**Noname manuscript No.**
(will be inserted by the editor)

# On the existence of a normal approximation to the distribution of the ratio of two independent normal random variables

**Eloísa Díaz-Francés · Francisco J. Rubio**

Received: date / Accepted: date

**Abstract** The distribution of the ratio of two independent normal random variables $X$ and $Y$ is heavy tailed and has no moments. The shape of its density can be unimodal, bimodal, symmetric, asymmetric, and/or even similar to a normal distribution close to its mode. To our knowledge, conditions for a reasonable normal approximation to the distribution of $Z = X/Y$ have been presented in scientific literature only through simulations and empirical results. A proof of the existence of a proposed normal approximation to the distribution of $Z$, in an interval $I$ centered at $\beta = E(X)/E(Y)$, is given here for the case where both $X$ and $Y$ are independent, have positive means, and their coefficients of variation fulfill some conditions. In addition, a graphical informative way of assessing the closeness of the distribution of a particular ratio $X/Y$ to the proposed normal approximation is suggested by means of a Receiver Operating Characteristic (ROC) curve.

**Mathematics Subject Classification (2000)** MSC 62E17

**Keywords** Coefficient of variation · ratio of normal means · ROC curve.

## 1 Introduction

The distribution of the ratio of independent normal variables $X$ and $Y$ with positive means arises naturally in many scientific areas such as cytometry (Lisák and Doležel, 1998; Sklar, 2005; Watson, 1992), physiology (Kuethe et al., 2000), risk analysis (Hayya et al., 1975), DNA microarrays (Brody et al., 2002), and others. The random variable $Z = X/Y$ has no finite moments and its distribution $F_Z$ is heavy tailed (Hinkley, 1969; Marsaglia, 2006). Both the density and distribution of $Z$ have complicated expressions that are given in Section 2. The shape of $F_Z$

Eloísa Díaz-Francés
Centro de Investigación en Matemáticas (CIMAT), A.P. 402; Guanajuato 36000, México.
Tel.: +52 473 7327155, Fax: +52 473 7325749
E-mail: diazfran@cimat.mx

Francisco J. Rubio
University of Warwick, Department of Statistics, Coventry, CV4 7AL.

can be bimodal, asymmetric, symmetric, and even close to a normal distribution, depending largely on the values of the coefficient of variation of $Y$.

In many applications, the parameter of interest is the ratio of means $\beta = E(X)/E(Y)$ in its own right. Due to the complicated expressions of the density and distribution of $Z$, it has been of interest to approximate $F_Z$ with a normal distribution with location parameter $\beta$ in order to simplify making inferences about $\beta$ (Watson, 1992; Palomino et al., 1999; see Díaz-Francés and Sprott, 2001, for a discussion about this practice in flow cytometry). In addition the issue of having paired normal observations can complicate the separate estimation about $\beta$ due to the large number of additional parameters to be eliminated (see Schneeweiss et al., 1987; Chamberlin and Sprott, 1987). However, the conditions on the parameters of $F_Z$ in order to determine whether a normal distribution can approximate $F_Z$ reasonably well, have only been given empirically after performing simulations or exploratory studies in scientific literature (Marsaglia, 2006; Hayya et al., 1975). Since $F_Z$ is heavy tailed, it is not possible to approximate it reasonably well on the whole real line with a normal distribution. However, we will show in Section 4 that given a normal variable $X$ satisfying some conditions on the parameters, there exists a normal distribution that approximates well $F_Z$, corresponding to a certain $Y$ and $Z = X/Y$ , within an interval centered at $\beta$.

In Sections 5 and 6, some examples and an application that exhibit practical ways of using this result are provided. The goal is to check whether for a given pair of normal variables $X$ and $Y$ the proposed normal approximation to their ratio is reasonable or not. One way of doing this is to compare visually the plots of the density functions of $Z$ and the corresponding normal approximation. The plot of the corresponding ROC curve is quite useful as well for comparing the distributions of Z and the normal approximation.

## 2 The Distribution of $Z$

Consider the case of two independent normal variables $X$ and $Y$ with strictly positive means and variances $(\mu_x, \sigma_x^2)$ and $(\mu_y, \sigma_y^2)$, respectively. The case where their coefficients of variation, $\delta_x = \sigma_x/\mu_x$, $\delta_y = \sigma_y/\mu_y$, are smaller than one will be considered here. The joint distribution of $X$ and $Y$ depends only on four parameters $(\mu_x, \sigma_x, \mu_y, \sigma_y)$ in this case since the correlation between $X$ and $Y$ is zero.

Therefore, the joint density of $Y$ and $Z = X/Y$ can be obtained from that of $X$ and $Y$, by the change of variable theorem, and it will depend as well on the same four parameters. Consider the following one to one convenient reparametrization,

$$(\mu_x, \sigma_x, \mu_y, \sigma_y) \longleftrightarrow (\beta, \rho, \delta_y, \sigma_x),$$

where $\beta = \mu_x/\mu_y$, $\rho = \sigma_y/\sigma_x$. The joint density of $Y$ and $Z$ can be factored as

$$f_{Y,Z}(y, z; \beta, \rho, \delta_y, \sigma_x) = f_Z(z; \beta, \rho, \delta_y)\ f_{Y|Z}(y \mid z; \beta, \rho, \delta_y, \sigma_x),$$

where $f_Z(z; \beta, \rho, \delta_y)$ is the marginal density function of $Z$. This density $f_Z$ depends only on the three identifiable parameters $(\beta, \rho, \delta_y)$ and can be expressed as

$$f_Z\left(z;\beta,\rho,\delta_y\right)=\frac{\rho}{\pi(1+\rho^2z^2)}\exp\left[-\frac{\left(\rho^2\beta^2+1\right)}{2\delta_y^2}\right]\left\{1+\sqrt{\frac{\pi}{2}}q\,\mathrm{erf}\left(\frac{q}{\sqrt{2}}\right)\exp\left(\frac{q^2}{2}\right)\right\}, \tag{1}$$

where

$$q=\frac{\left(1+\beta\rho^2z\right)}{\delta_y\sqrt{1+\rho^2z^2}}.$$

Marsaglia (2006) presented an expression with a similar structure but for the density of $Z^*=\rho Z$.

The following equivalent expression was given by Kuethe et al. (2000) and has been used in physiology applications. It is given here as well in terms of the quantity $q$ for simplicity,

$$\begin{aligned}f_Z\left(z;\beta,\rho,\delta_y\right)=\frac{\rho}{\pi(1+\rho^2z^2)}&\left\{\exp\left[-\frac{\left(\rho^2\beta^2+1\right)}{2\delta_y^2}\right]\right.\\&\left.+\sqrt{\frac{\pi}{2}}q\,\mathrm{erf}\left(\frac{q}{\sqrt{2}}\right)\exp\left[-\frac{\rho^2\left(z-\beta\right)^2}{2\delta_y^2\left(1+\rho^2z^2\right)}\right]\right\}.\end{aligned} \tag{2}$$

The distribution function of $Z$, $F_Z$ can be obtained as in Hinkley (1969), from the bivariate normal cumulative distribution function, or by numerical integration of this density.

The coefficient of variation $\delta_y$ plays the role of a shape parameter of $f_Z$. This parameter also determines the probability that $Y$ takes negative values since

$$F_Z\left(0\right)=P\left[Y\leq 0\right]=\Phi\left(-\frac{1}{\delta_y}\right), \tag{3}$$

where $\Phi$ is the standard normal distribution. In order for this probability to be negligible, say smaller than a given small positive value $h>0$, an upper bound must be imposed on $\delta_y$ such as

$$\delta_y\leq-\Phi^{-1}\left(h\right)^{-1}. \tag{4}$$

Therefore, depending on what is considered to be a negligible probability, the value of $h$ can be selected and through (4) an upper bound for $\delta_y$ is determined. The reverse holds as well, every time an upper bound is assigned to $\delta_y$, a corresponding value of $h$ is implicitly being set. As examples, if $\delta_y\leq 0.43$, this bound corresponds to $h=0.01$ and $\delta_y\leq 0.32$ to $h=0.001$.

The cases where $F_Z$ can be close to a normal distribution have been established empirically or through simulations in several works (Merrill, 1928; Marsaglia, 1965; Kuethe et al., 2000). Kuethe et al. (2000) mentioned that normal approximations to $Z$ are good whenever $\delta_y\leq 0.1$ (this corresponds to a probability of observing negative values of $Y$ smaller or equal to $h=7.6\times10^{-24}$). Marsaglia (2006) provides as a practical rule that if $a<2.256$ and $b>4$, then $Z$ is approximately

normally distributed, where $a = \delta_x^{-1}$ and $b = \delta_y^{-1}$; Marsaglia's rule is equivalent to requiring that $\delta_x > 0.443$ and $\delta_y < 0.25$ (with corresponding $h = 3.2 \times 10^{-5}$). Marsaglia actually considers a slightly different ratio $Z^*$ of normal variables, but the relationship with the ratio considered here is that $Z^* = \rho Z$, since we are assuming here that $X$ and $Y$ are not correlated. Hayya et al. (1975) give the rule of thumb that if $\delta_x \geq 0.19$, and $\delta_y \leq 0.09$ (with associated $h = 5.5 \times 10^{-29}$), then that $F_Z$ is close to a normal distribution. Geary (1930) stated that if $\delta_y \leq 1/3$, $(h = 0.0013)$, then the normal approximation to the distribution of $Z^*$ (and consequently of $Z$) is reasonable. Though these results establish different bounds for the coefficients of variation, all of them coincide in requiring that $\delta_y$ should be sufficiently small.

The fact that the coefficient of variation $\delta_y$ determines the probability of observing negative values of $Y$, as stated in (3), has an important effect on the distribution of $Z$. Note that

$$\lim_{\delta_y \to 0} P\left[Y \leq 0\right] = \lim_{\delta_y \to 0} \Phi\left(-\frac{1}{\delta_y}\right) = 0. \tag{5}$$

The following result presented by Hinkley (1969) helps to understand why the magnitude of $\delta_y$ plays such an important role in determining the shape of the distribution $F_Z$. Hinkley defined the following function of $z$,

$$F^*\left(z\right) = \Phi\left(\frac{z\mu_y - \mu_x}{\sqrt{\sigma_x^2 + z^2\sigma_y^2}}\right) = \Phi\left(\frac{z - \beta}{\delta_y\sqrt{\rho^{-2} + z^2}}\right) = \Phi\left(\frac{z - \beta}{\sqrt{\delta_x^2\beta^2 + z^2\delta_y^2}}\right), \tag{6}$$

and noted that $F_Z\left(z\right)$ could be expressed as the sum of $F^*\left(z\right)$ plus other terms that involved the probability of $Y$ being negative and that were negligible when $\delta_y$ was sufficiently small. Thus he obtained an upper bound for the absolute difference of $F_Z$ and $F^*$ in terms of the coefficient of variation of $Y$,

$$\left|F_Z\left(z;\beta,\rho,\delta_y\right) - F^*\left(z\right)\right| \leq \Phi\left(-\frac{1}{\delta_y}\right).$$

Finally he proved that $F^*\left(z\right)$ converges uniformly to $F_Z\left(z\right)$ whenever the coefficient of variation $\delta_y$ tends to zero. However note that $F^*\left(z\right)$ is not normal. Moreover it is not even a distribution since $F^*\left(-\infty\right) \neq 0$ and $F^*\left(\infty\right) \neq 1$ if $\delta_y > 0$. The normal approximation to $F_Z$ presented in the following section overcomes these deficiencies. Nevertheless, Hinkley's result is crucial for the proof of the existence of a normal approximation to $F_Z$ that will be stated as the theorem of Section 4.

## 3 The Proposed Normal Approximation to $F_Z$

Both reparametrizations $(\beta, \delta_x, \delta_y)$ and $(\beta, \rho, \delta_y)$ will be used for $F_Z$, indistinctly from here on, for convenience. The first parametrization involves both coefficients of variation explicitly, and as mentioned, these parameters have been used in literature to determine when a normal approximation to $F_Z$ might be reasonable.

However, the second parametrization permits to give algebraic expressions in a simpler way.

The normal approximation that is being proposed here was obtained after considering the Taylor series expansion of the function $h(X,Y) = X/Y = Z$ of the ratio of two independent normal random variables $X$ and $Y$ with positive means about the point $(\mu_x, \mu_y)$. Considering only up to the first order term of the expansion and taking expectation on both sides of this approximation, the expected value of $Z$, if it were finite, could be approximated with

$$\beta = \frac{\mu_x}{\mu_y}. \tag{7}$$

In a similar way, a second order approximation to the variance of $Z$, if it were finite, could be

$$\sigma_z^2 = \beta^2 \left(\delta_x^2 + \delta_y^2\right) = \delta_y^2 \left(\rho^{-2} + \beta^2\right). \tag{8}$$

So a natural proposal for a normal approximation to $F_Z$ is the following normal distribution function with mean $\beta$, variance $\sigma_z^2$, and distribution

$$G(z; \beta, \rho, \delta_y) = \Phi\left(\frac{z-\beta}{\beta\sqrt{\delta_x^2 + \delta_y^2}}\right) = \Phi\left(\frac{z-\beta}{\delta_y\sqrt{\rho^{-2} + \beta^2}}\right). \tag{9}$$

The distribution $G$ is proper in contrast to Hinkley's $F^*(w)$. However, note the similarity of the arguments of $\Phi$ in both expressions, the difference being only in the denominator; $G$ considers $\beta^2$ and $F^*$ involves $z^2$.

## 4 Existence of a Normal Approximation to $F_Z$

**Theorem 1** *Let $X$ be a normal random variable with positive mean $\mu_x$, variance $\sigma_x^2$ and coefficient of variation $\delta_x = \sigma_x/\mu_x$ such that $0 < \delta_x < \lambda \leq 1$, where $\lambda$ is a known constant. For every $\varepsilon > 0$, there exists $\gamma(\varepsilon) \in \left(0, \sqrt{\lambda^2 - \delta_x^2}\right)$ and also a normal random variable $Y$ independent of $X$, with positive mean $\mu_y$, variance $\sigma_y^2$ and coefficient of variation $\delta_y = \sigma_y/\mu_y$ that satisfy the conditions,*

$$0 < \delta_y \leq \gamma(\varepsilon) \leq \sqrt{\lambda^2 - \delta_x^2} < \lambda, \tag{10}$$

*for which the following result holds.*

*Any $z$ that belongs to the interval*

$$I = \left[\beta - \frac{\sigma_z}{\lambda}, \beta + \frac{\sigma_z}{\lambda}\right], \tag{11}$$

*where $\beta = \mu_x/\mu_y$, and $\sigma_z = \beta\sqrt{\delta_x^2 + \delta_y^2} = \delta_y\sqrt{\rho^{-2} + \beta^2}$, satisfies that*

$$|G(z) - F_Z(z)| < \varepsilon, \tag{12}$$

*where $G(z)$ is the distribution function of a normal random variable with mean $\beta$, variance $\sigma_z^2$, as given in (9), and $F_Z$ is the distribution function of $Z = X/Y$.*

Note that the restriction $\delta_y < \sqrt{\lambda^2 - \delta_x^2}$, which is equivalent to $\lambda^{-1}\sqrt{\delta_y^2 + \delta_x^2} < 1$, guarantees that the left endpoint of the interval $I$ and all points within it are always positive. That is, the normal approximation holds only for positive values of $z$.

Figure 1 exhibits on the parameter space of the coefficients of variation $(\delta_x, \delta_y)$ the segment of the line where the $\delta_y$ of the random variable $Y$ in Theorem 1 could be, for a given $\delta_x^*$. Note that once a given $Y$ fulfills the closeness between the corresponding $G$ to $F_Z$, any other $Y^*$ with smaller coefficient of variation will satisfy this result too. The proof of Theorem 1 is given in the Appendix.

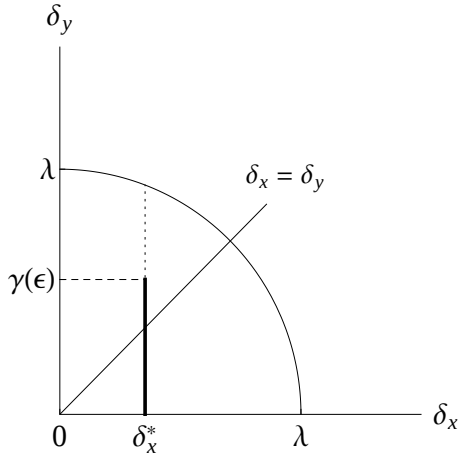



**Fig. 1** The thick vertical black line indicates pairs of values $(\delta_x, \delta_y)$ for which the normal approximation is close to $F_Z$ within the interval $I$, for a given $\delta_x^*$.

## 5 Some Examples

A very useful graphical way of comparing any pair of distributions $F$ and $G$ is provided by the Relative Operating Characteristic (ROC) curve, which as presented in Kotz et al. (2003) consists of the points $[G(z), F(z)]$ for $z \in \mathbb{R}$. Equivalently, the ROC curve is the plot of the function $ROC(t) = F\left[G^{-1}(t)\right]$ for $t \in [0,1]$. The ROC curve lies within the unit square, $[0,1] \times [0,1]$. Closeness to the 45 degree line, $F(z) = G(z)$, indicates that the distributions being compared are close to each other, and viceversa, departures from this line indicate differences between the distributions. A pair of lines at a given distance from the equality line were marked in the plots as a reference.

The additional plots of the densities and distributions of the normal approximation and of the given $Z$ are very informative too. These plots can show for a given pair of variables $X$, $Y$, and values $(\varepsilon, \lambda)$, according to Theorem 1, whether the proposed normal approximation is close or not to the actual distribution of $Z$. Note that the length of the interval $I$ where the approximation stated in Theorem 1 holds depends inversely on the size of $\lambda$.

Five examples are given here where the involved parameters are set to certain fixed values with the purpose of illustrating several interesting cases that can arise. Different values of $\delta_x$ and $\delta_y$ were selected for the same values of $\lambda = 0.5$, $\varepsilon = 0.03$, and $\beta = 2$. For each case, it will be checked whether the proposed normal approximation, corresponding to those fixed values of the parameters, is close or not to $F_Z$.

In Table 1, the corresponding values of $a = 1/\delta_x$ and $b = 1/\delta_y$ are given as well; these are used in Marsaglia's empirical rule (2006; see his Figure 4) for describing situations where it would not be possible to find a reasonable normal approximation to $F_Z$. Here, the first three examples show for the same $\delta_x = 0.2$, cases where the normal approximation to $F_Z$ is bad, reasonable, and very good, according to the corresponding value of $\delta_y$. Figure 2 shows the plots of the densities, the distributions, and the ROC curve of $F_Z$ and $G$ for Example 1. Figures 3 to 6 show only, for brevity, the plots of densities and the ROC curves for the remaining examples. A band of width 0.03 is marked in dashes above and below the solid line $F_Z = G$ in the ROC curve plots. The midpoint as well as the endpoints of interval $I$ are marked with asterisks over the ROC curve. The endpoints of $I$ are also marked over the normal approximation density on all plots.

Examples 2 and 3 contradict Marsaglia's empirical rule which would have predicted that the normal approximation would not hold. In contrast it is shown here that the proposed normal approximation is quite close to $F_Z$, within the interval $I$. Note that for this $X$, any other $Y$ with $\delta_y \leq 0.15$ also yields a good normal approximation $G$.

Example 4 shows a case where the approximation is bad even if Geary's rule of thumb holds (that $\delta_y \leq 1/3$ for a good normal approximation); see Figure 5. The associated $Y$ for such an $X$ would need to have $\delta_y \leq 0.19$ for a reasonable normal approximation, as shown in Example 5 and Figure 6.

Overall, for $\varepsilon = 0.03$ it seems that the recommendation given in Kuethe et al. (2000) that $\delta_y < 0.1$ for a good normal approximation, seems to work well for the values of $\delta_x$ that we explored. However for smaller values of $\varepsilon$, a smaller upper bound for $\delta_y$ might be required.

| Example | $\delta_x$ | $\delta_y$ | $a$ | $b$ | Normal Approximation |
|---|---|---|---|---|---|
| 1 | 0.2 | 0.3 | 5 | 3.3 | Bad |
| 2 | 0.2 | 0.15 | 5 | 6.7 | Reasonable |
| 3 | 0.2 | 0.05 | 5 | 20 | Good |
| 4 | 0.49 | 0.25 | 2.04 | 4 | Bad |
| 5 | 0.49 | 0.19 | 2.04 | 5.26 | Reasonable |

Table 1. Examples of normal approximations to $F_Z$.

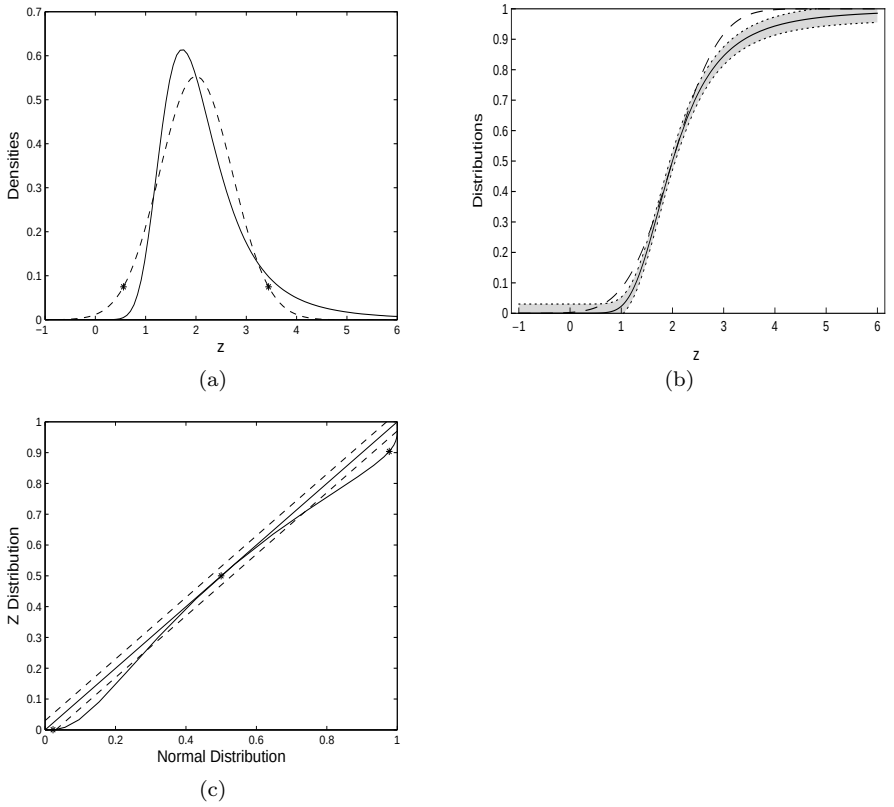



**Fig. 2** Example 1. A case of a bad normal approximation (dashes) to $F_Z$ (solid line); (a) Density functions; (b) Distribution functions, $\pm 0.03$ band in dots; (c) ROC curve (solid), $\pm 0.03$ band in dashes.

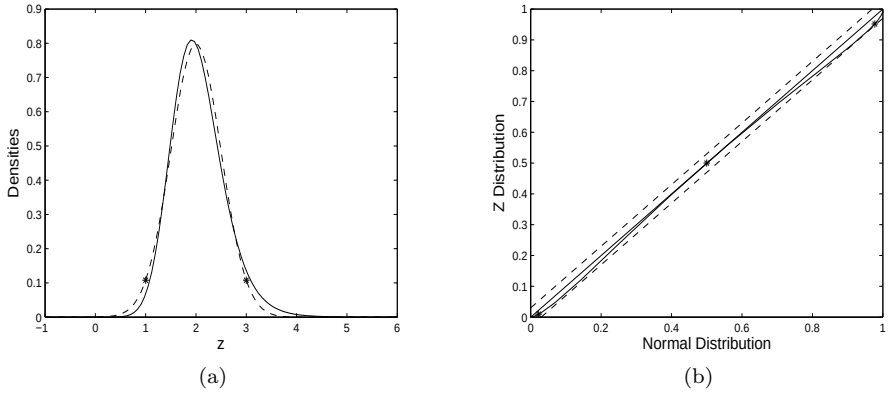
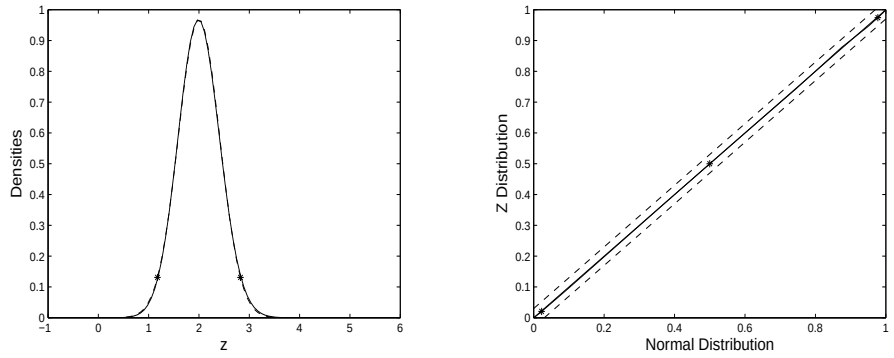
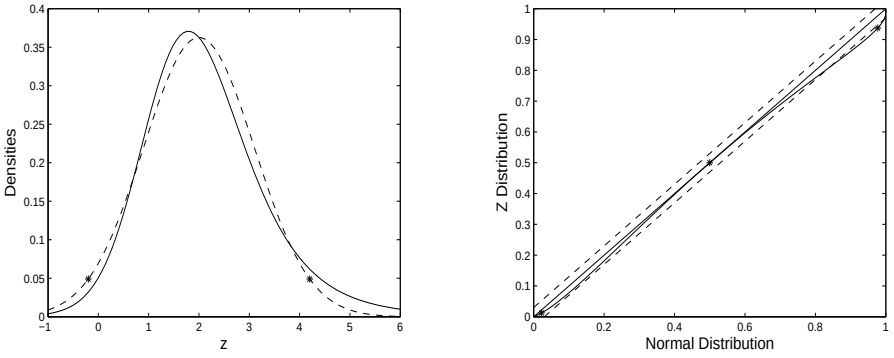



**Fig. 3** Example 2. A case where the normal approximation (dashes) to $F_Z$ (solid line) is reasonable; (a) Density functions; (b) ROC curve (solid), $\pm 0.03$ band in dashes.



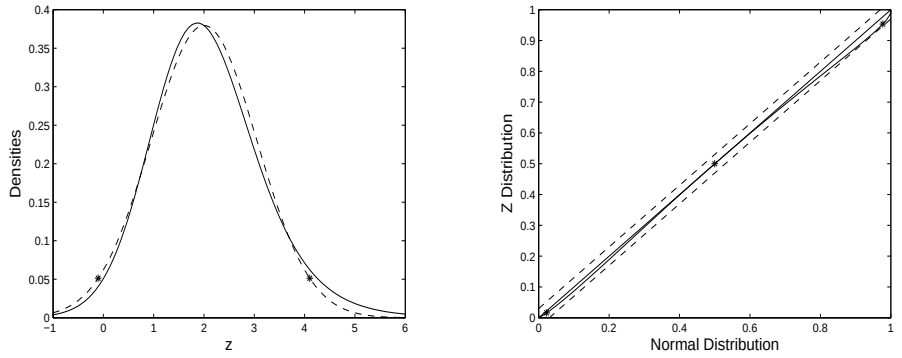

(a) (b)

**Fig. 4** Example 3. A case where the normal approximation (dashes) to $F_Z$ (solid line) is very good; (a) Density functions; (b) ROC curve (solid), $\pm 0.03$ band in dashes.




(a) (b)

**Fig. 5** Example 4. Case of a bad normal approximation to $F_Z$; (a) Density functions, $f_Z$ (solid line) and normal approximation (in dashes); (b) ROC curve (solid), $\pm 0.03$ band in dashes.




(a) (b)

**Fig. 6** Example 5. Case of a reasonable normal approximation to $F_Z$. (a) Density functions, $f_Z$ (solid line) and normal approximation (in dashes); (b) ROC curve (solid), $\pm 0.03$ band in dashes.

## 6 Practical Application

Paired independent normal observations consisting of flow cytometer measurements of an experimental plant $(X)$ with unknown nuclear DNA content and the corresponding control plant $(Y)$, with a well known nuclear DNA content, were taken for eight adult experimental plants. The goal of the experiment was to estimate the nuclear genome size of the experimental plant. More details are given in Díaz-Francés and Sprott (2001) where proper statistical models are discussed to obtain inferences about the common ratio $\beta$ of the normal means for all pairs. The ratio $\beta$ is proportional to the DNA nuclear content of the plant of interest. Those methods were applied to the present data set and other observations to estimate successfully the DNA content of this experimental plant as part of the project "Programa General de Apoyo y Desarrollo Tecnológico a la Cadena Productiva Agave-Tequila" sponsored by the Mexican Government and the Consejo Regulador del Tequila, where CIMAT participated providing statistical consulting.

Here the mentioned paired data set, given below in Table 2, will be used to exemplify a case where the proposed normal approximation of Section 3 is practically indistinguishable from the distribution $F_Z$ within an interval $I$ that is also given explicitly below. The maximum likelihood estimates of the parameters obtained from the pairs $(x_i, y_i)$, $i = 1, ..., 8$, will be used to calculate the proposed normal approximation and $f_Z$ in order to show their proximity.. These are: $\hat{\mu}_x = 77.125$, $\hat{\mu}_y = 50.154$, $\hat{\sigma}_x = 2.104$, $\hat{\sigma}_y = 1.040$, $\hat{\beta} = 1.5378$, $\hat{\rho} = 0.496$, $\hat{\delta}_x = 0.0273$ , $\hat{\delta}_y = 0.0208$, $\hat{\sigma}_z = 0.0528$.

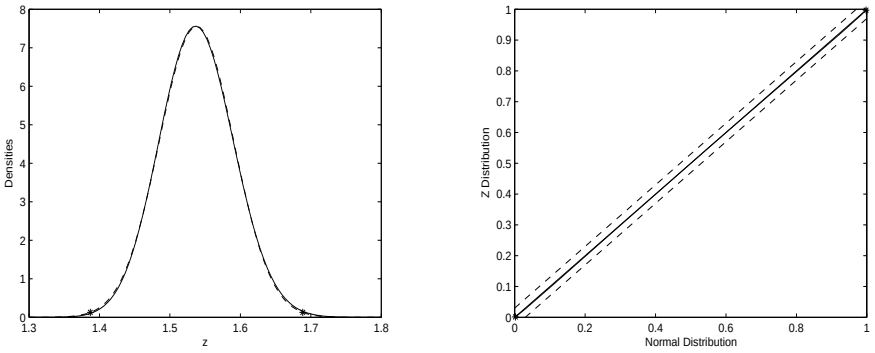
| Pair | Experimental plant | Control plant |
|---|---|---|
| 1 | 80.773 | 51.668 |
| 2 | 78.293 | 49.928 |
| 3 | 75.424 | 49.768 |
| 4 | 74.535 | 48.214 |
| 5 | 79.052 | 51.033 |
| 6 | 75.106 | 49.478 |
| 7 | 78.188 | 51.240 |
| 8 | 75.632 | 49.904 |

Table 2. Paired data for eight adult plants.

For example for the value $\lambda = 0.35$ (which is smaller than one as required in Theorem 1) and for the mentioned estimates, the interval $I$ of (11) is $I = [1.387, 1.688]$. As an example, for a value of $\varepsilon = 0.0038$, there is an associated value of $\gamma(\varepsilon) = 0.02094$ (obtained by trial and error), such that if $\delta_y \leq \gamma(\varepsilon)$, as in (10), then the inequality (12) of Theorem 1 holds for any $z$ within $I$. Since the estimate $\hat{\delta}_y$ is smaller than this value of $\gamma(\varepsilon)$, consequently it will fulfill the inequality (12) of Theorem 1 within this interval.

Note that the estimated value $\hat{\delta}_y$ is fixed, so that given the selected $\lambda = 0.35$ there is a smallest value of $\varepsilon$ for which Theorem 1 holds for the associated interval $I$. This value is precisely $\varepsilon = 0.0038$. Theorem 1 certainly holds for any other value of $\varepsilon$, smaller than this value, but then another random variable $Y$ different from the one considered here, with a smaller $\delta_y$ would be the one that would fulfill Theorem 1.

Figure 7a. shows the densities of the above mentioned normal approximation and $f_Z$; the interval $I$ is marked with asterisks over these curves that almost overlap everywhere, indicating that the normal approximation is very close to $f_Z$. Figure 7b. shows the corresponding ROC curve which practically overlaps the 45 degree line, indicating the closeness of these distributions. A band of an arbitrary width 0.03 is marked in dashes, above and below the diagonal line. Verifying the closeness of a normal distribution to $F_Z$ is not made in cytometry to our knowledge, even when a normal distribution is frequently used to simplify inferences about $\beta$.




(a) (b)

**Fig. 7** Paired data of adult experimental plants: (a) Estimated normal approximation of Section 3 (in dashes) and $f_Z$ (solid line); (b) Corresponding ROC curve (solid line), $\pm$ 0.03 band (in dashes).

## 7 Conclusions

It has been proven here that for any normal random variable $X$ with positive mean and coefficient of variation $\delta_x \leq 1$, there exists another independent normal variable $Y$ with positive mean and a small coefficient of variation, fulfilling some conditions, such that their ratio $Z$ can be well approximated with the proposed normal distribution in Section 3 within a given interval. Once any such $Y$ has this property, any other $Y^*$ with a smaller coefficient of variation will satisfy this result as well.

It was also shown here how the ROC curves are useful to check the proximity or discrepancy between the proposed normal approximation and the distribution of $Z$, for any given pair of independent normal random variables $X$, $Y$ with positive means.

**Acknowledgments** The authors would like to thank the referee and the editor for their insightful comments, as well as Dr. David A. Sprott for fruitful discussions. The authors also thank the Consejo Regulador del Tequila for facilitating the data

set of Section 6. Finally, the second author thanks CIMAT for the support that was given while working on this manuscript.

## Appendix

For the proof of Theorem 1, the parametrization $(\beta, \rho, \delta_y)$ will be used first for the sake of simpler algebraic expressions. However, the alternative parametrization $(\beta, \delta_x, \delta_y)$ will be used when taking limits when $\delta_y$ tends to zero.

Consider the arguments of the functions $F^*$ given in (6) and of the proposed normal approximation $G$ given in (9) for $z > 0$, and define them as

$$h_1(z) = \frac{z-\beta}{\delta_y\sqrt{\rho^{-2}+z^2}}, \text{ and } h_2(z) = \frac{z-\beta}{\delta_y\sqrt{\rho^{-2}+\beta^2}},$$

respectively. That is, $G(z) = \Phi[h_2(z)]$, and $F^*(z) = \Phi[h_1(z)]$. Also, define the difference between these two functions as

$$w(z) = h_2(z) - h_1(z) = \frac{(z-\beta)}{\delta_y}\left(\frac{1}{\sqrt{\rho^{-2}+\beta^2}} - \frac{1}{\sqrt{\rho^{-2}+z^2}}\right).$$

Notice again that $h_1$ and $h_2$ differ only in a term within the square root of their denominator.

The outline of the proof is as follows. First, it will be proved that for all $z \in I$, $h_2(z) \geq h_1(z)$, or equivalently that $w(z) \geq 0$.

Second, it will be proved that $w(z)$ is decreasing in $[\beta - \sigma_z/\lambda, \beta)$ and increasing in $[\beta, \beta + \sigma_z/\lambda]$. For that purpose its first derivative will be proved to be negative in the former interval, while positive in the latter interval. This will imply then that $w(z)$ achieves its maximum, precisely at one of the endpoints of the interval $I$, since it takes a minimum and a value of zero at its midpoint where $w(\beta) = 0$. Since the right endpoint of $I$ has a larger value of $w(z)$ than the left endpoint, therefore the largest difference between $h_1$ and $h_2$ in the interval $I$ is achieved precisely at the right endpoint.

Third it will be proved that this maximum difference can be made as small as desired by making $\delta_y$ sufficiently small. By doing this, a bound for the difference between $G(z)$ and $F^*(z)$ is found for all $z$ in the interval $I$. Since Hinkley (1969) proved that $F_Z$ converges uniformly to $F^*(z)$ as well, then another upper bound exists for the difference between these two functions that can be made as small as desired by controlling $\delta_y$.

Finally, applying the triangle inequality to the differences between the three functions of interest, $G(z)$, $F^*(z)$ and $F_Z(z)$, it is proved that the difference between the proposed normal approximation $G$ and $F_Z$ can be made as small as desired.

**Proof:**

As mentioned, it will be proved first that for all $z \in I$, $h_2(z) \geq h_1(z)$, or equivalently that $w(z) \geq 0$. This holds because for $z$ in $[\beta - \sigma_z/\lambda, \beta)$, the fact that $z < \beta$ implies that

$$\frac{1}{\sqrt{\rho^{-2}+\beta^2}} \leq \frac{1}{\sqrt{\rho^{-2}+z^2}}. \tag{13}$$

Since $(z-\beta)$ is negative in this interval, multiplying both sides of the above inequality yields that $h_2(z) \geq h_1(z)$. Now, for $z$ in $[\beta, \beta+\sigma_z/\lambda]$, the reverse inequality holds for (13), and since the term $(z-\beta)$ is positive, after multiplying both sides of the inequality, one obtains again $h_2(z) \geq h_1(z)$. Therefore, for all $z \in I$, $w(z) \geq 0$; the equality is achieved at $\beta$.

For the second part of the proof, note that the derivative of $w(z)$ is

$$\frac{d\,w(z)}{dz} = \frac{dh_2(z)}{dz} - \frac{dh_1(z)}{dz} = \frac{1}{\delta_y}\left[\frac{1}{\sqrt{\rho^{-2}+\beta^2}} - \frac{\rho^{-2}+z\beta}{(\rho^{-2}+z^2)^{3/2}}\right].$$

For $z$ in $[\beta-\sigma_z/\lambda, \beta)$, since $z<\beta$,

$$\begin{aligned}\frac{d\,w(z)}{dz} &< \frac{1}{\delta_y}\left[\frac{1}{\sqrt{\rho^{-2}+\beta^2}} - \frac{\rho^{-2}+z^2}{(\rho^{-2}+z^2)^{3/2}}\right] \\ &= \frac{1}{\delta_y}\left[\frac{1}{\sqrt{\rho^{-2}+\beta^2}} - \frac{1}{\sqrt{\rho^{-2}+z^2}}\right] < 0.\end{aligned}$$

For $z$ in $(\beta, \beta+\sigma_z/\lambda]$, since $z>\beta$,

$$\begin{aligned}\frac{d\,w(z)}{dz} &> \frac{1}{\delta_y}\left[\frac{1}{\sqrt{\rho^{-2}+\beta^2}} - \frac{\rho^{-2}+z\beta}{(\rho^{-2}+z\beta)^{3/2}}\right] \\ &= \frac{1}{\delta_y}\left[\frac{1}{\sqrt{\rho^{-2}+\beta^2}} - \frac{1}{\sqrt{\rho^{-2}+z\beta}}\right] > 0.\end{aligned}$$

Therefore $w(z)$ is decreasing in $[\beta-\sigma_z/\lambda, \beta)$, since it has a negative derivative in this interval, and increasing in $(\beta, \beta+\sigma_z/\lambda]$ because of having a positive derivative here. By simplifying the following inequality, it is straightforward to prove that

$$w(\beta-\sigma_z/\lambda) \leq w(\beta+\sigma_z/\lambda).$$

Therefore, $w(z)$ reaches a maximum at the right endpoint of the interval $I$, at $z=\beta+\sigma_z/\lambda$.

For the third part of the proof, this maximum difference at the right endpoint is the largest. For convenience, it will be expressed now as a function of the parameters $(\beta, \delta_x, \delta_y)$,

$$w(\beta+\sigma_z/\lambda) = A(\delta_y, \delta_x) = \frac{1}{\lambda} - \frac{1}{\lambda}\frac{\sqrt{\delta_x^2+\delta_y^2}}{\sqrt{\delta_x^2+\delta_y^2\left(1+\frac{1}{\lambda}\sqrt{\delta_x^2+\delta_y^2}\right)^2}}.$$

Note that this quantity approaches zero when $\delta_y$ tends to zero; that is,

$$\lim_{\delta_y \to 0} A(\delta_y, \delta_x) = 0.$$

The definition of this limit implies that for every $\varepsilon_1 > 0$, there exists a value $\eta(\varepsilon_1) > 0$ such that if $\delta_y < \eta(\varepsilon_1)$, then

$$w(\beta + \sigma_z/\lambda) = A(\delta_y, \delta_x) < \varepsilon_1.$$

Therefore the following holds for such $\delta_y$ and $z \in I$,

$$0 \leq h_2(z) - h_1(z) \leq w(\beta + \sigma_z/\lambda) < \varepsilon_1. \tag{14}$$

Then for $z \in I$, note that $\Phi[h_2(z)] \geq \Phi[h_1(z)]$ since the standard normal distribution is an increasing function and $h_2(z) \geq h_1(z)$. Also, for the same reason, expression (14) implies that $\Phi[h_1(z)] > \Phi[h_2(z) - \varepsilon_1]$. Then,

$$\begin{aligned} \left|G(z) - F^*(z)\right| &= \Phi[h_2(z)] - \Phi[h_1(z)] < \Phi[h_2(z)] - \Phi[h_2(z) - \varepsilon_1] \\ &= \int_{h_2(z)-\varepsilon_1}^{h_2(z)} \phi(t)\, dt \overset{(*)}{=} \phi(t_0)\, \varepsilon_1 \leq \frac{\varepsilon_1}{\sqrt{2\pi}}. \end{aligned} \tag{15}$$

The last equality $(*)$ in (15) is obtained after applying the integral mean value theorem to this expression, where $\phi(t)$ is the standard normal density; the value $t_0$ is the point in the interval $(h_2(z) - \varepsilon_1, h_2(z))$ for which the mean value theorem holds. The last inequality holds because the maximum of $\phi(t)$ is $1/\sqrt{2\pi}$. Note that we have obtained in this way an upper bound for the difference between the proposed normal approximation $G(z)$ and Hinkley's approximation $F^*(z)$ that can be made as small as desired by making $\delta_y$ sufficiently small. That is,

$$\left|G(z) - F^*(z)\right| < \frac{\varepsilon_1}{\sqrt{2\pi}}. \tag{16}$$

Hinkley (1969) proved that $F_Z$ converges uniformly to $F^*$ when $\delta_y \to 0$. Therefore, for every $\varepsilon_2 > 0$, there exists $\nu(\varepsilon_2)$ such that if $\delta_y < \nu(\varepsilon_2)$, then

$$\left|F^*(z) - F_Z(z)\right| < \varepsilon_2. \tag{17}$$

For the final part of the proof note that for every $\varepsilon > 0$, if $0 < \delta_y < \gamma = \min\left\{\eta(\varepsilon_1), \nu(\varepsilon_2), \sqrt{\lambda^2 - \delta_x^2}\right\}$ and $z \in I$, then by the triangle inequality,

$$|G(z) - F_Z(z)| \leq \left|G(z) - F^*(z)\right| + \left|F^*(z) - F_Z(z)\right| \leq \frac{\varepsilon_1}{\sqrt{2\pi}} + \varepsilon_2 = \varepsilon,$$

for suitable selections of $\varepsilon_1$ and $\varepsilon_2$ that satisfy that $\varepsilon = (\varepsilon_1/\sqrt{2\pi} + \varepsilon_2)$.

So finally, for every $\varepsilon > 0$, the proposed normal approximation $G(z)$ can be as close as desired to $F_Z(z)$ in the interval $I$, for a sufficiently small value of $\delta_y$. Therefore the proof is complete.

## References

Brody JP, Williams BA, Wold BJ and Quake SR (2002) Significance and statistical errors in the analysis of DNA microarray data. P Natl Acad Sci USA 99: 12975–12978

Chamberlin SR and Sprott DA (1987) Some logical aspects of the linear functional relationship. Stat Pap 28: 291–299

Díaz-Francés E and Sprott DA (2001) Statistical analysis of nuclear genome size of plants with flow cytometer data. Cytometry 45: 244–249

Geary, R. C. (1930) The frequency distribution of the quotient of two normal variates. J Roy Stat Soc B Met 93: 442–446

Hayya J, Armstrong D and Gressis N (1975) A note on the ratio of two normally distributed variables. Manage Sci 21: 1338–1341

Hinkley DV (1969) On the ratio of two correlated normal variables. Biometrika 56: 635–639

Kotz S, Lumelskii Y and Pensky M (2003) The Stress-Strength Model and Its Generalizations: Theory and Applications. World Scientific Publishing, River Edge NJ

Kuethe DO, Caprihan A, Gach HM, Lowe IJ and Fukushima E (2000). Imaging obstructed ventilation with NMR using inert fluorinated gases. J Appl Physiol 88: 2279–2286

Lisák MA and Doležel J (1998). Estimation of nuclear DNA content in Sesleria (Poaceae). Caryologia 52: 123–132

Marsaglia G (1965). Ratios of normal variables and ratios of sums of uniform variables. J Am Stat Assoc 60: 193–204

Marsaglia G (2006). Ratios of normal variables. J Stat Softw 16: 1–10

Merrill AS (1928). Frequency distribution of an index when both the components follow the normal law. Biometrika 20A: 53–63

Palomino G, Doležel J, Cid R, Brunner I, Méndez I and Rubluo A (1999). Nuclear genome stability of mammillaria San-Angelensis (Cactaceae) regenerants. Plant Sci 141: 191-200

Sklar LA (2005). Flow Cytometry for Biotechnology. Oxford University Press, New York

Schneeweiss H, Sprott DA and Viveros R (1987) An approximate conditional analysis of the linear functional relationship. Stat Papers 28: 183–202

Watson JV (1992). Flow Cytometry Data Analysis: Basic Concepts and Statistics. Cambridge University Press, New York