CONFIDENCE INTERVALS FOR THE COEFFICIENT OF VARIATION

Mark E. Payton

CONFIDENCE INTERVALS FOR THE COEFFICIENT OF VARIATION

Mark E. Payton
Department of Statistics
Oklahoma State University
Stillwater, OK 74078

ABSTRACT

The coefficient of variation (CV), defined as the ratio of the standard deviation to the mean, is often used in experimental situations. The exact distribution of the sample CV from a normally distributed population is complicated and obtaining a confidence interval for the population CV in this situation would require using the non-central t distribution and sequential techniques (Koopmans, et al., 1964). This paper explores the use of approximate distributions in determining confidence limits for the CV. The gamma distribution is used to model data appropriate for the calculation of the CV. A Monte Carlo simulation is performed to evaluate the effectiveness of four different intervals developed in this paper. A data set from a forestry experiment is analyzed using one of these techniques.

INTRODUCTION

Researchers in many fields use the coefficient of variation (CV) as a measure of relative variability. The sample CV is defined as the ratio of the sample standard deviation to the sample mean. Though the CV is calculated using sample values, rarely do practitioners express confidence limits associated with the CV. As Warren (1982) states: "workers will treat the sample coefficient of variation as if it were an absolute quantity. Inferences based on this measure of variability may then be questionable."

Rarely is it discussed what type of data is necessary for calculating the coefficient of variation. Zar (1984, p. 32) states that the CV should be used only for ratio-scale data since the CV is itself a ratio measure.

The coefficient of variation is not defined for a sample mean equal to zero, and it is unreliable for small sample means relative to the sample standard deviation. Using only ratio-scale data in calculating a CV alleviates this problem. Since negative values do not occur with ratio-scale data, a small sample mean would not occur along with a large sample standard deviation. The coefficient of variation could also be used as an ad-hoc test for normality for approximately ratio-scale data since large values for the CV would indicate highly skewed populations.

Researchers often use the CV to describe the variability associated with an experiment or a set of data. For example, Rusydi (1996) examined different sampling techniques for estimating the density and volume of teak plantations in Indonesia. Figure 1 displays the 6-tree sampling technique, where the density and volume are estimated by using the six closest trees from a randomly chosen point within a plot. Many techniques are compared, including n-tree sampling that ranges from 3-tree through 10-tree methods. A complicating factor in this study is that differing soil conditions will yield differing densities and volumes of teak. These are defined as "strata" in this study, the higher strata yielding larger volumes and densities. The mean, standard deviation, and coefficient of variation were calculated for each technique and stratum combination.

If one is interested in comparing the variabilities of the different sampling techniques at a given stratum, the sample standard deviations should be compared. Using the CV would be biased in favor of the techniques that overestimated the mean. However, if comparing variabilities of different strata for a given sampling technique was of interest, then the CV should be used since these strata all have different means. Refer to the first three columns of Table 1 (the remainder of Table 1 will be referred to later in this paper). This table displays a portion of the data from Rusydi (1996), where pairwise comparisons of CVs among the strata are desired for the 5-tree sampling method (this particular sampling method is the only one displayed for this paper; however, all sampling methods could be analyzed in this fashion). This can be accomplished by the calculation of confidence intervals for the coefficient of variation for each stratum at this particular sampling technique.

This paper explores techniques for obtaining a confidence interval for the population CV (or the ratio of the population standard deviation over the population mean). A variable expressed by McKay (1932) that is approximately Chi-square distributed is used to obtain an approximate $(1-\alpha)100\%$ confidence interval for σ/μ. It has been shown that McKay's approximation works well for normal data (Iglewicz and Myers (1970) and Iglewicz, et al. (1968)). The gamma distribution is used in this paper for modeling the data utilized in the calculation of the sample CV. Previous work by Bain and Engelhardt (1975) and Glaser (1976) is used to obtain approximate distributions of the ratio of the arithmetic mean to the geometric mean, and these approximations are used as pivotal quantities to obtain the aforementioned confidence intervals for the population CV. Simulation studies are performed to compare these techniques to that of McKay (1932) and to ascertain the effectiveness of the intervals calculated.

RECENT RELATED WORK

Other papers that have very recently been published address this issue. Vangel (1996) compared four confidence intervals for the CV: McKay (1932), Modified McKay (a method different from what is referred to as McKay-Modified later in this paper), David (1949), and an interval Vangel referred to as the "naive" approach. The "naive" confidence interval is an interval for the population standard deviation divided by the sample mean. Vangel recommends using the Modified McKay method. In other related work, Gupta and Ma (1996) derived a test for the equality of coefficients of variation from k normally distributed populations.

McKAY'S APPROXIMATION

McKay found that the quantity

$$\frac{nc^2(1+(\sigma/\mu)^2)}{(1+c^2)(\sigma/\mu)^2},$$

where c is the sample coefficient of variation, has approximately a Chi-square distribution with n-1 degrees of freedom. Iglewicz et al. (1968) and Iglewicz and Myers (1970) both state that this approximation works well under conditions of a normally distributed population and for values of the population CV under 0.3. The resulting $(1-\alpha)100\%$ confidence interval using McKay's approximation is

$$\left[\{\chi^2_U\ (1+c^2)/nc^2\ -\ 1\}^{-0.5},\ \{\chi^2_L\ (1+c^2)/nc^2\ -\ 1\}^{-0.5}\right],$$

where χ^2_L and χ^2_u refer to the lower and upper $\alpha/2$ percentiles of the Chi-square distribution with n-1 degrees of freedom, respectively. The formula for the confidence interval could be simplified by removing the quantity -1 from both the lower and upper confidence limits. This interval will be referred to as "McKay-Modified" later in this paper. One should note at this point that this method is not the same method Vangel (1996) refers to as "Modified McKay" in his paper. Vangel's modification complicated McKay's approximation. The modification proposed in this paper simplifies it. Both the McKay and McKay-Modified intervals will be studied further by simulation.

USING THE GAMMA DISTRIBUTION TO MODEL THE COEFFICIENT OF VARIATION

The gamma distribution can be used to model ratio-scale data that is appropriate for the calculation of the coefficient of variation. Consider the gamma distribution with parameters λ and r. The probability density function (p.d.f.) is of the form

$$f(x;\lambda,r) = \lambda^r\ x^{r-1}\ \exp(-\lambda x)/\Gamma(r); \qquad 0<x<\infty$$

$$\lambda>0,\ r>0.$$

The mean of the distribution is r/λ, and the variance is r/λ^2. The population coefficient of variation is therefore $r^{-1/2}$. If one wishes to set confidence limits for the population CV, it suffices to place confidence limits on r.

The arithmetic mean and geometric mean are complete sufficient statistics for the parameters of the gamma distribution. Assume X is gamma distributed. Define U as the log of the ratio of the arithmetic mean to the geometric mean, i.e.,

$$U = \log \frac{(\Sigma x_i)/n}{(\Pi x_i)^{1/n}} .$$

Bain and Engelhardt (1975) showed that the quantity $2 \cdot k(c) \cdot rnU$ is Chi-square distributed with $\nu(r) \cdot (n-1)$ degrees of freedom, where k is a function (given below) of the sample coefficient of variation, c, and ν is approximately equal to the quantity $1 + r/(r + 8.6\sqrt{r} + 18.49)$. This function is very close to one for values of the sample CV less than 0.5, so the degrees of freedom for the Chi-square approximation will be simplified to the quantity n-1. Using the above, a $(1-\alpha)100\%$ confidence interval for the population CV can be stated as

$$\left[\{2nU \cdot k(c)/\chi^2_U\}^{0.5}, \{2nU \cdot k(c)/\chi^2_L\}^{0.5} \right],$$

where $k(c) = 6/(c^2+6)$.

Glaser (1976) reported that the quantity 2rnU has approximately a Chi-square distribution with n-1 degrees of freedom as r approaches infinity if the x_i, $i = 1, \ldots, n$ are sampled from a gamma distribution. This result can be used to obtain the following confidence interval for the population CV:

$$\left[\{2nU/\chi^2_U\}^{0.5}, \{2nU/\chi^2_L\}^{0.5} \right].$$

Since this confidence interval utilizes a limiting distribution for large r, this interval's effectiveness would be questionable for relatively large CV's.

SIMULATION RESULTS USING NORMALLY DISTRIBUTED DATA

Ten thousand random samples of size n were generated from a normal population using PC SAS. The sample size n were 10, 30, and 50. The normal populations had population CV's of 0.05, 0.15, and 0.25. Confidence levels used for the confidence intervals were 99%, 95%, and 90%. Table 2 displays the results of the simulation, where Mc, Mc-M, BE, and Gl stand for McKay, McKay-Modified, Bain and Engelhardt, and Glaser, respectively. The boldface numbers represent simulations that fell within 95% tolerance limits calculated using the normal approximation to the binomial. The average lengths of these intervals were examined (data not shown) and the McKay's-Modified, Glaser, and Bain and Engelhardt intervals were all very close to each other. The McKay's interval tended to be longer than the other three on average.

The results of the simulation study indicate that the confidence intervals created using Glaser's or Bain and Engelhardt's approximation are adequate for small coefficients of variation. However, McKay's and the modified McKay's approximations are uniformly more accurate over the range of values for the CV in attaining the desired level of confidence. In fact, the modified McKay's approximation performs well regardless of level of confidence, sample size, or population CV.

SIMULATION RESULTS USING GAMMA DISTRIBUTED DATA

The simulation study performed above with the samples coming from normal distributions was repeated with the samples originating from gamma distributions. This was performed to determine whether the distribution of the data affects the effectiveness of the methods in question. Table 2 contains the results of these simulations.

As expected, the procedures that utilized the gamma distribution in their formation (Bain and Engelhardt, Glaser) performed well in the simulations involving samples from a gamma. However, intervals using McKay's and McKay's-Modified approximations performed adequately, with the possible exception of the combination of small population CV's (0.05 - 0.10) and small samples (n=10). This would suggest that McKay's and the McKay's-Modified approximations are robust against departures from normality, an aspect of the approximation that has yet to be fully determined.

CONFIDENCE INTERVALS FOR FORESTRY DATA

Table 1 reports the results of calculating 85% confidence intervals for the coefficient of variation using the McKay-Modified method for each strata for the 5-tree sampling technique. Pairwise comparisons of the strata are made by comparing the individual confidence intervals to see if they overlap or not. No formal test of hypothesis is attempted here; the CI-overlap method is being used to approximate one. The significance level of 85% was used to obtain pairwise comparisons with approximate comparisonwise error rates of 5%. Had 95% confidence intervals been used on the individual stratum, the resulting error rates of the pairwise comparisons would have very small (approximately 1%).

SUMMARY

McKay's and the modified McKay's approximations are adequate for arriving at confidence intervals for the population coefficient of variation. It is recommended that the following formula be used when a confidence interval for the CV is desired:

$$\left[\{\chi^2_U\ (1+c^2)/nc^2\}^{-0.5},\ \{\chi^2_L\ (1+c^2)/nc^2\}^{-0.5}\right],$$

which is the interval using McKay's approximation without the quantity -1 (referred to as "McKay-Modified" in this paper). This interval is accurate over most combinations of confidence level, sample size, and population CV. Surprisingly, the confidence interval has been shown to be robust for data from gamma distributed populations.

ACKNOWLEDGMENTS

The author would like to thank Sandra Caldwell of Paris Junior College for preliminary work on this topic. as well as Dr. Tom Lynch and R. Rusydi of the Oklahoma State University Department of Forestry for providing data and a suitable setting for application of the procedures in this paper. The author would also like to thank the referee for his time and effort in providing a helpful critique that improved this paper greatly.

REFERENCES

Bain, L.J. and Engelhardt, M. (1975). "A Two-Moment Chi-Square Approximation for the Statistic Log($\bar{x}$/~x)". *Journal of the American Statistical Association*, 70, 948-950.

David, F.N. (1949), "Note on the Application of Fisher's k-statistics". *Biometrika*, 36, 383-393.

Glaser, R.E. (1976). "The Ratio of the Geometric Mean to the Arithmetic Mean for a Random Sample from a Gamma Distribution". *Journal of the American Statistical Association*, 71, 480-487.

Gupta, R.C. and Ma, S. (1996). "Testing the Equality of the Coefficient of Variation in k Normal Populations". *Communications in Statistics*, 25(1), 115-132.

Iglewicz, B. and Myers, R.H. (1970). "Comparisons of Approximations to the Percentage Points of the Sample Coefficient of Variation". *Technometrics*, 12, 166-169.

Iglewicz, B., Myers, R.H., and Howe, R.B. (1968). "On the Percentage Points of the Sample Coefficient of Variation". *Biometrika*, 55, 580-581.

Koopmans, L.H., Owen, D.B. and Rosenblatt, J.I. (1964). "Confidence Intervals for the Coefficient of Variation for the Normal and Lognormal Distributions". *Biometrika*, 51, 25-39.

McKay, A.T. (1932). "Distributions of the Coefficient of Variation and the Extended t Distribution". *Journal of the Royal Statistical Association*, 95, 695-698.

Rusydi, R. (1996). "Comparison of Inventory Methods for Teak Plantations at Three Forest Management Areas in East Java, Indonesia". M.S. Thesis, Oklahoma State University Department of Forestry.

Vangel, M.G. (1996). "Confidence Intervals for a Normal Coefficient of Variation". *The American Statistician*, 50(1), 21-26.

Warren, W.G. (1982). "On the Adequacy of the Chi-squared Approximation for the Coefficient of Variation". *Communications in Statistics - Simulation and Computation*, 11(6), 659-666.

Zar, J.H. (1984). Biostatistical Analysis (2nd Ed.). Prentice Hall, Inc., Englewood, Cliffs, NJ.

Table 1

Sample sizes, CV's and 85% confidence intervals associated with the five strata for the 5-tree sampling method from Rusydi (1996). McKay's-Modified method was used for calculating the confidence intervals. Letters denote significant differences among strata.

Strata	Sample size	CV	85% CI
1	18	17.3 c	(13.5, 25.2)
2	30	8.8 ab	(7.3, 11.7)
3	24	8.6 ab	(7.0, 12.0)
4	12	5.1 a	(3.9, 8.6)
5	9	12.5 bc	(9.2, 23.7)

Table 2
Observed confidence levels from 10000 samples simulated from normal and gamma distributions.

			normal		CV=0.05		gamma	
Sample size:		10	30	50		10	30	50
99%	Mc	0.9840	0.9875	**0.9896**		0.9869	**0.9898**	0.9877
	Mc-M	0.9841	0.9878	**0.9902**		0.9870	**0.9899**	**0.9881**
	BE	**0.9894**	**0.9886**	**0.9896**		**0.9913**	**0.9908**	**0.9882**
	Gl	**0.9893**	**0.9884**	**0.9895**		**0.9913**	**0.9908**	0.9880
95%	Mc	0.9319	0.9453	0.9455		0.9346	0.9455	**0.9491**
	Mc-M	0.9324	0.9457	**0.9459**		0.9351	0.9457	**0.9491**
	BE	0.9450	0.9457	0.9453		**0.9484**	**0.9479**	**0.9504**
	Gl	0.9449	0.9457	0.9451		**0.9484**	**0.9480**	**0.9504**
90%	Mc	0.8746	**0.8947**	0.8928		0.8811	**0.8954**	**0.8994**
	Mc-M	0.8755	**0.8960**	0.8934		0.8826	**0.8961**	**0.9000**
	BE	0.8917	**0.9003**	**0.8944**		**0.9025**	**0.9011**	**0.9040**
	Gl	0.8917	**0.9000**	**0.8942**		**0.9023**	**0.9009**	**0.9039**

			normal		CV=0.15		gamma	
Sample size:		10	30	50		10	30	50
99%	Mc	0.9833	**0.9884**	**0.9894**		0.9866	**0.9896**	**0.9908**
	Mc-M	0.9852	**0.9897**	**0.9905**		**0.9881**	**0.9904**	**0.9916**
	BE	0.9819	0.9802	0.9825		**0.9889**	**0.9896**	**0.9905**
	Gl	0.9812	0.9797	0.9820		**0.9886**	**0.9894**	**0.9901**
95%	Mc	0.9381	**0.9466**	**0.9461**		0.9401	**0.9485**	**0.9472**
	Mc-M	0.9413	**0.9487**	**0.9481**		0.9440	**0.9524**	**0.9498**
	BE	0.9367	0.9299	0.9254		**0.9520**	**0.9508**	**0.9474**
	Gl	0.9354	0.9278	0.9238		**0.9509**	**0.9494**	**0.9468**
90%	Mc	0.8802	**0.8967**	**0.8953**		0.8896	**0.8983**	**0.8992**
	Mc-M	0.8846	**0.9017**	**0.8994**		**0.8952**	**0.9013**	0.9064
	BE	0.8824	0.8770	0.8661		**0.9020**	**0.8981**	**0.9019**
	Gl	0.8813	0.8743	0.8641		**0.9006**	**0.8968**	**0.9014**

			normal		CV=0.25		gamma	
Sample size:		10	30	50		10	30	50
99%	Mc	0.9861	**0.9891**	**0.9897**		**0.9895**	**0.9908**	**0.9898**
	Mc-M	**0.9896**	**0.9904**	**0.9911**		**0.9918**	**0.9916**	**0.9899**
	BE	0.9605	0.9460	0.9355		**0.9899**	**0.9904**	**0.9881**
	Gl	0.9574	0.9415	0.9304		**0.9892**	**0.9891**	0.9868
95%	Mc	0.9370	0.9455	**0.9517**		**0.9493**	0.9561	**0.9525**
	Mc-M	**0.9484**	**0.9514**	**0.9525**		0.9584	0.9582	0.9567
	BE	0.9025	0.8715	0.8603		**0.9528**	**0.9504**	**0.9526**
	Gl	0.8981	0.8646	0.8503		**0.9492**	**0.9469**	**0.9491**
90%	Mc	0.8779	**0.8951**	0.8891		**0.9041**	0.9072	0.9165
	Mc-M	0.8927	**0.9033**	**0.8970**		0.9179	0.9110	0.9128
	BE	0.8380	0.8002	0.7745		0.9078	**0.9007**	0.9094
	Gl	0.8317	0.7924	0.7640		**0.9026**	**0.8963**	**0.9026**

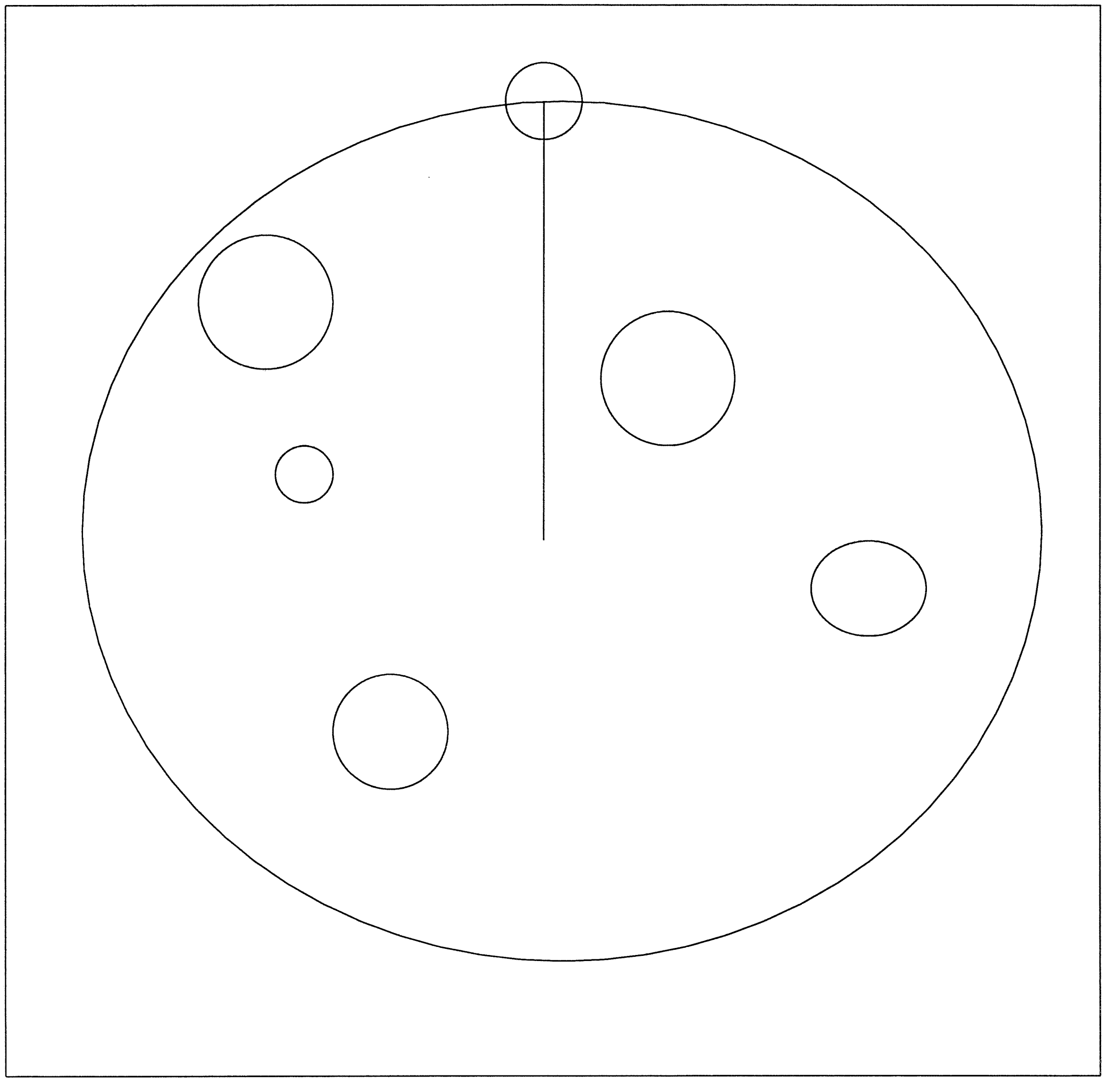

Figure 1. Determination of radius, sample area and trees included in a 6-tree sampling method