6. Distribuzioni statistiche doppie

6.1 Distribuzioni doppie di frequenze

Come si è visto nel paragrafo 2, di solito, lo studio statistico quantitativo di un fenomeno viene svolto rilevando contemporaneamente più caratteri per ciascuna unità statistica che compone il collettivo studiato. Ciò è necessario se si vuole cercare di spiegare il fenomeno attraverso nessi, interconnessioni, relazioni, fra diverse caratteristiche del collettivo in esame. Perciò, dopo aver esaminato i caratteri ad uno ad uno, ed avere determinato per i caratteri quantitativi media, variabilità e forma, è utile approfondire lo studio mettendo in relazione fra loro i caratteri che più appaiono importanti per la spiegazione del fenomeno. È perciò opportuno procedere costruendo le distribuzioni unitarie e quelle di frequenze che si ottengono considerando contemporaneamente il "modo di essere" di ciascuna unità rispetto a due caratteri.

Con riferimento ai dati della tabella 1, può avere interesse, dapprima, conoscere più a fondo il collettivo degli studenti rispetto alla scuola da cui provengono e alla riuscita o meno all'esame di statistica, per vedere se il percorso didattico seguito a scuola può avere influenza sull'esito all'esame di statistica. Fatto ciò può divenire interessante mettere direttamente in relazione i risultati dell'esame di statistica di coloro che l'hanno superato con il loro voto alla maturità.

Entrambe le analisi proposte richiedono la costruzione di distribuzioni doppie. In particolare per studiare la relazione tra tipo di maturità e superamento o meno dell'esame di statistica, occorre costruire innanzitutto la distribuzione doppia di frequenze rispetto ai due caratteri.

Poiché dalla distribuzione unitaria contenuta nella tabella 1 è difficile cogliere tale relazione, si organizzano i dati in modo da separare tra loro i diversi tipi di maturità e da evidenziare contemporaneamente se gli studenti hanno o non hanno superato l'esame di statistica. Si procede così ad effettuare ciò che, in termini tecnici statistici, si indica come classificazione del collettivo degli studenti contemporaneamente secondo due caratteri, in questo caso: tipo di maturità e voto.

Si ottiene in tale modo la seguente tabella doppia di frequenze[1].

Tab. 22 - Studenti presenti alla lezione di statistica del 6-10-1994
per tipo di maturità ed esito dell'esame di statistica

Esito dell'esame di statistica	CL	SC	TC	TI	A	Totale
Superato	7	41	9	9	5	71
Non superato	6	12	13	4	3	38
Totale	13	53	22	13	8	109

La tabella 22 contiene una distribuzione doppia di frequenze, o tabella di contingenza. In essa, per ogni possibile combinazione delle modalità delle due variabili, è indicata la corrispondente frequenza. Così 7 indica il numero di studenti che contemporaneamente hanno superato l'esame di statistica ed hanno la maturità classica; 3 è il numero di studenti che non hanno superato l'esame ed hanno un tipo di maturità indicata con altro, e così via. Le possibili combinazioni di modalità dei due caratteri sono 2x5=10. Nella tabella 22 esistono 2 distribuzioni semplici condizionate di riga e 5 distribuzioni semplici condizionate di colonna ed inoltre 2 distribuzioni semplici marginali, una di riga ed una di colonna.

[1] La tabella è il risultato dell'operazione di spoglio che consiste nel predisporre una tavola doppia con tante caselle quanto è il prodotto delle modalità dei due caratteri, nel nostro caso, 2x5=10 caselle, e nell'inserire in ogni casella una tacca ogni volta che una unità statistica presenta contemporaneamente la combinazione delle modalità dei due caratteri che la casella incrocia. La tabella si otterrà alla fine dello spoglio contando le tacche di ogni casella. Per le ultime cinque unità lo spoglio è il seguente:

Prospetto di spoglio dei caratteri: tipo di maturità ed esito dell'esame di statistica

Esito dell'esame di statistica	CL	SC	TC	TI	A
Superato		⌊			|
Non superato		⌊			

In dettaglio, le distribuzioni semplici condizionate di riga sono

CL	SC	TC	TI	A	
7	41	9	9	5	71

e

CL	SC	TC	TI	A	
6	12	13	4	3	38

ossia rispettivamente la distribuzione di coloro che hanno e non hanno superato l'esame per tipo di maturità.

Le distribuzioni condizionate di colonna sono 5.

Superato	Non superato	
7	6	13

Superato	Non superato	
41	12	53

Superato	Non superato	
9	13	22

Superato	Non superato	
9	4	13

Superato	Non superato	
5	3	8

Si tratta delle distribuzioni, rispetto all'esito, degli studenti che hanno conseguito rispettivamente la maturità classica, scientifica, e così via.

Le marginali sono sempre e solo 2, qualunque sia la dimensione della tabella. In questo caso, si tratta della distribuzione degli studenti alla maturità:

CL	SC	TC	TI	A	
13	53	22	13	8	109

e della distribuzione degli studenti rispetto all'esito dell'esame di statistica:

Superato	Non superato	
71	38	109

Occorre fare attenzione al fatto che, nota la distribuzione doppia le due distribuzioni marginali sono definite, mentre non è vero il viceversa. È anzi opportuno sapere che date due distribuzioni semplici con uguale numerosità n, le possibili associazioni di ogni unità dell'una con ciascuna unità dell'altra sono n!.

Si può dare generalità a quanto si è esaminato, esprimendo in modo simbolico la distribuzione doppia di frequenze. Si ha allora:

Tab. 23 - Distribuzione doppia di frequenze o tabella di contingenza.

Variabile X	Variabile Y						Totale
	y_1	y_2	...	y_j	...	y_c	
x_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1c}	$n_{1\bullet}$
x_2	n_{21}	n_{22}	...	n_{2j}	...	n_{2c}	$n_{2\bullet}$
...	...	...	...	...	...	...	...
x_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{ic}	$n_{i\bullet}$
...	...	...	...	...	...	...	...
x_r	n_{r1}	n_{r2}	...	n_{rj}	...	n_{rc}	$n_{r\bullet}$
Totale	$n_{\bullet 1}$	$n_{\bullet 2}$	...	$n_{\bullet j}$	...	$n_{\bullet c}$	n

dove n_{ij} , i=1, ..., r e j = 1, ..., c, è il numero di unità statistiche che presentano contemporaneamente le modalità x_i e y_j; $n_{i\bullet}=\sum_{j=1}^{c} n_{ij}$; $n_{\bullet j}=\sum_{i=1}^{r} n_{ij}$; e dove:

$$Y/x_i = \begin{pmatrix} y_1 & \dots & y_j & \dots & y_c \\ n_{i1} & \dots & n_{ij} & \dots & n_{ic} \end{pmatrix} n_{i\bullet} \qquad \text{(per i= 1,..., r)}$$

è la distribuzione di Y condizionata ad x_i, ossia una delle r distribuzioni di Y condizionate ad X;

$$X/y_j = \begin{pmatrix} x_1 & \dots & x_i & \dots & x_r \\ n_{1j} & \dots & n_{ij} & \dots & n_{rj} \end{pmatrix} n_{\bullet j} \qquad \text{(per j= 1,..., c)}$$

è una delle c distribuzioni di X (condizionate) ad Y;

$$\begin{pmatrix} y_1 & \dots & y_j & \dots & y_c \\ n_{\cdot 1} & \dots & n_{\bullet j} & \dots & n_{\cdot c} \end{pmatrix} n$$

è la distribuzione marginale della variabile Y;

$$\begin{pmatrix} x_1 & \dots & x_i & \dots & x_r \\ n_{1\cdot} & \dots & n_{i\bullet} & \dots & n_{r\cdot} \end{pmatrix} n$$

è la distribuzione marginale della variabile X.

6.2. Connessione e sue misure: relazioni statistiche fra caratteri qualunque ne sia la natura.

Rispetto alla tabella 22, ci si era chiesti se vi è relazione fra i due caratteri voto alla maturità ed esito all'esame di statistica, ossia se, cambiando tipo di maturità, cambia la distribuzione degli studenti rispetto all'esito dell'esame di statistica, o viceversa se cambiando l'esito dell'esame di statistica cambia la distribuzione degli studenti secondo la maturità. A prima vista la risposta è sì. Ciò che si cerca, tuttavia, è una risposta "tecnica oggettiva" ed essa è fornita dalla cosiddetta "misura del chi-quadrato" dovuta a K. Pearson.

Facendo riferimento in generale alla distribuzione doppia della tabella 23, si dice che "non c'è connessione" fra le due variabili X ed Y se al variare delle modalità dell'una la distribuzione condizionata delle modalità dell'altra non varia (rispetto ai dati, se al variare dell'esito non cambia la distribuzione del tipo di maturità condizionata all'esito). Poichè le distribuzioni condizionate hanno frequenza totale diversa (in tabella 22, 71 e 38 rispettivamente) va tenuto conto di ciò, facendo ricorso alle frequenze relative. Sicchè, ammettere che, ad esempio, Y/x_i è distribuita come Y/x_h, significa ipotizzare che, per ogni modalità j del carattere Y (j=1,2,....,c), è:

$$\frac{n_{i1}}{n_{i\bullet}} = \frac{n_{h1}}{n_{h\bullet}};.....;\frac{n_{ij}}{n_{i\bullet}} = \frac{n_{hj}}{n_{h\bullet}};.....;\frac{n_{ic}}{n_{i\bullet}} = \frac{n_{hc}}{n_{h\bullet}} \qquad [1].$$

Poichè la [1], in assenza di connessione è valida qualunque sia la coppia i,h presa in esame (per i,h= 1, ..., r), si ha in generale che:

$$\frac{n_{1j}}{n_{1\bullet}} = \frac{n_{2j}}{n_{2\bullet}} =.... \frac{n_{ij}}{n_{i\bullet}} = \frac{n_{hj}}{n_{h\bullet}} =... \frac{n_{rj}}{n_{r\bullet}} \qquad [2].$$

Applicando il componendo alla [2], si ha:

$$\frac{\sum_{i=1}^{r} n_{ij}}{\sum_{i=1}^{r} n_{i\bullet}} = \frac{n_{ij}}{n_{i\bullet}},$$

ossia:

$$\frac{n_{\bullet j}}{n} = \frac{n_{ij}}{n_{i\bullet}} \quad \forall i, j \qquad [3].$$

La [3] è una relazione chiave per le conseguenze che da essa discendono.

1) Se le Y/x_i (i=1, ..., r) sono distribuite tutte nello stesso modo, in statistica si dice che sono "simili", allora esse sono simili alla distribuzione marginale della variabile Y. Quindi, per giudicare circa l'assenza di connessione è sufficiente confrontare ogni Y/x_i alla distribuzione di Y, istituendo r confronti.

2) Cambiando i medi si ottiene

$$\frac{n_{i\bullet}}{n} = \frac{n_{ij}}{n_{\bullet j}} \qquad \forall i, j$$

ossia, in assenza di connessione di Y ad X, anche le distribuzioni X/y_j (j=1, ..., c) sono simili fra di loro e simili alla marginale X, ossia l'assenza di connessione è mutua.

3) Mettendo in evidenza n_{ij} si ha:

$$n_{ij} = \frac{n_{i\bullet} n_{\bullet j}}{n} = n^*_{ij}$$

che fornisce il valore teorico della frequenza nella casella ij in assenza di connessione, d'ora in poi indicato con n^*_{ij}.

4) La trasformazione

$$\frac{n^*_{ij}}{n} = \frac{n_{i\bullet}}{n}\frac{n_{\bullet j}}{n} = f_{i\bullet} f_{\bullet j}$$

mette in evidenza che la frequenza teorica relativa della casella i,j coincide con quella ottenuta per via probabilistica, secondo lo schema teorico di indipendenza (che sarà trattato nel seguito), perciò assenza di connessione ed indipendenza sono sinonimi.

Poiché il valore n^*_{ij}, per ogni coppia i,j, è la frequenza teorica in assenza di connessione (o per quanto detto al punto 4), in situazione di indipendenza), quanto più ci si allontana da esso, tanto più la connessione è grande, da qui l'indice di connessione proposto da K. Pearson, ossia la "misura del chi-quadrato":

$$D^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}} \qquad [4].$$

L'indice D^2 è sempre positivo, è mutuo (misura la connessione fra X ed Y) e il suo ordine di grandezza dipende da n [2], infatti:

$$D^2 = n\left(\sum_{ij} \frac{n^2_{ij}}{n_{i\bullet} n_{\bullet j}} - 1\right) \qquad [5].$$

Pertanto l'indice

$$\frac{D^2}{n} = \Delta^2 = \sum_{ij} \frac{n^2_{ij}}{n_{i\bullet} n_{\bullet j}} - 1 \qquad [6]$$

non dipende da n.

Ma qual è il massimo di Δ^2, ossia qual è la situazione di connessione massima o di connessione perfetta a cui l'indice fa riferimento?

Riprendendo i dati, la connessione dell'esito al tipo di maturità sarebbe massima se ad ogni tipo di maturità corrispondesse un solo esito, ad esempio se la distribuzione fosse:

---

[2] Trasformando la [4] si ottiene:

$$\sum_{ij} \frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}} = \sum_{ij} \frac{n^2_{ij} - 2n_{ij}n^*_{ij} + n^{*2}_{ij}}{n^*_{ij}} = \sum_{ij} \frac{n^2_{ij}}{n^*_{ij}} - 2\sum_{ij} n_{ij} + \sum_{ij} n^*_{ij} =$$

$$= \sum_{ij} n \frac{n^2_{ij}}{n_{i\cdot} n_{\cdot j}} - n = n\left(\sum_{ij} \frac{n^2_{ij}}{n_{i\cdot} n_{\cdot j}} - 1\right)$$

dato che $\sum_{ij} n_{ij} = \sum_{ij} n^*_{ij} = n$.

Esito dell'esame di statistica	CL	SC	TC	TI	A	Totale
Superato		71				71
Non superato	6		20	9	3	38

In questo caso, sapendo che gli studenti hanno la maturità scientifica, si saprebbe anche che hanno superato l'esame di statistica, mentre per ciascuna delle altre maturità l'esito sarebbe negativo. In questa situazione, tuttavia, non vi è connessione massima della maturità dato l'esito, poichè, noto l'esito, non si conosce esattamente la maturità. Infatti, mentre per gli studenti che hanno superato l'esame si sa che hanno la maturità scientifica, per quelli che non l'hanno superato non si conosce esattamente la maturità posseduta, a meno che non si modifichi la tabella, ad esempio così:

Esito dell'esame di statistica	SC	Non SC	Totale
Superato	71		71
Non superato		38	38

Dunque la connessione perfetta e mutua si può realizzare solo in una tabella quadrata, cioè quando r=c.
In tal caso è possibile che ad ogni modalità della variabile X si associ una ed una sola modalità di Y e viceversa, come ad esempio nella tabella 24.

Tabella 24 - Tabella quadrata di connessione massima.

X	Y			Totale
	y_1	y_2	y_3	
x_1	a			a
x_2			c	c
x_3		b		b
Totale	a	b	c	a+b+c

E' semplice verificare che, in tal caso è Δ^2=r-1.

Se, invece, r≠c si può realizzare o massima connessione di X ad Y se r<c, oppure massima connessione di Y ad X se c<r. In tale situazione max Δ^2=min(r-1,c-1).

Per misurare la connessione si può pertanto utilizzare:

$$0 \le d^2 = \frac{D^2}{n\left[\min(r-1, c-1)\right]} \le 1 \qquad [7],$$

dove d^2 è l'indice medio di contingenza. Esso vale 0 se e solo se vi è assenza di connessione. Quando vale 1 occorre distinguere 3 diverse situazioni. Se r=c , fra X ed Y vi è connessione perfetta e mutua; se r<c vi è perfetta connessione di X ad Y, se r>c vi è perfetta connessione di Y ad X.

Se d^2=0, nota una modalità della X nulla si può dire sulla corrispondente distribuzione condizionata della Y, e viceversa, poichè per ogni casella è $n_{ij} = n_{ij}^* \ \forall i, j$. Se d^2=1 ed r=c, la connessione è massima e mutua e quindi, nota una qualsiasi modalità di X, la Y assume una ed una sola modalità e viceversa. In generale, quanto più d^2 è grande tanto più le variabili sono connesse e la conoscenza della modalità assunta da una variabile è informativa sulla distribuzione delle modalità assunte dell'altra variabile.

Ritornando ai dati della tabella 22, si è in grado di costruire la tabella di connessione nulla, cioè quella per la quale è $n_{ij}^* = \frac{n_{i\bullet} n_{\bullet j}}{n}$.

Tabella 25 - Tabella di connessione nulla fra tipo di maturità
ed esito all'esame di statistica[3].

Esito dell'esame di statistica	Tipo di maturità					Totale
	CL	SC	TC	TI	A	
Superato	$\frac{71\cdot 13}{109}=8{,}468_{[7]}$	$\frac{71\cdot 53}{109}=34{,}523_{[41]}$	$\frac{71\cdot 22}{109}=14{,}330_{[9]}$	$\frac{71\cdot 13}{109}=8{,}468_{[9]}$	$\frac{71\cdot 8}{109}=5{,}211_{[5]}$	71
Non superato	$\frac{38\cdot 13}{109}=4{,}532_{[6]}$	$\frac{38\cdot 53}{109}=18{,}477_{[12]}$	$\frac{38\cdot 22}{109}=7{,}670_{[13]}$	$\frac{38\cdot 13}{109}=4{,}532_{[4]}$	$\frac{38\cdot 8}{109}=2{,}789_{[3]}$	38
Totale	13	53	22	13	8	109

Si hanno ora tutti gli elementi per calcolare D^2

$$D^2=\frac{(7-8{,}468)^2}{8{,}468}+\frac{(41-34{,}523)^2}{34{,}523}+\frac{(9-14{,}330)^2}{14{,}330}+\frac{(9-8{,}468)^2}{8{,}468}+\frac{(5-5{,}211)^2}{5{,}211}+$$
$$+\frac{(6-4{,}532)^2}{4{,}532}+\frac{(12-18{,}477)^2}{18{,}477}+\frac{(13-7{,}670)^2}{7{,}670}+\frac{(4-4{,}532)^2}{4{,}532}+\frac{(3-2{,}789)^2}{2{,}789}=\quad .$$
$$=0{,}254+1{,}215+1{,}982+0{,}033+0{,}009+$$
$$+0{,}476+2{,}270+3{,}704+0{,}062+0{,}016=10{,}021$$

Per d^2, si ottiene:

$$d^2=\frac{10{,}021}{109}=0{,}092\,.$$

Tenuto conto che $0\le d^2\le 1$,vi è bassa connessione fra le variabili. La tabella 25 mostra, tuttavia, che gli studenti del liceo scientifico che hanno superato l'esame di statistica sono di più di quanto si sarebbe atteso in un'associazione casuale;viceversa, gli studenti dell'istituo tecnico commerciale che hanno fallito nella prova sono quasi il doppio di quanto ci si sarebbe atteso.

La formula di calcolo utilizzata consente di mettere in evidenza il contributo di ogni casella a D^2. Così ad esempio sono i tecnici commerciali che non hanno superato l'esame ad aver dato il contributo maggiore, pari al 36,962% (3,704x100/10,021); altro contributo importante viene fornito dagli studenti dello scientifico che non hanno superato l'esame, esso è pari al 22,652% (2,270x100/10,021).

Il calcolo di D^2 risulta più rapido utilizzando la [5], soprattutto in presenza di caselle vuote, essa tuttavia non consente di evidenziare il contributo a D^2 di ciascuna casella.

Si osserva che per calcolare D^2 e d^2 sono state utilizzate solo le frequenze n_{ij}, $n_{i\bullet}$, $n_{\bullet j}$, non si è pertanto tenuto conto delle modalità e della natura dei caratteri trattati, perciò le misure di connessione proposte possono essere calcolate, purchè si abbia a disposizione una tabella di contingenza (o tabella doppia di frequenze). Poiché l'indice dipende dal numero delle modalità in riga ed in colonna, è opportuno tener conto di ciò quando si costruisce la tabella doppia, limitandone opportunamente il numero.

6.3. Relazioni statistiche tra caratteri entrambi quantitativi.

Dopo aver osservato che il tipo di maturità non è del tutto ininfluente rispetto all'esito dell'esame di statistica, visto che per ogni studente si è rilevato il voto alla maturità e il voto con cui è stato superato l'esame di statistica, può sorgere la domanda se esiste qualche forma di associazione fra i due caratteri. Per 70[4] studenti si conoscono entrambe le informazioni quantitative: voto alla maturità e voto all'esame di statistica (tabella 26), è possibile perciò fare la rappresentazione grafica della distribuzione doppia delle 70 unità statistiche su un piano cartesiano ortogonale, disponendo, ad esempio, il voto alla maturità sull'asse delle ascisse (variabile X), e il voto all'esame di statistica sull'asse delle ordinate (variabile Y).

---

[3] Il numero entro parentesi quadra è la frequenza empirica ottenuta dall'operazione di spoglio.

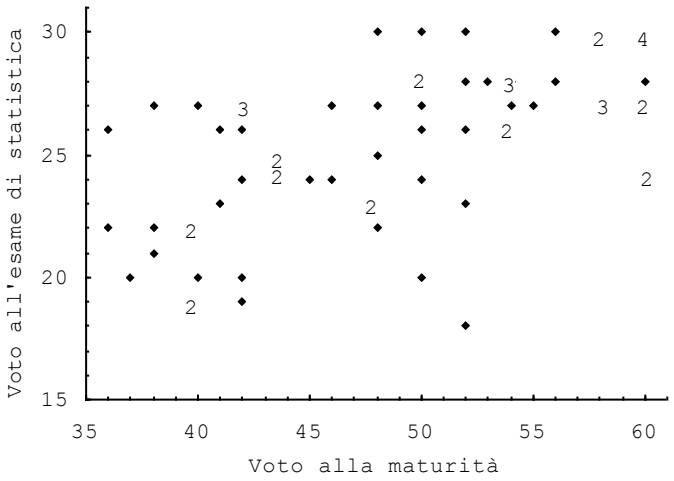
[4] Come già osservato per l'unità 41 della tabella 1. non si dispone del dato che riguarda il voto alla maturità, poiché non è possibile colmare questa lacuna, l'unità è state eliminata da questa analisi. Inoltre il voto 30 e lode riportato dalle unità 92 e 96 è stato ricodificato in 30.

Tabella 26 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95 e di cui è nota l'informazione completa

N° d'ordine	Tipo di maturità	Voto alla maturità	Esito all'esame di statistica	N° d'ordine	Tipo di maturità	Voto alla maturità	Esito all'esame di statistica
1	SC	50	28	36	TI	48	23
2	SC	38	21	37	SC	44	25
3	SC	42	24	38	TC	50	24
4	SC	44	24	39	CL	50	27
5	SC	37	20	40	CL	60	27
6	SC	44	25	41	CL	42	27
7	AL	42	27	42	CL	46	24
8	TI	48	27	43	SC	40	20
9	SC	60	24	44	SC	60	30
10	SC	44	24	45	CL	52	26
11	TI	50	26	46	TI	48	25
12	AL	60	30	47	SC	54	26
13	TC	58	27	48	SC	60	30
14	SC	60	28	49	SC	54	27
15	SC	54	26	50	TC	50	30
16	TI	46	27	51	SC	42	19
17	SC	50	20	52	SC	41	23
18	SC	53	28	53	TI	52	23
19	SC	38	27	54	SC	42	26
20	SC	58	30	55	AL	40	22
21	SC	48	23	56	TI	52	18
22	SC	45	24	57	SC	60	24
23	SC	42	27	58	SC	54	28
24	TC	58	30	59	SC	60	30
25	SC	52	28	60	CL	36	22
26	SC	41	26	61	TC	48	30
27	SC	36	26	62	TC	52	30
28	SC	54	28	63	TC	60	27
29	SC	40	27	64	SC	54	28
30	AL	40	19	65	SC	58	27
31	SC	55	27	66	TC	40	22
32	TI	42	20	67	SC	58	27
33	CL	56	30	68	SC	56	28
34	TC	40	19	69	AL	50	28
35	SC	38	22	70	SC	48	22

Con questo procedimento si ottiene la cosiddetta nuvola dei punti, ogni punto della quale rappresenta la coppia ordinata (x_i, y_i) che corrisponde all'unità statistica i-esima. Nel grafico ottenuto (Figura 5) osserviamo che vi sono unità con la stessa coppia di modalità, ad esempio la 30 e la 34 che hanno entrambe coordinate (40;19) e nel grafico sono rapprentate dal numero 2; le unità 7, 23, 41, hanno tutte coordinate (42;27) e sono individuate dal numero 3, e così via. Il simbolo ♦ individua, invece, punti di frequenza unitaria.

Figura 5 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95
per voto alla maturità e all'esame di statistica



La figura 5 mostra che non esiste una "legge" che mette in relazione le due variabili, tuttavia la disposizione dei punti nel piano segnala una tendenza generale all'associazione fra la variabile X e la variabile Y. Le domande che si pongono, a questo punto, sono due:

1) qual è l'intensità della relazione ed il suo verso?

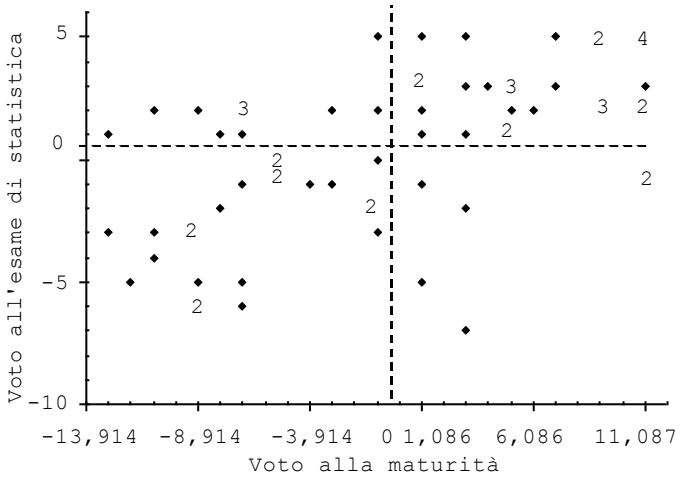
2) ammesso che la variabile X, voto alla maturità, possa servire a "spiegare" la variabile Y, voto all'esame di statistica, come si può rappresentare tale relazione, utilizzando strumenti analitici?

Il primo problema si risolve con lo studio della correlazione. Il secondo con lo studio della regressione. Essi verranno affrontati in successione.

6.3.1. Correlazione.

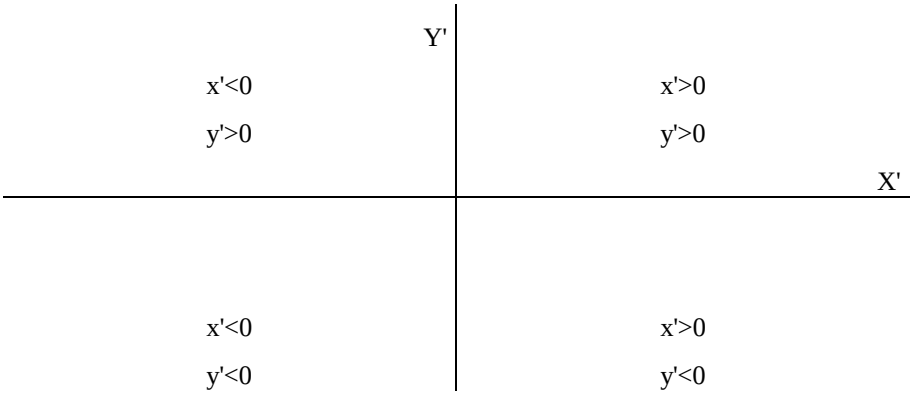
a) I punti del grafico della figura 5 mostrano che al crescere della variabile X: "voto alla maturità" tende a crescere anche la variabile Y: "voto all'esame di statistica". Quando i punti del grafico mostrano una situazione di questo tipo si dice che fra le due variabili c'è "correlazione positiva" o "concordanza". Se, ma non è qui il caso, al crescere di una variabile l'altra tende a decrescere si parla di "correlazione negativa" o di "discordanza" fra le variabili. La correlazione fra le variabili è dunque simmetrica e dotata di un segno. Per misurarla è necessario un indice che rispetti queste due caratteristiche. Per ottenerlo si trasla l'origine degli assi nel punto di coordinate $(\overline{x};\overline{y})$, così facendo la nuvola dei punti non muta, cambia invece il sistema di riferimento che diventa X'=X-$\overline{x}$ e Y'=Y-$\overline{y}$, dove X' e Y' sono dette variabili scarto (figura 6).

Figura 6 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95
per voto alla maturità e all'esame di statistica (scarti dalla media)



La traslazione degli assi consente di dare una interpretazione statistica della posizione che i punti possono assumere nei quattro quadranti del piano, dove, come noto, i segni delle coordinate si dispongono come indicato nella figura 7.

Figura 7 - Quadranti del piano (X',Y')



I punti del I e del III quadrante hanno "coordinate concordi", poiché contengono unità che presentano entrambe modalità o sopra la rispettiva media (I quadrante) o sotto la rispettiva media (III quadrante). Se i "punti della nuvola" si dispongono con maggiore frequenza in questi quadranti (come è il caso dei dati che si stanno analizzando) si dice che fra le variabili vi è "concordanza" o, anche, "correlazione positiva", poichè al crescere dell'una mediamente cresce anche l'altra.

Viceversa i punti del II e del IV quadrante hanno "coordinate discordi", poiché, per ciascuna unità, delle due modalità l'una è sopra e l'altra è sotto la rispettiva media. Se i punti della nuvola si dispongono prevalentemente in questi

quadranti fra le variabili vi è "discordanza" o, anche, "correlazione negativa", poichè al crescere di una variabile l'altra tendenzialmente decresce.

Per quanto si è detto, un indice del tipo:

$$s_{XY} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n}$$

è adatto a misurare la correlazione, poichè rispetta la simmetria del concetto, ed inoltre s_{XY} può essere o positivo o negativo a seconda che prevalga l'associazione fra scarti concordi o fra scarti discordi. Tuttavia s_{XY}, noto come covarianza, non è adimensionale, ma dipende dall'unità di misura in cui sono espresse le variabili, e ciò rende impossibili i confronti fra distribuzioni doppie. Questo problema viene superato tenendo conto della nota diseguaglianza di Cauchy-Schwarz:

$$\left(\frac{\sum_i a_i \cdot b_i}{n} \right)^2 \le \frac{\sum_i a_i^2}{n} \frac{\sum_i b_i^2}{n},$$

applicando la quale si ha

$$s_{XY} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n} \le \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n} \frac{\sum_i (y_i - \bar{y})^2}{n}} = s_X \cdot s_Y,$$

dove s_X ed s_Y sono, lo scarto quadratico medio, rispettivamente, della variabile X e della Y.

Per misurare la correlazione si utilizza, pertanto, l'indice:

$$r = \frac{s_{XY}}{s_X s_Y}$$

noto come coefficiente di correlazione di Bravais-Pearson. L'indice r risponde ai requisiti richiesti ad un indice per misurare la correlazione, infatti si tratta di un indice simmetrico, dotato di segno, che non dipende né dall'ordine medio di grandezza delle due variabili, né dalla loro unità di misura.

Poichè $-1 \le r \le 1$, se è positivo, $0 < r \le 1$, vi è concordanza fra i caratteri; se è negativo, $-1 \le r < 0$, vi è discordanza.

Va osservato che $|r| = 1$ se e solo se le due variabili X' eY' sono direttamente proporzionali, ossia se i punti della nuvola sono allineati. In tale situazione, infatti, $x_i - \bar{x} = k(y_i - \bar{y})\ \forall$ i, con il coefficiente di proporzionalità positivo o negativo. Di conseguenza si ha:

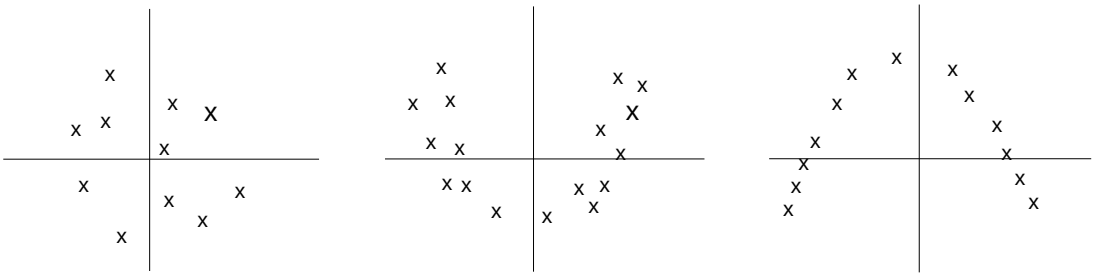
$$r = \frac{\frac{1}{n}\sum_i k(y_i - \bar{y})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_i [k(y_i - \bar{y})]^2 \frac{1}{n}\sum_i (y_i - \bar{y})^2}} =$$

$$= \frac{k\sum_i (y_i - \bar{y})^2}{\sqrt{k^2\left[\sum_i (y_i - \bar{y})^2\right]^2}} = \frac{k}{|k|} = (\text{segno } k)1$$

ossia r=+1 se il coefficiente di proporzionalità è positivo, r=-1 se è negativo.E' immediato verificare il viceversa.

Poiché è legato alla linearità della relazione fra X ed Y, r è anche detto coefficiente di correlazione lineare. In questi termini è meglio comprensibile l'interpretazione di r=0. Per r=0, si dice che vi è "indifferenza", tra le variabili non vi è infatti in tal caso né concordanza né discordanza, ma la somma dei prodotti degli scarti concordi si compensa con la somma dei prodotti degli scarti discordi. E' la situazione che si verifica, ad esempio, nelle nuvole dei punti rappresentate in figura 8. In tutti i grafici r=0, ma il significato è ben diverso.Nella figura 8.1, la forma rotondeggiante della nuvola dei punti mostra che non vi è connessione fra le variabili, pertanto non vi può essere correlazione; i grafici delle figure 8.2 e 8.3 mostrano, invece, una situazione in cui esiste connessione fra X ed Y, ma r non segnala correlazione poichè il legame fra le variabili non è lineare, né monotono.

Figura 8 - r=0.



8.1 8.2 8.3

r si può trovare scritto in diversi modi[5]:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{Codev(X,Y)}{\sqrt{Dev(X)Dev(Y)}} \qquad [8],$$

oppure, scrivendo più in chiaro l'ultima espressione, che è quella utilizzata per il calcolo:

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_i x_i^2 - n\bar{x}^2\right)\left(\sum_i y_i^2 - n\bar{y}^2\right)}} \qquad [9]$$

dato che vale la relazione:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - n\bar{x}\bar{y}\ ^{6}. \qquad [10]$$

Infine, se si usa la notazione:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{1}{n}\sum_i \frac{x_i - \bar{x}}{s_X}\frac{y_i - \bar{y}}{s_Y} \quad ,$$

dove $(x_i - \bar{x})/s_X$ e $(y_i - \bar{y})/s_Y$ sono rispettivamente gli scarti standardizzati delle due variabili, si evidenzia che il coefficiente di correlazione lineare non varia se viene calcolato sulle variabili, sui loro scarti, sui loro scarti standardizzati.

Ritornando all'esame dei dati della tabella 26, si voglia calcolare la correlazione esistente fra voto alla maturità e voto all'esame di statistica, ossia l'intensità e il verso del legame fra le due variabili. Dalla figura 6. si sa che r>0, per

---

[5] Si tenga presente che valgono le seguenti relazioni:

$$s_{XY} = Cov(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{Codev(X,Y)}{n}$$

e che

$$s_x^2 = Var(X) = \frac{\sum_i (x_i - \bar{x})^2}{n} = \frac{Dev(X)}{n} = \frac{\sum_i x_i^2 - n\bar{x}^2}{n}$$

[6] È infatti:

$$\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \sum_i x_i y_i - \bar{y}\sum_i x_i - \bar{x}\sum_i y_i + n\bar{x}\bar{y} =$$
$$= \sum_i x_i y_i - n\bar{y}\bar{x} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum_i x_i y_i - n\bar{x}\bar{y}.$$

calcolarne il valore, applicando la formula [9], è opportuno predisporre i dati come nella tabella 27 dalla quale si ottiene con facilità:

$$\bar{x} = \frac{3424}{70} = 48{,}914$$

$$\bar{y} = \frac{1786}{70} = 25{,}514$$

$$\sum_i x_i y_i = 88330$$

$$\sum_i y_i^2 = 46312$$

$$\sum_i x_i^2 = 171270$$

Tabella 27 - Calcoli per la determinazione di r.

x_i	x_i^2	y_i	y_i^2	$x_i y_i$	x_i	x_i^2	y_i	y_i^2	$x_i y_i$
52	2704	18	324	936	42	1764	26	676	1092
40	1600	19	361	760	42	1764	27	729	1134
40	1600	19	361	760	48	2304	27	729	1296
42	1764	19	361	798	58	3364	27	729	1566
37	1369	20	400	740	46	2116	27	729	1242
50	2500	20	400	1000	38	1444	27	729	1026
42	1764	20	400	840	42	1764	27	729	1134
40	1600	20	400	800	40	1600	27	729	1080
38	1444	21	441	798	55	3025	27	729	1485
38	1444	22	484	836	50	2500	27	729	1350
40	1600	22	484	880	60	3600	27	729	1620
36	1296	22	484	792	42	1764	27	729	1134
40	1600	22	484	880	54	2916	27	729	1458
48	2304	22	484	1056	60	3600	27	729	1620
48	2304	23	529	1104	58	3364	27	729	1566
48	2304	23	529	1104	58	3364	27	729	1566
41	1681	23	529	943	50	2500	28	784	1400
52	2704	23	529	1196	60	3600	28	784	1680
42	1764	24	576	1008	53	2809	28	784	1484
44	1936	24	576	1056	52	2704	28	784	1456
60	3600	24	576	1440	54	2916	28	784	1512
44	1936	24	576	1056	54	2916	28	784	1512
45	2025	24	576	1080	54	2916	28	784	1512
50	2500	24	576	1200	56	3136	28	784	1568
46	2116	24	576	1104	50	2500	28	784	1400
60	3600	24	576	1440	60	3600	30	900	1800
44	1936	25	625	1100	58	3364	30	900	1740
44	1936	25	625	1100	58	3364	30	900	1740
48	2304	25	625	1200	56	3136	30	900	1680
50	2500	26	676	1300	60	3600	30	900	1800
54	2916	26	676	1404	60	3600	30	900	1800
41	1681	26	676	1066	50	2500	30	900	1500
36	1296	26	676	936	52	2704	30	900	1560
52	2704	26	676	1352	60	3600	31	961	1860
54	2916	26	676	1404	48	2304	31	961	1488
				Totale	3424	171270	1786	46312	88330

Sostituendo nella [9] si ha pertanto:

$$r = \frac{88330 - 70 \cdot 48{,}914 \cdot 25{,}514}{\sqrt{\left(171270 - 70 \cdot 48{,}914^2\right)\left(46312 - 70 \cdot 25{,}514^2\right)}} =$$

$$= \frac{970{,}574}{\sqrt{3789{,}442 \cdot 744{,}506}} = 0{,}578$$

Fra voto alla maturità e voto all'esame di statistica vi è, nel collettivo degli studenti esaminato, correlazione positiva, pari al 57,8% di quella massima che si realizzerebbe qualora vi fosse proporzionalità diretta fra le due rispettive variabili scarto. La concordanza è dunque evidente.

aa) Finora sono stati trattati dati organizzati in una tabella unità-variabili nella quale, per ogni unità statistica, era nota la modalità portata, o presentata, da ciascuna variabile. Non sempre i dati disponibili assumono questa forma, e ci si può trovare nella necessità di calcolare la correlazione fra variabili di cui è nota la distribuzione doppia di frequenze. In tal caso le formule di calcolo di r vanno opportunamente modificate. In particolare, tenuto conto della tabella 23, si ha:

$$r = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}(x_i-\bar{x})(y_j-\bar{y})n_{ij}}{\sqrt{\sum_{i=1}^{r}(x_i-\bar{x})^2 n_{i\bullet}\sum_{j=1}^{c}(y_j-\bar{y})^2 n_{\bullet j}}} = \frac{\sum_i\sum_j x_i y_j n_{ij} - n\bar{x}\bar{y}}{\sqrt{\left(\sum_i x_i^2 n_{i\bullet} - n\bar{x}^2\right)\left(\sum_j y_j^2 n_{\bullet j} - n\bar{y}^2\right)}} \qquad [11]$$

Classificando i dati della tabella 26 si è costruita la distribuzione doppia di frequenze per voto alla maturità e voto all'esame di statistica (tabella 28[7]).

Tabella 28 - Studenti che hanno superato l'esame di statistica nell'a.a. 94.95 per per classe di voto alla maturità e classe di voto all'esame di statistica

	Classe di voto all'esame di statistica (Y)			
Classe di voto alla maturità (X)	18 ⊢ 23	24 ⊢ 27	28 ⊢ 30	Totale
36 ⊢ 43	12	9	0	21
44 ⊢ 54	6	16	10	32
55 ⊢ 60	0	8	9	17
Totale	18	33	19	70

Per calcolare r si utilizza la formula [11].
Il voto medio all'esame di statistica si ottiene nel modo noto:

$$\bar{y} = \frac{\sum_{j=1}^{3} y_j n_{\bullet j}}{n} = \frac{20{,}5\cdot 18 + 25{,}5\cdot 33 + 29\cdot 19}{70} = 25{,}164$$

Il voto medio alla maturità è, analogamente, dato da:

$$\bar{x} = \frac{\sum_{i=1}^{3} x_i n_{i\bullet}}{n} = \frac{39{,}5\cdot 21 + 49\cdot 32 + 57{,}5\cdot 17}{70} = 48{,}214\,.$$

E' inoltre: $\sum_{i=1}^{3} x_i^2 n_{i\bullet} = 165803{,}5$ e $\sum_{j=1}^{3} y_j^2 n_{\bullet j} = 45001{,}75$.

[7] Il raggruppamento delle modalità in classi è effettuato secondo una scelta soggettiva, ma influente sui risultati successivi dell'analisi statistica, che ne rimane condizionata. Non esiste al momento nessun criterio generale per risolvere questo problema, affidato anche al buon senso dell'operatore e alla sua conoscenza del fenomeno studiato. Qui, in particolare, si è sfruttata la figura 5 che suggerisce di ripartire i voti di maturità in 3 classi (bassi: 36 ⊢ 43, medi: 44 ⊢ 54, alti: 55 ⊢ 60) e la conoscenza della classificazione dei voti universitari secondo il docente che esamina gli studenti (bassi: 18 ⊢ 23, medi: 24 ⊢ 27, alti: 28 ⊢ 30).

Si calcola ora $\sum_{i,j} x_i y_j n_{ij}$, utilizzando la tabella 29, ogni casella i,j della quale contiene il prodotto $x_i y_j n_{ij}$:

Tabella 29 - Tabella dei prodotti $x_i y_j n_{ij}$.

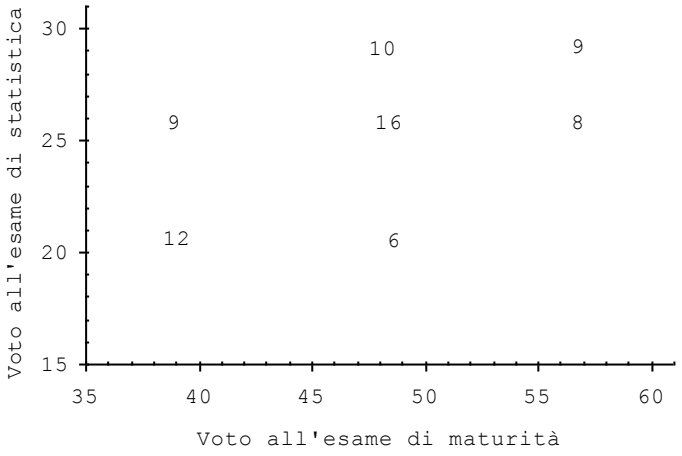
Xi	Y_j			Totale
	20,5	25,5	29	
39,5	9717	9065,25	0	18782,25
49	6027	19992	14210	40229
57,5	0	11730	15007,5	26737,5
Totale	15744	40787,25	29217,5	85748,75

Si può ora calcolare r:

$$r = \frac{85748{,}75 - 70 \cdot 25{,}164 \cdot 48{,}214}{\sqrt{(165803{,}5 - 70 \cdot 2324{,}59)(45001{,}75 - 70 \cdot 633{,}227)}} = \frac{820{,}753}{\sqrt{3082{,}214 \cdot 675{,}867}} = 0{,}569 .$$

Si osserva che il valore di r ottenuto è diverso da quello calcolato in precedenza con dati del tipo unità-variabili. Ciò è giustificato dal raggruppamento delle modalità in classi, il che rende uguali dal punto di vista classificatorio unità fra di loro diverse, ed inoltre richiede, per il calcolo, di ipotizzare che, per ogni casella ij, la frequenza sia attribuita, rispettivamente, al valore centrale della classi i-esima della X e j-esima della Y. Anche dal punto di vista grafico si può evidenziare la diversità fra la distribuzione doppia unità-variabili (figura 5) e la distribuzione doppia di frequenze (figura 9). In particolare, nel secondo grafico le irregolarità della nuvola dei punti si attenuano rispetto al primo ed emerge meglio la forma ellittica della nuvola.

Figura 9 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95
per voto alla maturità e all'esame di statistica (valori centrali delle classi)



6.3.2 Regressione.

I dati esaminati mostrano, dunque, concordanza (il 57,8% per la distribuzione doppia unità-variabili) fra voto alla maturità e voto all'esame di statistica. Essendo entrambi le variabili quantitative, è possibile concettualmente procedere oltre, e cercare una "espressione" che le leghi fra loro. Così, se si considera il voto all'esame di maturità come "antecedente causale", come "variabile esplicativa" rispetto al voto all'esame di statistica ha senso chiedersi di trovare una relazione funzionale fra voto all'esame di statistica e voto alla maturità, considerando il primo come variabile dipendente dal secondo, che assume il ruolo di variabile indipendente o esplicativa. Viceversa, se si considera come esplicativo il voto all'esame di statistica. In modo più formale, si può cercare una relazione del tipo:

$$y = f(x)$$

oppure, se ha senso:

$$x = f(y)\,.$$

Il problema non ha soluzione univoca, essa dipende da come si specifica f e dal modo in cui si individuano i parametri della f prescelta.

La funzione più semplice con cui legare due variabili è la retta, che ha al massimo due parametri, entrambi facilmente interpretabili: l'intercetta e il coefficiente angolare. La retta è dunque la proposta più elementare per specificare f e le due variabili si possono "legare" fra loro con le funzioni:

$$y = a + bx \qquad [12]$$

oppure:

$$x = a'+b'y \qquad [13].$$

Poichè la [12] e la [13] individuano ciascuna una famiglia di rette, è necessario disporre di uno strumento che consenta di "estrarre" quella particolare relazione lineare che è la più "adatta" a rappresentare la nuvola dei punti empirici osservati. Per passare dalla "nuvola empirica" alla "relazione analitica teorica" espressa mediante una retta, nota in statistica come retta di regressione, si utilizza il metodo di interpolazione dei minimi quadrati.

Ammesso di dover interpolare con una retta n punti (x_i,y_i), il metodo dei minimi quadrati perviene alla determinazione dei coefficienti ponendo la condizione:

$$F(a,b) = \sum_{i=1}^{n}\left[y_i - (a + bx_i)\right]^2 = \min_{a,b} \qquad [14]$$

o

$$F(a',b') = \sum_{i=1}^{n}\left[x_i - (a'+b'y_i)\right]^2 = \min_{a',b'}\,. \qquad [15]$$

La prima relazione richiede di individuare a e b in modo che la somma dei quadrati delle differenze fra le ordinate osservate y_i, e le corrispondenti ordinate teoriche $y_{(i)} = a + bx_i$ sia minima, da cui il nome del metodo. La seconda relazione si legge in modo analogo alla prima.

Ammesso di voler interpolare y=a+bx, utilizzando la [14] si ha:

$$F(a,b) = \sum_i \left(y_i^2 + b^2x_i^2 + a^2 - 2bx_iy_i - 2ay_i + 2abx_i\right) =$$

$$= \sum_i y_i^2 + b^2\sum_i x_i^2 + na^2 - 2b\sum_i x_iy_i - 2a\sum_i y_i + 2ab\sum_i x_i \qquad [16]$$

Poiché $\sum_i y_i = n\bar{y}$ e $\sum_i x_i = n\bar{x}$ dalla [16] si ottiene:

$$F(a,b) = \sum_i y_i^2 + b^2\sum_i x_i^2 + na^2 - 2b\sum_i x_iy_i - 2an\bar{y} + 2abn\bar{x}$$

dalla quale raccogliendo n si ottiene:

$$n\left[a^2 - 2a(\bar{y} - b\bar{x})\right] - 2b\sum_i x_iy_i + \sum_i y_i^2 + b^2\sum_i x_i^2 =$$

$$= n\left[a - (\bar{y} - b\bar{x})\right]^2 - n(\bar{y} - b\bar{x})^2 - 2b\sum_i x_iy_i + \sum_i y_i^2 + b^2\sum_i x_i^2\,.$$

Per minimizzare F(a,b), fissato b occorre che:

$$a - (\bar{y} - b\bar{x}) = 0 \qquad [17],$$

da cui:

$$a = \bar{y} - b\bar{x}.$$

La [16] può anche essere anche trasformata in:

$$\sum_i x_i^2 \left[b^2 - 2b\left(\frac{\sum_i x_i y_i}{\sum_i x_i^2} - a\frac{\sum_i x_i}{\sum_i x_i^2} \right) \right] + \sum_i y_i^2 + na^2 - 2a\sum_i y_i \ ,$$

ossia:

$$\sum_i x_i^2 \left[b - \left(\frac{\sum_i x_i y_i}{\sum_i x_i^2} - a\frac{\sum_i x_i}{\sum_i x_i^2} \right) \right]^2 - \sum_i x_i^2 \left[\frac{\sum_i x_i y_i}{\sum_i x_i^2} - a\frac{\sum_i x_i}{\sum_i x_i^2} \right]^2 + \sum_i y_i^2 + na^2 - 2a\sum_i y_i$$

che è minima, fissato a, se:

$$b - \left(\frac{\sum_i x_i y_i}{\sum_i x_i^2} - a\frac{\sum_i x_i}{\sum_i x_i^2} \right) = 0$$

ossia se:

$$b\sum_i x_i^2 - \left(\sum_i x_i y_i - an\bar{x} \right) = 0 \qquad [18]$$

Riunendo la [17] e la [18], si ottiene il seguente sistema lineare a due equazioni nelle incognite a e b:

$$\begin{cases} a + \bar{x}b = \bar{y} \\ n\bar{x}a + \sum_i x_i^2 b = \sum_i x_i y_i \end{cases} \qquad [19]^8,$$

dal quale si può ottenere b applicando la regola di Cramer:

$$b^* = \frac{\begin{vmatrix} 1 & \bar{y} \\ n\bar{x} & \sum_i x_i y_i \end{vmatrix}}{\begin{vmatrix} 1 & \bar{x} \\ n\bar{x} & \sum_i x_i^2 \end{vmatrix}} = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} \qquad [20].$$

Per la [10], il numeratore della [20] è Codev(X,Y); il denominatore è Dev(X), pertanto si può scrivere:

$$b^* = \frac{\sum_i x_i y_i - n\bar{x}\bar{y}}{\sum_i x_i^2 - n\bar{x}^2} = \frac{\text{Codev}(X,Y)}{\text{Dev}(X)} = \frac{\text{Cov}(X,Y)}{\text{Var}(X)} \qquad [21].$$

b^*, il coefficiente angolare della retta, prende il nome di coefficiente di regressione lineare. Esso varia fra $-\infty$ e $+\infty$ ed è espresso nell'unità di misura di Y rapportata all'unità di misura di X. Esprime la variazione subita da Y ad una variazione unitaria di X. Se b^* è positivo, la retta interpolata è crescente; se è negativo, è decrescente.

a^* si ottiene dalla prima equazione della [19] per sostituzione, si ha infatti:

---

[8] Lo stesso risultato si può ottenere utilizzando il concetto di derivata parziale di una funzione di due variabili, strumento che non è generalmente a disposizione degli studenti della scuola media superiore.

$$a^* = \bar{y} - \frac{Codev(X,Y)}{Dev(X)}\bar{x}$$

La retta interpolatrice diviene allora:

$$y = \frac{Codev(XY)}{Dev(X)}x + \bar{y} - \frac{Codev(X,Y)}{Dev(X)}\bar{x}$$

ossia:

$$y = \frac{Codev(X,Y)}{Dev(X)}\left(x - \bar{x}\right) + \bar{y} \qquad [22]$$

Se anziché operare sulle variabili X ed Y si opera sugli scarti $X'=X-\bar{x}$ e $Y'=Y-\bar{y}$, traslando l'origine nel baricentro, si ha:

$$y - \bar{y} = \frac{Codev(X,Y)}{Dev(X)}(x - \bar{x}) \qquad [23]$$

ossia,

$$y' = \frac{Codev(XY)}{Dev(X)}x'.$$

Si ottiene in tal modo una retta passante per l'origine il cui coefficiente angolare è il coefficiente di regressione. Si può anche osservare che il coefficiente di regressione calcolato sulle variabili X ed Y è identico a quello ottenuto con gli scarti X' e Y'.

Con un procedimento analogo a quello sin qui utilizzato si ottiene la retta che esprime la variabile X in funzione di Y, la cui equazione è:

$$x = \frac{Codev(X,Y)}{Dev(Y)}\left(y - \bar{y}\right) + \bar{x}$$

con riferimento alle variabili X e Y, e

$$x' = \frac{Codev(X,Y)}{Dev(Y)}y'$$

con riferimento agli scarti[9].

Si osserva che se $b^*=0$, si ha $a^*=\bar{y}$ e $y=\bar{y}$, se $b'^*=0$, $a'^*=\bar{x}$ e $x=\bar{x}$, perciò in questa situazione, che si realizza se e solo se Codev(X,Y)=0, le due rette di regressione sono parallele agli assi (o con essi coincidenti se si opera con gli scarti), perpendicolari fra loro e si incontrano nel baricentro.

Infine, i coefficienti delle due rette hanno lo stesso segno, che dipende solo da Codev(X,Y).

Si osserva che è possibile interpolare la nuvola degli n punti con una costante, ossia con y=c. In questo caso la condizione posta dal metodo dei minimi quadrati diviene $F(c) = \sum_i (y_i - c)^2 = \min_c$. Il problema dell'individuazione di c è stato di fatto già risolto nel momento in cui si è mostrato che la media aritmetica rende minima la somma degli scarti al quadrato e perciò $c = \bar{y}$ e di conseguenza $y = \bar{y}$. Il risultato per x è analogo.

È anche possibile verificare tale risultato per altra via. Infatti per y=c si può scrivere $F(c) = \sum_i y_i^{\,2} - 2c\sum_i y_i + nc^2$, che è l'equazione di una parabola in c che volge la concavità verso l'alto quando $\sum_i y_i$ è positiva, come è nelle applicazioni statistiche. Tale parabola ha il suo minimo nel vertice, di coordinate:

$$\left(-\frac{-\sum_i y_i}{n}; -\frac{\left(\sum_i y_i\right)^2 - n\sum_i y_i^{\,2}}{n} \right),$$

[9] E' utile osservare che b* e b'* hanno lo stesso numeratore, mentre al denominatore vi è la devianza della variabile indipendente (X per b*; Y per b'*).

ossia $\left(\bar{y};\sigma_y{}^2\right)$. Si conclude perciò che il minimo di F(c) si ha per $c=\bar{y}$, ed esso coincide con la varianza.

6.3.3 Calcolo delle rette di regressione.

a) I dati della tabella 26 consentono di esprimere mediante una funzione lineare il legame fra X ed Y, assumendo il voto alla maturità (X) come "antecedente causale", come "variabile esplicativa", come "variabile indipendente", rispetto al voto all'esame di statistica (Y), "variabile risposta", "variabile dipendente".

La retta interpolatrice è, per la [23]:

$$y=\frac{\text{Codev}(X,Y)}{\text{Dev}(X)}\left(x-\bar{x}\right)+\bar{y}\,.$$

Tutti gli elementi della formula sono già stati calcolati per ottenere r e perciò, sostituendo si ha:

$$y=\frac{970{,}574}{3789{,}442}\left(x-48{,}914\right)+25{,}514$$

da cui:

$$y=0{,}256x+12{,}986\,.$$

La retta è dunque ascendente, come la disposizione dei punti chiaramente indicava. Ad una variazione unitaria di X corrisponde sulla retta una variazione delle stesso segno, di ammontare pari a 0,256. La costante poi assicura circa 13 a tutti gli studenti che hanno superato l'esame di statistica, a cui si deve aggiungere circa 1/4 del voto di maturità.

E' anche possibile interpolare la retta che esprime il voto alla maturità in funzione del voto all'esame di statistica. Si ha allora:

$$x=\frac{\text{Codev}(X,Y)}{\text{Dev}(Y)}\left(y-\bar{y}\right)+\bar{x}$$

ossia:

$$x=\frac{970{,}574}{744{,}506}\left(y-25{,}514\right)+48{,}914$$

e perciò

$$x=1{,}304y+15{,}653\,.$$

La relazione permette, ad esempio, di valutare a quale voto di maturità corrisponde, con riferimento al collettivo in esame, un voto di statistica pari a 24, il voto dello studente 41 di cui non si conosce il voto alla maturità. Si ha:

$$x=1{,}304\cdot 24+15{,}653\approx 47\,.$$

aa) Come già si è visto per la correlazione, anche nel caso della regressione può verificarsi la necessità di operare su una tabella doppia di frequenze. In tale situazione le formule vanno opportunamente trasformate, in particolare si ha:

$$b^*=\frac{\sum_{i,j}x_iy_jn_{ij}-n\bar{x}\bar{y}}{\sum_i x_i^2n_{i\bullet}-n\bar{x}^2};b'^*=\frac{\sum_{i,j}x_iy_jn_{ij}-n\bar{x}\bar{y}}{\sum_j y_j^2n_{\bullet j}-n\bar{y}^2}\,.$$

Gli elementi della formula sono già stati tutti calcolati per ottenere r, si ha pertanto:

$$y=\frac{820{,}753}{3082{,}214}\left(x-48{,}214\right)+25{,}164$$

$$y=0{,}266x+12{,}339$$

e

$$x=\frac{820{,}753}{675{,}867}\left(y-25{,}164\right)+48{,}214$$

.

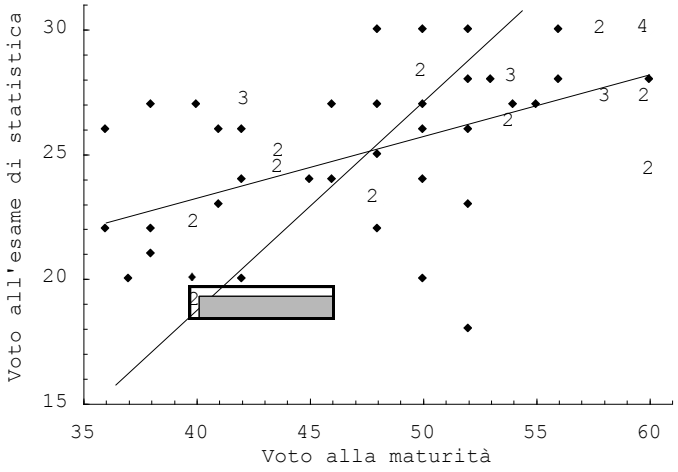
$$x=1{,}214y+17{,}665$$

Anche per le rette di regressione è evidente la diversità dei risultati numerici ottenuti operando su una distribuzione doppia unità-variabili da quelli ricavati operando su una distribuzione doppia di frequenze. La giustificazione di ciò è la stessa fornita analizzando la correlazione.

6.3.4 Rappresentazione grafica della nuvola dei punti, delle rette di regressione e proprietà delle rette di regressione.

E' anche possibile sovrapporre alla nuvola dei punti le due rette di regressione interpolate (figura 10).

Figura 10 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95
per voto alla maturità e all'esame di statistica e rette di regressione



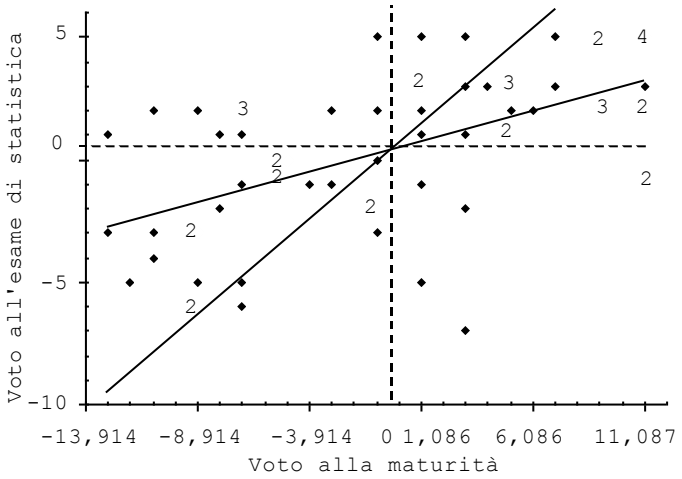
Dal punto di vista geometrico vi sono diverse osservazioni da fare.

La retta che esprime X in funzione di Y è più inclinata sull'asse delle X rispetto alla retta che esprime Y in funzione di X, ciò è vero sempre, e le due rette non si possono mai scambiare fra loro.

Le rette si incontrano nel baricentro. Infatti il punto $(\bar{x},\bar{y})$ soddisfa entrambe.

Traslando l'origine nel baricentro si ottiene la figura 11, che mette in evidenza che le rette ascendenti, quando si utilizzano gli scarti sono contenute nel I e nel III quadrante. Se le rette fossero state discendenti sarebbero state nel II e nel IV quadrante.

Figura 11 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95
per voto alla maturità e all'esame di statistica e rette di regressione (scarti dalle medie)



Non è difficile dal punto di vista empirico intuire che, se la nuvola dei punti degenera in una successione di punti allineati ($|r|=1$),le due rette di regressione coincidono, ciò significa, dal punto di vista fenomenico, che fra gli scarti delle variabili esaminate vi è proporzionalità diretta e perciò:

$$y_i - y_h = b^*(x_i - x_h)\forall i,h$$

$$x_i - x_h = b'^*(y_i - y_h)\forall i,h$$

Ciò richiede che $b'^* = 1/b^*$.

Se si opera con gli scarti standardizzati, ricordando che essi sono rispettivamente: $\frac{x-\bar{x}}{s_X}$ e $\frac{y-\bar{y}}{s_Y}$ e tenendo conto che $\frac{Codev(X,Y)}{Dev(X)} = \frac{Cov(X,Y)}{Var(X)}$ e che $Var(X) = s_X^2$, si può scrivere la [23] nella forma:

$$y-\bar{y} = \frac{Cov(X,Y)}{s_X^2}(x-\bar{x})$$

da cui, dividendo ambo i membri per s_Y, si ottiene

$$\frac{y-\bar{y}}{s_Y} = \frac{Cov(X,Y)}{s_X s_Y}\frac{x-\bar{x}}{s_X}$$

ma poiché per la [8] $\frac{Cov(X,Y)}{s_X s_Y} = \frac{s_{XY}}{s_X s_Y} = r$, segue che: $\frac{y-\bar{y}}{s_Y} = r\frac{x-\bar{x}}{s_X}$ [24].

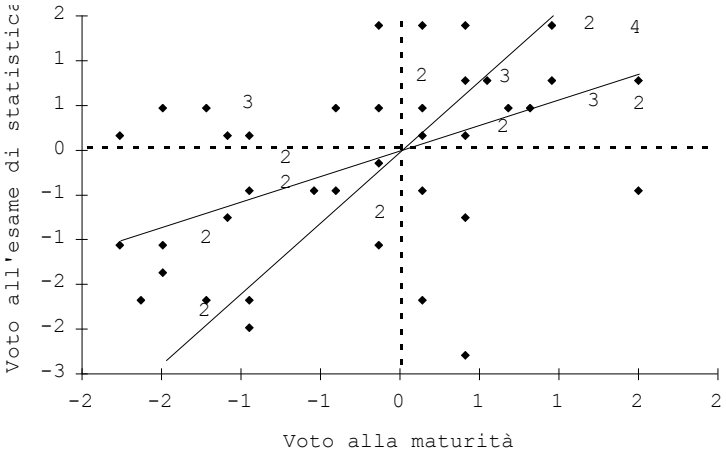
Anche la retta $x-\bar{x} = \frac{Cov(X,Y)}{Var(Y)}(y-\bar{y})$, dividendo ambo i membri per s_X, si trasforma in

$$\frac{x-\bar{x}}{s_X} = \frac{Cov(X,Y)}{s_X s_Y}\frac{y-\bar{y}}{s_Y}, \text{ ossia: } \frac{x-\bar{x}}{s_X} = r\frac{y-\bar{y}}{s_Y} \quad [25].$$

Perciò le rette [24] e [25] passano per l'origine ed hanno lo stesso coefficiente angolare, che coincide con r. Le rette aventi come variabili gli scarti standardizzati sono dunque simmetriche rispetto alla bisettrice del I e del III quadrante se r è positivo; rispetto alla bisettrice del II e del IV quadrante se r è negativo. Inoltre si può osservare che in questa situazione particolare i coefficienti di regressione sono compresi fra -1 e +1. Quando r=+1 le rette evidentemente coincidono con la bisettrice del I e del III quadrante, quando r=-1 le rette coincidono, invece, con la bisettrice del II e del IV.

Ritornando ai dati, la figura 12 mostra la nuvola e le rette di regressione ottenute con le variabili standardizzate: $X'' = \frac{X-49{,}914}{7{,}358}$ e $Y'' = \frac{Y-42{,}56}{3{,}261}$, tenuto conto che le rette hanno rispettivamente equazione: $x'' = 0{,}578y''$ e $y'' = 0{,}578x''$, essendo r=0,578.

Figura 12 - Studenti che hanno superato l'esame di statistica nell'a.a. 94-95per voto alla maturità e all'esame di statistica e rette di regressione (scarti standardizzati)



Quando si usano gli scarti standardizzati, è possibile valutare sul grafico l'intensità del legame lineare fra X ed Y. Infatti quanto più le rette di regressione si avvicinano agli assi, tanto meno vi è legame lineare fra le variabili, quanto più le rette si avvicinano fra loro e alla bisettrice del I e III quadrante, tanto più il legame lineare è forte, ed è perfetto quando le rette si sovrappongono.

6.3.5. Da r a r^2. Analisi dell'adattamento.

r è stato introdotto come misura della interrelazione fra variabili, ossia come misura della intensità della loro "struttura associativa", è stato poi ritrovato anche come coefficiente di regressione lineare nel caso particolare in cui si utilizzano scarti standardizzati, ma analizzando ulteriormente le formule lo si può ottenere anche per altra via.

Si considerino i 2 coefficienti di regressione:

$$b^* = \frac{Cov(X,Y)}{Var(X)};\ \ b'^* = \frac{Cov(X,Y)}{Var(Y)},$$

di essi è possibile costruire una sintesi attraverso la media geometrica. Si ha allora:

$$\sqrt{b^* b'^*} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = r,$$

perciò:

$$-1 \le \sqrt{b^* b'^*} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = r \le +1,$$

ma è anche, ovviamente:

$$0 \le b^* b'^* = \frac{Cov^2(X,Y)}{Var(X)Var(Y)} = r^2 \le +1 \qquad [26].$$

Dalla [26] si ricava che $r^2=r=0$ se e solo se Cov(XY)=0, di conseguenza è anche b*=b'*=0, ossia i due coefficienti di regressione si annullano contemporaneamente. Ciò si può verificare, come già osservato, o in una situazione di "indifferenza" fra X ed Y, o in una situazione di "indipendenza" (assenza di connessione). Le 2 rette di regressione che si ottengono sono fra loro perpendicolari, passano per il baricentro e sono o parallele agli assi (se si utilizzano le variabili X ed Y) o con essi coincidenti (se si utilizzano gli scarti o gli scarti standardizzati).

Dalla [26] si ottiene anche un'altra relazione già nota, infatti se $r^2=1$ segue che $b'^* = 1/b^*$ e le due rette coincidono.

In generale, il variare di r^2 fra 0 e 1 segnala il variare della relazione fra le variabili da una situazione di "indifferenza" o di "indipendenza" ad una situazione di linearità perfetta.

Vi è poi un'altra interpretazione di r^2. Riprendendo la [26] e moltiplicando e dividendo per Var(X) si ottiene:

$$r^2 = \frac{Cov^2(X,Y)}{Var^2(X)}\frac{Var(X)}{Var(Y)} = b^{*2}\frac{Var(X)}{Var(Y)} =$$

$$= \frac{b^{*2}\frac{1}{n}\sum_i (x_i - \bar{x})^2}{Var(Y)} = \frac{\frac{1}{n}\sum_i \left(b^* x_i - b^*\bar{x}\right)^2}{Var(Y)} =$$

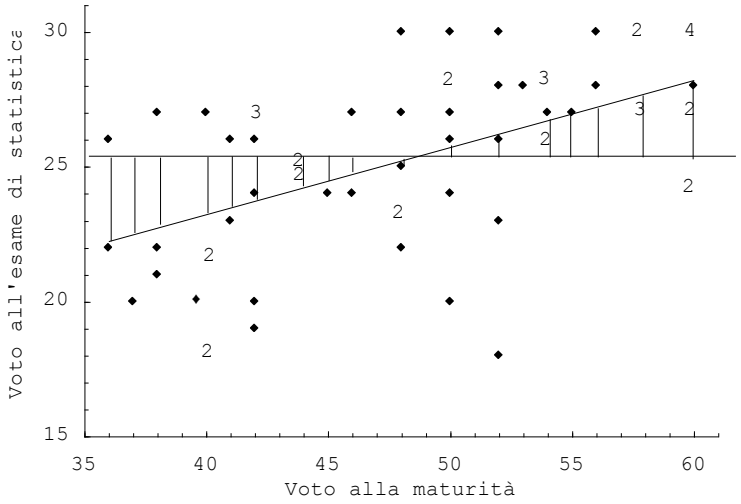
$$= \frac{\frac{1}{n}\sum_i \left[a^* + b^* x_i - \left(a^* + b^*\bar{x}\right)\right]^2}{Var(Y)}$$

Poiché, per la prima equazione del sistema [19] $a^* + b^*\bar{x} = \bar{y}$, ed inoltre, come noto, $a^* + b^* x_i = y_i^*$, si ha:

$$r^2 = \frac{\frac{1}{n}\sum_i \left(y_i^* - \bar{y}\right)^2}{Var(Y)} = \frac{\frac{1}{n}\sum_i \left(y_i^* - \bar{y}\right)^2}{\frac{1}{n}\sum_i \left(y_i - \bar{y}\right)^2} = \frac{\sum_i \left(y_i^* - \bar{y}\right)^2}{\sum_i \left(y_i - \bar{y}\right)^2} \qquad [27]$$

L'importante risultato ottenuto dalla [27] può essere interpretato con l'aiuto di un grafico (figura 13).

Figura 13 - Nuvola dei punti, retta di regressione e contributi alla varianza di regressione



$y_i^* - \bar{y}$ è rappresentato graficamente dalla differenza fra l'ordinata alla retta di regressione nel punto x_i e l'ordinata $\bar{y}$. Perciò $\sum_i \left(y_i^* - \bar{y}\right)^2$ è la somma dei quadrati degli n segmenti compresi fra la retta di regressione e la retta $y = \bar{y}$, uno per ciascun x_i (i=1,2,3,.....,n).

Poiché per il metodo dei minimi quadrati é (cfr. la prima equazione della [19]),

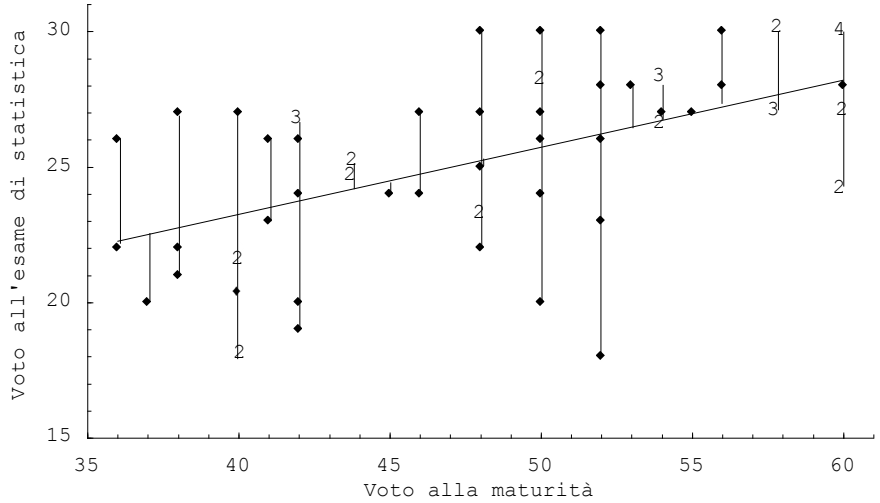
$$\bar{y} = \frac{\sum_i y_i}{n} = a^* + b^* \sum_i \frac{x_i}{n} = \frac{\sum_i \left(a^* + b^* x_i\right)}{n} = \frac{\sum_i y_i^*}{n} \qquad [28]$$

la quantità:

$$\frac{1}{n} \sum_i \left(y_i^* - \bar{y}\right)^2 \qquad [29]$$

è una varianza, essa è nota come varianza di regressione lineare, Var(L), ed esprime la varianza dovuta, spiegata, giustificata, dalla relazione lineare introdotta fra X ed Y.

Figura 14 - Nuvola dei punti, retta di regressione e residui



Dalla figura 13. emerge, tuttavia, che oltre alla dispersione dei valori teorici attorno alla loro media, esiste anche una dispersione dei punti della nuvola attorno alla retta $y = a^* + b^* x$. Ciò è rappresentato in figura 14.

La differenza fra il valore empirico osservato e il corrispondente valore teorico, $y_i - y_i^*$, è detto residuo. Poiché la media dei residui è nulla, per la [28], la quantità:

$$\frac{1}{n}\sum_i \left(y_i - y_i^*\right)^2 \qquad [30]$$

è anch'essa una varianza, ed è nota come varianza residua, Var(R).

Si può mostrare[10] che:

$$\frac{1}{n}\sum_i \left(y_i^* - \bar{y}\right)^2 + \frac{1}{n}\sum_i \left(y_i - y_i^*\right)^2 = \frac{1}{n}\sum_i \left(y_i - \bar{y}\right)^2,$$

ossia che:

$$\sum_i \left(y_i^* - \bar{y}\right)^2 + \sum_i \left(y_i - y_i^*\right)^2 = \sum_i \left(y_i - \bar{y}\right)^2,$$

da cui:

$$\text{Dev(L)} + \text{Dev(R)} = \text{Dev(T)} \qquad [31].$$

La [31] è una formula di scomposizione della devianza ed indica che la devianza del carattere assunto come dipendente è spiegata in parte dalla variabile X, tramite la relazione lineare che lega i caratteri in esame, mentre la parte residua rimane non spiegata..

---

[10] Infatti

$$\begin{aligned}
&\sum_i \left(y_i^* - \bar{y}\right)^2 + \sum_i \left(y_i - y_i^*\right)^2 = \\
&= \sum_i y_i^{*2} + n\bar{y}^2 - 2\bar{y}\sum_i y_i^* + \sum_i y_i^2 + \sum_i y_i^{*2} - 2\sum_i y_i y_i^* = \\
&= 2\sum_i y_i^{*2} + n\bar{y}^2 - 2n\bar{y}^2 + \sum_i y_i^2 - 2\sum_i y_i y_i^* = \\
&= \sum_i y_i^2 - n\bar{y}^2 - 2\left[\sum_i y_i^*\left(y_i - y_i^*\right)\right] = \\
&= \sum_i \left(y_i - \bar{y}\right)^2
\end{aligned}$$

poiché $\sum_i y_i^*\left(y_i - y_i^*\right) = 0$. Per dimostrarlo occorre tener conto che: $y_i^* = a^* + b^* x_i$, e che $a^* = \bar{y} - b^*\bar{x}$. Si ottiene pertanto:

$$\begin{aligned}
&\sum_i y_i^*\left(y_i - y_i^*\right) = \sum_i \left(a^* + b^* x_i\right)\left(y_i - a^* - b^* x_i\right) = \\
&= a^*\sum_i y_i + b^*\sum_i x_i y_i - na^{*2} - a^* b^*\sum_i x_i - a^* b^*\sum_i x_i - b^{*2}\sum_i x_i^2 = \\
&= na^*\bar{y} + b^*\sum_i x_i y_i - na^{*2} - 2na^* b^*\bar{x} - b^{*2}\sum_i x_i^2 = \\
&= n\left(\bar{y} - b^*\bar{x}\right)\bar{y} + b^*\sum_i x_i y_i - n\left(\bar{y} - b^*\bar{x}\right)^2 - 2n\left(\bar{y} - b^*\bar{x}\right)b^*\bar{x} - b^{*2}\sum_i x_i^2 = \\
&= b^*\left(\sum_i x_i y_i - n\bar{x}\bar{y}\right) - b^{*2}\left(\sum_i x_i^2 - n\bar{x}^2\right) = \\
&= b^*\left[\text{Codev(XY)} - b^*\text{Dev(X)}\right] = \\
&= b^*\left[\text{Codev(XY)} - \frac{\text{Codev(XY)}}{\text{Dev(X)}}\text{Dev(X)}\right] = 0
\end{aligned}$$

Poiché è anche $\sum_i y_i^*\left(y_i - y_i^*\right) = \sum_i \left(y_i - y_i^*\right)\left(y_i^* - \bar{y}\right) = 0$, consegue che non c'è correlazione fra valori teorici e residui.

Si osserva che ad un risultato analogo si sarebbe pervenuti operando rispetto alla retta $x = a'^{*} + b'^{*}y$.

Ritornando ad r^2, se ne può ora dare l'interpretazione statistica, dal momento che per la [27] è:

$$0 \le r^2 = \frac{Dev(L)}{Dev(T)} = 1 - \frac{Dev(R)}{Dev(T)} \le 1,$$

r^2, esprimendo il rapporto fra la devianza di regressione e la devianza totale, indica quanta parte della devianza totale della Y è spiegata dalla relazione di dipendenza lineare introdotta fra le variabili. In tale contesto interpretativo r^2 prende il nome di indice di determinazione lineare.

In particolare:

$r^2=0$ se e solo se Dev(L)=0, ossia se $y_i^* = \bar{y}\ \forall i$ e quindi la retta di regressione coincide con $y = \bar{y}$, se si opera sulle variabili X ed Y, o con gli assi, se si opera con gli scarti o gli scarti standardizzati.

$r^2=1$ se e solo se Dev(R)=0, ossia se tutti i punti della nuvola sono allineati e la retta di regressione esprime perfettamente la relazione esistente fra X ed Y.

L'interpretazione è analoga se si esprime X in funzione di Y.

Ritornando ai dati, si osserva che, per la nuvola dei punti "studenti classificati contemporaneamente secondo il voto alla maturità e voto all'esame di statistica", $r^2 = 0{,}578^2 = 0{,}334$, ossia la devianza totale del voto di statistica è spiegata per il 33,4% dalla dipendenza lineare del voto di statistica dal voto alla maturità, ma si può egualmente dire che la devianza del voto all'esame di statistica è spiegata dalla relazione lineare col voto alla maturità per il 33,4%. Resta dunque ben il 66,6% di devianza non spiegata. A questo punto sta al ricercatore decidere se ritenersi soddisfatto della relazione trovata fra voto alla maturità e voto all'esame di statistica, o se al contrario ricercare altre possibili variabili per spiegare la variabilità residua, e proseguire di conseguenza l'indagine.

Nell'uso di r e delle rette di regressione e nella corrispondente interpretazione, occorre cautela.
Dall'esposizione fatta emergono, infatti, alcune considerazioni di ordine generale sull'uso di r e delle rette di regressione.

Non si deve dimenticare che il concetto di "causa" è uno fra i più discussi in filosofia. La dipendenza, intesa come nesso di causalità fra due variabili, può essere ipotizzata ed introdotta solo sulla base dello studio attento del fenomeno analizzato. Dunque: correlazione non è sinonimo di causalità e neppure di dipendenza.

Ammesso che esista una relazione, un nesso, r ne riguarda uno particolare: quello lineare. L'esame grafico dei dati è pertanto indispensabile per rendersi conto della "forma" della relazione.

La rappresentazione grafica dei dati è necessaria anche per analizzare i residui $y_i - y_i^*$. Infatti, perché la retta possa essere proposta per interpolare la relazione fra le variabili occorre che i residui si alternino in modo casuale, non devono cioè esserci residui positivi prevalentemente da una parte del grafico e residui negativi prevalentemente dall'altra; che l'andamento dei residui non dipenda dalla variabile indipendente e che i residui abbiano variabilità pressoché costante al variare della variabile indipendente stessa.

Se tutto ciò non si verifica, l'uso di r e delle rette di regressione è improprio e può essere fuorviante.