Pak.j.stat.oper.res. Vol.20 No.2 2024 pp 157-170 DOI: http://dx.doi.org/10.18187/pjsor.v20i2.4240

Pakistan Journal of Statistics and Operation Research

Sample size determination when the parameter of interest is the coefficient of variation under normality for the data

Jorge Alberto Achcar[1], Emerson Barili[2*]

*Corresponding author

1. Medical School, University of São Paulo, Ribeirão Preto - São Paulo, Brazil, achcar@fmrp.usp.br
2. Medical School, University of São Paulo, Ribeirão Preto - São Paulo, Brazil, ebarili2@gmail.com

Abstract

This study considers classical and Bayesian inference approaches for the coefficient of variation under normality for the data, especially on the determination of the sample size of a random sample needed in the second stage of an experiment. This topic has been explored by many authors in the last decades. The first goal of the study is to present simple formulations to get the inferences of interest for the coefficient of variation under normality and the usual frequentist approach based on the asymptotic normality of the maximum likelihood estimators for the mean and standard deviation of the normal distribution and using the delta method to get the inferences of interest for the coefficient of variation. Simple hypothesis tests and determination of the sample size are discussed under the frequentist approach.The second goal of the study is to present a sample size determination under a Bayesian approach, where it is assumed a Jeffreys non-informative prior distribution of the parameters of the normal distribution assumed for the data and using standard Markov Chain Monte Carlo (MCMC) methods to get the posterior summaries of interest.

Key Words: coefficient of variation; normal distribution; sample determination; delta method; Bayesian approach; MCMC methods.

1. Introduction

The coefficient of variation (CV) defined by $CV = \sigma/\mu$, where σ is the standard deviation and μ is the mean, is a statistical measure of the dispersion of data around the mean also known as relative standard deviation (RSD). The coefficient of variation provides a relatively simple tool to compare different data series which is an advantage for the use of standard deviation which is measured in the context of the mean of the data. In some areas of application, as finance for example, the CV is used in investment selection where the CV is linked to the risk-to-reward ratio of investment. Other areas of applications of the CV are given in engineering or physics linked to quality assurance studies and in neuroscience.

An advantage for the use the CV in applications is that the actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number. A disadvantages for the use of the CV is observed when the mean value is close to zero. In this situation, the coefficient of variation will approach infinity and is therefore sensitive to small changes in the mean.

Assuming a random sample of size n given by the vector $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ of a normal distribution $N(\mu, \sigma^2)$, a bised estimator of CV is given by $\widehat{(CV)} = s/\overline{X}$ where s^2 is the sample variance defined by $(n-1)s^2 =$

$\sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2$ and $\overline{X}$ is the sample mean given by $n\overline{X}=\sum_{j=1}^{n}X_j$. An unbiased estimator (Sokal and Rohlf, 1995) is given by $\widehat{CV}^*=(1+1/4n)\widehat{CV}$.

The sampling distribution of the CV was introduced by Hendricks and Robey (1936) which could be useful in the construction of hypothesis tests or confidence intervals is given by the probability density function,

$$f(y)=\frac{2}{\Gamma\left(\frac{n-1}{2}\right)}\exp\left[-\frac{ny^2}{2\left(1+y^2\right)\left(\frac{\sigma^2}{\mu}\right)}\right]\frac{y^{n-2}}{\left(1+y^2\right)^{n/2}}\sum_{i=0}^{n-1}\frac{(n-1)!n^{i/2}\Gamma\left(\frac{n-1}{2}\right)}{(n-1-i)!i!2^{\frac{i}{2}}\left(1+y^2\right)\left(\frac{\sigma}{\mu}\right)^i} \tag{1}$$

where y denotes the CV.

Statistical inference for the coefficient of variation in normally distributed data is often based on McKay's chi-square approximation for the coefficient of variation ((Iglewicz and Myers, 1970); (Bennett, 1976); (Vangel, 1996); (Feltz and Miller, 1996); (Forkman, 2009); (Krishnamoorthy and Lee, 2014)).

In clinical trials applications, Connett and Lee (1990) introduced the estimation of the coefficient of variation from laboratory analysis of split specimens for quality control where an explicit statistical model was proposed for the coefficient of variation for laboratory analyses of constituents of blood, serum, saliva, or other specimens.

(Lehmann and Lehmann, (1986)) also assuming a random sample of size n of a normal distribution $N(\mu,\sigma^2)$, derived the sample distribution of the CV to get an exact method for the construction of a confidence interval for CV based on a non-central t-distribution for $\overline{X}\sqrt{n}/s$ with n-1 degrees of freedom and non-centrality parameter $\mu\sqrt{n}/\sigma$.

Other likelihood inference procedures were also proposed in the literature to construct confidence intervals for the CV (see Barndorff-Nielsen, (1986); Barndorff-Nielsen, (1991); Pierce and Peters, (1992); Reid, (1995)).

The coefficient of variation is also often used as a measure of precision and reproducibility of data in medical and biological sciences. (Tian, 2005) considers the problem of making inference about the common population coefficient of variation when it is a priori suspected that several independent samples are from populations with a common coefficient of variation.

Many other studies were also introduced in the literature. (Ahmed, (1995); Ahmed, (2002)) introduced a pooling methodology for the coefficient of variation. (Bennett, (1978)) introduced likelihood ratio (LR) tests for homogeneity of coefficients of variation in repeated samples. (Doornbos and Dijkstra, (1983)) introduced a multisample test for the equality of coefficients of variation in normal populations. (Forkman and Verrill, (2008)) derived the distribution of McKay's approximation for the coefficient of variation. (Fung and Tsang, (1998)) introduced a simulation study comparing tests for the equality of coefficients of variation. (Jafari and Kazemi, (2013)) considered a parametric bootstrap approach for the equality of coefficients of variation. (Jafari, (2015)) introduced inferences on the coefficients of variation in a multivariate normal population. (Gupta and Ma, (1996)) considered testing the equality of coefficients of variation in k normal populations. (Mahmoudvand and Hassani, (2009)) introduced two new confidence intervals for the coefficient of variation in a normal distribution. (Edward Miller, (1991)) studied asymptotic test statistics for coefficients of variation. (Subrahmanya Nairy and Aruna Rao, (2003)) derived tests of coefficient of variation of normal populations. (Pardo and Pardo, (2000)) used Renyi's divergence to test for the equality of the coefficient of variation. (Verrill, (2003)) derived confidence bounds for normal and lognormal distribution coefficients of variation.(Verrill and Johnson, (2007)) presented confidence bounds and hypothesis tests for normal distribution coefficients of variation. (Wong and Wu, (2002)) studied small sample asymptotic inference for the coefficient of variation considering normal and nonnormal models.

2. Inference under a classical approach

Let us assume a random sample of size n given by the vector $\mathbf{X}=(X_1,X_2,\cdots,X_n)$ of a normal $N(\mu,\sigma^2)$. The likelihood function for μ and σ is given by,

$$L(\mu,\sigma)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}\left(X_i-\mu\right)^2\right\} \tag{2}$$

That is,

$$L(\mu,\sigma) \propto \left(\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right\}$$

The logarithm of the likelihood function is given by,

$$l(\mu,\sigma) \propto n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(X_i-\mu\right)^2 \tag{3}$$

From the equations $\partial l/\partial\sigma = 0$ and $\partial l/\partial\mu = 0$, we get the maximum likelihood estimators (MLE) for μ and σ given respectively, by, $\widehat{\mu} = \sum_{j=1}^{n} X_j/n$ (a unbiased estimator) and $\widehat{\sigma}^2 = \sum_{j=1}^{n}\left(X_j-\overline{X}\right)^2$ (a biased estimator). The MLE estimator of σ is the positive square root of $\widehat{\sigma}^2$. From the invariance property of MLE we also get the MLE for the CV, $\theta = \sigma/\mu$ given by $\widehat{\theta} = \widehat{\sigma}/\widehat{\mu}$. It is important to point out that when the transformation of the parameters is one-to-one, the invariance property of maximum lilelihood estimators is a standard inference result. The invariance property is also extended to arbitrary transformations of the parameters (see Casella and Berger, (2021)).
The second derivatives of the log-likelihood function $l(\mu,\sigma)$ with respect to μ and σ are given by,

$$\begin{aligned}
\frac{\partial^2 l}{\partial\sigma^2} &= \frac{n}{\sigma^2} - \frac{3}{\sigma^4}\sum_{i=1}^{n}(X_i-\mu)^2 \\
\frac{\partial^2 l}{\partial\mu^2} &= -\frac{n}{\sigma^2} \\
\frac{\partial^2 l}{\partial\sigma\partial\mu} &= -\frac{3}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2
\end{aligned} \tag{4}$$

The Fisher information matrix $I(\mu,\sigma)$ is obtained from $E[-\partial^2 l/\partial\mu^2] = n/\sigma^2$, $E[-\partial^2 l/\partial\sigma^2] = 2n/\sigma^2$ and $E[-\partial^2 l/\partial\mu\partial\sigma] = 0$.That is,

$$I(\mu,\sigma) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{pmatrix} \tag{5}$$

The MLE $\widehat{\mu}$ and $\widehat{\sigma}$ have a asymptotic bivariate normal distribution for large values of n given by, $N(\boldsymbol{v}, I^{-1}(\widehat{\boldsymbol{v}}))$, where $\boldsymbol{v} = (\mu,\sigma^2)$, where, $I^{-1}(\boldsymbol{v})$ is given by,

$$I(\mu,\sigma) = \begin{pmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{\sigma^2}{2n} \end{pmatrix} \tag{6}$$

Considering that the coefficient of variation $CV = \sigma/\mu$ is the parameter of interest, we could consider a reparametrization $\theta = \sigma/\mu$ and $\tau = \sigma$. With these parameter trasformations, the likelihood function for θ and τ is given by,

$$L(\theta,\tau) \propto (\tau^2)^{-n/2} \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{n}\left(X_i-\frac{\tau}{\theta}\right)^2\right\} \tag{7}$$

Assuming this reparameterization, we could get the MLE for θ and τ, using numerical methods to solve the equations $\partial l/\partial\theta = 0$ and $\partial l/\partial\tau = 0$, and from the Fisher information matrix for θ and τ, we could get the asymptotic bivariate normal distribution for $\widehat{\theta}$ and $\widehat{\tau}$. From the asymptotic bivariate normal distribution for $\widehat{\theta}$ and $\widehat{\tau}$, we could obtain the

asymptotic univariate distribution for the MLE for the $CV = \theta = \sigma/\mu$ (a laborious solution) to obtain a sample distribution to be used to get confidence intervals and hypothesis tests for θ. Alternatively, we propose in this study, the use of the "delta method" to get the asymptotic univariate sample distribution for $\widehat{\theta} = \widehat{\sigma}/\widehat{\mu}$.

The delta method is a well known probability/statistics result concerning the approximate probability distribution for a function of an asymptotically normal statistical estimator from the knowledge of the limiting variance of that estimator (Doob, (1935); Ver Hoef, (2012)). The accuracy of the inference results could be not be the better compared to other existing approaches introduced in the literature but it is simple to be used in practical work as we show in this study.

2.1. Delta method for the bivariate case

Let us assume a random variable X with a probability distribution with density f(x; $\boldsymbol{\theta}$) where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ are the parameters of the distribution, and $\widehat{\theta_1}$ and $\widehat{\theta_2}$ are the MLE of θ_1 and θ_2. Thus, for large values of n, $\widehat{\boldsymbol{\theta}} = (\widehat{\theta_1}, \widehat{\theta_2})$ has an asymptotic bivariate normal distribution $N_2(\theta_1, \theta_2)$, $I^{-1}(\boldsymbol{\theta})$, where $I(\boldsymbol{\theta})$ is the Fisher information matrix.

Let us assume a function of θ_1 and θ_2, given by $g(\theta_1, \theta_2)$. Also assume that the usual standard regularity conditions are satisfied. Thus,

$$g(\widehat{\theta_1}, \widehat{\theta_2}) \sim N(g(\theta_1, \theta_2), \sigma_{12}^2) \tag{8}$$

where,

$$\sigma_{12}^2 = \sigma_1^2(\partial g/\partial\theta_1)^2 + 2\sigma_{12}(\partial g/\partial\theta_1)(\partial g/\partial\theta_2) + \sigma_2^2(\partial g/\partial\theta_2) \tag{9}$$

and the derivatives are calculated in the maximum likelihood estimators $(\widehat{\theta_1}, \widehat{\theta_2})$ and,

$$I(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \tag{10}$$

(See Lehmann and Lehmann, (1986)).

2.2. Hypothesis tests

Assuming that the data $(X_1, X_2, \cdots, X_n)$ is a random sample of a normal distribution $N(\mu, \sigma^2)$, we have the vector parameter $\boldsymbol{\theta} = (\theta_1, \theta_2)$ where $\theta_1 = \mu$ and $\theta_2 = \sigma$. For large sample sizes, the maximum likelihood estimator for θ has an asymptotic normal distribution $\widehat{\theta} \sim N(\theta, I^{-1}(\boldsymbol{\theta}))$, where $I^{-1}(\boldsymbol{\theta})$ is given by (6), that is, $\sigma_1^2 = \sigma_\mu^2 = \sigma^2/n, \sigma_2^2 = \sigma_\sigma^2 = \sigma^2/2n$ and $\sigma_{12} = 0$ in (10). From, (9) we have $\sigma_{\mu\sigma}^2 = \sigma_\mu^2(\partial g/\partial\mu)^2 + \sigma_\sigma^2(\partial g/\partial\sigma)^2$ where $g(\mu, \sigma) = \sigma/\mu$. The first partial derivatives of g with respect to μ and σ are given, respectively, by $\partial g/\partial\mu = -\sigma/\mu^2$ and $\partial g/\partial\sigma = 1/\mu$. Thus,

$$g(\widehat{\mu}, \widehat{\sigma}) = \frac{\widehat{\sigma}}{\widehat{\mu}} \sim N\left(\frac{\sigma}{\mu}, \sigma_{\mu\sigma}^2\right) \tag{11}$$

where $\sigma_{\mu\sigma}^2 = (\sigma^2/n)(-\sigma/\mu^2)^2 + (\sigma^2/2n)(1/\mu)^2 = (\sigma^2/n\mu^2)(\sigma^2/\mu^2 + 1/2)$.

Let us assume the simple hypothesis tests given by, $H_0 : \sigma/\mu = a_0$ versus the alternative hypothesis $H_1 : \sigma/\mu = a_1$, where a_0 and a_1 are known constants $(a_0 < a_1)$. The rejection region is given: reject $H_0 : \sigma/\mu = a_0$ if $\widehat{\sigma}/\widehat{\mu} > k$ where k is found by fixing a significance level α. That is,

$$\begin{aligned}\alpha &= P(\text{Type I error}) \\ &= P\left(\widehat{\sigma}/\widehat{\mu} \geq K \mid H_0 : \sigma/\mu = a_0\right) \\ &= \left(\frac{\frac{\widehat{\sigma}}{\widehat{\mu}} - a_0}{\sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]}}\right) \\ &= P(Z \geq z_\alpha)\end{aligned} \tag{12}$$

where

$$z_\alpha = \frac{k - a_0}{\sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]}} \tag{13}$$

Thus, we reject $H_0 : \sigma/\mu = a_0$ in a fixed significant level α, if $\widehat{\sigma}/\widehat{\mu} > k$, where,

$$k = a_0 + z_\alpha \sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]} \tag{14}$$

2.3. Sample size determination under a classical inference approach

In many studies (for example clinical trials), the sample sizes are small where at the beginning of an experiment, a small number n_1 of units are put to the test and with the information of this first stage, we are interested in determining the number of units n_2 to estimate the CV parameter with a fixed accuracy. In general, under the classical approach we use the power function or in the case of a simple null hypothesis versus a simple alternative hypothesis we fix the probabilities of type I error and type II error to estimate the sample size needed in a second stage. From (12), the asymptotic sampling distribution to test $H_0 : \sigma/\mu = a_0$ versus $H_1 : \sigma/\mu = a_1 (a_0 < a_1)$ is given by,

$$Z = \frac{\frac{\widehat{\sigma}}{\widehat{\mu}} - a_0}{\sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]}} \tag{15}$$

With the α significance level fixed (a small value), we have from (12) and (13):

$$Z_\alpha = \frac{k - a_0}{\sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]}} \tag{16}$$

where $P(Z \geq z_\alpha) = \alpha$ and Z has the standard normal distribution, $Z \sim N(0, 1)$.

Under the alternative hypothesis $H_1 : \sigma/\mu = a_1$, assuming a fixed value for β (a small value) for the probability of type II error, we have,

$$\begin{aligned}\beta &= P(\text{Type II error}) \\ &= P(\widehat{\sigma}/\widehat{\mu} < k \mid H_1 : \sigma/\mu = a_1) \\ &= \frac{\dfrac{\widehat{\sigma}}{\widehat{\mu}} - a_1}{\sqrt{\dfrac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\dfrac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \dfrac{1}{2}\right]}} > \frac{k-1}{\sqrt{\dfrac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\dfrac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \dfrac{1}{2}\right]}} \\ &= P(Z < -z_\beta)\end{aligned} \tag{17}$$

That is,

$$-z_\beta = \frac{k - a_1}{\sqrt{\dfrac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\dfrac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \dfrac{1}{2}\right]}} \tag{18}$$

From (16) and (18), we get (elimination of k) the equation,

$$a_0 + z_\alpha\sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]} = a_1 - z_\beta\sqrt{\frac{\widehat{\sigma}^2}{n\widehat{\mu}^2}\left[\frac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \frac{1}{2}\right]} \tag{19}$$

Solving the equation (19) for n, we get the needed sample size for the second stage,

$$n = \frac{\left(z_\alpha + z_\beta\right)^2 \widehat{\sigma}^2\left[\dfrac{\widehat{\sigma}^2}{\widehat{\mu}^2} + \dfrac{1}{2}\right]}{\left(a_1 - a_0\right)^2 \widehat{\mu}^2} \tag{20}$$

where z_α , z_β, a_0 and a_1 are known constants and $\widehat{\sigma}$ and $\widehat{\mu}$ are MLE of σ and $\widehat{\mu}$ obtained in the first stage.

3. Inference under a Bayesian approach

If the parameter of interest is the CV given by $\theta = \sigma/\mu$, assuming a random sample of size n of a normal distribution $N(\mu, \sigma^2)$, we assume a joint Jeffreys non-informative prior distribution (Jeffreys, (1946) Kass and Wasserman; (1996); Lee, (2012)) for the parameters μ and σ given by,

$$\pi(\mu, \sigma) \ \propto \ \sqrt{\det\left[I(\mu, \sigma)\right]} \tag{21}$$

where $I(\mu, \sigma)$ is the Fisher information matrix (4).

That is,

$$\pi(\mu, \sigma) \ \propto \ \frac{1}{\sigma^2} \tag{22}$$

for $-\infty < \mu < \infty$ and $\sigma > 0$.

The joint posterior distribution for μ and σ is obtained using the Bayes formula (Box and Tiao, (1973)) by combining the prior distribution 22) with the likelihood function (2), that is,

$$\pi(\mu,\sigma \mid \text{data}) \;\propto\; \sigma^{-(n+2)} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\} \tag{23}$$

for $-\infty < \mu < \infty$ and $\sigma > 0$.

The posterior summaries of interest are obtained using Markov Chain Monte Carlo (MCMC) simulation methods as the popular Gibbs sampling algorithm or the Metropolis-Hastings algorithm (Gelfand and Smith, (1990); Chib and Greenberg, (1995)) using the free existing OpenBUGS software (Lunn et al., (2000)). Since the OpenBUGS software only requires the likelihood function and the prior distributions for each parameter of the model, we do not present here all conditional posterior distributions $p(\theta_j/\boldsymbol{\theta}_{(j)}, \text{data})$, where $\boldsymbol{\theta}_{(j)}$ denotes the vector of all k parameters of the model except $\theta_j, j = 1, 2, \cdots, k$ needed for the Gibbs sampling or Metropolis-Hastings algorithms (see for example, Bernardo and Smith, (1994)).

In our case we get Monte Carlo estimates for the posterior means for μ and σ assuming a quadratic loss function. With the same generated Gibbs samples for μ and σ, we also get simultaneously Monte Carlo estimates for the CV given by $\theta = \sigma/\mu$ (point estimate and 95% credible interval for θ).

Observe that the joint posterior distribution for μ and σ (23) could be given by,

$$\pi(\mu,\sigma \mid \text{data}) \;\propto\; \sigma^{-(n+2)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n(\overline{X}-\mu)^2\right]\right\} \tag{24}$$

where $\overline{X} = \sum_{j=1}^{n} X_j/n$ and $s^2 = \sum_{j=1}^{n}(X_j - \overline{X})^2/(n-1)$ (sample mean and sample variance).

Thus the joint posterior distribution $\pi(\mu, \sigma \mid \text{data})$ is given by a product of an inverse gamma distribution $IG[(n-1)/2, (n-1)s^2/2]$ and a normal distribution $N(\overline{X}, \sigma^2/n)$.

3.1. Sample size determination under a Bayesian inference approach

Usually under a Bayesian approach we consider credible intervals to get the inferences of interest. In this way, we could obtain the joint posterior distribution for a transformation of variables given by $\theta = \sigma/\mu$ and $\tau = \sigma$ from the joint posterior distribution $\pi(\mu, \sigma \mid \text{data})$. From the joint posterior distribution $\pi(\theta, \tau \mid \text{data})$ we could get the marginal posterior distribution $\pi(\theta \mid \text{data})$ integrating out τ in the joint distribution $\pi(\theta, \tau \mid \text{data})$. The sample size in a second stage of the study is obtained considering a fixed length of the $(1-\alpha)\%$ credible interval for the CV parameter $\theta = \sigma/\mu$.

A great simplification is obtanined using MCMC methods. In this way, we could consider the joint posterior distribution for μ and $\sigma(23)$ as a prior distribution in a second stage of the experiment given by,

$$\begin{aligned} \mu \mid \sigma &\sim N(\overline{X}, \sigma^2/n) \\ \sigma &\sim IG[(n-1)/2, (n-1)s^2/2] \end{aligned} \tag{25}$$

The Bayesian procedure introduced in this study to estimate the sample size n_2 needed to have the lenght of a $(1-\alpha)\%$ credibility interval for θ, equal to a fixed value L, is given as follows:

- Simulate a random sample of fixed size n_2 from a normal distribution $N(b, c^2)$ where b is the Monte Carlo Bayesian estimate for the posterior mean for μ obtained in the first stage and c is the Monte Carlo Bayesian estimate for the posterior mean for σ obtained in the first stage.

- With the obtained simulated random sample for the second stage, assume as informative prior distributions for μ and σ, the posterior distributions given by (25). Use MCMC methods to get the posterior summaries of interest for the parameters μ, σ and $\theta = \sigma/\mu$.

- Calculate the length of the credibility interval for $\theta = \sigma/\mu$. If the length is close to the fixed L value, use the n_2 value as the estimated sample size needed for the second stage.

- If the obtained lenght of a $(1-\alpha)\%$ credible interval for θ obtained above is larger than L, simulate other random sample with a larger value for n_2 from the normal distribution $N(b, c^2)$, and repeat the above procedure. If the lenght of the $(1-\alpha)\%$ credible interval for θ is close to the fixed L value, use the assumed n_2 value as the estimated sample size needed for the second stage.

- Do this procedure until we get the lenght of a $(1-\alpha)\%$ credible interval for θ close (or small) to the fixed L value and use the final n_2 value as the sample size needed for the second stage.

4. Applications

4.1. A simulated data set with 50 observations

Table 1 shows 50 observations simulated from a normal distribution with mean $\mu = 2.8$ and standard deviation $\sigma = 1.8$ (sample mean $\overline{X} = 2.764$ and sample standard deviation $s = 1.801$).

Table 1: Simulated data (n=50)

4.95207	4.80491	1.23403	4.36515	3.30947	2.71657	3.11351	0.29848
-0.56406	1.09614	2.20693	4.42328	3.81848	1.30984	1.46244	3.31177
1.21853	1.83959	3.09909	2.19130	2.54555	6.41784	3.91336	4.58615
3.50832	4.10011	5.64794	0.98992	7.00364	1.83242	2.44093	4.83615
1.85789	0.89509	4.68334	5.52501	1.86927	-0.85424	2.74159	2.62888
3.21413	2.71115	1.71850	1.45538	-0.01146	3.63528	-0.31207	4.45551
1.78194	2.75154						

The MLE for μ and σ are given respectively by $\widehat{\mu} = 2.76359$ and $\widehat{\sigma} = 1.78247$. Let us assume the simple hypothesis tests given by, $H_0 : \sigma/\mu = 0.5$ versus the alternative hypothesis $H_1 : \sigma/\mu = 1$ assuming a significant level $\alpha = 0.5$, that is, $z_{0.5} = 1.645$. We reject $H_0 : \sigma/\mu = 0.5$, if $\widehat{\sigma}/\widehat{\mu} > k$, where, k is given in (14). That is, $k = 0.5 + 1.645(0.0872996) = 0.643608$. Since $\widehat{\sigma}/\widehat{\mu} = 1.78247/2.764 = 0.644984 > 0.643608$, we should reject $H_0 : \sigma/\mu = 0.5$.

- Determination of sample size under a frequentist approach

Under the null hypothesis, $H_0 : \sigma/\mu = 0.5$ we want to estimate the sample size under the alternative hypothesis $H_1 : \sigma/\mu = a_1$ for different values of a_1, assuming $\alpha = 0.5$ and $\beta = 0.025$, that is, $z_\alpha = z_{0.05} = 1.645$ and $z_\beta = z_{0.025} = 1.96$ (see section 2.3). From (20) we get the estimated sample size n for the second stage of the experiment. If $a_1 = 0.7$, we get from (20),

$$n = \frac{(1.645 + 1.96)^2 (1.78247)^2 \left[\left(\frac{1.78247}{2.764} \right)^2 + 0.5 \right]}{(0.7 - 0.5)^2 (2.764)^2} = 123.754$$

Thus, $n \approx 124$. Table 2 shows the estimated sample sizes for different values of a_1 in the alternative hypothesis.

Table 2: Sample sizes for different values of a_1

a_1	n
0.6	$495.0.14 \approx 495$
0.7	$123.754 \approx 124$
0.8	$55.016 \approx 55$
1	$19.8006 \approx 20$

- Determination of sample size under a Bayesian approach

Under a Bayesian approach we get the posterior summaries for the joint posterior distribution (23) considering the 50 simulated observations of a normal distribution with mean 2.7 and standard-deviation 1.8 given in Table 1, assuming a non-informative prior for the mean μ and the standard deviation σ from 1000 generated Gibbs samples (burn in of size 10000 samples discarded to eliminate the effect of the initial values in the iterative procedure and taking 1000 additional samples) using the OpenBUGS software. Table 3 shows the Monte Carlo estimates of the posterior means, posterior standard deviations and 95% credible interval for each parameter.

Table 3: Posterior summaries (first stage)

	mean	sd	Lower 95% ci	Upper 95% ci
$\boldsymbol{\mu}$	2.7530	0.2591	2.2710	3.2840
$\boldsymbol{\sigma}$	1.8240	0.1910	1.4920	2.2480
$\theta = \boldsymbol{\sigma}/\boldsymbol{\mu}$	0.6686	0.0969	0.5116	0.8987

From the results of Table 3, the length of the 95% credibility interval for the CV, denoted by θ is given by $L = 0.8987 - 0.5116 = 0.3871$.

To find the estimated sample size for a second stage with a fixed value for the length L of the 95% credibility interval for θ, we simulate from a normal distribution with mean 2.753 and standard-deviation 1.824 (information of the first stage) samples with different sample sizes until to get approximately the specified fixed length of interest.In the second stage we assume the informative prior distributions (25) for the parameters μ and σ (information of the first stage).

Let us assume that our objective is to find the sample size n_2 for an additional sample having a length L for the 95% credibility interval for θ given by $L = 0.2$.

Considering $n_2 = 50$ we generate 50 new observations from a normal distribution with mean 2.753 and standard-deviation 1.824 (information of the first stage). Assuming the prior distribution (25) for the parameters μ and σ and the same MCMC scheme used in the first stage using the OpenBUGS software, Table 4 shows the posterior summaries of interest.

Table 4: Posterior summaries (second stage with $n_2 = 50$)

	mean	sd	Lower 95% ci	Upper 95% ci
$\boldsymbol{\mu}$	2.8090	0.2391	2.3340	3.2430
$\boldsymbol{\sigma}$	1.7650	0.1766	1.4580	2.1580
$\theta = \boldsymbol{\sigma}/\boldsymbol{\mu}$	0.6334	0.0865	0.4938	0.8234

From the results of Table 4, the length of the 95% credibility interval for the CV, denoted by θ is given by $L = 0.8234 - 0.4938 = 0.3296$.We follow this approach until we get the value n_2 to have the length L for the 95% credibility interval for θ close to $L = 0.2$.
Table 5 shows the obtained lengths L for the 95% credibility interval for θ assuming different values of n_2.

Table 5: Lengths L for the 95% credibility interval for θ assuming different values of n_2

n_2	L
50	0.3296
120	0.2607
150	0.2567
200	0.1997

From the results of Table 5, we need $n_2 = 200$ additional observations to have a 95% credible interval forθ close to 0.2.

4.2. A simulated data set with 20 observations

Table 1 shows 20 observations simulated from a normal distribution with mean $\mu = 5.0$ and standard deviation $\sigma = 3$ (sample mean $\overline{X} = 5.457$ and sample standard deviation $s = 3.132$).

Table 6: Simulated data (n=20)

9.6457	4.2293	3.8995	7.7238	9.7525	6.5453
5.1468	0.3453	9.1116	6.3097	4.6842	-1.1669
6.4149	2.3411	4.2487	10.6894	5.0358	3.5913
2.8369	7.7601				

The MLE for μ and σ are given respectively by $\widehat{\mu} = 5.45724$ and $\widehat{\sigma} = 3.05226$. Considering the simple hypothesis tests given by, $H_0 : \theta = \sigma/\mu = 0.6$ versus the alternative hypothesis $H_1 : \sigma/\mu = 0.7$ assuming a significant level $\alpha = 0.5$, that is, $z_{0.5} = 1.645$. We reject $H_0 : \sigma/\mu = 0.5$, if $\widehat{\sigma}/\widehat{\mu} > k$, where k is given in (14). That is, $k = 0.6 + 1.645(0.181752) = 0.898983$. Since $\widehat{\sigma}/\widehat{\mu} = 3.05226/5.45724 = 0.559305 < k$, we do not reject $H_0 : \theta = 0.6$ assuming the significance level $\alpha = 0.5$.

- Determination of sample size under a frequentist approach

Under the null hypothesis, $H_0 : \sigma/\mu = 0.60$ we want to estimate the sample size under the alternative hypothesis $H_1 : \sigma/\mu = a_1$ for different values of a_1, assuming $\alpha = 0.5$ and $\beta = 0.025$, that is, $z_\alpha = z_{0.05} = 1.645$ and $z_\beta = z_{0.025} = 1.96$ (see section 2.3). From (20) we get the estimated sample size n for the second stage of the experiment. If $a_1 = 0.61$, we get from (20),

$$n = \frac{(1.645 + 1.96)^2(3.05226)^2\left[\left(\frac{3.05226}{5.45724}\right)^2 + 0.5\right]}{(0.61 - 0.60)^2(5.45724)^2} = 330.448$$

Thus, $n \approx 331$. Table 7 shows the estimated sample sizes for different values of a_1 in the alternative hypothesis.

Table 7: Sample sizes for different values of a_1

a_1	n
0.61	$330.448 \approx 331$
0.62	$165.224 \approx 165$
0.63	$110.149 \approx 110$
0.64	$82.6119 \approx 83$
0.65	$66.0895 \approx 66$
0.70	$33.0448 \approx 33$

- Determination of a sample size under a Bayesian approach

Under a Bayesian approach we get the posterior summaries for the joint posterior distribution (23) assuming a non-informative prior for the mean μ and the standard deviation σ from 1000 generated Gibbs samples (burn in of size 10000 samples discarded to eliminate the effect of the initial values in the iterative procedure and taking 1000 additional samples) using the OpenBUGS software. Table 8 shows the Monte Carlo estimates of the posterior means, posterior standard deviations and 95% credible interval for each parameter.

Table 8: Posterior summaries (first stage with $n_2 = 50$)

	mean	sd	Lower 95% ci	Upper 95% ci
$\boldsymbol{\mu}$	5.4520	0.7345	3.9680	6.9500
$\boldsymbol{\sigma}$	3.2560	0.5373	2.4310	4.4900
$\theta = \boldsymbol{\sigma}/\boldsymbol{\mu}$	0.6091	0.1390	0.4162	0.9694

From the results of Table 8, the length of the 95% credibility interval for the CV, denoted by θ is given by $L = 0.9694 - 0.4162 = 0.5532$. To find the estimated sample size for a second stage with a fixed value for the length L of the 95% credibility interval for θ, we simulate from a normal distribution with mean 5.452 and standard-deviation 3.256 (information of the first stage) samples with different sample sizes until to get approximately the specified fixed length of interest.In the second stage we assume the informative prior distributions (25) for the parameters μ and σ (information of the first stage).
Let us assume that our objective is to find the sample size n_2 for an additional sample have a length L for the 95% credible interval for θ given by $L = 0.2$.
Considering $n_2 = 50$, we generate 50 new observations from a normal distribution with mean 5.452 and standard-deviation 3.256. Assuming the prior distribution (25) for the parameters μ and σ and the same MCMC scheme used in the first stage using the OpenBUGS software, Table 9 shows the posterior summaries of interest.

Table 9: Posterior summaries (second stage with $n_2 = 50$)

	mean	sd	Lower 95% ci	Upper 95% ci
$\boldsymbol{\mu}$	2.8430	0.3171	2.2100	3.4860
$\boldsymbol{\sigma}$	2.2540	0.2668	1.7960	2.8400
$\theta = \boldsymbol{\sigma}/\boldsymbol{\mu}$	0.8028	0.1340	0.5928	1.1100

From the results of Table 9, the length of the 95% credibility interval for the CV, denoted by θ is given by $L = 1.1100 - 0.5928 = 0.5172$. We follow this approach until to get the value n_2 to have the length L for the 95% credibility interval for θ close to $L = 0.2$.

Table 10 shows the obtained lengths L for the 95% credibility interval for θ assuming different values of n_2.

Table 10: Lengths L for the 95% credibility interval for θ (CV) assuming different values of n_2

n_2	L
50	0.5172
80	0.2546
100	0.2261
110	0.1996

From the results of Table 10 we observe that it is needed $n_2 = 110$ additional observations to have a 95% credible interval for θ close to 0.2.

5. Conclusions

This study presented simple inference formulations to get classical or Bayesian inference results for the coefficient of variation assuming data from a normal distribution with mean μ and standard deviation σ. We observe from a review of the literature on this subject, that there is a very large number of studies to find sample sizes in a second stage of an experiment assuming as the parameter of interest, the coefficient of variation $\theta = \sigma/\mu$. Many of these studies have as their main goal, theoretical formulations to get better accurate inference results for the coefficient of variation (CV). Sometimes the proposed methodology presented in the literature is not simple to be use in applications. In our study, the main goal was to introduce simple sample size formulas to be used in applications.

In the first part of the study, we presented simple formulations to get the inferences of interest for the coefficient of variation under normality for the data and using a standard frequentist approach based on the asymptotic normality of the maximum likelihood estimators for the mean μ and standard deviation σ of the normal distribution and using the delta method to get the inferences of interest for the coefficient of variation. Simple hypothesis tests and determination of the sample size are discussed under the frequentist approach. These results also could be generalized for compound hypotheses.

In the second part of the study we introduced a procedure for sample size determination under a Bayesian approach, where it is assumed a Jeffreys non-informative prior distribution of the parameters of the normal distribution assumed for the data and using standard Markov Chain Monte Carlo (MCMC) methods to get the Monte Carlo estimates for the coefficient of variation $\theta = \sigma/\mu$ in a first stage of a experiment. Assuming the obtained posterior distribution in the first stage as an informative prior in the second stage of an experiment we proposed a simple way to get a sample size to get a fixed accuracy (fixed length of a credible interval for the parameter $\theta = \sigma/\mu$) of the posterior distribution. The use of MCMC methods under a Bayesian approach leads to great simplification in obtaining Monte Carlo estimators for the parameters of the data distribution and also for functions of the original parameters, such as the coefficient of variation. Thus, point estimators or credible interval estimators with a specified probability are obtained in a simple way without the need to use a sampling distribution of the estimator of $\theta = \sigma/\mu$ (CV). Thus, the sample size determination under a Bayesian approach to have a fixed length for de credible interval for the CV is obtained in a trivial way as observed in this study. These results could be of great interest in different areas of application when the parameter of interest is the coefficient of variation.

References

1. Ahmed, S. (1995). A pooling methodology for coefficient of variation. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 57–75.

2. Ahmed, S. (2002). Simultaneous estimation of coefficients of variation. *Journal of Statistical Planning and Inference*, 104(1):31–51.
3. Barndorff-Nielsen, O. E. (1986). Infereni on full or partial parameters based on the standardized signed log likelihood ratio. *Biometrika*, 73(2):307–322.
4. Barndorff-Nielsen, O. E. (1991). Modified signed log likelihood ratio. *Biometrika*, 78(3):557–563.
5. Bennett, B. (1976). On an approximate test for homogeneity of coefficients of variation. In *Contribution to applied statistics*, pages 169–171. Springer.
6. Bennett, B. (1978). Lr tests for homogeneity of coefficients of variation in repeated samples. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 400–405.
7. Bernardo, J. and Smith, A. (1994). Bayesian theory wiley. *New York*, 49.
8. Box, G. E. and Tiao, G. C. (1973). Bayesian inference in statistical analysis, addision-wesley. *Reading, MA*.
9. Casella, G. and Berger, R. L. (2021). *Statistical inference*. Cengage Learning.
10. Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335.
11. Connett, J. E. and Lee, W. W. (1990). Estimation of the coefficient of variation from laboratory analysis of split specimens for quality control in clinical trials. *controlled Clinical trials*, 11(1):24–36.
12. Doob, J. L. (1935). The limiting distributions of certain statistics. *The Annals of Mathematical Statistics*, 6(3):160–169.
13. Doornbos, R. and Dijkstra, J. (1983). A multi sample test for the equality of coefficients of variation in normal populations. *Communications in Statistics-Simulation and Computation*, 12(2):147–158.
14. Edward Miller, G. (1991). Asymptotic test statistics for coefficients of variation. *Communications in Statistics-Theory and Methods*, 20(10):3351–3363.
15. Feltz, C. J. and Miller, G. E. (1996). An asymptotic test for the equality of coefficients of variation from k populations. *Statistics in medicine*, 15(6):647–658.
16. Forkman, J. (2009). Estimator and tests for common coefficients of variation in normal distributions. *Communications in Statistics—Theory and Methods*, 38(2):233–251.
17. Forkman, J. and Verrill, S. (2008). The distribution of mckay's approximation for the coefficient of variation. *Statistics & probability letters*, 78(1):10–14.
18. Fung, W. and Tsang, T. (1998). A simulation study comparing tests for the equality of coefficients of variation. *Statistics in Medicine*, 17(17):2003–2014.
19. Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, 85(410):398–409.
20. Gupta, R. C. and Ma, S. (1996). Testing the equality of coefficients of variation in k normal populations. *Communications in Statistics-Theory and Methods*, 25(1):115–132.
21. Hendricks, W. A. and Robey, K. W. (1936). The sampling distribution of the coefficient of variation. *The Annals of Mathematical Statistics*, 7(3):129–132.
22. Iglewicz, B. and Myers, R. H. (1970). Comparisons of approximations to the percentage points of the sample coeffcient of variation. *Technometrics*, 12(1):166–169.
23. Jafari, A. A. (2015). Inferences on the coefficients of variation in a multivariate normal population. *Communications in Statistics-Theory and Methods*, 44(12):2630–2643.
24. Jafari, A. A. and Kazemi, M. R. (2013). A parametric bootstrap approach for the equality of coefficients of variation. *Computational Statistics*, 28(6):2621–2639.
25. Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461.
26. Kass, R. E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American statistical Association*, 91(435):1343–1370.
27. Krishnamoorthy, K. and Lee, M. (2014). Improved tests for the equality of normal coefficients of variation. *Computational Statistics*, 29(1):215–232.
28. Lee, P. M. (2012). *Bayesian Statistics: An Introdution*. John Wiley & Sons.
29. Lehmann, E. L. and Lehmann, E. (1986). *Testing statistical hypotheses*, volume 2. Springer.
30. Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). Winbugs-a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and computing*, 10(4):325–337.
31. Mahmoudvand, R. and Hassani, H. (2009). Two new confidence intervals for the coefficient of variation in a normal distribution. *Journal of applied statistics*, 36(4):429–442.

32. Pardo, M. and Pardo, J. (2000). Use of rényi's divergence to test for the equality of the coefficients of variation. *Journal of computational and applied mathematics*, 116(1):93–104.
33. Pierce, D. A. and Peters, D. (1992). Practical use of higher order asymptotics for multiparameter exponential families. *Journal of the Royal Statistical Society: Series B (Methodological)*, 54(3):701–725.
34. Reid, N. (1995). The roles of conditioning in inference. *Statistical Science*, 10(2):138–157.
35. Sokal, R. and Rohlf, F. (1995). Biometry. 3rd ednew york wh freeman and company.
36. Subrahmanya Nairy, K. and Aruna Rao, K. (2003). Tests of coefficients of variation of normal population. *Communications in Statistics-Simulation and Computation*, 32(3):641–661.
37. Tian, L. (2005). Inferences on the common coefficient of variation. *Statistics in medicine*, 24(14):2213–2220.
38. Vangel, M. G. (1996). Confidence intervals for a normal coefficient of variation. *The American Statistician*, 50(1):21–26.
39. Ver Hoef, J. M. (2012). Who invented the delta method? *The American Statistician*, 66(2):124–127.
40. Verrill, S. (2003). Confidence bounds for normal and lognormal distribution coefficients of variation. *Res. Pap. FPL-RP-609. Madison, WI: US Department of Agriculture, Forest Service, Forest Products Laboratory: 13 pages*, 609.
41. Verrill, S. and Johnson, R. A. (2007). Confidence bounds and hypothesis tests for normal distribution coefficients of variation. *Communications in Statistics—Theory and Methods*, 36(12):2187–2206.
42. Wong, A. and Wu, J. (2002). Small sample asymptotic inference for the coefficient of variation: normal and nonnormal models. *Journal of Statistical Planning and Inference*, 104(1):73–82.