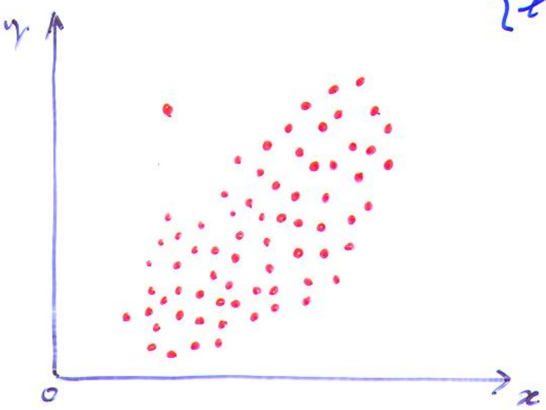
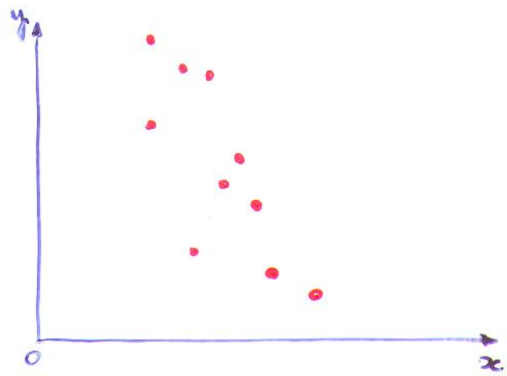
Note su regressione lineare

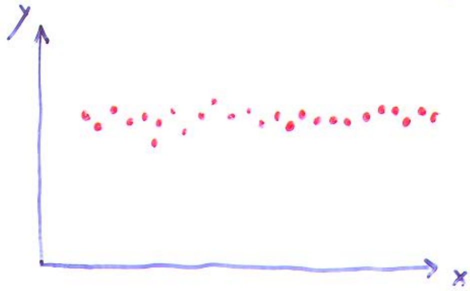
- Determinazione dell'**esistenza** di una **relazione funzionale** tra due G.F. tramite il **coefficiente di correlazione lineare** r

→ correlazione $\begin{Bmatrix}\text{positiva}\\\text{negativa}\end{Bmatrix}$ = al crescere di una G.F. $\begin{Bmatrix}\text{cresce anche l'altra G.F.}\\\text{l'altra G.F. diminuisce}\end{Bmatrix}$





→ correlazione nulla o assenza di correlazione ≈ al crescere di una G.F. l'altra varia in modo puramente casuale

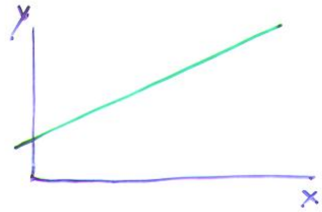


$$\boxed{r = \frac{\sum_{1}^{N}{}_i (x_i - \langle x \rangle)(y_i - \langle y \rangle)}{\sqrt{\left[\sum_{1}^{N}{}_i (x_i - \langle x \rangle)^2\right]\left[\sum_{1}^{N}{}_i (y_i - \langle y \rangle)^2\right]}}} = \frac{\sigma(x,y)}{\sigma(x)\cdot\sigma(y)}$$

$-1 \leq r \leq +1$ correlazione $\begin{cases}\text{positiva se } 0 < r \leq +1\\ \text{nulla » } r = 0\\ \text{negativa » } -1 \leq r < 0\end{cases} \Rightarrow \begin{cases} r \approx +1 \text{ ok}\\ r \approx 0 \text{ ?? TABELLE}\\ r \approx -1 \text{ ok}\end{cases}$

132

Coefficiente di correlazione nel caso di legame lineare tra le due variabili x, y



$$y = ax + b \qquad a = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$$

$$x = a'y + b' \qquad a' = \frac{N\sum xy - \sum x \sum y}{N\sum y^2 - (\sum y)^2}$$

$$a'y = x - b'$$

$$y = \frac{1}{a'}x - \frac{b'}{a'} \longrightarrow \frac{1}{a'} = a \rightarrow \boxed{aa' = 1}$$

nel caso di correlazione perfetta!

$$\sqrt{a \cdot a'} = \sqrt{\frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2} \cdot \frac{N\sum xy - \sum x \sum y}{N\sum y^2 - (\sum y)^2}} =$$

$$= \frac{N \cdot (\sum x \cdot y) - (\sum x)(\sum y)}{\sqrt{\left[N\sum x^2 - (\sum x)^2\right] \cdot \left[N\sum y^2 - (\sum y)^2\right]}} = \underbrace{\frac{\sum_1^N{}_i (x_i - \langle x\rangle)(y_i - \langle y\rangle)}{\sqrt{\left(\sum_1^N{}_i (x_i - \langle x\rangle)^2\right)\left(\sum_1^N{}_i (y_i - \langle y\rangle)^2\right)}}}_{\text{è proprio il coefficiente di correlazione } r}$$

$$\sum (x - \langle x\rangle)(y - \langle y\rangle) = \sum \left(xy - x\langle y\rangle - y\langle x\rangle + \langle x\rangle\langle y\rangle\right) =$$

$$= \sum xy - \frac{\sum y}{N}\sum x - \frac{\sum x}{N}\sum y + \frac{\sum x}{N}\frac{\sum y}{\cancel{N}} \cdot \cancel{N} =$$

$$= \sum xy - \frac{\sum x \sum y}{N} = \frac{1}{N}\left(N\sum xy - \sum x \sum y\right)$$

$$\sum (x - \langle x\rangle)^2 = \sum \left(x^2 + \langle x\rangle^2 - 2\langle x\rangle x\right) = \sum x^2 + N\left(\frac{\sum x}{N}\right)^2 - 2\frac{\sum x}{N}\sum x =$$

$$= \sum x^2 - \frac{(\sum x)^2}{N} = \frac{1}{N}\left(N\sum x^2 - (\sum x)^2\right)$$

$$\sum (y - \langle y\rangle)^2 = \ldots\ldots = \frac{1}{N}\left(N\sum y^2 - (\sum y)^2\right)$$

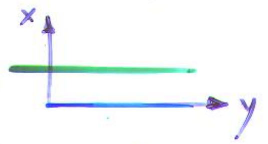
133

$$\boxed{r = \sqrt{aa'}}$$

± 1 nel caso di perfetta correlazione
poiché per una retta $aa' = 1$
in modo rigoroso $\begin{cases} y = ax + b \\ x = a'y + b' \end{cases}$

0 nel caso di scorrelazione perfetta
poiché $a = 0$ implica

$a' = 0$ implica



in entrambi i casi, al variare della x, la y non cambia.

- $N = 20$ con $r = 0{,}21$ c'è correlazione tra x e y?

$P_{20}(|r| \geq 0{,}2) = 40\%$ $\Rightarrow$ la frase: x, y sono scorrelati ha una probabilità del 40% di essere vera.

- $N = 20$ con $r = 0{,}60$

$P_{20}(|r| \geq 0.6) = 0{,}5\%$ $\Rightarrow$ la frase: x, y sono correlati ha una probabilità $100 - 0{,}5 = 99{,}5\%$ di essere vera.

Osservazione sul coefficiente di correlazione r :

$$\boxed{r = \frac{\sigma(x,y)}{\sigma(x)\cdot\sigma(y)}} = \frac{\sum_1^N{}_i (x_i - \langle x\rangle)(y_i - \langle y\rangle)}{\sqrt{\sum_1^N{}_i (x_i - \langle x\rangle)^2}\sqrt{\sum_1^N{}_i (y_i - \langle y\rangle)^2}}$$

- La covarianza $\sigma(x,y)$ è una misura della correlazione lineare tra due variabili casuali x ed y.

- Se variabili x, y perfettamente correlate in modo lineare

  $\rightarrow \quad |r| = 1$

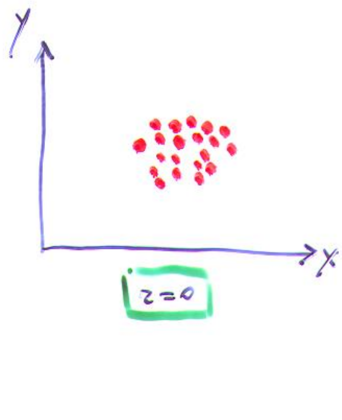
- Se variabili x, y perfettamente indipendenti

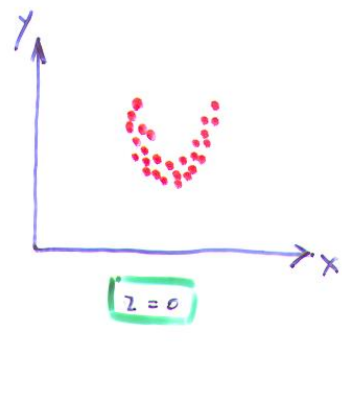
  $\rightarrow \quad r = 0$

- Se $r = 0$

  $\rightarrow$ x, y sono solo linearmente indipendenti

  infatti potrebbero essere correlate parabolicamente ($y = ax^2 + bx + c$) ed avrebbero ancora $r = 0$!!





135

APPENDICE C

Probabilità per i Coefficienti di Correlazione lineare

La bontà con cui N punti $(x_1, y_1), \ldots, (x_N, y_N)$ si adattano ad una linea retta è indicata dal coefficiente di correlazione lineare

$$\boxed{r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2\right]^{1/2}},}$$

che è sempre compreso nell'intervallo $-1 \leqslant r \leqslant 1$. Valori di r vicini a ± 1 indicano una buona correlazione lineare; valori vicini a 0 indicano poca o nessuna correlazione.

Una misura più quantitativa dell'adattamento si può trovare usando la Tabella C. Per ogni definito r_O, $\boxed{P_N(|r| \geqslant |r_O|)}$ è la probabilità che N misure di due variabili incorrelate diano un coefficiente r grande quanto r_O. Così se otteniamo un coefficiente r_O per cui $P_N(|r| \geqslant |r_O|)$ è piccola, allora è corrispondentemente improbabile che le nostre variabili siano incorrelate; cioè, è indicata una correlazione. In particolare, se $P_N(|r| \geqslant |r_O|) \leqslant 5$ percento, la correlazione è chiamata "significativa"; se è minore dell'1 percento, la correlazione è chiamata "altamente significativa".

Per esempio, la probabilità che 20 misure ($N = 20$) di due variabili incorrelate diano $|r| \geqslant 0.5$ è data nella tabella come 2.5 percento. Così se 20 misure danno $r = 0.5$, dovremmo avere evidenza "significativa" di una correlazione lineare tra le due variabili. Per ulteriore discussione, vedi Sezioni 9.3 ÷ 9.5.

I valori in Tabella C sono stati calcolati dall'integrale

$$P_N(|r| \geqslant |r_o|) = \frac{2\Gamma[(N-1)/2]}{\sqrt{\pi}\,\Gamma[(N-2)/2]} \int_{|r_o|}^{1} (1 - r^2)^{(N-4)/4}\, dr.$$

Vedi, per esempio, E.M. Pugh e G.H. Winslow, "The Analysis of Physical Measurements" (Addison-Wesley, 1966), Sezione 12-8.

Tabella C. **La probabilità percentuale $P_N(|r| \geqslant r_O)$ che N misure di due variabili incorrelate diano un coefficiente di correlazione con $|r| \geqslant r_O$, come una funzione di N ed r_O. (I bianchi indicano probabilità minori di 0.05 percento).**

N	r_o: 0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
3	100	94	87	81	74	67	59	51	41	29	0
4	100	90	80	70	60	50	40	30	20	10	0
5	100	87	75	62	50	39	28	19	10	3.7	0
6	100	85	70	56	43	31	21	12	5.6	1.4	0
7	100	83	67	51	37	25	15	8.0	3.1	0.6	0
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0
9	100	80	61	43	29	17	8.8	3.6	1.0	0.1	0
10	100	78	58	40	25	14	6.7	2.4	0.5	—	0
11	100	77	56	37	22	12	5.1	1.6	0.3	—	0
12	100	76	53	34	20	9.8	3.9	1.1	0.2	—	0
13	100	75	51	32	18	8.2	3.0	0.8	0.1	—	0
14	100	73	49	30	16	6.9	2.3	0.5	0.1	—	0
15	100	72	47	28	14	5.8	1.8	0.4	—	—	0
16	100	71	46	26	12	4.9	1.4	0.3	—	—	0
17	100	70	44	24	11	4.1	1.1	0.2	—	—	0
18	100	69	43	23	10	3.5	0.8	0.1	—	—	0
19	100	68	41	21	9.0	2.9	0.7	0.1	—	—	0
N → 20	100	67	40	20	8.1	2.5	0.5	0.1	—	—	0
25	100	63	34	15	4.8	1.1	0.2	—	—	—	0
30	100	60	29	11	2.9	0.5	—	—	—	—	0
35	100	57	25	8.0	1.7	0.2	—	—	—	—	0
40	100	54	22	6.0	1.1	0.1	—	—	—	—	0
45	100	51	19	4.5	0.6	—	—	—	—	—	0
	0	.05	.1	.15	.2	.25	.3	.35	.4	.45	
50	100	73	49	30	16	8.0	3.4	1.3	0.4	0.1	
60	100	70	45	25	13	5.4	2.0	0.6	0.2	—	
70	100	68	41	22	9.7	3.7	1.2	0.3	0.1	—	
80	100	66	38	18	7.5	2.5	0.7	0.1	—	—	
90	100	64	35	16	5.9	1.7	0.4	0.1	—	—	
100	100	62	32	14	4.6	1.2	0.2	—	—	—	

Bassa probabilità (<5%) che per $r > r_0$ i valori di r siano ottenibili in assenza di correlazione

Osservazione: $Y = Y(x_1, x_2)$

$$\boxed{\sigma^2(Y) = \left(\frac{\partial Y}{\partial x_1}\right)^2 \sigma^2(x_1) + \left(\frac{\partial Y}{\partial x_2}\right)^2 \sigma^2(x_2) + 2\left(\frac{\partial Y}{\partial x_1}\right)\left(\frac{\partial Y}{\partial x_2}\right) \sigma(x_1, x_2)}$$

- In funzione del suo segno e della sua entità il termine $\left(\frac{\partial Y}{\partial x_1}\right)\left(\frac{\partial Y}{\partial x_2}\right) \cdot \sigma(x_1, x_2)$ potrà aumentare o diminuire in modo consistente l'entità dell'incertezza statistica su Y

- In linea generale la correlazione compare quando da un unico insieme di dati vengono estratti due o più parametri: questi potranno (forse) risultare correlati pur essendo perfettamente scorrelati i dati di partenza!

- Un esempio di ciò è dato dai parametri di un fit ...... per fortuna ciò non accade con i due estimatori tipici per:

  a) media $\langle x \rangle = \sum_1^N i \, x_i / N$

  b) deviazione standard $\sigma = \sqrt{\dfrac{\sum_1^N i (x_i - \langle x \rangle)^2}{N-1}}$

In altri termini per essi il coefficiente di correlazione lineare è nullo.

$$\boxed{\sigma(x_1, x_2) = \frac{1}{N} \sum_1^N i \, (x_{1i} - \langle x_1 \rangle)(x_{2i} - \langle x_2 \rangle)}$$