

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

<sup>3</sup> Delacre Marie<sup>1</sup>

<sup>4</sup> <sup>1</sup> ULB

<sup>5</sup> Author Note

<sup>6</sup> 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**

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

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

where both populations follow a normal distribution with mean  $\mu_j$  in the  $j^{th}$  population ( $j=1,2$ ) and common standard deviation  $\sigma$ . 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,  $\sigma$  is estimated by pooling both sample standard deviations ( $S_1$  and  $S_2$ ):

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

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

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

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

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

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

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

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

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

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

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

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

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

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

44 bias and the variance. For example, when population variances are equal across groups and  
 45 sample sizes are unequal, we will see below that the lower  $n_j$ , the larger the magnitude of the  
 46 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  
 47 positive, the smaller the sample size, the larger the positive correlation. At the same time,  
 48 we know that increasing the sample size decrease the bias. This is a nice example of  
 49 situations where two factors might have an opposite action on bias.

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

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

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

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

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

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

68 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  
 69 bottom plots in Figure 4).

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

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

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

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

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

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

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

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

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

95 The correlation between  $\bar{X}_1 - \bar{X}_2$  and respectively  $SD_1$ ,  $SD_2$ , the standardizer of  
96 Cohen's  $d_s^*$ , the standardizer of Shieh's  $d_s$  and the standardizer of Cohen's  $d_s$  are  
97 summarized in Table 1:

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

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

104 This also explain that when computing a standardizer that takes both  $S_1$  and  $S_2$  into  
105 account, it results in a standardizer that is correlated with  $\bar{X}_1 - \bar{X}_2$  (see Figures 16 and 17).  
106 The correlation between the mean difference ( $\bar{X}_1 - \bar{X}_2$ ) and respectively the standardizer of  
107 Shieh's  $d_s$ , Cohen's  $d_s^*$  and Cohen's  $d_s$  will have the same sign as the correlation between  
108 ( $\bar{X}_1 - \bar{X}_2$ ) and the larger  $SD$ . Table 2 summarizes the sign of the correlation between  
109  $\bar{X}_1 - \bar{X}_2$  and respectively  $SD_1$ ,  $SD_2$  and the three standardizers taking both  $SD_1$  and  $SD_2$   
110 into account (see "Others" in the Table).

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, 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 $n_1 = n_2$				
	$SD_1$ : positive			$SD_1$ : negative
	$SD_2$ : negative			$SD_2$ : 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 $n_1 > n_2$				
	$SD_1$ : positive			$SD_1$ : negative
	$SD_2$ : negative			$SD_2$ : 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 $n_1 < n_2$				
	$SD_1$ : positive			$SD_1$ : negative
	$SD_2$ : negative			$SD_2$ : 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 ( $SD_1, SD_2$  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.*

<b>population distribution</b>		
	<i>right-skewed</i>	<i>left-skewed</i>
When $\sigma_1 = \sigma_2$	$SD_1$ : <i>positive</i> $SD_2$ : <i>negative</i> Others: <i>null</i>	$SD_1$ : <i>negative</i> $SD_2$ : <i>positive</i> Others: <i>null</i>
When $\sigma_1 > \sigma_2$	$SD_1$ : <i>positive</i> $SD_2$ : <i>negative</i> Others: <i>positive</i>	$SD_1$ : <i>negative</i> $SD_2$ : <i>positive</i> Others: <i>negative</i>
When $\sigma_1 < \sigma_2$	$SD_1$ : <i>positive</i> $SD_2$ : <i>negative</i> Others: <i>negative</i>	$SD_1$ : <i>negative</i> $SD_2$ : <i>positive</i> Others: <i>positive</i>

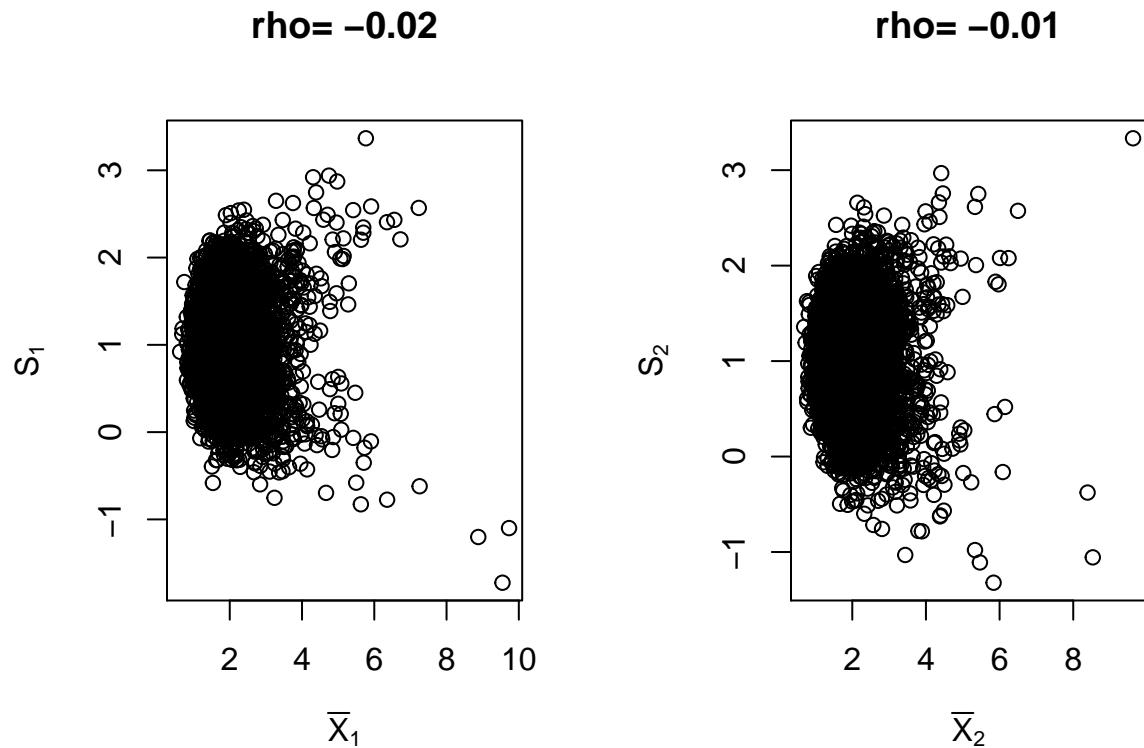


Figure 1.  $S_j$  as a function of  $\bar{X}_j$  ( $j=1,2$ ), when samples are extracted from symmetric distributions ( $\gamma_1 = 0$ )

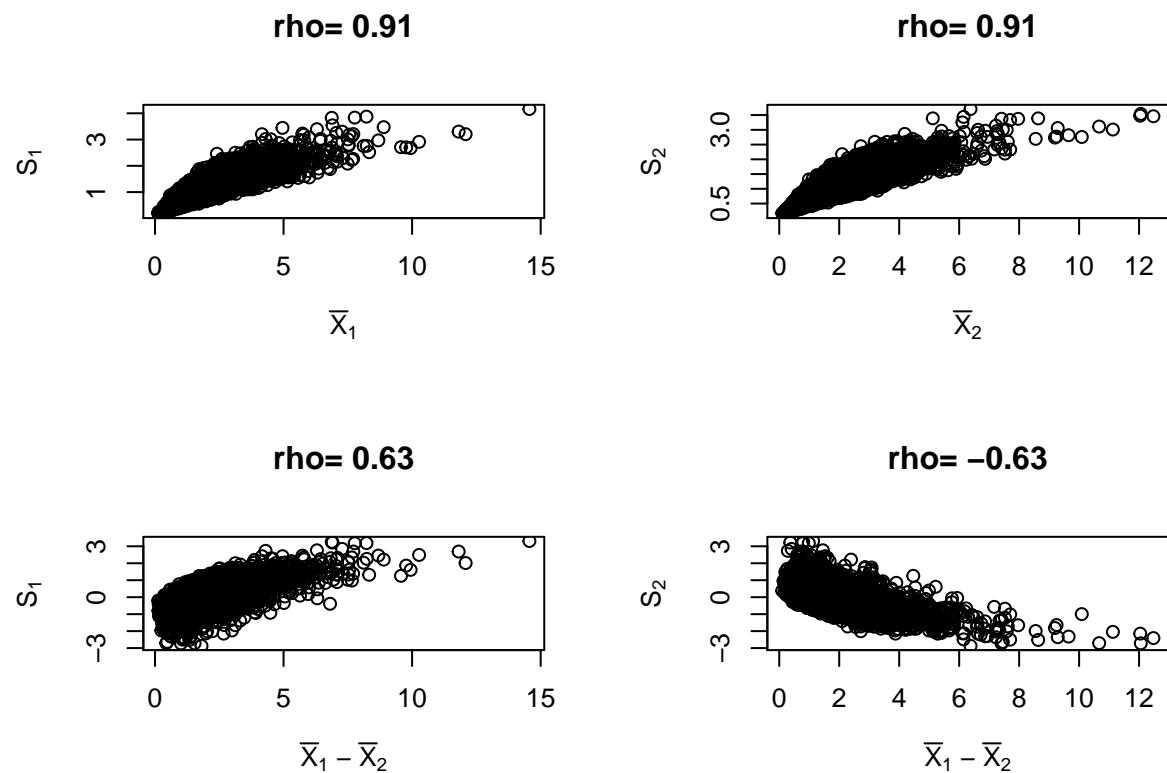
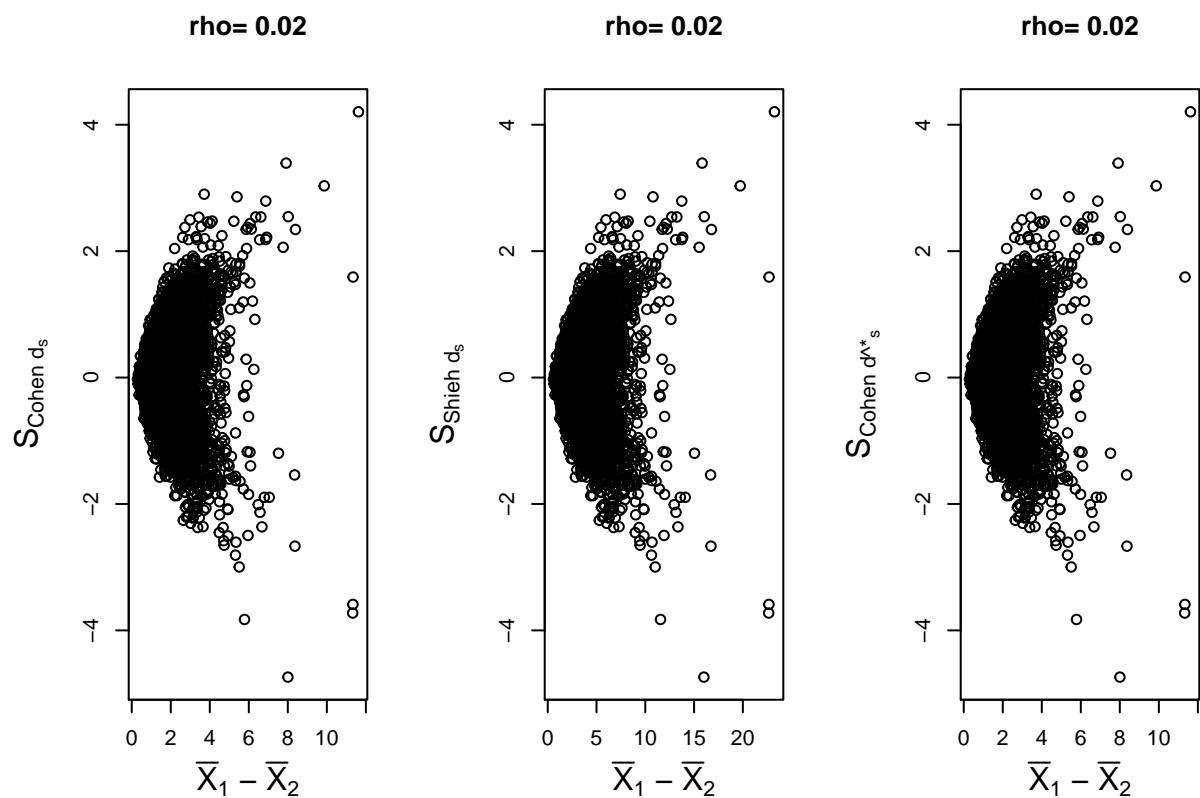


Figure 2.  $S_j$  ( $j=1,2$ ) as a function  $\bar{X}_j$  (top plots) or  $\bar{X}_1 - \bar{X}_2$  (bottom plots), when samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ )



*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 ( $\gamma_1 = 6.32$ )

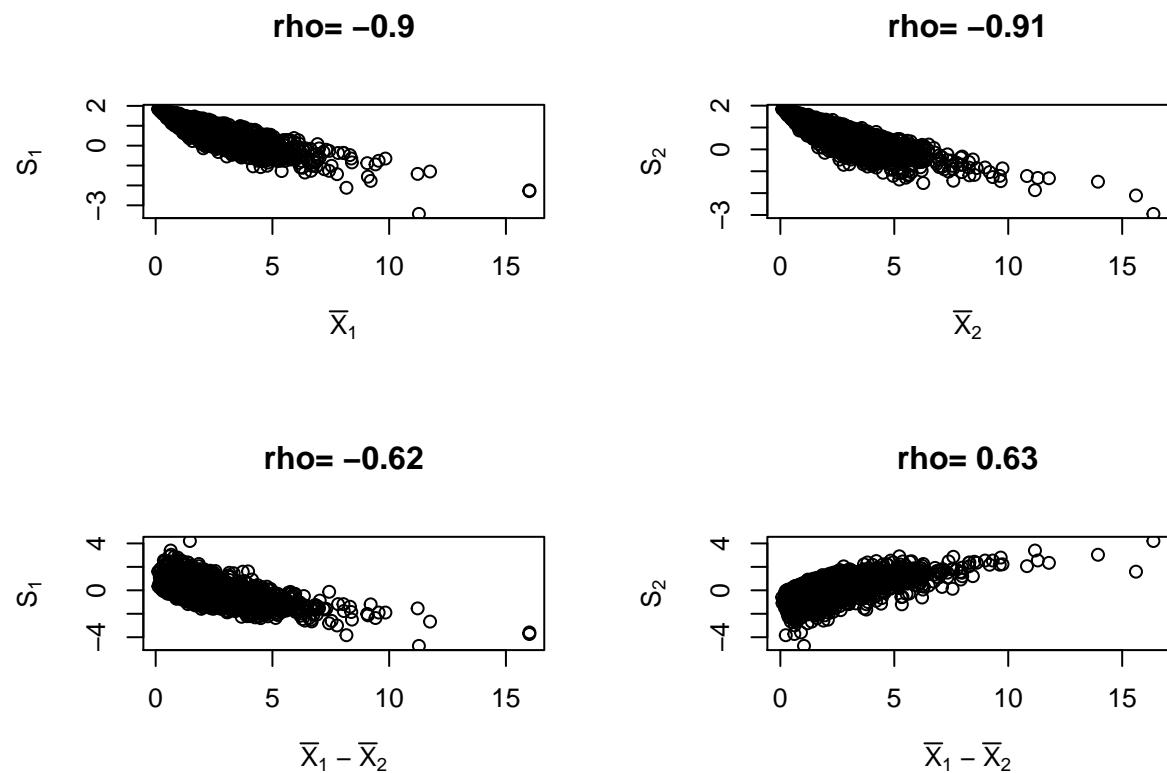


Figure 4.  $S_j$  ( $j=1,2$ ) as a function  $\bar{X}_j$  (top plots) or  $\bar{X}_1 - \bar{X}_2$  (bottom plots), when samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ )

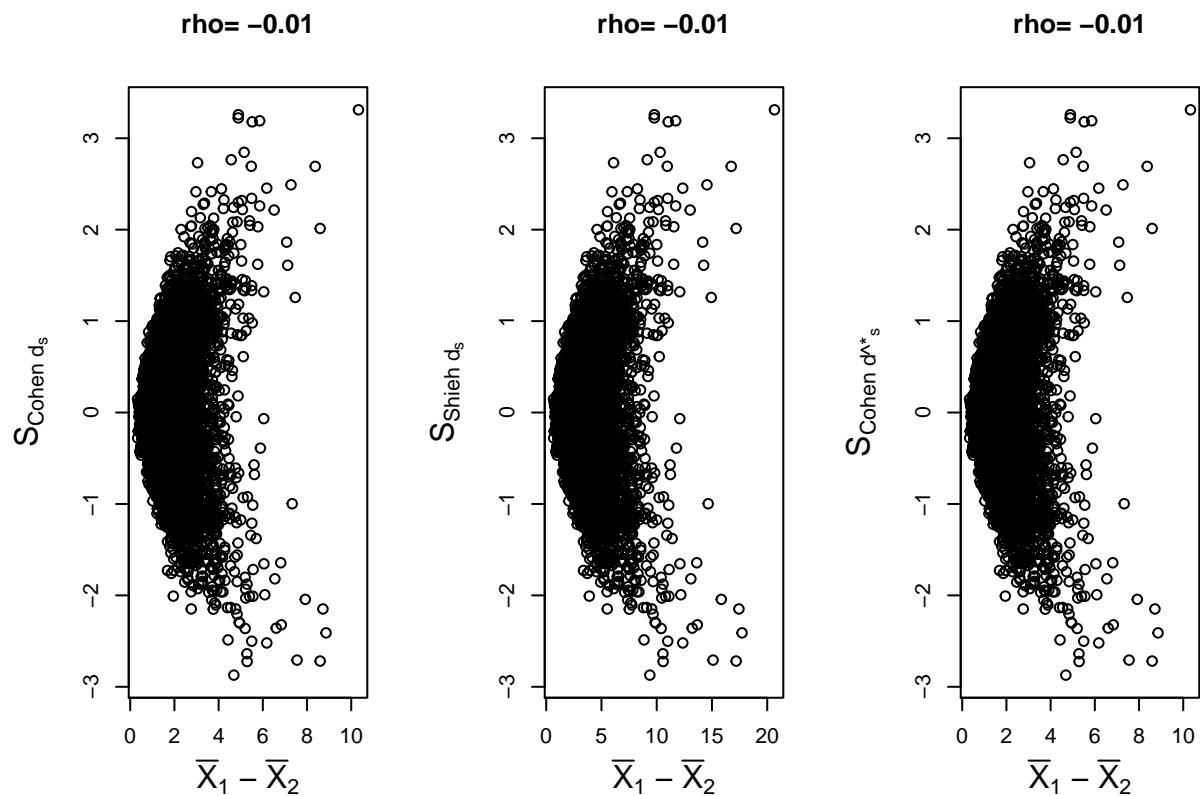


Figure 5.  $S_{\text{Glass}'s} d_s$ ,  $S_{\text{Shieh}'s} d_s$  and  $S_{\text{Cohen}'s} d_s^*$  as a function of the mean difference ( $\bar{X}_1 - \bar{X}_2$ ), when samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ )

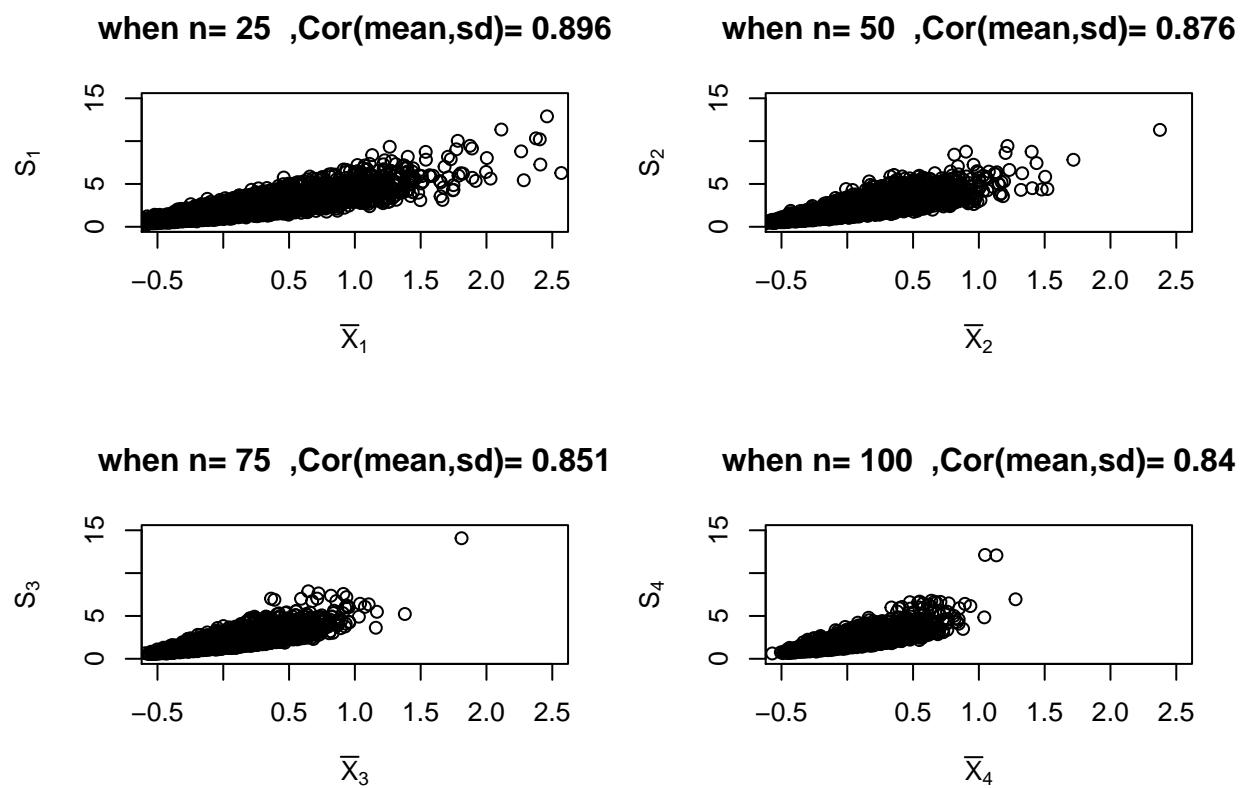


Figure 6. correlation between  $S_j$  and  $\bar{X}_j$  when  $n = 25, 50, 75$  or  $100$  and samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ )

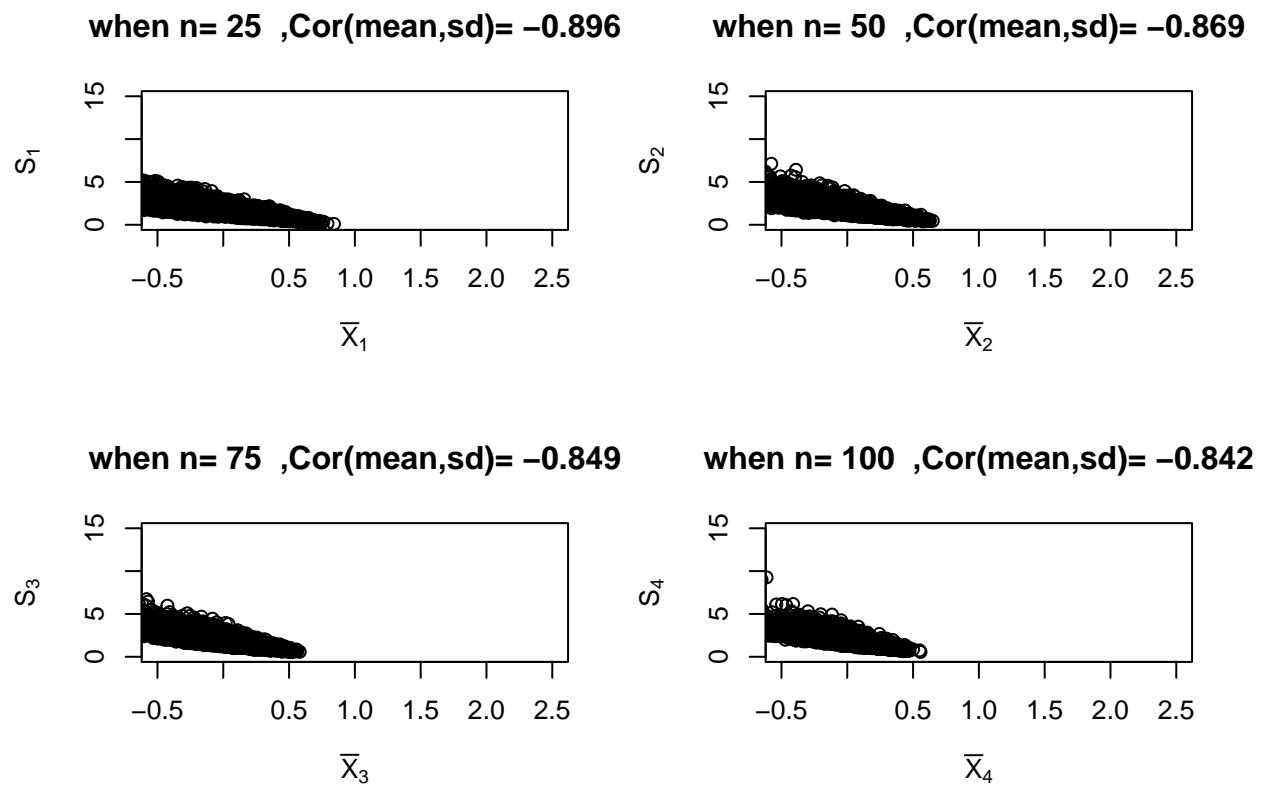


Figure 7. correlation between  $S_j$  and  $\bar{X}_j$  when  $n = 25, 50, 75$  or  $100$  and samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ )

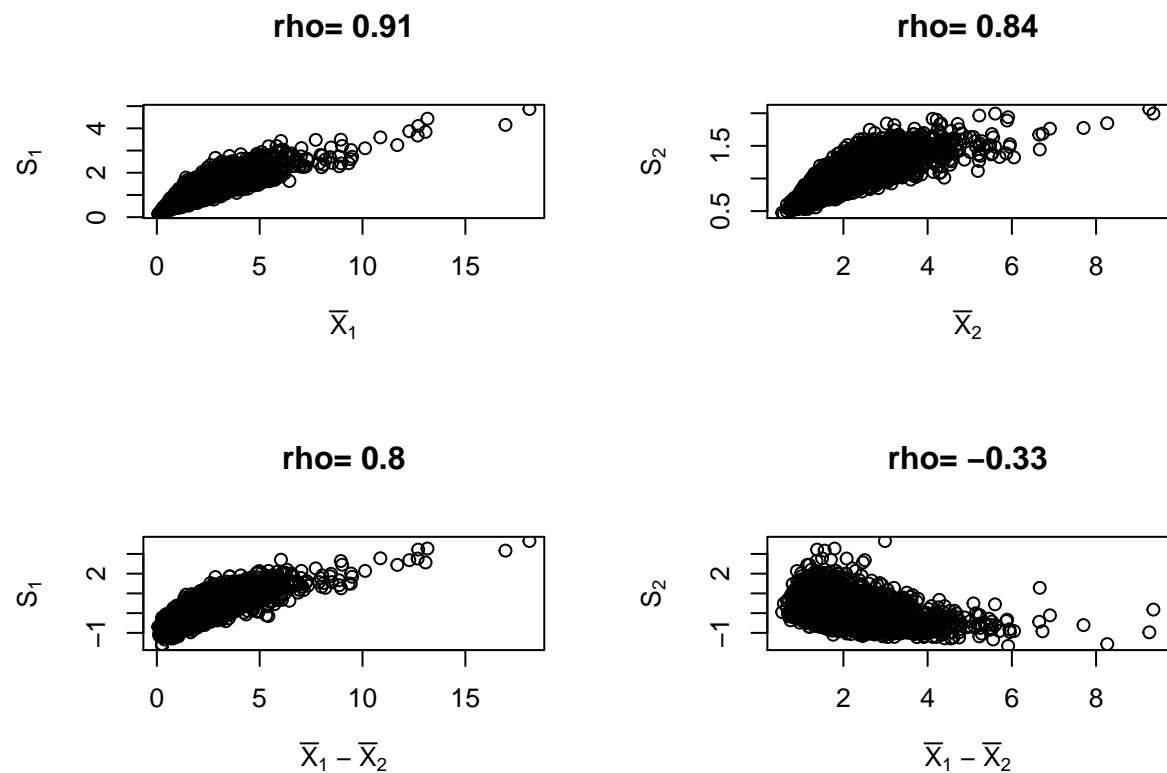


Figure 8.  $S_j$  ( $j=1,2$ ) as a function  $\bar{X}_j$  (top plots) or  $\bar{X}_1 - \bar{X}_2$  (bottom plots), when samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ ), with  $n_1=20$  and  $n_2=100$

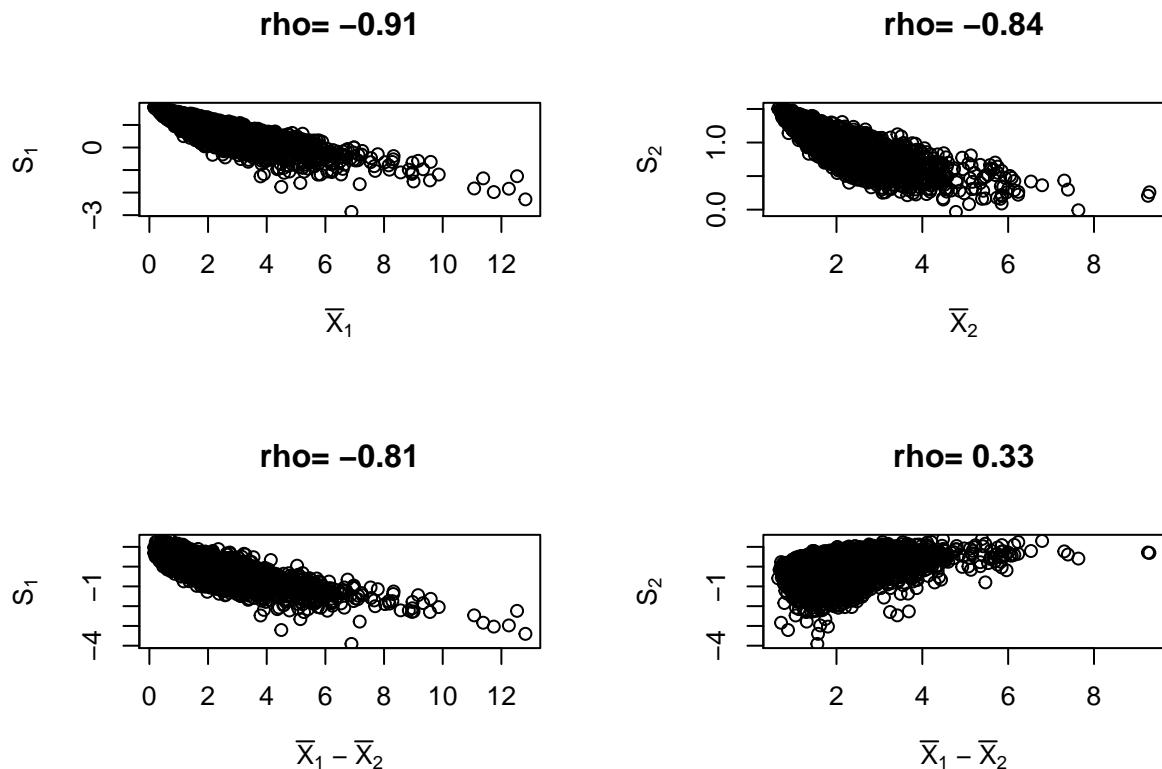


Figure 9.  $S_j$  ( $j=1,2$ ) as a function  $\bar{X}_j$  (top plots) or  $\bar{X}_1 - \bar{X}_2$  (bottom plots), when samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ ), with  $n_1=20$  and  $n_2=100$

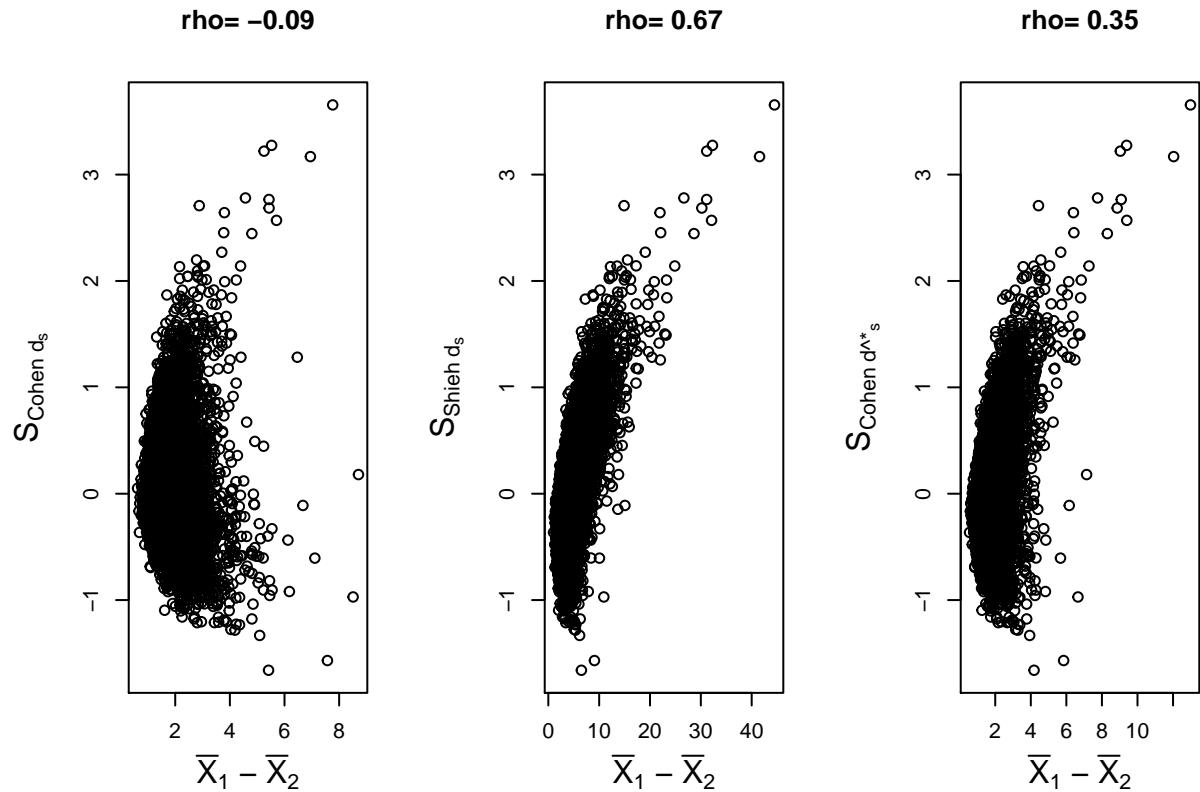


Figure 10.  $S_{Cohen's\ ds}$ ,  $S_{Shieh's\ ds}$  and  $S_{Cohen's\ d^*s}$  as a function of the mean difference ( $\bar{X}_1 - \bar{X}_2$ ), when samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ , with  $n_1=20$  and  $n_2=100$ )

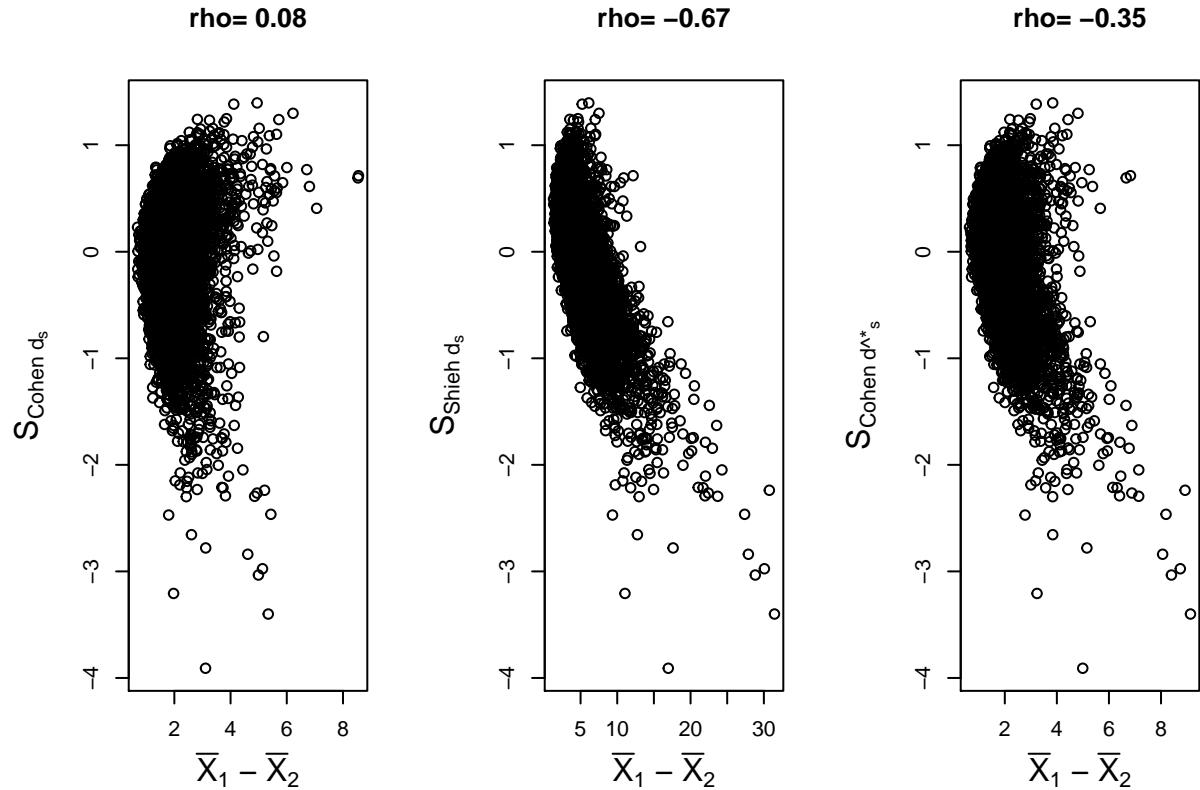


Figure 11.  $S_{Cohen's\ ds}$ ,  $S_{Shieh's\ ds}$  and  $S_{Cohen's\ ds^*}$  as a function of the mean difference ( $\bar{X}_1 - \bar{X}_2$ ), when samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ ), with  $n_1=20$  and  $n_2=100$

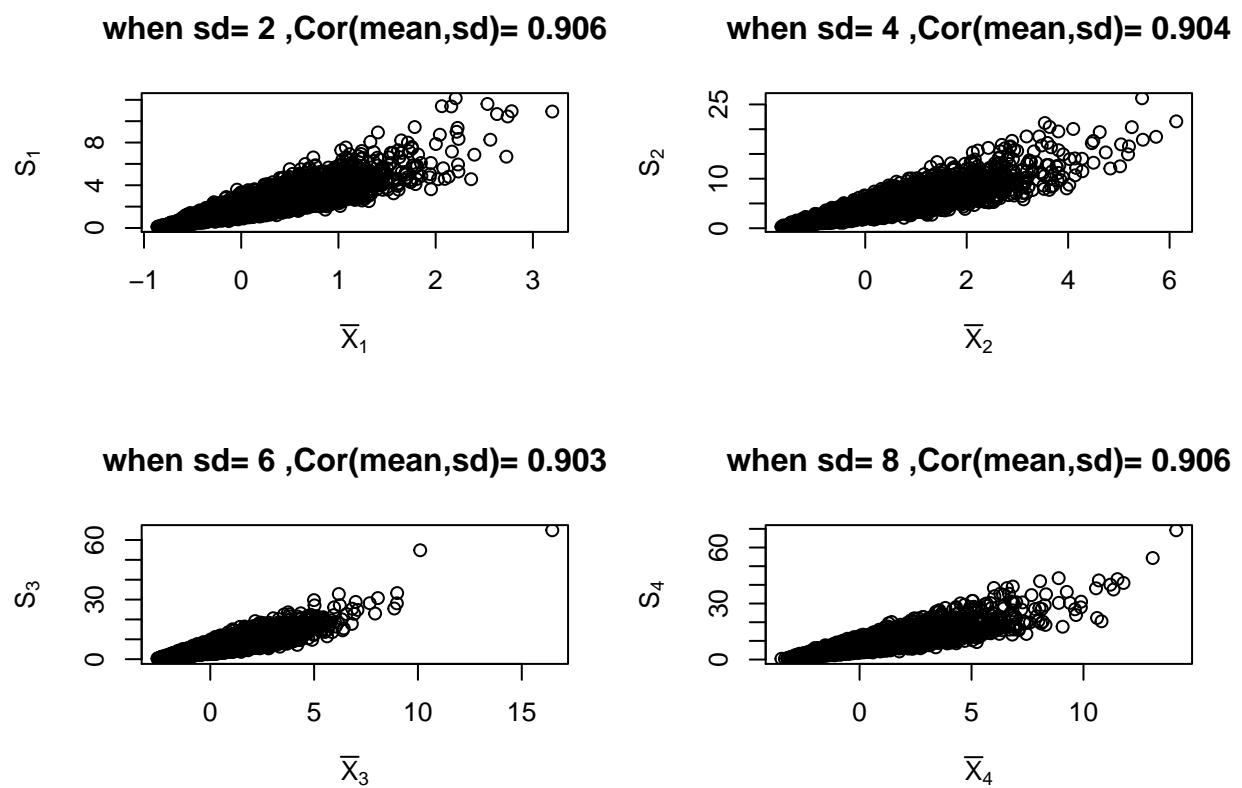


Figure 12. correlation between  $S_j$  and  $\bar{X}_j$  when  $SD = 2, 4, 6$  or  $8$  and samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ )

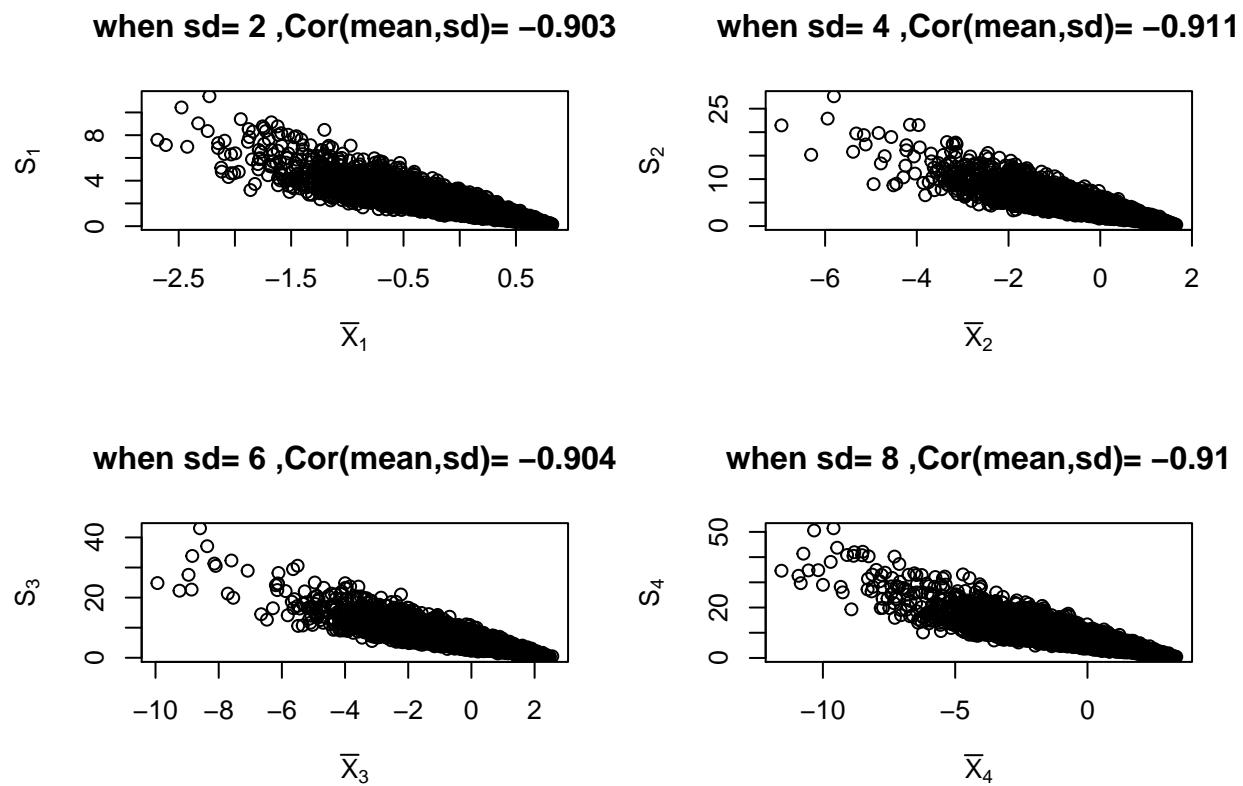


Figure 13. correlation between  $S_j$  and  $\bar{X}_j$  when  $SD = 2, 4, 6$  or  $8$  and samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ )

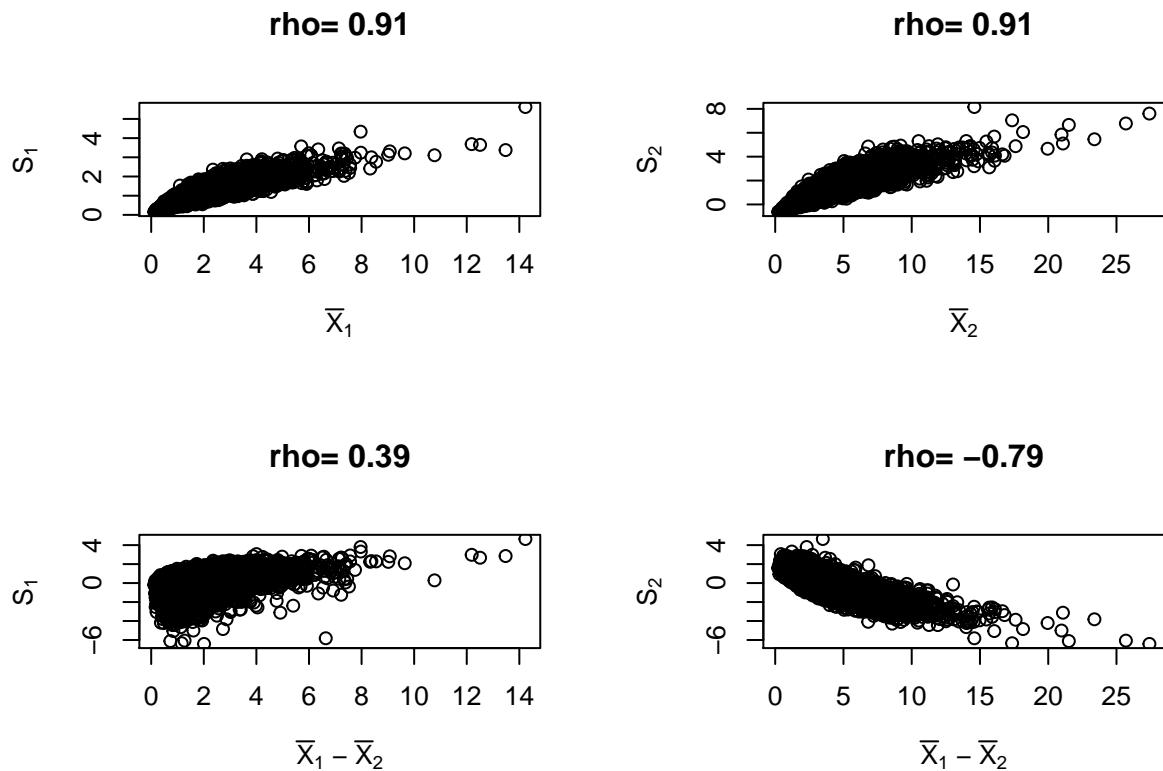


Figure 14.  $S_j$  ( $j=1,2$ ) as a function  $\bar{X}_j$  (top plots) or  $\bar{X}_1 - \bar{X}_2$  (bottom plots), when samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ ), with  $SD_1=2$  and  $SD_2=4$

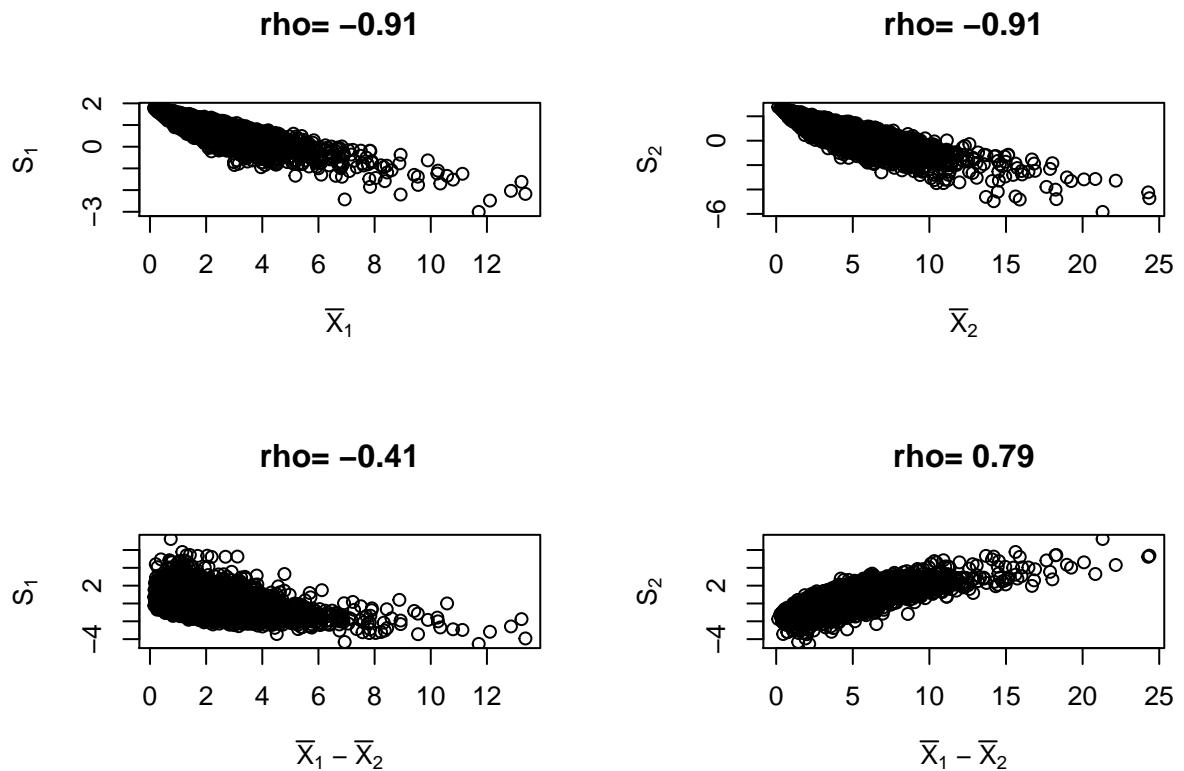


Figure 15.  $S_j$  ( $j=1,2$ ) as a function  $\bar{X}_j$  (top plots) or  $\bar{X}_1 - \bar{X}_2$  (bottom plots), when samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ ), with  $SD_1=2$  and  $SD_2=4$

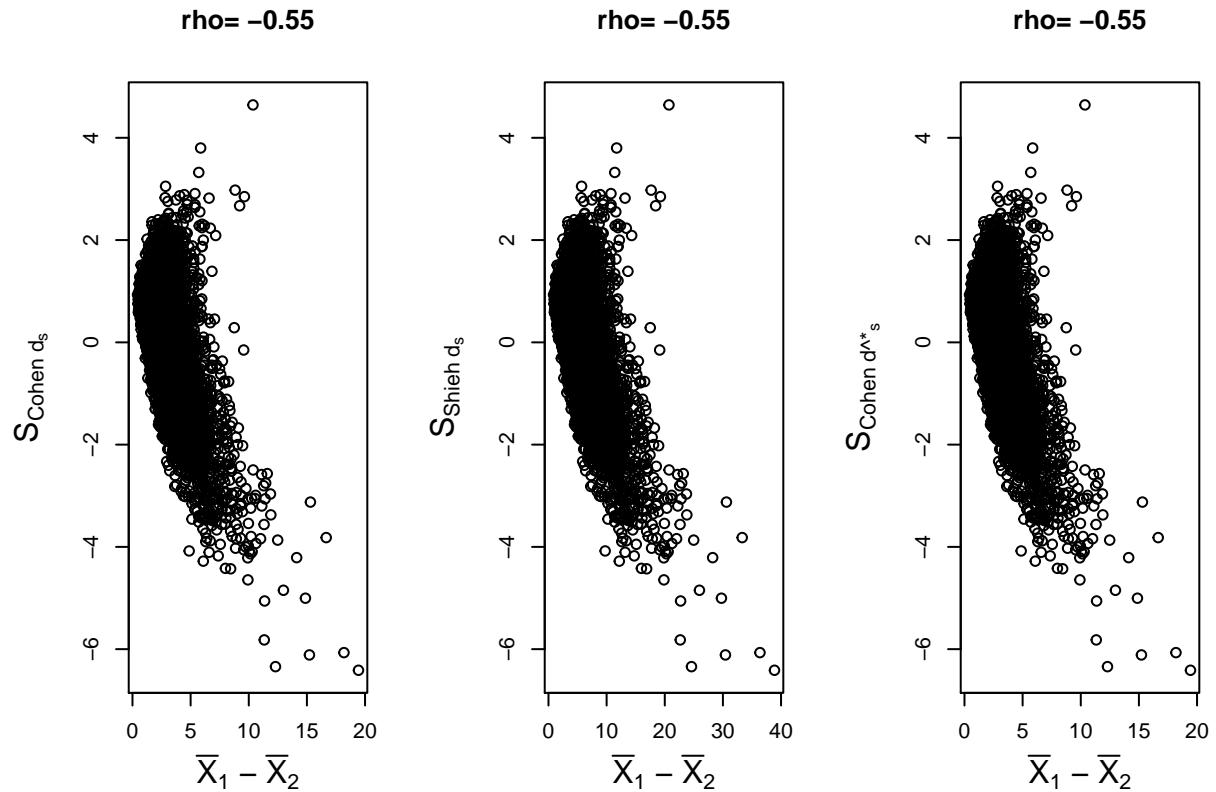


Figure 16.  $S_{Cohen's\ ds}$ ,  $S_{Shieh's\ ds}$  and  $S_{Cohen\ d^*s}$  as a function of the mean difference  $(\bar{X}_1 - \bar{X}_2)$ , when samples are extracted from right skewed distributions ( $\gamma_1 = 6.32$ ), with  $SD_1=2$  and  $SD_2=4$

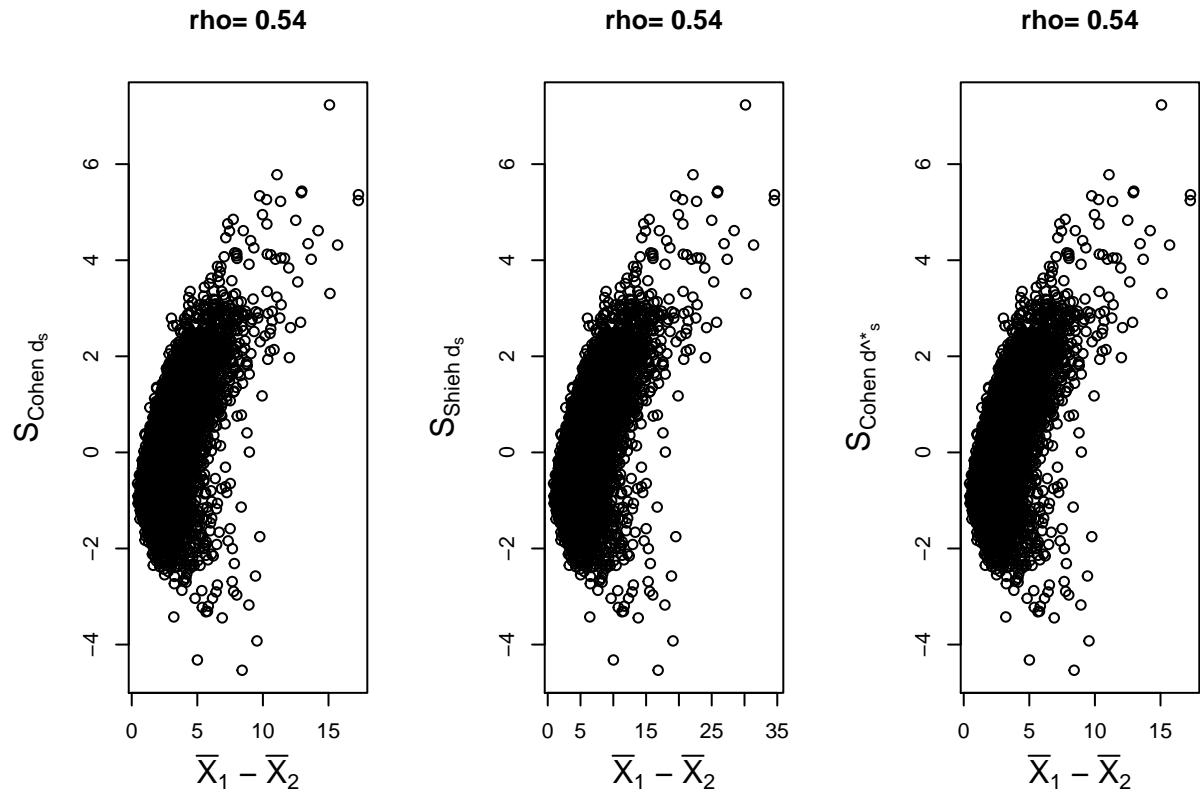


Figure 17.  $S_{Cohen's\,d_s}$ ,  $S_{Shieh's\,d_s}$  and  $S_{Cohen's\,d_s^*}$  as a function of the mean difference  $(\bar{X}_1 - \bar{X}_2)$ , when samples are extracted from left skewed distributions ( $\gamma_1 = -6.32$ ), with  $SD_1=2$  and  $SD_2=4$