The Distribution Law of Sharma's Coefficient of Variation

Tovohery Josoa Michel[1], Totohasina André[2], Feno Daniel Rajaonasy[3]

[1]Thematic Doctoral School "Science, Culture, Society and Development" of the University of Toamasina – Madagascar

[2]Department of Mathematics and Computer Science, École Normale Supérieure pour l'Enseignement Technique, University of Antsiranana– Madagascar

[3]Department of Mathematics, Computer Science and Applications, Faculté de Droit, d'Économie, de Gestion, et de Mathématiques, Informatique et Applications, University of Toamasina – Madagascar

Abstract: *Pearson's coefficient of variation (CV) has now become one of the widely used measures of dispersion among researchers, despite its weaknesses and misleading nature. Faced with this problem, in 2011, Sharma and his research team attempted to find an alternative coefficient of dispersion via variance normalization. Although this new measure of dispersion appears to be better and more efficient than the Pearson coefficient of variation, its use is still limited. Thus, this article contributes to the study of the Sharma's coefficient of variation (CVS) distribution law in order to promote its application to different scientific fields.*

Keywords: Coefficient of variation, Gumbel distribution, Fisher distribution, Student distribution, product of two random variables

1. Introduction

Pearson's coefficient of variation (CV) is a measure of dispersion that is of concern to researchers today. This fact is characterized by a wide application of this measure in several scientific fields: currently, laboratories use it to calculate the uncertainty of measurement results [1]; Agricultural researchers use it to certify spreaders [2], professional technicians in production plants use it to make very sophisticated control charts to be able to act on the quality of products [3] and [4] ; medical researchers use it to evaluate or to quantify the pain felt by patients [5] and to process ultrasound images [6] ; psychology researchers have found its application in the choice of a decision under risk [7]; renewable energy researchers take advantage of it to measure the variability of global solar radiation between two different sites [8], etc.

However, some researchers point out that this measure poses a huge problem. It could be misleading in several cases, especially when the real variable under study takes its mean in the vicinity of 0 [9]. In addition, its mathematical interpretation is problematic. Therefore, in 2011, Sharma and his research team proposed a new standardized coefficient of variation that has better characteristics than the Pearson CV [10].

Despite the effectiveness of this new Sharma dispersion measure, it is not yet exploited by today's researchers. Thus, it is useful to study the distribution law of this normalized dispersion measure to promote its use in various fields. Hence, this context motivates our current interest in this paper.

The rest of our paper is organized as follows: section 2 will give the exact distribution law and the approximate distribution law of Sharma's coefficient of variation, and discuss their similarity as a function of the parameter of the exact distribution law; section 3 will end our presentation with a brief conclusion.

2. Materials and methods

In this subsection, we will first present the two measures of dispersion CV of Pearson and CVS and then we list the two distribution laws of the random variables that help us to easily find the exact distribution and the approximate distribution law of CVS.

2.1 Pearson's coefficient of variation

Definition 1. The Pearson's coefficient of variation of a series of values $X_1, \ldots, X_n$ is the quantity $CV(X)$ defined by:

$$CV(X) = \frac{S(X)}{\overline{X}} \qquad (1)$$

where $S(X) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ is the standard deviation of the series studied.

It seems difficult to find a direct mathematical interpretation of $CV(X)$ but the only interpretation shared by statisticians is: if $CV(X) > 0.3$, then the values of the series under study are scattered [11].

Proposition 1: The properties of the Pearson's coefficient of variation are as follows:
1) $CV(X)$ is not a normalized measure. According to [12], we have : $CV(X) \leq \sqrt{n-1}$.
2) $CV(X)$ is not stable by a linear transformation :
$$\forall\, a \in R, CV(X + a) \neq CV(X).$$
3) $CV(X)$ is stable by a multiplication with a scalar :
$$\forall\, k \in R, CV(k.X) = CV(X).$$
4) $CV(X)$ is not defined for a series with a mean of zero or very close to 0.
5) $CV(X)$ is influenced by the mean and gives two different values for two series with symmetric values.
Indeed, be, for example, $X = \{1\,;2\,;3\,;10\}$ and $Y = \{1\,;8\,;9\,;10\}$. We have : $\overline{X} = 4$, $\overline{Y} = 7$, $CV(X) = 0.88$ and $CV(Y) = 0.51$. This difference in coefficient of

variation is due to the difference in the mean. So, we have the property (5).
6) Distribution law of (X) : if X follows the normal distribution $N(\mu, \sigma)$ and if we suppose that $\delta = \frac{\sigma}{\mu}$, then $\frac{\sqrt{n}}{CV(X)}$ follows the non-centered Student's distribution law with $n - 1$ degrees of freedom and of decentering parameter equal to $\frac{\sqrt{n}}{\delta}$[3].

The properties (4) and (5) show that $CV(X)$ could be misleading. Furthermore, it is important to know that the variance is even better as a measure of dispersion than the Pearson coefficient of variation. This claim is further supported by the classification of heterohedasticity tests found in [13]. This work clearly shows that tests based on the comparison of variances are more sensitive in detecting possible heteroscedasticity in a linear regression model than those based on the comparison of Pearson's coefficients of variation.

To overcome this problem of Pearson's CV(X), in 2011, Sharma and his research team proposed in [14] a new coefficient of variation that seems better than the $CV(X)$ of Pearson.

2.2 Sharma's coefficient of variation

Sharma and his research team wanted to find a coefficient of variation with the properties of the variance. Thus, in order to normalize the variance, they applied the variance increase property found by Muilwijk in 1966 [14]: for any real random variable X, we have : $0 \leq S^2(X) \leq [Max(X) - X][X - MinX]$. Therefore, we have:
$$0 \leq \frac{S(X)}{\sqrt{[Max(X) - \overline{X}][\overline{X} - Min(X)]}} \leq 1.$$

Definition 2. The Sharma coefficient of variation of a real random variable X is the quantity $CVS(X)$ defined by:

$$CVS(X) = \begin{cases} \dfrac{S(X)}{\sqrt{[\,Max(X) - \overline{X}][\overline{X} - Min(X)]}}, & \text{if } Max(X) \neq \overline{X} \text{ or } Min(X) \neq \overline{X}; \\[2mm] 0, & \text{otherwise.} \end{cases} \quad (2)$$

Thus, $CVS(X)$ is the normalized standard deviation of X.

Proposition 2. The properties of the Sharma's coefficient of variation are as follows [14] :
1) $CVS(X)$ is not influenced by the mean. It measures the dispersion of two series with symmetrical values by the same value.
   Let's take the previous example again: let $X = \{1 ; 2 ; 3 ; 10\}$ and $Y = \{1 ; 8 ; 9 ; 10\}$. We have : $\overline{X} = 4$, $\overline{Y} = 7$, $CVS(X) = 0.83$ and $CVS(Y) = 0.83$.
2) $CVS(X)$ is normalized: For any real random variable X, $0 \leq CVS(X) \leq 1$.
3) $CVS(X)$ is stable by a linear transformation : $\forall a \in R, CVS(X + a) = CVS(X)$.
4) $CVS(X)$ is stable by a multiplication with a scalar : $\forall k \in R, CVS(k.X) = CVS(X)$.

Corollary : $\forall (k, a) \in R^2, CVS(k.X + a) = CVS(X)$.

Indeed, after evaluating the 12 existing measures of dispersion in the literature with the 8 existing assessment criteria, Tovohery et al. have shown that the Sharma's coefficient of variation is the best [16].

2.3 Distribution law of the product of two variables following a Gumbel distribution

In 2005, Nadarajah gave in [9] the distribution law of random variables $S = X + Y$, $D = X - Y$, $P = XY$ and $R = X/Y$, such that X and Y are real random variables following the bivariate Gumbel distribution. In this article, the author has even given the quantile tables of these four random variables. We recall below the necessary tricks for the distribution law of $P = XY$. If X and Y are random variables whose joint density function is defined by:
$$\forall \, x > 0, y > 0 \text{ and } 0 \leq \theta < 1, f(x, y) = [(1 + \theta x)(1 + \theta y) - \theta] \times \exp[-(x + y + \theta xy)] \quad (3)$$

and the joint distribution function is defined by :
$$\forall \, x > 0, y > 0 \text{ and } 0 \leq \theta < 1, F(x, y) = \exp[-(x + y + \theta xy)] \quad (4)$$
then, (X, Y) is said to be a random variable following the bivariate Gumbel distribution.

Theorem 1. If the joint distribution of X and Y is defined according to (3), then the density function of $P = XY$ is :
$$f_P(p) = 2.\left[\theta\sqrt{p}\, K_1\left(2\sqrt{p}\right) + (1 - \theta + \theta^2 p)K_0\left(2\sqrt{p}\right) + \theta\sqrt{p}K_{-1}\left(2\sqrt{p}\right)\right].\exp(-\theta p) \quad (5)$$
where $0 \leq p < \infty$ and
$$K_\nu(x) = \frac{\sqrt{\pi}x^\nu}{2^\nu \Gamma(\nu + 1/2)} \int_0^{+\infty} \exp(-xt)(t^2 - 1)^{\nu - 1/2} dt.$$

The density function (5) can also be written as :
$$f_P(p) = \exp(-\theta.p)\,[\,\theta.I_0 + (1 - \theta + \theta^2.p)I_1 + \theta.p.I_2\,] \quad (6)$$
where $I_k = \int_0^{+\infty} \frac{1}{x^k} \exp\left(-x - \frac{p}{x}\right) dx.$
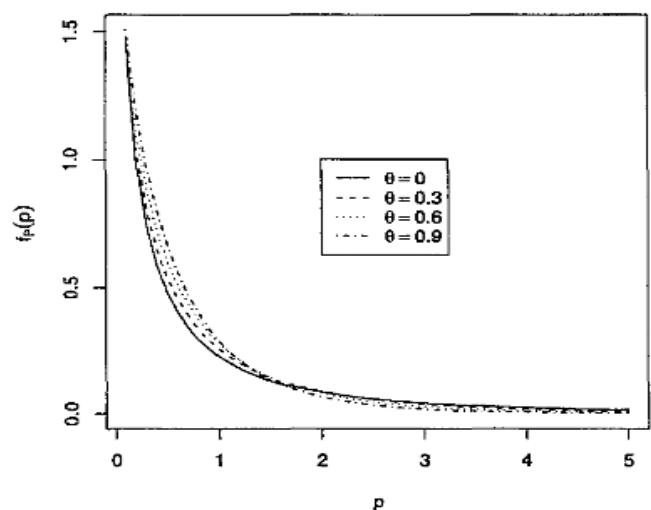

Figure 1: Curve of function (3) for $\theta = 0$, $\theta = 0.3$, $\theta = 0.6$ and $\theta = 0.9$. (Source: [9, p. 503])

Theorem 2: If the joint distribution of X and Y is defined according to (3), then the moment of order n of the random variable $P = XY$ is defined by :

$$E(P^n) = \frac{(n!)^2}{\theta^n} \phi\left(n, 1, \frac{1}{\theta}\right), \text{for } n = 1,2,3 \qquad \dots (7)$$

where $\phi(a, b, x) = \frac{\Gamma(b-1)}{\Gamma(a)} x^{1-b} G(a - b + 1; 2 - b; x) +$

$\frac{\Gamma(1-b)}{\Gamma((a-b+1))} G(a, b, x)$ and $G(\alpha, \beta, x) = \sum_{k=0}^{+\infty} \frac{(\alpha)_k}{(\beta)_k} \frac{x^k}{k!}$ and
$(n)_k = n(n + 1) \dots (n + k - 1)$.

2.4 Distribution law of the product of two random variables following a Student distribution

In 1980, Wallgren already studied in [10] the distribution law of a product of two random variables following Student distribution by introducing the random variable $W = \frac{XY}{S^2} = \left(\frac{X}{S}\right)\left(\frac{Y}{S}\right)$, such that X and Y respectively follow the normal distribution $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$. We note by ρ the linear correlation coefficient of X and Y, and by $f(w, \rho)$ the density function of W. Wallgren mentioned the following properties [10, p. 997]:
a) $f(w, \rho) = f(-w, -\rho)$;
b) In the case where $\mu_1 = \mu_2 = 0$, if $n \to +\infty$ and $\rho \to 1^-$ then the distribution of W tends to the Fisher distribution with 1 and n degrees of freedom, which is denoted by $F_{(1,n)}$.

3. Results

The previous findings allow us to have the two results mentioned in the two following proposals.

Proposition 3. The exact distribution law of $\frac{n}{CVS^2(X)(n+1)}$ is defined by the density function (5), and its moments of order n are given by equation (7).

Proof: Let $E_n = x_1, \dots, x_n$ a sample of n values of a real random variable X following the normal distribution of mean μ and standard deviation σ, noted $N(\mu, \sigma)$. Then, the random variables $Y_1 = X - \overline{X}$ and $Y_2 = \overline{X} - X$ follow the normal distribution $N\left(0, \sigma\sqrt{\frac{n+1}{n}}\right)$. Thus, the random variables $U_1 = \frac{Y_1\sqrt{n}}{\sigma\sqrt{n+1}} = \frac{(X-\overline{X})\sqrt{n}}{\sigma\sqrt{n+1}}$ and $U_2 = \frac{Y_2\sqrt{n}}{\sigma\sqrt{n+1}} = \frac{(\overline{X}-X)\sqrt{n}}{\sigma\sqrt{n+1}}$ follow the normal distribution $N(0,1)$.

Since the random variable $\frac{nS^2}{\sigma^2}$ follows the Chi-square distribution law with n degrees of freedom, then the random variables $Z_i = \frac{U_i}{(\sqrt{n} s/\sigma)/\sqrt{n}}, i = 1,2$, follow the Student distribution with n degrees of freedom. After simplification, we have : $Z_1 = \frac{(X-\overline{X})\sqrt{n}}{S\sqrt{n+1}}$ and $Z_2 = \frac{(\overline{X}-X)\sqrt{n}}{S\sqrt{n+1}}$. We can notice that $W_1 = Max(Z_1) = \frac{Max(X)-\overline{X}}{S}\sqrt{\frac{n}{n+1}}$ and $Min(Z_2) = \frac{\overline{X}-Min(X)}{S}\sqrt{\frac{n}{n+1}}$. Thus, we have :

$$W_1 \cdot W_2 = \frac{n}{CVS^2(X)(n + 1)} \qquad (8)$$

It is known that "*Student distribution belong in majority to the domain of attraction of Fréchet distribution*" [15, p.11]. "*But it can be shown that if a distribution F belongs to the domain of attraction of Weibull distribution or Fréchet distribution, then it also belongs to Gumbel's domain of attraction. This is one of the reasons why the Gumbel distribution is frequently used in applications. It is also for this reason that the Gumbel distribution is used as a reference for defining the null hypothesis in diagnostic tests of asymptotic distribution laws of maxima*" [15, p.15]. Thus, we can say without doubt that the random variables W_1 and W_2 all follow Gumbel distribution for n sufficiently large. Then, the random variable $W_1 \cdot W_2 = \frac{n}{CVS^2(X)(n+1)}$ is a product of two random variables following Gumbel distribution. So, we have the proposition 3.

Proposition 4. The random variable $\frac{n}{CVS^2(X)(n+1)}$ approximately follows Fisher distribution $F_{(1,n)}$ for n sufficiently large.

Proof: Indeed, the most widely used approximation of the Gumbel distribution is the normal distribution [17] and [18]. Moreover, if the two random variables X and Y follow the same distribution $N(0, \sigma)$, where σ is very small, then the distribution law of $Z = Max(X, Y)$ can be approximated by the normal distribution $N(0, \sigma)$ [17].

As X follows the normal distribution $N(\mu, \sigma)$, then, if we suppose that $Max(X)$ and $Min(X)$ follow approximately the normal distribution $N(\mu, \sigma)$, then $V_1 = Max(X - \overline{X}) = Max(X) - \overline{X}$ and $V_2 = Min(\overline{X} - X) = \overline{X} - Min(X)$ follow the normal distribution $N(0, \sigma\sqrt{(n + 1)/n})$. If we pose $D = S(X) \cdot \sqrt{(n + 1)/n}$, then we have a random variable identical to the one studied by Wallgren in [10] : $P = \left(\frac{V_1}{D}\right)\left(\frac{V_2}{D}\right) = W_1 \cdot W_2$.

However, we have : $Z_1 = \frac{(X-\overline{X})\sqrt{n}}{S\sqrt{n+1}}$ and $Z_2 = \frac{(\overline{X}-X)\sqrt{n}}{S\sqrt{n+1}}$. Moreover, we have : $W_1 = Max(Z_1)$ and $W_2 = Min(Z_2)$. But we have: for any real random variable U, $Min(U) = -Max(-U)$. We have : $Z_1 = -Z_2$. Then we have : $W_2 = Min(Z_2) = -Max(-Z_2) = -Max(Z_1) = -W_1$. Therefore, the linear correlation coefficient of W_1 and W_2 tends to -1^+. Then, the linear correlation coefficient of W_1 and $-W_2$ tends to 1^-. Thus, for n sufficiently large, the distribution law of $-J = -W_1 \cdot W_2$ can be approximated by Fisher distribution $F_{(1, n)}$. On the other hand, according to the first property of the distribution function of the product of two random variables following Student distribution quoted above $(f(w, \rho) = f(-w, -\rho))$, we deduce that $f(-J, 1^-) = f(J, -1^+)$. This equality allows us to confirm that we can approximate the distribution of $J = W_1 \cdot W_2 = \frac{n}{CVS^2(X)(n+1)}$ by Fisher distribution $F_{(1, n)}$ for n sufficiently large. Thus, we have the proposition 4.

4. Discussion

In this section, we will discuss, using Kolmogorov's goodness-of-fit test, the similarity of the exact distribution law and the approximate distribution law of the variable $W_1 \cdot W_2 = \frac{n}{CVS^2(X)(n+1)}$ according to the values of the parameter θ of the density function (5).

The Kolmogorov's goodness-of-fit test consists in comparing the maximum of the differences between the distribution function F_P of the product of two random variables following Gumbel distribution and the distribution function $F_{1,n}$ of the Fisher distribution with 1 and n degrees of freedom with a critical value already tabulated. Thus, we consider the maximum deviation:

$$D_{max} = \max\left(\left|F_p(x_i) - F_{1,n}(x_i)\right|\right), i = \overline{1, m} \qquad (9)$$

If $D_{max} \leq$ critical_value(α), then we accept at the confidence level $(1 - \alpha)100\%$ that F_p is approximately equal to $F_{1,n}$.

The following Table 1 gives the values of D_{max}, values obtained according to the sample size n and the parameter θ of the product distribution law of two random variables following Gumbel distribution.

Table 1: Value of D_{max} obtained according to the size n of the sample and the parameter θ of the product of two variables according to Gumbel distribution

$\frac{\theta}{n}$	0.01	0.3	0.6	0.9	Critical values $\alpha = 0.05$
15	0.08104	0.1116	0.15315	0.19198	0.33760
30	0.08114	0.11182	0.15315	0.19203	0.24170
50	0.08127	0.11188	0.15316	**0.19226**	0.18841
80	0.08134	0.11188	**0.15316**	**0.19237**	0.14960
100	0.08134	0.11188	**0.15315**	**0.19237**	0.13403

According to Table 1, we can conclude that Fisher distribution $F_{1;n}$ can approximate the distribution law of the product of the two random variables following Gumbel distribution for $\theta \leq 0.3$. Moreover, Figures 2, 3 and 4 further illustrate this statement.
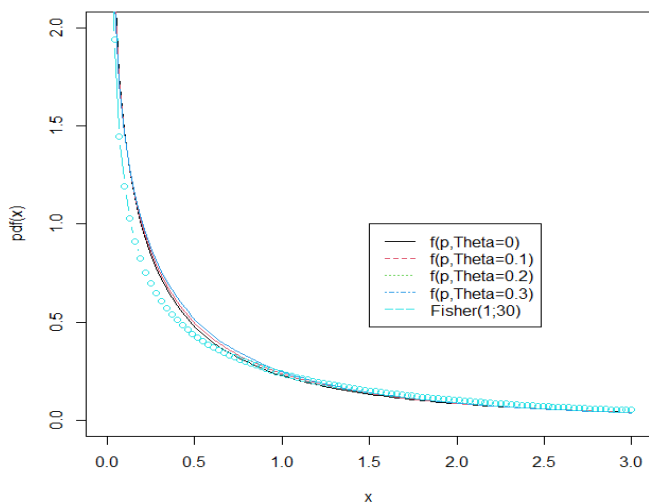

Figure 2: Exact density function of $W_1.W_2$ for $\theta = \{0 \; ; \; 0.1 \; ; 0.2 \; ; 0.3\}$ and density function of the Fisher distribution with 1 and 30 degrees of freedom.
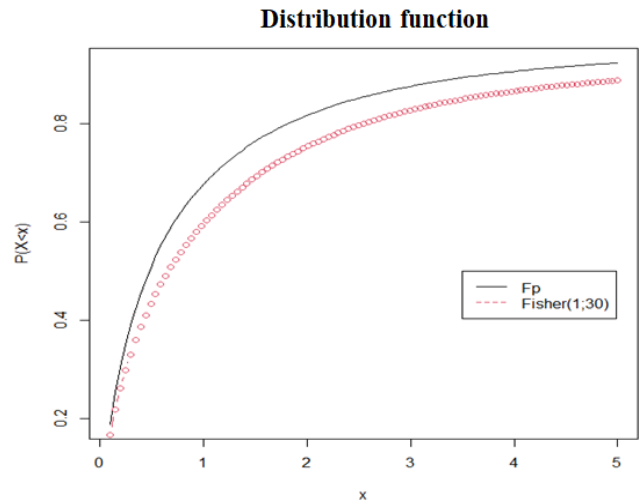

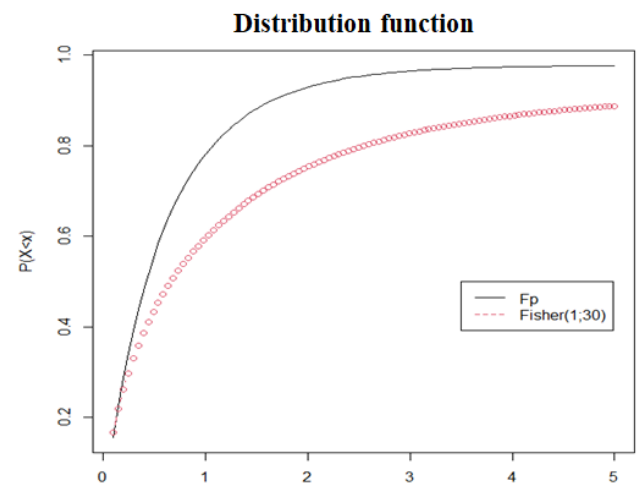**Figure 3:** Distribution function F_P obtained from equation (4) and $F_{1;30}$ for $\theta = 0.01$

Figure 4: Distribution function F_P obtained from equation(4) and $F_{1;30}$ for $\theta = 0.9$

5. Conclusion

To conclude this analysis, we are led to conclude that the exact distribution law of $J = \frac{n}{CVS^2(X)(n+1)}$, where CVS is the Sharma's coefficient of variation, is the distribution law of the product of two random variables following the Gumbel distribution. The quantiles of this distribution law are already tabulated by Nadarajah in [9].

Due to the complexity of this exact distribution law, we tried to find the approximate distribution law of J. Thus, we have seen that the approximate distribution law of J is the Fisher distribution $F_{1,n}$ for n sufficiently large and for $\theta \leq 0.3$.

The logical continuation of our work is to find a test of equality of two or more coefficients of variation of Sharma, in order to find an application of this measure to the comparison of the dispersions of two different populations, to the test of heteroscedasticity as what we proposed in **[13]** with the Pearson CV, and to the test of change of structure of a linear regression model. Then, it is also important to find the application of this measure in the field of quality control: design of a control chart of type $(\bar{X}, CVS)$.

References

[1] Centre d'expertise en analyse environnementale du Québec. Guide d'estimation de l'incertitude des mesures pour les analyses chimiques. DR-12-INC, Quebec, Ministère du Développement durable, de l'Environnement et de la Lutte contre les changements climatiques. 2016.

[2] S. Villette, E. Piron and D. Miclet. Hybrid centrifugal spreading model to study the fertilizer spatial distribution and its assessment using the transverse coefficient of variation.Science Direct, Computers and Electronics in Agriculture. 2017.

[3] A. Achouri. Cartes de contrôle pour le coefficient de variation. Thèse de Doctorat, l'Université de Nantes Angers Le Mans. 2014.

[4] A.J.X. Lim, M. B.C. Khoo, W.L. Teoh and A. Haq. Run sum chart for monitoring multivariate coefficient of variation. Science Direct, Computers & Industrial Engineering, 2017.

[5] D.J. Charier, D. Zantour, V. Pichot, F. Chouchou, J. M. Barthelemy, F. Roche and S. B.Molliex. Assessing Pain Using the Variation Coefficient of Pupillary Diameter. Science Direct, The Journal of Pain, 2017.

[6] J. J. J. Babu and G. F. Sudha. Adaptive speckle reduction in ultrasound images using fuzzy logic on Coefficient of Variation. Science Direct, Biomedical Signal Processing and Control, 2015.

[7] J. C. Cox and V. Sadiraj On the coefficient of variation as a criterion for decision under risk. Science Direct, Journal of Mathematical Psychology, 2010.

[8] R. Calif, T. Soubdhan. On the use of the coefficient of variation to measure spatial and temporal correlation of global solar radiation. Science Direct, On the use of the coefficient of variation to measure spatial and temporal correlation of global solar radiation, 2015.

[9] S. Nadarajah. Sums, products, and ratios for the bivariate Gumbel distribution. Mathematical and computer modelling - Science direct. 2005. Doi: 10.1016/j.mcm.2005.02.003

[10] C. M. Wallgren. The distribution of the product of two correlated t variates. Journal of the american statistical association Vol. 75, no. 372, pp. 996 - 1000. 1980. http://www.jstor.org/stable/2287194 .

[11] C.E.Brown. Coefficient of variation. Applied statistics in geohydrology and related sciences. Springer. P. 155 – 157. 1998. DOI: 10.1007/978-3-642-80328-4_13.

[12] W. Reh and B. Scheffier.Significance Tests and Confidence Intervals for Coefficients of Variation. The statistical software newsletter, pp. 449 - 452. 1996. https://doi.org/10.1016/0167-9473(96)83707-8

[13] J. M. Tovohery, A. Totohasina and D. R. Feno. Application of equality test of coefficients of variation to the heteroskedasticity test. American Journal of Computational Mathematics, 10, pp. 73 - 89, 2020. https://doi.org/10.4236/ajcm.2020.101005.

[14] R. Sharma, R.G. Shandil and G. Kapoor. A note on Karl Pearson's coefficient of dispersion. Himachal Pradesh University Journal, 2011. https://pdfs.semanticscholar. org/229c/6e64c38e0b39773b03b82f458d8779e1f638.p df or https://studylib.net/ doc/8762538/a-note-on-karl-pearson-s-coefficient-of-dispersionkarl-pearson-s-coefficient-of-dispersion

[15] S. Avouyi – Dovi et D. Guegan. Valeursextrêmes et sériestemporelles : application à la finance. Document de recherche EPEE. Université d'Evry. 2001.

[16] J. M. Tovohery, A. Totohasina and D. R. Feno. Une technique d'extraction des règles d'association quantitatives basée sur le coefficient de variation. Analyse Statistique Implicative, ASI11, Belfort, France, page 79 – 95. 2021. ISBN : 978-2-9562045-5-8

[17] S. Nadarajah and S. Kotz. Exact distribution of the Max/Min of two gaussian random variables. IEEE transactions on very large scale integration (VLSI) systems, vol. 16, No. 2. February, p. 210 - 212. 2008. Doi : 10.1109/TVLSI.2007.912191.

[18] R. D. Gupta and R. C. Gupta. Analyzing skewed data by power normal model. 2007 Doi: 10.1007/s11749-006-0030-x