Course 3 :
Processing quantitative bivariate data
-
Correlation

20 november 2012

Introduction

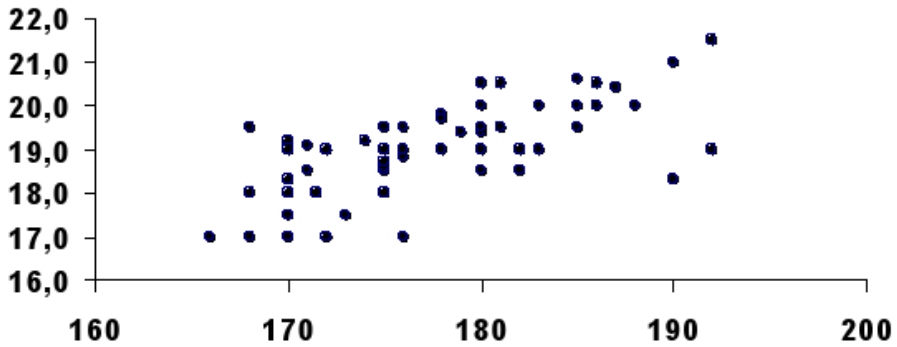
Introduction

When the data is described by two values (ie random variables), generaly, we are interested to know the possible statistical link between these two variables.

The problem

- Is there a *statistical link* between the both variabls X and Y ?
  - The value of X depend upon the value of Y ?
- What is the strength (weight) of the relationship between X and Y ?
  - If I know perfectly X, how accurately can I deduct Y (and vice versa) ?
  - What is the correlation between X and Y ?
- What is the *numerical relation* between X and Y ?
  - If I know X well, which is the function that allows me to estimate Y (and vice versa) ?
  - Quelle est la régression entre X et Y ?

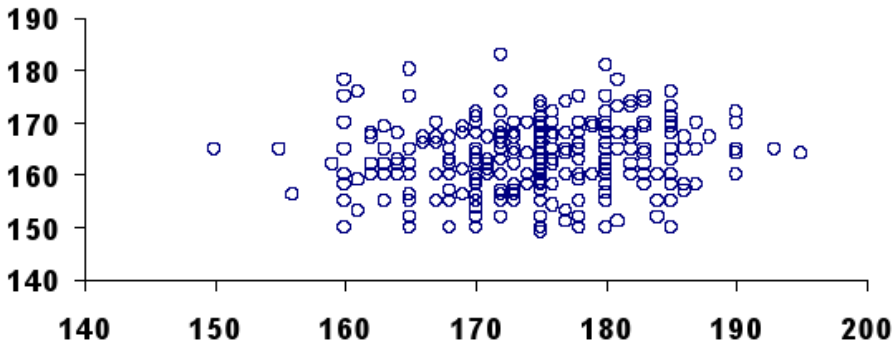
Examples



Bird egg size depending on the size of the ponder

- There is a clear link between the two variables.
- The possibility to make a regression in order to estimate the approximate size of an egg by the ponder's size and vice versa (Predictive Analytics course).

Examples



Respective weight of man and woman in a familly

- Here there is no link between the two variables.
- No regression...

The interest

- The study of these relationships is often very useful for the understanding of a phenomenon.
  - Understand the relationships between different aspects of the phenomenon.
  - Discover redundancy in the description of a phenomenon (dimensionality reduction).

- In this course we will study linear relationships between variables.
  - They are simple to analyze.
  - It is often easier to apply transformations on the values of variables to be reduced to the linear case.

Covariance

Introduction

Covariance measure a link between two variables.

- Then data are far from a mean of a variable, and it is also far from the mean o f the second variable, when the covariance between the two variables is high.
- When two variables *move* in the same direction, their covariance is *positive*.
- When two variables *move* oppositely, their covariance is *negative*. Finally, when two variables are completely

  independent of each other, their covariance is *null*.
  For example the annual production of corn in Peru and the fluctuation of the temperature at the surface of Mars.

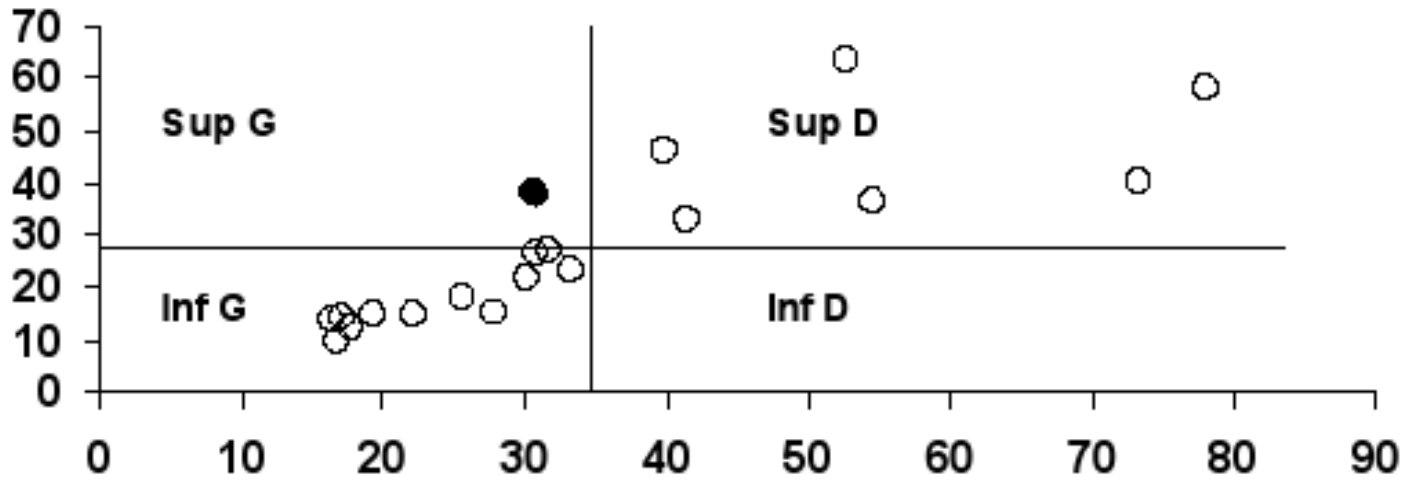
Computation

The formula for the covariance is very close to the variance formula :

The covariance between X and Y

$$Cov(X, Y) = \frac{1}{n \text{ ou } n-1} \sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)$$

Computation



Interpretation

Attention

The value of the covariance depends entirely on the selected units for the variables X and Y !

- The interpretation is very difficult.
- We need to find a criterion that does not depend on these units.

Correlation Coefficient

Summary

We want an index that gives the same information as the covariance, but does not depend on the units of the variables :

This is **Correlation**

Computation

- The maximum possible covariance between two variables (when both values are changing in exactly the same way) is equal to the product of their standard deviations.
- So just divide the covariance by the product of their standard deviations to obtain a unit index between -1 and 1.

COrrelation between X and Y

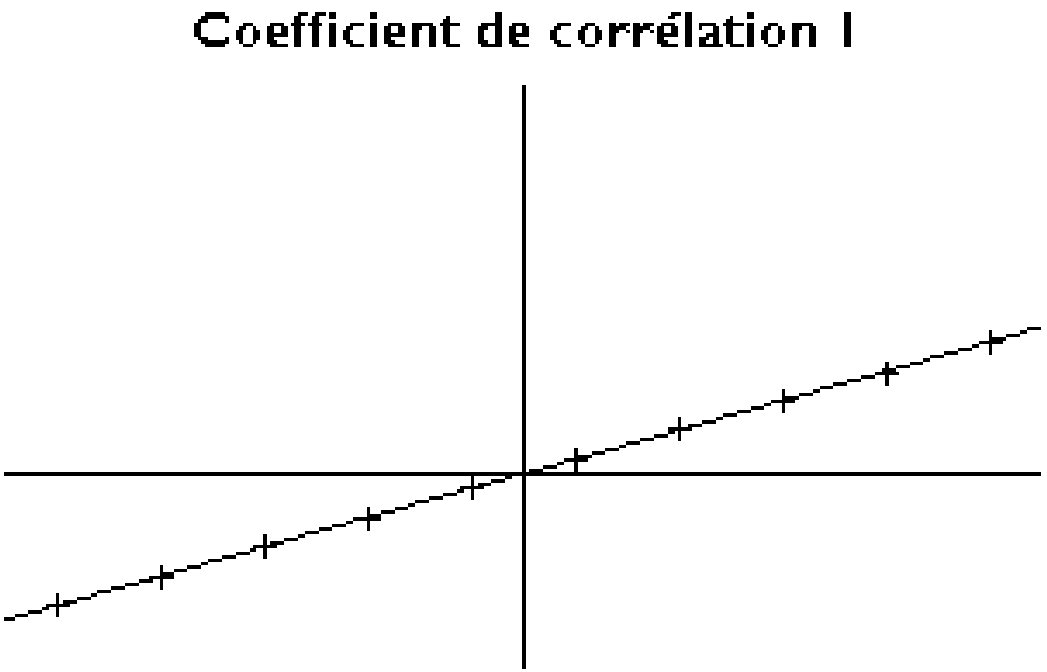
$$r = \frac{Cov(X, Y)}{s_x s_y}$$

Interpretation

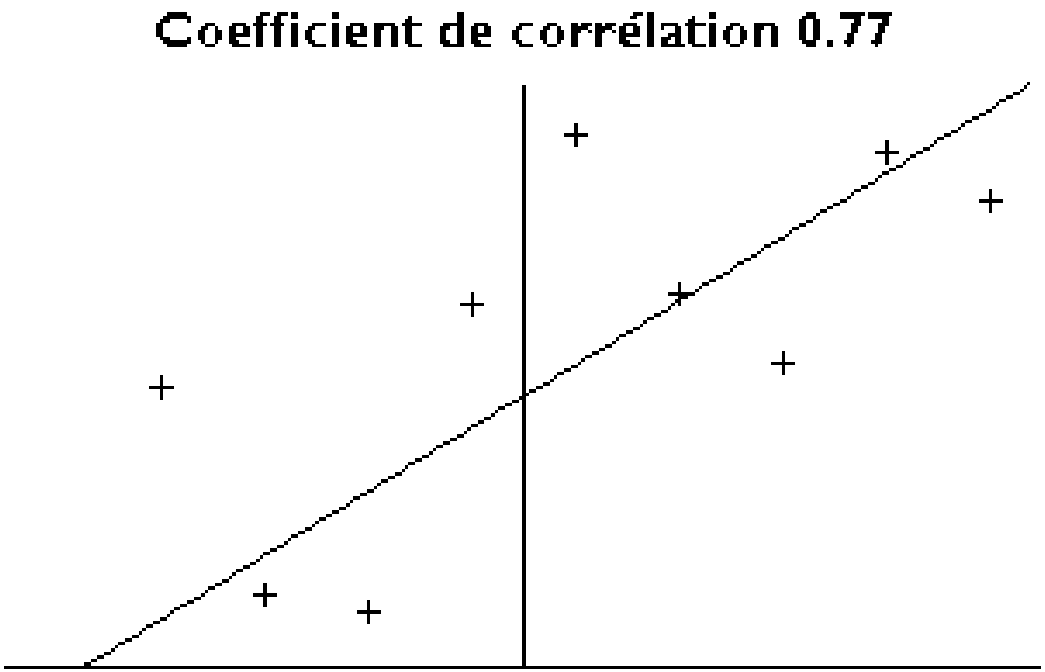
The value of the correlation is independent of the selected units for the variables X and Y.

- When *r* is close to 0, the relationship between the two variables is very low (and vice versa).
- When *r* is positive, the relationship between the two variables is positive.
- When *r* is negative, the relationship between the two variables is negative.

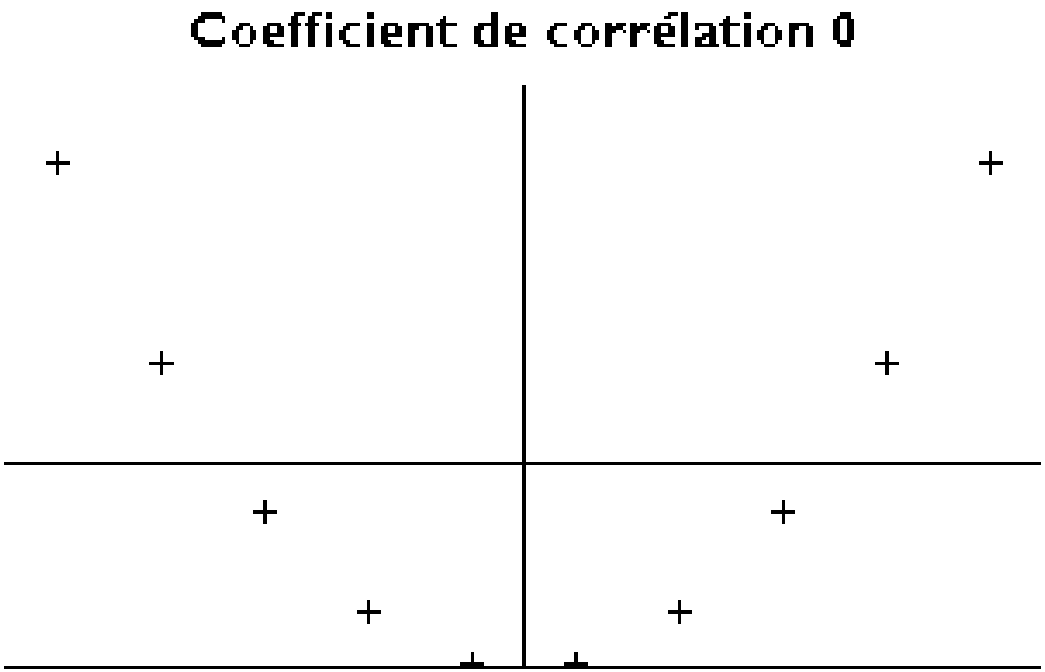
Interpretation



Interpretation



Interpretation



Interpretation

Attention

- Find a correlation (even ”high”) between the variable X and variable Y does not show at all that there is a causal link between X and Y (or between Y and X). For example : the number of sunburns as a function of temperature or of the number of persons ; computer power / cost for 20 years.
- The interpretation of *r* is unintuitive (r = -0.6 it is a strong correlation ?), There is a better index.

Coefficient of determination

Summary

The **coefficient of determination** is the proportion of the variance of Y, which disappears if X is fixed (or vice versa). The interpretation of the results becomes very easy.

Computation

- The coefficient of determination is the square of the correlation coefficient.
- r^2 is between 0 and 1, that is a proportion (or percentage).

Determination coefficient between X and Y

$$r^2 = \left(\frac{Cov(X, Y)}{s_X s_Y} \right)^2$$

Interpretation

The coefficient of determination is independent of the chosen units for the variables X and Y.

- When r^2 is close to 0, the relationship between the two variables is very low. Know perfectly X provides no information on Y.
- When r^2 is close to 1, the relationship between the two variables is very strong. Know perfectly X greatly reduces the variability (and thus the extent of probable values) of Y.

Interpretation

Example

- Let X and Y as $r = -0,6$.
- The correlation is negative.
- Is it strong ?
- $r^2 = 0.36$, ie that 36% of Y information is contained in X.
- It's not bad, but it also means that 64% of the Y information can not be known from X.
- The estimation of Y from X is unreliable.

Interpretation

Attention

- Now we are able to assess the strength of the bond between X and Y.
- We would now determine the function (here a linear equation) to estimate the best values of Y from those of X (or vice versa).
- So we will do a linear regression on the data (Predictive Analytics course).

Confidence intervals

Summary

As always when computing indices from samples, calculate a confidence index for estimating the reliability of the results. This is also used for the correlation coefficient, coefficient of determination and the parameters of the regression line.

Computation

IC_{95} of *rand* r^2

- If X (resp. Y) is fixed, the distribution of Y (resp. X) follows a normal distribution (often difficult to assess) :
  - Then $Z = \frac{ln(1+r)-ln(1-r)}{2}$ follows a normal distribution
  - With $s_Z = \sqrt{1/(n-3)}$
  - Then $Z_{inf} = Z - 1,96 s_Z$ et $Z_{sup} = Z + 1,96 s_Z$
  - Where $IC_{95}(r) = [\frac{e^{2Z_{inf}}-1}{e^{2Z_{inf}}+1} \ ; \ \frac{e^{2Z_{sup}}-1}{e^{2Z_{sup}}+1}]$
- **Otherwise :The bootstrap method still works !**
- For the coefficient of determination : $IC_{95}(r^2) = (IC_{95}(r))^2$

Interpretation

interpretation of the confidence interval is identical to what we saw for mean and standard deviations.If 0 is in the confidence

interval of r and r^2, the correlation is not significant (we can not say that there is a link between X and Y).But, this is not

because the correlation is significant that it has any interest (... use r^2).

- Examples

Example 1

Suppose a random sample of four pharmaceutical firms with follows research spending X and profits Y (in millions of dollars)

X	Y
40	50
40	60
30	40
50	50

Correlation coefficient ?

Example 1

Remarks :

1. The correlation coefficient gives us information on the existence of a linear relationship between the two considered variables.

A correlation coefficient of zero does not mean the absence of any relationship between the two variables. There may be a nonlinear relationship between them.

2. Do not confuse correlation and causality.

A good correlation between two variables can reveal a causal relationship between them, but not necessarily.

Example 1

”Correlation does not imply causation” is a phrase used in science and statistics to emphasize that a correlation between two variables does not necessarily imply that one causes the other. The opposite assumption, that correlation proves causation, is one of several questionable cause logical fallacies (error in reasoning) by which two events that occur together are taken to have a cause-and-effect relationship.
Examples of illogically inferring causation from correlation : Since the 1950s, both the atmospheric CO2 level and obesity levels have increased sharply. Hence, atmospheric CO2 causes obesity.
Richer populations tend to eat more food and consume more energy

Example 1

If we compare the life of individuals in the quantity of drugs for the heart they have absorbed, there will probably be a negative correlation. It would be unwise to conclude that taking heart drugs shortens the lives of individuals (in fact, in this case, the correlation index is a common cause of heart disease).