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

Delacre Marie[1]

[1] ULB

Correlations between the sample means 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 groups scores means are compared. The population effect size is defined as

$$\delta = \frac{\mu_1 - \mu_2}{\sigma} \tag{1}$$

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 (σ). When the equality of variances assumption is met, σ is estimated by pooling both samples standard deviations (S_1 and S_2):

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

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:

$$S_{Glass's\ d_s} = S_c \tag{3}$$

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

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

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

$$S_{Cohen's\ d'_s} = \sqrt{\frac{(S_1^2 + S_2^2)}{2}} \tag{5}$$

As we previously mentioned, using these formulas implies meeting 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 distribution, correlations might occur between the sample means difference ($\bar{X}_1 - \bar{X}_2$) and standardizers (σ). Studying when these correlations occur is the main goal of this appendix. To this end, we will distinguish 4 situations, as a function of the sample sizes ratio $\left(\frac{n_1}{n_2} = 1\ vs. \frac{n_1}{n_2} \neq 1\right)$ and the population SD-ratio $\left(\frac{\sigma_1}{\sigma_2} = 1\ vs. \frac{\sigma_1}{\sigma_2} \neq 1\right)$, but before that, we will briefly introduce the impact of 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 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.

GRAPHIQUE POUR L'EXPLIQUER.

Note: I mentioned "all other parameters being equal", 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 decrease the bias. This is a nice example of situations where two factors might have an opposite action on bias.

Correlations between the mean difference ($\bar{X}_1 - \bar{X}_2$) 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 $\bar{X}_1 - \bar{X}_2$ and 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 why 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; note that 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 distribution are right-skewed (left-skewed), as illustrated in the two bottom plots of Figures 8 and 9).

This might also explain why the standardizers of Shieh's d_s and Cohen's d'_s are this time **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 doesn't 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 S_1 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 $\bar{X}_1 - \bar{X}_2$ and respectively SD_1, SD_2, the standardizer of Hedge's g'_s and Shieh's g_s and the standardizer of Hedge's g_s are summarized in Table 2:

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 remain the same for any population SD (σ). However, the magnitude of the correlation between S_j and $\bar{X}_1 - \bar{X}_2$ differ: it is stronger in the sample extracted from the larger population variance.

This also explain that when computing a standardizer taking 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 1 summarizes the sign of the correlation between $\bar{X}_1 - \bar{X}_2$ and respectively SD_1, SD_2 and the three standardizers taking both SD_1 and SD_2 into account (see “Others” in the Table).

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

When distributions are right-skewed

We already know that Glass's d_s using SD_1 will have the smallest bias (and that Glass's d_s using SD_2 as standardizer will have the largest one) when:

- $n_1 > n_2$ (condition b)
- $\sigma_1 < \sigma_2$ (condition c) There is therefore no surprises that the most extreme differences between both Glass's estimators (in favour of using SD_1) occurs when there is a negative pairing between n and SD, and $n_1 > n_2$.

When distributions are left-skewed

We already know that Glass's d_s using SD_2 will have the smallest bias (and that Glass's d_s using SD_1 as standardizer will have the largest one) when:

- $n_1 < n_2$ (condition b)
- $\sigma_1 > \sigma_2$ (condition c) There is therefore no surprises that the most extreme differences between both Glass's estimators (in favour of using SD_2) occurs when there is a negative pairing between n and SD, and $n_1 < n_2$.

Table 1

Correlation between standardizers (SD_1,SD_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, as a function of the n-ratio.

	population distribution	
	right-skewed	*left-skewed*
When $n_1 = n_2$	SD_1: *positive*	SD_1: *negative*
	SD_2: *negative*	SD_2: *positive*
	Others: *null*	Others: *null*
When $n_1 > n_2$	SD_1: *positive*	SD_1: *negative*
	SD_2: *negative*	SD_2: *positive*
	Others: *negative*	Others: *positive*
	$S_{Cohen's\ d_s}$: *null*	$S_{Cohen's\ d_s}$: *null*
When $n_1 < n_2$	SD_1: *positive*	SD_1: *negative*
	SD_2: *negative*	SD_2: *positive*
	Others: *positive*	Others: *negative*
	$S_{Cohen's\ d_s}$: *null*	$S_{Cohen's\ d_s}$: *null*

Table 2

Correlation between standardizers (SD_1,SD_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 $\sigma_1 = \sigma_2$	SD_1: *positive*	SD_1: *negative*
	SD_2: *negative*	SD_2: *positive*
	Others: *null*	Others: *null*
When $\sigma_1 > \sigma_2$	SD_1: *positive*	SD_1: *negative*
	SD_2: *negative*	SD_2: *positive*
	Others: *positive*	Others: *negative*
When $\sigma_1 < \sigma_2$	SD_1: *positive*	SD_1: *negative*
	SD_2: *negative*	SD_2: *positive*
	Others: *negative*	Others: *positive*

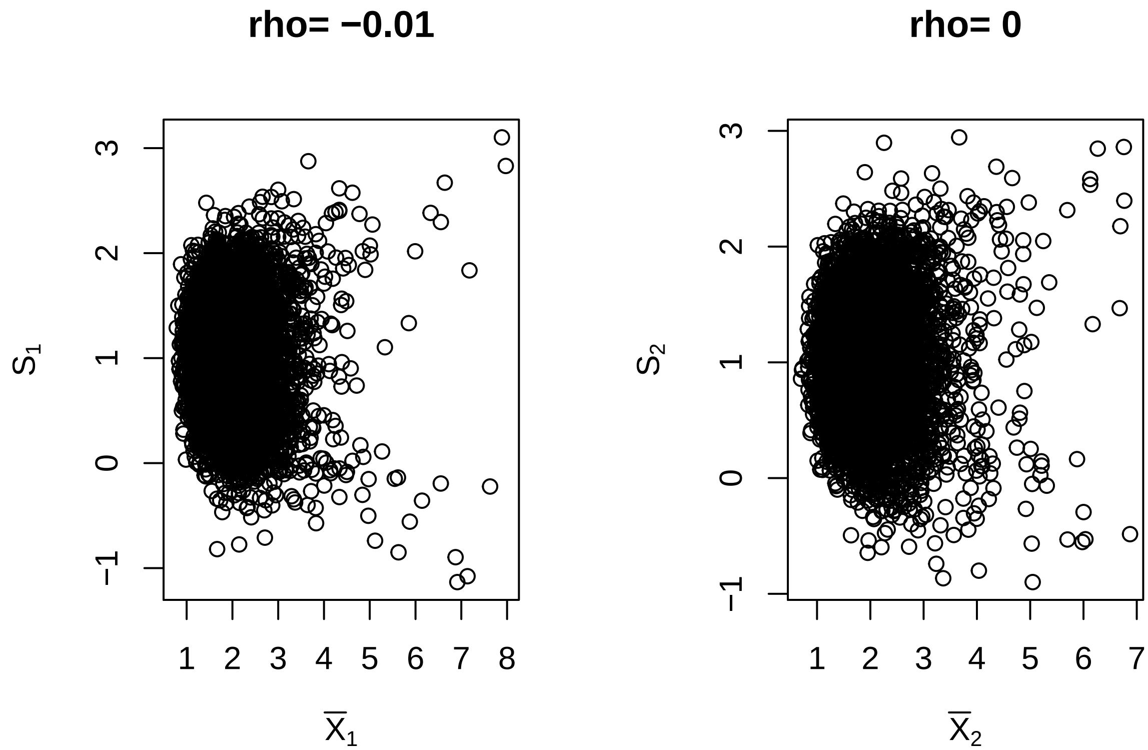

Figure 1. S_j as a function of $\bar{X}_j$ (j=1,2), when samples are extracted from symmetric distributions ($\gamma_1 = 0$)

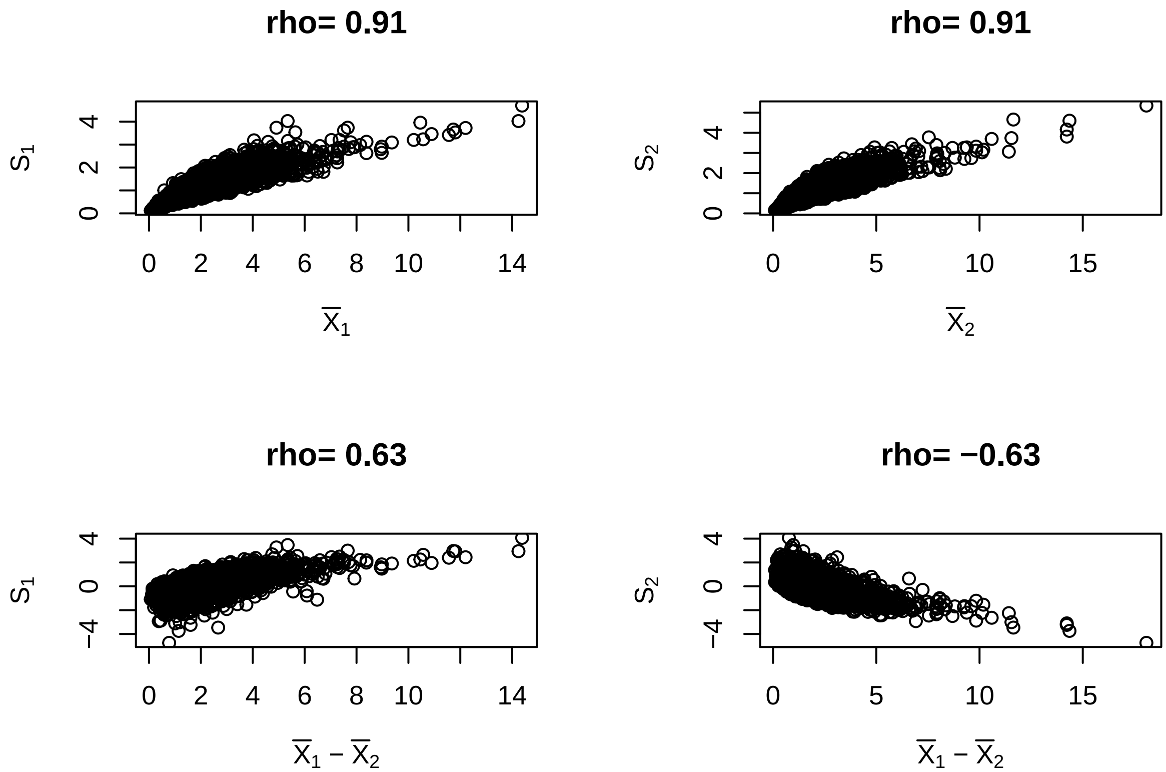

Figure 2. S_j (j=1,2) as a function $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$; top plots)

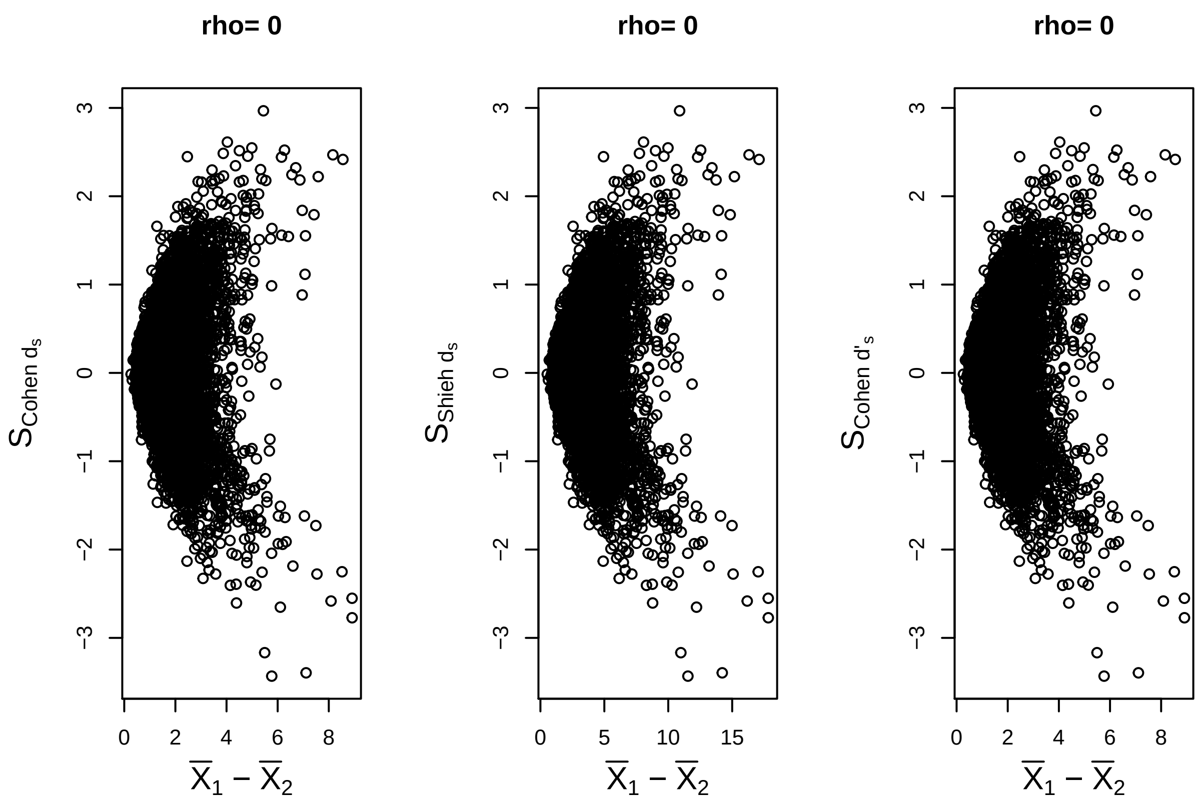

Figure 3. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d'_s}$ as a function of the means difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from right skewed distributions $(\gamma_1 = 6.32)$

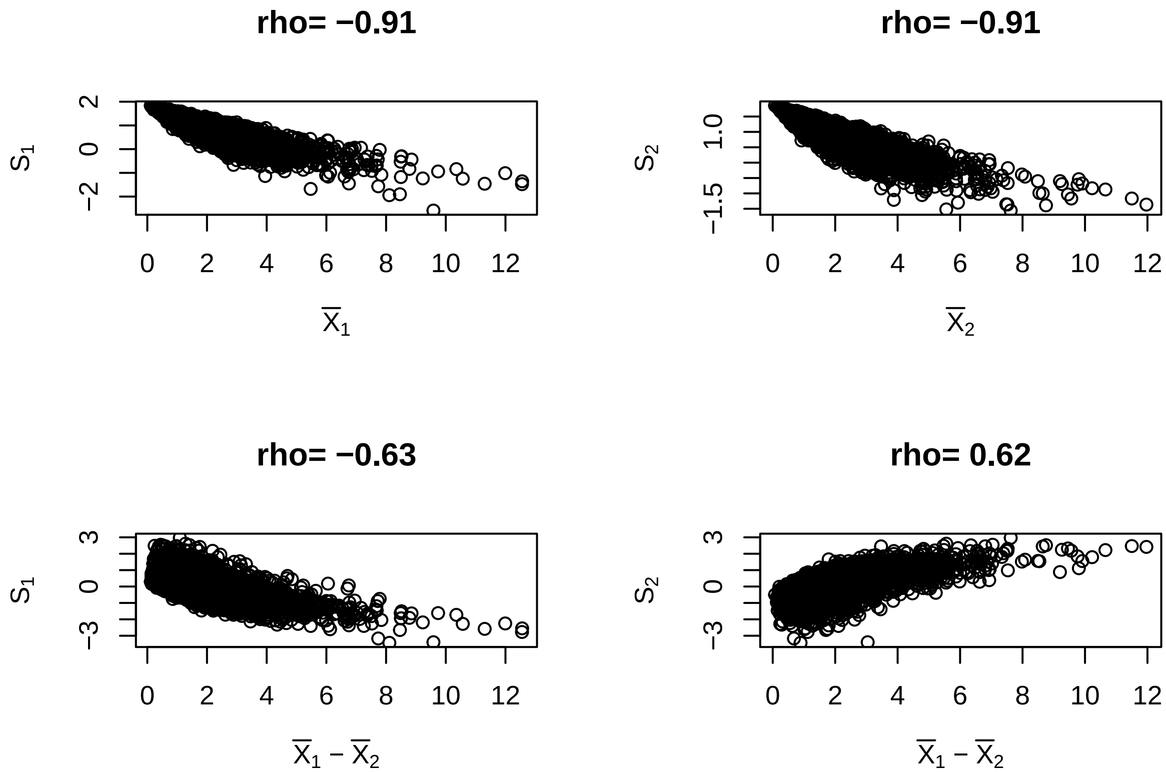

Figure 4. S_j (j=1,2) as a function $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$; top plots)

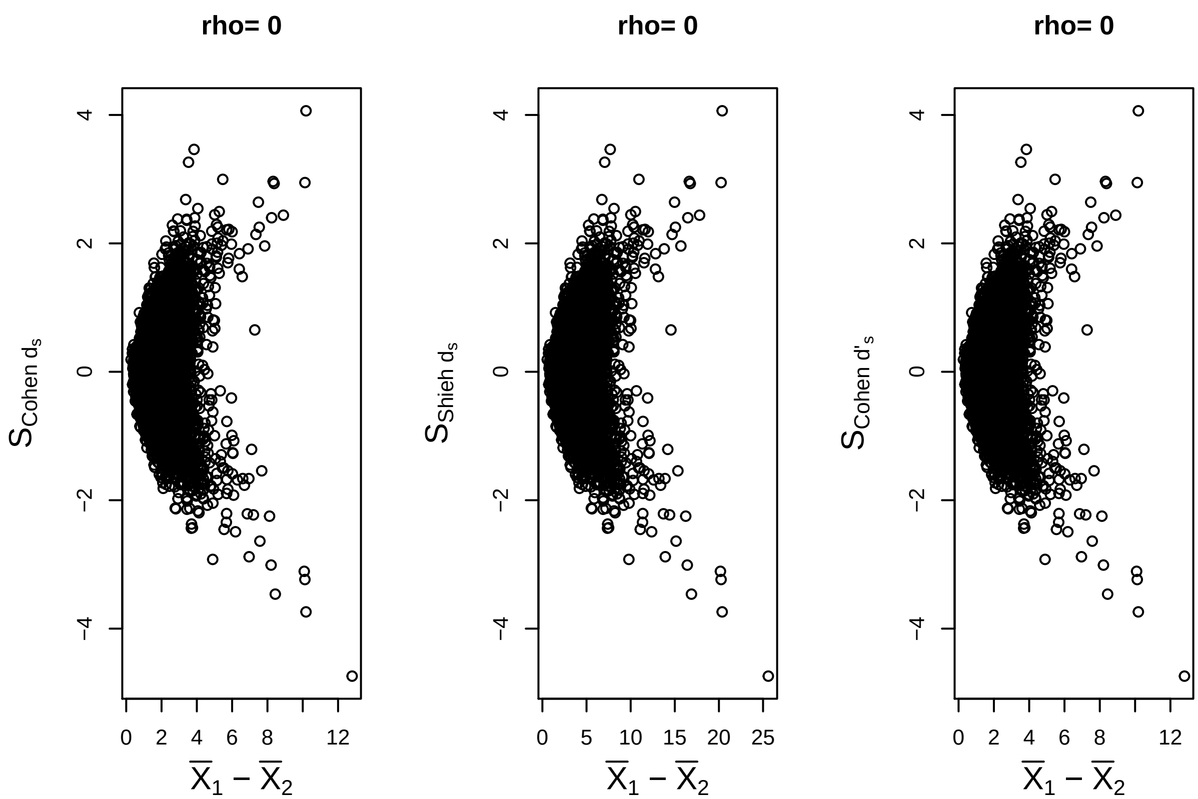

Figure 5. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d'_s}$ as a function of the means difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions $(\gamma_1 = -6.32)$

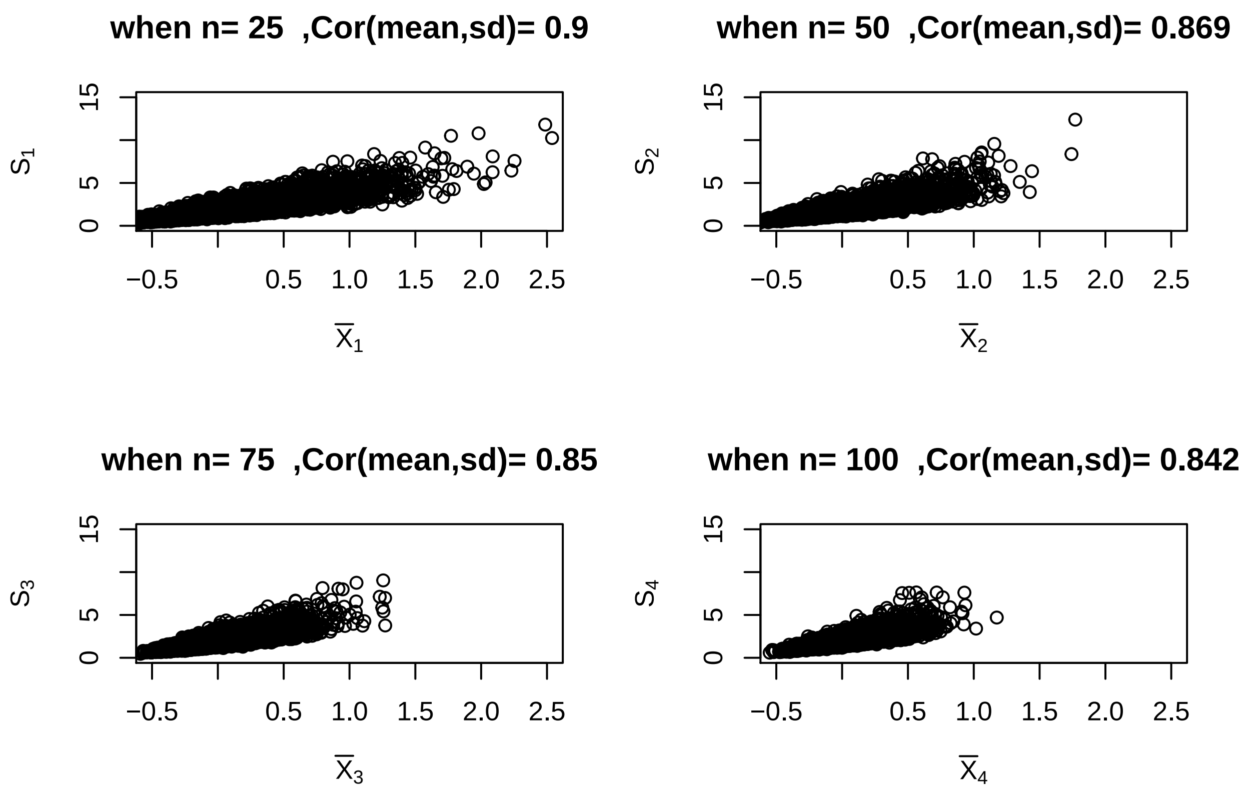

Figure 6. correlation between S_j and $\bar{X}_j$ when n = 25, 50, 75 or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

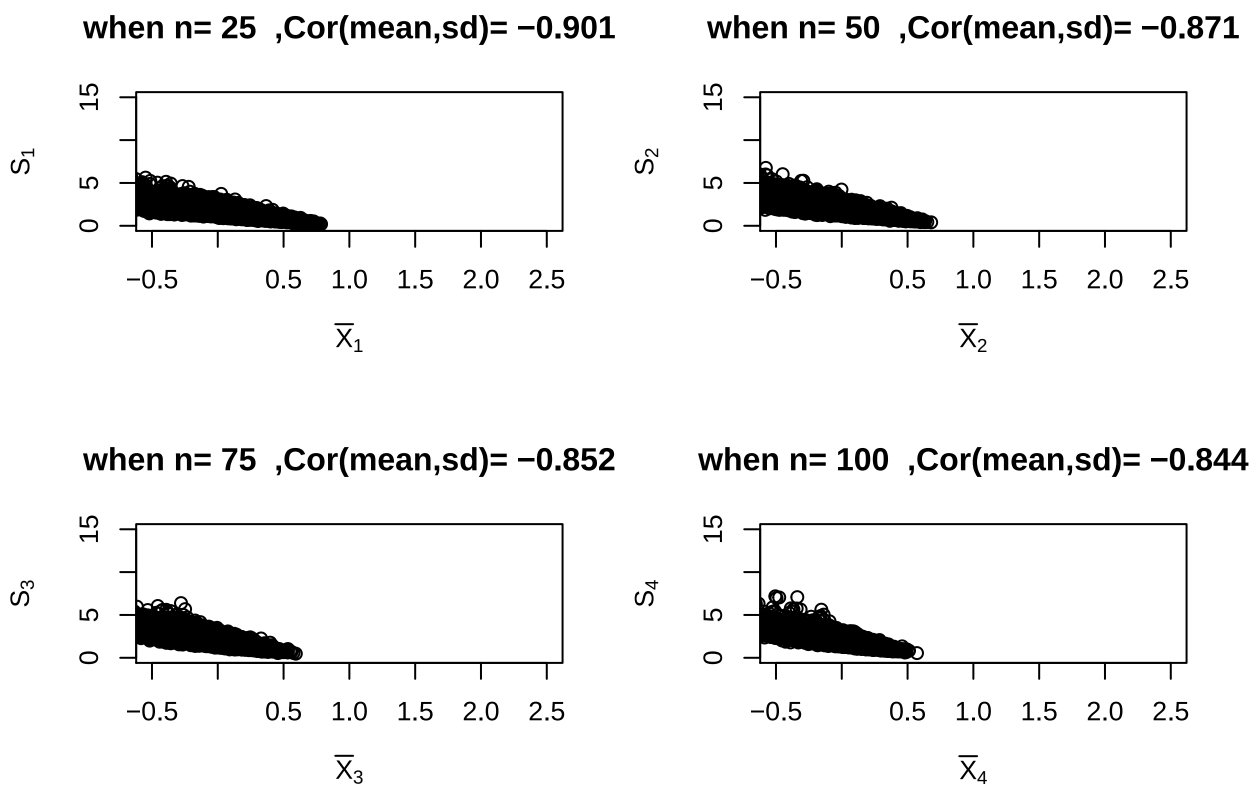

Figure 7. correlation between S_j and $\bar{X}_j$ when n = 25, 50, 75 or 100 and samples are extracted from right left distributions ($\gamma_1 = -6.32$)

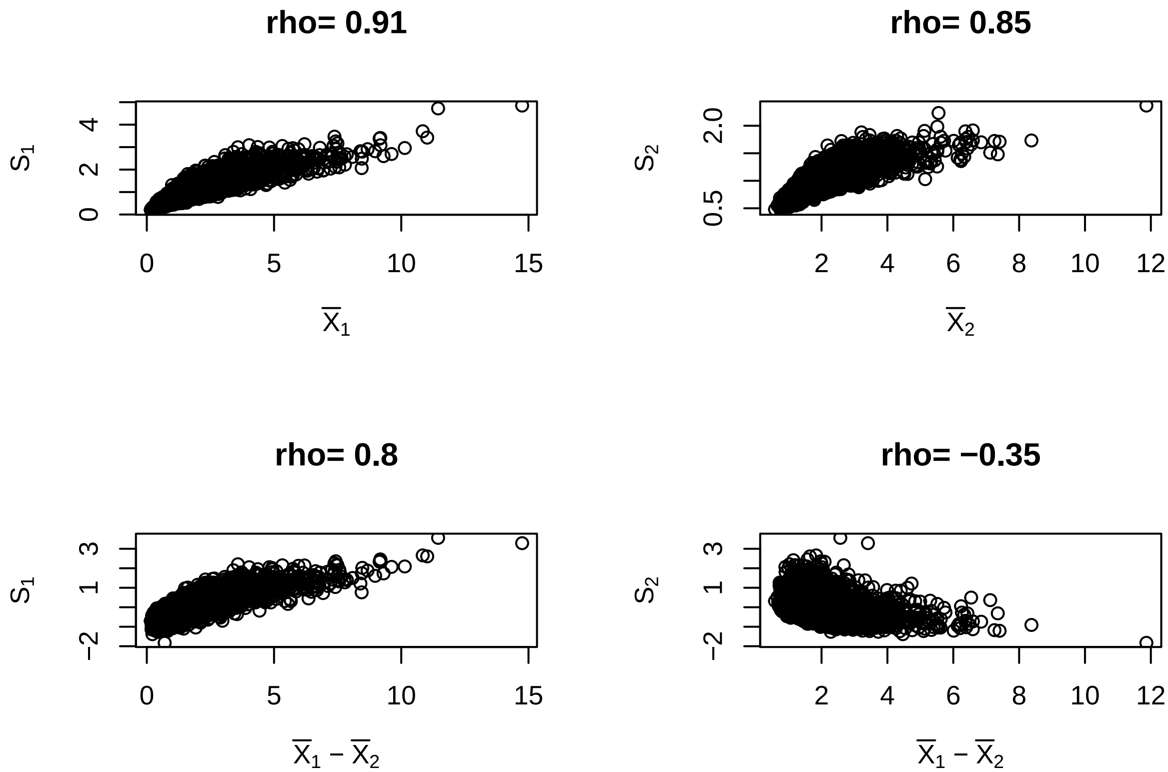

Figure 8. S_j (j=1,2) as a function $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$; top plots), with n1=20 and n2=100

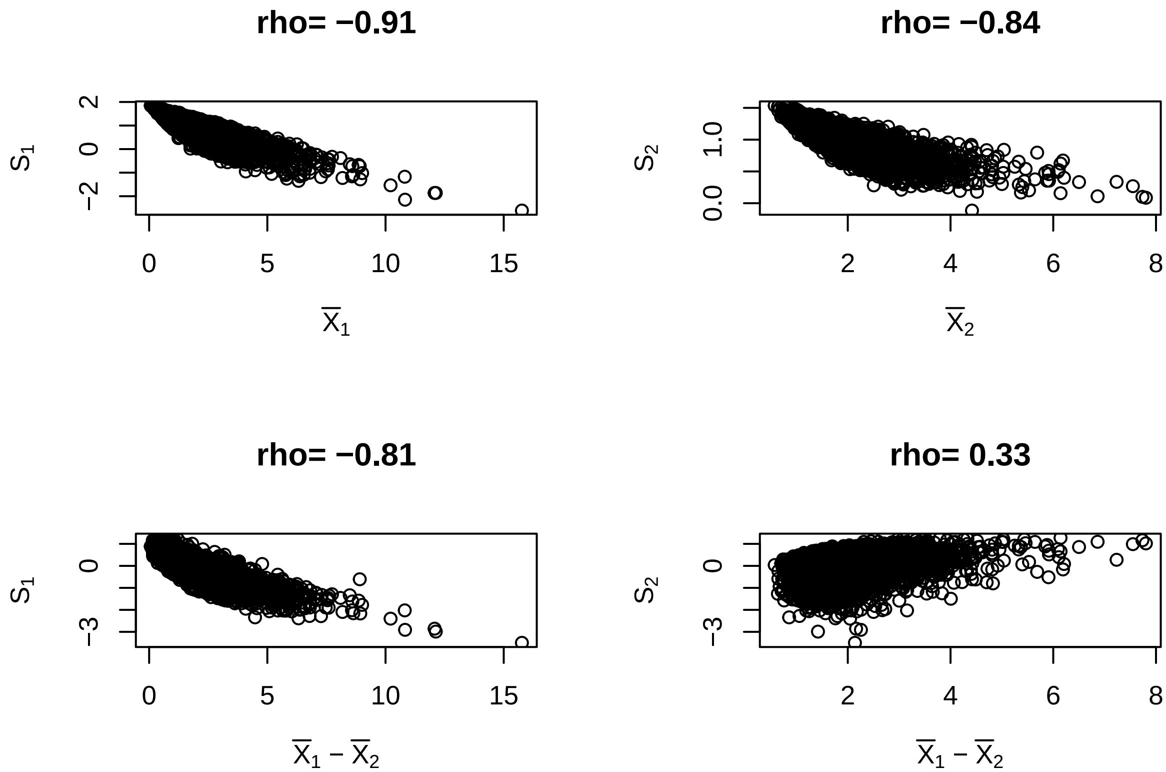

Figure 9. S_j (j=1,2) as a function $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$; top plots), with n1=20 and n2=100

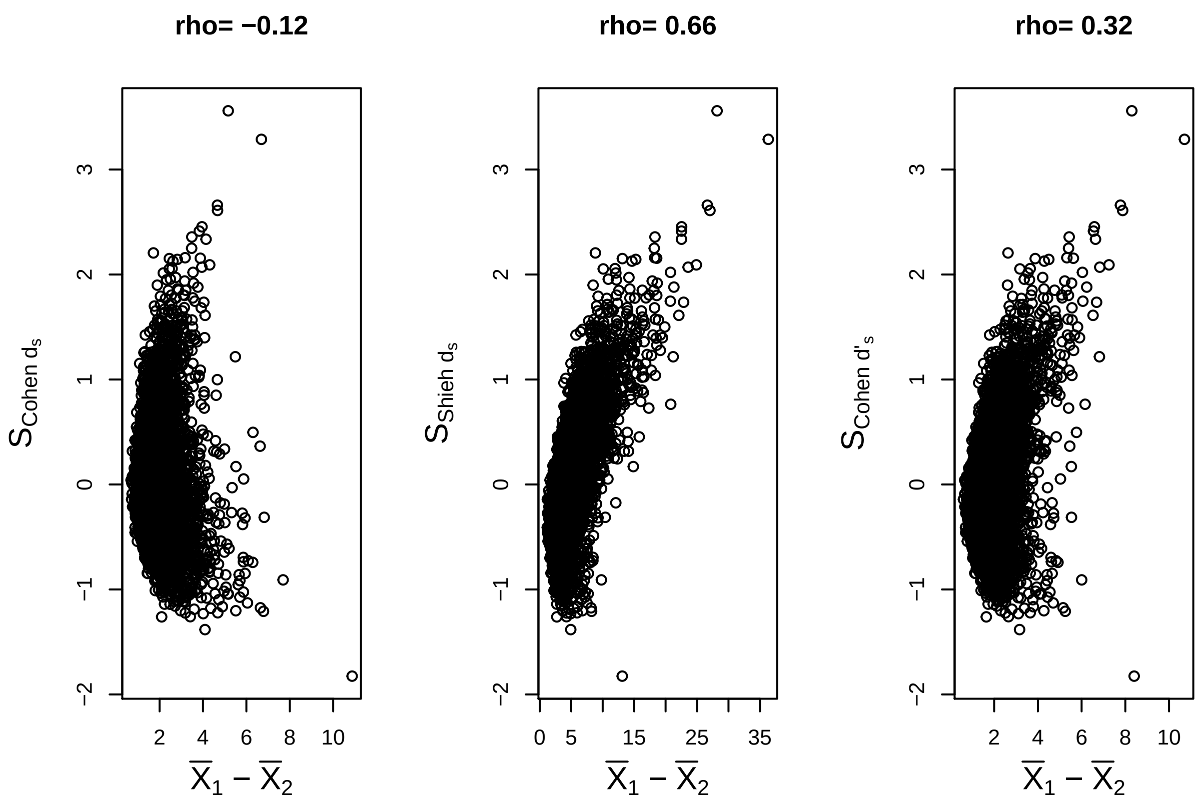

Figure 10. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d'_s}$ as a function of the means difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from right skewed distributions $(\gamma_1 = 6.32)$

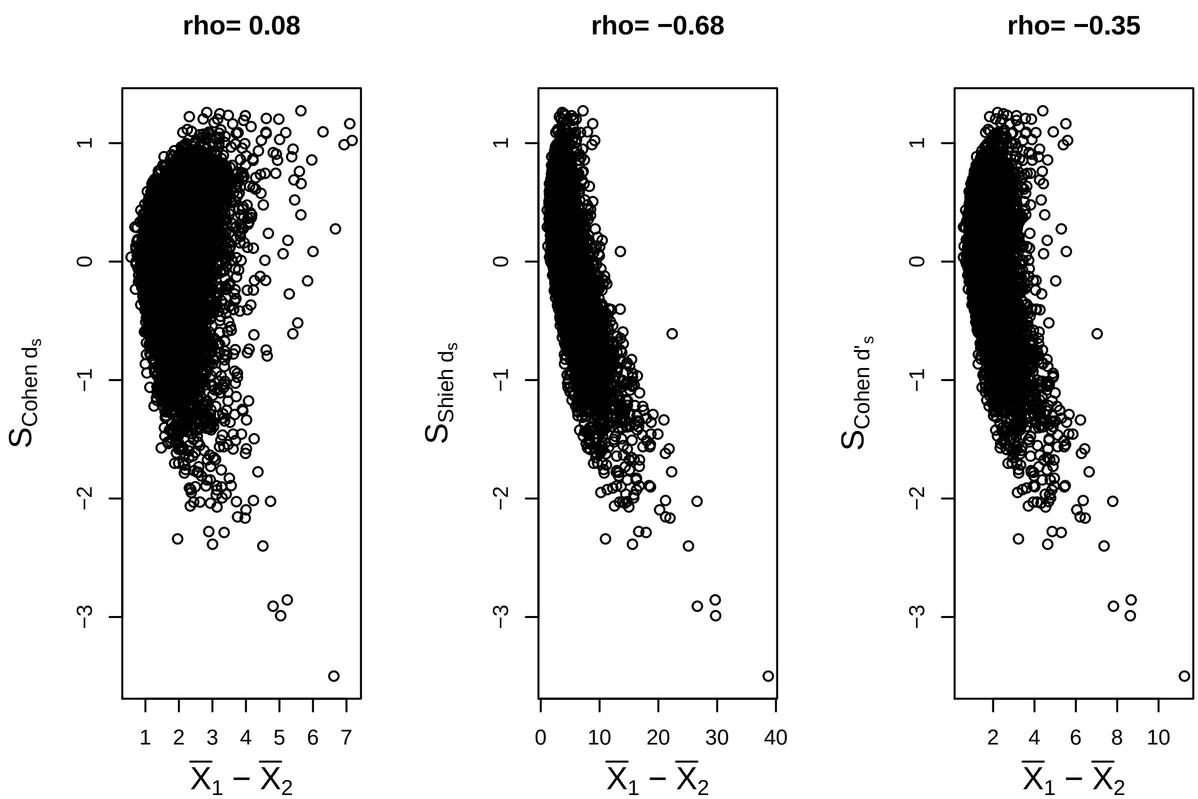

Figure 11. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d'_s}$ as a function of the means difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions $(\gamma_1 = -6.32)$

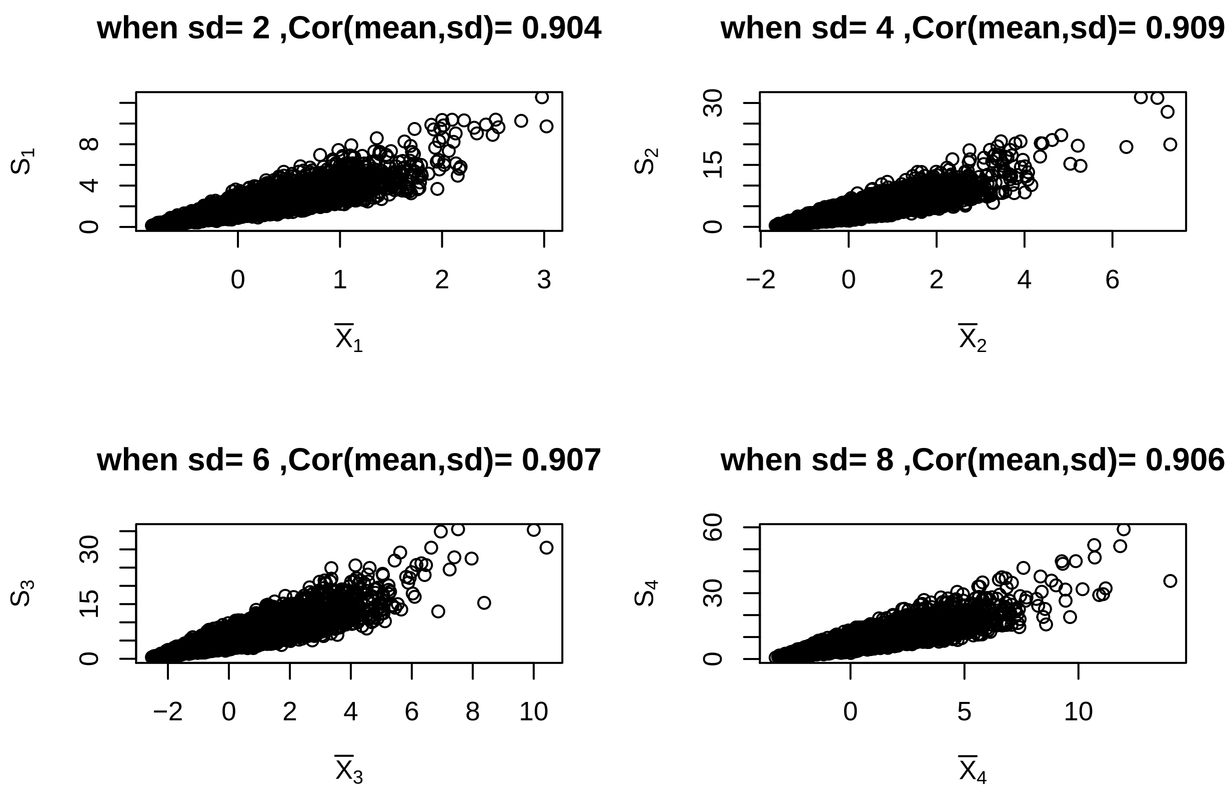

Figure 12. correlation between S_j and $\bar{X}_j$ when n = 25, 50, 75 or 100 and samples are extracted from right skewed distributions ($\gamma_1 = 6.32$)

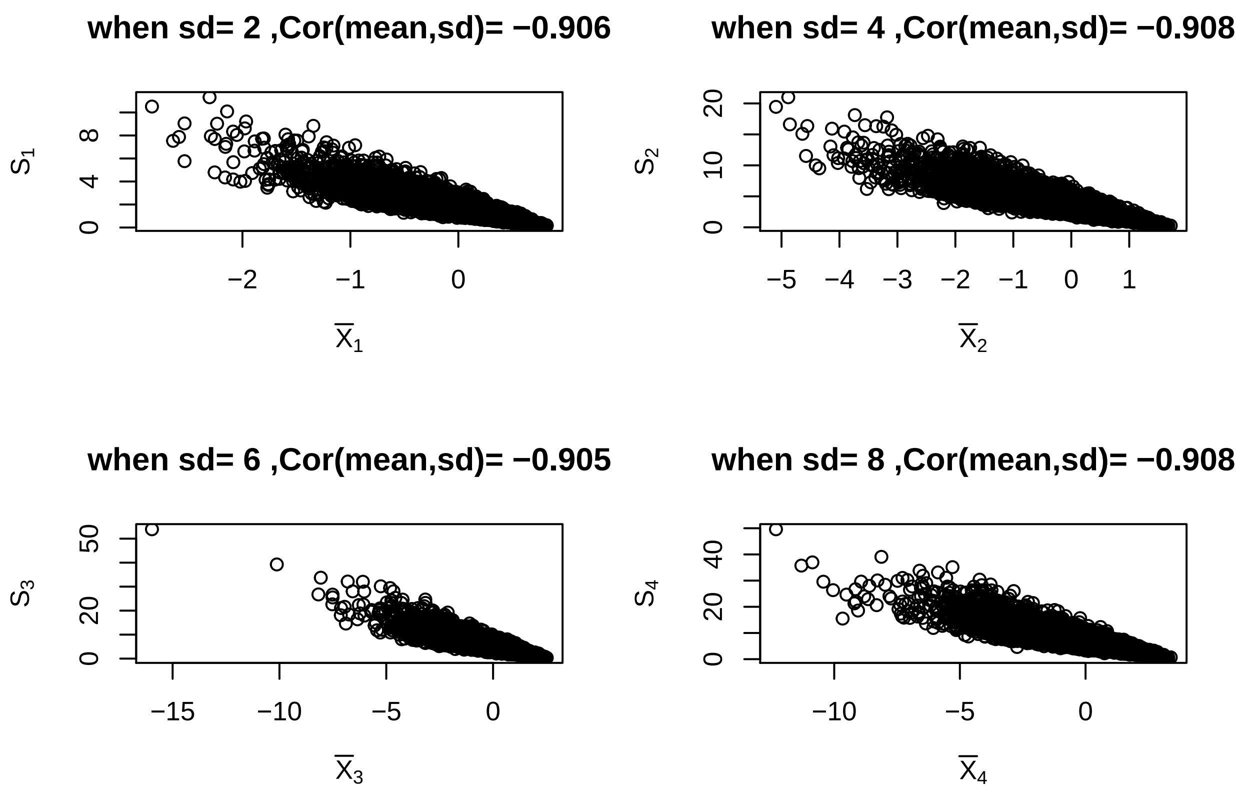

Figure 13. correlation between S_j and $\bar{X}_j$ when n = 25, 50, 75 or 100 and samples are extracted from left skewed distributions ($\gamma_1 = -6.32$)

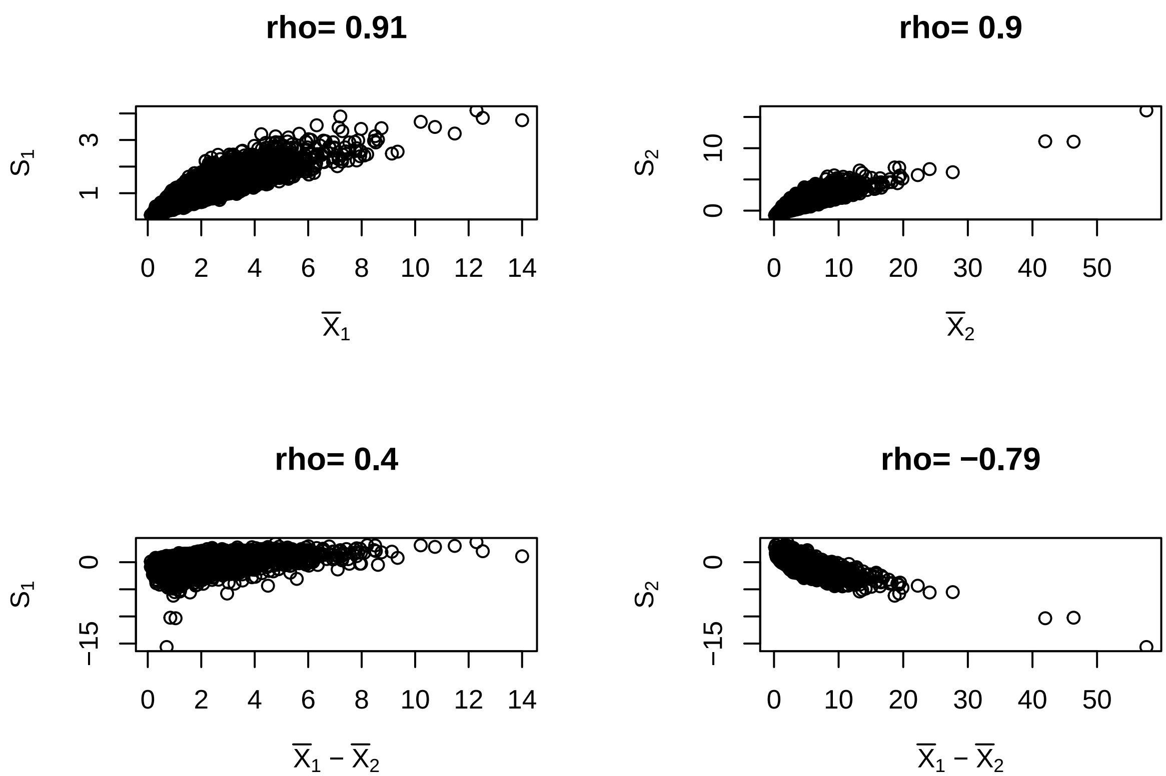

Figure 14. S_j (j=1,2) as a function $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from right skewed distributions ($\gamma_1 = 6.32$; top plots), with n1=20 and n2=100

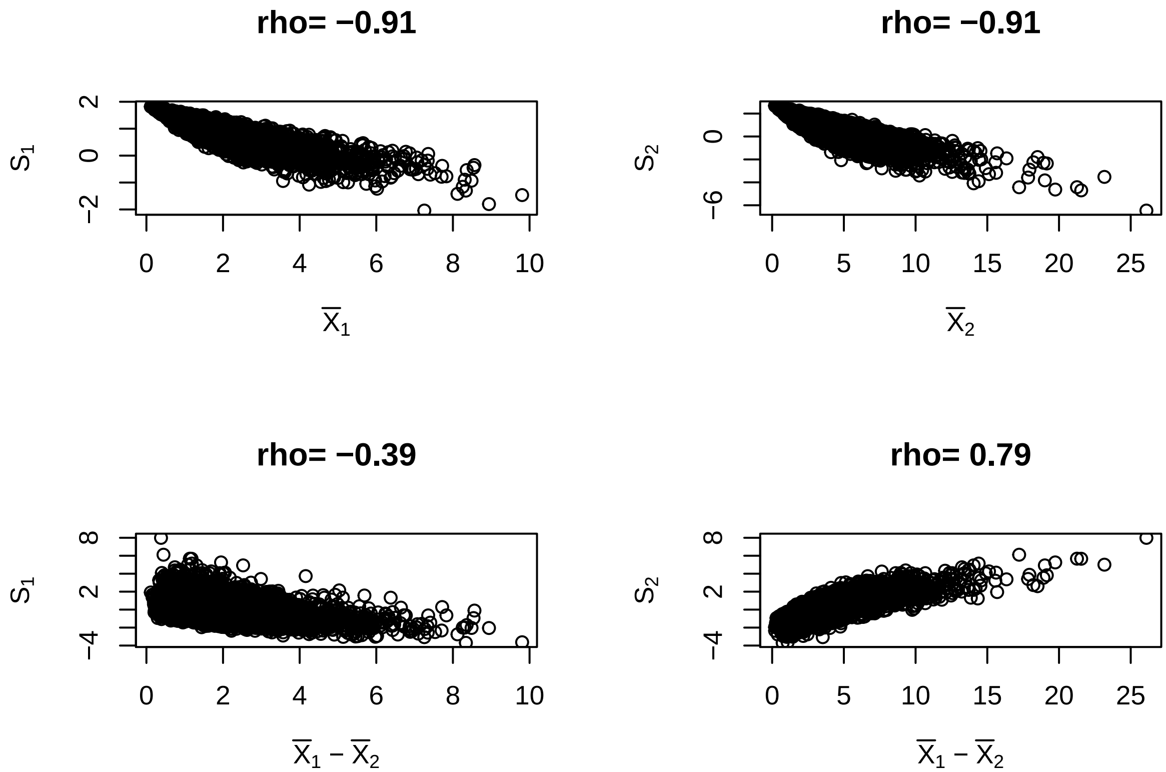

Figure 15. S_j (j=1,2) as a function $\bar{X}_j$ (top plots) or $\bar{X}_1 - \bar{X}_2$ (bottom plots), when samples are extracted from left skewed distributions ($\gamma_1 = -6.32$; top plots), with n1=20 and n2=100

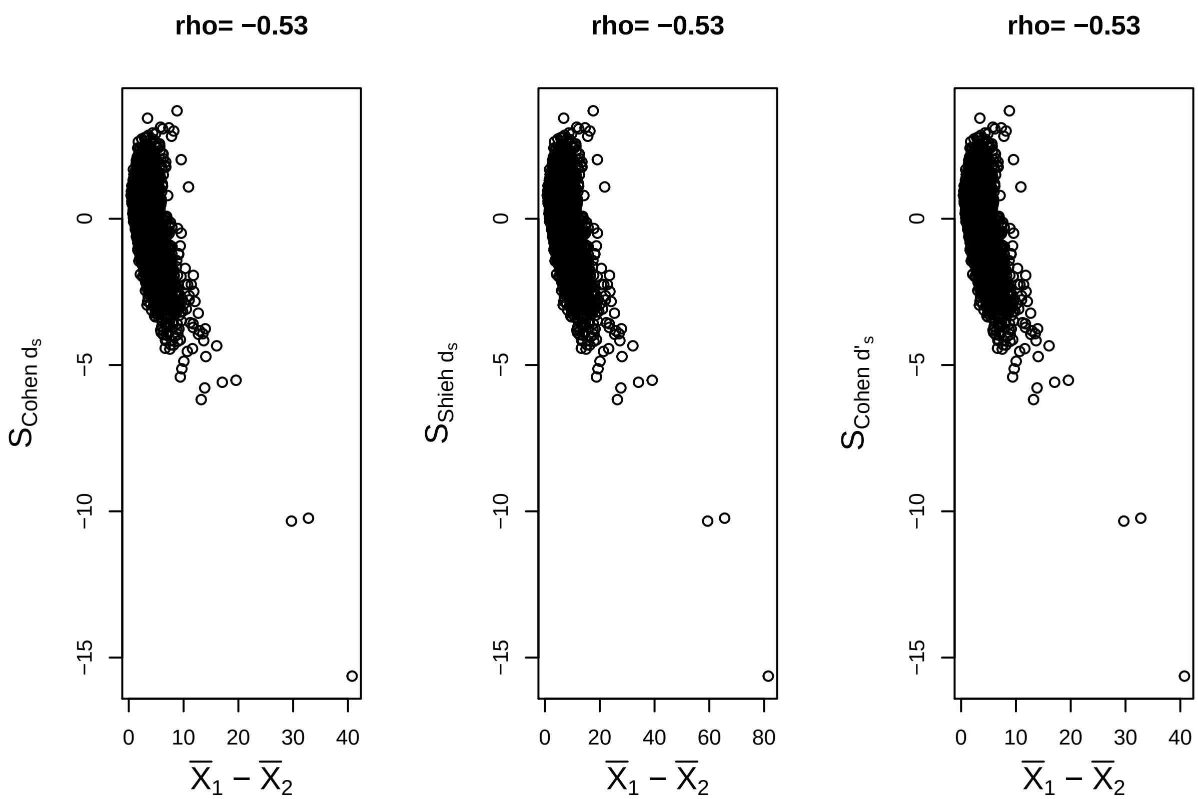

Figure 16. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d'_s}$ as a function of the means difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from right skewed distributions $(\gamma_1 = 6.32)$

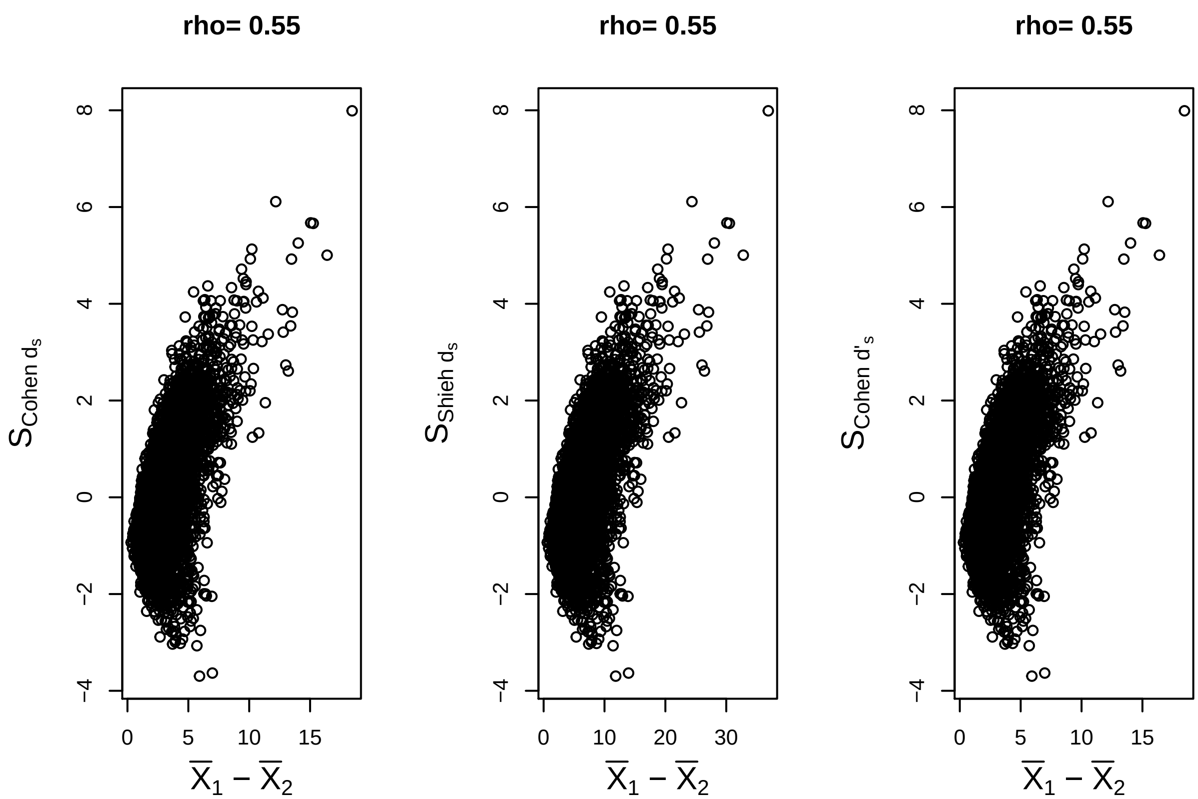

Figure 17. $S_{Glass's\ d_s}$, $S_{Shieh's\ d_s}$ and $S_{Cohen's\ d'_s}$ as a function of the means difference $(\bar{X}_1 - \bar{X}_2)$, when samples are extracted from left skewed distributions $(\gamma_1 = -6.32)$