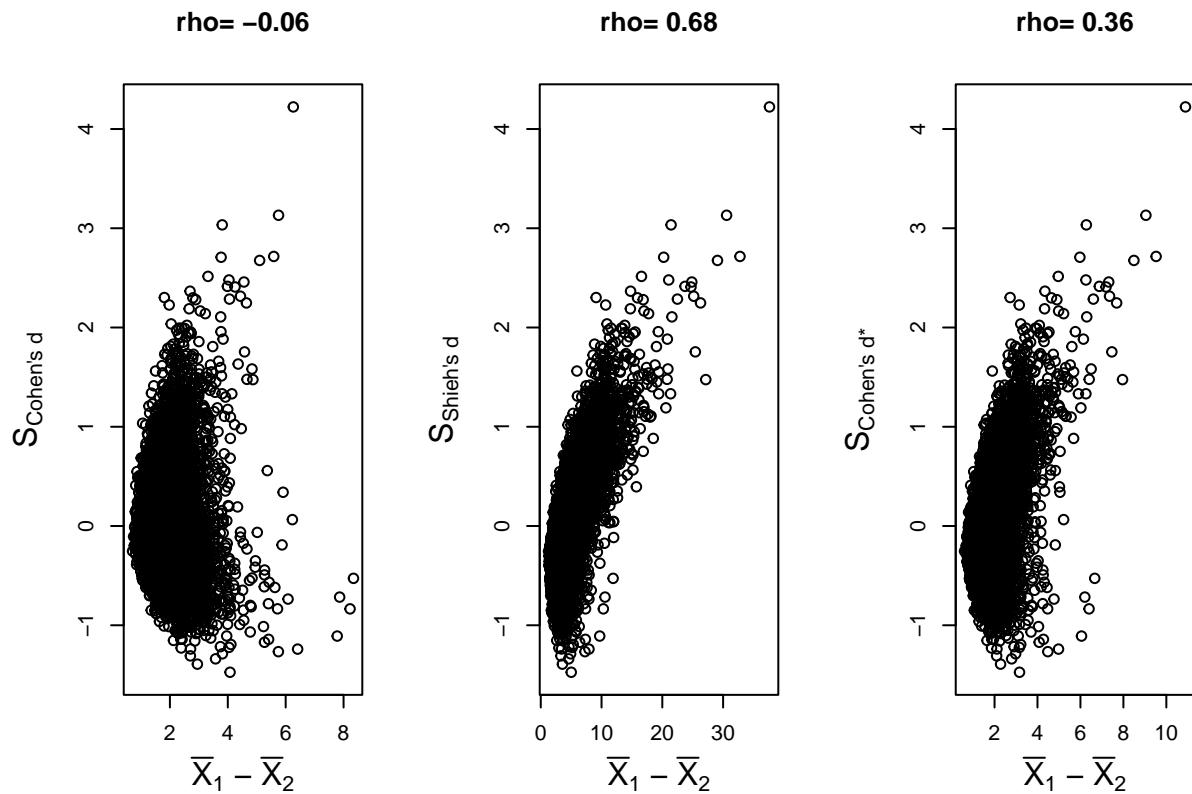


Figure 10. $S_{Cohen's\ d}$, $S_{Shieh's\ d}$ and $S_{Cohen's\ d^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$, with $n_1=20$ and $n_2=100$)

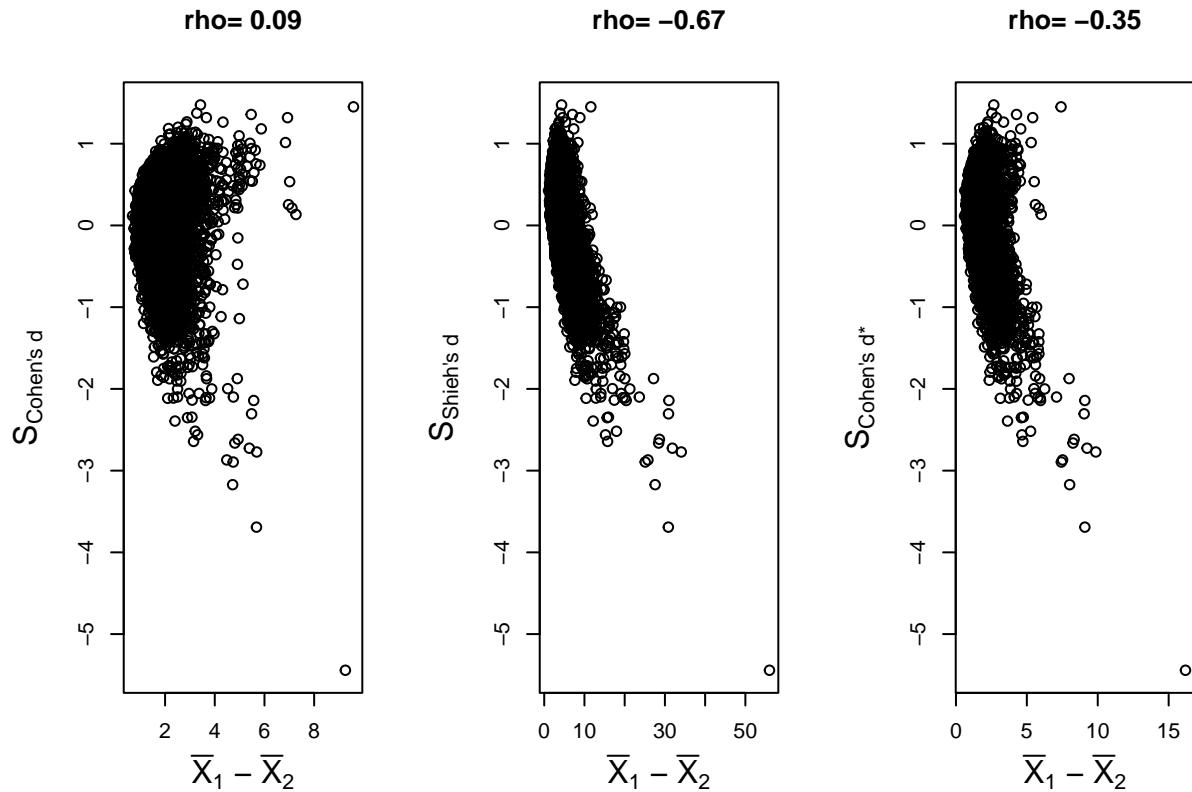


Figure 11. $S_{Cohen's\ d}$, $S_{Shieh's\ d}$ and $S_{Cohen's\ d^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $n_1=20$ and $n_2=100$

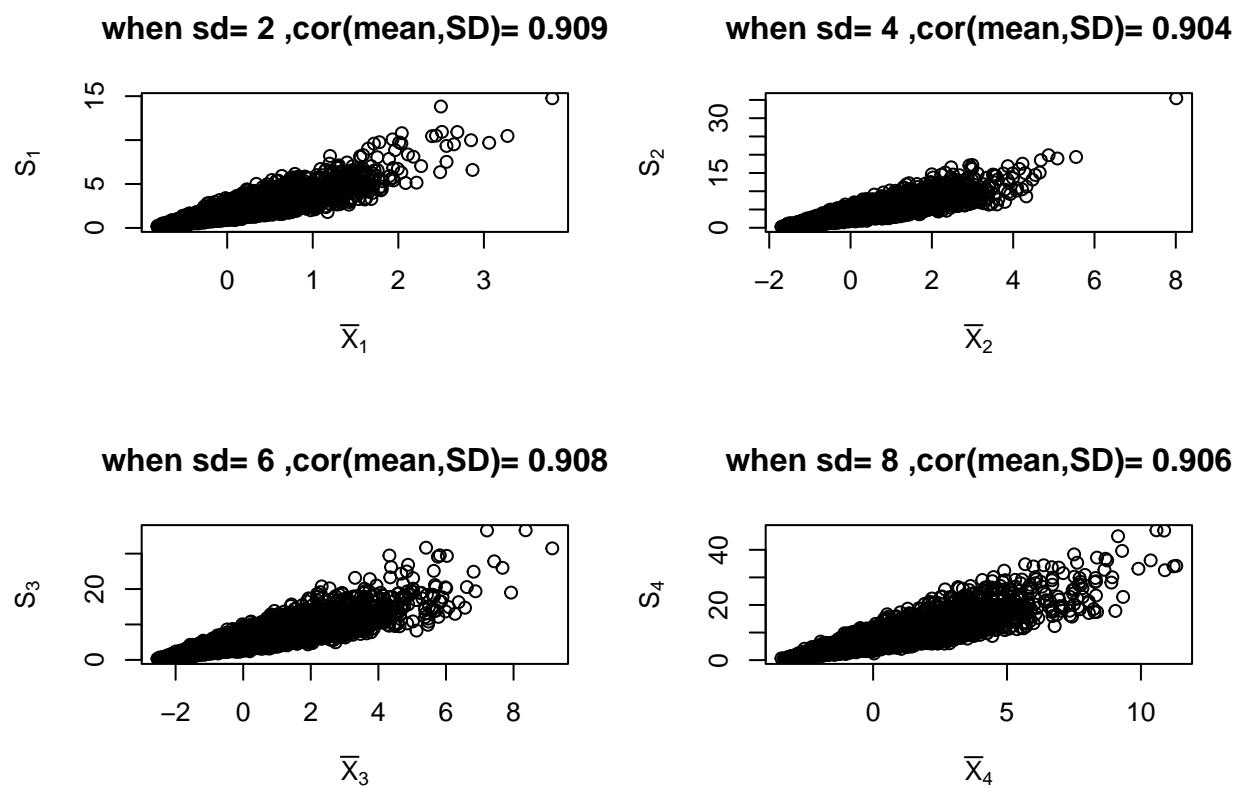


Figure 12. Correlation between S_j and \bar{X}_j when $SD = 2, 4, 6$ or 8 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

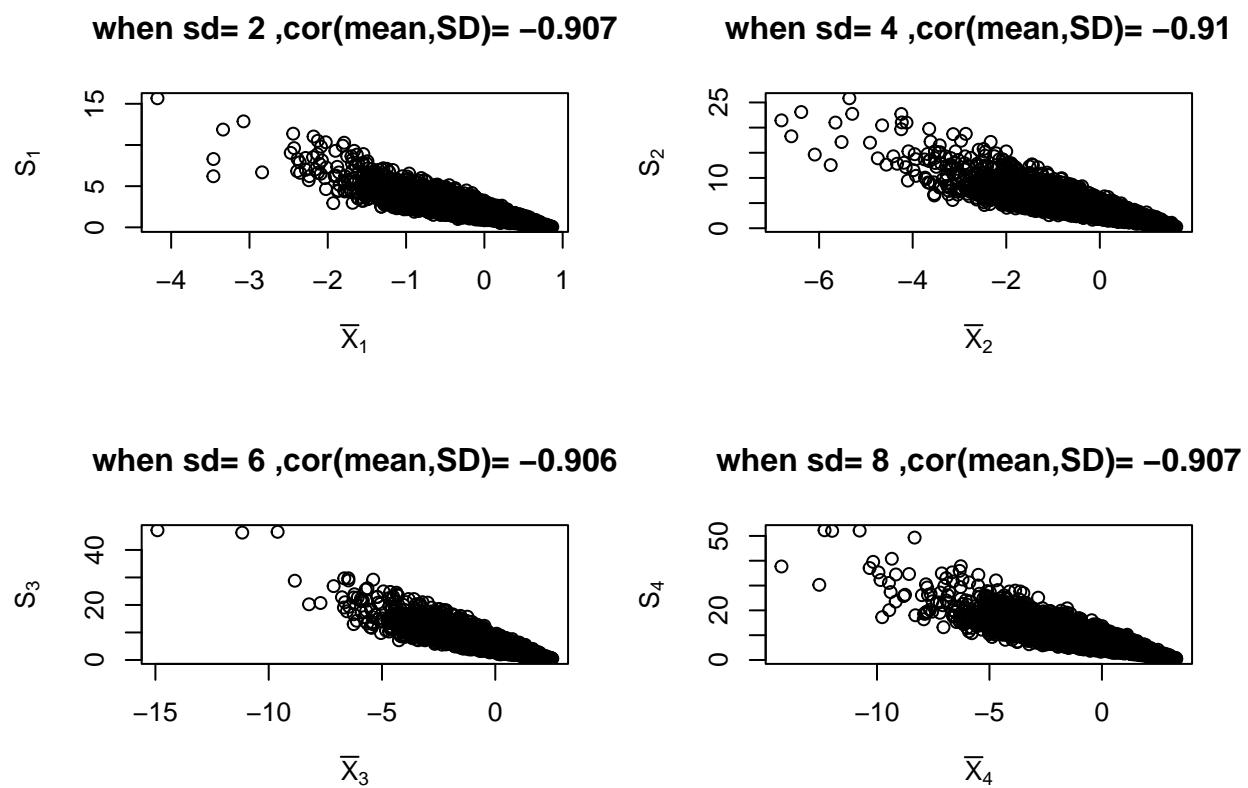


Figure 13. Correlation between S_j and \bar{X}_j when $SD = 2, 4, 6$ or 8 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

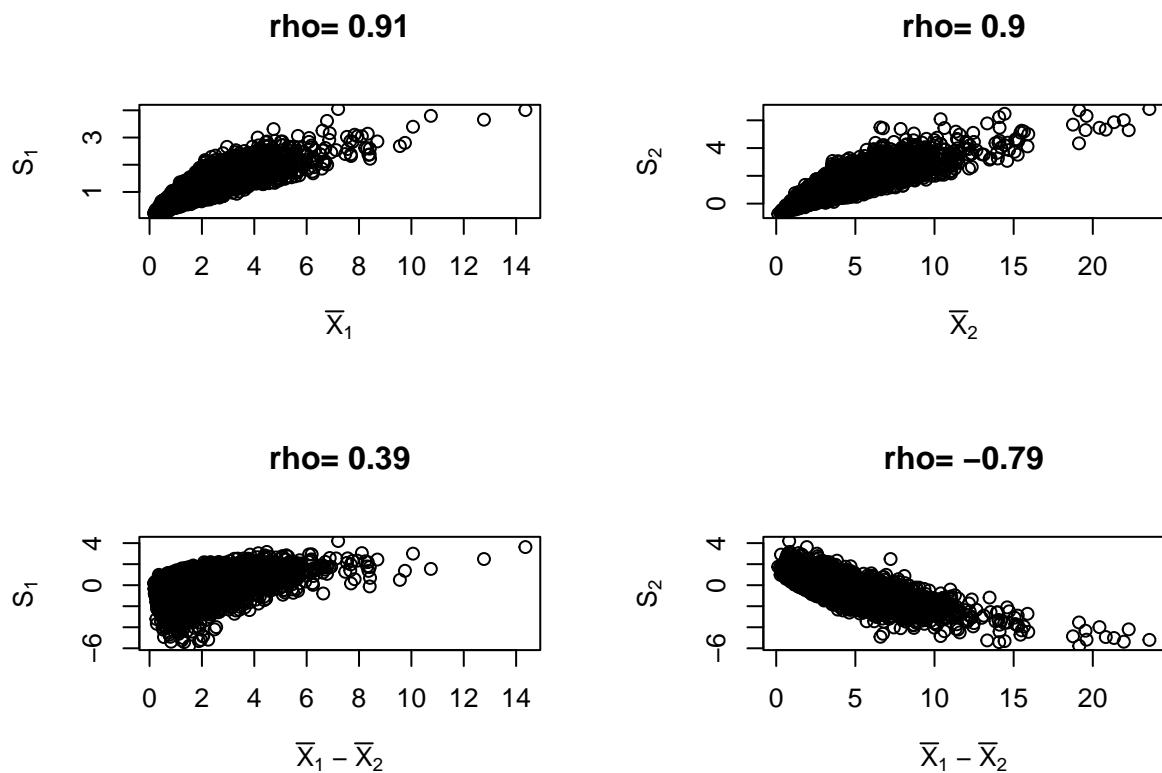


Figure 14. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $S_1=2$ and $S_2=4$

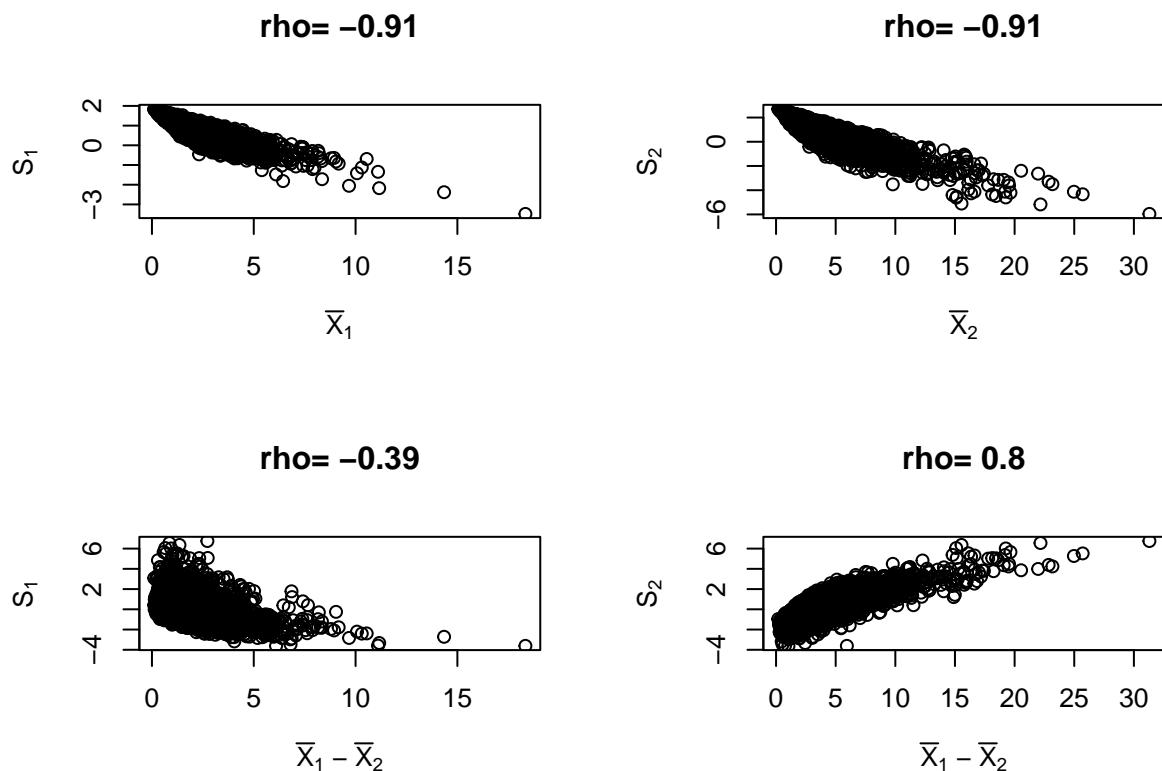


Figure 15. S_j ($j=1,2$) as a function of \bar{X}_j (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $S_1=2$ and $S_2=4$

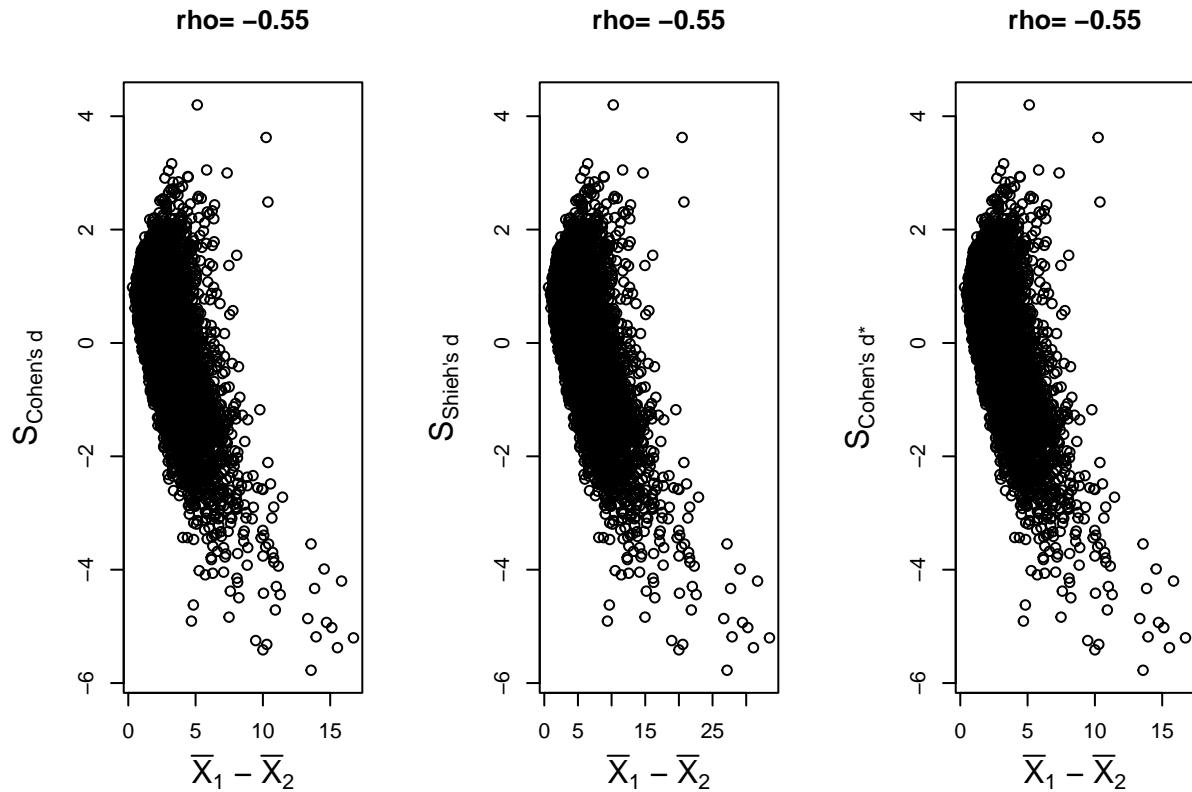


Figure 16. $S_{\text{Cohen's } d}$, $S_{\text{Shieh's } d}$ and $S_{\text{Cohen's } d^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with $S_1=2$ and $S_2=4$

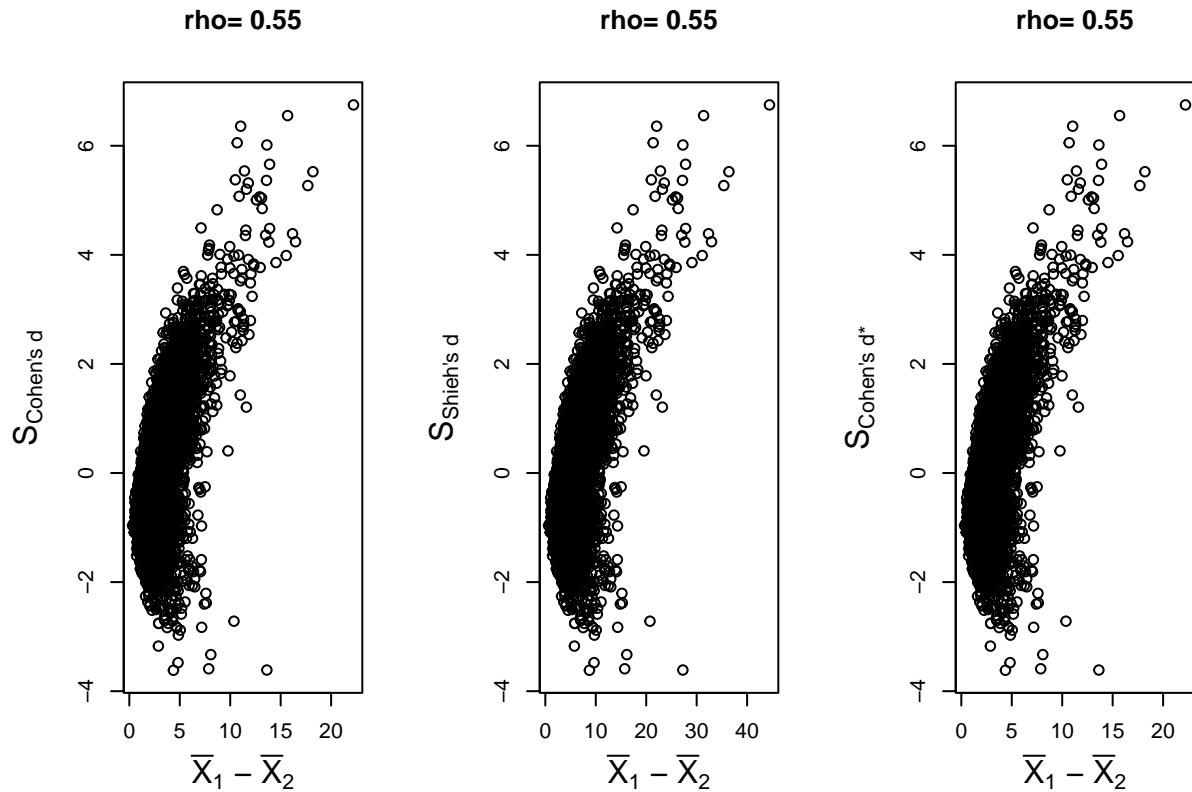


Figure 17. $S_{Cohen's\ d}$, $S_{Shieh's\ d}$ and $S_{Cohen's\ d^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with $S_1=2$ and $S_2=4$