COURSE

1 Review on Bivariate Statistics

The objective of this statistical study is to analyze, within the same population of N individuals, two different characteristics (or different modalities) and to determine whether there exists a link or correlation between these two variables. Examples of possible relationships between the following variables: height and age; diabetes and weight; cholesterol level and diet; ecological niche and population; sunlight and plant growth; toxin and metabolic reaction; survival and pollution; effects and doses; organ 1 and organ 2; organ and biological function; ...

1 Bivariate Statistical Series

Definition

A statistical series with two variables (or bivariate statistical series) is a statistical series in which two characteristics are studied simultaneously.

Example 01

For a car model, the fuel consumption (in L/100 km) was recorded at different speeds (in km/h) in fifth gear:

Speed x_i(inkm/h)	60	70	90	110	130	150
Consumption y_i(inL/100km)	3	3.1	3.7	4.7	6	9

1.1.1 Scatter Plot

Definition

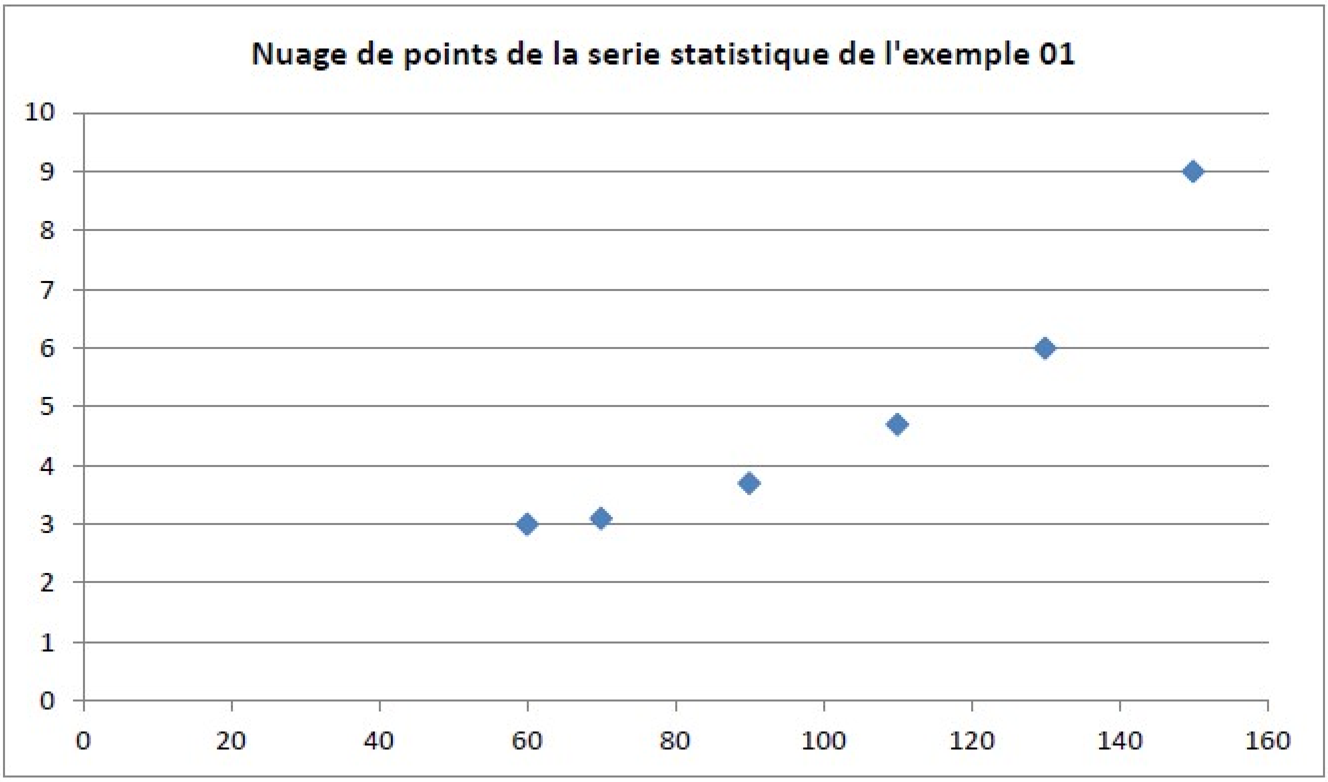
In an orthogonal coordinate system, the set of points M_i with coordinates (x_i, y_i) constitutes the scatter plot associated with the bivariate statistical series.

1.1.2 Marginal Means

Definition

$\bar{x}$ represents the mean of x_i:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{N}\sum_{i=1}^{n} x_i$$



$\bar{y}$ represents the mean of y_i:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{N} = \frac{1}{N}\sum_{i=1}^{n} y_i$$

Example

The marginal means from Example 01 are:

$$\bar{x} = \frac{60 + 70 + 90 + 110 + 130 + 150}{6} = 101.66$$
$$\bar{y} = \frac{3 + 3.1 + 3.7 + 4.7 + 6 + 9}{6} = 4.91$$

1.1.3 Covariance

Definition

The covariance of x and y is defined as the number

$$\mathrm{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \left(\frac{1}{N}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y}$$

Recall

The variance of the variable x is:

$$\boldsymbol{V}(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \left(\frac{1}{N}\sum_{i=1}^{n}{x_i}^2\right) - \bar{x}^2 = \operatorname{cov}(x,x)$$

The variance of the variable y is:

$$\boldsymbol{V}(\boldsymbol{y}) = \frac{1}{\boldsymbol{N}}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \left(\frac{\mathbf{1}}{\boldsymbol{N}}\sum_{i=1}^{n}{y_i}^2\right) - \bar{y}^2 = \operatorname{cov}(\boldsymbol{y},\boldsymbol{y})$$

It is used to calculate the standard deviation: $\sigma(x) = \sqrt{\boldsymbol{V}(\boldsymbol{x})}, \sigma(y) = \sqrt{\boldsymbol{V}(\boldsymbol{y})}$.

Example

Compute in Example 01: $\operatorname{cov}(\boldsymbol{x},\boldsymbol{y}), \mathbf{cov}(\boldsymbol{x},\boldsymbol{x}), \mathbf{cov}(\boldsymbol{y},\boldsymbol{y}), \sigma(x), \sigma(y)$.
We have:

							Sum
x_i	60	70	90	110	130	150	
y_i	3	3.1	3.7	4.7	6	9	
x_iy_i	180	217	333	517	780	1350	3377
${x_i}^2$	3600	4900	8100	12100	16900	22500	68100
${y_i}^2$	9	9.61	13.69	22.09	36	81	171.39

$$\bar{x} = 101.66, \quad \bar{y} = 4.91$$

$$\operatorname{cov}(\boldsymbol{x},\boldsymbol{y}) = \left(\frac{1}{N}\sum_{i=1}^{n}x_iy_i\right) - \bar{x}\bar{y} = \frac{3377}{6} - 499.15 = 63.68$$

$$\boldsymbol{V}(\boldsymbol{x}) = \left(\frac{1}{N}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 = \frac{\mathbf{68100}}{6} - (\mathbf{101.66})^2 = \mathbf{1015.2444}$$

$$\boldsymbol{V}(\boldsymbol{y}) = \left(\frac{1}{N}\sum_{i=1}^{n}y_i^2\right) - \bar{y}^2 = \frac{\mathbf{171.39}}{\mathbf{6}} - (\mathbf{4.91})^2 = \mathbf{4.4569}$$

$$\sigma(x) = \sqrt{\boldsymbol{V}(\boldsymbol{x})} = \sqrt{\mathbf{1015.2444}} = 31.86$$

$$\sigma(y) = \sqrt{\boldsymbol{V}(\boldsymbol{y})} = \sqrt{\mathbf{4.4569}} = 2.11$$

Theorem

1. The regression line D of Y with respect to X has the equation $D(Y/X):\ Y = aX+b$, where:

$$a = \frac{\operatorname{cov}(x,y)}{V(x)}$$

and $\quad b = \bar{Y} - a\bar{X}$.
2. The regression line D of X with respect to Y has the equation $D(X/Y):\ X = a'Y+b'$, where:

$$a' = \frac{\operatorname{cov}(x,y)}{V(y)}$$

and $\quad b' = \bar{X} - a'\bar{Y}$.

Example

Compute in Example 01 the regression line D of Y with respect to X.
We have:

$$\bar{x} = 101.66, \bar{y} = 4.91, \operatorname{cov}(x, y) = 63.68, \quad V(x) = 1015.2444, \quad V(y) = 4.4569.$$

1. $D(Y/X) : Y = aX + b$

$$a = \frac{\operatorname{cov}(x, y)}{V(x)} = 0.0627, \quad b = \bar{Y} - a\bar{X} = -1.46.$$

Therefore $D(Y/X) : Y = aX + b = 0.0627X - 1.46$
2. $D(X/Y) : X = a'Y + b'$

$$a' = \frac{\operatorname{cov}(x, y)}{V(y)} = 14.287, \quad b' = \bar{X} - a'\bar{Y} = 31.51$$

Therefore $D(X/Y) : X = a'Y + b' = 14.287Y + 31.51$

1.1.4 Linear Correlation Coefficient

Definition

The linear correlation coefficient of a bivariate statistical series x and y is the number r defined by:

$$r = \frac{\operatorname{cov}(x, y)}{\sqrt{V(x)}\sqrt{V(y)}} = \frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)}$$

Remark

1. $-1 \leq r \leq 1$.
2. If $r = 1$ or $r = -1$, then there is a perfect positive or negative correlation between X and Y, and all the points $(\mathrm{x_i}, \mathrm{y_i})$ lie on the regression line.
A positive correlation means that an increase in X causes an increase in Y.
A negative correlation means that an increase in X causes a decrease in Y (or vice versa).
3. If $r = 0$, then there is no correlation between X and Y, and the points (x_i, y_i) are scattered randomly.
4. If $0 < r < 1$, then there is a weak, moderate, or strong positive correlation between X and Y.
5. If $-1 < r < 0$, then there is a weak, moderate, or strong negative correlation between X and Y.

Example

Compute in Example 01 the linear correlation coefficient.
We have $\operatorname{cov}(x, y) = 63.68, \sigma(x) = 31.86, \sigma(y) = 2.11$.
Therefore:

$$r = \frac{\operatorname{cov}(x, y)}{\sigma(x)\sigma(y)} = 0.947$$

Hence, there is a strong positive correlation between X and Y.