Chapter 2

Bivariate Statistical Series

0.1 Introduction

In the preceding chapter, we discussed methods to summarize and represent information related to a single variable. However, an individual can be analyzed using multiple characteristics (or variables). For example, examining employees through their seniority and education level, or monitoring a child's growth by observing both weight and height. Subsequently, we introduce the comprehensive study of relationships between two variables, specifically limiting our discussion to the case of two variables. Consider a population where

$$Z : \Omega \to \mathbb{R}^2$$

is defined by

$$w \mapsto Z(w) = (X(w), Y(w))$$

or directly as

$$(X, Y) : \Omega \to \mathbb{R}^2$$

with

$$w \mapsto (X(w), Y(w))$$

.

Definition 1. *In this context, Z is considered a two-dimensional statistical variable with $Card(\Omega) = N$, where N is a finite integer. The pair (X, Y) is referred to as the statistical variable pair.*

Example 1.

- *In a sample of 200 households, we observe simultaneously the number of children X and the number of rooms Y.*
- *In a sample of 20 households, the monthly income X in Dinars and the monthly expenditures Y are observed.*
- *Among randomly selected students from a second-year civil engineering class, we observe the scores for 'math3' X and statistics Y.*
- *A company conducts a study on the relationship between monthly advertising expenditures X and the volume of sales Y achieved.*

0.2 Representation of Bivariate Statistical Series

Bivariate statistical series can be presented in two ways.

0.2.1 First Presentation

For each w_i, we associate the pair (x_i, y_i), i.e., $w_i \mapsto (x_i, y_i)$. The data is arranged as shown in the following table:

We will consistently use the following notations: $x_i := X(w_i)$ and $y_i := Y(w_i)$.

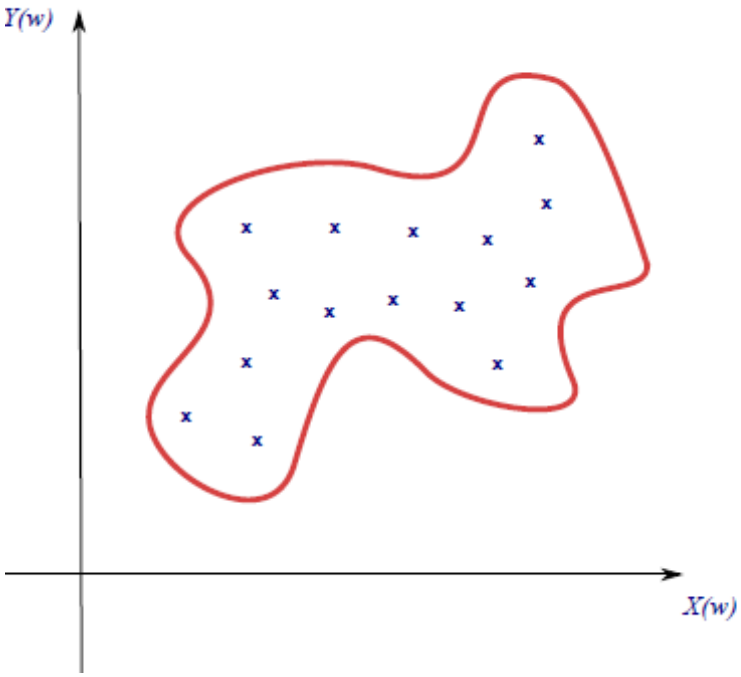
w	w_1	w_2	...	w_N
Variable X	$X(w_1)$	$X(w_2)$	...	$X(w_N)$
Variable Y	$Y(w_1)$	$Y(w_2)$	...	$Y(w_N)$

Example 2. *Consider a set of 8 students. We have the following table:*

w	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
$X(w)$	8	2	6	6	11	10	7	2
$Y(w)$	9	10	11	7	14	16	12	5

Here, X represents the number of hours each student spent preparing for the statistics exam, and Y represents the score out of 20 obtained by each student on the exam.

Remark 0.2.1. *We can translate the associated table into a figure known as a "scatter plot" or "scatter diagram". This representation is achieved by plotting each pair of observations (x_i, y_j) as a point on a cartesian coordinate system.*



0.2.2 Second Presentation

Let the statistical variable Z be given by the pair (X, Y). Let $x_1, \ldots, x_k$ and $y_1, \ldots, y_l$ be the values taken by X and Y respectively. In this case, we define the values of Z as follows:

for i ranging from 1 to k and for j ranging from 1 to l,

$$z_{ij} := (x_i, y_j).$$

The statistical variable Z takes $k \times l$ values. During this study, we use the following double-entry table (or contingency table) whether discrete or continuous:

Table 1: Contingency Table of X and Y

X \Y	**[L'1,L'2[or y1**	**...**	**[L'i,L'i+1[or yi**	**...**	**Marginale X**
[L1,L2[ou x1	n_{11} or f_{11}	...	n_{1j} or f_{1j}	...	$n_{1\cdot}$ or $f_{1\cdot}$
...	...	...	...	...	...
[Li,Li+1[ou xi	n_{i1} or f_{i1}	...	n_{ij} or f_{ij}	...	$n_{i\cdot}$ or $f_{i\cdot}$
...	...	...	...	...	...
[Lk,Lk+1[ou xk	n_{k1} or f_{k1}	...	n_{kj} or f_{kj}	...	$n_{k\cdot}$ or $f_{k\cdot}$
Marginale Y	$n_{\cdot 1}$ or $f_{\cdot 1}$	...	$n_{\cdot j}$ or $f_{\cdot j}$	...	N

$$N = \sum_{j=1}^{l}\sum_{i=1}^{k} n_{ij}, = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}.$$

Remark 0.2.2. *The calculation or development of this double series is given by*

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = n_{11} + n_{12} + n_{13} + \cdots + n_{1l}$$

$$+n_{21} + n_{22} + n_{23} + \cdots + n_{2l} + \cdots + n_{k1}$$

$$+n_{k2} + n_{k3} + \cdots + n_{kl}.$$

Remark 0.2.3. *We have the following property,*

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = 1.$$

0.3 Marginal Laws

On the margin of the contingency table, we can extract data solely with respect to X and solely with respect to Y (see the previously established contingency table).

1. **Marginal frequencies with respect to** Y: For each column $j = 1, \ldots, l$,

$$n_{\cdot j} := \sum_{i=1}^{k} n_{ij},$$

and

$$f_{\cdot j} := \frac{n_{\cdot j}}{N} = \sum_{i=1}^{k} f_{ij}.$$

2. **Marginal frequencies with respect to** X: For each row $i = 1, \ldots, k$,

$$n_{i\cdot} := \sum_{j=1}^{l} n_{ij},$$

and

$$f_{i\cdot} := \frac{n_{i\cdot}}{N} = \sum_{j=1}^{l} f_{ij}.$$

Remark 0.3.1. *We have the following properties:*

- $\sum_{i=1}^{k} n_{i\cdot} = \sum_{j=1}^{l} n_{\cdot j} = N$,
- $\sum_{i=1}^{k} f_{i\cdot} = \sum_{j=1}^{l} f_{\cdot j} = 1$.

0.4 Numerical Description

In the case of a two-dimensional statistical variable X and Y, the means are given respectively by:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{k} n_{i\cdot} x_i = \sum_{i=1}^{k} f_{i\cdot} x_i \quad \text{(mean of } X\text{)},$$

and

$$\bar{Y} = \frac{1}{N} \sum_{j=1}^{l} n_{\cdot j} y_j = \sum_{j=1}^{l} f_{\cdot j} y_j \quad \text{(mean of } Y\text{)}.$$

In the continuous case, x_i and y_j represent respectively the center of the classes for X and Y, that is,

$$x_i = \frac{L_{i+1} + L_i}{2} \quad \text{and} \quad y_j = \frac{L_{j+1} + L_j}{2}.$$

We now define the variance of X and the variance of Y as follows,

$$\mathrm{Var}(X) := x^2 - \bar{X}^2, \quad \text{with} \quad x^2 := \frac{1}{N}\sum_{i=1}^{k} n_{i.}x_i^2 = \sum_{i=1}^{k} f_{i.}x_i^2,$$

$$\mathrm{Var}(Y) := y^2 - \bar{Y}^2, \quad \text{with} \quad y^2 := \frac{1}{N}\sum_{j=1}^{l} n_{.j}y_j^2 = \sum_{j=1}^{l} f_{.j}y_j^2.$$

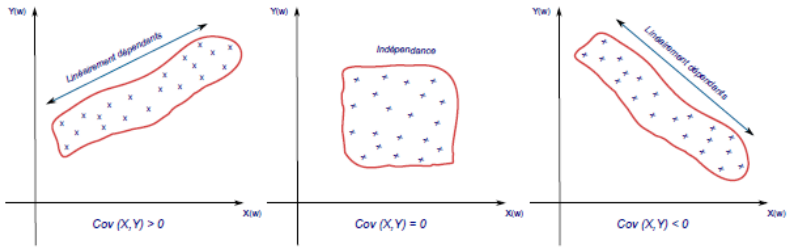
The standard deviations of X and Y are given, respectively, by

$$\sigma_X := \sqrt{\mathrm{Var}(X)} \quad \text{and} \quad \sigma_Y := \sqrt{\mathrm{Var}(Y)}.$$

0.5 Notion of Covariance

COVARIANCE.PNG

We denote by $\mathrm{Cov}(X, Y)$ the covariance between the variables X and Y. Covariance is a parameter that indicates the variability of X in relation to Y.



Covariance is calculated by the following expression

$$\mathrm{Cov}(X, Y) = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}x_i y_j - \bar{X}\bar{Y}$$

0.6 Linear Regression (Adjustment)

In cases where a significant linear relationship can be demonstrated between two continuous quantitative variables X and Y (the outline of the point cloud is stretched in one direction), one might seek to formalize the average relationship linking these two variables using a linear equation that summarizes this relationship. This approach is referred to as linear adjustment.

0.6.1 Correlation Coefficien

Definition 2. *The quantity*

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

is called the correlation coefficient.

Remark 0.6.1. ***Correlation Coefficient***

The correlation coefficient r ranges between $[-1, 1]$, or equivalently,

$$|r| \leq 1.$$

The correlation coefficient r_{XY} measures the degree of linear association between X and Y. We observe the following two characteristics:

- *The closer the absolute value of r is to 1, the stronger the linear relationship between X and Y.*

- *The closer the absolute value of r is to 0, the weaker the linear relationship between X and Y.*

0.6.2 Regression Line

The idea is to transform a scatter plot into a line. This line should be as close as possible to each of the points. Therefore, the objective is to minimize the discrepancies between the points and the line.

For this purpose, the method of least squares is used. This method aims to explain a scatter plot by a line that relates Y to X, that is,

$$Y = aX + b,$$

such that the distance between the scatter plot and the line is minimized.

For the regression line equation, the coefficient a is calculated as:

$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)},$$

and the intercept b is calculated as:

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