Brazilian Electronic Journal of Mathematics, v.2 - n.4, jul/dez, 2021. ISSN: 2675-1313

Note on the coefficient of variation properties [1]

Nota sobre as propriedades do coeficiente de variação

Carla Santos
CMA-Center of Mathematics and its Applications-FCT-New
University of Lisbon and Polytechnic Institute of Beja, Portugal
carla.santos@ipbeja.pt
ORCID: 0000-0002-0077-1249

Cristina Dias
CMA-Center of Mathematics and its Applications-FCT-New
University of Lisbon and Polytechnic Institute of Portalegre,
Portugal
cpsd@ipportalegre.pt
ORCID: 0000-0001-6350-5610

Abstract. The extent to which the values within the dataset differ from one another and from the mean value itself is revealed through measures of variability. One common measure of variability is the coefficient of variation, which expresses the standard deviation as a proportion of the mean and does not depend on the unit scales. Exploiting its potential as a unitless measure of variability, the coefficient of variation has been used as risk sensitivity measure, to represent the reliability of trials, and for other purposes. In different frameworks, the coefficient of variation can be used when a single sample is considered, but also when comparing distributions. In general, the assumptions for the use of the coefficient of variation are related to the nature of data, nevertheless, some properties of this coefficient lead to limitations on its suitability in certain situations. In this work we present a comparative review of the coefficient of variation properties and those of Dodd's corrected coefficient of variation.

Keywords. Bounds. Dodd's corrected coefficient of variation. Dispersion.

Resumo. Para determinar o grau com que os valores de um conjunto de dados diferem uns dos outros recorre-se a medidas de variabilidade. Uma dessas medidas é o coeficiente de variação, que expressa o desvio-padrão como uma proporção da média, não dependendo da ordem de grandeza da variável. Tirando partido do seu potencial como medida de variabilidade adimensional, o coeficiente de variação tem sido usado como medida de sensibilidade

[1]This work is funded by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of the project UIDB/00297/2020 (Center for Mathematics and Applications).

ao risco, para representar a variabilidade de ensaios, e para outros fins. Nos diferentes enquadramentos, o coeficiente de variação pode ser usado quando é considerada uma única amostra, mas também para a comparação de distribuições. Em geral, os pressupostos para o uso do coeficiente de variação assentam no tipo de dados, não obstante, algumas propriedades deste coeficiente limitam a sua adequação a certas situações. Neste trabalho apresentamos uma revisão comparativa das propriedades do coeficiente de variação com as propriedades de uma das suas alternativas, o coeficiente de variação corrigido de Dodd.

Palavras-chave. Coeficiente de variação corrigido de Dodd. Dispersão. Limites.

Mathematics Subject Classification (MSC): 62F10.

1 Introduction

Statistics plays a vital role in scientific research and development, providing methods to visualize and summarize data and to produce analysis and predictions resorting to numerical evidence. In this framework, measures of central tendency and measures of variability (dispersion) are features of great importance. To obtain a single value that describes a data set, measures of central tendency are used. Although, under different conditions, some measures of central tendency become more appropriate for use than others, the mean is the most popular measure of central tendency, due to the important properties it has. Regarding the usefulness of the measures of central tendency, it is important to highlight that the typical values of a frequency distribution of a variable, represented in measures of central tendency, do not provide information on important aspects of the distribution, and that it is not known whether these measures are, in fact, representative of the values of the data set, therefore, it is essential to resort to other types of measures. The extent to which the values within the dataset differ from one another and from the mean value itself is revealed through measures of variability. To know the magnitude of the differences between the values of a data set in relation to its mean, the variance, also called mean square deviation, is the most obvious measure, since it provides the average of the squares of that differences. Considering that the variance has the disadvantage of being expressed in square units of the variable under study, its square root, the standard deviation, stands out for its ease of interpretation. However, when comparing the variability of several datasets, whose values are expressed in different measurement units, the standard deviation is not appropriate, being convenient to use a relative variability measure. One common variability measure is the coefficient of variation, which expresses the standard deviation as

a proportion of the mean and does not depend on the unit scales. The coefficient of variation has important applications in research in agriculture, industry, medical and social sciences, education, and many other fields, and has been applied for several purposes. For example, [10], [15], [3] and [20] used the coefficient of variation as measure of risk sensitivity, [17] employed it for assessing variability in agricultural experiments, [6] applied it to represent the reliability of trials, and [4], [2], [13] and [14] considered it in the assessment of the accuracy of experiments.

Despite the usefulness of the coefficient of variation, it cannot be applied in a generalized way, and it has appropriate meaning only if certain requirements are met. The coefficient of variation is only suitable for variables that are measured on ratio scales with absolute zero, [11]. If all the observations are non-negative, a null mean would occur only in the trivial case in which all the observation are zero, and then the coefficient of variation is undefined [9]. When the variable has positive and negative values, and the mean is close to zero, the coefficient of variation can be misleading [18]. It cannot be used to compare extremely distinct magnitudes [5].

In addition to the applicability conditions of the coefficient of variation, related to the nature of data under study, questions have arisen regarding the coefficient of variation properties, which triggered the appearance of alternative measures to overcome limitations.

In this work we develop a comparative approach between the properties of the common coefficient of variation and the properties of Dodd's corrected coefficient of variation, introduced by Stuart Dodd, in 1952, (see [9]). Due its characteristics, Dodd's corrected coefficient of variation is often considered a good alternative to the common coefficient of variation, however it also has its limitations and restrictions on their suitability in certain contexts, since some situations require variation measures whose upper bound depends on sample size (e.g.[19]) and others where this dependency is not desirable (e.g.[12]).

2 Coefficient of variation definition and bounds

The coefficient of variation (C_V) is a relative variability measure, expressing the dispersion of data values around the mean.

Let us consider a positive random variable X, with mean μ , $\mu \neq 0$, and standard deviation σ . The population coefficient of variation of X is,

$$\gamma_V = \frac{\sigma}{\mu} \tag{1}$$

In a single sample, with observations $x_1, x_2, \cdots, x_n$, with $x_j \geq 0$, $j = 1, \cdots, n$, the

coefficient of variation, C_V, is

$$C_V = \frac{s}{\overline{x}} \tag{2}$$

where

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \tag{3}$$

$\overline{x} \neq 0$, and

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{n} (x_j - \overline{x})^2} \tag{4}$$

are the mean and the standard deviation of the observations, respectively.

The standard deviation given by Equation (4) can be made unbiased through Bessel's correction, this is, using $n-1$ instead of n in the denominator,

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n} (x_j - \overline{x})^2}$$

When n is very large, Bessel's correction can be neglected, since it becomes approximately 1.

From this point on, we will only consider the case where n is very large, since similar results would be obtained when this is not the case.

Regarding to the limit values of the coefficient of variation, the lower and upper bounds are reached when all values of the variable are equal (minimum) and all values except one are null (maximum) (see, for example, [7]).

Proposition 1. *When $x_j = a$, $j = 1, \cdots, n$, $C_V = 0$.*

Proof. A sample in which $x_j = 1$, $j = 1, \cdots, n$, has mean value

$$\overline{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{na}{n} = a$$

and the standard deviation,

$$s = \sqrt{\frac{\sum_{j=1}^{n} (x_j - a)^2}{n}} = \sqrt{\frac{\sum_{j=1}^{n} (a - a)^2}{n}} = 0$$

so, its coefficient of variation is

$$C_V = \frac{s}{\overline{x}} = 0$$

■

Proposition 2. *When $x_j = 0$, $j = 1, \cdots, n$, $j \neq i$, $x_i = b$, with $b \neq 0$, and n is very large, $C_V = \sqrt{n-1}$.*

Proof. A sample in which $x_j = 0$, $j = 1, \cdots, n$, $j \neq i$ and $x_i = b$, with $b \neq 0$, has mean value

$$\overline{x} = \frac{\sum_{j=1}^{n} x_j}{n} = \frac{(n-1) \times 0 + b}{n} = \frac{b}{n}$$

and the standard deviation

$$s = \sqrt{\frac{\sum_{j=1}^{n} \left(x_j - \frac{b}{n}\right)^2}{n}} = \sqrt{\frac{(n-1)\left(0 - \frac{b}{n}\right)^2 + \left(b - \frac{b}{n}\right)^2}{n}} =$$

$$= \sqrt{\frac{(n-1)\left(-\frac{b}{n}\right)^2 + \left(b - \frac{b}{n}\right)^2}{n}} = \sqrt{\frac{(n-1)\frac{b^2}{n^2} + b^2 - 2\frac{b^2}{n} + \frac{b^2}{n^2}}{n}} =$$

$$= \sqrt{\frac{n\frac{b^2}{n^2} + b^2 - 2\frac{b^2}{n}}{n}} = \sqrt{b^2\left(\frac{1}{n} - \frac{1}{n^2}\right)} = b\frac{\sqrt{n-1}}{n}$$

so, its coefficient of variation is

$$C_V = \frac{s}{\overline{x}} = \frac{b\frac{\sqrt{n-1}}{n}}{\frac{b}{n}} = \sqrt{n-1}.$$

■

Remark: For small samples, considering Bessel's correction for the standard deviation, the upper bound of the coefficient of variation is $\sqrt{n}$.

3 Coefficient of variation properties

Proposition 3. *C_V is scale invariant.*

Proof. Given two random variables X and $Y = kX$, with k a positive constant, we have $\overline{y} = k\overline{x}$ and $s_y = ks_x$, then

$$C_V(Y) = \frac{ks_x}{k\overline{x}} = \frac{s_x}{\overline{x}} = C_V(X)$$

so, when all data points increased (decreased) by the same proportion, their relative differences remain the same, so the value of C_V remains the same too.

■

Proposition 4. *C_V is sensitive to location.*

Proof. Given two random variables X and $Z = X + c$, with c a non-null constant, we have $\overline{z} = \overline{x} + c$ and $s_z = s_x$, then

$$C_V(Z) = \frac{s_x}{\overline{x} + c} \neq C_V(X)$$

so, the coefficient of variation C_V is not invariant to transformations that involve adding (or subtracting) a constant. ■

Proposition 5. *C_V is sample-size invariant.*

Proof. Let us consider two samples A and B, whose difference is only in size. The equality between their coefficients of variation comes immediately from the definition of C_V, since $\overline{x}_A = \overline{x}_B$ and $s_A = s_B$. ■

Addressing the acceptability of the measures of relative variation, in social and economic phenomena, [1] imposes, in addition to the scale invariance, the principle of transfers. This principle states that the value of a measure of inequality increases when resources are transferred from a poor person to a richer person.

Proposition 6. *C_V is sensitive to transfers.*

Proof. Let us consider x_j and x_k, two values of the variable X such that $x_j \leq x_k, j \neq k$, and the transfer of α units ($\alpha > 0$) from x_j to x_k, with no change in all other x_i, $i = 1, \cdots, n, i \neq j$ and $i \neq k$.

Before the transfer, the mean is

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_j + \cdots + x_k + \cdots + x_n}{n}$$

the standard deviation is

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_j - \overline{x})^2 + \cdots + (x_k - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n}}$$

and the coefficient of variation is

$$C_V = \frac{s}{\overline{x}}$$

After the transfer, the mean is

$$\overline{x}^* = \frac{x_1 + x_2 + \cdots + x_j - \alpha + \cdots + x_k + \alpha + \cdots + x_n}{n}$$

the standard deviation is

$$s^* = \sqrt{\frac{(x_1-\overline{x})^2+(x_2-\overline{x})^2+\cdots+(x_j-\alpha-\overline{x})^2+\cdots+(x_k+\alpha-\overline{x})^2+\cdots+(x_n-\overline{x})^2}{n}}$$

and the coefficient of variation is

$$C_V^* = \frac{s^*}{\overline{x}^*}.$$

It is easy to see that $\overline{x} = \overline{x}^*$ and $n^2({s^*}^2 - s^2) = 2\alpha(x_k - x_j) + 2\alpha^2$. So, C_V is sensitive to transfers, since

$${C_V^*}^2 - {C_V}^2 = \beta[\alpha(x_k - x_j) + \alpha^2] \tag{5}$$

with $\beta = \frac{2}{n^2\overline{x}^2}$, and ${C_V^*}^2 = {C_V}^2$ only if $\alpha = 0$. ■

Proposition 7. *C_V verify the principle of transfers.*

Proof. From Equation (5) it is easy to see that C_V increases when $x_j < x_k$, $j \neq k$, and there is a transfer of α units ($\alpha > 0$) from x_j to x_k. ■

4 Dodd's corrected coefficient of variation

Expressing the coefficient of variation as a percentage,

$$C_V = \frac{s}{\overline{x}} \times 100 \tag{6}$$

is common, and sometimes referred to as percentage relative standard deviation (e.g. [8]). Although percentage representation is widely used, some literature claims it can lead to potential misinterpretations since C_V value can exceed 1, and expressing it as a percentage we can get percentage values greater than one hundred. These misinterpretations can be overcome by setting C_V upper bound at 1. Considering that C_V vary in their maximum possible values with n, Stuart Dodd introduced, in 1952, the corrected coefficient of variation [9], given by

$$C_{Vcorr} = \frac{C_V}{\sqrt{n-1}} \tag{7}$$

Proposition 8. *Dodd's corrected coefficient of variation, C_{Vcorr}, varies from 0 to 1.*

Proof. Since the limit values for C_V are 0 and $\sqrt{n-1}$, for C_{Vcorr} we have

$$0 \leq C_{Vcorr} \leq 1 \tag{8}$$

Using C_{Vcorr} ensures that the upper limit of the dispersion measure does not depend on the sample size. In addition, it is easy to see that, like C_V, also C_{Vcorr} is scale invariant, sensitive to location, and sensitive to transfers. ■

Proposition 9. *Dodd's corrected coefficient of variation, C_{Vcorr}, is scale invariant.*

Proof. Given two random variables X and $Y = kX$, with k a positive constant, we have $\overline{y} = k\overline{x}$ and $s_y = ks_x$, then

$$C_{vcorr}(Y) = \frac{C_V(Y)}{\sqrt{n-1}} = \frac{\frac{ks_x}{k\overline{x}}}{\sqrt{n-1}} = \frac{\frac{s_x}{\overline{x}}}{\sqrt{n-1}} = C_{Vcorr}(X)$$

so, when all data points increased (decreased) by the same proportion their relative differences remain the same, so the value of C_{Vcorr} remains the same too. ■

Proposition 10. *Dodd's corrected coefficient of variation, C_{Vcorr} is sensitive to location.*

Proof. Given two random variables X and $Z = X + c$, with c a non-null constant, we have $\overline{z} = \overline{x} + c$ and $s_z = s_x$, then

$$C_{Vcorr}(Z) = \frac{C_V(Z)}{\sqrt{n-1}} = \frac{\frac{s_x}{\overline{x}+c}}{\sqrt{n-1}} \neq C_{Vcorr}(X)$$

so, C_{Vcorr} is not invariant to transformations that involve adding (or subtracting) a constant. ■

Proposition 11. *Dodd's corrected coefficient of variation, C_{Vcorr} is sensitive to transfers.*

Proof. Let us consider x_j and x_k, two values of the variable X such that $x_j \leq x_k, j \neq k$, and the transfer of α units ($\alpha > 0$) from x_j to x_k, with no change in all other x_i, $i = 1, \cdots, n, i \neq j$ and $i \neq k$. Before the transfer,

$$C_{Vcorr} = \frac{\frac{s_x}{\overline{x}}}{\sqrt{n-1}}$$

After the transfer,

$$C^*_{Vcorr} = \frac{\frac{s^*}{\overline{x}^*}}{\sqrt{n-1}}.$$

Since $\overline{x} = \overline{x}^*$ and $n^2(s^{*2} - s^2) = 2\alpha(x_k - x_j) + 2\alpha^2$, C_{Vcorr} is sensitive to transfers, because

$$C^*_{Vcorr}{}^2 - C_{Vcorr}{}^2 = \frac{C^*_V{}^2 - C_V{}^2}{n-1} = \frac{\beta[\alpha(x_k - x_j) + \alpha^2]}{n-1}$$

with $\beta = \frac{2}{n^2\overline{x}^2}$, and $C^*_V{}^2 = C_V{}^2$ only if $\alpha = 0$. ■

Proposition 12. *Dodd's corrected coefficient of variation, C_{Vcorr} verify the principle of transfers.*

Proof. From Equation (5), when there is a transfer of α units ($\alpha > 0$) from x_j to x_k, with $x_j < x_k, j \neq k$, C_V increases, so C_{Vcorr} increases too. ■

All these properties refer to the case where a single sample is considered, however, as noted by [16], Dodd's corrected coefficient of variation, C_{Vcorr} is not suitable for comparative purpose, as can be seen in the next example.

Example 1. *Let us consider two samples that only differ in size:*
Sample A: $x_1, x_2, \cdots, x_n$,
Sample B: $x_1, x_1, x_2, x_2, \cdots, x_n, x_n$,
where n *is very large.*

The mean values are, respectively,

$$\overline{x}_A = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

and

$$\overline{x}_B = \frac{2x_1 + 2x_2 + \cdots + 2x_n}{2n} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \overline{x}_A.$$

And the standard deviations are

$$s_A = \sqrt{\frac{(x_1 - \overline{x}_A)^2 + (x_2 - \overline{x}_A)^2 + \cdots + (x_n - \overline{x}_A)^2}{n}},$$

and, since $\overline{x}_B = \overline{x}_A$,

$$s_B = \sqrt{\frac{(x_1 - \overline{x}_A)^2 + (x_1 - \overline{x}_A)^2 + (x_2 - \overline{x}_A)^2 + (x_2 - \overline{x}_A)^2 + \cdots + (x_n - \overline{x}_A)^2 + (x_n - \overline{x}_A)^2}{n}} =$$

$$= \sqrt{\frac{2(x_1 - \overline{x}_A)^2 + 2(x_2 - \overline{x}_A)^2 + \cdots + 2(x_n - \overline{x}_A)^2}{2n}}$$

Now, since the corrected coefficients of variation, of samples A and B, are

$$C_{VcorrA} = \frac{1}{\sqrt{n-1}} \frac{s_A}{\overline{x}_A}$$

and

$$C_{VcorrB} = \frac{1}{\sqrt{2n-1}} \frac{s_B}{\overline{x}_B},$$

with $\overline{x}_A = \overline{x}_B$ *and* $s_A = s_B$, *we have*

$$C_{VcorrA} \neq C_{VcorrB}.$$

So, considering what was described on Example 1, we can now establish that

Proposition 13. *Dodd's corrected coefficient of variation,* C_{Vcorr}, *is sample-size sensitive.*

5 Conclusion

Expressing the standard deviation as a proportion of the mean, we obtain a unitless measure of variability - the coefficient of variation. In addition to the limitations of applicability of the coefficient of variation related to the nature of data, there are situations in which the fact that the upper limit of the coefficient of variation depends on the sample size constitutes a disadvantage. Dodd's corrected coefficient of variation is an alternative to the common coefficient of variation that allows to overcome this disadvantage of dependence on sample size, since it varies from 0 to 1, and keep the same properties of scale invariant, sensitivity to location, and sensitivity to transfers. However, despite sharing these important properties of the common coefficient of variation, there are also situations in which Dodd's corrected coefficient of variation adequacy is restricted, for example, it is not effective for sample comparison due to its sample-size sensitivity.

References

[1] ALLISON, P. D. Measures of Inequality. **American Sociological Review** v. 43, p. 865-80, 1978.

[2] AMARAL, A.; MUNIZ, J.; SOUZA, M. Avaliação do coeficiente de variação como medida de precisão na experimentação com citros. **Pesquisa Agropecuária Brasileira** Brasília, v. 32, n. 12, p. 1221-1225,1997.

[3] COX, J.C.; SADIRAJ, V. On the Coefficient of Variation as a Measure of Risk Sensitivity. **Behavioral & Experimental Economics (Editor's Choice) eJournal,** 2011.

[4] GOMEZ, K. A., GOMEZ, A. A. **Statistical Procedures for Agricultural Research.** 2nd ed. New York: John Wiley and Sons, Inc., 1984.

[5] GRANER, E.A. **Estatística.** São Paulo: Melhoramentos, 1996.

[6] HOPKINS, W. G. Measures of reliability in sports medicine and science. **Sports Med.,** v. 30, p. 1-15, 2000.

[7] KATSNELSON, J.; KOTZ, S. On the upper limits of some measures of variability. **Archiv für Meteorologie, Geophysik und Bioklimatologie,** Series B, v. 8, p. 103–107, 1957.

[8] MAGNUSSON, B.; NÄYKKI, T.; HOVIND, H.; KRYSELL, M.; SAHLIN, E. **Handbook for calculation of measurement uncertainty in environmental laboratories,** Nordtest Report TR 537 (ed. 4), 2017.

[9] MARTIN, J.; GRAY, L. Measurement of Relative Variation: Sociological Examples. **American Sociological Review,** v. 36, p. 496-502, 1971.

[10] MILLER, E. G.; KARSON, M. J. Testing the equality of two coefficient of variation. **American Statistical Association: Proceedings of the Business and Economics Section**, Part v. 1, p. 278-283, 1977.

[11] MUELLER, J. H. **Statistical reasoning in sociology.** Houghton Mifflin ed., 1977.

[12] RAY, J. ; SINGER, J. Measuring the Concentration of Power in the International System. **Sociological Methods & Research,** 1(4): 403-437, 1973.

[13] REED, G. F., LYNN, F.; MEADE, B. D. Use of coefficient of variation in assessing variability of quantitative assays. **Clinical & Diagnostic Laboratory Immunology,** v. 9, p. 1235-1239, 2002.

[14] ROMANO, F. L.; AMBROSANO, G.; MAGNANI, M. B.; NOUER, D. Analysis of the coefficient of variation in shear and tensile bond strength tests. **Journal of Applied Oral Science,** v. 13, n. 3, p. 243-246, 2005.

[15] SHAFIR, S. Risk-sensitive foraging: The effect of relative variability. **Oikos,** v. 88, p. 663-669, 2000.

[16] SMITHSON, M. On relative dispersion: A new solution for some old problems. **Quality and Quantity,** v. 16, p. 261-271, 1982.

[17] SNEDECOR, G.W.;COCHRAN, W.G. **Statistical Methods.** 8th Edition, Iowa State University Press, Ames, 1989.

[18] SPIEGEL, M. R.; STEPHENS, L. J. **Schaum's Outline of Theory and Problems of Statistics.** 4 th ed. New York: McGraw-Hill, 2008.

[19] THEIL, H. **Economics and Information Theory**Studies in Mathematical and Managerial Economics, v. 7, Amsterdam: North-Holland, 1967.

[20] WEBER, E. U., SHAFIR, S., BLAIS, A. Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. **Psychological Review,** v. 111, p. 430-445, 2004.

Submetido em 4 nov. 2020
Aceito em 19 fev. 2021