Mathematical study of Glass's d

Marie Delacre[1]

[1] Université Libre de Bruxelles, Service of Analysis of the Data (SAD), Bruxelles, Belgium

Author Note

I would like to thank Matt Williams and Thom Baguley for their helpful insights in order to undertand the phenomenon explained in this appendix.

Correspondence concerning this article should be addressed to Marie Delacre, CP191, avenue F.D. Roosevelt 50, 1050 Bruxelles. E-mail: marie.delacre@ulb.ac.be

Abstract

Keywords: keywords

Word count: X

Mathematical study of Glass's d

When two samples are extracted from distributions with identical shapes, with $\sigma_1 = \sigma_2$ and $n_1 = n_2$

When population distributions are symmetric (i.e. $\gamma_1 = 0$), the sampling distribution of glass's d_s is the same, whatever one chooses s_1 or s_2 as standardizer. As an example, in Figure 1, we plotted the sampling distribution of both measures of glass's d_s when two samples of 20 subjects are extracted from two symmetric distributions where $\gamma_1 = 0$,$\gamma_2 = 95.75$, $\sigma_1 = \sigma_2 = 1$ and $\mu_2 = 0$. μ_1 is either 0 or 1, depending on the plot. One can see that in the two plots, distributions of glass's d_S using s_1 and s_2 as standardiser are superimposed. [1]

However, when population distributions are skewed (i.e. $\gamma_1 \neq 0$), the sampling distribution of glass's d_s varies as a function of the chosen standardizer, as illustrated in Figure 2.

It might seem surprising, or even counter-intuitive, as in all plots, s_1 and s_2 are both estimates of the same population standard deviation (σ), based on the same number of observations (as $n_1 = n_2$), but this phenomenon can be mathematically explained. In the following section, we will provides detailed informations to understand the results plotted in Figure 2.

[1] Anytime the mean difference is null, the sampling distribution of glass will be symmetric. On the other side, anytime the mean difference is non null, the sampling distriution of glass's ds will be skewed: right-skewed if the mean difference is positive, left-skewed if the mean difference is negative. The right skewed distribution is therefore only due to the sense of the mean difference

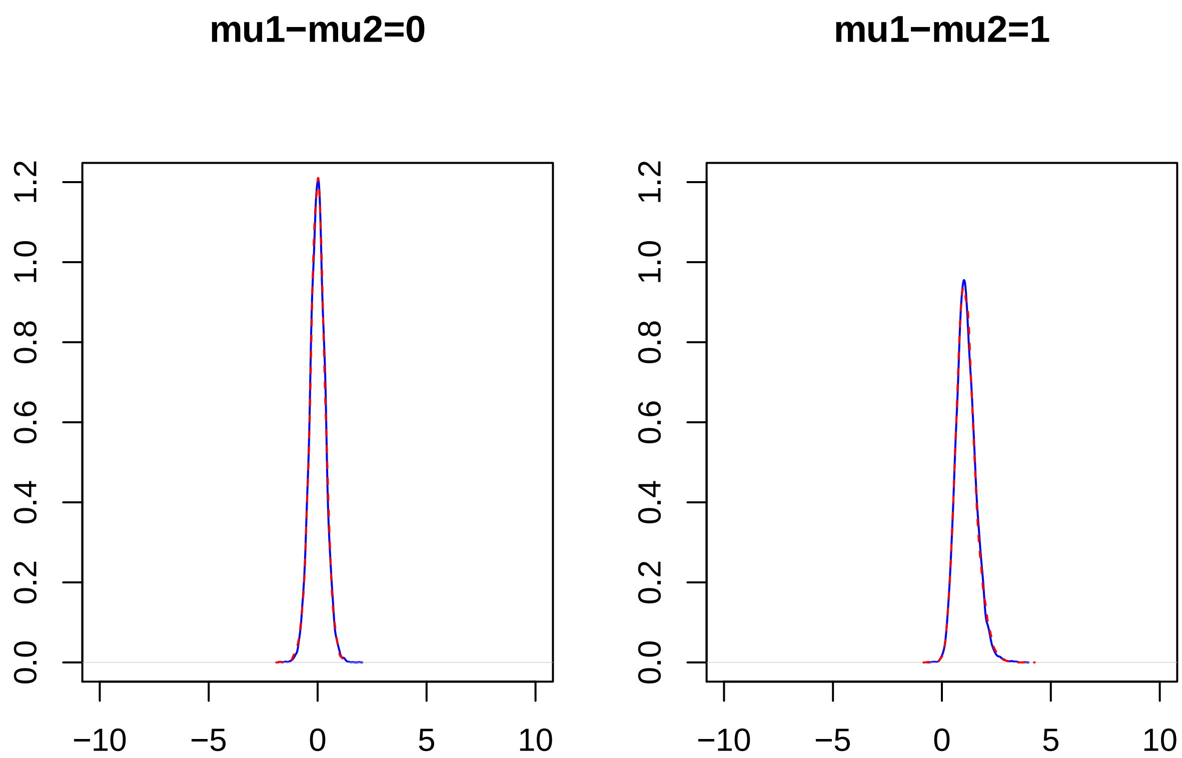


Figure 1. Comparison of Glass's ds when choosing either s1 (blue line) or s2 (red dotted line) as standardizer, with s1=standard deviation of the first sample and s2=standard deviation of the second sample, when n1=n2=20 and both samples are extracted from a distribution where G1 =0, G2=95.75 and sigma=1

When both samples are extracted from a common right-skewed distribution ($\mu_1 - \mu_2 = 0$) (top right plot in Figure 2)

We will first study the configuration where both samples are extracted from a right-skewed distribution where $\mu = 0$, $\sigma = 1$, $\gamma_1 = 6.32$ and $\gamma_2 = 95.75$. Because this distribution is right-skewed, the sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ will also be right-skewed. However, because $\bar{X}_1$ and $\bar{X}_2$ are identically distributed, $\bar{X}_1 - \bar{X}_2$ will follow a symmetric distribution, as illustrated in Figure 3 (right plot). More specifically, the distribution will be symmetrically centered around $\mu_1 - \mu_2 = 0$ (i.e. the green area in the

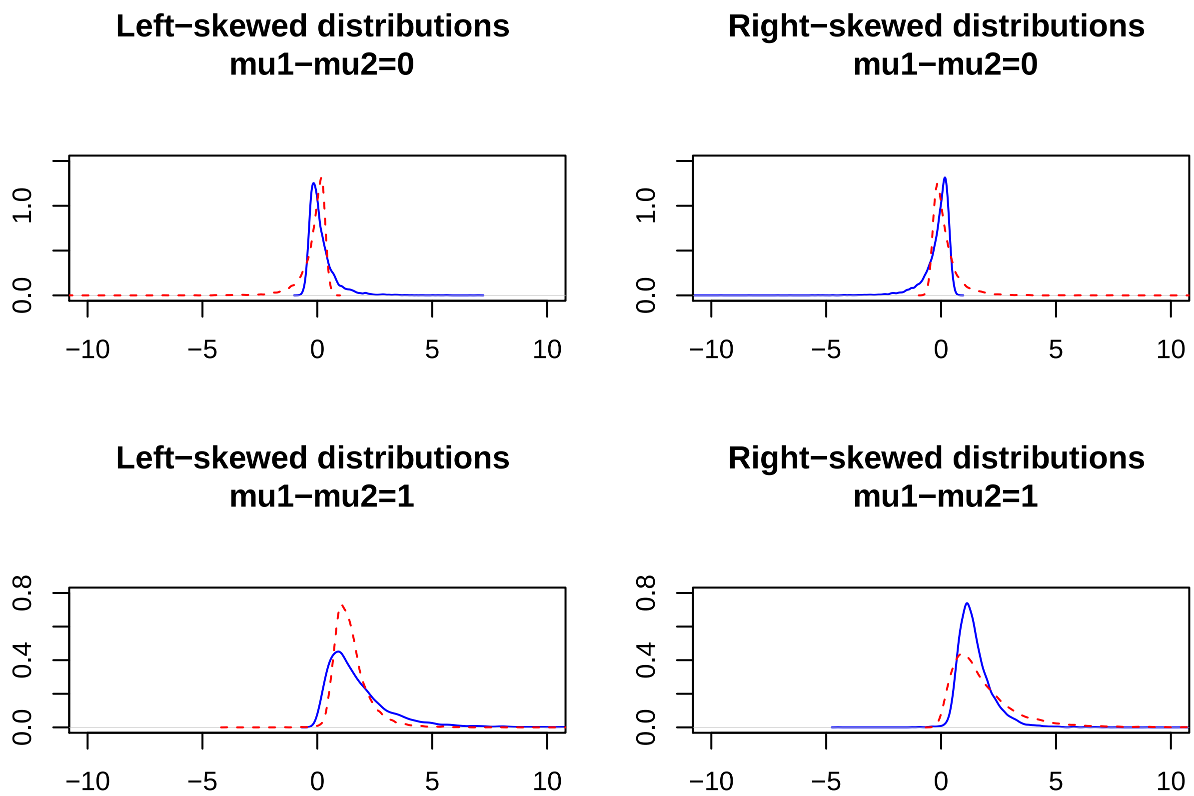


Figure 2. Comparison of Glass's ds when choosing either sd1 (blue line) or sd2 (red dotted line) as standardizer when n1=n2=20 and both samples are extracted from a distribution where sigma=1, G2=95.75, G1 is either -6.32 (left) or 6.32 (right). In all cases, the second sample is extracted from a population distribution where mu2=0. First sample is extracted from a population distribution where mu1 is either 0 (top) of 1 (bottom)

right plot in Figure 3 is the reflexion of the blue area, with the vertical line being the line of reflexion). It means that:

- Half the mean difference estimates will be positive (i.e. $\bar{X}_1 - \bar{X}_2 > 0$; see green area) and the other half will be negative (i.e. $\bar{X}_1 - \bar{X}_2 < 0$; see blue area).

- For a constant k, $|(\mu_1 - \mu_2) - k| = |(\mu_1 - \mu_2) + k|$

Because we compute the mean difference as the mean estimate of the first sample minus the mean estimate of the second sample, there is a positive correlation between $\bar{X}_1$

and $\bar{X}_1 - \bar{X}_2$, and a negative correlation between $\bar{X}_2$ and $\bar{X}_1 - \bar{X}_2$ (correlations would be trivially reversed if we computed $\bar{X}_2 - \bar{X}_1$ instead of $\bar{X}_1 - \bar{X}_2$).

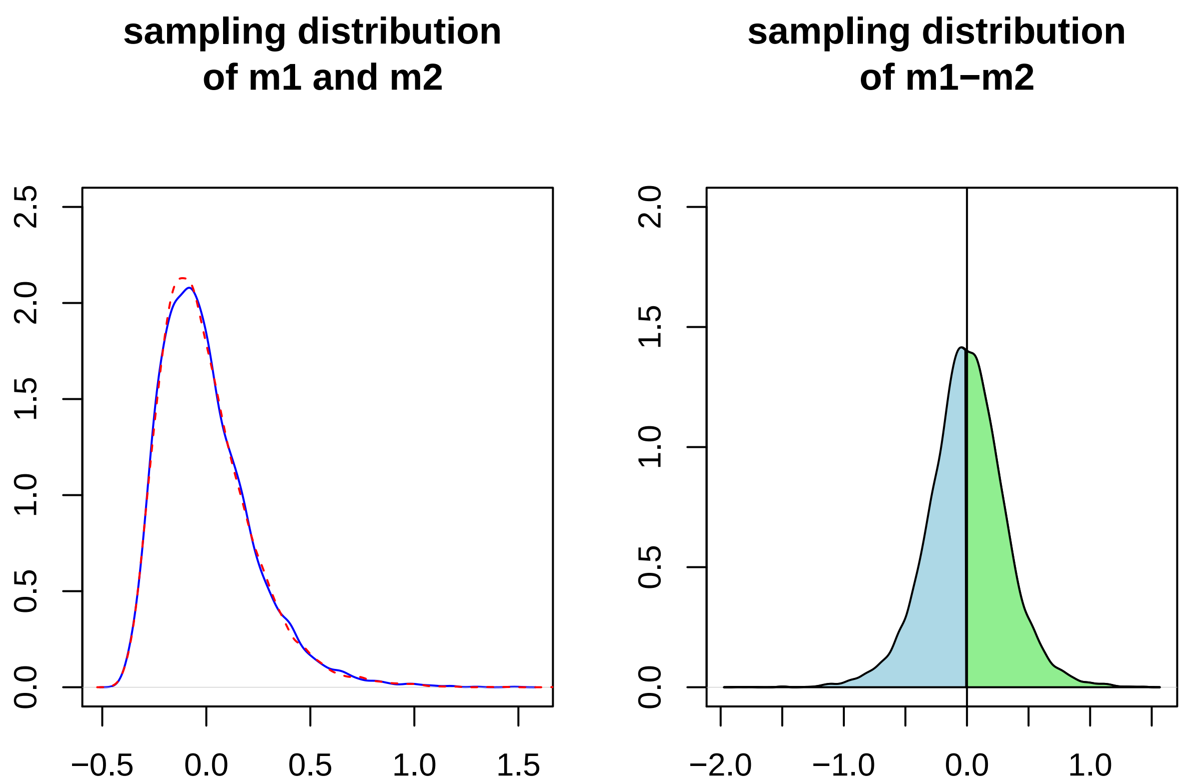


Figure 3. Sampling distribution of m1 (blue line in left plot), m2 (red dotted line in left plot), and m1-m2 (right plot), when m1 and m2 are estimates of the mean of a population distribution where mu=0, sigma=1,G1=6.32 and G2=95.75, with n1=n2=20

The sampling distributions of s_1 and s_2 are right-skewed, because estimates of the standard deviation are bounded: they can be very large, but never below 0. Moreover, as s_1 and s_2 are estimates of the same population standard deviation σ, based on the same sample size, of course, the sampling distributions of s_1 and s_2 will be identical, as illustrated in Figure 4.

Therefore, how to explain the different sampling distributions of glass's d_s, as a function of the standardizer? This is due to the fact that when distributions are skewed,

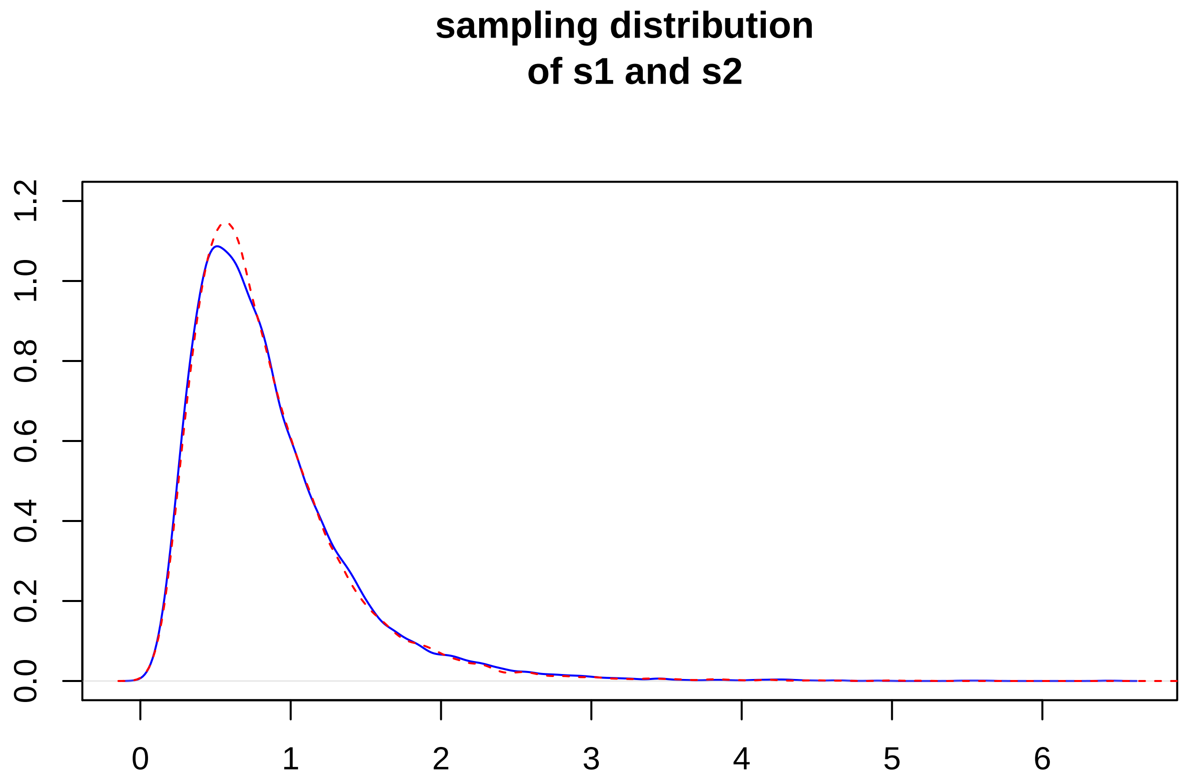


Figure 4. Sampling distribution of s1 (blue line) and s2 (red dotted line), when s1 and s2 are estimates of the standard deviation of a population distribution where mu=0, sigma=1,G1=6.32 and G2=95.75, with n1=n2=20

there is a non-nul correlation between $\bar{X}$ and s (see Zhang, 2007). More specifically, when distributions are right-skewed, there is a **positive** correlation between $\bar{X}$ and s, as illustrated in the left plots in Figure 5.

First, consider the glass's d_s estimate using s_1 as standardiser. We already mentioned that there is a *positive* correlation between $\bar{X}_1$ and $\bar{X}_1 - \bar{X}_2$ ($cor(\bar{X}_1, \bar{X}_1 - \bar{X}_2) > 0$). Because there is also a positive correlation between $\bar{X}_1$ and s_1 ($cor(\bar{X}_1, s_1) > 0$), it results in a **positive** correlation between $\bar{X}_1 - \bar{X}_2$ and s_1 ($cor(\bar{X}_1 - \bar{X}_2, s_1) > 0$; see top right plot in Figure 5): when moving from the left to the right in the right plot in Figure 3, s_1 get larger.

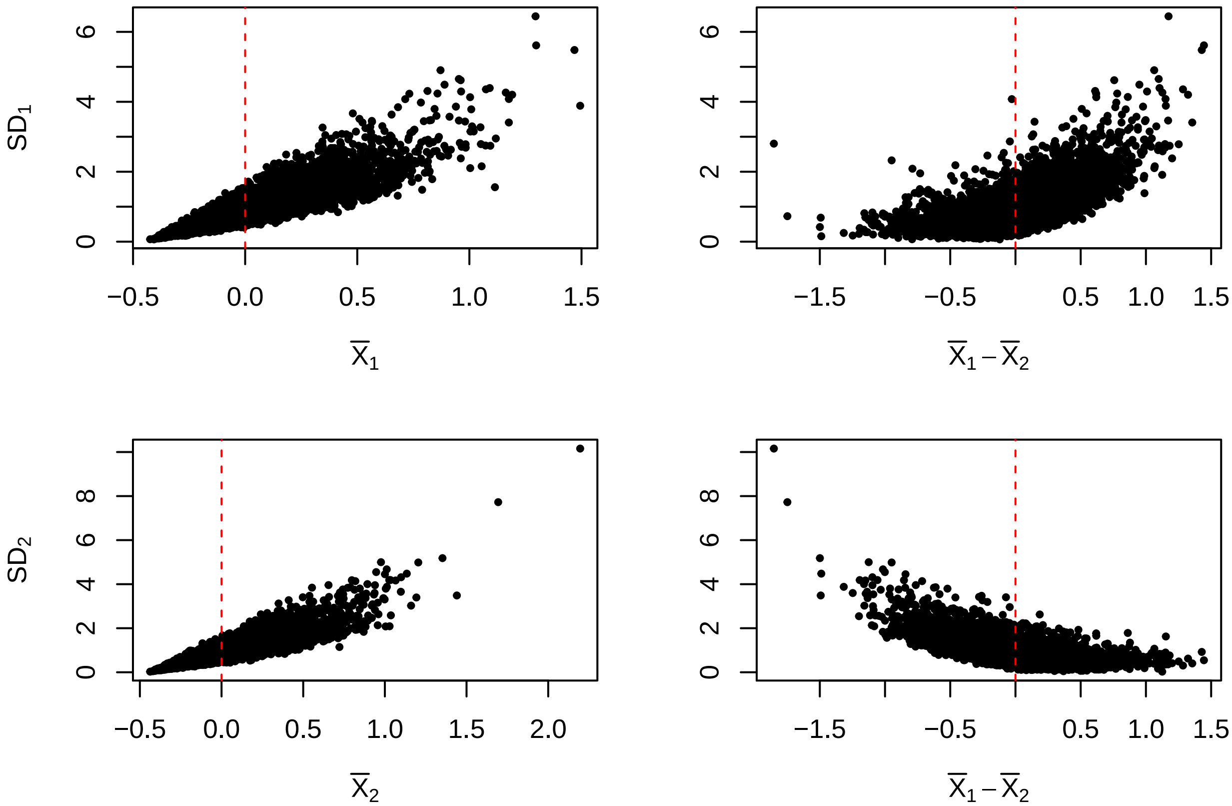


Figure 5. sample mean against sample standard deviation (left) and sample mean difference against sample standard deviation (right), for 100,000 iterations of sampling from two population distributions where mu=0, sigma=1,G1=6.32 and G2=95.75, with n1=n2=20

As a consequence, for many positive numbers k, $-k$ will be divided by a smaller positive value (resulting in a larger ratio) than $+k$, resulting in a left-skewed sampling distribution of glass's d_S.[2] Importantly, while the median of the sampling distribution of glass's d_s is 0, as expected (because the sampling distributions of $\bar{X}_1 - \bar{X}_1$ is centered around 0), the mean will be a little lower (i.e. -0.17), meaning that glass's d_s is negatively biased.

When considering s_2 as standardiser, because there is a *negative* correlation between $\bar{X}_2$

[2] This is not true for each value of k, because the correlation between the mean difference and standard deviation estimates is not perfect, but the larger k (i.e. the more extreme the mean difference estimate), the larger the probability that it is true.

and $\bar{X}_1 - \bar{X}_2$, there is also a **negative** correlation between $\bar{X}_1 - \bar{X}_2$ and s_2 (see bottom right plot in Figure 5): when moving from the left to the right in the right plot in Figure 3, s_2 get lower. In other word, for many positive numbers k, $-k$ will be divided by a larger positive value (resulting in a larger ratio) than $+k$, resulting in a right-skewed sampling distribution of glass's d_S. This time, while the median of the sampling distribution of glass's d_s is still 0, the mean will be a little larger (i.e. 0.17), meaning that glass's d_s is positively biased.

When both samples are extracted from a common left-skewed distribution ($\mu_1 - \mu_2 = 0$) (top left plot in Figure 2)

When distributions are left-skewed, one observes the opposite: there is a **negative** correlation between $\bar{X}$ and s, as illustrated in the left plots in Figure 6.

and therefore, when moving from the left to the right in the right plot in Figure 3, s_1 get lower $(cor(\bar{X}_1, s_1) < 0 \; and \; cor(\bar{X}_1, \bar{X}_1 - \bar{X}_2 > 0) \rightarrow cor(\bar{X}_1 - \bar{X}_2, s_1) < 0)$ and s_2 get larger $(cor(\bar{X}_2, s_2) < 0 \; and \; cor(\bar{X}_2, \bar{X}_1 - \bar{X}_2 < 0) \rightarrow cor(\bar{X}_1 - \bar{X}_2, s_2) > 0)$. As a consequence, glass's d_S will be positively biased when using s_1 as a standardiser, and negatively biased when using s_2 as a standardiser.

When samples are extracted from skewed distributions, with $\mu_1 - \mu_2 = 1$ (bottom plot in Figure 2)

We will first consider the example where both samples are extracted from right-skewed distributions with μ_1 and μ_2 being respectively 1 and 0, and other moments of the population distributions being equal: $\sigma = 1$, $\gamma_1 = 6.32$ and $\gamma_2 = 95.75$ (see bottom right plot in Figure 2).

Of course, the sampling distributions of $\bar{X}_1$ and $\bar{X}_2$ are not superimposed anymore, because $\bar{X}_1$ will be centered around $\mu_1 = 1$, and $\bar{X}_2$ will be centered around $\mu_2 = 0$. However, except for the mean, all other moments of both distributions (i.e. γ_1, γ_2 and σ)

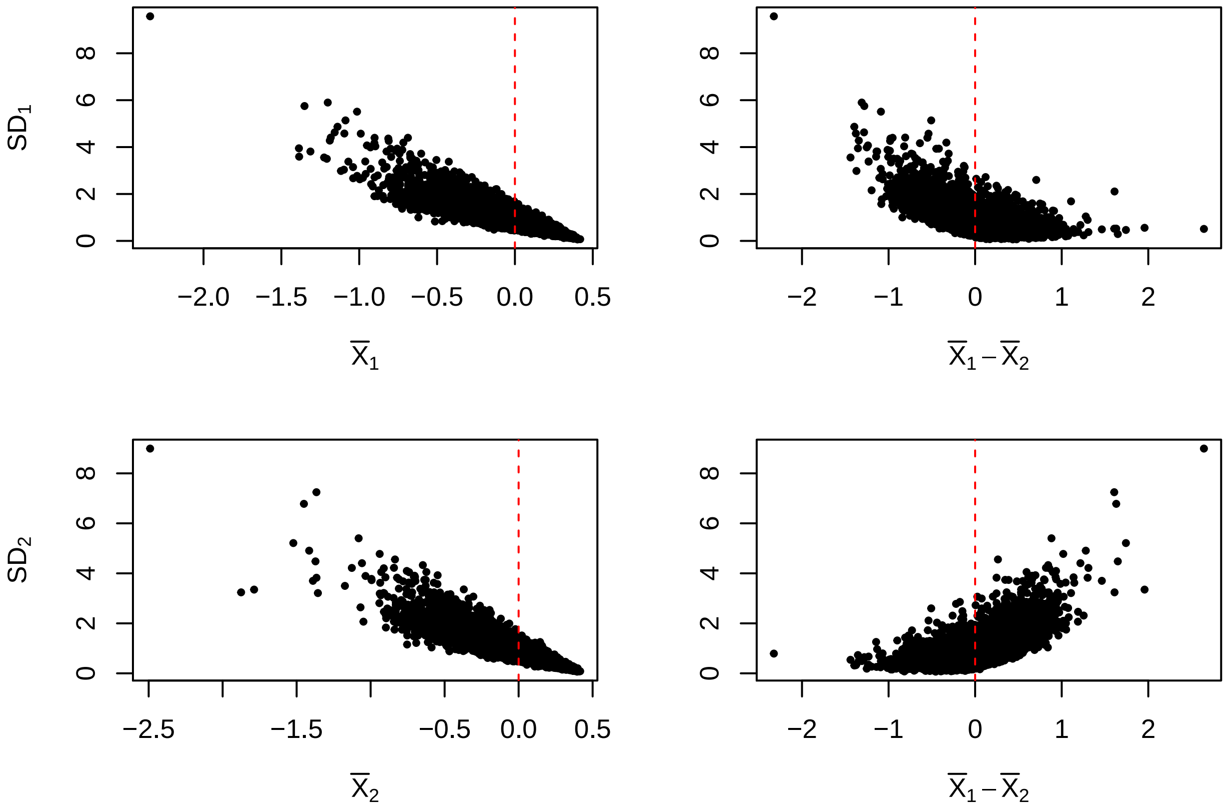


Figure 6. sample mean against sample standard deviation (left) and sample mean difference against sample standard deviation (right), for 100,000 iterations of sampling from two population distributions where mu=0, sigma=1,G1=-6.32 and G2=95.75, with n1=n2=20

remain identical (see left plot in Figure 7) and therefore, the sampling distribution of $\bar{X}_1 - \bar{X}_2$ still follow a symmetric distribution, as illustrated in the right plot in Figure 7.

When $\mu_1 - \mu_2$ was nul, comparing the magnitude of glass's d_S when $\bar{X}_1 - \bar{X}_2 = (\mu_1 - \mu_2) \pm k$ was only a function of the denominator (as $|(\mu_1 - \mu_2) - k| = |(\mu_1 - \mu_2) + k|$). When $\mu_1 - \mu_2 \neq 0$, it is a function of both numerator and denominator. For example, when $\mu_1 - \mu_2 = 1$, only about 0.40% of the mean estimates are negative, meaning that almost all mean difference estimates will be positive (so will be glass's d_s estimates). When computing glass's d_s using s_1 as standardizer, the mean difference estimates that are close of 0 will be divided by a smaller standard deviation

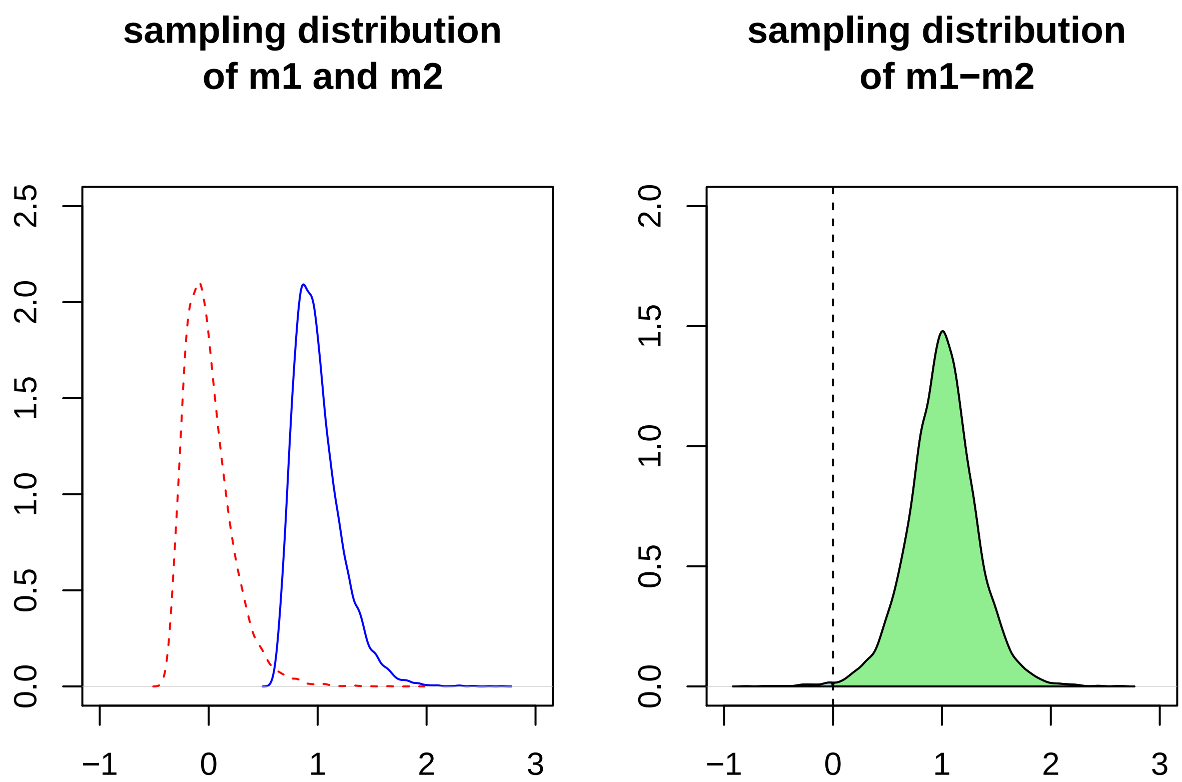


Figure 7. Sampling distribution of m1 (blue line in left plot), m2 (red dotted line in left plot), and m1-m2 (right plot), when m1 is the mean of a sample extracted from a population distribution where mu=1 and m2 is the mean of a sample extracted from a population distribution where mu2=0. Except for the mean, all other moments of both populations distributions are identical, i.e. sigma=1,G1=6.32 and G2=95.75, with n1=n2=20

estimate that larger mean difference estimates (as $cor(\bar{X}_1 - \bar{X}_2, s_1) > 0$). On the other side, when computing glass's d_s using s_2 as standardizer, the mean difference estimates that are very small will be divided by a larger standard deviation estimate than large mean difference estimates (as $cor(\bar{X}_1 - \bar{X}_2, s_2) < 0$). It is therefore not surprising that the sampling distribution of glass's d_s is more skewed and variable when using s_2 rather than s_1 as standardizer.[3]. When distributions are extracted from a left-skewed distribution (bottom left

[3] Of course, when the mean difference is negative, this is the opposite: the sampling distribution of glass's d_s

in Figure 2), this is exactly the opposite.

When two samples are extracted from distributions with identical shapes, and $n_1 \neq n_2$

When population distributions are symmetric (i.e. $\gamma_1 = 0$), the sampling distribution of glass's d_s is only a function of the sample size of the group from which standardizer is computed (because $\sigma_1 = \sigma_2$). Using the SD of the smallest group as standardiser results in a more biased and variable measure of Glass's d_S, whatever it is the first or the second group, as illustrated in Figure 8.

We know that the sampling distribution of $\bar{X}_1 - \bar{X}_2$ is always symmetric, but that the sampling distribution of any standard deviation is right-skewed (because a standard deviation can never be below 0). The lower the sample size, the more skewed the sampling distribution of the standard deviation and the larger it's variance. As a consequence, the more skewed the sampling distribution of glass's d_s).

When distributions are skewed, the skewness of the glass's d_s distribution (and therefore, the bias and variance of glass's d_s) depends on two parameters: the size of the group from which standardizer is computed (as with symmetric distributions), but also the correlation between $\bar{X}_1 - \bar{X}_2$ and s_j. As long as $\mu_1 - \mu_2$ is positive (or negative), the more skewed and variable estimation of glass's d_s will occur when s_j is simultaneously negatively (or positively) correlated with $\bar{X}_1 - \bar{X}_2$ and computed based on the smallest sample size. On the other side, the best estimation (less biased and variable) will occur when s_j is simultaneously positively (or negatively) correlated with $\bar{X}_1 - \bar{X}_2$ and computed based on the largest sample size. This is illustrated in Figure **??** in four plots where samples are extracted from either a right-skewed distribution ($\gamma_1 = 6.32$; right) or a left-skewed distribution ($\gamma_1 = -6.32$; left), and with a positive (top) or negative (bottom) population

---

will be left-skewed, and will be more skewed and variable when using s_1 rather than s_2 as standardizer

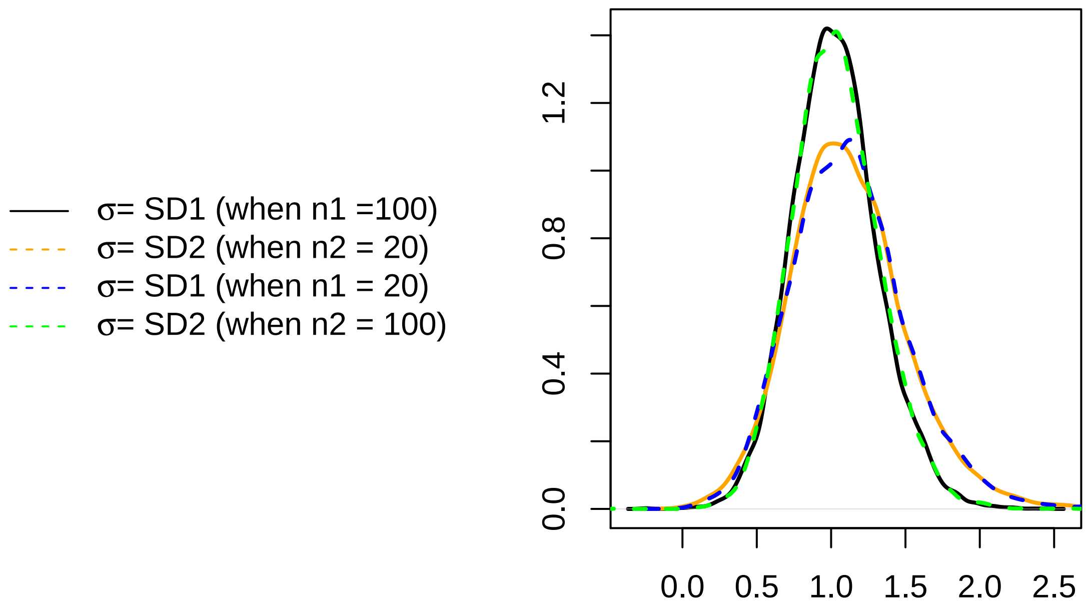
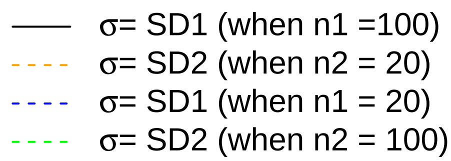
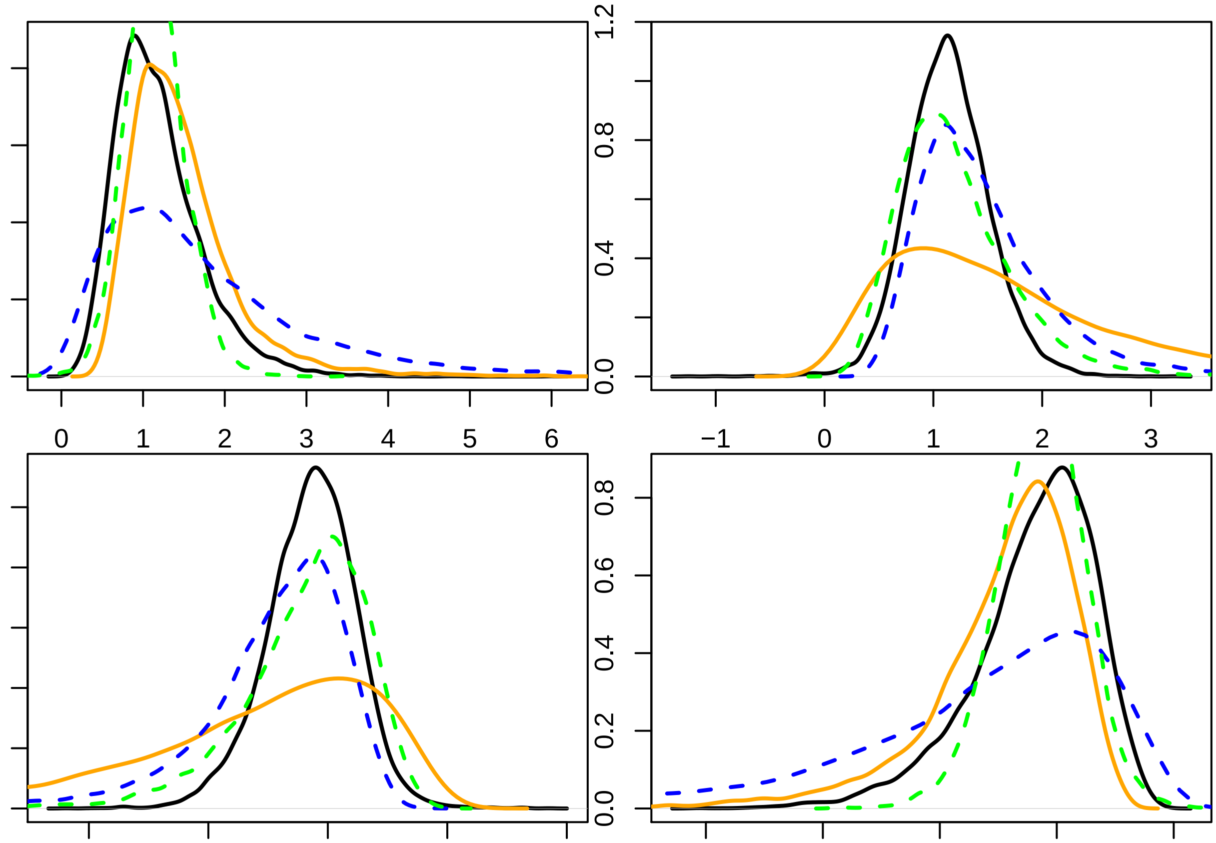


Figure 8. Comparison of Glass's ds when choosing either sd1 or sd2 as standardizer when both samples are extracted from a distribution where sigma=1, m1=1,m2=0, G1=0 and G2=95.75, and either n1=100 and n2=20, or n1=20 and n2=100

132 mean difference.


σ= SD1 (when n1 =100)
σ= SD2 (when n2 = 20)
σ= SD1 (when n1 = 20)
σ= SD2 (when n2 = 100)



In the two top plots in Figure **??**, $\bar{X}_1 - \bar{X}_2$ is positive. When distributions are right-skewed (top right), choosing s_2 (i.e. the SD that is negatively correlated with $\bar{X}_1 - \bar{X}_2$) associated with the smallest sample size (i.e. $n_2 = 20$) will result in the most biased and variable estimation of glass's d_s. The best estimation will occur when choosing s_1 associated with the largest sample size (i.e. $n_1 = 100$). On the other side, when distributions are left-skewed (top left), choosing s_1 (i.e. the SD that is negatively correlated with $\bar{X}_1 - \bar{X}_2$) associated with the smallest sample size (i.e. $n_1 = 20$) will result in the most biased and variable estimation of glass's d_s. The less biased and variable glass d_s will occur when s_2 is chosen as standardizer and $n_2 = 100$.

In the two bottom plots in Figure **??**, $\bar{X}_1 - \bar{X}_2$ is negative. When distributions are right-skewed (bottom right), choosing s_1 (i.e. the SD that is positively correlated with $\bar{X}_1 - \bar{X}_2$) associated with the smallest sample size (i.e. $n_1 = 20$) will result in the most biased and variable estimation of glass's d_s. The best estimation will occur when choosing s_2

associated with the largest sample size (i.e. $n_2 = 100$). On the other side, when distributions are left-skewed (top left), choosing s_2 (i.e. the SD that is positively correlated with $\bar{X}_1 - \bar{X}_2$) associated with the smallest sample size (i.e. $n_2 = 20$) will result in the most biased and variable estimation of glass's d_s. The less biased and variable glass d_s will occur when s_1 is chosen as standardizer and $n_1 = 100$.

Note: when computing an effect size estimates where standardizer is computed based on both SD_1 and SD_2, the correlation between standardizer and $\bar{X}_1 - \bar{X}_2$ is null.

When two samples are extracted from distributions with identical shapes, with $\sigma_1 \neq \sigma_2$ and $n_1 = n_2$

It is interesting to study the sampling distribution of s, $\bar{X}_1 - \bar{X}_2$ and glass's d_s (using either SD_1 or SD_2 as standardiser), as a function of σ, for different kind of distributions, and more specifically, the variance of their sampling distribution.

The larger σ, the larger the standard deviation of $\bar{X}_1 - \bar{X}_2$. In the left plot in Figure 9, we plotted $sd(\bar{X}_1 - \bar{X}_2)$ against σ for different kind of distributions underlying the data. We can see that the standard deviation of the sample mean difference does not depend on the distribution (it is the same for all represented distributions underlying the data). However, it is always a function of σ, as we can see in the right plot in Figure 10, where we plotted $\frac{sd(\bar{X}_1 - \bar{X}_2)}{\sigma}$. This value is constant for any value of σ.

As for the sample mean difference, the larger σ, the larger the standard deviation of the sampling distribution of s. In the left plot in Figure 10, we plotted $sd(s)$ against σ for different kind of distributions underlying the data. We can see that unlike the standard deviation of $\bar{X}_1 - \bar{X}_2$, the standard deviation of S is larger when distributions are symmetric with high kurtosis than when distributions are normal, and that the variance of s is even larger when distributions are skewed. Again, it is also always a function of σ, as we can see in the right plot in Figure 10, where we plotted $\frac{sd(s)}{\sigma}$. This value is constant for any value of σ.

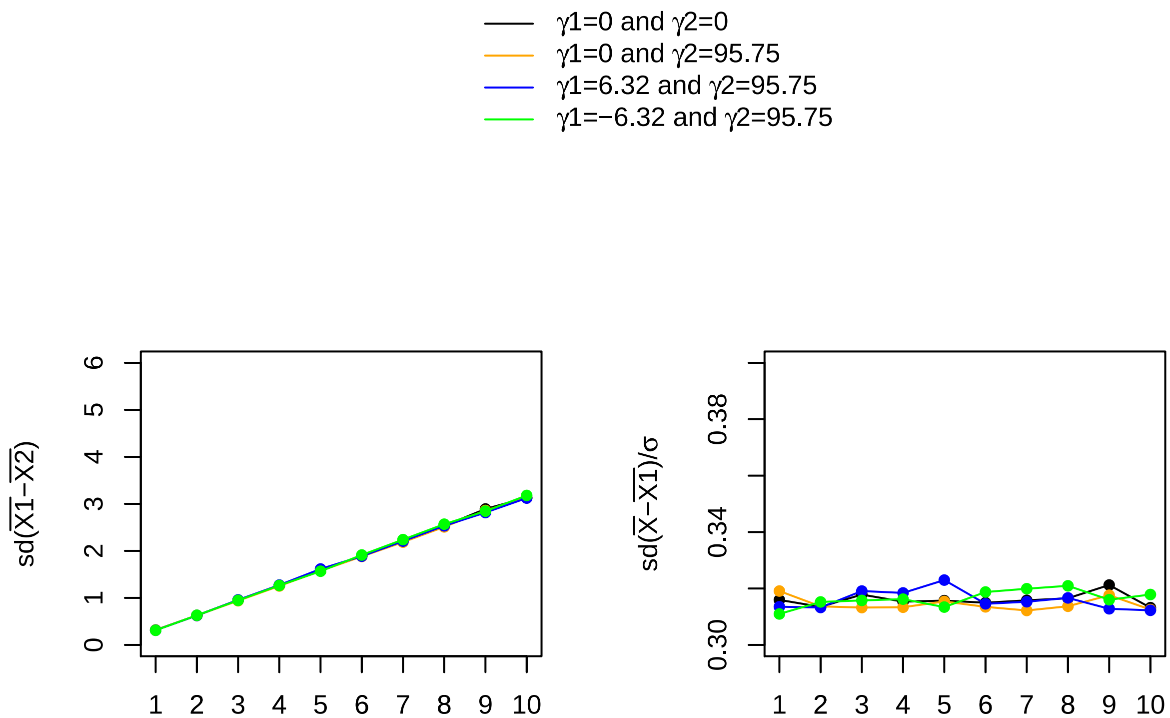


Figure 9. standard deviation of the sampling distribution of s, as a function of sigma, for difference distributions underlying the data

Rem que pour le test, j'avais reproduit le script mais pour voir le sd d'une moyenne d'échantillon, et comme pour la différence de moyenne, ça ne dépend PAS de la distribution (strange, mais c'est comme ça).

- L'écart-type de la distribution d'échantillonnage de la moyenne ne dépend PAS de la distribution (ce n'est qu'une fonction de sigma)
- L'écart-type de la distribution d'échantillonnage de SD dépend d la distribution par contre!

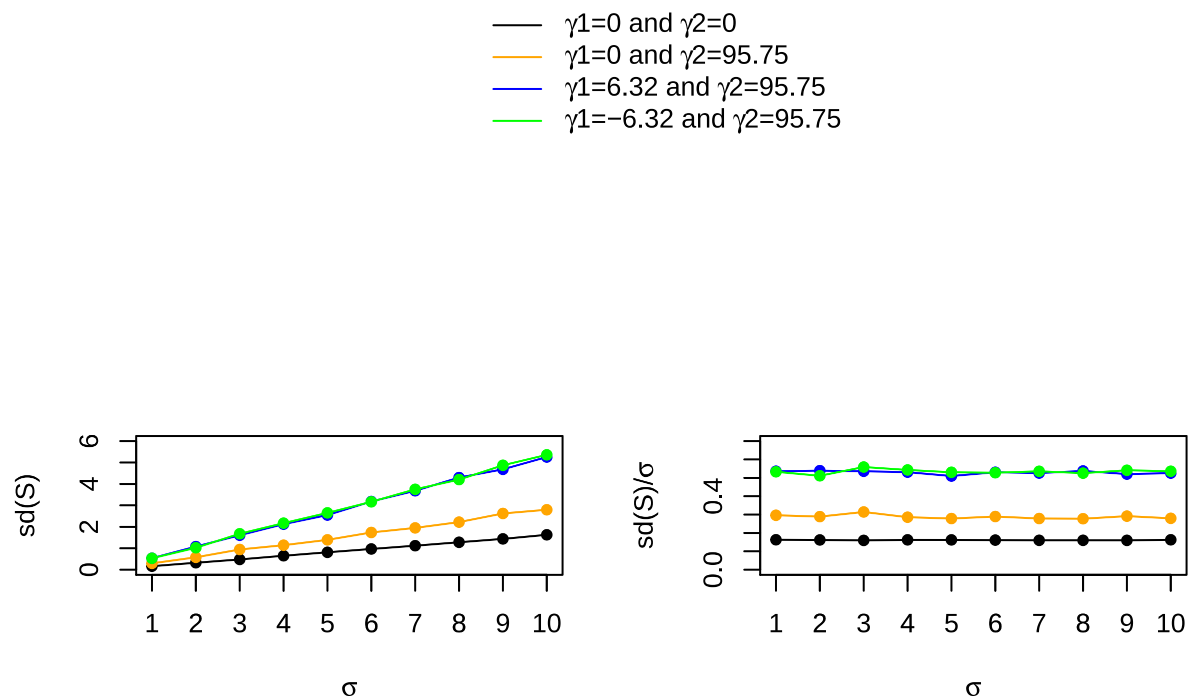


Figure 10. standard deviation of the sampling distribution of s, as a function of sigma, for difference distributions underlying the data

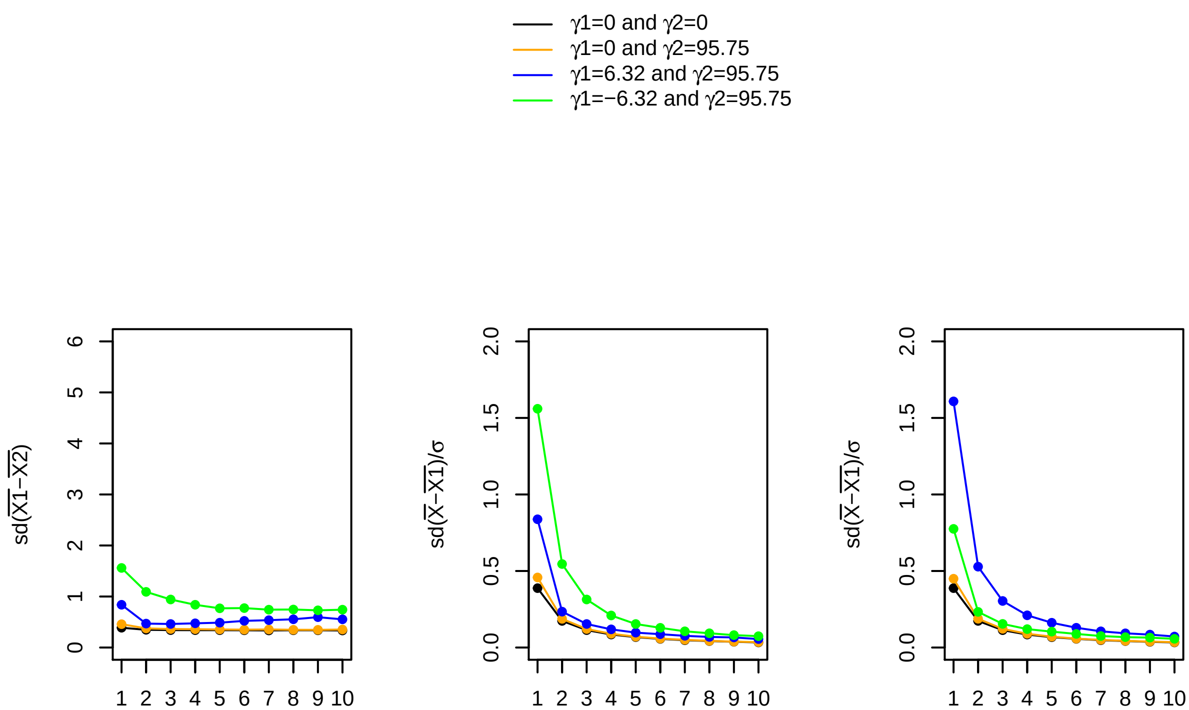


Figure 11. standard deviation of the sampling distribution of s, as a function of sigma, for difference distributions underlying the data

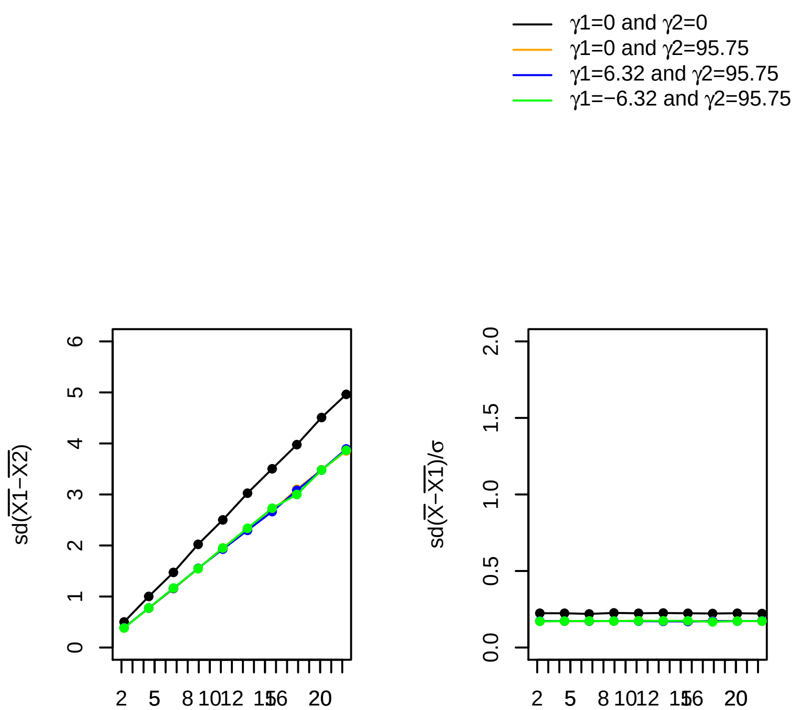


Figure 12. standard deviation of the sampling distribution of s, as a function of sigma, for difference distributions underlying the data