[STAT-01] Descriptive statistics

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 the discussion with the **mean**.

Let $\mathbf{x}$ be an n-vector, with components x_1, x_2, $\ldots$, 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 female inhabitant may be 1.7, but none of those inhabitants is expected to have 1.7 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 $\mathbf{x} \le \mathbf{y}$, then $\mathrm{E}[\mathbf{x}] \le \mathrm{E}[\mathbf{y}]$.
- 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, $h\big(\mathrm{E}[\mathbf{x}]\big) \le \mathrm{E}\big[h(\mathbf{x})\big]$. 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 my remember Jensen's inequality from the Mathematics course. Exercise A consists in proving the last of these properties.

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. It is the **median** which has this property.

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 4, with 3/4 of the observations on the left and 1/4 on the right. The median is 2.5, which splits this set of observations in two halves.

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, $\operatorname{var}[a\mathbf{x}] = a^2 \operatorname{var}[\mathbf{x}]$.

¶ In general, $\operatorname{var}[\mathbf{x} + \mathbf{y}] \neq \operatorname{var}[\mathbf{x}] + \operatorname{var}[\mathbf{y}]$. This is discussed in the next section.

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.

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 of reducing the variables to a common scale, which allows 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 norm of $\mathbf{x}$. So, the covariance inherit the algebraic properties of the dot product, and the variance those of the squared norm. 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

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

Note that $\mathrm{var}[\mathbf{x}+\mathbf{y}] = \mathrm{var}[\mathbf{x}] + \mathrm{var}[\mathbf{y}]$ if and only if $\mathrm{cov}[\mathbf{x},\mathbf{y}] = 0$, 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:

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

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

$$s^2_{x+y} = s^2_x + s^2_y + 2\,r\,s_x\,s_y.$$

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. When $r < 0$, 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. We arrange multivariate data as a **data matrix $\mathbf{X}$**, 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 $\mathbf{X}$. Then, 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 $\mathbf{y} = \mathbf{X}\mathbf{a}$, 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 (n, k)-matrix $\mathbf{X}$ into a data (n, m)-matrix $\mathbf{Y}$ using a (k, m)-matrix $\mathbf{A}$, we get

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

Correlation matrices are defined in a similar way:

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

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).

The regression line

Let me refresh, in the rest of this lecture, the basics of the regression line. Take a set of n joint observations on two variables X and Y, putting them as the columns $\mathbf{x}$ and $\mathbf{y}$ of a data matrix, and the coefficients b_0 and b_1 of the linear regression equation (of Y on X).

The expression $y = b_0 + b_1 x$ can be taken as the equation of a line, which we call the **regression line**. b_1 is the **slope** and b_0 the **intercept** or constant. In this case, the OLS formulas are reduced to simple expressions:

$$b_1 = \frac{\sum \big(x_i - \bar{x}\big)\big(y_i - \bar{y}\big)}{\sum \big(x_i - \bar{x}\big)^2}\,, \qquad b_0 = \bar{y} - b_1\bar{x}.$$

It follows from the formula of the intercept that $\bar{y} = b_0 + b_1\bar{x}$, meaning that the regression line crosses the average point $\big(\bar{x}, \bar{y}\big)$. This is equivalent to the sum (and the mean) of the residuals being equal to zero. A consequence of this is that we can write the equation of the regression line as $y - \bar{y} = b_1\big(x - \bar{x}\big)$. This intercept-free version of the regression equation is practical for some analyses.

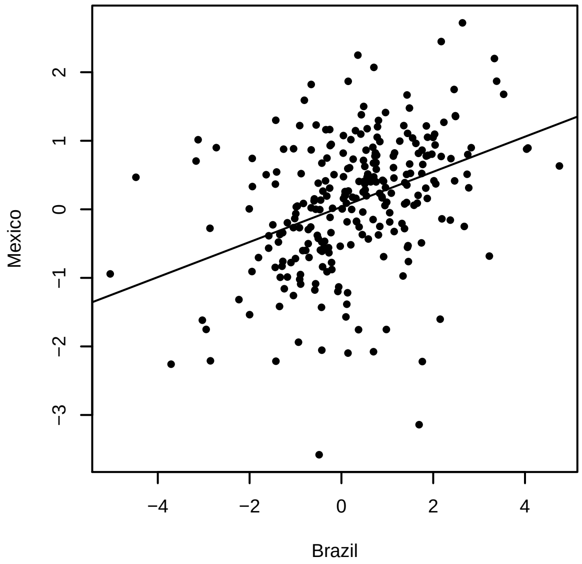

Figure 1. Regression line (Example 1)

Regression and correlation

The formulas related to the regression line become more compact if we introduce standard deviations and correlations, which make sense if we take the set of points as a bivariate sample. Dividing by $n - 1$ the numerator and denominator in the expression of the slope, we get

$$b_1 = \frac{s_{xy}}{s_x^2} = r\,\frac{s_y}{s_x}.$$

This tells us that the sample correlation is a standardized regression slope. If $\mathbf{x}$ and $\mathbf{y}$ have unit variance, the slope is equal to the correlation. Now, we can write the regression equation as

$$\frac{y - \bar{y}}{s_y} = r\,\frac{x - \bar{x}}{s_x}\,,$$

which shows that, if both variables are standardized, the slope coincides with the correlation and the intercept is zero. The former is no longer true in multiple regression, where **standardized regression coefficients** are not correlations, although, in most cases, they look as if they were.

Example 1. The `bramex` data set contains the daily returns of the Brazil and Mexico MSCI indexes. It has been extracted from the Datastream database and covers the whole year 2003, with a total of 261 observations (no data in week-ends).

The daily returns are derived from the index values as follows. If p_t is the value of a particular index at day t, the daily return at this day is given by $r_t = p_t/p_{t-1} - 1$. The returns used here come in percentage scale.

Let me denote by $\mathbf{x}$ the vector of Brazil returns and by $\mathbf{y}$ that of Mexico returns. The means are $\bar{x} = 0.273$ and $\bar{y} = 0.105$. The covariance and correlation matrices are, respectively,

$$\mathbf{S} = \begin{bmatrix} 2.120 & 0.543 \\ 0.543 & 0.934 \end{bmatrix}, \qquad \mathbf{R} = \begin{bmatrix} 1 & 0.386 \\ 0.386 & 1 \end{bmatrix}.$$

To calculate the regression line of Mexico on Brazil, I use the formulas of the regression coefficients:

$$b_1 = 0.386\sqrt{\frac{2.120}{0.934}} = 0.256, \qquad b_0 = 0.105 - 0.256 \times 0.273 = 0.036.$$

I have thus obtained the equation of the line for the regression of Mexico on Brazil,

$$\texttt{mex} = 0.036 + 0.256\,\texttt{bra}.$$

We can see in Figure 1 a scatterplot of these data, with the regression line superimposed.

¶ Source: MA Canela & E Pedreira (2012), Modelling dependence in Latin American markets using copula functions, *Journal of Emerging Markets Finance* **11**, 231–270.

The R-squared statistic

I turn now to general linear regression, where the correlation issue is a bit more complex. I start from the equation

$$\mathbf{y} = b_0\mathbf{1} + b_1\mathbf{x}_1 + \cdots + b_k\mathbf{x}_k + \mathbf{e},$$

which is, in fact, a decomposition of $\mathbf{y}$ into two orthogonal vectors. Statisticians prefer to write this as

$$\mathbf{y} - \bar{y} = b_1\big(\mathbf{x}_1 - \bar{x}_1\big) + \cdots + b_k\big(\mathbf{x}_k - \bar{x}_k\big) + \mathbf{e},$$

which is also an orthogonal decomposition. Then, Pythagoras theorem gives us

$$\big\|\mathbf{y} - \bar{y}\big\|^2 = \big\|b_1\big(\mathbf{x}_1 - \bar{x}_1\big) + \cdots + b_k\big(\mathbf{x}_k - \bar{x}_k\big)\big\|^2 + \big\|\mathbf{e}\big\|^2.$$

The left side of the equation is

$$\big\|\mathbf{y} - \bar{y}\big\|^2 = \sum_{i=1}^{n}\big(y_i - \bar{y}\big)^2.$$

Statisticians call this a **sum of squares**. More specifically, it would be the **total sum of squares**, SST. The first term on the right side is the **sum of squares explained by the regression**, SSE. The last term is the residual sum of squares,

$$\mathrm{SSR} = \sum e_i^2.$$

So, the orthogonal decomposition becomes a formula involving sums of squares, SST = SSE + SSR. This type of formula is called, in Statistics, an **ANOVA decomposition** (this will be later explained). The **residual sum of squares** SSR is a measure of the **goodness-of-fit**. But, in practice, it is difficult to interpret, since it depends on the units of Y and the number of points. So, we use the **R-squared statistic**, defined as follows,

$$R^2 = \frac{\mathrm{SSE}}{\mathrm{SST}}\,.$$

R^2 is taken as the percentage of variation explained by the regression. It is not hard to see that R^2 is the square of the correlation between $\mathbf{y}$ and the vector of **predicted values** $b_0\mathbf{1}+b_1\mathbf{x}_1+\cdots+b_k\mathbf{x}_k$, which is called the **multiple correlation**. An R-squared value which is close to 1 is taken as an indication of good fit. Nevertheless, how good is the fit that we can expect for a particular type of data is something that we learn only with practice, so it is better to skip general specifications of threshold values for the R-squared statistic.

Homework

A. 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 $\bar{x}$.

B. Although the idea of the financial performance of a firm may seem obvious, there is no consensus on how to measure it. Two well known measures are the **return on equity** (ROE) and the **return on assets** (ROA). The ROE measures a firm's efficiency at generating profits from every unit of shareholders' equity. The ROA tells us how profitable the firm's assets are in generating revenue, or, more specifically, how many dollars of earnings it derives from each dollar of assets it controls.

The `roeroa` data set covers a wide range of industries. It contains the ROE and the ROA of 426 firms for the year 2000, derived from public sources. The ROE has been calculated as net income over total equity, and the ROA as operating income over total assets. Perform a regression analysis. What is your conclusion?

C. The `indianbanks` data set contains data on daily opening prices of five Indian banks in the National Stock Exchange (NSE), from 2002-08-12 to 2013-12-31, extracted from Yahoo Finance India. Calculate the correlation matrix of the daily returns and discuss.