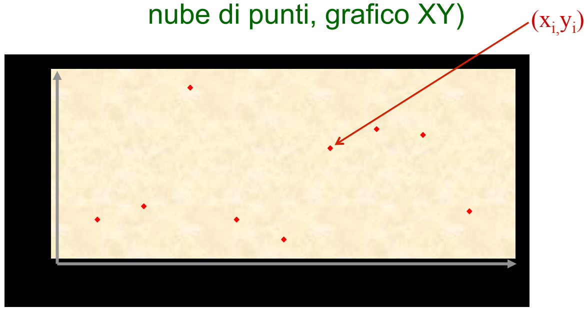
La Standardizzazione di variabili

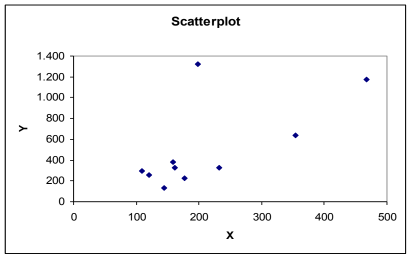
In caso di variabili quantitative possiamo rappresentare graficamente la variabile doppia con
uno SCATTER PLOT (diagramma a dispersione, nube di punti, grafico XY)



Scatterplot **o diagramma a dispersione**

- Analisi grafica della relazione tra due variabili numeriche

osserv.	(X)	(Y)
1	121	260
2	109	292
3	233	323
4	199	1.320
5	354	640
6	145	135
7	467	1.176
8	177	225
9	161	326
10	158	378



L'uso di variabili standardizzate

- Siano X e Y due variabili numeriche (non necessariamente con la stessa scala di misura)
- La standardizzazione fa si che abbiamo

$$Z(X) = \frac{X - \mu_X}{\sigma_X} \qquad Z(Y) = \frac{Y - \mu_Y}{\sigma_Y}$$

Variabile standardizzata

$$X = x_1, x_2, x_3, \ldots, x_n \Rightarrow Z = Z(X)$$

$$Z(X) = \frac{X - \mu}{\sigma}$$

X	$Z = Z(X)$
x_1	$\frac{x_1 - \mu}{\sigma} = z_1$
x_2	$\frac{x_2 - \mu}{\sigma} = z_2$
x_3	$\frac{x_3 - \mu}{\sigma} = z_3$

Proprietà

- $\mu = 0$

$$\mu = \frac{1}{n}\sum z_i = \frac{1}{n}\left(\frac{x_i - \mu}{\sigma}\right) = \frac{1}{n}\frac{1}{\sigma}\sum (X_i - \mu) = \frac{0}{n\sigma} = 0$$

- $\sigma^2 = 1$

$$\sigma^2 = \frac{1}{n}(z_i - \mu_z)^2 = \frac{1}{n}\sum z_i^2 = \frac{1}{n}\left(\frac{x_i - \mu}{\sigma}\right)^2 = \frac{1}{n}\frac{1}{\sigma^2}\sum (x_i - \mu)^2 = \frac{1}{\sigma^2}\sigma^2 = 1$$

Come si standardizzano i dati

(X)	(Y)
121	260
109	292
233	323
199	1.320
354	640
145	135
467	1.176
177	225
161	326
158	378

$$Z(X) = \frac{X - \mu_X}{\sigma_X}$$

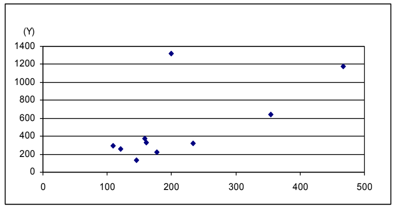
$$Z(Y) = \frac{Y - \mu_Y}{\sigma_Y}$$

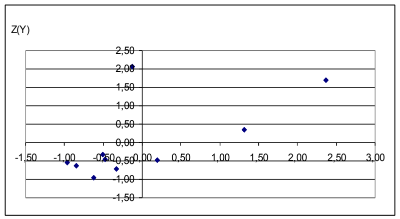
Media X =	212,4
Media Y =	507,5
Varianza X =	11592
Varianza Y =	153462
Sqm X =	107,67
Sqm Y =	391,74

Z(X)	Z(Y)
-0,85	-0,63
-0,96	-0,55
0,19	-0,47
-0,12	2,07
1,32	0,34
-0,63	-0,95
2,36	1,71
-0,33	-0,72
-0,48	-0,46
-0,51	-0,33

$$\frac{225 - 507{,}5}{391{,}74}$$

$$\frac{121 - 212{,}4}{107{,}67}$$





Se rappresentiamo graficamente sia le distribuzioni originali X e Y

che quelle standardizzate Z(X) e Z(Y)

notiamo che la relazione tra i dati non muta ma cambia Il centro di riferimento (le medie di Z(X) e Z(Y) sono pari a 0)