THE THEORY OF THE INTERNALLY STUDENTIZED RANGE DISTRIBUTION REVISITED

Daniel Furtado FERREIRA[1]
Lucas Monteiro CHAVES[2]
Devanil Jaques de SOUZA[1]

- ABSTRACT: The present paper intends to revisit the distribution of the ratio of the range to the sample standard deviation, known as the distribution of the internally studentized range, in the normal case. This distribution has its importance recognized in several areas, as quality control and inference, for testing the lack of homogeneity of the data or kurtosis. An alternative distribution to the one presented by David et al. (1954), based on the distribution of the maximum, is proposed. We exhibit a detailed proof for the distribution of the internally studentized range in the normal case and sample size 3. We also provide a new result: the distribution for the uniform case with sample of size 3.

- KEYWORDS: Order statistics; closest of three; normal distribution; uniform distribution.

1 Introduction

Consider two independent random samples, both of size n, of a normal distribution. Let $X_{(1)}$ be the minimum, $X_{(n)}$ be the maximum, and $W = X_{(n)} - X_{(1)}$ be the range of the first sample and S the standard deviation of the second sample. Then S and W are independent and the distribution of the quantity $U = W/S$ is very well known, usually called externally studentized range distribution, and is very important in statistical inference. This situation occurs naturally in experimental statistics. Note that the assumed two samples is only a resource

[1] Universidade Federal de Lavras - UFLA, Departamento de Estatística, CEP: 37.200-000, Lavras, MG, Brazil. E-mail: *danielff@des.ufla.br; devaniljaques@des.ufla.br*

[2] Universidade Federal de Lavras - UFLA, Departamento de Ciências Exatas, CEP: 37.200-000, Lavras, MG, Brazil. E-mail: *lucas@dex.ufla.br*

to guarantee the independence (RENCHER; SCHAALJE, 2008). The externally studentized range distribution is also intensively applied in multiple comparisons problems, as Tukey, Duncan and Student-Newman-Keuls's tests (HINKELMANN; KEMPTHORNE, 2007, p. 224). Case S and W are computed in the same sample, they are not independent anymore and have a complicated joint distribution. The exact distribution of internally studentized range $U = W/S$ is not known (DAVID; HARTLEY; PEARSON, 1954; CURRIE, 1980) and will be the focus of this work.

David, Hartley e Pearson (1954) argue that approximations of this distribution has been efficient both in the exploration of homogeneity of data or departure from normality and in testing for kurtosis. They had studied it using two approximations. Firstly, they calculated the first four moments of $U = W/S$ and approached the true distribution by the Pearson's curves. Secondly, they related the distribution of $U = W/S$ to the Student t distribution and used the late to approach upper tail quantiles of the distribution of $U = W/S$. After the work of David, Hartley e Pearson (1954) very few was done to improve the knowledge of this distribution. Thomson (1955) exhibit, without proof, the distribution in the normal case and samples of size 3. Currie (1980), considering parent normal distribution and using geometric arguments, obtained the distribution for sample of size 4.

This work may be outlined as: in section 2 we review the theory of the internally studentized range distribution and exhibit detailed proofs for all the important and original theoretical mathematical statistics used in David, Hartley e Pearson (1954)'s article. We also provide a new approximation to the distribution of the internally studentized range for normal random samples in subsection 2.2 and a Monte Carlo approach in subsection 2.3. In section 3 our approximations are compared with that presented by David, Hartley e Pearson (1954). In section 4 we exhibit a detailed proof for the distribution of the internally studentized range, originally obtained by Thomson (1955), in the normal case with sample of size 3. We also provide a new result: the distribution, in the uniform case, for samples of size 3. Finally, in section 5, we show an application of the internally studentized range. We present a simple new normality test and the evaluation of its performance.

2 The theory of the internally studentized range distribution

Let X_1, X_2, …, X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2. Let $X_{(1)}$, $X_{(2)}$, …, $X_{(n)}$ be the order statistics for this sample, where $X_{(j)}$ is the jth smallest order statistics. Thus $X_{(1)}$ is the minimum and $X_{(n)}$ is the maximum order statistics. The sample range is defined as the difference

$$W = X_{(n)} - X_{(1)}.$$

Consider the sample mean

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We want do study the statistics: the studentized range $U = (X_{(n)} - X_{(1)})/S = W/S$, $(X_{(n)} - \bar{X})/S$ and $(X_{(1)} - \bar{X})/S$. These statistics are important, for example, in analysing outliers.

Proposition 2.1. *$(X_{(n)} - \bar{X})/S$, $(X_{(1)} - \bar{X})/S$ and $W/S = (X_{(n)} - X_{(1)})/S$ are independent of S, where all those statistics are functions of a random sample X_1, X_2, ..., X_n from a normal distribution with mean μ and variance σ^2.*

Proof. We present here a proof based on Basu's theorem which states that any ancillary statistic is independent of any minimal sufficient statistic.

The quantity $(X_i - \bar{X})/S$, $i = 1, 2, \ldots, n$, is invariant for linear transformation $\alpha X + \beta$, $\alpha > 0$, and therefore it is an ancillary statistic. As $(X_{(i)} - \bar{X})/S$ depends only on $(X_i - \bar{X})/S$, it is also ancillary. By Basu's theorem (CASELLA; BERGER, 2002, p. 287) it is independent of the minimal sufficient statistics $\bar{X}$ and S. An alternative proof is presented in Appendix A. □

As the moments of W and S are known, this result allows an easy calculation of the moments of $U = W/S$:

$$\begin{aligned} E\left[W^r\right] =& E\left[\frac{W^r}{S^r}S^r\right] = E\left[\left(\frac{W}{S}\right)^r S^r\right] = E\left[\left(\frac{W}{S}\right)^r\right]E\left[S^r\right] \\ \rightarrow& E\left[\left(\frac{W}{S}\right)^r\right] = \frac{E\left[W^r\right]}{E\left[S^r\right]}. \end{aligned}$$

Let's consider now

$$U' = (X_j - X_k)/S,$$

where $(j,\ k)$ is any pair of elements in $\{1,\ 2,\ \ldots,\ n\}$ with $j \neq k$. The random variable U' is related to Student's t distribution. To see this we need a very trick identity in sum of squares.

Proposition 2.2. $(n-1)s^2 = \frac{1}{2}(x_j - x_k)^2 + \sum_{i\neq j,k}^{n}(x_i - \bar{x}')^2 + \frac{2(n-2)}{n}(\bar{x}' - \bar{x}'')^2$, *where*

$$\bar{x}' = \frac{\sum_{i\neq j,k}^{n} x_i}{n-2} \qquad \textit{and} \qquad \bar{x}'' = \frac{x_j + x_k}{2}.$$

Proof. An algebraic proof, that the authors could not find anywhere, is left to the appendix A. We present here a new geometric proof. We suppose, without loss of generality, $j = 1$ and $k = 2$.

In Figure 1 and in what follows we use the notations: $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)^\top$, $\bar{\boldsymbol{x}} = (\bar{x}, \bar{x}, \cdots, \bar{x})^\top$, $\boldsymbol{\eta} = (\bar{x}'', \bar{x}'', x_3, \cdots, x_n)^\top$ and $\boldsymbol{\xi} = (\bar{x}'', \bar{x}'', \bar{x}', \cdots, \bar{x}')^\top$

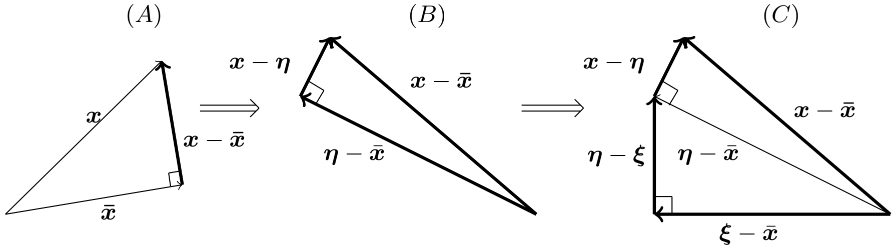
(A)
(B)
(C)
$\boldsymbol{x}$
$\bar{\boldsymbol{x}}$
$\boldsymbol{x} - \bar{\boldsymbol{x}}$
$\Longrightarrow$
$\boldsymbol{x} - \boldsymbol{\eta}$
$\boldsymbol{x} - \bar{\boldsymbol{x}}$
$\boldsymbol{\eta} - \bar{\boldsymbol{x}}$
$\Longrightarrow$
$\boldsymbol{x} - \boldsymbol{\eta}$
$\boldsymbol{x} - \bar{\boldsymbol{x}}$
$\boldsymbol{\eta} - \boldsymbol{\xi}$
$\boldsymbol{\eta} - \bar{\boldsymbol{x}}$
$\boldsymbol{\xi} - \bar{\boldsymbol{x}}$

Figure 1 - Geometric construction of the partition of $\boldsymbol{x} - \bar{\boldsymbol{x}}$.

Figure 1(A) just shows the vectors $\boldsymbol{x}$, $\bar{\boldsymbol{x}}$ and the diference $\boldsymbol{x} - \bar{\boldsymbol{x}}$. Figure 1(B) shows the decomposition of $\boldsymbol{x} - \bar{\boldsymbol{x}}$ as the sum of the vectors $\boldsymbol{x} - \boldsymbol{\eta}$ and $\boldsymbol{\eta} - \bar{\boldsymbol{x}}$. Figure 1(C) decomposes the vector $\boldsymbol{\eta} - \bar{\boldsymbol{x}}$ as the sum of $\boldsymbol{\eta} - \boldsymbol{\xi}$ and $\boldsymbol{\xi} - \bar{\boldsymbol{x}}$, what allows us to write $\boldsymbol{x} - \bar{\boldsymbol{x}} = (\boldsymbol{x} - \boldsymbol{\eta}) + (\boldsymbol{\eta} - \boldsymbol{\xi}) + (\boldsymbol{\xi} - \bar{\boldsymbol{x}})$. Now, if we prove the orthogonality of $\boldsymbol{x} - \boldsymbol{\eta}$ and $\boldsymbol{\eta} - \bar{\boldsymbol{x}}$ we may write $\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 = \|\boldsymbol{x} - \boldsymbol{\eta}\|^2 + \|\boldsymbol{\eta} - \bar{\boldsymbol{x}}\|^2$; if we prove the orthogonality of $\boldsymbol{\eta} - \boldsymbol{\xi}$ and $\boldsymbol{\xi} - \bar{\boldsymbol{x}}$, we may also write $\|\boldsymbol{\eta} - \bar{\boldsymbol{x}}\|^2 = \|\boldsymbol{\eta} - \boldsymbol{\xi}\|^2 + \|\boldsymbol{\xi} - \bar{\boldsymbol{x}}\|^2$. Putting all together we may, finally, write:

$$\|\boldsymbol{x} - \bar{\boldsymbol{x}}\|^2 = \|\boldsymbol{x} - \boldsymbol{\eta}\|^2 + \|\boldsymbol{\eta} - \boldsymbol{\xi}\|^2 + \|\boldsymbol{\xi} - \bar{\boldsymbol{x}}\|^2$$

To show the orthogonality of $\boldsymbol{x} - \boldsymbol{\eta}$ and $\boldsymbol{\eta} - \bar{\boldsymbol{x}}$ we observe that

$$\begin{aligned}
(\boldsymbol{x} - \boldsymbol{\eta}) \boldsymbol{\cdot} (\boldsymbol{\eta} - \bar{\boldsymbol{x}}) = & \left[(x_1, x_2, \cdots, x_n)^\top - (\bar{x}'', \bar{x}'', x_3, \cdots, x_n)^\top\right] \boldsymbol{\cdot} \\
& \qquad \left[(\bar{x}'', \bar{x}'', x_3, \cdots, x_n)^\top - (\bar{x}, \bar{x}, \cdots, \bar{x})^\top\right] \\
= & (\bar{x}'' - \bar{x}) \left(x_1 + x_2 - 2\frac{x_1 + x_2}{2}\right) \\
= & 0
\end{aligned}$$

and, so, the vectors $\boldsymbol{x} - \boldsymbol{\eta}$ and $\boldsymbol{\eta} - \bar{\boldsymbol{x}}$ are orthogonal. In a similar way one shows that $(\boldsymbol{\eta} - \boldsymbol{\xi}) \boldsymbol{\cdot} (\boldsymbol{\xi} - \bar{\boldsymbol{x}}) = 0$ and, so, $\boldsymbol{\eta} - \boldsymbol{\xi}$ and $\boldsymbol{\xi} - \bar{\boldsymbol{x}}$ are also orthogonal. Therefore

$$\begin{aligned}
& (x_1, x_2, \cdots, x_n)^\top - (\bar{x}, \bar{x}, \cdots, \bar{x})^\top = (x_1, x_2, \cdots, x_n)^\top - \\
& - (\bar{x}'', \bar{x}'', x_3, \cdots, x_n)^\top + (\bar{x}'', \bar{x}'', x_3, \cdots, x_n)^\top - \\
& - (\bar{x}'', \bar{x}'', \bar{x}', \cdots, \bar{x}')^\top + (\bar{x}'', \bar{x}'', \bar{x}', \cdots, \bar{x}')^\top - (\bar{x}, \bar{x}, \cdots, \bar{x})^\top
\end{aligned}$$

and

$$\left\|(x_1,x_2,\cdots,x_n)^\top-(\bar{x},\bar{x},\cdots,\bar{x})^\top\right\|^2=\left\|(x_1,x_2,\cdots,x_n)^\top-\left(\bar{x}'',\bar{x}'',x_3,\cdots,x_n\right)^\top\right\|^2+$$
$$\left\|\left(\bar{x}'',\bar{x}'',x_3,\cdots,x_n\right)^\top-\left(\bar{x}'',\bar{x}'',\bar{x}',\cdots,\bar{x}'\right)^\top\right\|^2+\left\|\left(\bar{x}'',\bar{x}'',\bar{x}',\cdots,\bar{x}'\right)^\top-(\bar{x},\cdots,\bar{x})^\top\right\|^2$$

or

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\underbrace{\sum_{i=1}^{2}\left(x_i-\bar{x}''\right)^2}_{(*)}+\sum_{i=3}^{n}\left(x_i-\bar{x}'\right)^2+\underbrace{2\left(\bar{x}''-\bar{x}\right)^2+(n-2)\left(\bar{x}'-\bar{x}\right)^2}_{(**)}.$$

Observe that

$$\begin{aligned}(*)=&\sum_{i=1}^{2}(x_i-\bar{x}'')^2\\=&\left(x_1-\frac{x_1+x_2}{2}\right)^2+\left(x_2-\frac{x_1+x_2}{2}\right)^2=\left(\frac{x_1-x_2}{2}\right)^2+\left(\frac{x_2-x_1}{2}\right)^2\\=&\frac{1}{2}(x_2-x_1)^2.\end{aligned}$$

Taking

$$\bar{x}=\frac{2\left(\frac{x_1+x_2}{2}\right)+(n-2)\left(\frac{\sum_{i=3}^{n}x_i}{n-2}\right)}{n}=\frac{2\bar{x}''+(n-2)\bar{x}'}{n},$$

the expression $(**)$ simplifies as:

$$\begin{aligned}(**)=&2\left[\bar{x}''-\frac{2\bar{x}''+(n-2)\bar{x}'}{n}\right]^2+\\&+(n-2)\left[\bar{x}'-\frac{2\bar{x}''+(n-2)\bar{x}'}{n}\right]^2\\=&\frac{2(n-2)}{n}\left(\bar{x}'-\bar{x}''\right)^2,\end{aligned}$$

then, the result follows:

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=(n-1)s^2=\frac{1}{2}(x_2-x_1)^2+\sum_{i=3}^{n}(x_i-\bar{x}')^2+\frac{2(n-2)}{n}\left(\bar{x}'-\bar{x}''\right)^2.$$

□

2.1 An approximation for the upper tail of the internally studentized range distribution

Taking (i,j) instead of $(1,2)$ and considering random variables, the last result may be written as

$$(n-1)S^2 = \frac{1}{2}\left(X_j - X_k\right)^2 + \sum_{i\neq j,k}^{n}\left(X_i - \bar{X}'\right)^2 + \frac{2(n-2)}{n}\left(\bar{X}' - \bar{X}''\right)^2.$$

We show below that the rank of left side quadratic form is $n-1$ and the ranks of the right side quadratic forms are, respectively, 1, $n-3$ and 1. Those quadratic forms are all invariant for the translation $X_i - \mu$, $i = 1,\ 2,\ \ldots,\ n$. Dividing both sides by σ^2,

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{2}\left(\frac{X_j - X_k}{\sigma}\right)^2 + \sum_{i\neq j,k}^{n}\left(\frac{X_i - \bar{X}'}{\sigma}\right)^2 + \frac{2(n-2)}{n}\left(\frac{\bar{X}' - \bar{X}''}{\sigma}\right)^2$$

is equivalent to dividing each X_i by σ^2. So, the variables in the quadratic form may be taken as standard normal.

Proposition 2.3. *For $j \neq k$ and $n \geq 3$, the variate $U' = (X_j - X_k)/S$ is distributed as*

$$\frac{T\sqrt{2(n-1)}}{\sqrt{T^2 + n - 2}}$$

where T has a Student's t distribution with $n-2$ degrees of freedom.

Proof. Let

$$\begin{aligned} U' &= \frac{X_j - X_k}{S} = \frac{X_j - X_k}{\sqrt{S^2}} \\ &= \frac{X_j - X_k}{\sqrt{\frac{1}{n-1}\left[\frac{1}{2}\left(X_j - X_k\right)^2 + \sum_{i\neq j,k}^{n}\left(X_i - \bar{X}'\right)^2 + \frac{2(n-2)}{n}\left(\bar{X}' - \bar{X}''\right)^2\right]}} \\ &= \frac{(X_j - X_k)\sqrt{n-1}}{\sqrt{\frac{1}{2}\left(X_j - X_k\right)^2 + \sum_{i\neq j,k}^{n}\left(X_i - \bar{X}'\right)^2 + \frac{2(n-2)}{n}\left(\bar{X}' - \bar{X}''\right)^2}}, \end{aligned}$$

$$\mathcal{Q} = \sum_{i\neq j,k}^{n}\left(X_i - \bar{X}'\right)^2 + \frac{2(n-2)\left(\bar{X}' - \bar{X}''\right)^2}{n}$$

and

$$
\begin{aligned}
U' =& \frac{(X_j - X_k)\sqrt{n-1}}{\sqrt{\frac{1}{2}\left(X_j - X_k\right)^2 + \mathcal{Q}}} = \frac{[(X_j - X_k)\sqrt{n-1}]/\sigma}{\sqrt{\frac{\mathcal{Q}}{(n-2)\sigma^2}\left[\frac{0.5(X_j-X_k)^2/\sigma^2}{\frac{\mathcal{Q}}{(n-2)\sigma^2}} + n - 2\right]}} \\
=& \frac{\left(\frac{X_j - X_k}{\sqrt{2}\sigma}\right)\sqrt{2(n-1)}}{\sqrt{\frac{\mathcal{Q}}{(n-2)\sigma^2}}} \frac{1}{\sqrt{\frac{0.5(X_j-X_k)^2/\sigma^2}{\frac{\mathcal{Q}}{(n-2)\sigma^2}} + n - 2}} = \frac{T\sqrt{2(n-1)}}{\sqrt{T^2 + n - 2}}
\end{aligned}
$$

with

$$
T = \frac{\left(\dfrac{X_j - X_k}{\sqrt{2}\sigma}\right)}{\sqrt{\dfrac{\mathcal{Q}}{(n-2)\sigma^2}}}.
$$

To derive the distribution of T we show that $\mathcal{Q}$ has a chi-square distribution with $n-2$ degrees of freedom. For this we need the classical result:

Fisher-Cochran Theorem (RAO, 2002, p. 185): Consider n independent standard normal variables $X_i \sim N(0, 1)$. Let $Q_1, Q_2, \ldots, Q_k$ be quadratic forms with ranks $n_1, n_2, \ldots, n_k$, respectively, such that

$$
\boldsymbol{X}^\top \boldsymbol{X} = Q_1 + Q_2 + \cdots + Q_k,
$$

where $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^\top$. Then, a necessary and sufficient condition that $Q_i \sim \chi^2_{n_i}$ and the Q_i's are independent is $n = \sum_{i=1}^k n_i$.

The decomposition of the corrected total sum of squares

$$
(n-1)S^2 = \frac{1}{2}\left(X_j - X_k\right)^2 + \sum_{i \neq j,k}^{n} \left(X_i - \bar{X}'\right)^2 + \frac{2(n-2)}{n}\left(\bar{X}' - \bar{X}''\right)^2,
$$

can be expressed as

$$
\begin{aligned}
(n-1)S^2 = \boldsymbol{x}^\top \boldsymbol{P} \boldsymbol{x} =& \boldsymbol{x}^\top \boldsymbol{P}_1 \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{P}_2 \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{P}_3 \boldsymbol{x} \\
=& Q_1 + Q_2 + Q_3,
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{P} =&\boldsymbol{I} - \frac{1}{n}\boldsymbol{J} = \begin{bmatrix} \frac{n-1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & \frac{n-1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \cdots & \frac{n-1}{n} \end{bmatrix} \\
\boldsymbol{P}_1 =&\frac{1}{2}\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \\
\boldsymbol{P}_2 =&\begin{bmatrix} 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{n-3}{n-2} & -\frac{1}{n-2} & \cdots & -\frac{1}{n-2} \\ 0 & 0 & -\frac{1}{n-2} & \frac{n-3}{n-2} & \cdots & -\frac{1}{n-2} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & -\frac{1}{n-2} & -\frac{1}{n-2} & \cdots & \frac{n-3}{n-2} \end{bmatrix} \\
\boldsymbol{P}_3 =&\begin{bmatrix} \frac{n-2}{2n} & \frac{n-2}{2n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \frac{n-2}{2n} & \frac{n-2}{2n} & -\frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & -\frac{1}{n} & \frac{2}{n(n-2)} & \frac{2}{n(n-2)} & \cdots & \frac{2}{n(n-2)} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{n} & -\frac{1}{n} & \frac{2}{n(n-2)} & \frac{2}{n(n-2)} & \cdots & \frac{2}{n(n-2)} \end{bmatrix}
\end{aligned}
$$

are symmetric and idempotent matrices, and $\boldsymbol{J}$ is a $n \times n$ unit matrix and their ranks are their traces, that is, rank($\boldsymbol{P}$)= tr($\boldsymbol{P}$) $= n-1$, rank($\boldsymbol{P}_1$)= tr($\boldsymbol{P}_1$) $= 1$, rank($\boldsymbol{P}_2$)= tr($\boldsymbol{P}_2$) $= n-3$ and rank($\boldsymbol{P}_3$)= tr($\boldsymbol{P}_3$) $= 1$. As $\mathcal{Q} = \boldsymbol{x}^\top\boldsymbol{P}_2\boldsymbol{x} + \boldsymbol{x}^\top\boldsymbol{P}_3\boldsymbol{x}$, by Fisher-Cochran theorem, $\mathcal{Q}/\sigma^2 \sim \chi^2_{n-2}$.

For the independence, note that $\mathcal{Q}$ is a function that depends on X_1, X_2, $\ldots$, X_{j-1}, X_{j+1}, $\ldots$, X_{k-1}, X_{k+1}, $\ldots$, X_n and on X_j and X_k only through the sum $X_j + X_k$ and, therefore, we have to prove only that $X_j - X_k$ and $X_j + X_k$ are independently distributed. Since both variables are normal, as a consequence of being linear combinations of normal variables, it suffices to show that their covariance is zero. That is

$$
\begin{aligned}
\mathrm{Cov}(X_j - X_k, X_j + X_k) =&\mathrm{Cov}\,(X_j, X_j) + \mathrm{Cov}(X_j, X_k) - \mathrm{Cov}(X_k, X_j) - \mathrm{Cov}\,(X_k, X_k) \\
=&\mathrm{Var}(X_j) - \mathrm{Var}(X_k) = 0.
\end{aligned}
$$

So, $\mathcal{Q}$ and $X_j - X_k$ are independently distributed and, therefore, $\mathrm{T} \sim t_{n-2}$. $\square$

If now we consider a particular sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\top$, there are $n(n-1)$ different choices of j and k, each pair defining a value of U'. These $n(n-1)$ values may be arranged in descending order and denoted by

$$u'_{[1]} \geq u'_{[2]} \geq \cdots \geq u'_{[n(n-1)]}.$$

Observe that $u'_{[1]} = (x_{(n)} - x_{(1)})/s$ is the internally studentized range of that particular sample. This construction defines the random variables $U'_{[1]}$, $U'_{[2]}$, $\ldots$, $U'_{[n(n-1)]}$. The random variable U' may be written as a mixture of the variables $U'_{[i]}$, $i = 1, 2, \ldots, n(n-1)$. As j and k are arbitrary in the definition of U', we choose uniformly a number i in $\{1, 2, \ldots, n(n-1)\}$ and take the variable $U'_{[i]}$. This can be expressed formally as:

$$U' = Z_1 U'_{[1]} + Z_2 U'_{[2]} + \cdots + Z_{n(n-1)} U'_{[n(n-1)]},$$

where $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_{n(n-1)})^\top$ is a multivariate random variable (ANDERSON, 2003) assuming values in the sample space $\Omega = \{\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_{n(n-1)}\}$, uniformly, with $\boldsymbol{\omega}_i = (0, 0, \ldots, 1, \ldots, 0)^\top$, a vector $n(n-1) \times 1$ that has 1 in the i-th position and zeroes elsewhere. So, it is easy to see that the distribution of U' may be expressed as the (Florescu and Tudor, 2014) mixture:

$$f_{U'}(u) = \frac{1}{n(n-1)} \sum_{i=1}^{n(n-1)} f_{U'_{[i]}}(u).$$

We know $f_{U'}(.)$ and we are interested in the distribution $f_{U'_{[1]}}(.)$. The idea, then, is, if there is no overlap between $f_{U'_{[1]}}(.)$ and $f_{U'_{[2]}}(.)$, the upper quantiles of $U'_{[1]}$ can be approximated by the upper quantiles of U'. For this we have the proposition,

Proposition 2.4. *The maximum value of $U'_{[2]}$ is $\sqrt{3(n-1)/2}$.*

Proof. A detailed proof may be found in the appendix of the original article (David et al., 1954). For the sake of completeness, we outline it here. The key feature of the statistic $U' = (X_j - X_k)/S$ is its invariance for transformations of the form $\alpha X + \beta$, that is, it does not depend on origin or scale. So, we may assume, without loss of generality, that, for any sample $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\top$ arranged in ascending order of magnitude, $x_1 = 0$, $x_{n-1} = 1$. So, the maximum of $u_{[2]} = (x_{n-1} - x_1)/s$ $= 1/s$, can be found by minimizing s, restricted to samples with $x_1 = 0$, $x_{n-1} = 1$, in three steps. First consider all samples with a given x_n. Then the possible values for $\bar{x}$ are in the interval $[(1 + x_n)/n; \infty)$, and $\bar{x} = (1 + x_n)/n$ corresponds to the configuration

$$\begin{array}{ccc}
x_1, \ldots, x_{n-2} & x_{n-1} & x_n \\
\downarrow & \downarrow & \downarrow \\
0 & 1 & \text{fixed}
\end{array}$$

Now, in the set of all samples with a given $\bar{x}$, clearly the minimum of s is attained with the configuration

$$\begin{array}{cccc} x_1 & x_2, \dots, x_{n-2} & x_{n-1} & x_n \\ \downarrow & \downarrow & \downarrow & \downarrow \\ 0 & \bar{x} & 1 & \text{fixed} \end{array}$$

Then we may consider only samples with the above configuration. Among them, it is clear that the minimum s corresponds to $x_n = 1$ and, in this case, the minimum in given by the configuration

$$\begin{array}{ccc} x_1 & x_2, \dots, x_{n-2} & x_{n-1}, x_n \\ \downarrow & \downarrow & \downarrow \\ 0 & 2/3 & 1 \end{array}$$

It is clear that the minimum of the s is attained when $x_2 = x_3 = \ldots = x_{n-2} = \bar{x}$. In this case $s^2 = 2/[3(n-1)]$ and

$$u'_{[2]} = \frac{1-0}{\sqrt{\frac{2}{3(n-1)}}} = \sqrt{\frac{3(n-1)}{2}}.$$

□

So, as $f_{U'_{[2]}}(u) = 0$ for $u > \sqrt{3(n-1)/2}$, and

$$f_{U'}(u) = \frac{1}{n(n-1)} \sum_{i=1}^{n(n-1)} f_{U'_{[i]}}(u)$$

for $u > \sqrt{3(n-1)/2}$, the contribution for $f_{U'}(u)$ is given only by $U'_{[1]}$, that is, $f_{U'}(u) = 1/[n(n-1)] f_{U'_{[1]}}(u)$. It follows that upper quantile ξ_p of $f_{U'_{[1]}}(u)$ coincides with the upper quantile $u_{p/[n(n-1)]}$ of $f_{U'}(u)$, which, as

$$U' = g(T) = \frac{T\sqrt{2(n-1)}}{\sqrt{T^2 + n - 2}} \tag{1}$$

is an strictly increasing function, is equal to the function $g(.)$ applied to the quantile $t_{p/[n(n-1)]}$ of a Student t distribution with $n-2$ degrees of freedom. For quantiles near the quantity $\sqrt{3(n-1)/2}$,

$$\xi_p^2 = u^2_{\frac{p}{n(n-1)}} = 2(n-1)\frac{t^2_{p/[n(n-1)];n-2}}{n-2+t^2_{p/[n(n-1)];n-2}} \tag{2}$$

is still a good approximation, where $t_{p/[n(n-1)];n-2}$ is the $100p/[n(n-1)]\%$ upper quantile of the Student's t distribution with $n-2$ degrees of freedom. This expression is valid only for computing upper quantiles.

The distribution of U has upper bound of $\sqrt{2(n-1)}$ and lower bound of $2\sqrt{(n-1)/n}$, case n is even, and $2\sqrt{n/(n+1)}$, case n is odd. Note the overlap of $U'_{[1]}$ and $U'_{[2]}$ distributions in the interval above. Thomson (1955) justifies these limits only by the exhibition of the corresponding sampling configurations. As the authors could not find a proof for this anywhere, it is given here:

The upper bound: Consider the sample $x_1 \leq x_2 \leq \ldots \leq x_n$, with $x_1 = 0$ and $x_n = 1$. Then $w = 1$ and the upper bound for $U = W/S$ depends only on S. In this case,

$$
\begin{aligned}
\bar{x} =& \frac{1 + \sum\limits_{j=2}^{n-1} x_j}{n}, \\
s^2 =& \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})^2 \\
=& \frac{1}{n-1} \left[\bar{x}^2 + \sum_{j=2}^{n-1} (x_j - \bar{x})^2 + (1 - \bar{x})^2 \right] \text{and, for } i = 2, \ldots, n-1, \\
\frac{\partial s^2}{\partial x_i} =& \frac{1}{n-1} \left[\frac{2}{n}\bar{x} - \frac{2}{n} \sum_{\substack{j=2 \\ j \neq i}}^{n-1} (x_j - \bar{x}) + 2\,(x_i - \bar{x}) \left(1 - \frac{1}{n}\right) - \frac{2}{n}\,(1 - \bar{x}) \right].
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial s^2}{\partial x_i} = 0 \quad & \Rightarrow \quad 0 = \bar{x} - \sum_{\substack{j=2 \\ j \neq i}}^{n-1} (x_j - \bar{x}) + (x_i - \bar{x})\,(n-1) - (1 - \bar{x}) \\
& \Rightarrow \quad \bar{x} + (n-3)\,\bar{x} + (n-1)\,x_i + \bar{x} = 1 + \sum_{\substack{j=2 \\ j \neq i}}^{n-1} x_j + (n-1)\,\bar{x} \\
& \Rightarrow \quad (n-1)\,x_i = 1 + \sum_{\substack{j=2 \\ j \neq i}}^{n-1} x_j \quad \Rightarrow \quad x_i = \bar{x}.
\end{aligned}
$$

$$
\frac{\partial^2 s^2}{\partial x_i^2} = \frac{1}{n-1} \left[\frac{2}{n^2} + \frac{2(n-3)}{n^2} + 2\left(1 - \frac{1}{n}\right)^2 + 2\left(-\frac{1}{n}\right)^2 \right] = \frac{2}{n} > 0.
$$

Also, $\partial^2 s^2 / \partial x_i \partial x_k = 0$, $i \neq k$, then the Hessian matrix is a diagonal matrix with element $h_{ii} = 2/n$. So, the determinant of the Hessian matrix is $(2/n)^{n-2} > 0$, indicating that the s^2 has a minimum value for $x_i = \bar{x}$. Therefore, the configuration that maximizes $U = W/S$ is $x_i = \bar{x}$, $i = 2, \ldots, n-1$. In this case, $\bar{x} = 1/2$ and $s = \sqrt{1/[2(n-1)]}$ and, so, $U = W/S \leq \sqrt{2(n-1)}$.

For another proof for the upper bound, consider (1):

$$\frac{T\sqrt{2\,(n-1)}}{\sqrt{T^2+n-2}} = \frac{\sqrt{2\,(n-1)}}{\sqrt{1+\frac{n-2}{T^2}}} < \sqrt{2\,(n-1)}, \qquad (n>2).$$

The lower bound: As s^2 has only one null derivative in the interior of the hypercube $0 \leq x_2 \leq \ldots \leq x_{n-1} \leq 1$, other existing extremes will be at vertices, that is, $x_i = 0,\, 1,\, i = 2, \ldots, n-1$. Suppose k 1's and $(n-2-k)$ 0's, $k = 0, \ldots, n-2$. In this case $\bar{x} = (1+k)/n$ and

$$\begin{aligned} s^2 =& \frac{1}{n-1}\sum_{j=1}^{n}\left(x_j - \frac{1+k}{n}\right)^2 \\ =& \frac{1}{n-1}\left[\left(\frac{1+k}{n}\right)^2 + (k+1)\left(1-\frac{1+k}{n}\right)^2 + (n-2-k)\left(\frac{1+k}{n}\right)^2\right] \\ =& \frac{1}{n-1}\left[(k+1) - n\left(\frac{1+k}{n}\right)^2\right]. \end{aligned}$$

Supposing k continuous and differentiating

$$\frac{d}{dk}s^2 = 0 \qquad \Rightarrow \qquad 0 = \frac{1}{n-1}\left[1-2\left(\frac{1+k}{n}\right)\right] \qquad \Rightarrow \qquad k = \frac{n}{2}-1.$$

Case n is odd then $n = 2\alpha$, where α is an integer, $k = \frac{2\alpha}{2} - 1 = \alpha - 1 \qquad \Rightarrow$ $k = \frac{n}{2} - 1$.

Thus

$$S^2 \leq \frac{1}{n-1}\left[\frac{n}{2} - n\left(\frac{1}{2}\right)^2\right] = \frac{n}{4\,(n-1)}.$$

and

$$U = \frac{W}{S} \geq \frac{1}{\sqrt{\frac{n}{4(n-1)}}} = 2\sqrt{\frac{n-1}{n}}.$$

Case n is even then $n = 2\alpha+1$ $k = \frac{2\alpha+1}{2} - 1 = \alpha - \frac{1}{2} \approx \alpha \qquad \Rightarrow \; k = \frac{n-1}{2}$.

So

$$S^2 \leq \frac{1}{n-1}\left[\left(\frac{n-1}{2}+1\right) - n\left(\frac{1+\frac{n-1}{2}}{n}\right)^2\right] = \frac{n+1}{4n}$$

and

$$U = \frac{W}{S} \geq \frac{1}{\sqrt{\frac{n+1}{4n}}} = 2\sqrt{\frac{n}{n+1}}.$$

2.2 A new approximation to the upper tail based on the maximum of U'

Our approach to the distribution of $U = W/S$ is based on the distribution of the maximum of U' and should be used only to obtain upper quantiles. Since we have $n(n-1)$ statistics, the distribution of the maximum is

$$P(U \leq u) = P\left(\max(U') \leq u\right) = \left[F_{U'}(u)\right]^{n(n-1)}.$$

Following the proof of proposition 2.4, we consider that different U''s are identical and nearly independent, since their distributions have a small overlap.

If we use the relationship given by (1) we get

$$P(U \leq u) = \left[F_T\left(\frac{\sqrt{n-2}u}{\sqrt{2(n-1)-u^2}}\right)\right]^{n(n-1)}, \qquad (3)$$

where $F_T(t)$ is the cumulative distribution function of the Student's t with $n-2$ degrees of freedom.

Upper quantile of the distribution of U can be obtained by inverting (3) for a given probability $1-\alpha$. Thus, considering $p = (1-\alpha)^{1/[n(n-1)]}$, we can obtain quantiles $u_{_{1-\alpha}}$ by

$$u^2_{_{1-\alpha}} = 2(n-1)\frac{t^2_{p;n-2}}{n-2+t^2_{p;n-2}}, \qquad (4)$$

where $t_{p;n-2}$ is the $100p\%$ quantile from the Student's t distribution with $n-2$ degrees of freedom.

2.3 Approximation based on Monte Carlo simulations

The above approximations are valid only when computing upper quantiles of the internally studentized range. Pearson's curves, used to approach the true distribution, must be obtained for each sample size and they are not available in David, Hartley e Pearson (1954). Also, only some combinations of lower percentile points and sample sizes are available, with other lower points obtained by interpolations. To overcome this limitations we propose here an approach based on Monte Carlo simulations. For this, a random sample X_1, X_2, …, X_n of size n from the standard normal distribution is generated by the Box-Müller algorithm. Then we construct a realization of $U = W/S$, denoted by u. This process is repeated to obtain a Monte Carlo sample of size N from the distribution of U. The sample size N should be chosen to achieve some previously fixed precision. The error bound of the Monte Carlo simulation is proportional to $1/\sqrt{N}$. Let u_1, u_2, …, u_N be this sample. To compute the cumulative distribution function for a given value ξ, we compute

$$p = \frac{\sum_{i=1}^{N} I(u_i \leq \xi)}{N}, \qquad (5)$$

where $I(u_i \leq \xi)$ is the indicator function that returns 1, if $u_i \leq \xi$, and 0, otherwise.

Given the cumulative distribution p, we arrange the elements in ascending order and pick up the Np-th element of the resulting vector. Since Np is not necessarily an integer we define $j = \lfloor Np \rfloor$, the largest integer less than or equal to Np. Considering, in this context, that $u_{(1)}$, $u_{(2)}$, $\ldots$, $u_{(N)}$ represents the internally studentized range Monte Carlo sample of size N, arranged in ascending order, the quantile ξ_p is computed by

$$\xi_p = u_{(j)}, \text{ where } j = \lfloor Np \rfloor. \qquad (6)$$

The R (R Core Team, 2018) codes to generate random samples of the internally studentized range distribution and to compute cumulative distribution functions (5) and quantiles (6) are available in the appendix B. Illustrative examples are presented below. In the next section, we compare our Monte Carlo approach with the other two approaches with respect to chosen sample sizes and percentiles. Besides that, we show some lower tail quantiles from the distribution of the internally studentized range.

3 Comparison of approximations

In a first stage we obtain quantiles from both proposed approximations to compare their accuracy with that of the approach of David, Hartley e Pearson (1954). Quantiles from the approximations (2) and (4) are compared for different combinations of n and cumulative probability $1-\alpha$. Values of $n = 3(1)20$, $20(5)50$, $50(10)100$, 500, $1,000$, $10,0000$ and $1-\alpha = 0.900$, 0.950, 0.990 and 0.995 are considered. Results are shown in Table 1. There is a very good agreement between both approximations, even for small sample sizes. They should be used only for determining upper tails, since they are based on quantities that are not independently distributed and the lower quantiles are very inaccurate.

The upper and lower quantiles from the above approximations are compared to those from Monte Carlo simulations, considering some combinations of $1-\alpha$ and n (Table 2). Lower quantiles are computed for $1-\alpha = 0.005$, 0.01, 0.05 and 0.10. The results for the upper quantiles should be compared to those showed in Table 1. There is a very accurate agreement among all approximations, specially for small sample sizes ($n \leq 20$). So, for large n, the Monte Carlo simulations show larger differences for the upper quantiles among the three methods. It could be considered that the differences are due to the Monte Carlo errors. For example, with $n = 100$, using $N = 100,000$ Monte Carlo simulations, the 95% upper quantile is 5.909 (Table 2). If we use $N = 3,000,000$ monte Carlo simulations, the 95% upper quantile is 5.903, what is neither close to the (DAVID; HARTLEY; PEARSON, 1954) approximation, 5.990, nor to our approximation, 5.983.

For very large samples, we expect that the range and the standard deviation are asymptotically independently distributed. Therefore, we can use the externally studentized range distribution to computes approximate quantiles. For $n = 10,000$, using this approach, we find that the 95% upper quantile is 8.481. This value is

Table 1 - Percentage points of the distribution of the ratio of range to the standard deviation $U = W/S$ in sample size of n from a normal population using approximations of David, Hartley e Pearson (1954) and of the maximum of the studentized sample pair differences

	David et al.				Distribution of the Maximum			
n	0.90	0.95	0.99	0.995	0.90	0.95	0.99	0.995
3	1.997	1.999	2.000	2.000	1.997	1.999	2.000	2.000
4	2.409	2.429	2.445	2.447	2.407	2.429	2.445	2.447
5	2.712	2.755	2.803	2.813	2.708	2.754	2.803	2.813
6	2.949	3.012	3.095	3.115	2.944	3.010	3.095	3.115
7	3.143	3.222	3.338	3.369	3.137	3.220	3.338	3.369
8	3.308	3.399	3.543	3.585	3.300	3.396	3.543	3.585
9	3.449	3.552	3.720	3.771	3.441	3.548	3.720	3.771
10	3.574	3.685	3.875	3.935	3.565	3.681	3.874	3.935
11	3.685	3.803	4.012	4.079	3.675	3.799	4.011	4.079
12	3.785	3.909	4.134	4.208	3.774	3.905	4.133	4.208
13	3.875	4.005	4.244	4.325	3.864	4.001	4.243	4.325
14	3.958	4.093	4.344	4.431	3.947	4.088	4.343	4.430
15	4.034	4.173	4.435	4.527	4.023	4.168	4.435	4.527
16	4.104	4.247	4.519	4.616	4.093	4.242	4.519	4.615
17	4.170	4.316	4.597	4.698	4.158	4.311	4.596	4.697
18	4.231	4.380	4.669	4.774	4.219	4.375	4.668	4.773
19	4.288	4.440	4.737	4.844	4.276	4.435	4.736	4.844
20	4.342	4.496	4.800	4.911	4.330	4.491	4.799	4.910
25	4.571	4.734	5.064	5.187	4.558	4.728	5.063	5.187
30	4.751	4.921	5.268	5.401	4.738	4.915	5.267	5.400
35	4.899	5.073	5.433	5.573	4.886	5.067	5.432	5.572
40	5.024	5.201	5.571	5.715	5.010	5.194	5.570	5.715
45	5.131	5.311	5.688	5.837	5.117	5.304	5.687	5.836
50	5.226	5.407	5.790	5.942	5.212	5.400	5.789	5.941
60	5.384	5.568	5.960	6.116	5.370	5.562	5.959	6.116
70	5.515	5.700	6.098	6.257	5.500	5.693	6.097	6.257
80	5.624	5.811	6.213	6.375	5.610	5.804	6.212	6.374
90	5.719	5.906	6.311	6.475	5.705	5.899	6.310	6.474
100	5.802	5.990	6.397	6.562	5.787	5.983	6.396	6.561
500	6.905	7.087	7.492	7.660	6.891	7.081	7.491	7.659
1,000	7.309	7.485	7.880	8.044	7.295	7.479	7.879	8.043
10,000	8.475	8.633	8.988	9.137	8.463	8.627	8.987	9.136

closer to the Monte Carlo quantile, showing that this approach is more accurate than the other two above. A second advantage of this method is the possibility to compute lower quantiles, what is not possible with the two previous methods. Finally, it should be noticed that the number of Monte Carlo simulations in Table 2 was reduced to $N = 30,000$, due to computer memory limitation.

Table 2 - Percentage points of the distribution of the ratio of range to the standard deviation $U = W/S$ in sample size of n from a normal population using Monte Carlo approximation with $N = 100,000$

	Lower quantiles				Upper quantiles			
n	0.005	0.01	0.05	0.10	0.90	0.95	0.99	0.995
3	1.739	1.737	1.758	1.782	1.997	1.999	2.000	2.000
4	1.817	1.851	1.979	2.054	2.408	2.429	2.445	2.447
5	1.989	2.023	2.140	2.220	2.712	2.754	2.804	2.813
6	2.114	2.156	2.281	2.360	2.948	3.012	3.095	3.115
7	2.219	2.263	2.401	2.485	3.145	3.224	3.341	3.368
8	2.309	2.355	2.503	2.591	3.308	3.401	3.547	3.588
9	2.385	2.437	2.592	2.685	3.450	3.552	3.723	3.772
10	2.459	2.513	2.673	2.769	3.575	3.682	3.876	3.940
15	2.745	2.799	2.969	3.074	4.026	4.171	4.437	4.525
20	2.927	2.987	3.177	3.289	4.318	4.487	4.796	4.910
25	3.082	3.139	3.335	3.454	4.535	4.714	5.064	5.184
30	3.207	3.274	3.472	3.590	4.698	4.893	5.247	5.389
35	3.312	3.376	3.578	3.698	4.840	5.043	5.427	5.563
40	3.399	3.473	3.675	3.798	4.954	5.157	5.560	5.711
45	3.479	3.546	3.754	3.878	5.055	5.266	5.687	5.831
50	3.550	3.622	3.833	3.957	5.144	5.353	5.775	5.935
60	3.671	3.742	3.955	4.082	5.291	5.511	5.935	6.090
70	3.774	3.851	4.065	4.193	5.410	5.632	6.074	6.244
80	3.867	3.945	4.158	4.288	5.516	5.736	6.179	6.344
90	3.951	4.023	4.237	4.366	5.606	5.829	6.288	6.443
100	4.013	4.084	4.312	4.443	5.684	5.909	6.354	6.519
500	5.069	5.143	5.360	5.491	6.721	6.952	7.421	7.610
1,000	5.498	5.576	5.792	5.916	7.112	7.336	7.791	7.983
10,000*	6.826	6.895	7.084	7.196	8.268	8.467	8.875	9.044

* $N = 30,000$.

4 Exact distributions

The exact distribution for the studentized range is known only for sample sizes $n = 2, 3, 4$. For $n = 2$,

$$\frac{W}{S} = \frac{X_{(2)} - X_{(1)}}{\sqrt{\left(X_1 - \bar{X}\right)^2 + \left(X_2 - \bar{X}\right)^2}} = \frac{X_{(2)} - X_{(1)}}{\sqrt{\left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2}} = \sqrt{2},$$

with any parent distribution.

For $n = 3$, Thomson (1955), based in Lieblein (1952), presents, without the proof, the expression of the distribution for the normal case. As the authors could not find that proof anywhere, it is presented here, for the normal case. For the uniform distribution, a new result was obtained and was showed in the sequence of the normal case.

The normal case: Suppose $X_{(1)} < X_{(2)} < X_{(3)}$ an ordered sample from the same normal population. As W/S is invariant for translation and homothety, we may consider samples with configuration $0 < Y = (X_{(2)} - X_{(1)})/(X_{(3)} - X_{(1)}) < 1$. If now, following the work of Lieblein (1952), we denote by (X'', X') the closest of the pairs $\left(X_{(1)}, X_{(2)}\right)$ and $\left(X_{(2)}, X_{(3)}\right)$ we may define

$$Y_1 = X' - X'' = \begin{cases} Y & \text{case } X_{(2)} - X_{(1)} \leq X_{(3)} - X_{(2)} \\ 1 - Y & \text{case } X_{(2)} - X_{(1)} > X_{(3)} - X_{(2)}. \end{cases}$$

Observe that $0 < Y_1 < 1/2$ and each value of Y_1 may came from two different configurations, $\left(X_{(1)}, X_{(2)}, X_{(3)}\right)$ or $\left(X^*_{(1)}, X^*_{(2)}, X^*_{(3)}\right)$, with $X^*_{(3)} - X^*_{(1)} = X_{(3)} - X_{(1)}$ and $X_{(3)} - X_{(2)} > X_{(2)} - X_{(1)} = X^*_{(3)} - X^*_{(2)} < X^*_{(2)} - X^*_{(1)}$. In this case,

$$\begin{aligned} \frac{W}{S} =& \frac{1}{\sqrt{\frac{1}{2}\left[\left(\frac{1+Y_1}{3}\right)^2 + \left(Y_1 - \frac{1+Y_1}{3}\right)^2 + \left(1 - \frac{1+Y_1}{3}\right)^2\right]}} \\ =& \frac{3\sqrt{2}}{\sqrt{(1+Y_1)^2 + (2Y_1 - 1)^2 + (2 - Y_1)^2}} \\ =& \frac{3\sqrt{2}}{\sqrt{1 + 2Y_1 + Y_1^2 + 4Y_1^2 - 4Y_1 + 1 + 4 - 4Y_1 + Y_1^2}} = \frac{3\sqrt{2}}{\sqrt{6}\sqrt{1 - Y_1 + Y_1^2}} \\ =& \frac{\sqrt{3}}{\sqrt{1 - Y_1 + Y_1^2}} = g\left(Y_1\right). \end{aligned}$$

Thus

$$z = \frac{\sqrt{3}}{\sqrt{1 - y_1 + y_1^2}} \qquad \Rightarrow \qquad 0 = 1 - y_1 + y_1^2 - \frac{3}{z^2}$$

and

$$\begin{aligned} \Rightarrow \qquad y_1 = g^{-1}\left(z\right) =& \frac{1 - \sqrt{1 - 4\left(1 - \frac{3}{z^2}\right)}}{2} = \frac{1 - \sqrt{3}\sqrt{4 - z^2}}{2}. \\ \frac{d}{dz} y_1 =& -\frac{1}{4}\left[1 - 4\left(1 - \frac{3}{z^2}\right)\right]^{-\frac{1}{2}}\left(-24\, z^{-3}\right) \\ =& \frac{2\sqrt{3}}{z^2\sqrt{4 - z^2}}. \end{aligned}$$

Therefore,

$$\begin{aligned} f_{W/S}\left(z\right) =& \left|\frac{d}{dz}g^{-1}\left(z\right)\right| f_{Y_1}\left(g^{-1}\left(z\right)\right) \\ =& \frac{2\sqrt{3}}{z^2\sqrt{4-z^2}} \quad f_{Y_1}\left(\frac{1}{2}-\frac{\sqrt{3}}{2}\frac{\sqrt{4-z^2}}{z}\right). \end{aligned}$$

According to Lieblein (1952), case $n=3$, $f_{Y_1}\left(y_1\right) = 3\sqrt{3}/[\pi\left(1-y_1+y_1^2\right)]$. So, for $n=3$,

$$\begin{aligned} f_{W/S}\ \left(z\right) =& \frac{2\sqrt{3}}{z^2\sqrt{4-z^2}} \frac{3\sqrt{3}}{\pi\left[1-\left(\frac{1}{2}-\frac{\sqrt{3}}{2}\frac{\sqrt{4-z^2}}{z}\right)+\left(\frac{1}{2}-\frac{\sqrt{3}}{2}\frac{\sqrt{4-z^2}}{z}\right)^2\right]} \\ =& \frac{6}{\pi\sqrt{4-z^2}}, \\ F_{W/S}\ \left(z\right) =& \int\limits_{\sqrt{3}}^{z} \frac{6}{\pi\sqrt{4-u^2}}du \\ =& 1-\frac{6}{\pi}\arccos\left(\frac{z}{2}\right) \qquad\qquad \text{and, finally,} \\ z =& 2\cos\left[\frac{\pi}{6}(1-p)\right], \end{aligned}$$

as obtained by Thomson (1955), where p is the cumulative probability.

For $n=3$, a new result was obtained for the exact distribution in the uniform case. The distribution followed and the proof are presented below.

The uniform case: As already shown in the normal case, for $n=3$,

$$\begin{aligned} \frac{W}{S} =& \frac{\sqrt{3}}{\sqrt{1-Y_1+Y_1^2}} = g\left(Y_1\right), \\ f_{W/S}\left(z\right) =& \left|\frac{d}{dz}g^{-1}\left(z\right)\right| f_{Y_1}\left(g^{-1}\left(z\right)\right). \end{aligned}$$

As Lieblein (1952) showed that Y_1 is uniformly distributed in the interval $\left(0,\frac{1}{2}\right)$,

$$f_{W/S}\ \left(z\right) = 2\left|\frac{d}{dz}g^{-1}\left(z\right)\right| = \frac{4\sqrt{3}}{z^2\sqrt{4-z^2}}, \qquad\qquad \left(\sqrt{3}<z<2\right),$$

and with cumulative distribution function and quantile function given by

$$\begin{aligned} F_{W/S}(z) =& 1-\frac{\sqrt{3}\sqrt{4-z^2}}{z}, \\ z =& \frac{2\sqrt{3}}{\sqrt{3+(1-p)^2}}, \end{aligned}$$

where $0<p<1$ is the cumulative probability.

5 Application: a new test for normality

A new normality test is proposed, as already suggested by David, Hartley e Pearson (1954). Under normality, the observed U statistic $u = w/s$ is a typical value of the internally range distribution, where w and s are the observed values of the sample range and standard deviation, respectively. Therefore, given a random sample of size n, the u statistic value should be in the interval $[u_{\alpha/2}; u_{1-\alpha/2}]$, under the null hypothesis of normality, where u_p is the $100p\%$ quantile from the internally studentized range distribution, that is $P(u_{\alpha/2} \leq U \leq u_{1-\alpha/2}) = 1 - \alpha$. Hence, to test the normality hypothesis, the test statistic u is computed and the p-value is computed from the internally studentized range distribution by

$$p\text{-value} = 2 \min \left[P(U \leq u); 1 - P(U \leq u) \right],$$

since the distribution of U is asymmetric. This p-value is computed by simulation, as described in subsection 2.3.

The performance of the normality test based on the internally studentized range distribution was appraised by Monte Carlo simulations. Further, its relative performance is compared with the Shapiro-Wilk test using the toolkit of (ROYSTON, 1993). The proportions of rejections under H_0 were computed to evaluate the type I error rates of both tests, considering several sample sizes. The power was also evaluated for several sample sizes and probability distributions. The sample sizes considered were 50, 100 and 500. The non-normal distributions were the Student t with $\nu = 1$ and $\nu = 30$ degrees of freedom, the uniform, the beta with parameters $\alpha = 5$ and $\beta = 1$ and the Pearson VII distribution with $m = 2$, location parameter $\lambda = 0$ and scale parameter $\alpha = 5$. A total of $5,000$ Monte Carlo simulations were performed. Also, $30,000$ values were simulated to obtain the null distribution of the internally studentized range.

Type I error rates are shown in Table 3. The proposed internally studentized distribution test (ISRD) and the Shapiro-Wilk (SW) test are both exact, since their type error rates are equal to the nominal level α. The small differences were due to the Monte Carlo error. Therefore, the proposed test may have some advantage under some alternative distributions.

The power of the ISRD and SW tests are shown in Table 4. The performance of the ISRD test was very good for symmetric distributions like Student t and Pearson VII. For $n > 50$ its performance was almost equal of the SW test, considered one of the most powerful test. For the uniform case, where the distribution besides being symmetric is also platykurtic, the ISRD test performed much better, mainly for small sample sizes.

For the beta distribution, however, the ISRD test showed smaller power values than the SW. This distribution is very asymmetric, with the chosen parameters. The Pearson VII distribution, where our test was also less powerful, was a symmetric and leptokurtic distribution. We did not intend that our test performed better than the SW test. It was considered here for illustrative purpose only.

Table 3 - Type I error rates computed in $5,000$ Monte Carlo simulations of the internally studentized range distribution test (ISRD) and the Shapiro-Wilk test (SW) for normality, considering several n and α

		Tests	
α	n	ISRD	SW
0.05	50	0.0470	0.0498
	100	0.0504	0.0496
	500	0.0500	0.0484
0.10	50	0.1042	0.0990
	100	0.0958	0.0960
	500	0.1002	0.1026

Table 4 - Power computed in $5,000$ Monte Carlo simulations of the internally studentized range distribution test (ISRD) and the test of Shapiro-Wilk (SW) for normality, considering several alternative distributions, n and α

		$\alpha = 0.05$		$\alpha = 0.10$	
Distribution	n	ISRD	SW	ISRD	SW
t with $\nu = 1$	50	0.9678	0.9974	0.9808	0.9984
t with $\nu = 30$		0.0658	0.0766	0.1200	0.1330
Uniform		0.9542	0.7470	0.9852	0.8840
Beta		0.1882	0.9910	0.2792	0.9982
Pearson VII		0.7870	0.8620	0.8404	0.8962
t with $\nu = 1$	100	0.9982	1.0000	0.9988	1.0000
t with $\nu = 30$		0.0786	0.0798	0.1424	0.1378
Uniform		1.0000	0.9952	1.0000	0.9992
Beta		0.2742	1.0000	0.3672	1.0000
Pearson VII		0.9506	0.9834	0.9670	0.9886
t with $\nu = 1$	500	1.0000	1.0000	1.0000	1.0000
t with $\nu = 30$		0.1334	0.1394	0.2102	0.2110
Uniform		1.0000	1.0000	1.0000	1.0000
Beta		0.5608	1.0000	0.6630	1.0000
Pearson VII		1.0000	1.0000	1.0000	1.0000

Conclusions

The statistical mathematics of the theory of the internally studentized range is complex and plentiful of original ideas. For sample size 3, this problem is related to the distribution of the closest pair in a sample of three observations. The geometric approach has shown to be helpful in clarifying some aspects of the theory.

The approximation using the distribution of the maximum of the studentized sample pair differences were very accurate even for small sample sizes. It should be used only for determining upper tails.

Acknowledgments

Authors would like to thank the CNPq for the financial support, reviewers and editors for their suggestions.

FERREIRA, D. F.; CHAVES, L. M.; SOUZA, D. J. Teoria da amplitude *estudentizada* internamente revisitada. *Rev. Bras. Biom.,* Lavras, v.36, n.4, p.802-826, 2018.

- *RESUMO: O presente artigo pretende revisitar a distribuição da razão da amplitude pelo desvio padrão amostral, conhecida como distribuição da amplitude estudentizada internamente, no caso normal. Esta distribuição é importante em várias áreas do conhecimento, como controle de qualidade e inferência, nos testes de falta de homogeneidade dos dados e de curtose. Um distribuição alternativa a que foi apresentada por David et al. (1954), baseada na distribuição do máximo é proposta. Exibiu-se prova detalhada da distribuição da amplitude estudentizada internamente no caso normal e para amostras de tamanho* 3*. Também foi apresentado um novo resultado: a distribuição no caso uniforme para amostras de tamanho* 3*.*

- *PALAVRAS-CHAVE: Estatística de ordem; o mais próximo de três; distribuição normal; distribuição uniforme*

References

ANDERSON, T. W. *An introduction to multivariate statistical analysis.* 3.ed. Hoboken: John Wiley and Sons, 2003. 722p.

CASELLA, G.; BERGER, R. L. *Statistical Inference.* 2.ed. Duxbury: Thomson Learning, 2002. 660p.

CURRIE, I. D. On the distribution of the studentized range in a single normal sample. *Scandinavian Journal of Statistics*, v.7, n.3, p.150–154, 1980.

DAVID, H. A.; HARTLEY, H. O.; PEARSON, E. S. The distribution of the ratio, in a single normal sample, of range to standard deviation. *Biometrika*, v.41, n.3/4, p.482–493, 1954.

HINKELMANN, K.; KEMPTHORNE, O. *Design and Analysis of Experiments, Introduction to Experimental Design.* New York: John Wiley and Sons, 2007. 632p.

LIEBLEIN, J. Properties of certain statistics involving the closest pair in a sample of three observations. *Journal Research of the National Bureau of Standards*, v.48, n.3, p.255–268, 1952.

R CORE TEAM. *R: A Language and Environment for Statistical Computing.* Vienna, Austria, 2018. Disponível em: ⟨https://www.R-project.org/⟩.

RAO, C. R. *Linear Statistical Inference and its applications.* New York: John Wiley, 2002. 625p.

RENCHER, A. C.; SCHAALJE, G. B. *Linear models in statistics.* 2. ed. New York: John Wiley and Sons, 2008. 672p.

ROYSTON, J. P. A toolkit for testing for non-normality in complete and censored samples. *The Statistician*, London, v.42, n.1, p.37–43, 1993.

THOMSON, G. W. Bounds for the ratio of range to standard deviation. *Biometrika*, New York, v.42, n.1–2, p.268–269, 1955.

Received on 16.05.2017.

Approved after revised on 22.02.2018.

A An algebraic proof for the decomposition of the sum of squares

We want to proof the identity

$$(n-1)s^2 = \frac{1}{2}\left(x_j - x_k\right)^2 + \sum_{i \neq j,k}^{n} \left(x_i - \bar{x}'\right)^2 + \frac{2(n-2)}{n}\left(\bar{x}' - \bar{x}''\right)^2 .$$

Without loss of generality, we can consider $(x_j,\, x_k) = (x_{n-1},\, x_n)$. Thus

$$\begin{aligned}(n-1)s^2 =& \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n^2}\left[n(n-1)\sum_{i=1}^{n} x_i^2 - 2n\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} x_i x_\ell\right] \\ =& \frac{1}{2}\left(x_{n-1} - x_n\right)^2 + \chi^2\sigma^2,\end{aligned}$$

where

$$\begin{aligned}\chi^2\sigma^2 =& \frac{1}{n^2}\left[n(n-1)\sum_{i=1}^{n} x_i^2 - 2n\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} x_i x_\ell\right] - \frac{1}{2}\left(x_{n-1} - x_n\right)^2 \\ =& \frac{n-1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} x_i x_\ell + \frac{n-2}{2n}\left(x_{n-1}^2 + x_n^2\right) + x_{n-1}x_n\end{aligned}$$

resulting on

$$\begin{aligned}\chi^2\sigma^2 =& \frac{n-1}{n}\sum_{i=1}^{n-2} x_i^2 - \frac{2}{n}\sum_{i=1}^{n-1}\sum_{\ell=i+1}^{n} x_i x_\ell + \frac{n-2}{2n}\left(x_{n-1}^2 + x_n^2\right) \\ &+ \frac{n-2}{n} x_{n-1}x_n. \end{aligned} \tag{7}$$

Since, in this particular case,

$$\bar{x}' = \frac{1}{n-2}\sum_{i=1}^{n-2} x_i \qquad \text{and} \qquad \bar{x}'' = \frac{x_{n-1} + x_n}{2},$$

then

$$\begin{aligned}\chi^2\sigma^2 =& \sum_{i=1}^{n-2}\left(x_i - \bar{x}'\right)^2 + \frac{2(n-2)}{n}\left(\bar{x}' - \bar{x}''\right)^2 \\ =& \frac{1}{(n-2)^2}\left[(n-2)(n-3)\sum_{i=1}^{n-2} x_i^2 - 2(n-2)\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2} x_i x_\ell\right] + \\ &+ \frac{2(n-2)}{n}\left(\frac{1}{n-2}\sum_{i=1}^{n-2} x_i - \frac{x_{n-1} + x_n}{2}\right)^2\end{aligned}$$

$$
\begin{aligned}
=&\frac{n-3}{n-2}\sum_{i=1}^{n-2}x_i^2-\frac{2}{n-2}\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2}x_ix_\ell+\\
&\qquad\qquad+\frac{2(n-2)}{n}\left(\frac{1}{n-2}\sum_{i=1}^{n-2}x_i-\frac{x_{n-1}+x_n}{2}\right)^2\\
=&\frac{n-3}{n-2}\sum_{i=1}^{n-2}x_i^2-\frac{2}{n-2}\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2}x_ix_\ell+\\
&+\frac{2}{n(n-2)}\left(\sum_{i=1}^{n-2}x_i\right)^2+\frac{n-2}{2n}\left(x_{n-1}+x_n\right)^2-\frac{2}{n}\left(x_{n-1}+x_n\right)\sum_{i=1}^{n-2}x_i
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{n-3}{n-2}\sum_{i=1}^{n-2}x_i^2-\frac{2}{n-2}\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2}x_ix_\ell+\frac{2}{n(n-2)}\sum_{i=1}^{n-2}x_i^2+\\
&\qquad+\frac{4}{n(n-2)}\left(\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2}x_ix_\ell\right)+\frac{n-2}{2n}\left(x_{n-1}+x_n\right)^2-\\
&\qquad-\frac{2}{n}\left(x_{n-1}+x_n\right)\sum_{i=1}^{n-2}x_i
\end{aligned}
$$

$$
\begin{aligned}
=&\frac{n-1}{n}\sum_{i=1}^{n-2}x_i^2-\frac{2}{n}\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2}x_ix_\ell+\frac{n-2}{2n}\left(x_{n-1}+x_n\right)^2-\\
&\qquad-\frac{2\left(x_{n-1}+x_n\right)}{n}\sum_{i=1}^{n-2}x_i,
\end{aligned}
$$

resulting on

$$
\begin{aligned}
\chi^2\sigma^2=&\frac{n-1}{n}\sum_{i=1}^{n-2}x_i^2-\frac{2}{n}\sum_{i=1}^{n-3}\sum_{\ell=i+1}^{n-2}x_ix_\ell+\frac{n-2}{2n}\left(x_{n-1}^2+x_n^2\right)\\
&+\frac{n-2}{n}x_{n-1}x_n. \qquad (8)
\end{aligned}
$$

Since (7)=(8), the proof is complete.

B R codes

```
# random samples from  the normal internally studentized range
# given the sample size n, where N is the number of samples
# simulated
rISR <- function(N=1, n)
{
   sr <- function(x) return((max(x)-min(x))/sd(x))
   return(apply(matrix(rnorm(N * n), N, n), 1, sr))
}
# function to obtain cumulative probabilities  from   the
# internally studentized range by Monte Carlo simulations.
# Given N: the number of simulations, n: the sample size
# and the quantile q > 0.
pISR <- function(q, n, N=1)
{
   x <- rISR(N, n)
   return(length(x[x <= q]) / N)
}
# function    to   obtain    quantiles      from    the
# internally studentized range by Monte Carlo simulations.
# Given N: the number of simulations, n: the sample size
# and the percentile 0< p < 1.
qISR <- function(p, n, N=1)
{
  x <- sort(rISR(N, n))
  i <- trunc(N*p)
  if (any(i<=0)) i[i<=0] <- 1
  q <- x[i]
  return(q)
}
# examples
# random sample
N <- 100000
n <-   10
x <- rISR(N, n)
hist(x)
# cumulative probabilities
q <- 3.685
pISR(q, n, N)
# quantiles
p <- c(0.95, 0.05)
qISR(p, n, N)
```