cvcqv: A Package for Estimation of Relative Variability

by Maani Beigy

Abstract Coefficient of variation (*cv*) and coefficient of quartile variation (*cqv*) are widely used measures of relative dispersion which play descriptive and inferential roles (*e.g.,* reliability analysis, quality control, inequality measurement, and anomaly detection) in various fields such as biological and medical sciences, economics, actuarial sciences, etc. Since *cv* and *cqv* are unit-free, they are useful for comparing data from different distributions, data from different scales, or widely different means. However, to avoid their common misuses, confidence intervals (*CI*) are required. The **cvcqv** package provides a home for such tools. To our knowledge, the new `R` package **cvcqv** is the first `R` implementation of **cqv** as a robust variability measure, with almost all available methods for *CI* of *cv* and *cqv*. This paper elucidates this versatile functionality using reproducible examples on real datasets. Also, the new insights that **cvcqv**, alongside other `R` packages, brings into data science will be discussed.

Introduction

Researchers and practitioners in various fields use the coefficient of variation (*cv*) as a measure of relative variability (Panichkitkosolkul, 2013; Payton, 1996). *cv* is calculated as the ratio of the sample standard deviation (*sd*) to the sample mean ($\bar{x}$). However, *cv* is often misleading for variables with non-ratio scales (Payton, 1996), for homoscedastic data, and for variables without different magnitudes or units (Shechtman, 2013).

Robust statistical measurements such as coefficient of quartile variation (*cqv*) are better alternatives in non-normal distributions (Altunkaynak and Gamgam, 2018):

$$cqv = \left(\frac{q_3 - q_1}{q_3 + q_1}\right) \times 100$$

where q_3 and q_1 are the sample third quartile (*i.e.,* 75^{th} percentile) and first quartile (*i.e.,* 25^{th} percentile), respectively.

Almost always, we calculate *cv* and *cqv* from samples but the final objective is to generalize them as the populations' parameters (Albatineh et al., 2014). For example, one may be interested in comparing the variabilities of the time-varying measurements of a variable to detect anomalies such as extreme behaviors of customers or institutes (as in actuarial sciences). Or someone might inquire into whether a laboratory test or technique has sufficient inter-assay and intra-assay reliablity (Panichkitkosolkul, 2013; Payton, 1996). In such scenarios, variabilities calculated from samples are often biased and misleading (Sørensen, 2002; Payton, 1996). Therefore, various confidence intervals (*CI*) have been introduced to correctly estimate the relative variability.

This paper sets out to demonstrate the versatility of **cvcqv** package in a variety of data science tasks related to variability measurement. `R` (R Core Team, 2016) provides a strong asset for progress in this direction because it already contains functionality used in a variety of packages like **DescTools**, **MBESS**, **goeveg**, and **sjstats**. However, robust variability measures such as *cqv* has been missing in `R` for a long time. Moreover, the implementations of *CI* for *cv* have been limited to one or two methods. Lack of functions for the rigorous methods of calculation of *CI* for *cv* and *cqv*, though available in the statistical literature, was a major motivation to develop this package and explain its versatile functionality in this paper.

Package structure and functionality

The package can be installed and loaded as follows (see the package's README for dependencies and access to development versions):

```
install.packages("cvcqv")
```

```
library(cvcqv)
```

cvcqv depends on **dplyr** (Wickham et al., 2019) for using `nth()` function and imports **R6** (Chang, 2019) for `"R6"` classes, **SciViews** (Grosjean, 2018) for `ln()` function, **boot** (Canty and Ripley, 2019) for bootstrapping methods, and **MBESS** (Kelley, 2018) for noncentral distributions.

Core functions and classes

The functionality of the package is developed as both simple functions and `"R6"` classes, for sake of versatility, portability and efficiency:

- The R6 class `"SampleQuantiles"` to produce the sample quantiles corresponding to the given probabilities. It uses quantile function from the built-in `R` package **stats**, but provides an `"R6"` interface to be inherited for other classes.
- The R6 class `"BootCoefVar"` produces the bootstrap resampling for the *cv*. It uses boot.ci function from **boot**, but provides an `"R6"` interface to be inherited for child classes.
- The R6 class `"BootCoefQuartVar"` produces the bootstrap resampling for the *cqv*. It uses boot and boot.ci functions from **boot**, but provides an `"R6"` interface to be inherited for child classes.
- The R6 class `"CoefVar"` calculates the sample *cv*.
- The R6 class `"CoefQuartVar"` calculates the sample *cqv*.
- The R6 class `"CoefVarCI"` calculates *CI* for *cv*.
- The R6 class `"CoefQuartVarCI"` calculates *CI* for *cqv*.
- The function `cv_versatile` calculates *cv* and its various *CIs*.
- The function `cqv_versatile` calculates *cqv* and its various *CIs*.

R6 Objects Tree

```
SampleQuantiles
        ├──── BootCoefVar
        │       └────────── CoefVar
        │                      └────────── CoefVarCI
        │
        └──── BootCoefQuartVar
                └────────── CoefQuartVar
                               └────────── CoefQuartVarCI
```

Confidence Interval Methods

There are various methods for the calculation of *CI* for *cv* and *cqv*, which have been implemented in **cvcqv** package:

Table 1: Methods for calculation of *CI* for *cv* and *cqv*

cv	cqv
"kelley" (2018; 2007)	"bonett" (2006)
"mckay" (1932)	"norm" (2018)
"miller" (1991)	"basic" (2018)
"vangel" (1996)	"perc" (2018)
"mahmoudvand_hassani" (2009)	"bca" (2018)
"equal_tailed" (2013)	
"shortest_length" (2013)	
"normal_approximation" (2013)	
"norm" (2019; 1997)	
"basic" (2019; 1997)	
"perc" (2019; 1997)	
"bca" (2019; 1997)	

Solutions for real-world problems

This section contains examples on real-world data science problems:

Using cvcqv in Anomaly Detection

Bibliography

A. N. Albatineh, B. M. Kibria, M. L. Wilcox, and B. Zogheib. Confidence interval estimation for the population coefficient of variation using ranked set sampling: A simulation study. 41(4):733–751, 2014. ISSN 02664763. URL https://doi.org/10.1080/02664763.2013.847405. [p1]

B. Altunkaynak and H. Gamgam. Bootstrap confidence intervals for the coefficient of quartile variation. 0(0):1–9, 2018. ISSN 15324141. doi: 10.1080/03610918.2018.1435800. URL https://doi.org/10.1080/03610918.2018.1435800. [p1, 2]

D. G. Bonett. Confidence interval for a coefficient of quartile variation. 50(11):2953–2957, 2006. ISSN 01679473. URL https://doi.org/10.1016/j.csda.2005.05.007. [p2]

A. Canty and B. D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2019. R package version 1.3-22. [p1, 2]

W. Chang. *R6: Encapsulated Classes with Reference Semantics*, 2019. URL https://CRAN.R-project.org/package=R6. R package version 2.4.0. [p1]

A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Applications*. Cambridge University Press, Cambridge, 1997. URL https://doi.org/10.1017/CBO9780511802843. ISBN 0-521-57391-2. [p2]

G. Edward Miller. Asymptotic test statistics for coefficients of variation. 20(10):3351–3363, 1991. ISSN 0361-0926. [p2]

P. Grosjean. *SciViews-R*. UMONS, MONS, Belgium, 2018. URL http://www.sciviews.org/SciViews-R. [p1]

K. Kelley. Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. 39(4):755–766, 2007. ISSN 1554351X. URL https://doi.org/10.3758/BF03192966. [p2]

K. Kelley. *MBESS: The MBESS R Package*, 2018. URL https://CRAN.R-project.org/package=MBESS. R package version 4.4.3. [p1, 2]

R. Mahmoudvand and H. Hassani. Two new confidence intervals for the coefficient of variation in a normal distribution. 36(4):429–442, 2009. ISSN 0266-4763. URL https://doi.org/10.1080/02664760802474249. [p2]

A. T. McKay. Distribution of the Coefficient of Variation and the Extended" t" Distribution. 95(4):695–698, 1932. ISSN 0952-8385. [p2]

W. Panichkitkosolkul. Confidence Intervals for the Coefficient of Variation in a Normal Distribution with a Known Population Mean. 2013:1–11, 2013. ISSN 1687-952X. URL https://doi.org/10.1155/2013/324940. [p1, 2]

M. E. Payton. Confidence Intervals for the Coefficient of Variation. pages 82–90, 1996. URL https://doi.org/10.4148/2475-7772.1320. [p1]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016. URL https://www.R-project.org/. ISBN 3-900051-07-0. [p1]

O. Shechtman. *The Coefficient of Variation as an Index of Measurement Reliability*, pages 39–49. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-37131-8. URL https://doi.org/10.1007/978-3-642-37131-8_4. [p1]

J. B. Sørensen. The use and misuse of the coefficient of variation in organizational demography research. *Sociological methods & research*, 30(4):475–491, 2002. URL https://doi.org/10.1177/0049124102030004001. [p1]

M. G. Vangel. Confidence intervals for a normal coefficient of variation. 50(1):21–26, 1996. ISSN 0003-1305. [p2]

H. Wickham, R. François, L. Henry, and K. Müller. *dplyr: A Grammar of Data Manipulation*, 2019. URL https://CRAN.R-project.org/package=dplyr. R package version 0.8.3. [p1]

Maani Beigy
Department of Epidemiology and Biostatistics
School of Public Health
Tehran University of Medical Sciences
Tehran
Iran
ORCiD: 0000-0003-2963-3533
manibeygi@gmail.com
m-beigy@alumnus.tums.ac.ir