[STAT-01] Mean, variance and correlation

Miguel-Angel Canela
Associate Professor, IESE Business School

The mean

Summary statistics can be used for both descriptive and testing purposes. The main three summary statistics are the mean, the variance and the correlation, discussed in this lecture. I start with the **mean**.

Let $\mathbf{x}$ be an n-vector, with components x_1, x_2, …, x_n. The expectation operator E is defined as

$$\mathrm{E}[\mathbf{x}] = \frac{x_1 + \cdots + x_n}{n}\,.$$

The value returned by this operator, when applied to a vector, is the mean. In Statistics, the mean is frequently called **expectation** or expected value. Typically, the x_i's are values of a **variable** obtained in a **data collection** experience.

The mean is taken as a central value (see discussion below). Note that, although the mean is called sometimes expected value, it is not what you "expect" to observe, but the average of what you have already observed. For instance, in a population, the expected value of the number of children per male inhabitant may be 0.7, but none of those inhabitants is expected to have 0.7 children.

We denote the mean by $\bar{x}$. I use in this course both the expectation operator E and the "bar" notation, as they suit me in every case. The same notational approach is used for the variance, standard deviation, etc.

The properties of the expectation are easy to understand. Let me list some of them:

- If $\mathbf{x} \leq \mathbf{y}$, then $\mathrm{E}[\mathbf{x}] \leq \mathrm{E}[\mathbf{y}]$.
- If $\mathbf{x}$ is constant equal to a, then $\mathrm{E}[\mathbf{x}] = a$.
- If a and b are constants, $\mathrm{E}\big[a\mathbf{x} + b\mathbf{y}\big] = a\,\mathrm{E}[\mathbf{x}] + b\,\mathrm{E}[\mathbf{y}]$.
- **Jensen's inequality**. If h is a convex function, $h\big(\mathrm{E}[\mathbf{x}]\big) \leq \mathrm{E}\big[h(\mathbf{x})\big]$. If h is strictly convex (i.e. $h''(x) > 0$), the inequality is strict, except when $\mathbf{x}$ is constant.
- A measure such as $\mathrm{E}\big[(\mathbf{x} - a)^2\big]$ is called a **mean squared error** (MSE). The minimum MSE is attained when a coincides with the expectation of $\mathbf{x}$. In short,

$$\mathrm{E}[\mathbf{x}] = \arg\min_a \mathrm{E}\big[(\mathbf{x} - a)^2\big].$$

The first three properties are obvious. You are expected to know Jensen's inequality form the Mathematics course, but exercise A may help you to refresh it. Exercise B is the proof of the last property.

The median

The mean is not always the value "in the middle", so that one half of the observations are above the mean and the other half below. The value with this property is called the **median**.

To calculate the median, we sort the data. If n is odd, the mean is equal to the data point in the middle of the list, $x_{(n+1)/2}$. If n is even, the median is the midpoint of $x_{n/2}$ and $x_{n/2+1}$. How close is the mean to the median is taken as an indication of the "symmetry" of the data.

A supersimple example: $x_1 = 1$, $x_2 = 2$, $x_3 = 3$, $x_4 = 10$. The mean is 2.5, with 3/4 of the observations on the left and 1/4 on the right. The median is 2.5, providing a 50–50 split.

The variance and the standard deviation

The minimal description of the data should contain a central measure, such as the mean or the median, and a **dispersion measure**. The latter tells us about how concentrated around the central value the observations can be expected to be. Because of its mathematical properties, the **variance** is the preferred dispersion measure. The variance operator is defined as

$$\operatorname{var}[\mathbf{x}] = \frac{1}{n-1} \sum_{i=1}^{n} \big(x_i - \bar{x}\big)^2.$$

This operator returns the variance, usually denoted by s^2. If needed, one can use subscripts, as in s_x^2. The following properties of the variance result directly from those of the expectation:

- $\operatorname{var}[\mathbf{x}] = 0$ if and only if $\mathbf{x}$ is constant.
- For a constant, $\operatorname{var}[a\mathbf{x}] = a^2 \operatorname{var}[\mathbf{x}]$.

But, in general, $\operatorname{var}[\mathbf{x} + \mathbf{y}] \neq \operatorname{var}[\mathbf{x}] + \operatorname{var}[\mathbf{y}]$, as discussed below.

The **standard deviation** $\operatorname{sd}[\mathbf{x}]$ is the square root of the variance. It is denoted by s, eventually with a subscript.

Note that the standard deviation has the same units as the data, but the variance has not. If $\mathbf{x}$ comes in dollars, both $\bar{x}$ and s are in dollars, but s^2 is in squared dollars. So, we use variances in statistical analysis, but we report standard deviations, which are easier to interpret. You may wonder why we use squares to calculate the variance, taking later the square root to get the standard deviation. We do it by the same reason that we use the square root of a sum of squares to calculate the distance between two points in the space, that is, because of Pythagoras theorem. We will find this soon in this course.

The transformation

$$\mathbf{z} = \frac{\mathbf{x} - \bar{x}}{s}$$

is called **standardization**. It follows from the properties of the expectation that $\mathbf{z}$ has zero mean and unit variance. The letter z is frequently used for standardized variables. The terms z-transform, z-values and z-scores are also popular. The advantage of using z-scores is that all variables have a common scale, allowing for direct comparison of many statistics.

The covariance

Let $\mathbf{x}$ and $\mathbf{y}$ be n-vectors. The **covariance** of these vectors is defined as

$$\operatorname{cov}[\mathbf{x}, \mathbf{y}] = \frac{1}{n-1} \sum_{i=1}^{n} \big(x_i - \bar{x}\big)\big(y_i - \bar{y}\big).$$

The covariance is also denoted by s_{xy}. The covariance of $\mathbf{x}$ and $\mathbf{x}$ is the same as the variance of $\mathbf{x}$.

Note that, dropping the denominator $n - 1$, the covariance is just the dot product of the *centered* vectors $\mathbf{x} - \bar{x}$ and $\mathbf{y} - \bar{y}$, while the variance of $\mathbf{x}$ is the squared modulus of $\mathbf{x}$. So, covariances inherit the algebraic properties of products, and variances those of squares. For instance,

$$\operatorname{cov}\big[a\mathbf{x} + b\mathbf{y}, \mathbf{z}\big] = a \operatorname{cov}\big[\mathbf{x}, \mathbf{z}\big] + b \operatorname{cov}\big[\mathbf{y}, \mathbf{z}\big],$$

and

$$\operatorname{var}[\mathbf{x}+\mathbf{y}] = \operatorname{var}[\mathbf{x}] + \operatorname{var}[\mathbf{y}] + 2\operatorname{cov}[\mathbf{x},\mathbf{y}].$$

Note that $\operatorname{var}[\mathbf{x}+\mathbf{y}] = \operatorname{var}[\mathbf{x}] + \operatorname{var}[\mathbf{y}]$ if and only if $\operatorname{cov}[\mathbf{x},\mathbf{y}] = 0$, that is, when the product of $\mathbf{x}-\bar{x}$ and $\mathbf{y}-\bar{y}$ is zero. Mathematicians call this **orthogonality**. So, the variance is additive when there is orthogonality. Although statisticians occasionally use the term orthogonal to refer to this situation, they prefer to say **uncorrelated** (this term is explained in the next section). Standard deviations do not have this property.

The correlation

The interpretation of the sign of the covariance is direct. When $s_{xy} > 0$, high (resp. low) values of one variable occur jointly when high (resp. low) values of the other variable. With $s_{xy} < 0$, it is the other way round. But, since the covariance depends on the scale used for measuring the variables, the interpretation of its absolute value is not that easy. The practice favours a standardized version of the covariance. When the covariance is calculated for the standardized variables, it is called (linear) **correlation**. An equivalent definition is

$$r = \frac{s_{xy}}{s_x\, s_y}\,.$$

Cancelling out $n-1$ in this ratio, we have the product of $(\mathbf{x}-\bar{x})$ and $(\mathbf{y}-\bar{y})$ in the numerator and the norms in the denominator. So, the correlation is just the cosine of these two vectors. From what we know about that, we can easily get the following properties of correlation, which will not surprise you if you are acquainted with the regression line.

- The correlation has the same sign as the covariance.
- Always $-1 \le r \le 1$.
- $r = \pm 1$ when $(\mathbf{x}-\bar{x})$ and $(\mathbf{y}-\bar{y})$ are linearly dependent. This is equivalent to the existence of two constants a and b such that $\mathbf{y} = a + b\mathbf{x}$.
- Linear transformations do not affect the absolute value of the correlation:

$$\operatorname{cor}\big[a\mathbf{x}+b,\mathbf{y}\big] = \pm\operatorname{cor}\big[\mathbf{x},\mathbf{y}\big].$$

The formula of the variance of the sum can be written as

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2\,r\,s_x\,s_y.$$

So, when $r > 0$ the variance of the sum is higher than the sum of variances. This is intuitively clear: the two variables vary in the same direction, so we get extra variance for the sum. If $r < 0$, they vary in opposite directions, so the variance of the sum is less than the sum of the separate variances. When they are uncorrelated, the variance is additive. With a bit of thinking, you may discover that the standard deviation is *never* additive, since we always have $s_{x+y} < s_x + s_y$, except when one of the variables is constant.

Covariance matrices

Let me consider now n joint observations of k variables. This is called multivariate data. Example: the weight, the height and the cholesterol level of a sample of n individuals. I arrange the data as a **data matrix X**, with n rows and k columns, so that every row is a sample unit and every column is a variable. Let me denote by $\mathbf{x}_1$, $\mathbf{x}_2$, …, $\mathbf{x}_k$ the columns of $\mathbf{X}$.

The **covariance matrix S** (also called variance matrix) has, in row i and column j, the covariance $s_{ij} = \mathrm{cov}[\mathbf{x}_i, \mathbf{x}_j]$,

$$\mathbf{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1k} \\ s_{21} & s_{22} & \cdots & s_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ s_{k1} & s_{k2} & \cdots & s_{kk} \end{bmatrix}.$$

The covariance matrix is symmetric. It is also definite positive, as shown by the following argument. From the formula given for the variance of a sum, we can easily obtain a formula for the variance of a linear combination $\mathbf{y} = a_1\mathbf{x}_1 + \cdots + a_k\mathbf{x}_k$, which admits a compact expression in matrix notation. Indeed, putting $\mathbf{y} = \mathbf{a}^\mathsf{T}\mathbf{X}$, the formula is

$$\mathrm{var}[\mathbf{y}] = \mathbf{a}^\mathsf{T}\,\mathbf{S}\,\mathbf{a}.$$

A consequence of this formula is that a covariance matrix must be at least positive semidefinite, since $\mathrm{var}[\mathbf{y}] \geq 0$ for any linear combination $\mathbf{y}$. Moreover, it is positive definite, unless there is a non-trivial linear combination of the $\mathbf{x}$'s which is constant (only constants have null variance).

The above formula is easily extended: if we transform the data b(n, k)-matrix $\mathbf{X}$ into a dataq (n, m)-matrix $\mathbf{Y}$ using a (k, m)-matrix $\mathbf{A}$, we get

$$\mathrm{cov}\big[\mathbf{Y}] = \mathbf{A}\,\mathbf{S}\,\mathbf{A}^\mathsf{T}.$$

Correlation matrices are defined in a similar way:

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1k} \\ s_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{bmatrix}$$

Of course, a correlation matrix is a particular case of a covariance matrix, so anything said for covariance matrices applies to correlation matrices (not conversely).

Homework

A. For a set of positive observations, compare the logarithm of the mean with the mean of the logarithms. Which is higher? Why?

B. Let $x_1 < x_2 < \cdots < x_n$. Using differential calculus, prove that the value of a for which the sum of squared deviations $(x_i - a)^2$ is minimum is the mean $\bar{x}$. What do we get for the sum of the absolute deviations $|x_i - a|$?