TESINA DEL CORSO DI : METODI QUANTITATIVI PER L'ANALISI DEL TERRITORIO

" ANALISI STATISTICA SUGLI IMMOBILI UBICATI NEL COMUNE DI CAMPO SAN MARTINO (PD) "

docente CARLO GRUMENZONI
studente LINDA PERON
matricola 261356 - SPUT

- INTEGRAZIONE -

(FILE ESERCITAZIONE _ MATRICOLA 261356 . XLS)

INDICE

BIBLIOGRAFIA:

- S.Borra, A.di Ciaccio "Statistica, metodologie per le scienze economiche e sociali" McGraw-Hill
- Wonnacott e Wonnacott "Introduzione alla Statistica" F.Angeli
- Dati raccolti presso l'Ufficio Tecnico del Comune di Campo San Martino (PD)

ALTRE FONTI

- www.sister.agenziaentrate.it
- www.comune.camposanmartino.pd.it

---

1 - DATI

Supponiamo di voler indagare sulle caratteristiche abitative degli immobili ubicati nel Comune di Campo San Martino (PD). La popolazione di riferimento è un campione composto di 20 abitazioni iscritte al catasto Urbano alla data del 30/06/2009. Le unità statistiche sono le singole abitazioni. Le variabili di studio sono:

- X SUPERFICIE IMMOBILE → variabile quantitativa continua, espressa in metri quadrati
- Y NUMERO VANI IMMOBILE → variabile quantitativa discreta
- Z VALORE IMMOBILE → variabile quantitativa continua, espressa in migliaia di Euro

Il campione ha portato alla tabella multipla di dati riportata a lato.

	x	y	z
	Superficie (mq)	n. Vani	Valore (migliaia €)
1	45	3	95
2	65	4	110
3	85	5	150
4	58	4	98
5	60	4	105
6	82	5	143
7	78	4	130
8	89	5	150
9	64	4	110
10	150	6	190
11	162	6	200
12	82	5	145
13	55	3	90
14	63	4	95
15	142	6	180
16	178	6	210
17	180	7	250
18	60	4	110
19	75	4	125
20	95	5	183

È opportuno definire le classi considerando che i caratteri, soprattutto X e Z, presentano molte modalità distinte portando ad una difficoltà nella comprensione dei dati osservati - Avremo dunque, intervalli aperti a destra di egual ampiezza:

CARATTERE X: 0 ⊢ 50 | 50 ⊢ 100 | 100 ⊢ 150 | 150 ⊢ 200

CARATTERE Z: 0 ⊢ 50 | 50 ⊢ 100 | 100 ⊢ 150 | 150 ⊢ 200 | 200 ⊢ 250

CARATTERE Y: 2 ⊢ 4 | 4 ⊢ 6 | 6 ⊢ 8

2 - DISTRIBUZIONE DEI CARATTERI E LORO RAPPRESENTAZIONE

Per ottenere una maggior sintesi del fenomeno è possibile considerare per ogni carattere la frequenza con cui le diverse modalità sono state osservate - Avremo:

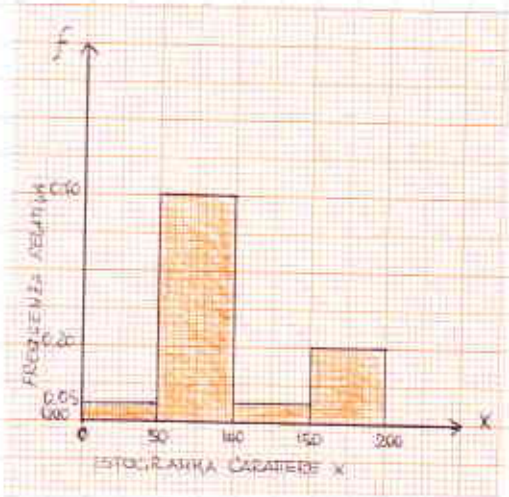
FREQUENZA ASSOLUTA (n_i) di una modalità di un carattere: numero di volte che questa viene osservata nel collettivo.

FREQUENZA RELATIVA (f_i) è il rapporto tra frequenza assoluta e numero del collettivo.

FREQUENZE CUMULATE è la somma della corrispondente frequenza e di tutte quelle relative alle classi precedenti.

2.1 CARATTERE X: SUPERFICIE IMMOBILE (MQ)

	c_i	n_i	f_i	N_i	F_i
0 ⊢ 50	25	1	0,05	1	0,05
50 ⊢ 100	75	14	0,70	15	0,75
100 ⊢ 150	125	1	0,05	16	0,80
150 ⊢ 200	175	4	0,20	20	1,00
		20	1,00		

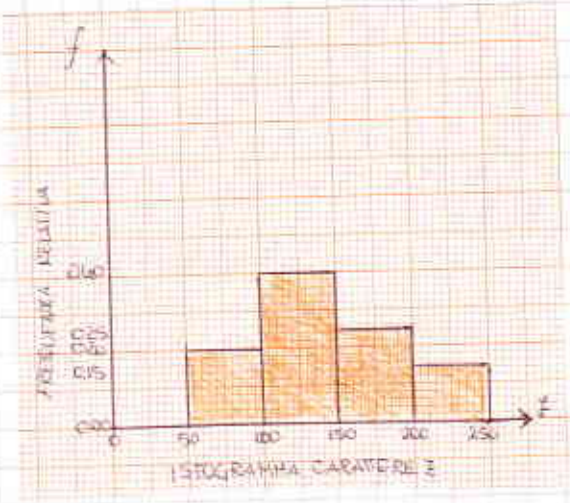


Dalla distribuzione frequenziale relativa si nota che il 70% degli immobili ha una superficie compresa tra 0 e 100 mq mentre il 75% ha una superficie

inferiore a 150 mq.
I dati ottenuti vengono rappresentati nell'istogramma a lato.

CARATTERE Z : VALORE IMMOBILE (MIGLIAIA DI €)

	c_i	n_i	f_i	N_i	F_i
0 ⊢ 50	25	0	0,00	0	0,00
50 ⊢ 100	75	4	0,20	4	0,20
100 ⊢ 150	125	8	0,40	12	0,60
150 ⊢ 200	175	5	0,25	17	0,85
200 ⊢ 250	225	3	0,15	20	1,00
		20	1,00		



Dalla distribuzione frequenziale si nota che il 40% degli immobili ha un valore compreso tra 100.000 e 150.000 € mentre il 60% di immobili ha un valore inferiore ai 150.000 €

(2.2) Supponiamo di analizzare congiuntamente i dati del carattere y e del carattere z tenendo conto della suddivisione in classi. Otteniamo della frequenze congiunte. La tabella è la seguente:

	2 ⊢ 4	4 ⊢ 6	6 ⊢ 8	
0 ⊢ 50	0	0	0	0
50 ⊢ 100	2	2	0	4
100 ⊢ 150	0	8		8
150 ⊢ 200	0	3	2	5
200 ⊢ 250	0	0	3	3
	2	13	5	20

3 - SINTESI DELLA DISTRIBUZIONE DI UN CARATTERE - LE MEDIE

Per descrivere le caratteristiche essenziali di un carattere, utilizzeremo degli indici di posizione. Tali indici si distinguono in:
MEDIE ANALITICHE calcolate mediante operazioni algebriche su valori del carattere. Sono MEDIA ARITMETICA MEDIA PONDERATA TRIMMED MEAN
MEDIE DI POSIZIONE calcolate mediante supposizioni logiche. Sono MODA MEDIANA
Si procederà al calcolo di tali indici per le tre variabili.

(3.1) MEDIA ARITMETICA : Si cerca il valore caratteristico attorno al quale si posizionano i fenomeni della distribuzione. Si ottiene mediante la somma dei valori osservati diviso il numero. L'indice è applicabile su caratteri quantitativi.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{x} = \frac{45+65+85+58+80+82+\dots}{20} = \frac{1868}{20} = 93{,}40 \text{ mq}$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad \bar{y} = \frac{3+4+5+4+4+5+4+5+\dots}{20} = \frac{94}{20} = 4{,}70 \text{ vani}$$

$$\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i \qquad \bar{z} = \frac{95+110+150+98+105+\dots}{20} = \frac{2869}{20} = 143{,}45 \text{ migliaia di €}$$

MEDIANA : è il valore della variabile di riferimento tale che, metà delle unità statistiche hanno modalità inferiori o uguali ad essa e metà superiore. L'indice è applicabile su caratteri quantitativi e qualitativi.
È necessario ordinare i valori in modo crescente. Prendo il valore posto in posizione centrale ma nel nostro caso abbiamo un numero pari di unità, pertanto avremo due medie. Con apposita formula si ricava la media esatta.

$$Me_x = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\left(\frac{n}{2}+1\right)}\right) \qquad Me_x = \frac{78+82}{2} = 80{,}00 \text{ mq}$$

$$Me_y = \frac{1}{2}(4+5) = 4{,}50 \text{ n° vani}$$

$$Me_z = \frac{1}{2}(130+143) = 136{,}50 \text{ migliaia di €}$$

xi	yi	zi
Superficie (mq)	n. Vani	Valore (migliaia €)
45	3	90
55	3	95
58	4	95
60	4	98
60	4	105
63	4	110
64	4	110
65	4	110
75	4	125
78	4	130 ← Mex Mey
82	5	143 Mez
82	5	145
85	5	150
89	5	150
95	5	180
142	6	183
150	6	190
162	6	200
178	6	210
180	7	250

Se la mediana divide a metà l'insieme delle unità ordinate per grandezza, possiamo anche immaginare di dividere la distribuzione in cento parti, ognuna delle quali contenente le stesse unità, chiamando i valori di suddivisione PERCENTILI. Quindi la mediana è considerata il 50-esimo percentile. Gli altri percentili più utilizzati sono il 25-esimo (PRIMO QUARTILE) e il 75 esimo (TERZO QUARTILE). Il primo e il terzo quartile individuano un intervallo di cinquanta unità statistiche e può essere considerato una misura della dispersione dei valori più frequentemente osservati. Avremo:

$$Q_{1x} = \frac{60+63}{2} = 61{,}50 \qquad Q_{3x} = \frac{95+142}{2} = 118{,}50$$

$$Q_{1y} = 4 \qquad Q_{3y} = \frac{5+6}{2}$$

$$Q_{1z} = \frac{105+110}{2} = 107{,}50 \qquad Q_{3z} = \frac{180+183}{2} = 181{,}50$$

MODA : è la modalità della distribuzione che si presenta con la massima frequenza. Tale indice è applicabile su caratteri quantitativi e qualitativi. La moda per il carattere z è 110, per il carattere y è 4 mentre per il carattere x ha uno scarso significato dato che i valori del carattere sono tutti diversi. In questo caso la moda ha senso se si considera il carattere suddiviso in classi. La classe modale per il carattere x è 50 ⊢ 100

CARATTERE x	n_i	f_i	
0 ⊢ 50	1	0,05	
50 ⊢ 100	14	0,70	→ CLASSE MODALE
100 ⊢ 150	1	0,05	
150 ⊢ 200	4	0,20	
	20	1,00	

(32) Analizziamo ora il carattere X suddiviso in classi; le classi con intervalli chiusi a sinistra ha ampiezza regolare - I dati sono ripresi dal PG. 2 :

CARATTERE X	c_i	n_i	f_i	N_i	F_i	
[0, 50)	25	0	0,00	0	0,00	
[50, 100)	75	4	0,20	4	0,20	
[100, 150)	125	8	0,40	12	0,60	→ CLASSE MODALE
[150, 200)	175	5	0,25	17	0,85	
[200; 250)	225	3	0,15	20	1,00	
		20	1,00			

MEDIA ARITHETICA:

$$\bar{x} = \frac{(25 \cdot 0) + (75 \cdot 4) + (8 \cdot 125) + (175 \cdot 5) + (225\ 3)}{20} = \frac{0 + 300 + 1000 + 875 + 675}{20} = \frac{2850}{20} = 142{,}5$$

MODA : la classe modale è 100 ⊢ 150 - Ha la maggior frequenza

MEDIANA : si deve trovare la classe dove si posiziona la mediana. Risulta necessaria la frequenza assoluta cumulata che mi aiuta a individuare la classe mediana : 100 ⊢ 150 - Il valore mediano si calcola con la seguente formula :

$$Me \approx Im + \left(\frac{0{,}5 - F_{m-1}}{F_m - F_{m-1}}\right) \cdot \Delta m$$

Im = estremo inferiore della classe mediana
F_{m-1} = frequenza relativa cumulata fino alla classe precedente a quella Mediana
F_m = frequenza relativa cumulata
Δm = ampiezza classe mediana

$$Me \approx 100 + \left(\frac{0{,}5 - 0{,}20}{0{,}60 - 0{,}20}\right) \cdot 50 = 137{,}50 \quad (\text{mila€})$$

Trovo i quartili Q1 e Q3 con formula simile alla precedente salvo il valore fisso che varia a seconda del quartile
La classe che contiene il primo quartile è 100 ⊢ 150 mentre quella che contiene il terzo quartile è 150 ⊢ 200 -

$$Q_1 \approx 100 + \left(\frac{0{,}25 - 0{,}20}{0{,}60 - 0{,}20}\right) \cdot 50 = 106{,}50 \quad \text{mila€}$$

$$Q_3 \approx 150 + \left(\frac{0{,}75 - 0{,}60}{0{,}85 - 0{,}60}\right) \cdot 50 = 180 \quad \text{mila€}$$

4_ SINTESI DELLA DISTRIBUZIONE DEI CARATTERI = LA VARIABILITA'

Si prosegue analizzando la variabilità del carattere. La variabilità ci consente di analizzare le caratteristiche di una distribuzione. Essa aumenta all'aumentare della diversità del carattere. La variabilità si misura mediante indici quali: VARIANZA - COEFFICIENTE DI VARIAZIONE - VARIANZA RELATIVA GINI = GINI RELATIVO.

(4.1) Per ciascun carattere calcolo gli indici sopra citati.

CARATTERE X: $\sigma^2_x = \frac{1}{n}\sum_{i=1}^{n}(x-\bar{x})^2$ dove $\bar{x} = 93{,}40$ mg

$$\sigma^2_x = \frac{(45-93{,}40)^2 + (65-93{,}40)^2 + (85-93{,}40)^2 + (58-93{,}40)^2 + (60-93{,}40)^2 + \dots}{20}$$

$$= \frac{35152{,}80}{20} = 1787{,}64 \text{ mg}^2$$

CARATTERE Y: $\sigma^2_y = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2$ dove $\bar{y} = 4{,}70$ vani medi

$$\sigma^2_y = \frac{(3-4{,}70)^2 + (4-4{,}70)^2 + (5-4{,}70)^2 + (4-4{,}70)^2 + (4-4{,}70)^2 + (5-4{,}70)^2 + \dots}{20}$$

$$= \frac{22{,}20}{20} = 1{,}11 \text{ vani}^2$$

CARATTERE Z: $\sigma^2_z = \frac{1}{n}\sum_{i=1}^{n}(z-\bar{z})^2$ dove $\bar{z} = 143{,}45$ migliaia €

$$\sigma^2_z = \frac{(95-143{,}45)^2 + (110-143{,}45)^2 + (150-143{,}45)^2 + (98-143{,}45)^2 + \dots}{20}$$

$$= \frac{39108{,}95}{20} = 1955{,}45 \text{ (migliaia €}^2\text{)}$$

La varianza non ha la stessa unità di misura dei valori di distribuzione. Si ricaverà la DEVIAZIONE STANDARD.

CARATTERE X $\sigma_x = \sqrt{\sigma^2_x}$ $\sigma_x = \sqrt{1787{,}64} = 42{,}28$

CARATTERE Y $\sigma_y = \sqrt{\sigma^2_y}$ $\sigma_y = \sqrt{1{,}11} = 1{,}05$

CARATTERE Z $\sigma_z = \sqrt{\sigma^2_z}$ $\sigma_z = \sqrt{1955{,}45} = 44{,}22$

La varianza e la deviazione standard sono indici di variabilità assoluta che risentono dell'ordine di grandezza e dell'unità di misura. Per un confronto è necessario portare il tutto ad una scala priva di unità di misura e che non risenta dell'ordine di grandezza. Si calcolerà il COEFFICIENTE DI VARIAZIONE:

CARATTERE X $CV_x = \frac{\sigma_x}{\bar{x}}$ $CV_x = \frac{42{,}28}{193{,}401} = 0{,}45$

CARATTERE Y $CV_y = \frac{\sigma_y}{\bar{y}}$ $CV_y = \frac{1{,}05}{14{,}701} = 0{,}22$

CARATTERE Z $CV_z = \frac{\sigma_z}{\bar{z}}$ $CV_z = \frac{44{,}22}{143{,}451} = 0{,}31$

Il COEFFICIENTE DI VARIAZIONE mostra che la variabilità del carattere y è minore rispetto il carattere z mentre il carattere x varia più degli altri. Lo studio della VARIANZA RELATIVA mi dà risultati coerenti con quelli ottenuti precedentemente:

CARATTERE X $\quad VR_x = \dfrac{\sigma^2_x}{(n-1)\,\bar{x}^2} \qquad VR_x = \dfrac{1787,64}{(20-1)\cdot 93,40^2} = 0,01$

CARATTERE Y $\quad VR_y = \dfrac{\sigma^2_y}{(n-1)\,\bar{x}^2} \qquad VR_y = \dfrac{1,11}{(20-1)\,93,40^2} = 0,0026$

CARATTERE Z $\quad VR_z = \dfrac{\sigma^2_z}{(n-1)\,\bar{z}^2} \qquad VR_z = \dfrac{1955,45}{(20-1)\,143,45^2} = 0,0050$

La varianza relativa ci dimostra una bassa variabilità dei caratteri presi in esame e riconferma quanto ricavato dal coefficiente di variazione.

(4.2) Analizzeremo la variabilità per un carattere diviso in classi.
Riproppongo la tabella delle frequenze riferita al carattere z e ricavo la varianza, deviazione standard, coefficiente di variazione e varianza relativa

CARATTERE z	cl	n_i	f_i
[0,50)	25	0	0,00
[50,100)	75	4	0,20
[100,150)	125	8	0,40
[150,200)	175	5	0,25
[200,250)	225	3	0,15
		20	1,00

VARIANZA:

$$\sigma^2_z = \frac{1}{n}\sum_{i=1}^{n}(z-\bar{z})^2 \cdot n_i \qquad \sigma^2_z = \sum_{i=1}^{n}(z-\bar{z})^2 \cdot f_i$$

dapprima calcolo la media, moltiplicando il valore centrale della classe con la frequenza relativa f_i

$$\bar{z} = (25\cdot 0)+(75\cdot 0,20)+(125\cdot 0,40)+(175\cdot 0,25)+(225\cdot 0,15) = 15+50+43,75+33,75 = 142,50$$

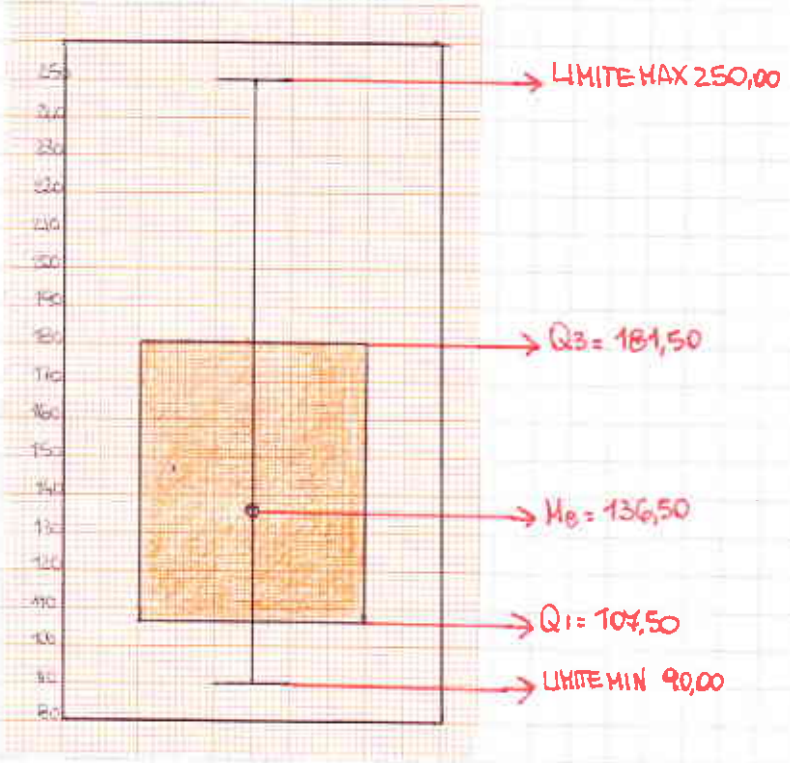
$$\sigma^2_z = (25-142,50)^2\cdot 0 + (75-142,50)^2\cdot 4 + (125-142,50)^2\cdot 8 + \ldots = 2318,75 \text{ migliaia di } €^2$$

DEVIAZIONE STANDARD: $\quad \sigma_z = \sqrt{\sigma^2_z} \qquad \sigma_z = \sqrt{2318,75} = 48,15$

COEFFICIENTE DI VARIAZIONE: $\quad CV_z = \dfrac{\sigma_z}{\bar{z}} \qquad CV_z = \dfrac{48,15}{142,50} = 0,3379$

VARIANZA RELATIVA: $\quad VR_z = \dfrac{\sigma^2_z}{(n-1)\,\bar{z}^2} \qquad VR_z = \dfrac{2318,75}{(20-1)\,142,50^2} = 0,0060$

(4.3) Si può rappresentare la distribuzione dei dati mediante un BOX PLOT. Si recuperano i dati precedentemente calcolati a FG 3, per il carattere z:
$Me_z = 136,50$ migliaia di €
$Q_{1z} = 107,50$ migliaia di € (frequenza percentuale cumulata pari al 25%)
$Q_{3z} = 181,50$ migliaia di € (frequenza percentuale cumulata pari al 75%)
La differenza interquartile è: $181,50 - 107,50 = 74,00$.
Il limite massimo è 250 mentre il limite minimo di 90.



(4.4) La concentrazione evidenzia in modo più efficace la variabilità del carattere trasferibile. Si considera la variabile z, carattere trasferibile. Il procedimento seguito è il seguente:

1. modalità ordinate in modo crescente z_i
2. calcolo delle somme cumulate, sommando ad una ad una con le precedenti (A_i)
3. si ricava Q_1 mediante il rapporto tra A_1 e la ricchezza totale
4. si ricava F_1 con il rapporto tra ricchezza totale e il numero di osservazioni

valore	z	Zi	Ai	Qi	Fi	Fi-Qi
1	95	90	90	0,0314	0,05	0,0186
2	110	95	185	0,0645	0,10	0,0355
3	150	95	280	0,0976	0,15	0,0524
4	98	98	378	0,1318	0,20	0,0682
5	105	105	483	0,1684	0,25	0,0816
6	143	110	593	0,2067	0,30	0,0933
7	130	110	703	0,2450	0,35	0,1050
8	150	110	813	0,2834	0,40	0,1166
9	110	125	938	0,3269	0,45	0,1231
10	190	130	1068	0,3723	0,50	0,1277
11	200	143	1211	0,4221	0,55	0,1279
12	145	145	1356	0,4726	0,60	0,1274
13	90	150	1506	0,5249	0,65	0,1251
14	95	150	1656	0,5772	0,70	0,1228
15	180	180	1836	0,6399	0,75	0,1101
16	210	183	2019	0,7037	0,80	0,0963
17	250	190	2209	0,7700	0,85	0,0800
18	110	200	2409	0,8397	0,90	0,0603
19	125	210	2619	0,9129	0,95	0,0371
20	183	250	2869	1,0000	1,00	0,0000
						1,7091

↓ SOMME CUMULATE

INDICE DI GINI:

$$G_n = \sum_{i=1}^{n} (F_i - Q_i)$$

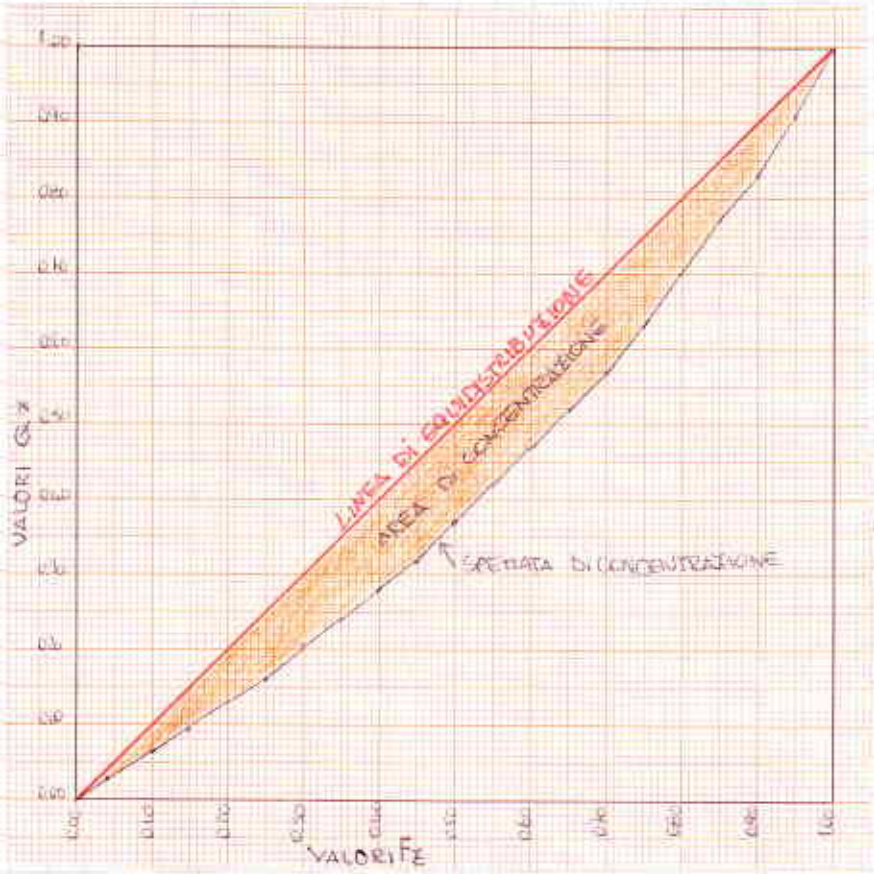
$G_n = 1,7091$

GINI RELATIVO

$$g = \frac{G_n}{G_{max}}$$

dove $G_{max} = \frac{n-1}{2}$

$$g = \frac{1,7091}{\frac{20-1}{2}} = 0,1799$$



Con le coppie di valori Qi e Fi presenti nella tabella precedentemente riportata (PG 7), si può costruire un grafico, dove sull'asse delle ascisse si rappresentano i valori Fi ed in quello delle ordinate i valori Qi. Per ogni coppia di valori si disegna un punto, e la loro unione crea una spezzata di concentrazione detta CURVA DI LORENZ.

(4.5)

Si vuole studiare il carattere Z suddiviso in classi.
Il procedimento seguito è
1. calcolo il valore Z come prodotto della frequenza assoluta e del valore centrale di classe
2. successivamente si dispone in ordine crescente (Zi)
3. calcolo delle somme cumulate (Ai)
4. Trovare Qi e Fi

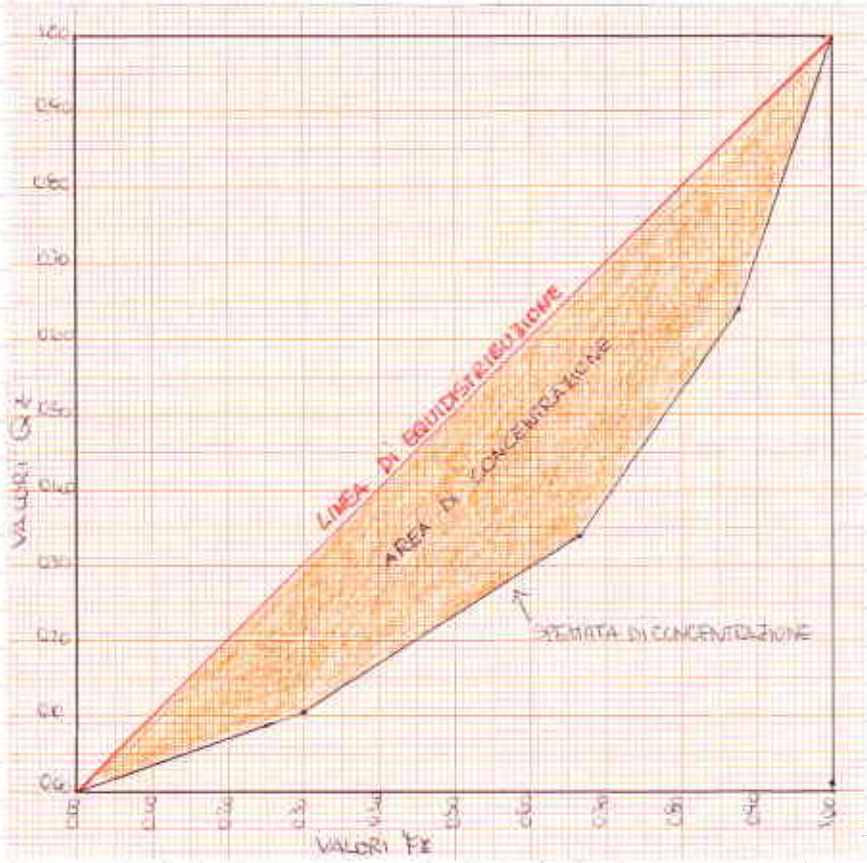
	c	n	z= z*n	Zi	Ai	Qi	Fi	Fi-Qi
[0,50)	25	0	0	0	0	0,0000	0,0000	0,0000
[50,100)	75	4	300	300	300	0,1053	0,3000	0,1947
[100,150)	125	8	1000	675	975	0,3421	0,6750	0,3329
[150,200)	175	5	875	875	1850	0,6491	0,8750	0,2259
[200,250]	225	3	675	1000	2850	1,0000	1,0000	0,0000
		20	2850					0,7535

INDICE DI GINI $\quad Gn = \sum_{i=1}^{n} (Fi - Qi)$

$Gn = 0,7535$

GINI RELATIVO $\quad g = \frac{Gn}{Gmax}$ $\quad$ dove $Gmax = \frac{2c-1}{2}$

$g = \frac{0,7535}{\frac{2c-1}{2}} = 0,0793$



(CURVA DI LORENZ PER IL CARATTERE Z SUDDIVISO IN CLASSI)

5 - ANALISI DELL'ASSOCIAZIONE TRA DUE CARATTERI:

A questo punto si analizzerà le relazioni che ogni carattere può avere con gli altri - Gli indici utilizzati per lo studio di tali relazioni si distinguono in: indici di associazione utilizzati per caratteri qualitativi e sono: Chi-QUADRATO χ^2, CONTINGENZA QUADRATICA Φ^2 - V DI CRAMER, e indici di correlazione utilizzati per caratteri quantitativi e sono: COVARIANZA σ_{xy} - COEFFICIENTE DI CORRELAZIONE LINEARE ρ_{xy} -

(5.1) Ritornando al nostro esercizio, si consideri le variabili y e z entrambe quantitative.

COVARIANZA $\sigma_{yz} = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})(z-\bar{z})$ dove $\bar{y} = 4,70$ $\bar{z} = 143,45$

$$\sigma_{yz} = \frac{1}{20}\left[(3-4,70)(95-143,50)+(4-4,70)(110-143,50)+(5-4,70)(150-143,50)+\dots\right]$$

$$= \frac{887,70}{20} = 44,34$$

σ_{yz} assume valori all'interno dell'intervallo: $-\sigma_y\sigma_z \leq \sigma_{zy} \leq \sigma_y\sigma_z$

CORRELAZIONE $\rho_{yz} = \frac{\sigma_{yz}}{\sigma_y\,\sigma_z}$ dove $\sigma_y = 1,05$ $\sigma_z = 44,22$

$$\rho_{yz} = \frac{44,39}{1,05\ 44,22} = 0,9560$$

ρ_{yz} ha un valore vicino a 1 _ Ciò dimostra che tra y e z vi è una spiccata dipendenza lineare - Poiché l'indice è positivo la relazione lineare è crescente

(5.2) Ora, si considerino le variabili y e x.

COVARIANZA: $\sigma_{yx} = \frac{1}{n}\sum(y-\bar{y})(x-\bar{x})$ oppure $\sigma_{yx} = \left(\frac{1}{n}\sum_{i=1}^{n} x \cdot y\right) - \bar{x}\cdot\bar{y}$

dove $\bar{x} = 93{,}40$ mq $\quad \bar{y} = 4{,}70$ vani medi

$$\sigma_{yx} = \frac{1}{20}\left(3-4{,}70\right)(45-93{,}40) + (4-4{,}70)(65-93{,}40) + (5-4{,}70)(85-93{,}40) + \dots$$
$$= \frac{829{,}40}{20} = 41{,}47$$

σ_{yx} assume valori all'interno dell'intervallo $-\sigma_y\sigma_x \leq \sigma_{yx} \leq \sigma_y\sigma_x$

CORRELAZIONE: $\rho_{yx} = \frac{\sigma_{yx}}{\sigma_y\,\sigma_x}$ dove $\sigma_y = 1{,}05 \quad \sigma_x = 42{,}28$

$$\rho_{yx} = \frac{41{,}47}{1{,}05\ 42{,}28} = 0{,}9341$$

Anche tra il carattere x e y c'è una spiccata dipendenza lineare. La relazione lineare risulta crescente vista la positività dell'indice.

(5.3) Si considerino le variabili x e z.

COVARIANZA $\quad \sigma_{xz} = \frac{1}{n}\sum(x-\bar{x})(z-\bar{z})$ oppure $\sigma_{xz} = \left(\frac{1}{n}\sum x\cdot z\right) - \bar{x}\cdot\bar{z}$

dove $\bar{x} = 93{,}40$ vani medi $\quad \bar{z} = 143{,}45$ migliaia di €

$$\sigma_{xz} = \frac{1}{20}\left(45-93{,}40\right)(95-143{,}45) + (65-93{,}40)(110-143{,}45) + \dots$$
$$= \frac{35\,474{,}50}{20} = 1773{,}77$$

σ_{xz} assume valori all'interno dell'intervallo $-\sigma_x\sigma_z \leq \sigma_{xz} \leq \sigma_x\sigma_z$

CORRELAZIONE: $\rho_{xz} = \frac{\sigma_{xz}}{\sigma_x\sigma_z}$ dove $\sigma_x = 42{,}28 \quad \sigma_z = 44{,}22$

$$\rho_{xz} = \frac{1773{,}77}{42{,}28 \cdot 44{,}22} = 0{,}9487$$

Tra il carattere x e il carattere z vi è dipendenza lineare. La relazione anche in questo caso risulta crescente vista la positività di ρ_{xy}.

(5.4) Si considerano i caratteri y e z suddivisi in classi. L'analisi dell'associazione tra questi due caratteri si è svolta nel seguente modo:

1. calcolo tabella delle frequenze congiunte
2. calcolo la tabella delle frequenze artificiali: $n^*_{zy} = \frac{n_z \cdot n_y}{n}$
3. trovo $\chi^2 = \sum_z\sum_y \frac{n_{zy}^2}{n^*_{zy}} - n$ o $\chi^2 = \sum_{i=1}^{z}\sum_{j=1}^{y} \frac{(n_{zy} - n^*_{zy})^2}{n^*_{zy}}$
4. ricavo la CONTINGENZA QUADRATICA
5. ricavo il V DI CRAMER

z \ y	[2,4)	[4,6)	[6,8)	tot.
[0,50)	0	0	0	0
[50,100)	2	3	0	5
[100,150)	0	6	1	7
[150,200)	0	2	2	4
[200,250)	0	0	4	4
tot.	2	11	7	20

(FREQUENZE CONGIUNTE)

z \ y	[2,4)	[4,6)	[6,8)	tot
[0,50)	0,00	0,00	0,00	0,00
[50,100)	0,40	2,60	1,00	4,00
[100,150)	0,80	5,20	2,00	8,00
[150,200)	0,50	3,25	1,25	5,00
[200,250)	0,30	1,95	0,75	3,00
tot	2,00	13,00	5,00	20,00

(TABELLA DI INDIPENDENZA STOCASTICA)

	[2,4)	[4,6)	[6,8)	
[0,50)	0,00	0,00	0,00	
[50,100)	6,40	0,14	1,00	
[100,150)	0,80	1,51	2,00	
[150,200)	0,50	0,02	0,45	
[200,225)	0,30	1,95	6,75	
	8,00	3,62	10,20	21,82

(TABELLA CONTINGENZA QUADRATICHE RELATIVE)

Quindi $\chi^2 = 21,82$, $\chi^2 \neq 0$ pertanto vi è associazione tra i due caratteri.

CONTINGENZA QUADRATICA: è l'indice che normalizza il CHI-QUADRATO:
$\Phi^2 = \chi^2/n$, quindi:

$$\Phi^2 = \frac{21,82}{20} = 1,091$$

V DI CRAMÉR: è un indice relativo che ci permette di capire chiaramente se vi è tra i due caratteri, dipendenza (se più vicino a 1) o indipendenza (se tendente allo 0).

$$V = \sqrt{\frac{\chi^2}{n \min[(H-1)(K-1)]}} \qquad 0 \leq V \leq 1$$

$$V = \sqrt{\frac{21,82}{20\ 2}} = 0,7385$$

Si conclude affermando che tra i due caratteri vi è una buona dipendenza.

Calcolo ora la correlazione che sussiste tra x e y suddivisi in classi; per entrambe le variabili si riscrive il valore centrale di classe e successivamente le corrispettive medie di $\bar{y}$ e $\bar{z}$.

CARATTERE y: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y \cdot n$ $\qquad \bar{y} = \frac{(2,5 \cdot 2) + (4,5 \cdot 13) + (6,5\ 5)}{20} = 4,80$

CARATTERE z: (recupero da PG6) $\bar{z} = 142,50$

Calcolo la VARIANZA

CARATTERE y: $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y-\bar{y})^2 n$ $\qquad \sigma_y^2 = \frac{(2,5-4,80)^2 \cdot 2 + (4,5-4,80)^2 \cdot 13 + (6,5-4,80)^2 \cdot 5}{20}$

$= 1,31$ vani^2

CARATTERE z: (recupero da PG6) $\sigma_z^2 = 2378,75$ migliaia di €^{2}

Calcolo la COVARIANZA:

$$\sigma_{zy} = \frac{1}{n}\sum_{i=1}^{z}\sum_{i=1}^{y}(z-\bar{z})(y-\bar{y}) \cdot n$$

$$\sigma_{zy} = \frac{1}{20}(25-142,50)(2,5-4,80) \cdot 0 + (25-142,50)(4,5-4,80) \cdot 0 + (25-142,50)(6,5-4,80) \cdot 0 + (75-142,50)(2,5-4,80) \cdot 2 + \dots = 44,75$$

z \ y	[2,4) 2,5	[4,6) 4,5	[6,8) 6,5	
[0,50) → 25	0	0	0	0
[50,100) → 75	2	2	0	4
[100,150) → 125	0	8	0	8
[150,200) → 175	0	3	2	5
[200,250) → 225	0	0	3	3
	2	13	5	20

CORRELAZIONE : $\rho_{zy} = \frac{\sigma_{zy}}{\sigma_z \sigma_y}$ $\rho_{zy} = \frac{44,75}{\sqrt{1,31} \cdot \sqrt{2318,75}} = 0,91$

Si può concludere che tra i due caratteri, divisi in classi, vi è una buona dipendenza. La relazione risulta lineare e crescente. Confrontando i dati di PG9 e quelli appena ottenuti si conclude che non vi è differenza nel considerare i dati in classi o grezzi

6 - STIMA CORRELAZIONE PARZIALE PER TRE VARIABILI :

(recupero dati da PG_9-10) COEFFICIENTI DI CORRELAZIONE TOTALI :
- Correlazione totale VANI - VALORE $\rho_{yz} = 0,9560$
- Correlazione totale VALORE - SUPERFICIE $\rho_{zx} = 0,9487$
- Correlazione totale VANI - SUPERFICIE $\rho_{yx} = 0,9341$

COEFFICIENTI DI CORRELAZIONE PARZIALE :

$$r_{zy \cdot x} = \frac{\rho_{zy} - (\rho_{zx} \cdot \rho_{yx})}{\sqrt{1-\rho_{yx}^2} \cdot \sqrt{1-\rho_{zx}^2}}$$

$$r_{zy \cdot x} = \frac{0,9560 - (0,9487 \cdot 0,9341)}{\sqrt{1-0,9560^2} \cdot \sqrt{1-0,9487^2}} = \frac{0,0698}{0,2934 \cdot 0,3162} = 0,7524$$

$$r_{zx \cdot y} = \frac{\rho_{zx} - (\rho_{zy} \cdot \rho_{xy})}{\sqrt{1-\rho_{xy}^2} \cdot \sqrt{1-\rho_{zy}^2}}$$

$$r_{zx \cdot y} = \frac{0,9487 - (0,9560 \cdot 0,9341)}{\sqrt{1-0,9341^2} \cdot \sqrt{1-0,9560^2}} = \frac{0,0557}{0,3570 \cdot 0,2934} = 0,5318$$

$$r_{yx \cdot z} = \frac{\rho_{yx} - (\rho_{yz} \cdot \rho_{xz})}{\sqrt{1-\rho_{xz}^2} \cdot \sqrt{1-\rho_{yz}^2}}$$

$$r_{yx \cdot z} = \frac{0,9341 - (0,9560 \cdot 0,9487)}{\sqrt{1-0,9487^2} \cdot \sqrt{1-0,9560^2}} = \frac{0,0271}{0,3162 \cdot 0,2934} = 0,2921$$

Si può notare che la correlazione tra le variabili VANI - SUPERFICIE è influenzata dall'effetto della variabile VALORE, infatti il $\rho_{yx \cdot z}$ tolto l'effetto della variabile valore è pari a 0,2921, mentre l'analogo coefficiente ottenuto senza togliere l'effetto della variabile VALORE vale $\rho_{yx} = 0,9341$ (PG10).
Considerazioni simili si possono fare anche osservando gli altri casi, tuttavia osservando i vari indicatori di correlazione parziale si può che la correlazione tra le variabili VALORE - VANI si mantiene ancora "abbastanza elevata" anche togliendo l'effetto della variabile SUPERFICIE (infatti passa da 0,9560 a 0,5318)

7 - PROBABILITÀ

Supponiamo che la variabile valore ξ sia distribuita secondo una variabile aleatoria normale.

IPOTESI DI NORMALITÀ: $\xi \sim N(\hat{\mu}; \hat{\sigma}^2)$

dove: $\hat{\mu} = \bar{\xi} = 143{,}45$

$\hat{\sigma}^2 = S^2_\xi = 1955{,}45$

Calcolo la PROBABILITÀ CHE IL VALORE SIA INFERIORE O UGUALE A 100.000 €

$P(\xi \leq 100.000)$

$P\left(Y \leq \frac{\xi - \mu}{\sigma}\right) \Rightarrow \quad P\left(Y \leq \frac{100 - 143{,}45}{\sqrt{1955{,}45}}\right)$ dove $Y \sim N(0;1)$

$P(Y \leq -0{,}982)$ tabella $\rightarrow 0{,}1635$

(7.1) Calcolo gli intervalli di confidenza per la media del carattere ξ; prima però mi trovo il valore dello stimatore della varianza campionaria corretta è

$$S^2_{c_\xi} = \frac{n}{n-1} \cdot S^2_\xi \rightarrow S^2_{c_\xi} = \frac{20}{20-1} \cdot 195{,}45 = 2058{,}37$$

L'intervallo di confidenza di dimensione 0,95 e quindi $d = 0{,}05$ (dove $1 - 0{,}95 = d$ mi da il livello di significatibilità) =

$$\left[\bar{\xi} - t_{d/2}\frac{S}{\sqrt{n}} \; ; \; \bar{\xi} + t_{d/2}\frac{S}{\sqrt{n}}\right] \rightarrow \left[143{,}45 \pm 2{,}09 \cdot \frac{\sqrt{2058{,}37}}{\sqrt{20}}\right]$$

$$\left[122{,}25 \; , \; 164{,}65\right]$$

L'intervallo di confidenza per la media di dimensione 0,95 è $[122{,}25, 164{,}65]$.

7.2 PROBLEMA TERRITORIALE DI APPLICAZIONE DEL MODELLO BINOMIALE:

Si consideri la variabile X definita come "numero di abitazioni del quartiere Q avente titolo di godimento L". Si pone n come numero di abitazioni del quartiere Q e con parametro incognito θ, la probabilità che nel quartiere ogni abitazione abbia un titolo di godimento L. Considerato tutte le abitazioni identiche e indipendenti rispetto al titolo di godimento, X si distribuisce secondo la variabile casuale binomiale con parametri n e θ.
Premesso ciò, la probabilità che nel quartiere Q vi siano x abitazioni aventi titolo di godimento L, assume il valore:

$$P(X=x) \rightarrow p(x) = \binom{n}{x} \theta^{x} (1-\theta)^{n-x}$$

posto, per esempio θ uguale a 0,05 ed n pari a 100, la probabilità che non vi siano abitazioni con titolo di godimento L è:

$$P(X=0) \rightarrow p(0) = \binom{100}{0} 0{,}05^{0} (1-0{,}05)^{100-0}$$

È pertanto poco probabile che nel quartiere Q non vi siano abitazioni con titolo di godimento L.

8 - REGRESSIONE LINEARE

Supponiamo che tra il carattere y e il carattere z vi sia una relazione di tipo lineare ossia che la variabile z possa essere scritta in funzione di y come segue $z = \beta_0 + \beta_1 y + \varepsilon$. Il valore di z in corrispondenza di y sarà $\hat{z} = \hat{\beta}_0 + \hat{\beta}_1 y$ dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime dei coefficienti di regressione. Considereremo pertanto la variabile y (NUMERO VANI) come esplicativa o indipendente e la variabile z (VALORE IMMOBILE) come dipendente o di risposta. I coefficienti di regressione sono

E.1

$$\hat{\beta}_1 = \frac{\sigma_{yz}}{\sigma_y^2} \qquad \hat{\beta}_1 = \frac{44,39}{1,11} = 39,99$$

$$\hat{\beta}_0 = \bar{z} - \hat{\beta}_1 \bar{y} \qquad \hat{\beta}_0 = 143,45 - 39,99 \cdot 4,70 = -44,503$$

Il COEFFICIENTE DI DETERMINAZIONE R^2_{zy}, pari a r^2_{zy}, è un indice che mi indica la bontà del modello.

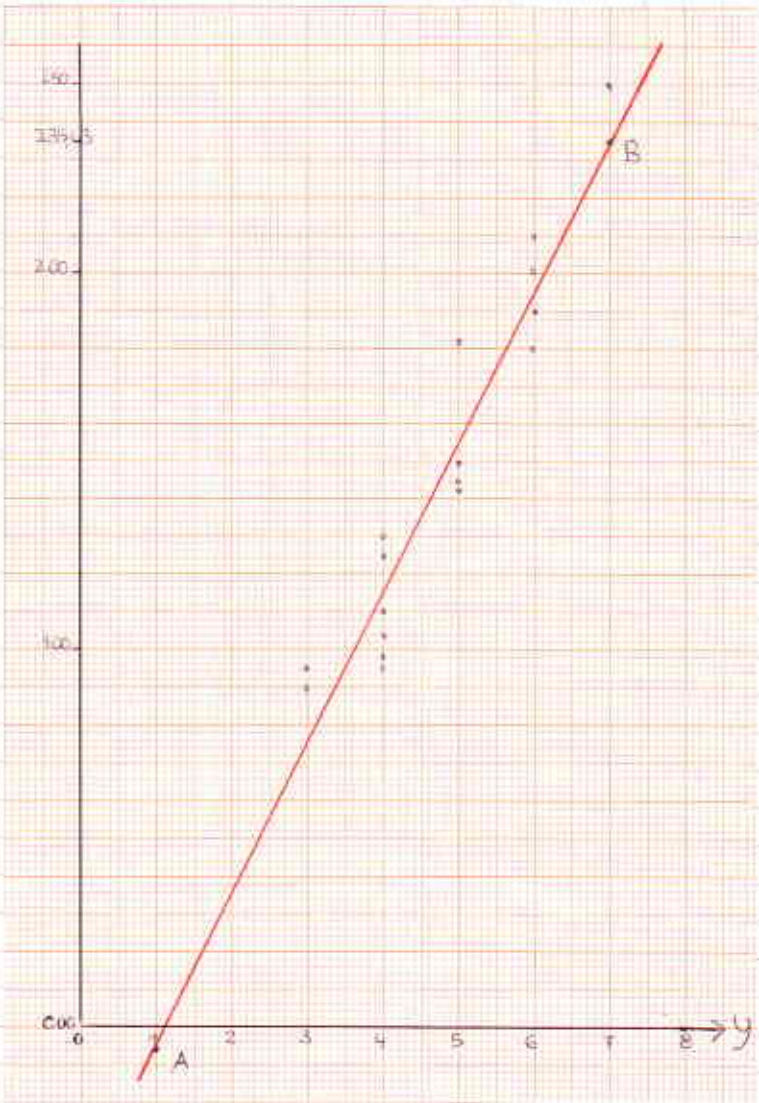
$$R^2_{zy} = \left(\frac{\sigma_{zy}}{\sigma_z \sigma_y}\right)^2 \qquad \sigma^2_{zy} = \left(\frac{44,39}{1,05 \cdot 44,22}\right)^2 = 0,9140$$

E.2 Considerando che il valore 0 identifica un modello pessimo mentre 1 un modello ottimo, nel nostro caso il coefficiente è pari a 0,9140 quindi discreto.
La RETTA DI REGRESSIONE è $z = \beta_0 + \beta_1 y + \varepsilon$
Ricavo due punti della retta:

y	z
1	-4,51
7	235,43

Avremo una retta di regressione lineare (il nostro modello) che passa per i punti di coordinate:
A(1 ; -4,51)
B(7 ; 235,43)

A lato è stato rappresentato il grafico della dispersione con la retta di regressione



E.3 Si determinano ora gli intervalli di confidenza per i parametri di regressione β_0 e β_1. Avremo:

$$z = \hat{\beta}_0 + \hat{\beta}_1 y + \varepsilon$$

dove: $\hat{\beta}_0 = -44,503$
$\hat{\beta}_1 = 39,99$
$n = 20$
$\bar{y} = 4,70$ vani medi
$\sigma_y = 1,11$

Calcolo la varianza dei residui:

$\hat{\sigma}^2_\varepsilon = S^2_\varepsilon - \hat{\beta}^2 \cdot S^2_y$ $\quad \sigma^2_\varepsilon = 1955{,}45 - 39{,}99^2 : 1{,}11 = 180{,}34$

La VARIANZA della stima di β_0 e di β_1 è:

$V(\hat{\beta}_0) = \hat{\sigma}^2_\varepsilon \left[\frac{1}{n} + \frac{\bar{x}^2}{n S^2_y} \right]$ $\quad V(\hat{\beta}_0) = 180{,}34 \left[\frac{1}{20} + \frac{4{,}70^2}{20 \cdot 1{,}11} \right] = 188{,}787$

$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2_\varepsilon}{n \cdot S^2_y}$ $\quad V(\beta_1) = \frac{180{,}34}{20 \cdot 1{,}11} = 8{,}123$

con la tavella dei valori critici della t di Student, posto $\alpha = 0{,}05$, quindi $\alpha/2 = 0{,}025$, e con gradi di libertà pari a 18, si avrà valore critico 2,1009.
l'intervallo di confidenza per β_0:

$\beta_0 \pm t_{\alpha/2} \sqrt{V(\beta_0)}$

$-44{,}503 \pm 2{,}101 \sqrt{188{,}787}$

$(-73{,}354 \; ; \; -15{,}619)$

(84) Verifico l'ipotesi che $\beta_1 = 0$. Calcolo la STATISTICA TEST:

$t = \frac{\hat{\beta}_1}{\sqrt{V(\beta)}}$ $\quad t = \frac{39{,}99}{\sqrt{8{,}123}} = 14{,}031$

Dal momento che $t_{\alpha/2} \to 2{,}1009$ mentre la STATISTICA TEST $t = 14{,}031$, quindi $t > t_{\alpha/2}$, rifiuto l'ipotesi che β_1 sia uguale a 0.

9. TEST DI INDIPENDENZA

Il test di indipendenza verifica l'ipotesi di indipendenza tra due variabili qualitative. È un test non parametrico cioè, che non si basa sulla distribuzione della popolazione. Considereremo le variabili y e z.
Il valore chi-quadrato trovato in precedenza (PG 11) è 21,82. Fisso il livello di significatività a $\alpha = 0{,}01$. Il valore critico per la distribuzione con 8 gradi di libertà è pari a 20,0902. I gradi di libertà calcolati sono $(H-1)(K-1) = (5-1)(3-1) = 8$

REGIONE DI RIFIUTO: $\chi^2 \geq \chi^2_\alpha$

$21{,}82 \geq 20{,}0902$

Quindi Rifiuto l'ipotesi di indipendenza tra i due caratteri. Si riconferma quanto descritto a PG 11.

(85) Ritornando all'analisi della regressione, si può fare una previsione sul modello trovato. Per esempio vogliamo sapere come varia se ipotizziamo un valore di y pari a 8. Quindi se $y = 8$, $z = \beta_0 + \beta_1 y \to z = -44{,}503 + 39{,}99 \cdot 8 = 275{,}43$
L'intervallo di confidenza per la previsione con $\alpha = 0{,}05 \to \alpha/2 = 0{,}025$ e 18 gradi di libertà è pari a:

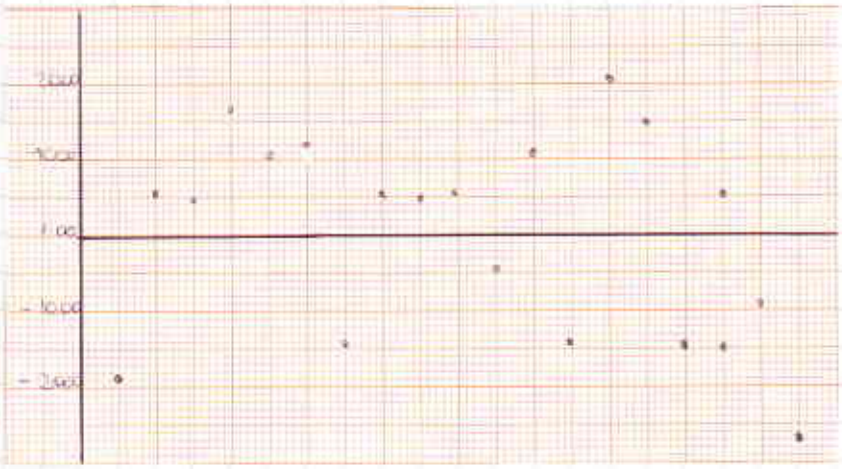
$\hat{z} \pm t_{\alpha/2} \sqrt{\sigma_\varepsilon^2}$

$275{,}43 \pm 2{,}101 \quad 13{,}441$

$(247{,}19 \; ; \; 303{,}67)$

10_ANALISI RESIDUI

La tecnica di "analisi dei residui" consente di controllare che le assunzioni poste con il modello di regressione siano vere. L'approccio grafico è riportato di seguito con la relativa tabella dei dati.



Il dato n° 20 presenta un residuo pari a -27,55 che è notevolmente maggiore rispetto agli altri. Ipoteticamente tale valore lo si può ricondurre ad un errore di rilevazione. Potrebbe essere utile riprendere il calcolo senza tener conto del dato n° 20. Tuttavia si rinuncia in quanto non è molto diverso dal resto.

	n. Vani	Valore (migliaia €)	valore teorico	residui	residui al quadrato	ε
1	3	95	75,47	-19,53	381,30	-1,38
2	4	110	115,46	5,46	29,81	0,39
3	5	150	155,45	5,45	29,66	0,38
4	4	98	115,46	17,46	304,83	1,23
5	4	105	115,46	10,46	109,40	0,74
6	5	143	155,45	12,45	154,90	0,88
7	4	130	115,46	-14,54	211,43	-1,03
8	5	150	155,45	5,45	29,66	0,38
9	4	110	115,46	5,46	29,81	0,39
10	6	190	195,43	5,43	29,51	0,38
11	6	200	195,43	-4,57	20,86	-0,32
12	5	145	155,45	10,45	109,12	0,74
13	3	90	75,47	-14,53	211,03	-1,03
14	4	95	115,46	20,46	418,59	1,44
15	6	180	195,43	15,43	238,16	1,09
16	6	210	195,43	-14,57	212,21	-1,03
17	7	250	235,42	-14,58	212,61	-1,03
18	4	110	115,46	5,46	29,81	0,39
19	4	125	115,46	-9,54	91,02	-0,67
20	5	183	155,45	-27,55	759,23	-1,94
Σ	**94,00**	**2869,00**	**2869,00**	**0,00**	**3612,95**	**0,00**
media	**4,70**	**143,45**	**143,45**	**0,00**		
var					**200,7192192**	