Coefficient of variation
From Wikipedia, the free encyclopedia

In probability theory and statistics, the coefficient of variation (CV) is a normalized measure of dispersion of a probability distribution. It is also known as unitized risk or the variation coefficient. The absolute value of the CV is sometimes known as relative standard deviation (RSD), which is expressed as a percentage.

Definition

The coefficient of variation (CV) is defined as the ratio of the standard deviation $\sigma$ to the mean $\mu$:

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

which is the inverse of one definition of the signal-to-noise ratio. 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. For example, most temperature scales are interval scales (e.g., Celsius, Fahrenheit etc.) that can take both positive and negative values. The Kelvin scale has an absolute null value, and no negative values can naturally occur. 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.

Often, laboratory values that are measured based on chromatographic methods are log-normally distributed. In this case, the CV would be constant over a large range of measurements, while SDs would vary depending on typical values that are being measured.

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 $\bar{x}$:
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\(^2\) for a sample of size \(n\) is:

\[
\hat{C}_v^* = \left(1 + \frac{1}{4n}\right)\hat{C}_v
\]

In many applications, it can be assumed that data are log-normally distributed (evidenced by the presence of skewness in the sampled data).\(^3\) In such cases, a more accurate estimate, derived from the properties of the log-normal distribution,\(^4\)[5][6] is defined as:

\[
\hat{C}_{vln} = \sqrt{e^{s_{ln}^2} - 1}
\]

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 \(\hat{C}_{vln}\) remains the same.\(^7\) This estimate is sometimes referred to as the “geometric coefficient of variation”\(^8\) in order to distinguish it from the simple estimate above. However, "geometric coefficient of variation" has also been defined\(^9\) as:

\[
GCV = e^{s_{ln}} - 1
\]

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.

**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\[^{10}\] to be

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

where the symbol \( \sum_{\prime} \) 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.

**Similar ratios**

Standardized moments are similar ratios, \( \mu_k / \sigma^k \), which are also dimensionless and scale invariant. The variance-to-mean ratio, \( \sigma^2 / \mu \) 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 \( \mu / \sigma \) is referred to as the signal to noise ratio.

- Relative standard deviation, \( |\sigma / \mu| \)
- Standardized moment, \( \mu_k / \sigma^k \)
- Variance-to-mean ratio, \( \sigma^2 / \mu \)
- Fano factor, \( \sigma_W^2 / \mu_W \) (windowed VMR)
- Signal-to-noise ratio, \( \mu / \sigma \) (in signal processing)
  - Signal-to-noise ratio (image processing)

**See also**

- Sampling (statistics)

**References**

1. ^ "What is the difference between ordinal, interval and ratio variables? Why should I care?"

en.wikipedia.org/wiki/Coefficient_of_variation


