7th International Conference on
Combinatorics, Cryptography, Computer Science and Computing
November. 16-17, 2022

Improved Confidence Interval and Hypothesis Testing for the Ratio of the Coefficients of Variation of Two Uncorrelated Populations

Zeynab Avazzadeh, Abbas Bahrampour, Mohammad Reza Mahmoudi

Department of Biostatistics and Epidemiology, Kerman University of Medical Sciences, Kerman, Iran

zeynab.evazzadeh@gmail.com, a_bahrampour@kmu.ac.ir, mahmoudi.m.r@fasau.ac.ir

ABSTRACT

One of the most accessible and useful statistical tools for comparing independent populations in different research areas is the coefficient of variation (CV). In this study, at first, the asymptotic distribution of the ratio of CVs of two uncorrelated populations is investigated. Then, the outputs are used to create a confidence interval and to establish a test of hypothesis about the CVs' ratio of the populations. The proposed approach is compared with an alternative method, showing its superiority and effectiveness.

KEYWORDS: Ratio of CVs; Test of Hypothesis; Symmetric Distributions; Asymmetric Distributions.

1 INTRODUCTION

According to the literature, three fundamental measures are used to explain a data set (random variable). These include central, shape and dispersion tendencies. By obtaining the value of the central tendency, we can know how a random variable is gathered around a central value. The mean, median and mode are the most used criteria to express the central tendency's measures. Criteria such as range, variance and standard deviation can be used to measure the dispersion of a random variable. In some literature, this is called the random variable distribution scale. Another need of statisticians is to know how a random variable is distributed, or to know its pattern shape, which can be addressed by the use of statistical measures such as kurtosis or skewness.

The coefficient of variation (CV) is obtained by dividing the population standard deviation by the population mean, $CV = \sigma / \mu$, which is an applicable and suitable statistic for evaluating relative variability. The CV is a free parameter that is used in many areas, such as agronomy, biology, engineering, finance, medicine, and many others, as an indicator of reliability or variability in some papers [1-3]. In many cases, relating standard deviation to the level of measurement is of great importance to researchers. For this reason, the CV is widely used to measure dispersion. When studying several independent populations, how their CVs are compared is essential. This becomes even more important when populations have skewed distributions. In practical matters, statisticians may be interested in comparing two independent populations' CVs to understand better the data structure.

2 MATERIALS AND METHODS

2.1 Asymptotic Outcomes

Let us consider two uncorrelated variables X and Y, with non-zero means of μ_X and μ_Y. Their finite i^{th} central moments are, respectively:

$$\mu_{iX} = E(X - \mu_X)^i, \quad \mu_{iY} = E(Y - \mu_Y)^i, \quad i \in \{2,3,4\}. \tag{1}$$

Furthermore, suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are samples that are identically and independently distributed from X and Y, respectively. According to what was mentioned in Section 1, given that CV_X and CV_Y are the CVs of X and Y, respectively, the parameter

$$\gamma = \frac{CV_Y}{CV_X} \tag{2}$$

is compelling for statistical inference.
Suppose that

$$m_{iX} = \frac{1}{m}\sum_{k=1}^{m}(x_k - \bar{x})^i, \qquad m_{iY} = \frac{1}{n}\sum_{k=1}^{n}(y_k - \bar{y})^i, \qquad i \in \{2,3,4\},$$

then, CV_X and CV_Y are consistently estimated [4-8] by $\widehat{CV}_X = \frac{\sqrt{m_{2X}}}{\bar{X}}$ and $\widehat{CV}_Y = \frac{\sqrt{m_{2Y}}}{\bar{Y}}$, respectively. Therefore, it follows that

$$\hat{\gamma} = \frac{\widehat{CV}_Y}{\widehat{CV}_X} \tag{3}$$

has the ability to estimate the parameter γ reasonably. For convenience, $n = m$ can be defined. At $n \neq m$, $n^* = min(m, n)$ can be used instead of m and n.

Lemma 1 If the mentioned assumptions are accepted, thus

$$\sqrt{n}\left(\widehat{CV}_X - CV_X\right) \xrightarrow{L} N(0, \delta_X^2), \quad as \quad n \to \infty,$$

where

$$\delta_X^2 = \left[\frac{\mu_{4X} - \mu_{2X}^2}{4\mu_X^2\mu_{2X}} - \frac{\mu_{3X}}{\mu_X^3} + \frac{\mu_{2X}^2}{\mu_{4X}}\right] \tag{4}$$

represents the asymptotic variance.

Proof. The proof sketch is given in [9,10].

Lemma 2 If the previously assumptions are accepted, thus

$$\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{\mathcal{L}} N(0, \lambda^2), \quad as \ n \to \infty,$$

where

$$\lambda^2 = \frac{1}{CV_X^2}(\gamma^2\delta_X^2 + \delta_Y^2), \tag{5}$$

and

$$\delta_Y^2 = \left[\frac{\mu_{4Y} - \mu_{2Y}^2}{4\mu_Y^2\mu_{2Y}} - \frac{\mu_{3Y}}{\mu_Y^3} + \frac{\mu_{2Y}^2}{\mu_{4Y}}\right]. \tag{6}$$

Proof of Lemma 2. The proof sketch is given in [10].

It can be observed that the asymptotic variance λ^2 depends on the unknown parameters CV_X, δ_X^2, δ_Y^2 and γ.

.

They used the Slutsky's theorem and proved that $\hat{\lambda} \xrightarrow{p} \lambda$, as $n \to \infty$.

In this work, we write the asymptotic variance λ^2 and estimate it using a robust estimator. The asymptotic variance λ^2 can be presented by

$$\lambda^2 = \frac{1}{\gamma^2}\left(\left(\frac{\delta_X}{CV_X}\right)^2 + \left(\frac{\delta_Y}{CV_Y}\right)^2\right).$$

where

$$\hat{\delta}_X^2 = \left[\frac{m_{4X} - m_{2X}^2}{4\bar{X}^2 m_{2X}} - \frac{m_{3X}}{\bar{X}^3} + \frac{m_{2X}^2}{m_{4X}}\right],$$

and

$$\hat{\delta}_Y^2 = \left[\frac{m_{4Y} - m_{2Y}^2}{4\bar{Y}^2 m_{2Y}} - \frac{m_{3Y}}{\bar{Y}^3} + \frac{m_{2Y}^2}{m_{4Y}}\right]$$

Therefore, we can construct the asymptotic distribution as

$$T_n = \sqrt{n}\left(\frac{\hat{\gamma} - \gamma}{\lambda}\right) = \sqrt{n}\left(\left(\frac{\delta_X}{CV_X}\right)^2 + \left(\frac{\delta_Y}{CV_Y}\right)^2\right)^{-1/2}\left(\frac{\hat{\gamma}}{\gamma} - 1\right) \xrightarrow{\mathcal{L}} N(0,1), \qquad as\ n \to \infty.$$

Theorem 1 If the previously assumptions are accepted, thus

$$T_n^* = \sqrt{n}\left(\frac{\hat{\gamma} - \gamma}{\hat{\lambda}_{opt}}\right) \xrightarrow{\mathcal{L}} N(0,1), \qquad as\ n \to \infty,$$

where

$$\hat{\lambda}_{opt}^2 = \frac{1}{\hat{\gamma}^2}\left(\left(\frac{S_X}{\widehat{CV}_X}\right)^2 + \left(\frac{S_Y}{\widehat{CV}_Y}\right)^2\right).$$

Proof of Theorem 1. Due to the Weak Law of Large Numbers, it is clear that

$\bar{X} \xrightarrow{p} \mu_X, \quad \bar{Y} \xrightarrow{p} \mu_Y, \qquad m_{iX} \xrightarrow{p} \mu_{iX}, \quad m_{iY} \xrightarrow{p} \mu_{iY}, \quad i \in \{2,3,4\},$
as $n \to \infty$.

As a result, by using the Slutsky's theorem, we know that $\hat{\lambda}_{opt} \xrightarrow{\mathrm{p}} \lambda$, as $\mathrm{n} \to \infty$. With the help of Lemma 2.2, the proof is completed.

2.1.1 Building the Confidence Interval

To construct an asymptotic confidence interval for γ we need a pivotal quantity for the parameter γ. Based on the previous results, T is a pivotal quantity for γ. Therefore, T can be used for this purpose, as follows:

$$\left(\frac{\hat{\gamma}}{1 + \frac{Z_{\alpha/2}}{\sqrt{n}}\left(\left(\frac{S_X}{\widehat{CV}_X}\right)^2 + \left(\frac{S_Y}{\widehat{CV}_Y}\right)^2\right)^{1/2}}, \frac{\hat{\gamma}}{1 - \frac{Z_{\alpha/2}}{\sqrt{n}}\left(\left(\frac{S_X}{\widehat{CV}_X}\right)^2 + \left(\frac{S_Y}{\widehat{CV}_Y}\right)^2\right)^{1/2}}\right) \tag{7}$$

2.2.1 Hypothesis Testing

In practical situations, scientists need to test the parameter γ. As particular instance, the null hypothesis $\mathrm{H}_0: \gamma = 1$ shows that the CVs are the same in both populations. To accomplish the hypothesis test $\mathrm{H}_0: \gamma = \gamma_0$, the test statistic

$$\mathrm{T}_0 = \sqrt{\mathrm{n}}\left(\left(\frac{\mathrm{S_X}}{\widehat{\mathrm{CV}}_\mathrm{X}}\right)^2 + \left(\frac{\mathrm{S_Y}}{\widehat{CV}_Y}\right)^2\right)^{-\frac{1}{2}}\left(\frac{\hat{\gamma}}{\gamma_0} - 1\right) \tag{8}$$

can be generally applied.

If the null hypothesis $\mathrm{H}_0: \gamma = \gamma_0$ is verified, thus the asymptotic distribution of T_0 is standard normal.

It should be noted that this method can be used for all distributions, because asymptotic methods have been used and the method includes all exponential or non-exponential distributions. Moreover, the method is also appropriate if the populations have different distributions.

3 SIMULATION STUDY

In this Section various simulated datasets are reviewed and analyzed to verify the accuracy of the theoretical findings. To simulate different samples of one symmetric and two asymmetric distributions, populations X and Y are considered. For this purpose, we selected the normal, and the gamma and beta distributions, respectively, with a variety of values of CV, $(\mathrm{CV_X}, \mathrm{CV_Y}) \in \{(1,1), (1,2), (2,3), (2,5)\}$, that is, identical to $\gamma \in \{1,2,1.5,2.5\}$.

Table 1. The values of CP of the proposed method, for various settings of parameters

	(m,n)									Mean Absolute Difference from 0.95
distribution	(CV_X, CV_Y)	(50,100)	(75,100)	(100,100)	(100,200)	(200,300)	(500,500)	(500,700)	(700,1000)	*(1000,1000)*	
Normal	(1,1)	0.947	0.949	0.949	0.949	0.951	0.950	0.949	0.950	0.950	0.001
	(1,2)	0.948	0.949	0.949	0.951	0.950	0.950	0.950	0.949	0.951	0.001
	(2,3)	0.947	0.948	0.948	0.949	0.950	0.950	0.950	0.950	0.950	0.001
	(2,5)	0.947	0.949	0.949	0.950	0.950	0.950	0.951	0.951	0.950	0.001
Gamma	(1,1)	0.947	0.950	0.950	0.951	0.950	0.950	0.950	0.950	0.951	0.001
	(1,2)	0.946	0.949	0.949	0.951	0.950	0.950	0.950	0.949	0.950	0.001
	(2,3)	0.946	0.949	0.950	0.950	0.949	0.950	0.951	0.949	0.950	0.001
	(2,5)	0.946	0.948	0.949	0.950	0.951	0.949	0.949	0.949	0.950	0.002
Beta	(1,1)	0.947	0.950	0.950	0.950	0.951	0.950	0.950	0.950	0.950	0.001
	(1,2)	0.947	0.949	0.949	0.949	0.949	0.950	0.951	0.951	0.949	0.001
	(2,3)	0.946	0.949	0.949	0.950	0.950	0.950	0.950	0.950	0.950	0.001
	(2,5)	0.948	0.949	0.950	0.950	0.949	0.950	0.951	0.950	0.950	0.001
Mean Absolute Difference from 0.95		0.003	0.001	0.002	0.001	0.001	0.001	0.001	0.001	0.002	0.001

From Table 1, we conclude that the method have been successful in controlling type I error.

Table 2. The CPU time needed for running the proposed method, for various settings of parameters

		(m,n)					
Distribution	(CV_X, CV_Y)	(50,100)	(75,100)	(100,200)	(200,300)	(500,700)	(700,1000)
Normal	(1,1)	7.00	8.00	9.00	16.00	43.00	57.00
	(1,2)	8.00	8.00	12.00	19.00	21.00	48.00
	(2,3)	8.00	8.00	10.00	15.00	38.00	62.00
	(2,5)	7.00	8.00	11.00	19.00	26.00	61.00
Gamma	(1,1)	8.00	9.00	11.00	16.00	26.00	58.00
	(1,2)	8.00	9.00	12.00	20.00	27.00	65.00
	(2,3)	7.00	8.00	11.00	17.00	23.00	52.00
	(2,5)	7.00	8.00	9.00	17.00	28.00	50.00
Beta	(1,1)	8.00	9.00	11.00	20.00	43.00	49.00
	(1,2)	7.00	9.00	12.00	15.00	42.00	57.00
	(2,3)	7.00	8.00	12.00	17.00	37.00	55.00
	(2,5)	7.00	8.00	9.00	15.00	22.00	63.00

From Table 1 we verify that the CP of the proposed method is close to the intended level ($1-\alpha = 0.95$). This is more visible when the sample size increases. As a result, the introduced approach controls type I error. This means that about 95% of the simulated confidence intervals contain true γ, so it is accepted that the proposed confidence is an asymptomatic confidence interval for γ. Moreover, since the CP of the proposed method is close to the intended level, our approach is robust. Table 2 shows the CPU time needed for running the proposed method, for various combinations of parameters. We verify that our approach is fast.

Table 3. The P-values to investigate the normality of the statistic T_0

		(m,n)								
Distribution	(CV_X, CV_Y)	(50,100)	(75,100)	(100,100)	(100,200)	(200,300)	(500,500)	(500,700)	(700,1000)	(1000,1000)
Normal	(1,1)	0.061	0.165	0.849	0.541	0.885	0.863	0.875	0.886	0.861
	(1,2)	0.255	0.224	0.871	0.559	0.892	0.861	0.895	0.901	0.870
	(2,3)	0.112	0.104	0.845	0.891	0.903	0.895	0.874	0.896	0.887
	(2,5)	0.099	0.163	0.855	0.435	0.898	0.858	0.895	0.890	0.844
Gamma	(1,1)	0.130	0.236	0.564	0.905	0.885	0.845	0.896	0.903	0.853
	(1,2)	0.119	0.099	0.590	0.892	0.808	0.851	0.884	0,981	0.845
	(2,3)	0.120	0.231	0.866	0.900	0.903	0.863	0.896	0.871	0.895
	(2,5)	0.310	0.490	0.787	0.902	0.908	0.857	0.893	0.891	0.866
Beta	(1,1)	0.079	0.410	0.856	0.249	0.904	0.863	0.843	0.601	0.841
	(1,2)	0.319	0.325	0.864	0.903	0.876	0.896	0.815	0.877	0.890
	(2,3)	0.216	0.086	0.857	0.853	0.891	0.850	0.912	0.908	0.866
	(2,5)	0.106	0.286	0.835	0.884	0.901	0.874	0.902	0.909	0.904

From Table 3, it can also be seen that our method is more powerful

4 CONCLUSION

To compare populations, the CV is a useful, convenient and simple tool. In countless cases, two populations have the same CV, despite different means and variances. One way to understand the data structure is to study the equality of the two separate populations' CVs. However, when the difference between the two CVs is slight, it does not provide a solid and valuable interpretation. For this reason, the CV ratio is used, which is more accurate. Asymptotic distribution was proposed in this paper, and then the test of hypothesis and the asymptotic confidence interval for the CV ratio of two independent populations were extracted. The findings showed that the probability of coverage is very close to the desired level when increasing the sample. Based on this, it can be concluded that the introduced methodology controls the type I error. In addition, CPU times confirmed that the proposed method does not involve excessive computational burden. Also, the normality of the introduced test statistics was confirmed using the normal Shaprio-Wilks test. According to the results, it was observed that the asymptotic approximation acted very well for the whole simulated data set.

5 ACKNOWLEDGEMENTS

A short acknowledgement section can be written between the conclusion and the references. Sponsorship and financial support acknowledgments should be included here. Acknowledging the contributions of other colleagues who are not included in the authorship of this paper is also added in this section. If no acknowledgement is necessary, this section should not appear in the paper.

REFERENCES

[1] Meng, Q.; Yan, L.; Chen, Y., Zhang, Q. Generation of Numerical Models of Anisotropic Columnar Jointed Rock Mass Using Modified Centroidal Voronoi Diagrams. Symmetry 2018, 10(11), 618.

[2] Aslam, M.; Aldosari, M.S. Inspection Strategy under Indeterminacy Based on Neutrosophic Coefficient of Variation. Symmetry 2019, 11(2), 193.
[3] Iglesias-Caamaño, M.; Carballo-López, J.; Álvarez-Yates, T.; Cuba-Dorado, A.; García-García, O. Intrasession Reliability of the Tests to Determine Lateral Asymmetry and Performance in Volleyball Players. Symmetry 2018, 10(9), 16.
[4] Yue, Z; Baleanu, D. Inference about the Ratio of the Coefficients of Variation of Two Independent Symmetric or Asymmetric Populations. Symmetry 2019, 11.
[5] Haghbin, H.; Mahmoudi, M.R.; Shishebor, Z. Large Sample Inference on the Ratio of Two Independent Binomial Proportions. Journal of Mathematical Extension 2011, 5(1), 87- 95.
[6] Mahmoudi, M.R.; Mahmoodi, M. Inference on the Ratio of Variances of Two Independent Populations. Journal of Mathematical Extension 2014, 7(2), 83-91.
[7] Mahmoudi, M.R; Mahmoodi, M. Inference on the Ratio of Correlations of Two Independent Populations. Journal of Mathematical Extension 2014, 7(4), 71-82.
[8] Mahmoudi, M.R.; Nasirzadeh, R.; Mohammadi, M. On the Ratio of Two Independent Skewnesses. Commun Stat Theory Methods 2019, 48(7)1721-1727.
[9] Mahmoudi, M.R.; Behboodian, J.; Maleki, M. Large Sample Inference about the Ratio of Means in Two Independent Populations. J Stat Theory Appl 2017, 16(3), 366-374.
[10] Ferguson, Thomas S. A Course in Large Sample Theory. Chapman & Hall: London, UK, 1996.