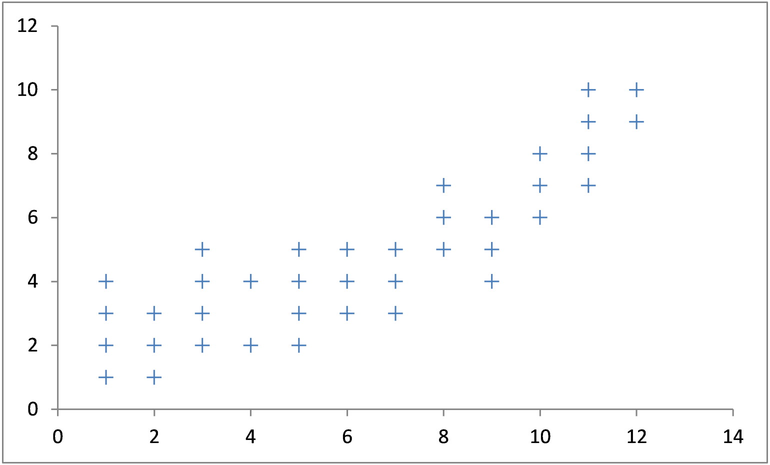
Let $\mathcal{P}$ be a population of total size n, on which we study two quantitative characteristics X and Y, we are interested in the relation between these two variables.

We start by defining the double statistical series of $\mathcal{P}$ for the characters X and Y

$$\begin{array}{ccc} \mathcal{P} & \longrightarrow & \mathbb{R}^2 \\ e_{ij} & \longmapsto & (x_i, y_j) \end{array}$$

A first idea to try to show the relation between X and Y is to plot the scatter plot associated with the double statistical series.



The scatter plot can have various shapes and these shapes will guide us in defining the notion of correlation.

- **First situation:** the Scatter plot can be formed of aligned points so X and Y are linked by a functional relation of the form $y = f(x)$.
- **Second situation:** the scatter plot is dispersed, the two observed values do not depend on each other, we say that the two characters X and Y are **independent**.
- **Third situation:** intermediate situation between independence and functional relationship.

X \ Y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_l	Marginal numbers
x_1	n_{11}	n_{12}		n_{1j}		n_{1l}	$n_{1\bullet}$
x_2	n_{21}	n_{22}		n_{2j}		n_{2l}	$n_{2\bullet}$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}		n_{ij}		n_{il}	$n_{i\bullet}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}		n_{kj}		n_{kl}	$n_{k\bullet}$
Marginal numbers	$n_{\bullet 1}$	$n_{\bullet 2}$	$\cdots$	$n_{\bullet j}$	$\cdots$	$n_{\bullet l}$	n

n_{ij} is the partial numbers of the couple (x_i, y_j) and $n = \sum_{i=1}^{k} \sum_{j=1}^{l} n_{ij}$
$n_{i\bullet}$ is the marginal numbers of x_i and $n_{i\bullet} = \sum_{j=1}^{l} n_{ij}$
$n_{\bullet j}$ is the marginal numbers of y_j and $n_{\bullet j} = \sum_{i=1}^{k} n_{ij}$.
$n = \sum_{i=1}^{k} n_{i\bullet} = \sum_{j=1}^{l} n_{\bullet j}$

Definition

The couple (X, Y) is statistically independent if we have $\forall i = 1, \cdots, k; j = 1, \cdots, l$

$$f_{ij} = \frac{n_{ij}}{n} = f_{i\bullet} \times f_{\bullet j} = \frac{n_{i\bullet}}{n} \times \frac{n_{\bullet j}}{n}$$

Definition

We call covariance of the variables X and Y and we note $Cov(X, Y)$, the number

$$\begin{aligned} Cov\left(X, Y\right) &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}\left(x_i - \overline{X}\right)\left(y_j - \overline{Y}\right) \\ &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \overline{XY}. \end{aligned}$$

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \; et \; \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j .$$

Remark

1. *The marginal means $\overline{X}$ and $\overline{Y}$ are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \ \ et \ \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

2. *If the variables X and Y are statistically independant then $Cov\left(X, Y\right) = 0$.*

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \ \text{et}\ \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

2. *If the variables* X *and* Y *are statistically independant then* $Cov\,(X, Y) = 0.$ *But the reciprocal is not true*

Remark

1. *The marginal means* $\overline{X}$ *and* $\overline{Y}$ *are given by*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{k} n_{i\bullet}x_i \ et \ \overline{Y} = \frac{1}{n}\sum_{j=1}^{l} n_{\bullet j}y_j.$$

2. *If the variables* X *and* Y *are statistically independant then* $Cov\left(X, Y\right) = 0$. *But the reciprocal is not true*

Definition

If $Cov(X, Y) = 0$ we say that the variables X and Y are uncorrelated.

Definition

We call the linear coefficient of correlation of the variables X and Y the number

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

1. The linear coefficient of correlation is invariant by change of origin and unit of measurement.
2. We have $-1 \leq \rho(X, Y) \leq 1$
3. If $\rho(X, Y) = 0$, the variables X and Y are uncorrelated.

Remark

If $\rho(X, Y) > 0$: X and Y evolve in the same direction

Definition

We call the linear coefficient of correlation of the variables X and Y the number

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}.$$

1. The linear coefficient of correlation is invariant by change of origin and unit of measurement.
2. We have $-1 \leq \rho(X, Y) \leq 1$
3. If $\rho(X, Y) = 0$, the variables X and Y are uncorrelated.

Remark

If $\rho(X, Y) > 0$: X and Y evolve in the same direction
If $\rho(X, Y) < 0$: X and Y evolve in the opposite direction

Definition

We call the linear coefficient of correlation of the variables X and Y the number

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

1. The linear coefficient of correlation is invariant by change of origin and unit of measurement.
2. We have $-1 \leq \rho(X,Y) \leq 1$
3. If $\rho(X,Y) = 0$, the variables X and Y are uncorrelated.

Remark

If $\rho(X,Y) > 0$: X and Y evolve in the same direction
If $\rho(X,Y) < 0$: X and Y evolve in the opposite direction
If $\rho(X,Y)$ is $\pm$ near of 1 : the correlation will be very good.

Determine all the parameters for X and Y from the following contingency table

X \ Y	5	7	9	11	13	$n_{i\bullet}$
1				1	4	5
2			2	7	1	10
4			9	1		10
6	2	8	6	1		17
9	5	2	1			8
$n_{\bullet j}$	7	10	18	10	5	50

$X \backslash Y$	5	7	9	11	13	$n_{i.}$	$n_{i.}x_i$	$n_{i.}x_i^2$	$\sum_j n_{ij}x_iy_j$
1				1	4	5	5	5	63
2			2	7	1	10	20	40	216
4			9	1		10	40	160	368
6	2	8	6	1		17	102	612	786
9	5	2	1			8	72	648	432
$n_{.j}$	7	10	18	10	5	**50**	**239**	**1465**	**1865**
$n_{.j}y_j$	35	70	162	110	65	**442**			
$n_{.j}y_j^2$	175	490	1458	1210	845	**4178**			
$\sum_i n_{ij}x_iy_j$	285	462	765	275	78	**1865**			

$$\begin{aligned}
\overline{X} &= \frac{239}{50} = 4.78;\ \overline{Y} = \frac{442}{50} = 8.84 \\
\sigma_X &= \sqrt{\frac{1465}{50} - 4.78^2} = \sqrt{6.4516} = 2.54 \\
\sigma_Y &= \sqrt{\frac{4178}{50} - 8.84^2} = \sqrt{5.4144} \approx 2.3269 \\
Cov\,(X, Y) &= \frac{1865}{50} - 4.78 \times 8.84 = -4.9552 \\
\rho\,(X, Y) &= \frac{-4.9552}{2.54 \times 2.3269} \approx -0.8384.
\end{aligned}$$

Since $\rho\,(X, Y) < 0$, then X and Y evolve in the opposite direction.

To have a general idea on the relation between two characters we study the conditional distributions, for that we are interested in the couples (x_i, y_j) where we fix one of the variables.

Definition

We call the conditional mean of the variable Y knowing x_i, the real number

$$\overline{Y}/x_i = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} y_j.$$

Definition

We call the regression curve of Y in X the broken curve that connects the points $\left(x_i, \overline{Y}/x_i\right)$.

In the same way we can construct the regression curve of X in Y.

Definition

We call the conditional mean of the variable X knowing y_j, the real number

$$\overline{X}/y_j = \frac{1}{n_{\cdot j}} \sum_{i=1}^{k} n_{ij} x_i.$$

Definition

We call the regression curve of X in Y the broken curve that connects the points $\left(\overline{X}/y_j, y_j\right)$.

We can also determine the conditional variance .

Definition

We call the conditional variance of the vairable Y knowing x_i, the real number

$$\sigma^2_{Y/x_i} = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} \left(y_j - \overline{Y}/x_i\right)^2 = \frac{1}{n_{i.}} \sum_{j=1}^{l} n_{ij} y_j^2 - \left(\overline{Y}/x_i\right)^2 .$$

Definition

We call th conditional variance of the vairable X knowing y_j, the real number

$$\sigma^2_{X/y_j} = \frac{1}{n_{.j}} \sum_{i=1}^{k} n_{ij} \left(x_i - \overline{X}/y_j\right)^2 = \frac{1}{n_{.j}} \sum_{i=1}^{l} n_{ij} x_i^2 - \left(\overline{X}/y_j\right)^2 .$$

We will use the conditional variance or conditional standard deviation to measure the strength of the relationship of Y with X, by drawing the broken curves connecting the points $\left(x_i, \overline{Y}/x_i - \sigma_{Y/x_i}\right)$ et $\left(x_i, \overline{Y}/x_i + \sigma_{Y/x_i}\right)$.

The band between these two curves is called the regression corridor of Y in X.

In the same way we can draw the regression corridor of X in Y by joining the points $\left(\overline{X}/y_j - \sigma_{X/y_j}, y_j\right)$ and $\left(\overline{X}/y_j + \sigma_{X/y_j}, y_j\right)$.