Approximate Confidence Intervals for Coefficients of Variation

by

James N. Arvesen

Purdue University

Department of Statistics
Division of Mathematical Sciences
Mimeograph Series #305

September 1972

Research supported in part by Office of Naval Research Contract N00014-67-A-0226-00014 at Purdue University. Reproduction in whole or part is permitted for any purpose of the United States Government.

Abstract

A confidence interval (or test) is obtained for the population coefficient of variation. The procedure is based on the jackknife. A Monte Carlo simulation compares it to the chi-square approximation of McKay (JRSS, 1932). An extension to two sample problems and an application is given.

1. Introduction

Let $X_1, \ldots, X_n$ be independent $\mathcal{N}(\mu, \sigma^2)$ random variables. Let $\beta = \sigma/|\mu|$, $\mu \neq 0$ denote the coefficient of variation. This paper reviews the classical confidence interval for β suggested by McKay [1932], and compares it to a confidence interval based on the jackknife technique.

Section 2 presents the situation described above, as well as a Monte Carlo simulation of McKay's result and the jackknife version. Section 3 generalizes these results to some two sample cases. Finally, Section 4 presents an application to a biological problem.

2. Confidence Intervals

If $\overline{X} = n^{-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$, McKay [1932] obtains the distribution of $b = s/|\overline{X}|$ as

$$f_b(t) = \left(\frac{n}{\sigma^2}\right)^{n/2} \frac{t^{n-2}}{2^{\frac{n-1}{2}} \sqrt{\pi}\, \Gamma\left(\frac{n-1}{2}\right)} \int_{-\infty}^{\infty} |x|^{n-1} e^{-n[t^2x^2 + (x-\mu)^2]/2\sigma^2} dx. \qquad (1.1)$$

In what follows, the maximum likelihood estimate (MLE) estimate of σ will be used (assuming normality). The more complicated unbiased estimate of σ could also be used.

Perhaps a more useful fact was McKay's approximation that if $\beta \leq .3$, $\frac{nb^2}{1+b^2} \frac{1+\beta^2}{\beta^2}$ has approximately a chi-square distribution with

(n-1) degrees of freedom. This approximation was further substantiated by Fieller [1932] and Pearson [1932]. Thus, if $\chi^2_{n-1;\alpha}$ denotes the upper α point of a chi-square distribution with n-1 degrees of freedom,

$$P\left(\chi^2_{n-1;1-\alpha} \leq \frac{nb^2(1+\beta^2)}{\beta^2(1+b^2)}\right) \cong 1 - \alpha. \tag{1.2}$$

Consequently, an approximate $(1-\alpha) \times 100\%$ confidence interval for β^2 is

$$\left[0, \left(\frac{\chi^2_{n-1;1-\alpha}(1+b^2)}{nb^2} - 1\right)^{-1}\right] .$$

Similar results hold for two-sided intervals, or tests of hypotheses.

Without the assumption of normality, this chi-square approximation is not valid. Under the normality assumption,

$$\text{var } (b) \cong \beta^2(1+2\beta^2)/(2n). \tag{1.3}$$

Without the assumption of normality, the situation is more complicated. If the parent population has v^{th} central moments μ_v, then

$$\text{var } (b) = \frac{\mu_1^2(\mu_4-\mu_2^2)-4\mu_1\mu_2\mu_3 + 4\mu_2^3}{4n\mu_1^4\mu_2}, \tag{1.4}$$

as shown in Cramér [1946]. Thus, assuming the skewness, μ_3, is zero, if the kurtosis is positive ($\mu_4 \geq 3\mu_2^2$), the confidence interval in (1.2) is too small (or has true level less than $(1-\alpha) \times 100\%$). The converse situation obtains if the kurtosis is negative.

The jackknife technique however, provides an approximate confidence interval (or test) for β regardless of the distribution assumptions.

Theorems 8 and 9 of Arvesen [1969] show this (asymptotically exact), as long as the parent population has finite fourth moments. In practice, with a finite number of observations, one would prefer to use the jackknife on a transformation of b. Let $c = 1/b$, $\gamma = 1/\beta$, then

$$\text{var}(c) = (2+\gamma^2)/2n \tag{1.5}$$

with the normality assumption. Hence an appropriate variance stabilizing transformation can be readily seen to be

$$\log\ (c+(2+c^2)^{\frac{1}{2}}). \tag{1.6}$$

Note that if c is large (b small) this is proportional to log (c). It was preferable to work with the statistic c rather than b as its distribution tends to be more symmetric. An earlier Monte Carlo result also bears this out.

Thus, following the notation in Arvesen [1969], if $\hat{\theta}$ denotes the jackknife estimate of $\log(\gamma+(2+\gamma^2)^{\frac{1}{2}})$ based on the estimate of $\log\ (c+(2+c^2)^{\frac{1}{2}})$, and $s^2_{\hat{\theta}}$ denotes the sum of squares of pseudo values, an approximate $(1-\alpha)$ x 100% level <u>lower</u> confidence interval for $\log\ (\gamma+(2+\gamma^2)^{\frac{1}{2}})$ is

$$[L = \hat{\theta}-Z_\alpha\ s_{\hat{\theta}}/\sqrt{n},\ \infty),$$

where Z_α is the upper α point of a standard normal. Hence an approximate $(1-\alpha)$ x 100% level <u>upper</u> confidence interval for β is given by

$$[0,\ (e^{2L}-2)/(2e^{L})]. \tag{1.7}$$

This assumes that in using the jakcknife, the number of groups, n, is large. In practice, n more than 30 seems to suffice. Otherwise the t-distribution is used.

A Monte Carlo simulation was run to assess these intervals. A sample of 25 independent pseudo random variables with means $\mu = 10,5,3,\frac{1}{3}$, 2 and 1, with standard deviation $\sigma = 1$ was generated. Thus $\beta = .1,.2,.3,.5$ and 1 for the parent population. The method of generation is described in Rubin [1972]. Each time a pseudo random variable is generated, it is transformed for the five β values. Three parent populations were chosen, normal, double exponential, and uniform. It was noted whether the interval in (1.2), or the jackknife interval covered the true β at the appropriate level. The transformation $c = 1/b$, as well as the more complicated transformation of (1.7) were used. Since $n = 25$, the upper α point of a t distribution with 24 degrees of freedom was used instead of the normal distribution. Of course, the practical difference is essentially nil, however slightly better results were obtained. Confidence levels of 90% and 95% were selected. The simulation was repeated 1000 times. Thus, a total of 75,000 pseudo random variables were generated. The results are given in tables I and II for the nominal 90% and 95% confidence intervals.

In examining tables I and II one notes the low coverage of McKay's interval for the double exponential, and the high coverage for the uniform. In addition, McKay's interval seems to give too much coverage for large values of β. The jackknife, at least as used in (1.7), performs quite admirably for all three distributions. From a practical point of view, a user would want to decide whether simply the log or no transformation of c is necessary.

3. Two sample cases

In Lorhding [1969], the following problem was treated. Given X_{ij} are independent $\mathcal{N}(\mu_i, \beta^2\mu_i)$, j=1,..,n the MLE of β is

$$\hat{\beta} = \frac{[(2(1+2b_1^2)^{\frac{1}{2}}(1+2b_2^2)^{\frac{1}{2}})((1+2b_1^2)(1+2b_2^2)-1)]^{\frac{1}{2}}}{(1+b_1^2)^{\frac{1}{2}}\ (1+b_2^2)^{\frac{1}{2}}}\ , \tag{2.1}$$

where $b_i = S_i/|\bar{X}_i|$, i=1,2, and $\bar{X}_i$, S_i^2 are the sample mean and MLE of the i^{th} population variance, i=1,2. Zeigler [1972] shows that $\hat{\beta} \doteq ((b_1^2+b_2^2))^{\frac{1}{2}}$, and thus for β^2 small, an appropriate transformation is log $\hat{\beta}$ upon which to base the jackknife. It is not clear whether another transformation would be preferable. Thus an approximate (1-α) x 100% level upper confidence interval for log β based on the jackknife would be $[0, \hat{\theta}+Z_\alpha s_{\hat{\theta}}/\sqrt{n}]$, and hence for β the interval would be $(0, \exp(\hat{\theta}+Z_\alpha s_{\hat{\theta}}/\sqrt{n})]$. This assumes a randomly selected pair X_{1j}, X_{2j} is deleted each time, j=1,...,n, and n is large.

A similar procedure would hold for the k sample case. Zeigler proposes

$$\hat{\beta} = \left[\frac{\sum_{i=1}^{k} \frac{b_i^2}{1+b_i^2}}{k(\frac{n-3}{n}) - \sum_{i=1}^{k} \frac{b_i^2}{1+b_i^2}}\right]^{\frac{1}{2}} \tag{2.2}$$

as an estimate in this case. He also shows that this estimate is essentially equivalent to Lohrding's if β is small ($\beta \leq .30$) and k = 2.

Another possibility in the two sample case is that one has

$(X_{1j}, X_{2j})'$ independent $\eta\left((\mu_1,\mu_2);\begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$, $j=1,\ldots,n$,

with $|\rho|< 1$, $\mu_i \neq 0$, $i = 1,2$. In this case the MLE of $\beta_i = \sigma_i/|\mu_i|$, are $b_i = S_i/|\overline{X}_i|$, $i=1,2$. The joint density of b_1, b_2 when ρ is unknown is undoubtedly very complicated. In fact, an approximation such as McKay's is probably also difficult when ρ is unknown. But a confidence interval for β_1/β_2 can readily be obtained using the jackknife.

Take as an initial estimate of $\theta = \log(\beta_1/\beta_2)$, the estimate $\log(b_1/b_2)$, and apply the jackknife technique. With $\hat{\theta}$, $s^2_{\hat{\theta}}$ as defined above, a $(1-\alpha) \times 100\%$ level two-sided confidence interval for β_1/β_2 is

$[\exp(\hat{\theta} - Z_{\alpha/2}\, s_{\hat{\theta}}/\sqrt{n}),\ \exp(\hat{\theta}+Z_{\alpha/2}\cdot s_{\hat{\theta}}/\sqrt{n})]$.

4. An Application

Consider the model

$$Y_{1ijk} = \mu_1 + \alpha_{1i} + \beta_{1j} + \gamma_{1ij} + \varepsilon_{1ijk},$$

$$Y_{2ijk} = \mu_2 + \alpha_{2i} + \beta_{2j} + \gamma_{2ij} + \varepsilon_{2ijk},$$

$i=1,\ldots, I$, $j=1,\ldots, J$, $k=1,\ldots, K_{ij}$,

$(\varepsilon_1,\varepsilon_2)'$ distributed independent $\eta\left((0,0), \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$. Let

$\beta_\ell = \sigma_\ell/|\mu_\ell|$, which has as MLE $b_\ell = s_\ell/|\overline{Y}_{\ell\ldots}|$, $\ell=1,2$. $\overline{Y}_{\ell\ldots}$ denotes the overall mean, and

$$s_\ell^2 = [\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K_{ij}} (Y_{\ell ijk} - Y_{\ell ij\cdot})^2]/(\sum_{i=1}^{I}\sum_{j=1}^{J} K_{ij}),\ \ell = 1,2.$$

In Ithakissios et al [1972], interest centered on the coefficient of variation for data measuring Cadmium uptake in the liver of rats. The data was recorded from an experimental design as given above with $I = 2$, $J = 4$, $K_{1j} = 5$, $K_{2j} = 7$. The I factor denotes whether the Cadmium injection was intravenous or intraperintoneal, J denotes a batch effect, and the K factor is replication. The data was recorded both for total Cadmium uptake, and for Cadmium uptake per gm. of liver tissue on each animal. Thus the data is (possibly) correlated.

The jackknife applied to these data yielded as a 99% confidence interval for $\log(\beta_1/\beta_2)$ the interval [-1.43,-.37]. Thus a 99% C.I. for β_1/β_2 is the interval [.24, .69].

Using the coefficient of variation as a measure of precision, one sees that more precise results are obtained recording the data on a whole body basis rather than on a per gm. basis. This leads one to conjecture that Cd. uptake in the liver of rats is not a function of organ size.

TABLE I - Empirical Coverage Probability

Nominal significance level = .90

	β	McKay's interval	c = 1/b	Interval of (1.7)
Normal Distribution	.1	.881	.884	.901
	.2	.884	.876	.902
	.3	.889	.875	.901
	.5	.895	.876	.900
	1.0	.926	.894	.905
Double exponential Distribution	.1	.770	.881	.899
	.2	.780	.880	.898
	.3	.790	.885	.902
	.5	.808	.895	.905
	1.0	.888	.904	.916
Uniform Distribution	.1	.966	.873	.891
	.2	.965	.875	.890
	.3	.965	.877	.891
	.5	.952	.883	.896
	1.0	.957	.909	.919

TABLE II - Empirical Coverage Probability

Nominal significance level = .95

	β	McKay's interval	c = 1/b	Interval of (1.7)
Normal Distribution	.1	.944	.934	.954
	.2	.942	.938	.954
	.3	.942	.935	.956
	.5	.947	.941	.956
	1.0	.974	.949	.960
Double exponential Distribution	.1	.840	.939	.957
	.2	.851	.939	.952
	.3	.854	.938	.956
	.5	.881	.942	.954
	1.0	.932	.954	.959
Uniform Distribution	.1	.987	.929	.947
	.2	.986	.931	.945
	.3	.983	.930	.941
	.5	.983	.935	.948
	1.0	.987	.954	.960

References

[1] Arvesen, J., "Jackknifing U-statistics," Ann. Math. Statist., 40, No. 6, (1969), 2076-2100.

[2] Cramér, H., Mathematical Methods of Statistics, Princeton University Press, Princeton (1946).

[3] Fieller, E., "A numerical test of the adequacy of A. T. McKay's approximation", Jour. Roy. Stat. Soc., 95

[4] Ithakissios, D., Kessler, W., and Arvesen, J., "Variability of Cd uptake in rats as affected by route of administration and manner of expressing results", Technical paper, Purdue University School of Pharmacy, (1972).

[5] Lohrding, R. K., "A test of equality of two normal population means assuming homogeneous coefficients of variation," Ann. Math. Statist., 40, No. 4, (1969), 1374-1385.

[6] McKay, A., "Distribution of the coefficient of variation and the extended 't' distribution", Jour. Roy. Stat. Soc., 95, (1932), 695-698.

[7] Pearson, K., "Comparison of A. T. McKay's approximation with experimental sampling results", Jour. Roy. Stat. Soc., 95, (1932) 703-704.

[8] Rubin, H., "Some fast methods of generating random variables with preassigned distributions," Purdue University Statistics Department Technical Report, (1972).

[9] Zeigler, R., "Estimators of coefficient of variation using k samples", Los Alamos Scientific Laboratory Preprint, (1972).

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Purdue University	Unclassified
	2b. GROUP

3. REPORT TITLE

Approximate Confidence Intervals for Coefficients of Variation

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report, September 1972

5. AUTHOR(S) (Last name, first name, initial)

Arvesen, James N.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
September 1972	8	9

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
N00014-67-A-0226-00014	Mimeo Series #305
b. PROJECT NO.	
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. AVAILABILITY/LIMITATION NOTICES

Distribution of this document is unlimited.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	Office of Naval Research Washington, D.C.

13. ABSTRACT

A confidence interval (or test) is obtained for the population coefficient of variation. The procedure is based on the jackknife. A Monte Carlo simulation compares it to the chi-square approximation of McKay (JRSS, 1932). An extension to two sample problems and an application is given.

DD FORM 1 JAN 64 1473

Unclassified
Security Classification