CENNI DI STATISTICA DESCRITTIVA BIVARIATA – DISTRIBUZIONI MARGINALI RETTA DI REGRESSIONE

Angela Donatiello

Fino ad ora ci siamo occupati di statistica univariata, ossia di analisi di dati provenienti dalla rilevazione di un singolo carattere su una data popolazione. L'indagine statistica può però estendersi alla rilevazione di più caratteri contemporaneamente, ad esempio, peso, età, città di nascita, sesso, etc… di una data popolazione di individui, in tal caso si parlerà di **statistica multivariata**.

In particolare, la **statistica bivariata** si occupa dell'analisi dei dati che si ottengono quando vengono rilevati *congiuntamente* due caratteri diversi.

L'obiettivo principale della statistica bivariata è quello di *mettere in luce le relazioni che intercorrono fra due caratteri rilevati sullo stesso collettivo di unità statistiche*.

Analizzeremo come:

- ✓ rappresentare due caratteri congiuntamente
- ✓ costruire tabelle a doppia entrata
- ✓ studiare alcuni possibili legami esistenti fra due caratteri

Distribuzioni congiunte e marginali

Supponiamo di aver osservato due caratteri X e Y su ciascuna delle n unità statistiche che compongono la popolazione.

Il risultato potrà essere prima sintetizzato in una **tabella dei dati grezzi**

Unità statistiche	Modalità di X rilevata	Modalità di Y rilevata
1	x_1	y_1
2	x_2	y_2
…	…	…
n	x_n	y_n

Esempio. Su una popolazione formata da 5 persone sono stati rilevati due caratteri: età (X) e città di nascita (Y)

Nome	Età (in anni)	Città di nascita
Giovanni	30	Milano
Andrea	35	Torino
Maria	32	Milano
Francesca	30	Milano
Amelia	32	Roma

Tali dati possono però essere organizzati in modo da renderli più leggibili e poter trarre delle informazioni dai dati stessi. E' utile a tal proposito costruire una **tabella a doppia entrata** in cui verranno riportate le **frequenze congiunte** con cui si presentano le coppie di modalità osservate sui caratteri studiati.

Carattere X presentato con le modalità $x_1, \ldots, x_k$
Carattere Y presentato con le modalità $y_1, \ldots, y_h$

Si costruisca una matrice con k + 1 righe e h + 1 colonne e si riportino:

- ✓ nella prima colonna le modalità $x_1, \ldots, x_k$ del carattere X;
- ✓ nella prima riga le modalità $y_1, \ldots, y_h$ del carattere Y
- ✓ nella casella di incrocio tra le modalità (i – esima riga e j – esima colonna) la frequenza assoluta con cui si presentano congiuntamente le modalità (x_i, y_j), indicata con il simbolo $f(x_i, y_j)$

Y / X	y_1	…	y_j	…	y_h
x_1	…	…	…	…	…
…	…	…	…	…	…
x_i	…	…	$f(x_i,y_j)$	…	…
…	…	…	…	…	…
x_k	…	…	…	…	…

Nell'esempio considerato

X \ Y	Milano	Torino	Roma
30	2	0	0
32	1	0	0
35	0	1	0

Tale tabella va completata con l'aggiunta di un'ultima riga e di un'ultima colonna dove si riportano le somme delle frequenze presenti su ciascuna riga (o colonna). Questa riga e questa colonna rappresentano le distribuzioni marginali dei due caratteri, ovvero le frequenze registrate nel caso in cui ogni carattere fosse stato rilevato singolarmente. Tali frequenze $f(x_1)$, ..., $f(x_k)$ e $f(y_1)$, ... , $f(y_h)$ vengono dette **frequenze marginali** di X e Y.

X \ Y	y_1	...	y_j	...	y_h	Totale
x_1	...	...	...	...	...	$f(x_1)$
...	...	...	...	...	...	...
x_i	...	...	$f(x_i,y_j)$	...	...	...
...	...	...	...	...	...	...
x_k	...	...	...	...	...	$f(x_k)$
Totale	$f(y_1)$	...	...	...	$f(y_h)$	n

Distribuzione marginale di X

Distribuzione marginale di Y

numero complessivo di unità del collettivo

Sono invece dette frequenze marginali relative i rapporti tra le frequenze marginali e il numero complessivo di unità del collettivo.

Distribuzioni condizionate

Se si fissa una singola modalità x_i (oppure y_j) , ossia si fissa l'attenzione su una singola riga (oppure su una singola colonna), ciò significa restringere l'analisi ad una sottopopolazione che presenta la modalità x_i (o y_j). Ciascuna di queste righe (colonne) è pertanto una distribuzione condizionata.

Y / X	Milano	Torino	Roma
30	2	0	0
32	1	0	0
35	0	1	0

Distribuzione condizionata di Y rispetto a x_1

X	Distribuzione condizionata $X\|y_1$	Distribuzione condizionata relativa
30	2	2/3
32	1	1/3
35	0	0
Totale	3	1

Indipendenza statistica

Def. Il carattere X si dirà **indipendente** dal carattere Y se le distribuzioni **condizionate relative** di X rispetto alla modalità di Y è uguale alla distribuzione **marginale relativa** di X.

Cosa vuol dire?

Se la distribuzione condizionata relativa di X rispetto a Y è uguale a quella marginale relativa, significa che il condizionamento del carattere X alle modalità del carattere Y è irrilevante.

Se i caratteri non sono indipendenti si dice che esiste tra essi una **connessione.**

In generale si prova che:

Teorema. Due caratteri X e Y, di cui sono state osservate le modalità $x_1, \ldots, x_k$ e $y_1, \ldots, y_h$ su una popolazione costituita da n unità, si dicono **indipendenti** se e solo se

$$f(x_i, y_j) = \frac{f(x_i) f(y_j)}{n} \qquad \forall i = 1,\ldots,k \quad \forall j = 1,\ldots,h$$

Ogni frequenza congiunta è uguale al prodotto delle frequenze marginali, diviso n.

L'indipendenza statistica è un **concetto simmetrico**: se X è statisticamente indipendente da Y anche Y è statisticamente indipendente da X e viceversa e su tale concetto si basa la dimostrazione del teorema suddetto.

- ✓ $\frac{f(x_i)f(y_j)}{n}$ sono dette frequenze teoriche di indipendenza
- ✓ Per verificare l'indipendenza dei caratteri si affianca alla tabella a doppia entrata delle frequenze assolute congiunte osservate una tabella delle frequenze teoriche di indipendenza. Affinché ci sia indipendenza le due tabelle devono coincidere.

Esempio.

X \ Y	y_1	y_2	Totale
x_1	15	10	25
x_2	7	8	15
x_3	8	2	10
Totale	30	20	50

Tabella osservata

X \ Y	y_1	y_2	Totale
x_1	$\frac{30\cdot 25}{50}=15$	$\frac{20\cdot 25}{50}=10$	25
x_2	$\frac{30\cdot 15}{50}=9$	$\frac{20\cdot 15}{50}=6$	15
x_3	$\frac{30\cdot 10}{50}=6$	$\frac{20\cdot 10}{50}=4$	10
Totale	30	20	50

Tabella teorica di indipendenza

Le due tabelle sono diverse, pertanto non c'è indipendenza, ma c'è connessione.

Per **valutare il grado di connessione** si costruisce la tabella delle contingenze facendo la differenza fra la tabella originale e quella di indipendenza.

Contingenza: $c(x_i, y_j) = f(x_i, y_j) - \frac{f(x_i) \cdot f(y_j)}{n} = f(x_i, y_j) - f^*(x_i, y_j)$

misura di quanto si discosta la tabella osservata da quella teorica

Osservazione: la somma delle contingenze è sempre nulla (può essere utile nell'applicazione pratica)

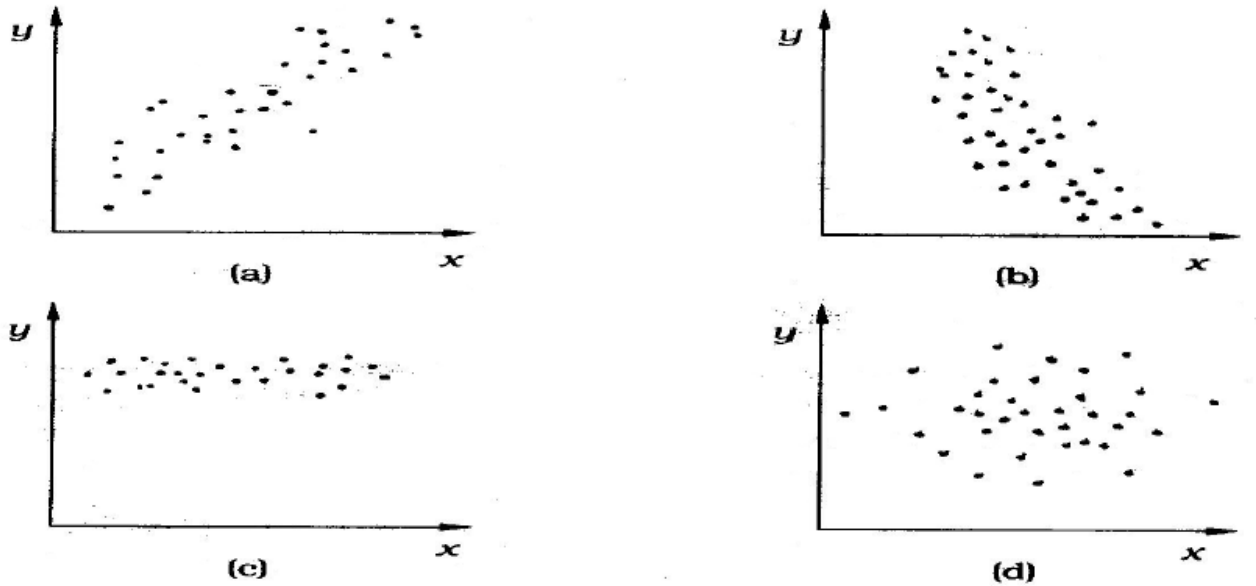
Indice chi – quadro o di Pearson: $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{c^2(x_i, y_j)}{f^*(x_i, y_j)}$

sintetizza in un unico indice tutte le contingenze

Si dimostra che: $\chi^2 = n\left(\sum_{i=1}^{k} \sum_{j=1}^{h} \frac{f^2(x_i, y_j)}{f(x_i) f(y_j)} - 1 \right)$ $\qquad \chi^2 \text{normalizzato} = \frac{\chi^2}{n \cdot \min\{k-1; h-1\}}$

Correlazione e regressione

Nel caso di una popolazione di individui adulti, supponiamo di voler cercare una eventuale relazione tra pressione arteriosa ed età. Numeriamo gli individui della popolazione da 1 ad n e associamo all'i - esimo individuo la coppia ordinata di numeri (xi; yi), dove xi denota la sua età (misurata per es. in anni) e yi denota la sua pressione arteriosa (misurata per es. in mm di Hg). In un sistema di coordinate cartesiane del piano, ogni coppia (xi; yi) individua un punto P_i, e il complesso degli n punti forma una specie di "nube". A seconda delle coppie di grandezze prese in esame, questa nube può presentare delle regolarità più o meno appariscenti.



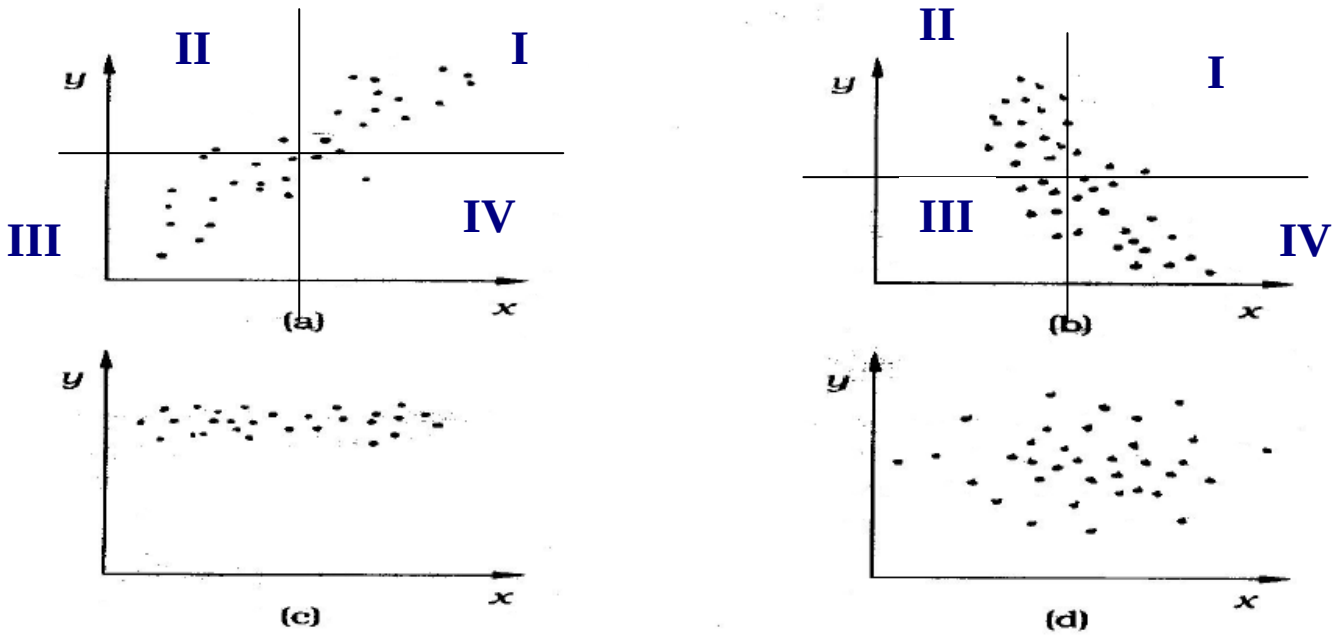
Se la nube è del tipo visualizzato in (a), si intuisce che al crescere dei valori di x anche i corrispondenti valori di y tendono a crescere (si parla allora di una concordanza o di una **correlazione positiva**); se invece la nube è del tipo visualizzato in (b), si intuisce che al crescere dei valori di x i corrispondenti valori di y tendono a diminuire (si parla allora di una discordanza o di una **correlazione negativa**); se la nube è del tipo visualizzato in (c), si intuisce che al crescere dei valori di x i valori di y si mantengono sostanzialmente costanti (si parla allora di **indifferenza** della y rispetto ad x). Infine, se la nube del tipo visualizzato in (d), si deve concludere che i dati a disposizione **non evidenziano alcuna correlazione** tra le due grandezze considerate.

Un indice utile per valutare la correlazione è la covarianza.

Covarianza: Siano X e Y due variabili statistiche rilevate congiuntamente su un collettivo di n unità. Siano $x_1, \ldots, x_n$ i valori osservati di X e e $y_1, \ldots, y_n$ i valori osservati di Y. Siano $\bar{x}$ e $\bar{y}$ le medie delle due variabili statistiche. Si chiama covarianza di X e Y, e si indica con il simbolo σ_{xy}, il numero definito come

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n}$$

Interpretiamo geometricamente la covarianza:



Il punto di intersezione delle parallele agli assi è ciò che può essere definito baricentro della "nuvola" di dati, in quanto è il punto di coordinate $(\bar{x}, \bar{y})$. Un dato generico avrà coordinate nel piano (x_i, y_i). Gli scarti $(x_i - \bar{x})$ e $(y_i - \bar{y})$ avranno segno che si distribuirà sui 4 quadranti individuati dalle rette parallele agli assi.

- ✓ Se la maggior parte dei prodotti degli scarti $(x_i - \bar{x})(y_i - \bar{y})$ sono positivi, allora i punti si collocheranno maggiormente nel I e III quadrante, la nuvola è perciò del tipo (a) e la **covarianza è positiva**

- ✓ Se la maggior parte dei prodotti degli scarti $(x_i - \bar{x})(y_i - \bar{y})$ sono negativi, allora i punti si collocheranno maggiormente nel II e IV quadrante, la nuvola sarà pertanto del tipo (b) e la **covarianza è negativa**

- ✓ Se i punti non sono correlati, in quanto sparpagliati (d) allora la **covarianza è nulla**

Formula abbreviata per il calcolo della covarianza

$$\sigma_{xy} = \frac{\sum_{i=1}^{n} x_i y_i}{n} - \bar{x} \cdot \bar{y}$$

Per stabilire se la correlazione è forte o debole si pone la covarianza a rapporto con il suo valore massimo.

Teorema. La covarianza di due variabili X e Y può assumere valori appartenenti all'intervallo

$$-\sigma_x\sigma_y \le \sigma_{xy} \le \sigma_x\sigma_y$$

dove σ_x e σ_y sono le deviazioni standard di X e Y.

Ricordiamo: deviazione standard di X

$$\sigma(X) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - M)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} s_i^2}{n}}$$

Coefficiente di correlazione lineare

Si chiama coefficiente di correlazione lineare di due variabili X e Y il numero definito come:

$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

- ✓ $-1 \le r \le 1$
- ✓ Stesso segno della covarianza, quindi r > 0 indica relazione lineare crescente, r < 0 indica relazione lineare decrescente, r quasi nullo indica che non c'è una relazione lineare
- ✓ $r = \pm 1$ allora c'è una perfetta relazione lineare

Calcolo del coefficiente di correlazione

In 4 supermercati sono stati valutati le superfici di esposizione (X) e il fatturato settimanale (Y)

X_i (m^2)	0,2	0,5	0,8	1
Y_i (migliaia di euro)	50	120	150	200

Determiniamo il coefficiente di correlazione

Xi	Yi	XiYi	Xi^2	Yi^2
0,2	50	10	0,04	2500
0,5	120	60	0,25	14400
0,8	150	120	0,64	22500
1	200	200	1	40000
som Xi	som Yi	som XiYi	som Xi^2	som Yi^2
2,5	520	390	1,93	79400

Calcoliamo tutti gli elementi utili:

$$\bar{x} = \frac{\sum x_i}{n} = 0{,}625 \qquad \bar{y} = \frac{\sum y_i}{n} = 130$$

Varianza della X $\sigma^2_x = \frac{\sum x_i^2}{n} - \bar{x}^2 = 0{,}091875$ Deviazione standard X $\sigma_x = \sqrt{\sigma^2_x}$

Varianza della Y $\sigma^2_y = \frac{\sum y_i^2}{n} - \bar{y}^2 = 2950$ Deviazione standard Y $\sigma_y = \sqrt{\sigma^2_y}$

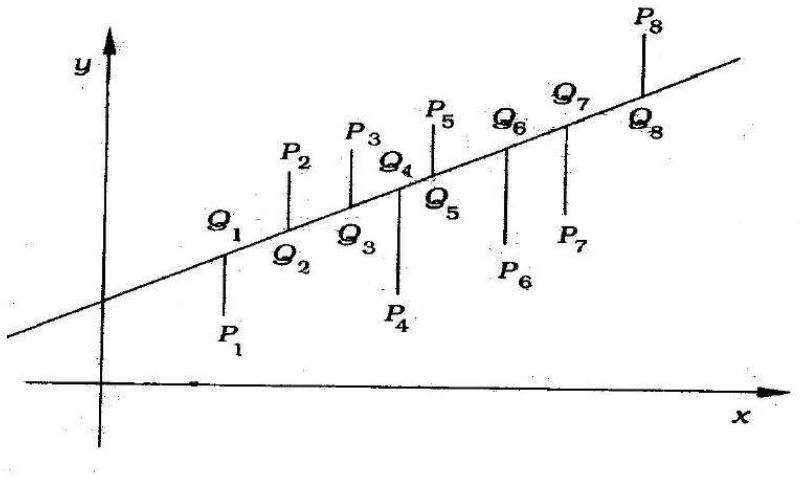
Covarianza $\sigma_{xy} = \frac{\sum x_i y_i}{n} - \bar{x} \cdot \bar{y} = 16{,}25$

Coefficiente di correlazione $r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = 0{,}987\ldots$

r è prossimo ad 1, pertanto c'è una forte correlazione lineare positiva

Retta di regressione

Quando si presume che tra due variabili x, y possa sussistere una relazione di dipendenza della y dalla x schematizzabile in termini matematici mediante una funzione lineare, si usa tracciare la cosiddetta retta di regressione, cioè la retta che meglio approssima la nube dei dati. Occorre però precisare ancora cosa si debba intendere per "migliore approssimazione".



1) Data una retta generica s, si congiungano i punti $P_i(x_i, y_i)$ della nube di dati con i punti $Q_i\ (x_i, y'_i)$ aventi stessa ascissa ma posti sulla retta

2) Si calcolino le lunghezze dei segmenti P_iQ_i.

$$P_iQ_i = |\, y_i - y'_i \,|$$

3)Si elevino al quadrato e se ne calcoli la somma $\sum_{i=1}^{n}(y_i - y'_i)^2$

Esista una posizione della retta che rende minima tale somma di quadrati, essa è detta appunto **retta dei minimi quadrati** o **retta di regressione**

$$y = a + bx$$

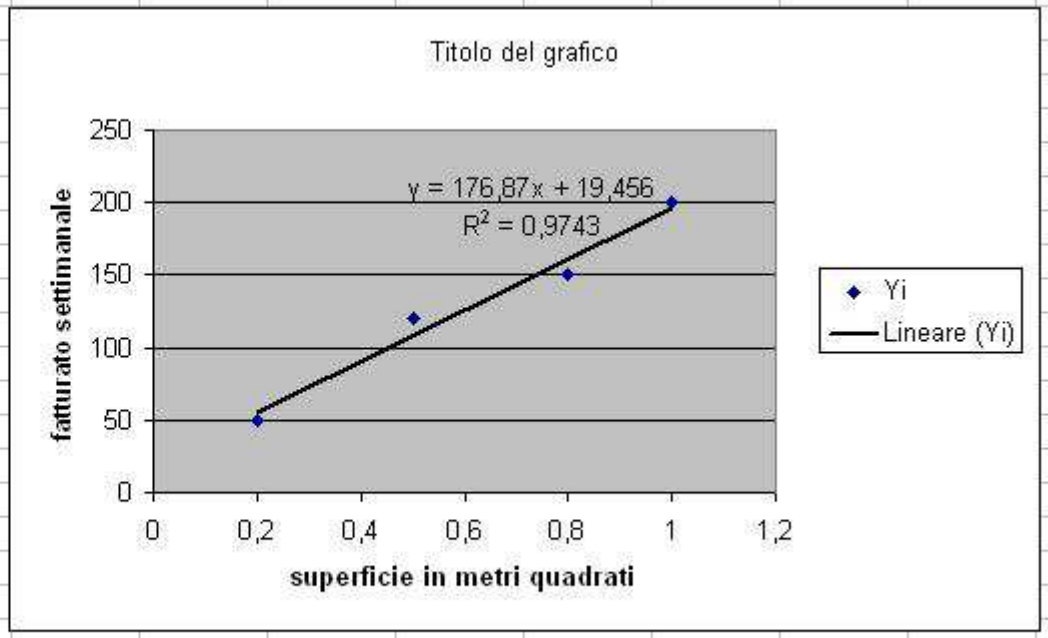
Con $$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad a = \overline{y} - b\overline{x}$$

In maniera più semplice si ha che:

date due variabili X e Y di valori medi $\overline{x}$ e $\overline{y}$, la **retta di regressione** che esprime Y in funzione della X è la retta che passa per il punto di coordinate ($\overline{x}$, $\overline{y}$) e che ha per coefficiente angolare m il coefficiente di regressione, ottenuto come rapprorto tra la covarianza e la varianza della X

$$y - \overline{y} = m(x - \overline{x}) \qquad m = \frac{\sigma_{xy}}{\sigma^2_x}$$

	A	B	C	D	E
1	Xi	0,2	0,5	0,8	1
2	Yi	50	120	150	200
3					
4	Xi	Yi	XiYi	Xi^2	Yi^2
5	0,2	50	10	0,04	2500
6	0,5	120	60	0,25	14400
7	0,8	150	120	0,64	22500
8	1	200	200	1	40000
9	som Xi	som Yi	som XiYi	som Xi^2	som Yi^2
10	2,5	520	390	1,93	79400
11					
12					
13	media X	0,625		media X ^2	0,390625
14	media Y	130		media Y ^2	16900
15	var X	0,091875		dev sta X	0,303109
16	Var Y	2950		dev sta Y	54,3139
17				sigmaXY	16,25
18				coeff corr	0,98706
19				r^2	0,974288
20					
21					
22	m=sigmaxy/sigmax^2		176,8707		



$$y = 176{,}87x + 19{,}46$$