ELEMENTARY STATISTICS COURSE

Alfredo Sánchez Alberca (asalber@ceu.es)

Feb 2016

Department of Applied Math and Statistics
CEU San Pablo

CEU Universidad San Pablo

LICENSE TERMS

CONTENTS

1. Regresión y Correlación

REGRESIÓN Y CORRELACIÓN

1. Regresión y Correlación

1.1 Joint frequency distribution

1.2 Covariance

In the last chapter we saw how to describe the distribution of a single variable in a sample. However, in most cases, studies require to describe several variables that are often related. For instance, a nutritional study should consider all the variables that could be related to the weight, as height, age, gender, smoking, diet, physic exercise, etc.

To understand a phenomenon that involve several variables is not enough to study every variable by its own. We have to study all the variables together to describe how they interact and the type of relation among them.

Usually in a *dependency study* there is a **dependent variable** Y that it is supposed to be influenced by a set of variables $X_1, \dots, X_n$ known as **independent variables**. The simpler case is a *simple dependency study* when there is only one independent variable, that is the case covered in this chapter.

To study the relation between two variables X and Y, we have to study the joint distribution of the **two-dimensional variable** (X, Y), whose values are pairs (x_i, y_j) where the first element is a value of X and the second a value of Y.

Definition (Joint sample frequencies)

Given a sample of n values and a two-dimensional variable (X, Y), for every value of the variable (x_i, y_j) is defined

- **Absolute frequency** n_{ij}: Is the number of times that the pair (x_i, y_j) appears in the sample.
- **Relative frequency** f_{ij}: Is the proportion of times that the pair (x_i, y_j) appears in the sample.

$$f_{ij} = \frac{n_{ij}}{n}$$

Watch out! For two-dimensional variables it make no sense cumulative frequencies.

The values of the two-dimensional variable with their frequencies is known as **joint frequency distribution**, and is represented in a joint frequency table.

$X \backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_q
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{iq}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}

The height (in cm) and weight (in kg) of a sample of 30 students is:

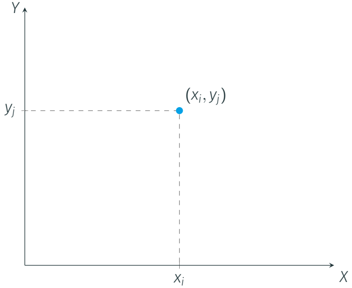
(179,85), (173,65), (181,71), (170,65), (158,51), (174,66),
(172,62), (166,60), (194,90), (185,75), (162,55), (187,78),
(198,109), (177,61), (178,70), (165,58), (154,50), (183,93),
(166,51), (171,65), (175,70), (182,60), (167,59), (169,62),
(172,70), (186,71), (172,54), (176,68),(168,67), (187,80).

The joint frequency table is

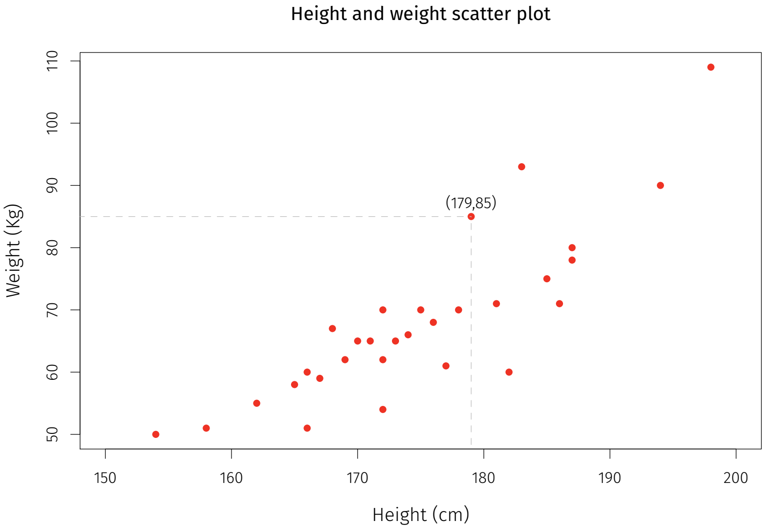
X/Y	$[50, 60)$	$[60, 70)$	$[70, 80)$	$[80, 90)$	$[90, 100)$	$[100, 110)$
$(150, 160]$	2	0	0	0	0	0
$(160, 170]$	4	4	0	0	0	0
$(170, 180]$	1	6	3	1	0	0
$(180, 190]$	0	1	4	1	1	0
$(190, 200]$	0	0	0	0	1	1

The joint frequency distribution can be represented graphically with a **scatter plot**, where data is displayed as a collections of points on a *XY* coordinate system.

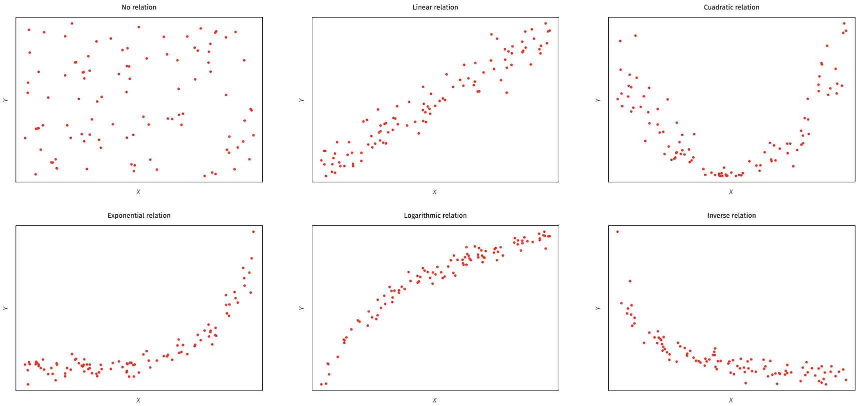
Usually the independent variable is represented in the X axis and the dependent variable in the Y axis. For every data pair (x_i, y_j) in the sample a dot is drawn on the plane with those coordinates.



The result is a set of points that usually is known as a *point cloud*.


Height and weight scatter plot
(179,85)
Weight (Kg)
50
60
70
80
90
100
110
150
160
170
180
190
200
Height (cm)

The shape of the point cloud in a scatter plot gives information about the type of relation between the variables.



The frequency distributions of each variable of the two-dimensional variable are known as **marginal frequency distributions**.

We can get the marginal frequency distributions from the joint frequency table summing frequencies by rows and columns.

$X\backslash Y$	y_1	$\cdots$	y_j	$\cdots$	y_q	n_x
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1q}	n_{x_1}
$\vdots$	$\vdots$	$\vdots$	$+$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$+$	n_{ij}	$+$	n_{iq}	n_{x_i}
$\vdots$	$\vdots$	$\vdots$	$+$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	n_{x_p}
n_y	n_{y_1}	$\cdots$	n_{y_j}	$\cdots$	n_{y_q}	n

The marginal frequency distributions for the previous sample of heights and weights are

X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$	n_x
$(150,160]$	2	0	0	0	0	0	2
$(160,170]$	4	4	0	0	0	0	8
$(170,180]$	1	6	3	1	0	0	11
$(180,190]$	0	1	4	1	1	0	7
$(190,200]$	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

and the corresponding statistics are

$$\bar{x} = 174.67 \text{ cm} \qquad s_x^2 = 102.06 \text{ cm}^2 \qquad s_x = 10.1 \text{ cm}$$
$$\bar{y} = 69.67 \text{ Kg} \qquad s_y^2 = 164.42 \text{ Kg}^2 \qquad s_y = 12.82 \text{ Kg}$$

To study the relation between two variables, we have to analyze the joint variation of them.