Coefficient of variation

From Wikipedia, the free encyclopedia

In probability theory and statistics, the **coefficient of variation** (**CV**) is a standardized measure of dispersion of a probability distribution or frequency distribution. It is defined as the ratio of the standard deviation σ to the mean μ. It is also known as **unitized risk**[1] or the **variation coefficient.** The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a percentage.

Contents

- 1 Definition
- 2 Estimation
  - 2.1 Log-normal data
- 3 Comparison to standard deviation
  - 3.1 Advantages
  - 3.2 Disadvantages
- 4 Applications
  - 4.1 Laboratory measures of intra and inter-assay CVs
- 5 Distribution
  - 5.1 Alternative
- 6 Similar ratios
- 7 See also
- 8 References

Definition

The coefficient of variation (CV) is defined as the ratio of the standard deviation σ to the mean μ:

$$c_v = \frac{\sigma}{\mu}$$

It shows the extent of variability in relation to mean of the population.

The coefficient of variation should be computed only for data measured on a ratio scale, as these are measurements that can only take non-negative values. The coefficient of variation may not have any meaning for data on an interval scale.[2] For example, most temperature scales are interval scales (e.g., Celsius, Fahrenheit etc.) that can take both positive and negative values, whereas the Kelvin scale has an absolute null value (i.e., will work in the complete absence of thermal energy), and negative values are nonsensical. Hence, the Kelvin scale is a ratio scale. While the standard deviation (SD) can be derived on both the Kelvin and the Celsius scale (with both leading to the same SDs), the CV is only relevant as a measure of relative variability for the Kelvin scale.

Measurements that are log-normally distributed exhibit stationary CV; in contrast, SD would vary depending on the expected value of measurements. This is the case for laboratory values that are measured based on chromatographic methods.

A nonparametric possibility is the quartile coefficient of dispersion, i.e. interquartile range $Q_3 - Q_1$ divided by the median Q_2.

Estimation

When only a sample of data from a population is available, the population CV can be estimated using the ratio of the sample standard deviation s to the sample mean $\bar{x}$:

$$\widehat{c_v} = \frac{s}{\bar{x}}$$

But this estimator, when applied to a small or moderately sized sample, tends to be too low: it is a biased estimator. For normally distributed data, an unbiased estimator[3] for a sample of size n is:

$$\widehat{c_v}^* = \left(1 + \frac{1}{4n}\right)\widehat{c_v}$$

Log-normal data

In many applications, it can be assumed that data are log-normally distributed (evidenced by the presence of skewness in the sampled data).[4] In such cases, a more accurate estimate, derived from the properties of the log-normal distribution,[5][6][7] is defined as:

$$\widehat{c_v}_{ln} = \sqrt{e^{s_{ln}{}^2} - 1}$$

where s_{ln} is the sample standard deviation of the data after a natural log transformation. (In the event that measurements are recorded using any other logarithmic base, b, their standard deviation s_b is converted to base e using $s_{ln} = s_b ln(b)$, and the formula for $\widehat{c_v}_{ln}$ remains the same.[8]) This estimate is sometimes referred to as the "geometric CV"[9][10] in order to distinguish it from the simple estimate above. However, "geometric coefficient of variation" has also been defined by Kirkwood[11] as:

$$GCV_K = e^{s_{ln}} - 1$$

This term was intended to be *analogous* to the coefficient of variation, for describing multiplicative variation in log-normal data, but this definition of GCV has no theoretical basis as an estimate of c_v itself.

For many practical purposes (such as sample size determination and calculation of confidence intervals) it is s_{ln} which is of most use in the context of log-normally distributed data. If necessary, this can be derived from an estimate of c_v or GCV by inverting the corresponding formula.

Comparison to standard deviation

Advantages

The coefficient of variation is useful because the standard deviation of data must always be understood in the context of the mean of the data. In contrast, the actual value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless number. For comparison between data sets with different units or widely different means, one should use the coefficient of variation instead of the standard deviation.

Disadvantages

- When the mean value is close to zero, the coefficient of variation will approach infinity and is therefore sensitive to small changes in the mean. This is often the case if the values do not originate from a ratio scale.
- Unlike the standard deviation, it cannot be used directly to construct confidence intervals for the mean.

Applications

The coefficient of variation is also common in applied probability fields such as renewal theory, queueing theory, and reliability theory. In these fields, the exponential distribution is often more important than the normal distribution. The standard deviation of an exponential distribution is equal to its mean, so its coefficient of variation is equal to 1.

Distributions with CV < 1 (such as an Erlang distribution) are considered low-variance, while those with CV > 1 (such as a hyper-exponential distribution) are considered high-variance. Some formulas in these fields are expressed using the **squared coefficient of variation**, often abbreviated SCV. In modeling, a variation of the CV is the CV(RMSD). Essentially the CV(RMSD) replaces the standard deviation term with the Root Mean Square Deviation (RMSD). While many natural processes indeed show a correlation between the average value and the amount of variation around it, accurate sensor devices need to be designed in such a way that the coefficient of variation is close to zero, i.e., yielding a constant absolute error over their working range.

Laboratory measures of intra and inter-assay CVs

CV measures are often used as quality controls for quantitative laboratory assays. While intra-assay and inter-assay CVs might be assumed to be calculated by simply averaging CV values across CV values for multiple samples within one assay or by averaging multiple inter-assay CV estimates, it has been suggested that these practices are incorrect and that a more complex computational process is required.[12]

Distribution

Provided that negative and small positive values of the sample mean occur with negligible frequency, the probability distribution of the coefficient of variation for a sample of size *n* has been shown by Hendricks and Robey[13] to be

$$dF_{c_v} = \frac{2}{\pi^{1/2}\Gamma\left(\frac{n-1}{2}\right)} e^{-\frac{n}{2\left(\frac{\sigma}{\mu}\right)^2}\frac{{c_v}^2}{1+{c_v}^2}} \frac{{c_v}^{n-2}}{(1+{c_v}^2)^{n/2}} \sum_{i=0}^{n-1}{}' \frac{(n-1)!\Gamma\left(\frac{n-i}{2}\right)}{(n-1-i)!i!} \frac{n^{i/2}}{2^{i/2}\left(\frac{\sigma}{\mu}\right)^i} \frac{1}{(1+{c_v}^2)^{i/2}} dc_v,$$

where the symbol $\sum{}'$ indicates that the summation is over only even values of *n*-1-*i*, i.e., if *n* is odd, sum over even values of *i* and if *n* is even, sum only over odd values of *i*.

This is useful, for instance, in the construction of hypothesis tests or confidence intervals. 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 [14][15][16][17]

Alternative

According to Liu (2012),[18] Lehmann (1986)[19] "also derived the sample distribution of CV in order to give an exact method for the construction of a confidence interval for CV;" it is based on a non-central t-distribution.

Similar ratios

Standardized moments are similar ratios, μ_k/σ^k where μ_k is the k^{th} moment about the mean, which are also dimensionless and scale invariant. The variance-to-mean ratio, σ^2/μ, is another similar ratio, but is not dimensionless, and hence not scale invariant. See Normalization (statistics) for further ratios.

In signal processing, particularly image processing, the reciprocal ratio μ/σ is referred to as the signal to noise ratio.

- Relative standard deviation, $|\sigma/\mu|$
- Efficiency, σ^2/μ^2
- Standardized moment, μ_k/σ^k
- Variance-to-mean ratio, σ^2/μ
- Fano factor, σ_W^2/μ_W (windowed VMR)
- Signal-to-noise ratio, μ/σ (in signal processing)
  - Signal-to-noise ratio (imaging)

See also

- Omega ratio
- Sampling (statistics)

References

1. ^ Broverman, Samuel A. (2001). *Actex study manual, Course 1, Examination of the Society of Actuaries, Exam 1 of the Casualty Actuarial Society* (http://books.google.de/books?id=qd4PAQAAMAAJ&q=%22unitized+risk%22&dq=%22unitized+risk%22&hl=en&sa=X&ei=tWKTU4byO8W2O_urgegN&ved=0CCgQ6AEwAA) (2001 ed. ed.). Winsted, CT: Actex Publications. p. 104. ISBN 9781566983969. Retrieved 7 June 2014.
2. ^ "What is the difference between ordinal, interval and ratio variables? Why should I care?" (http://www.graphpad.com/faq/viewfaq.cfm?faq=1089). GraphPad Software Inc. Retrieved 2008-02-22.
3. ^ Sokal RR & Rohlf FJ. *Biometry* (3rd Ed). New York: Freeman, 1995. p. 58. ISBN 0-7167-2411-1
4. ^ Limpert, Eckhard; Stahel, Werner A.; Abbt, Markus (2001). "Log-normal Distributions across the Sciences: Keys and Clues". *BioScience* **51**: 341–352. doi:10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2 (http://dx.doi.org/10.1641%2F0006-3568%282001%29051%5B0341%3ALNDATS%5D2.0.CO%3B2).
5. ^ Koopmans, L. H.; Owen, D. B.; Rosenblatt, J. I. (1964). "Confidence intervals for the coefficient of variation for the normal and log normal distributions". *Biometrika* **51**: 25. doi:10.1093/biomet/51.1-2.25 (http://dx.doi.org/10.1093%2Fbiomet%2F51.1-2.25).
6. ^ Diletti, E; Hauschke, D; Steinijans, VW (1992). "Sample size determination for bioequivalence assessment by means of confidence intervals". *International journal of clinical pharmacology, therapy, and toxicology*. 30 Suppl 1: S51–8. PMID 1601532 (https://www.ncbi.nlm.nih.gov/pubmed/1601532).
7. ^ Julious, Steven A.; Debarnot, Camille A. M. (2000). "Why Are Pharmacokinetic Data Summarized by Arithmetic Means?". *Journal of Biopharmaceutical Statistics* **10** (1): 55–71. doi:10.1081/BIP-100101013 (http://dx.doi.org/10.1081%2FBIP-100101013). PMID 10709801 (https://www.ncbi.nlm.nih.gov/pubmed/10709801).
8. ^ Reed JF, Lynn F, Meade BD. (2002) "Use of Coefficient of Variation in Assessing Variability of Quantitative Assays". *Clin Diagn Lab Immunol.* 9(6): 1235–1239. doi:10.1128/CDLI.9.6.1235-1239.2002 (http://dx.doi.org/10.1128%2FCDLI.9.6.1235-1239.2002)
9. ^ Sawant,S.; Mohan, N. (2011) "FAQ: Issues with Efficacy Analysis of Clinical Trial Data Using SAS" (http://pharmasug.org/proceedings/2011/PO/PharmaSUG-2011-PO08.pdf), *PharmaSUG2011*, Paper PO08
10. ^ Schiff MH, et al. Ann Rheum Dis 2014;0:1–3. doi:10.1136/annrheumdis-2014-205228 (http://dx.doi.org/10.1136%2Fannrheumdis-2014-205228)
11. ^ Kirkwood, TBL (1979). "Geometric means and measures of dispersion". *Biometrics* **35**: 908–9. doi:10.2307/2530139 (http://dx.doi.org/10.2307%2F2530139).
12. ^ Rodbard, D (October 1974). "Statistical quality control and routine data processing for radioimmunoassays and immunoradiometric assays.". *Clinical chemistry* **20** (10): 1255–70. PMID 4370388 (https://www.ncbi.nlm.nih.gov/pubmed/4370388).
13. ^ Hendricks, Walter A.; Robey, Kate W. (1936). "The Sampling Distribution of the Coefficient of Variation". *The Annals of Mathematical Statistics* **7** (3): 129–32. doi:10.1214/aoms/1177732503 (http://dx.doi.org/10.1214%2Faoms%2F1177732503). JSTOR 2957564 (https://www.jstor.org/stable/2957564).
14. ^ Iglevicz, Boris; Myers, Raymond (1970). "Comparisons of approximations to the percentage points of the sample coefficient of variation". *Technometrics* **12** (1): 166–169. JSTOR 1267363 (https://www.jstor.org/stable/1267363).
15. ^ Bennett, B. M. (1976). On an approximate test for homogeneity of coefficients of variation. In: Ziegler, W. J. (ed.) *Contributions to Applied Statistics Dedicated to A. Linder. Experentia Suppl.* 22, 169-171
16. ^ Vangel, Mark G. (1996). "Confidence intervals for a normal coefficient of variation". *The American Statistician* **50** (1): 21–26. doi:10.1080/00031305.1996.10473537 (http://dx.doi.org/10.1080%2F00031305.1996.10473537). JSTOR 2685039 (https://www.jstor.org/stable/2685039)..
17. ^ Forkman, Johannes. "Estimator and tests for common coefficients of variation in normal distributions" (http://pub.epsilon.slu.se/4489/1/forkman_j_110214.pdf). *Communications in Statistics - Theory and Methods* **38**. pp. 21–26.

Retrieved 2013-09-23.

18. ^ [1] (http://scholarworks.gsu.edu/cgi/viewcontent.cgi?article=1116&context=math_theses), p.3
19. ^ Lehmann, E. L. (1986), Testing Statistical Hypothesis. 2nd ed. New York: Wiley.

Categories: Theory of probability distributions | Statistical deviation and dispersion | Statistical ratios

---