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

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

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):

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

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

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

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

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

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

- 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).

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 ($\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2$) 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 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 positive, the smaller the sample size, the larger the positive correlation. At the same time, we know that increasing the sample size decreases the bias. This is a nice example of situations where two factors might have an opposite action on bias.

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

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

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

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

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

On the other hand, when distributions are left-skewed, there is a **negative** correlation between $\bar{X}_j$ and S_j (see the two top plots in Figure 4), resulting in a *negative* correlation between S_1 and $\bar{X}_1 - \bar{X}_2$ and in a *positive* correlation between S_2 and $\bar{X}_1 - \bar{X}_2$ (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 $cor(S_1, \bar{X}_1) \neq cor(S_2, \bar{X}_2)$, 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_s and Cohen's d_s^* are **correlated** with $\bar{X}_1 - \bar{X}_2$ (see Figures 10 and 11):

- When computing $S_{Cohen's\ d_s^*}$, 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 $S_{Cohen's\ d_s^*}$ and $\bar{X}_1 - \bar{X}_2$ is the same as the size of the correlation between $\bar{X}_1 - \bar{X}_2$ and the SD of the smallest sample;

- When computing $S_{Shieh's\ d_s}$, more weight is given to the SD of the smallest sample, it is therefore not really surprising to observe that the correlation between $S_{Shieh's\ d_s}$ and $\bar{X}_1 - \bar{X}_2$ is closer of the correlation between the SD of the smallest group and $\bar{X}_1 - \bar{X}_2$ (i.e. $|cor(S_{Shieh's\ d_s}, \bar{X}_1 - \bar{X}_2)| > |cor(S_{Cohen's\ d_s^*}, \bar{X}_1 - \bar{X}_2)|$);

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

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

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

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

This also explains that when computing a standardizer that takes both S_1 and S_2 into account, it results in a standardizer that is correlated with $\bar{X}_1 - \bar{X}_2$ (see Figures 16 and 17). The correlation between the mean difference $(\bar{X}_1 - \bar{X}_2)$ and respectively the standardizer of Shieh's d_s, Cohen's d_s^* and Cohen's d_s will have the same sign as the correlation between $(\bar{X}_1 - \bar{X}_2)$ and the larger SD. Table 2 summarizes the sign of the correlation between $\bar{X}_1 - \bar{X}_2$ and respectively S_1, S_2 and the three standardizers taking 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_2: *negative* $S_{Cohen's\ d_s}$: *null* $S_{Shieh's\ d_s}$: *null* $S_{Cohen's\ d_s^*}$: *null*	S_1: *negative* S_2: *positive* $S_{Cohen's\ d_s}$: *null* $S_{Shieh's\ d_s}$: *null* $S_{Cohen's\ d_s^*}$: *null*
When $n_1 > n_2$	S_1: *positive* S_2: *negative* $S_{Cohen's\ d_s}$: *null* $S_{Shieh's\ d_s}$: *negative* $S_{Cohen's\ d_s^*}$: *positive (but very small)*	S_1: *negative* S_2: *positive* $S_{Cohen's\ d_s}$: *null* $S_{Shieh's\ d_s}$: *positive* $S_{Cohen's\ d_s^*}$: *negative (but very small)*
When $n_1 < n_2$	S_1: *positive* S_2: *negative* $S_{Cohen's\ d_s}$: *negative (but very small)* $S_{Shieh's\ d_s}$: *positive*	S_1: *negative* S_2: *positive* $S_{Cohen's\ d_s}$: *positive (but very small)* $S_{Shieh's\ d_s}$: *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	
	right-skewed	*left-skewed*
When $\sigma_1 = \sigma_2$	S_1: *positive* S_2: *negative* Others: *null*	S_1: *negative* S_2: *positive* Others: *null*
When $\sigma_1 > \sigma_2$	S_1: *positive* S_2: *negative* Others: *positive*	S_1: *negative* S_2: *positive* Others: *negative*
When $\sigma_1 < \sigma_2$	S_1: *positive* S_2: *negative* Others: *negative*	S_1: *negative* S_2: *positive* Others: *positive*

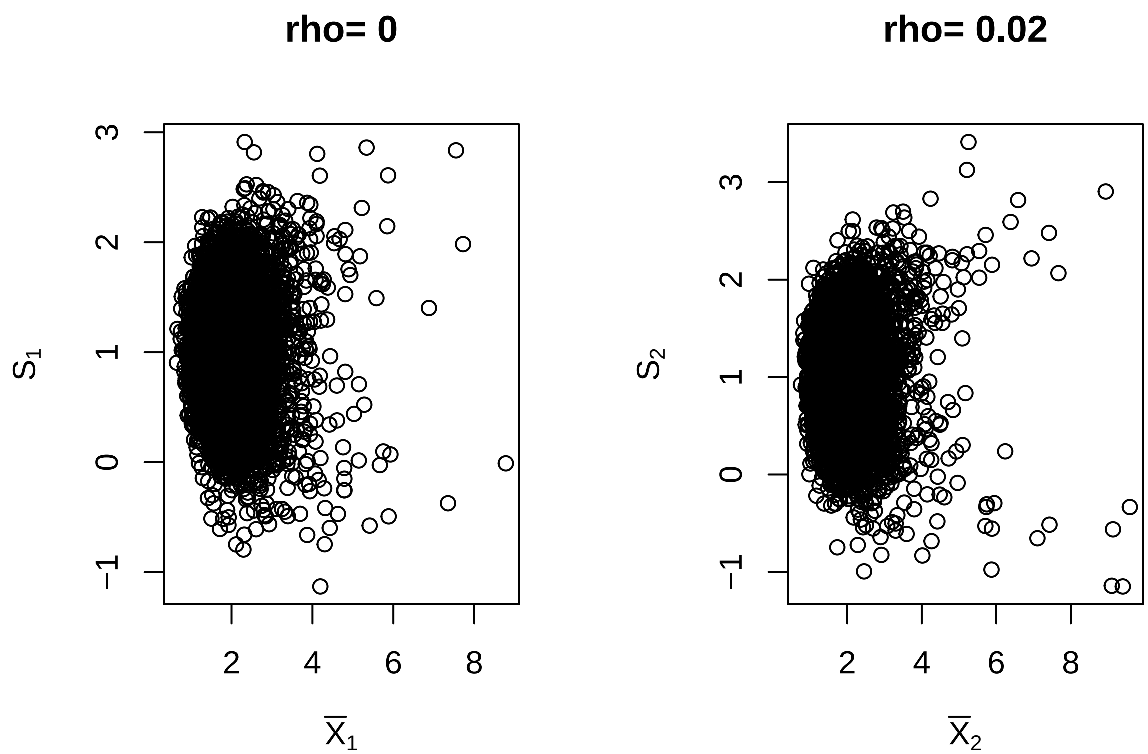

Figure 1. S_j as a function of $\bar{X}_j$ (j=1,2), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

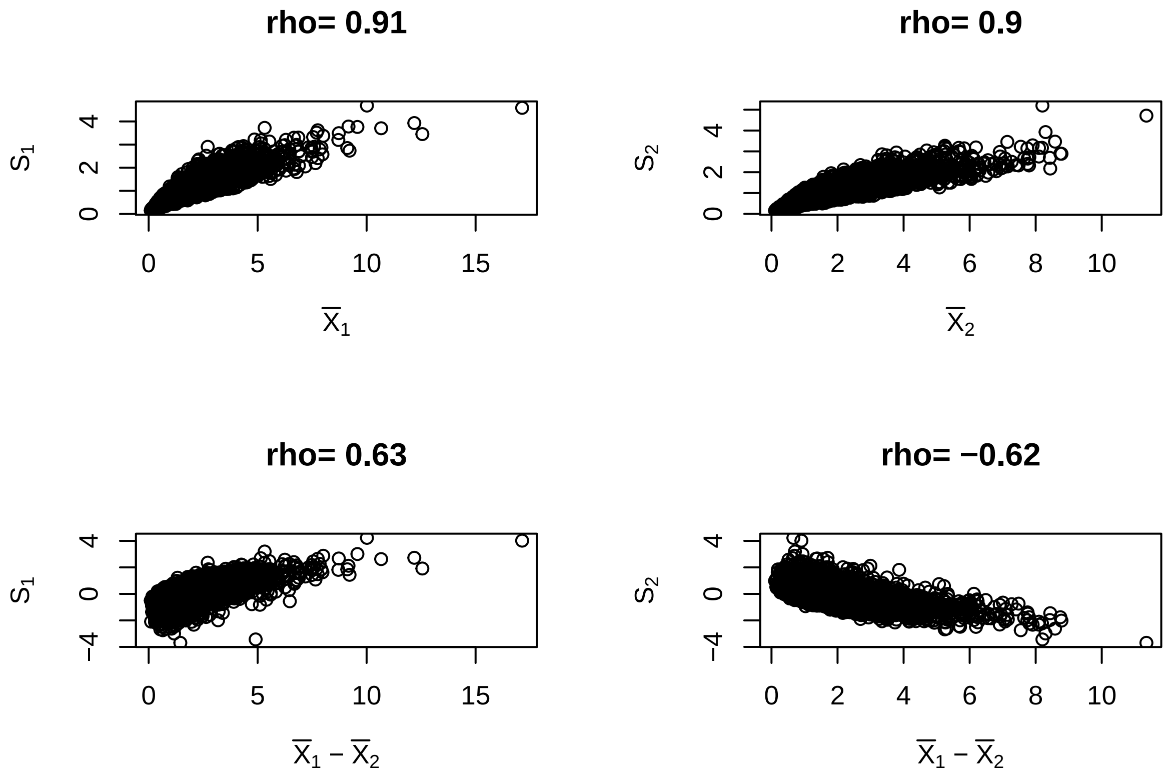

Figure 2. S_j (j=1,2) as a function of $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

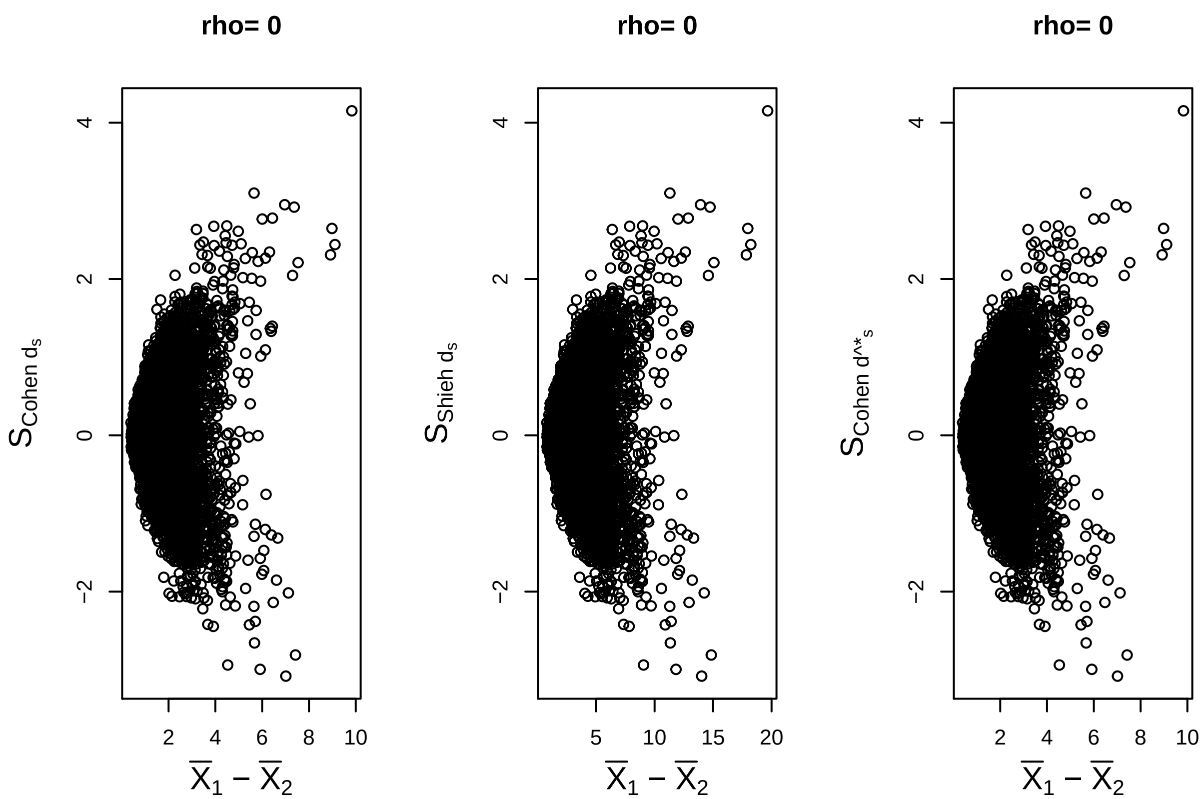

Figure 3. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d_s^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

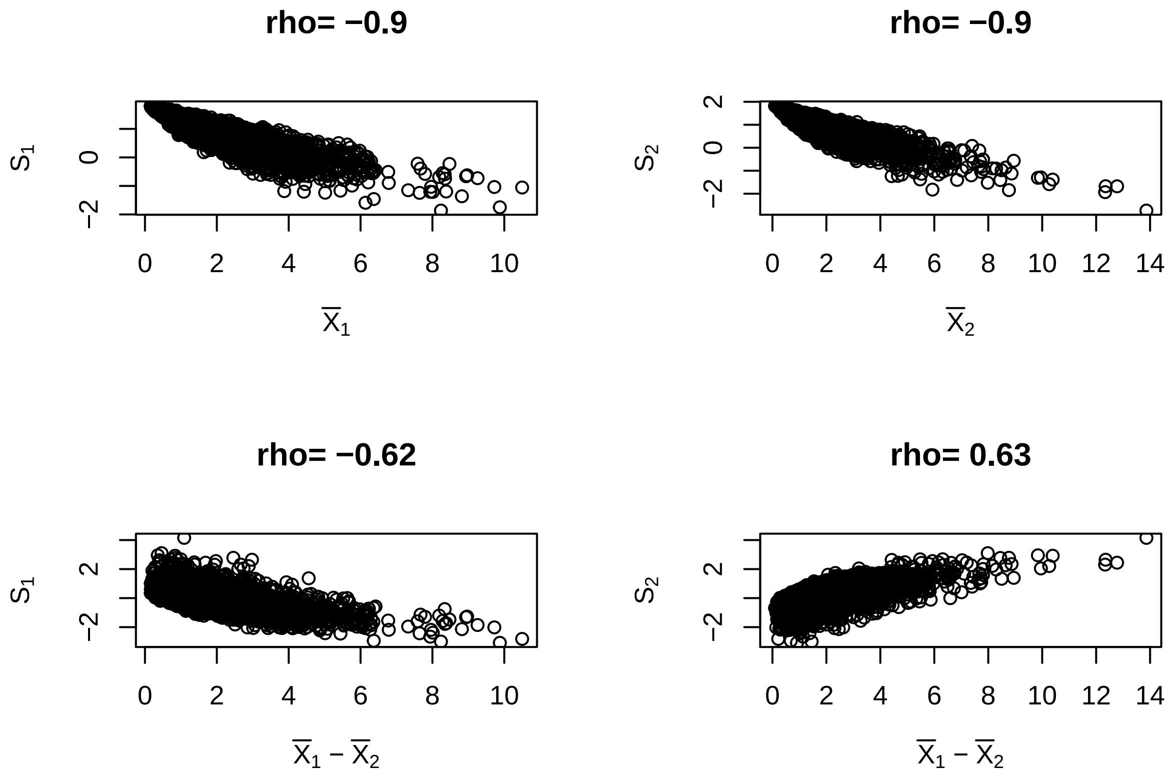

Figure 4. S_j (j=1,2) as a function of $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

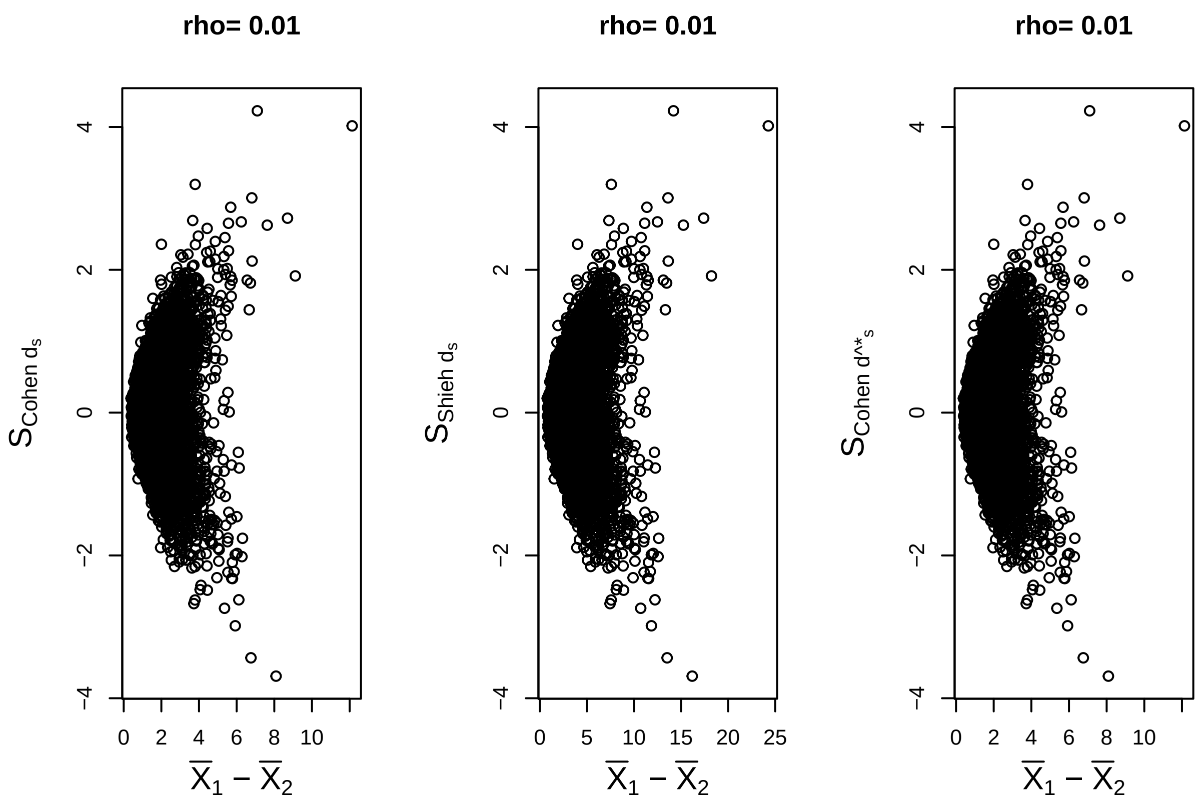

Figure 5. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d_s^*}$ as a function of the mean difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions $(\gamma_1 = -6.32)$

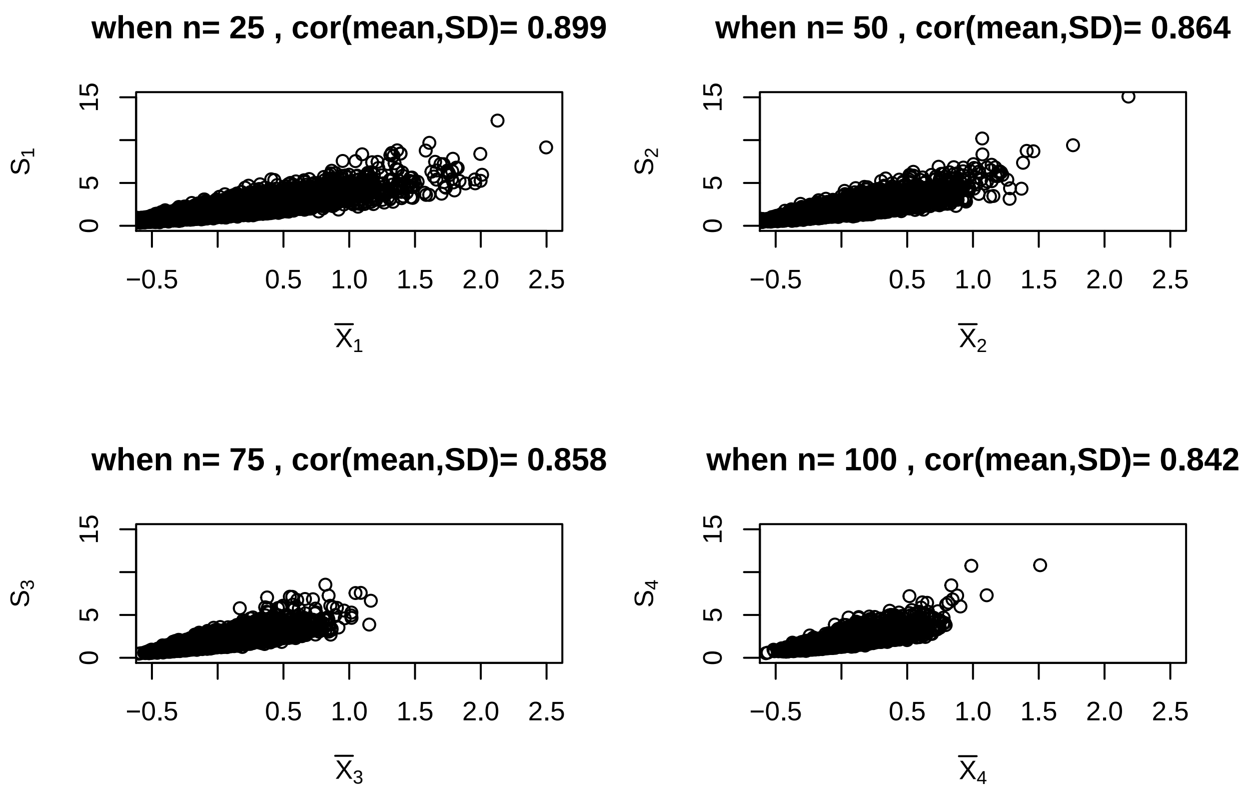

Figure 6. Correlation between S_j and $\bar{X}_j$ when n = 25, 50, 75 or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

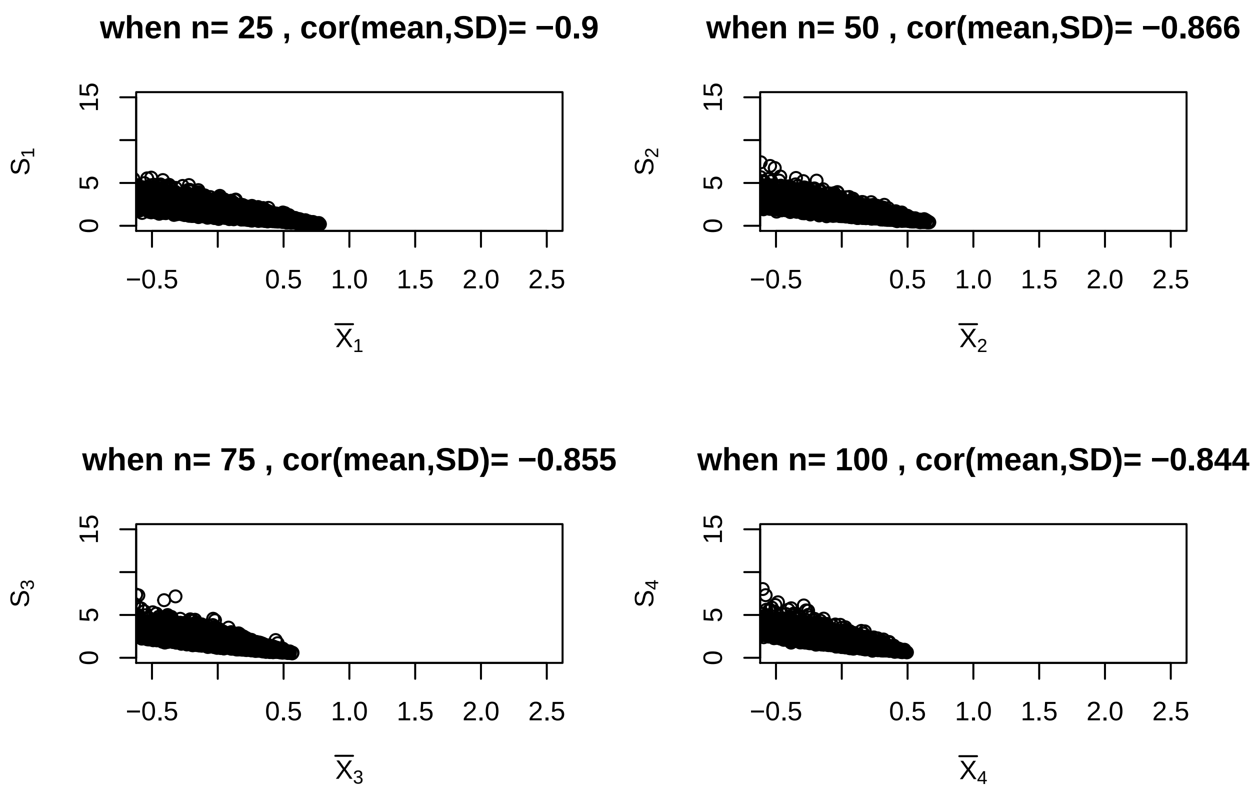

Figure 7. Correlation between S_j and $\bar{X}_j$ when $n = 25$, 50, 75 or 100 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

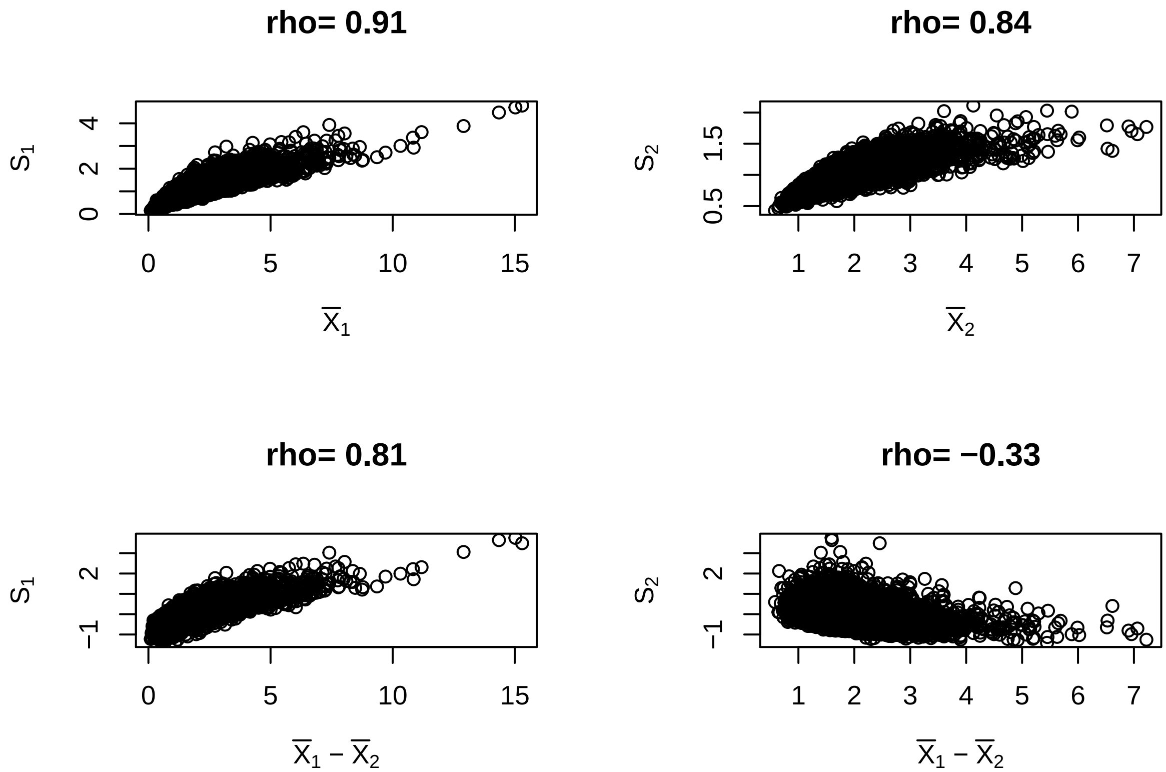

Figure 8. S_j (j=1,2) as a function of $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with n_1=20 and n_2=100

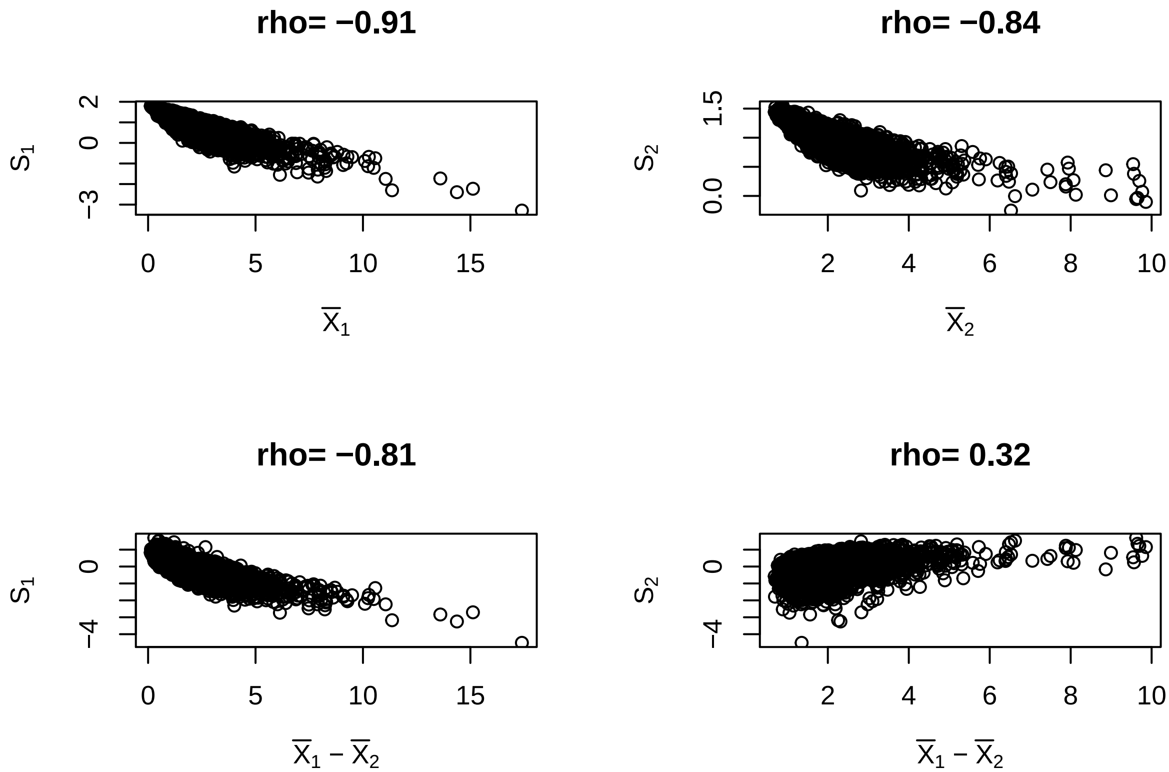

Figure 9. S_j (j=1,2) as a function of $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with n_1=20 and n_2=100

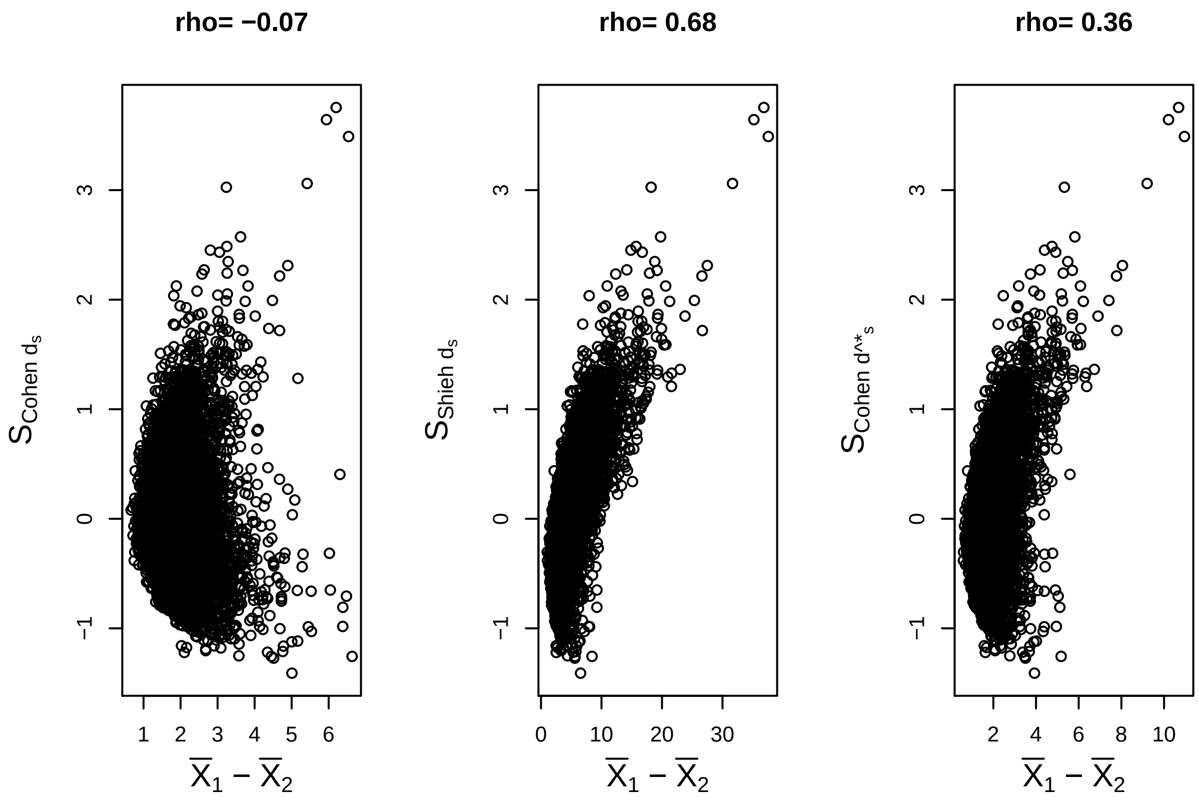

Figure 10. $S_{Cohen's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d_s^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$, with n_1=20 and n_2=100)

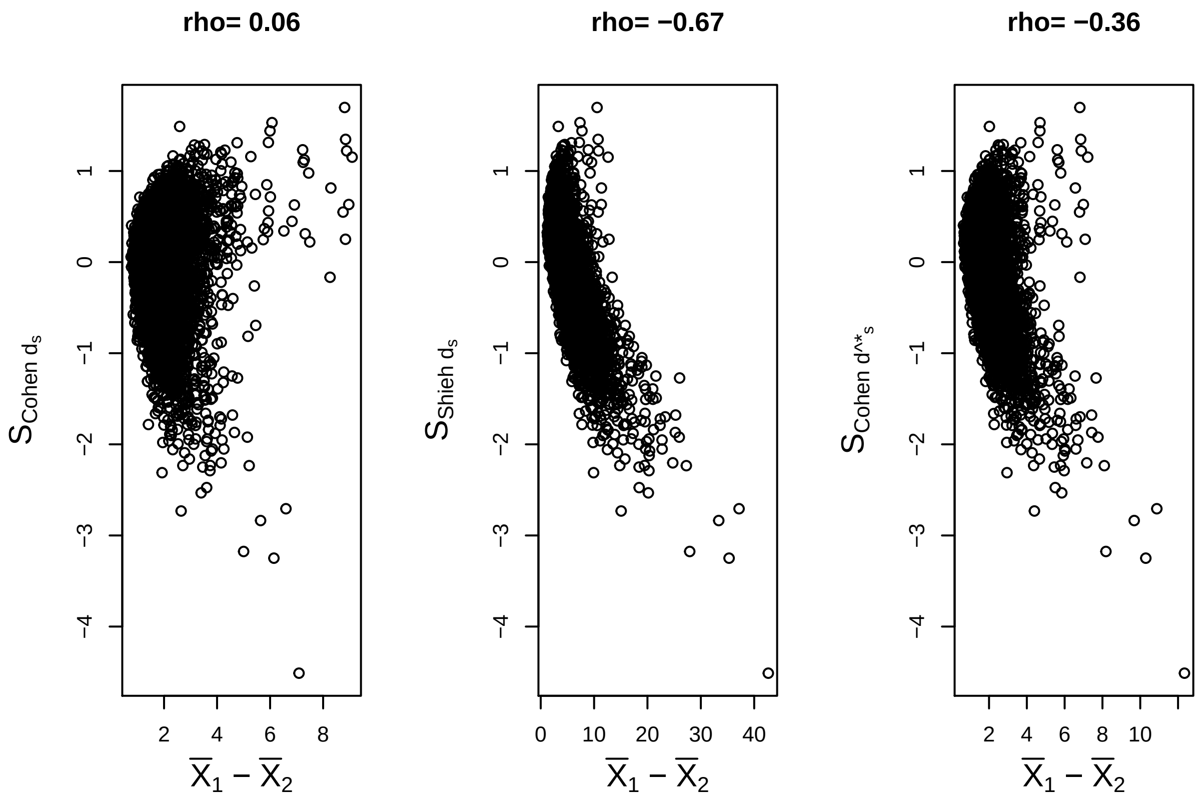

Figure 11. $S_{Cohen's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d_s^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with n_1=20 and n_2=100

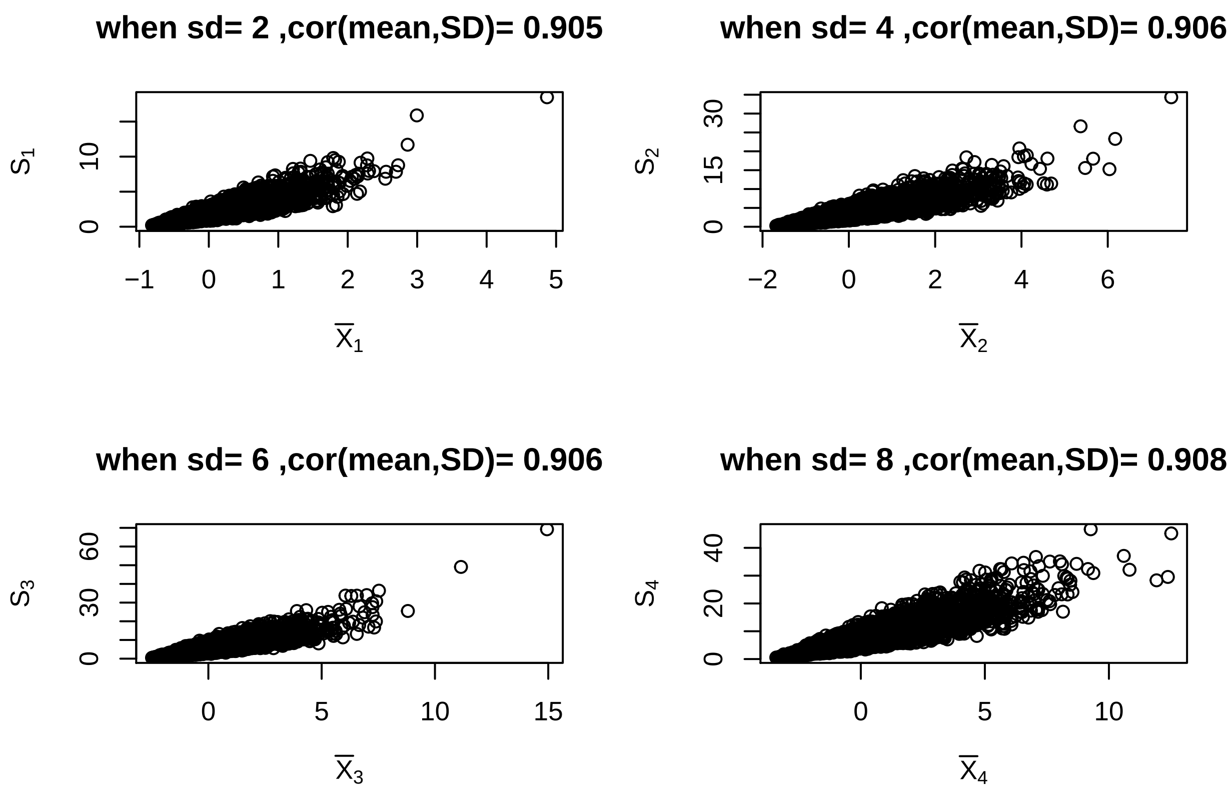

Figure 12. Correlation between S_j and $\bar{X}_j$ when $SD = 2$, 4, 6 or 8 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

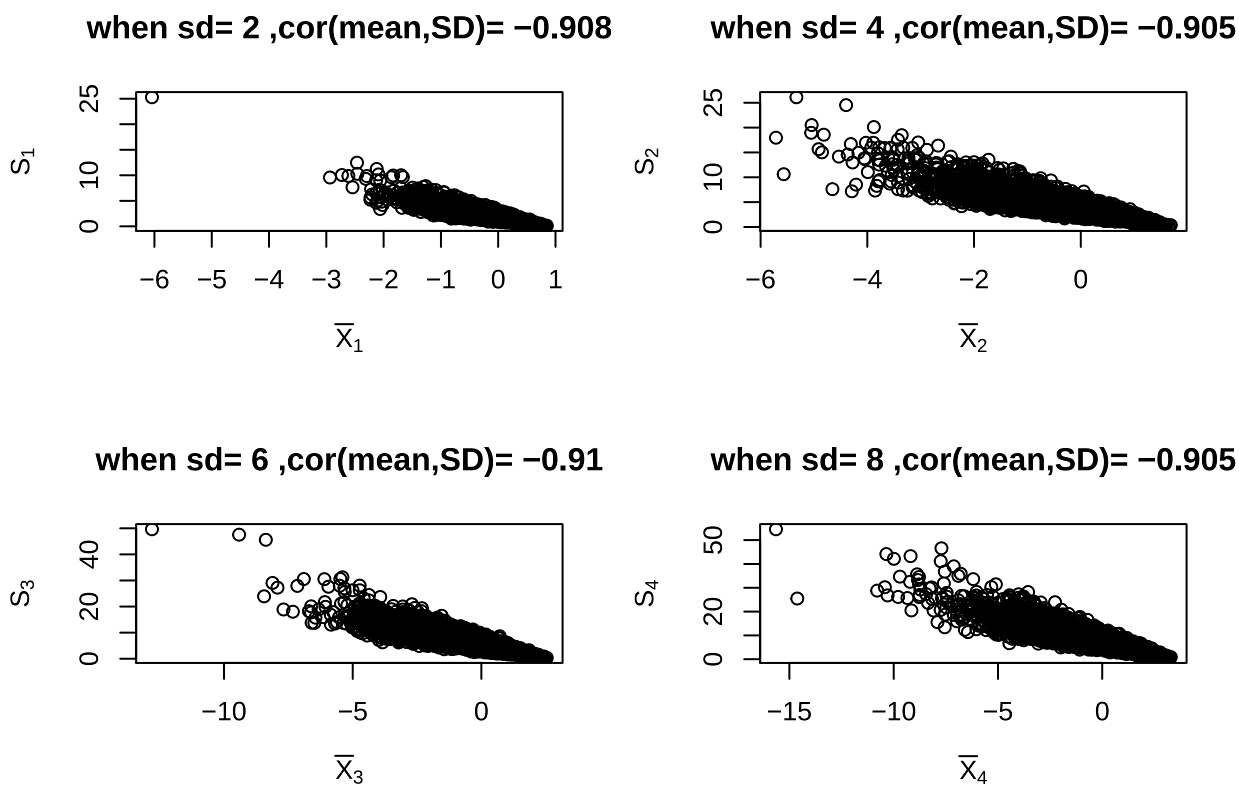

Figure 13. Correlation between S_j and $\bar{X}_j$ when $SD = 2$, 4, 6 or 8 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

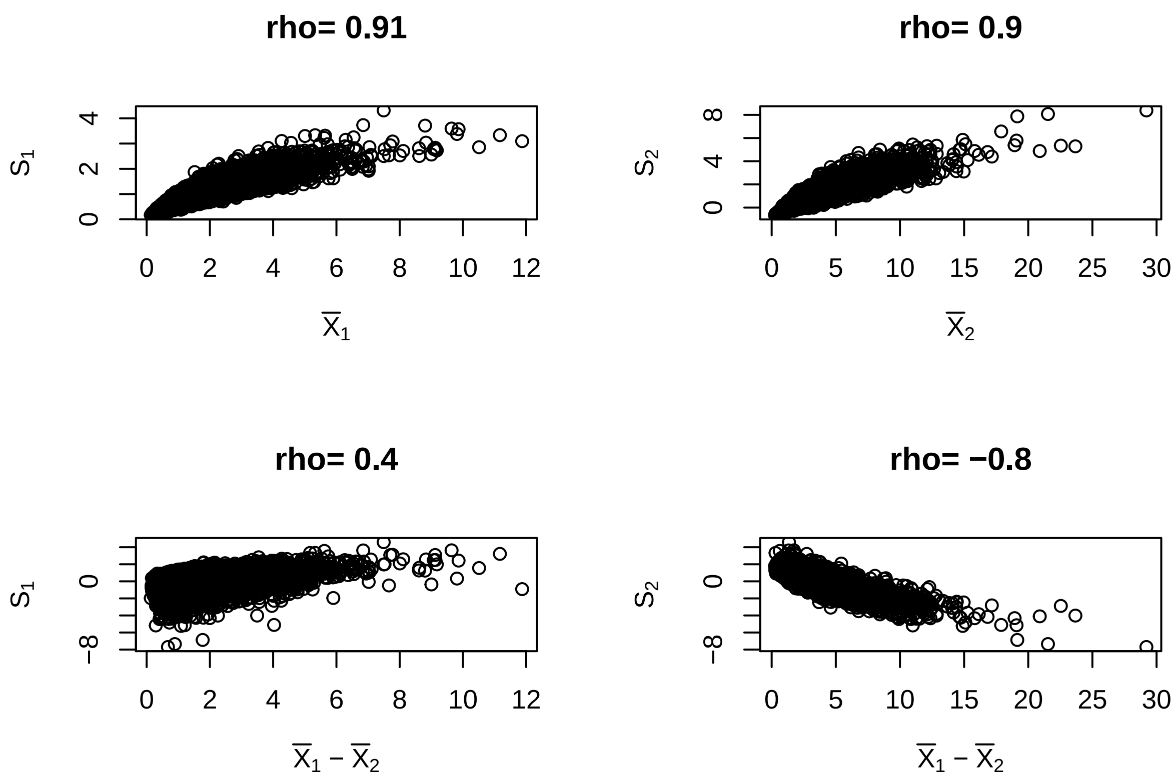

Figure 14. S_j (j=1,2) as a function of $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with S_1=2 and S_2=4

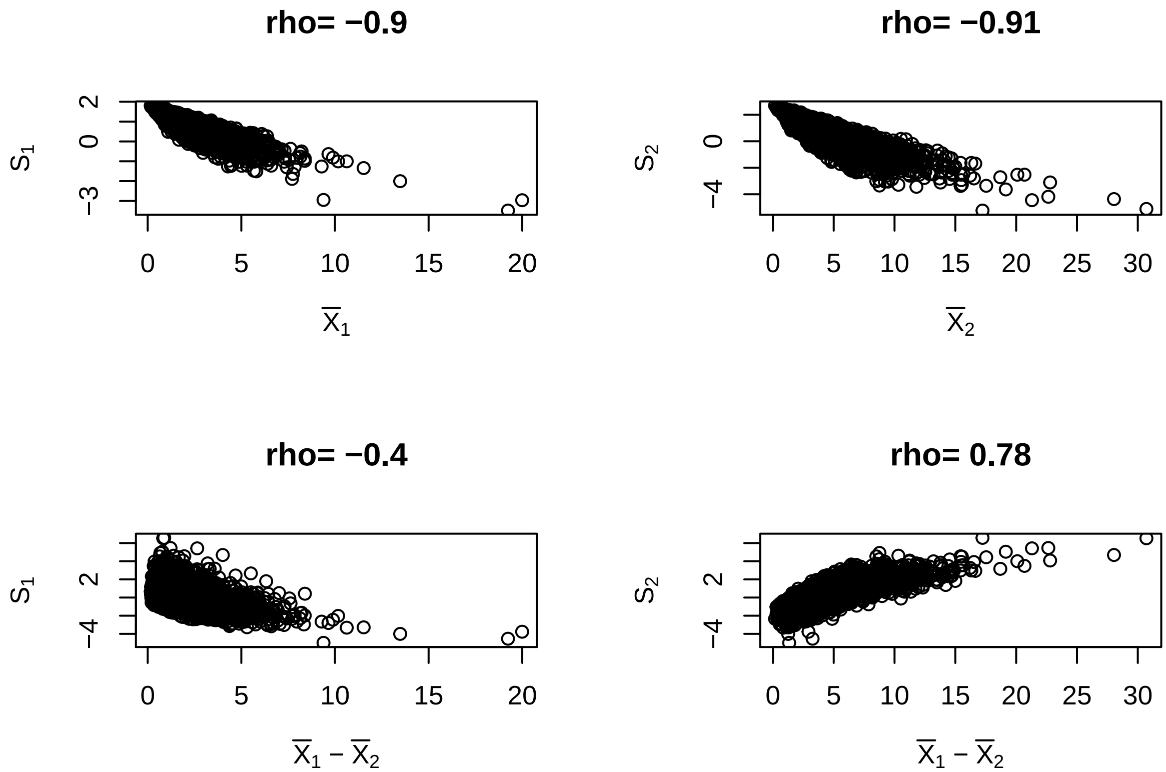

Figure 15. S_j (j=1,2) as a function of $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with S_1=2 and S_2=4

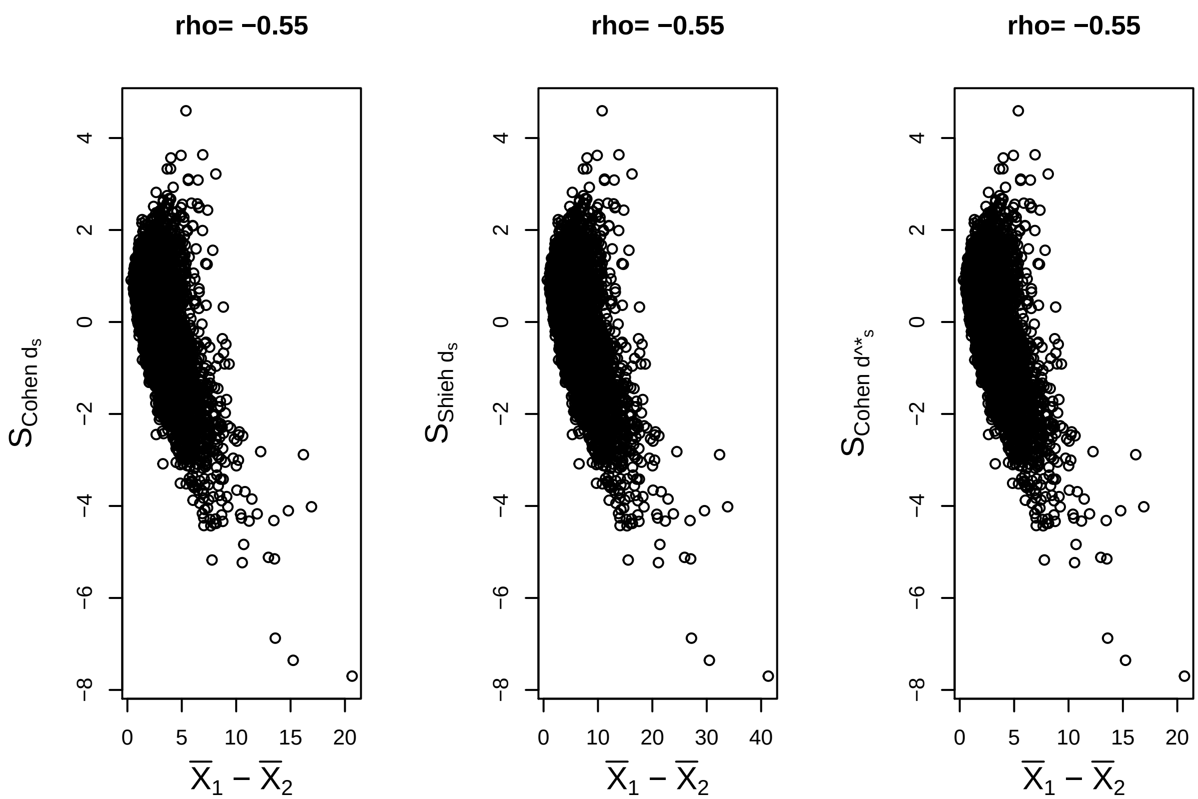

Figure 16. $S_{Cohen's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d_s^*}$ as a function of the mean difference ($\bar{X_1} - \bar{X_2}$), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$), with S_1=2 and S_2=4

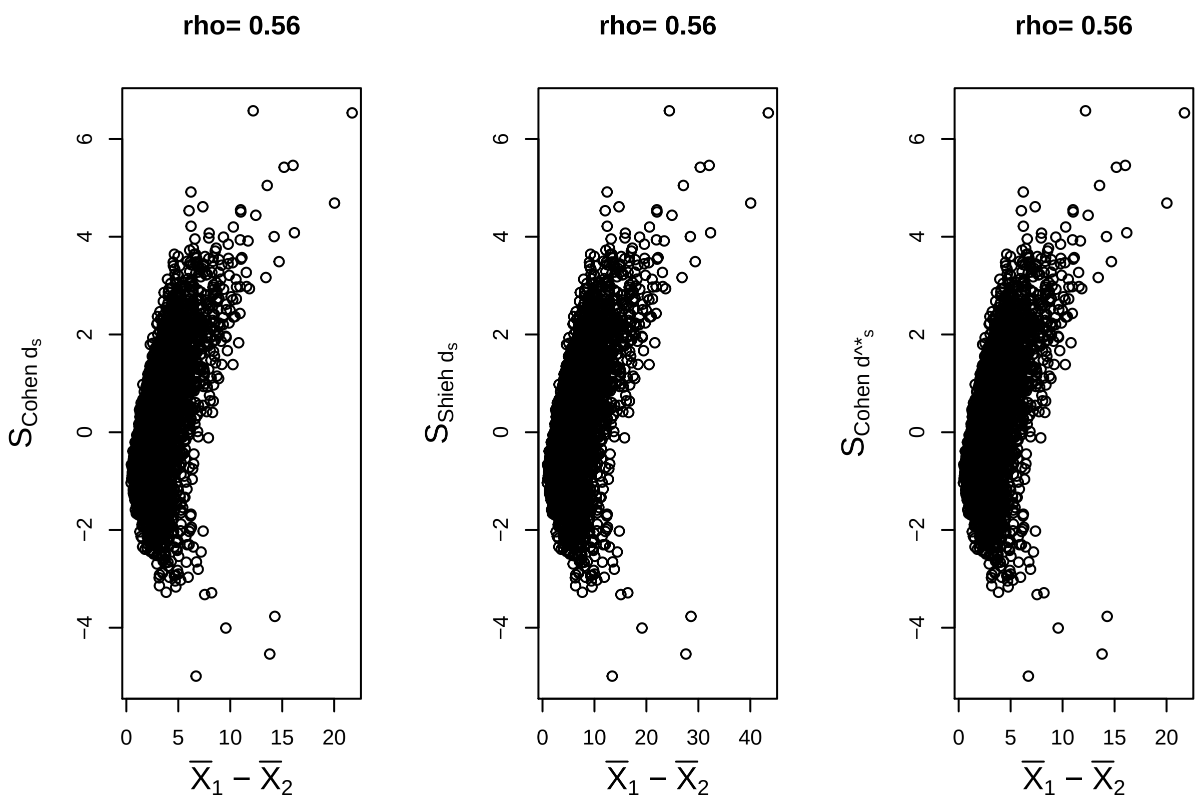

Figure 17. $S_{Cohen's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d_s^*}$ as a function of the mean difference ($\bar{X}_1 - \bar{X}_2$), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$), with S_1=2 and S_2=4