WIKIPEDIA

Sample mean and covariance

The **sample mean** (or "empirical mean") and the **sample covariance** are statistics computed from a sample of data on one or more random variables.

The sample mean is the average value (or mean value) of a sample of numbers taken from a larger population of numbers, where "population" indicates not number of people but the entirety of relevant data, whether collected or not. A sample of 40 companies' sales from the Fortune 500 might be used for convenience instead of looking at the population, all 500 companies' sales. The sample mean is used as an estimator for the population mean, the average value in the entire population, where the estimate is more likely to be close to the population mean if the sample is large and representative. The reliability of the sample mean is estimated using the standard error, which in turn is calculated using the variance of the sample. If the sample is random, the standard error falls with the size of the sample and the sample mean's distribution approaches the normal distribution as the sample size increases.

The term "sample mean" can also be used to refer to a vector of average values when the statistician is looking at the values of several variables in the sample, e.g. the sales, profits, and employees of a sample of Fortune 500 companies. In this case, there is not just a sample variance for each variable but a sample variance-covariance matrix (or simply covariance matrix) showing also the relationship between each pair of variables. This would be a 3×3 matrix when 3 variables are being considered. The sample covariance is useful in judging the reliability of the sample means as estimators and is also useful as an estimate of the population covariance matrix.

Due to their ease of calculation and other desirable characteristics, the sample mean and sample covariance are widely used in statistics to represent the location and dispersion of the distribution of values in the sample, and to estimate the values for the population.

Contents

Definition of the sample mean

The sample mean is the average of the values of a variable in a sample, which is the sum of those values divided by the number of values. Using mathematical notation, if a sample of N observations on variable X is taken from the population, the sample mean is:

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i.$$

Under this definition, if the sample (1, 4, 1) is taken from the population (1,1,3,4,0,2,1,0), then the sample mean is $\bar{x} = (1+4+1)/3 = 2$, as compared to the population mean of $\mu = (1+1+3+4+0+2+1+0)/8 = 12/8 = 1.5$. Even if a sample is random, it is rarely perfectly representative, and other samples would have other sample means even if the samples were all from the same population. The sample (2, 1, 0), for example, would have a sample mean of 1.

If the statistician is interested in *K* variables rather than one, each observation having a value for each of those *K* variables, the overall sample mean consists of *K* sample means for individual variables. Let x_{ij} be the *i*th independently drawn observation (*i*=1,...,*N*) on the *j*th random variable (*j*=1,...,*K*). These observations can be arranged into *N* column vectors, each with *K* entries, with the *K*×1 column vector giving the *i*-th observations of all variables being denoted $\mathbf{x}_i$ (*i*=1,...,*N*).

The **sample mean vector** $\bar{\mathbf{x}}$ is a column vector whose *j*-th element $\bar{x}_j$ is the average value of the *N* observations of the *j*th variable:

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}, \quad j = 1, \ldots, K.$$

Thus, the sample mean vector contains the average of the observations for each variable, and is written

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i = \begin{bmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_j \\ \vdots \\ \bar{x}_K \end{bmatrix}$$

Definition of sample covariance

The **sample covariance matrix** is a *K*-by-*K* matrix $\mathbf{Q} = [q_{jk}]$ with entries

$$q_{jk} = \frac{1}{N-1}\sum_{i=1}^{N} \left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right),$$

where q_{jk} is an estimate of the covariance between the jth variable and the kth variable of the population underlying the data. In terms of the observation vectors, the sample covariance is

$$\mathbf{Q} = \frac{1}{N-1}\sum_{i=1}^{N} (\mathbf{x}_i . - \bar{\mathbf{x}})(\mathbf{x}_i . - \bar{\mathbf{x}})^{\mathrm{T}},$$

Alternatively, arranging the observation vectors as the columns of a matrix, so that

$$\mathbf{F} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_N \end{bmatrix},$$

which is a matrix of *K* rows and *N* columns. Here, the sample covariance matrix can be computed as

$$\mathbf{Q} = \frac{1}{N-1}(\mathbf{F} - \bar{\mathbf{x}}\,\mathbf{1}_N^{\mathrm{T}})(\mathbf{F} - \bar{\mathbf{x}}\,\mathbf{1}_N^{\mathrm{T}})^{\mathrm{T}},$$

where $\mathbf{1}_N$ is an *N* by 1 vector of ones. If the observations are arranged as rows instead of columns, so $\bar{\mathbf{x}}$ is now a 1×*K* row vector and $\mathbf{M} = \mathbf{F}^{\mathrm{T}}$ is an *N*×*K* matrix whose column *j* is the vector of *N* observations on variable *j*, then applying transposes in the appropriate places yields

$$\mathbf{Q} = \frac{1}{N-1}(\mathbf{M} - \mathbf{1}_N\bar{\mathbf{x}})^{\mathrm{T}}(\mathbf{M} - \mathbf{1}_N\bar{\mathbf{x}}).$$

Like covariance matrices for random vector, sample covariance matrices are positive semi-definite. To prove it, note that for any matrix $\mathbf{A}$ the matrix $\mathbf{A}^T\mathbf{A}$ is positive semi-definite. Furthermore, a covariance matrix is positive definite if and only if the rank of the $\mathbf{x}_{i}. - \bar{\mathbf{x}}$ vectors is K.

Unbiasedness

The sample mean and the sample covariance matrix are unbiased estimates of the mean and the covariance matrix of the random vector $\mathbf{X}$, a row vector whose *j*th element (*j = 1, ..., K*) is one of the random variables.[1] The sample covariance matrix has $N-1$ in the denominator rather than N due to a variant of Bessel's correction: In short, the sample covariance relies on the difference between each observation and the sample mean, but the sample mean is slightly correlated with each observation since it is defined in terms of all observations. If the population mean $\operatorname{E}(\mathbf{X})$ is known, the analogous unbiased estimate

$$q_{jk} = \frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \operatorname{E}(X_j)\right)\left(x_{ik} - \operatorname{E}(X_k)\right),$$

using the population mean, has N in the denominator. This is an example of why in probability and statistics it is essential to distinguish between random variables (upper case letters) and realizations of the random variables (lower case letters).

The maximum likelihood estimate of the covariance

$$q_{jk} = \frac{1}{N}\sum_{i=1}^{N}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)$$

for the Gaussian distribution case has *N* in the denominator as well. The ratio of 1/*N* to 1/(*N* − 1) approaches 1 for large *N*, so the maximum likelihood estimate approximately equals the unbiased estimate when the sample is large.

Distribution of the sample mean

For each random variable, the sample mean is a good estimator of the population mean, where a "good" estimator is defined as being efficient and unbiased. Of course the estimator will likely not be the true value of the population mean since different samples drawn from the same distribution will give different sample means and hence different estimates of the true mean. Thus the sample mean is a random variable, not a constant, and consequently has its own distribution. For a random sample of *N* observations on the *j*th random variable, the sample mean's distribution itself has mean equal to the population mean $E(X_j)$ and variance equal to σ_j^2/N, where σ_j^2 is the population variance.

The arithmetic mean of a population, or population mean, is often denoted μ.[2] The sample mean $\bar{x}$ (the arithmetic mean of a sample of values drawn from the population) makes a good estimator of the population mean, as its expected value is equal to the population mean (that is, it is an unbiased estimator). The sample mean is a random variable, not a constant, since its calculated value will randomly differ depending on which members of the population are sampled, and consequently it will have its own distribution. For a random sample of n independent observations, the expected value of the sample mean is

$$\operatorname{E}(\bar{x}) = \mu$$

and the variance of the sample mean is

$$\operatorname{var}(\bar{x}) = \frac{\sigma^2}{n}.$$

If the samples are not independent, but correlated, then special care has to be taken in order to avoid the problem of pseudoreplication.

If the population is normally distributed, then the sample mean is normally distributed as follows:

$$\bar{x} \thicksim N\left\{\mu, \frac{\sigma^2}{n}\right\}.$$

If the population is not normally distributed, the sample mean is nonetheless approximately normally distributed if n is large and $\sigma^2/n < +\infty$. This is a consequence of the central limit theorem.

Weighted samples

In a weighted sample, each vector $\textbf{x}_i$ (each set of single observations on each of the K random variables) is assigned a weight $w_i \geq 0$. Without loss of generality, assume that the weights are normalized:

$$\sum_{i=1}^N w_i = 1.$$

(If they are not, divide the weights by their sum). Then the weighted mean vector $\mathbf{\bar{x}}$ is given by

$$\mathbf{\bar{x}} = \sum_{i=1}^N w_i \mathbf{x}_i.$$

and the elements q_{jk} of the weighted covariance matrix $\mathbf{Q}$ are [3]

$$q_{jk} = \frac{1}{1-\sum_{i=1}^N w_i^2} \sum_{i=1}^N w_i \left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right).$$

If all weights are the same, $w_i = 1/N$, the weighted mean and covariance reduce to the (biased) sample mean and covariance mentioned above.

Criticism

The sample mean and sample covariance are not robust statistics, meaning that they are sensitive to outliers. As robustness is often a desired trait, particularly in real-world applications, robust alternatives may prove desirable, notably quantile-based statistics such as the sample median for location,[4] and interquartile range (IQR) for dispersion. Other alternatives include trimming and Winsorising, as in the trimmed mean and the Winsorized mean.

See also

- Estimation of covariance matrices
- Scatter matrix
- Unbiased estimation of standard deviation

References

1. Richard Arnold Johnson; Dean W. Wichern (2007). *Applied Multivariate Statistical Analysis* (https://books.google.com/books?id=gFWcQgAACAAJ). Pearson Prentice Hall. ISBN 978-0-13-187715-3. Retrieved 10 August 2012.
2. Underhill, L.G.; Bradfield d. (1998) *Introstat*, Juta and Company Ltd. ISBN 0-7021-3838-X p. 181 (https://books.google.com/books?id=f6TlVjrSAsgC&lpg=PP1&pg=PA181#v=onepage&q&f=false)
3. Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. GNU Scientific Library - Reference manual, Version 2.6 (http://www.gnu.org/software/gsl/doc/html/), 2021. Section Statistics: Weighted Samples (https://web.archive.org/web/20210418031437/https://www.gnu.org/software/gsl/doc/html/statistics.html#weighted-samples)
4. The World Question Center 2006: The Sample Mean (http://www.edge.org/q2008/q08_16.html#kosko), Bart Kosko

Retrieved from "https://en.wikipedia.org/w/index.php?title=Sample_mean_and_covariance&oldid=1063504049"

This page was last edited on 3 January 2022, at 12:19 (UTC).