IS THE SAMPLE COEFFICIENT OF VARIATION A GOOD ESTIMATOR FOR THE POPULATION COEFFICIENT OF VARIATION?

Mahmoudvand, Rahim and Hassani, Hossein and Wilson, Rob

Payame Noor University, IRAN, Cardiff University, UK, Cardiff University, UK

7 September 2007

World Applied Sciences Journal 2 (5): 519-522, 2007
ISSN 1818-4952

Is the Sample Coefficient of Variation A Good Estimator for the Population Coefficient of Variation?

[1]Rahim Mahmoudvand, [2,3]Hossein Hassani and [4]Rob Wilson

[1]Group of Statistics, Payame Noor University of Toyserkan, Iran
[2]Cardiff School of Mathematics, Cardiff University, UK
[3]Central Bank of the Islamic Republic of Iran, Iran
[4]Cardiff School of Mathematics, Cardiff University, UK

Abstract: In this paper, we obtain bounds for the population Coefficient of Variation (CV) in Bernoulli, Discrete Uniform, Normal and Exponential distributions. We also show that the sample coefficient of variation (cv) is not an accurate estimator of the population CV in the above indicated distributions. Finally we provide some suggestions based on the Maximum Likelihood Estimation to improve the population CV estimate.

Key words: Coefficient of Variation (CV) . Estimator . Maximum Likelihood Estimation (MLE)

INTRODUCTION

The coefficient of variation is usually used as a measure of precision for the dispersion of data sets [1, 2] and is also often used to compare numerical distributions measured on different scales. Ostle [3] found that the population CV is an ideal device for comparing the variation in two series of data which are measured in two different units (e.g., a comparison of variation in height with variation in weight). Lewis [4] showed that the population CV may be used to compare the dispersion of series measured in different units and also that of series with the same units but running at different levels of magnitude. Similarly, population CVs have been used to evaluate results from different experiments involving the same units of measure, possibly conducted by different persons [5].

Scientists and researchers are often interested in obtaining estimations and confidence intervals on population coefficient of variations. Vangel [6] and Verril [7] both discuss confidence intervals for the population CV in Normal and Log-normal distributions. It must also be mentioned that the confidence intervals are based on a point estimator. Hence, the accuracy of the point estimator is important too. The theoretical investigation of properties of the sample coefficient of variation has a long history. Two noteworthy examples being Summers [8] and Albercher et al. [2], the latter providing limits for the sample coefficient of variation in a Normal distribution.

Here, we discuss whether the sample coefficient of variation, cv, is a good estimator for the population CV. In section 2, we obtain bounds for the population coefficient of variation for both discrete and continuous distributions namely, Bernoulli, Discrete Uniform, Exponential and Normal distributions. We also consider the accuracy of the sample cv as an estimator for the population CV in each of the above indicated distributions. In section 3, we examine whether the Maximum Likelihood Estimation (*MLE*) can be used as a better estimator for the population CV and finally conclusions are presented in Section 4.

BOUNDS FOR POPULATION CV AND ITS ESTIMATOR

The coefficient of variation of a distribution with mean μ and variance σ^2 is defined as σ/μ. If $\bar{X}$ and S^2 are the corresponding sample mean and variance, then the sample $cv = S/\bar{X}$ can be regarded as an estimator for the population CV. It can also be shown that this is an asymptotically unbiased estimator for the population CV [2]. In the following we investigate two discrete and two continuous distributions and discuss the accuracy of the sample cv as an estimator for the population CV. Note however, similar results to those obtained in this section can be determined for many other distributions.

Bernoulli distribution: Let X~Bemoulli (P), with $0<p<1$, then

Corresponding Author: R. Mahmoudvand, Group of Statistics, Payame Noor University of Toyserkan, Iran

$$CV^2 = \frac{\sigma^2}{\mu^2} = \frac{P(1-P)}{P^2} = \frac{1-P}{P}$$

From the above equation we obtain the following bounds for the population CV.

$$\begin{cases} 0 \le CV \le 1 & \text{if } P \ge 1/2 \\ 1 \le CV \le \infty & \text{if } P \le 1/2 \end{cases} \quad (1)$$

Let us now consider some sample points from this distribution. For example, consider the points (1,0,0,0,0,0,0,0,0,0)

From Bemoulli (P≥1/2). The value of the cv for this sample is 3 but according to (1) it should be less than 1. Similarly, consider the points (1,1,1,1,1,1,1,1,1,1) from Bemoulli (P≤1/2). The sample cv for this case is 0 but according to (1) it should be greater than 1. These are just two samples that show the sample cv does not provide a good estimation for the population CV in some situations. In the following, we compute the probability of violating (1) for the sample cv in general.

Let $X_1,\ldots,X_n \sim$ Bemoulli (P), where P≥1/2. Then the violation from (1) for the sample cv is:

$$P(cv > 1) = P(cv^2 > 1) = P\left(\frac{n}{n-1}\frac{1-\bar{X}}{\bar{X}} > 1\right)$$

$$= P\left(\sum_{i=1}^{n} X_i > \frac{n^2}{2n-1}\right) = \sum_{k=0}^{w}\binom{n}{k}P^k(1-P)^{n-k}$$

where, $w = [n^2/(2n-1)]$ and note also that $\bar{X}$, the sample mean, is used as an estimator for P. Table 1 shows the value of this probability of violation for different values of n and P. As can be seen from Table 1, the probability of generating poor estimators for the sample cv is decreased when the values of n and P increase. Nevertheless, this probability is still relatively high for some cases. For example, when P = 0.6 and n = 10, the probability is almost 0.37.

Discrete uniform distribution: Let X~DU{0,1,…,u-1}. The probability mass function for this distribution is given by:

$$f_X(x) = \frac{1}{u}, \quad x = 0,1,...,u-1$$

It follows that,

$$\text{Population CV} = \sqrt{(u+1)/(3(u-1))} \text{ and } \sqrt{3}/3 < CV \le 1 \quad (2)$$

Table 1: Probability of generating a poor estimate of the population CV

P =	0.6	0.7	0.8	0.9
n = 10	0.3669	0.1503	0.0328	0.0016
n = 20	0.2447	0.0480	0.0026	0.0000
n = 30	0.1754	0.0169	0.0002	0.0000
n = 40	0.1298	0.0063	0.0000	0.0000
n = 50	0.0978	0.0024	0.0000	0.0000

Table 2: Values of sample cv for the discrete uniform distribution on the set {0, 1, 2} of size 4

$\sum_{i=1}^{4} X_i$	Symbol sample	Number of similar samples	cv
0	(0,0,0,0)	1	-
1	(1,0,0,0)	4	2.00
2	(2,0,0,0)	4	2.00
2	(1,1,0,0)	6	1.15
3	(2,1,0,0)	12	1.28
3	(1,1,1,0)	4	0.67
4	(2,2,0,0)	6	1.15
4	(2,1,1,0)	12	0.82
4	(1,1,1,1)	1	0.00
5	(2,2,1,0)	12	0.77
5	(2,1,1,1)	4	0.40
6	(2,2,2,0)	4	0.67
6	(2,2,1,1)	6	0.38
7	(2,2,2,1)	4	0.29
8	(2,2,2,2)	1	0.00

Similar to the Bernoulli distribution, there are many samples for which the sample cv does not satisfy (2). For example, consider two samples (u,0,0,…,0) and (1,1,…,1) each of size n. The values of the cv for these two samples are $\sqrt{n}$ and 0 respectively, which lie outside the bounds obtained in (2).

We can also calculate the probability of a poor estimate of the population CV. For example, let X~DU{0,1,2}. Table 2 represents all samples from this distribution of size 4 together with the value of the sample cv in each case. According to Table 2, $P(\sqrt{3}/3 < cv \le 1) = 33/81 = 0.41$. That is, 59 percent of samples give an unacceptable estimation for the population CV.

Exponential distribution: Let X~exp(θ). For this distribution the value of the population CV = 1, for all θ>0. Ten thousand samples of size n = 5, 10 and 50 were generated from an exponential distribution with mean 1. Figure 1 shows the probability of frequency

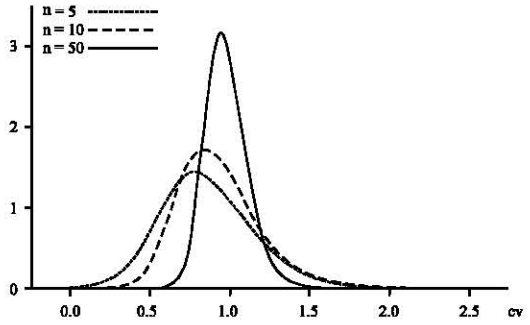

Fig. 1: Frequency plot for sample cv based on simulation

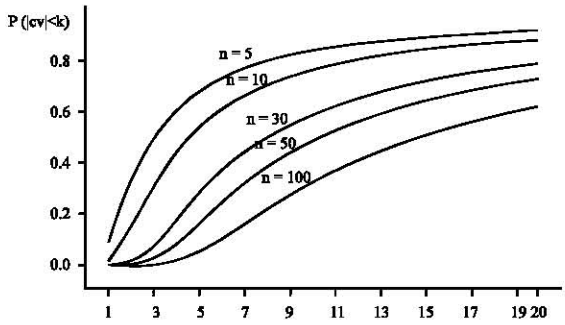

Fig. 2: P(|cv|<k) for different values k using the standard normal distribution

plot for the sample cv for the three different sample sizes. (The program for the simulation used in S-Plus is presented in appendix A). In Fig. 1 the vertical axis shows the probability density function of sample cv and horizontal axis shows the values of sample cv. It should be noted that the accuracy of this estimator increases as the values of n and P increase.

Normal distribution: Let $X \sim N(\mu, \sigma^2)$. In this distribution the population $CV=\sigma/\mu$ and there is no specific bound for this value [2, 7]. Here, we consider the accuracy of the sample cv to estimate the value of the population CV for the Standard Normal distribution.

Let $X_1,\ldots,X_n \sim N(0,1)$. It can be shown that:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}, \quad \frac{n(\bar{X})^2}{\sigma^2} \sim \chi^2_1$$

where S^2 is the sample variance and χ^2_v is the chi-square distribution with v degrees of freedom. It follows that:

$$\frac{\left(\frac{(n-1)S^2}{\sigma^2}\right)/(n-1)}{\left(\frac{n(\bar{X})^2}{\sigma^2}\right)/1} = \frac{S^2}{n(\bar{X})^2} = \frac{cv^2}{n} \sim F_{n-1,1}$$

where, $F_{m,n}$ represents the F distribution with m and n degrees of freedom. The value of the population CV for the Standard Normal distribution is infinity; however, there are many samples which produce a small finite value for the sample cv. For example, when n = 5 we have:

$$P(|cv| < 3) = P(F_{4,1} < 1.8) = 0.5$$

This suggests that in this case, almost 50 percent of samples lead to a very small estimation for the population CV. Figure 2 shows P(|cv|<k) for different values of n. Note also that this probability decreases as n increases for a fixed k. For example if k = 20, the probability is greater than 0.8 for n = 5 but is less than 0.6 when n = 100. This confirms that as n increases the sample cv provides a better estimate for the population CV.

MAXIMUM LIKELIHOOD ESTIMATION OF CV

Maximum Likelihood Estimation (MLE) is a popular statistical method used to make inferences about parameters of the underlying probability distribution from a given data set. As we observed in the previous sections, the sample cv provides a poor estimation for the population CV in many situations. We saw that the sample cv may lie outside the bounds obtained for the population CV, however, it should be noted that the range of the MLE coincides with that of the population parameters. Also MLEs are consistent estimators of their parameters and asymptotically efficient [9]. We now apply this property to estimate the parameters of each of the distributions examined above.

Bernoulli distribution: The MLE of P in the Bernoulli distribution is simply:

$$MLE(P) = \begin{cases} \min\{1/2, \bar{X}\} & , \ P \le 1/2 \\ \max\{1/2, \bar{X}\} & , \ P \ge 1/2 \end{cases}$$

We saw that the population CV for the Bernoulli distribution is $\sqrt{(1-P)/P}$. Using the invariance property of *MLE*s (if $\hat{\theta}$ is an MLE of θ and if g is a function, then $g(\hat{\theta})$ is an MLE of $g(\theta)$), we now see that the MLE for the population CV is:

$$MLE(CV) = \sqrt{(1 - MLE(P))/MLE(P)}$$

It follows that the bounds obtained now coincide with (1). (This result can also be confirmed for the previous samples examined.)

Discrete uniform distribution: The following is a good estimator for the population CV

$$MLE(CV) = \sqrt{(Y_n + 2)/(3Y_n)}$$

Where $Y_n = \max\{X_1,\ldots,X_n\}$. Notice that $\sqrt{3}/3 < MLE(CV) \leq 1$ which again coincides with the bounds for the CV obtained in (2) above. Thus this estimator doesn't have the defect of sample cv previously highlighted.

Exponential distribution: As was outlined above, the population CV for this distribution is always equal to 1. Similarly, the MLE(CV) = 1, which clearly provides a very good estimator in this case.

Normal distribution: The MLE of the population CV is as follows:

$$MLE(CV) = \frac{\left(\sum_{i=1}^{n}(X_i - X_n)^2\right)^{\frac{1}{2}}}{\sqrt{n}\bar{X}}$$

Note that the standard deviation is estimated with normalizing factor n instead of n-1, therefore the MLE of population CV is better than the sample cv for this distribution. Note also the discrepancy of these two estimators is very small as n increases. That is, these two estimators asymptotically coincide.

CONCLUSION

The coefficient of variation is one of the most popular measures for dispersion. This measure has many good features, but in some cases, using the coefficient of variation may lead to incorrect conclusions about empirical phenomena. In this paper we have demonstrated in many cases the sample cv provides a poor estimate of the population CV. Therefore the coefficient of variation should be used with care, if at all. The results of this paper show that using a MLE is better alternative to the classical estimator the population CV.

APPENDIX A

The S-plus program to generate ten thousand samples of sizes 5, 10 and 50 from an exponential distribution with mean 1 is:

```
dencvexp<-function(simulsize,samsize1,samsize2,samsiz3,param){
cv1<-array(dim=simulsize)
cv2<-array(dim=simulsize)
cv3<-array(dim=simulsize)
for(i in 1:simulsize){
x1<-rexp(samsize1,param)
x2<-rexp(samsize2,param)
x3<-rexp(samsize3,param)
cv1[i]<-sqrt(var(x1))/mean(x1)
cv2[i]<-sqrt(var(x2))/mean(x2)
cv3[i]<-sqrt(var(x3))/mean(x3)
}
guiPlot("density",DataSetValues=data.sheet(cv1,cv2,cv3))
}.
```

REFERENCES

1. Tian, L., 2005. Inferences on the common coefficient of variation. Statistics in Medicine, 24: 2213-2220.
2. Albercher, H, S.A. Ladoucette and J.L. Teogels, 2006. Asymptotic of the sample coefficient of variation and the sample dispersion, Submitted, available as KU. Leuven USC Report 2006-04.
3. Ostle, B., 1954. Statistics in research basic concepts and techniques for research workers. 1st Edn. Iowa State College Press, Ames, IA.
4. Lewis, E.E., 1963. Methods of statistical analysis in economics and business. 2nd Edn. Houghton Mifflin Co., Boston, MA.
5. Steel, R.G.D., J.H. Torrie and D.A. Dickey, 1997. Principles and procedures of statistics, a biometrical approach. 3rd Edn. McGraw Hill Book Co., New York, NY.
6. Vangel, M.G., 1996. Confidence intervals for a normal coefficient of variation, the American Statistician, 50: 21-26.
7. Verrill, S., 2003. Confidence bounds for normal and log-normal distribution coefficient of variation, Research Paper, EPL-RP-609, Madison, Wisconsin, U.S. workers. 1st Edn. Iowa State College Press, Ames, IA. Houghton Mifflin Co., Boston, MA.
8. Summers, R.D., 1965. An inequality for the sample coefficient of variation and an application to variables sampling. Technometrics, 7: 67-68.
9. Casella, G. and R.L. Berger, 2002. Statistical Inference. Pacific Grove, CA, Duxbury.