Empiria

Esercizi svolti cap. 6

Esercizio 1: distribuzione di frequenza e variabili

Viene svolta un'indagine su 20 famiglie. Per ciascuna si rileva il numero di componenti. Di seguito i dati:

1 3 2 5 4 2 2 3 3 2 3 4 4 3 2 5 4 3 3 1

a) Costruisci la distribuzione di frequenza, indicando le frequenze assolute, relative e le percentuali.
b) Qual è in questo caso l'*unità di analisi*?
c) Qual è la variabile? Di che tipologia di variabile si tratta?

Soluzione

a)

Numero componenti	n. famiglie (assolute)	f (relative)	%
1	2	0,1	10%
2	5	0,25	25%
3	7	0,35	35%
4	4	0,2	20%
5	2	0,1	10%
Tot	20	1	100%

b) L'unità di analisi è la famiglia
c) La variabile è il numero di componenti; si tratta di una variabile cardinale discreta.

Esercizio 2

In un'ora di tempo, una libreria sita al centro di Venezia emette 20 scontrini per i seguenti importi in Euro

10, 13, 13, 18, 18, 18, 19, 19, 20, 20, 20, 20, 22, 22, 23, 24, 24, 26, 26, 27.

a) Costruisci la distribuzione di frequenza, indicando le frequenze assolute, relative e le percentuali.
b) Qual è la variabile? Di che tipologia di variabile si tratta?
c) Qual è il prezzo medio orario degli scontrini?

Soluzione

a) Di seguito la tabella con le frequenze.

Importo scontrini in Euro	n (assolute)	f (relative)	% abitanti
10	1	0,05	5%
13	2	0,1	10%
18	3	0,15	15%
19	2	0,1	10%
20	4	0,2	20%
22	2	0,1	10%
23	1	0,05	5%
24	2	0,1	10%
26	2	0,1	10%
27	1	0,05	5%
Tot	20	1	100%

b) La variabile è "importo in Euro degli scontrini"; si tratta di una variabile cardinale discreta.
c) Il prezzo medio orario risulta di: 20,10 Euro.

(10x1+13x2+18x3+19x2+20x4+22x2+23x1+24x2+26x2+27x1)/20= 20,10

Esercizio 3

In una PMI lombarda con 70 dipendenti, l'ufficio delle Risorse Umane effettua una rilevazione dati sugli stipendi lordi delle diverse tipologie di personale. I dati vengono raccolti nella seguente tabella.

Inquadramento del personale	Stipendio in migliaia di Euro	Numero addetti
CEO	70	1
Quadri	20	4
Impiegati	16	10
Operai	12	25
Manovali	10	30

a) Che tipo di variabili sono rappresentate nella tabella?
b) Quali indici di posizione (tendenza centrale) è possibile individuare per quanto riguarda la variabile dello stipendio del personale? Effettuare i calcoli e indicare l'indice più adeguato a sintetizzare i dati.

Soluzione

a) La tipologia di personale è una variabile categoriale, nominale. L'importo dello stipendio è una variabile numerica (cardinale) continua a rapporti. Il numero di addetti è la frequenza assoluta rilevata per le due variabili.
b) Considerando che stiamo analizzando una variabile cardinale, possiamo calcolare tutti e tre gli indici di posizione/tendenza centrale, ossia moda, media e mediana.

Moda = 10 migliaia di Euro.

Mediana= è la semi-somma del caso 35 e 36 che in questa rilevazione assumono entrambe valore 12. La mediana degli stipendi è quindi di 12 migliaia di Euro.

Media:

Direttore	1x70	=	70
Capi Ufficio	4x20	=	80
Impiegati	10x16	=	160
Operai	25x12	=	300
Manovali	30x10	=	300

Media = (70+80+160+300+300) / 70 = 910 / 70 = 13

Lo stipendio medio del personale è dunque di 13 migliaia di Euro, poiché influenzato da un valore estremo (*outlier*, rappresentato dallo stipendio del CEO). Per cui, l'indice più opportuno (che non risente di questo effetto) è la mediana.

Esercizio 4: frequenze, indici di tendenza centrale

Agli abitanti di un piccolo comune del novarese viene chiesto di esprimere un giudizio su un nuovo servizio comunale, usando una scala di valutazione da 0 a 4 (0=pessimo, 4= ottimo).

Le risposte ottenute sono riassunte di seguito:

- Pessimo: 251 abitanti
- Insoddisfacente: 260 abitanti
- Sufficiente: 80 abitanti
- Buono: 154 abitanti
- Ottimo: 255 abitanti

a) Che tipologia di variabile stiamo analizzando? Come può essere trattato questo tipo di variabile?
b) Fornire le frequenze relative e percentuali, gli indici di posizione e di variabilità adeguati.

Soluzione

a) La variabile è di tipo ordinale, riferita a un giudizio valutativo. In questo caso quindi la variabile può essere *cardinalizzata*, ossia trasformata in numeri. Queste variabili vengono definite come *quasi-cardinali*.
b)

Giudizio espresso in numeri	n (assolute)	f (relative)	% abitanti	% cumulate
0	251	0,251	25%	25%
1	260	0,26	26%	51%
2	80	0,08	8%	59%
3	154	0,154	15%	74%
4	255	0,255	26%	100%
Tot	1000	1	100%	/

Moda = 1

Mediana = 1

Media = 1.90

Essendo trattata come una variabile numerica (quasi-cardinale), gli indici di variabilità che è possibile calcolare sono la varianza e la deviazione standard (scarto quadratico medio).

varianza: $s^2 = 2.43$

dev. std.: $s = 1.56$.

Esercizio 5: distribuzione di frequenza, indice di tendenza centrale e confronto della variabilità

Da una nota società di ricerca, vengono intervistati 36 milanesi, a cui viene chiesto il numero di vani presenti nella propria abitazione. Le risposte ottenute sono le seguenti:

1, 3, 4, 2, 2, 4, 5, 5, 1, 1, 2, 3, 4, 3, 2, 6, 6, 1, 2, 2, 3, 2, 1, 3, 4, 2, 3, 3, 3, 5, 6, 4, 2, 2, 4, 2.

a) Fornire le frequenze relative, percentuali e cumulate relative delle risposte ottenute.
b) Determinare gli indici di posizione (tendenza centrale) applicabili a questa variabile.
c) La medesima inchiesta è svolta anche a Roma, e le risposte fornite dagli intervistati hanno dato un valor medio uguale a 2,5 ed una varianza uguale a 3,6. Confrontare la variabilità relativa del numero di vani delle abitazioni nei due Comuni e commentare.

Soluzione

a)

Vani	f_i	p_i	P_i
1	5	5/36	5/36
2	11	11/36	16/36
3	8	8/36	24/36
4	6	6/36	30/36
5	3	3/36	33/36
6	3	3/36	36/36
Tot.	36	36/36	

b) media: x = 3

mediana: x^ = 3

moda: x˜ = 2

$s^2 = 2.11$

c) coefficiente di variazione milanesi (c.v. MI) = √2.11/3= 0.48
coefficiente di variazione romani (c.v. Roma) = √3.6/2.5 = 0.76

La variabilità è maggiore a Roma, sia in senso assoluto che relativo.

Esercizio 6: distribuzione di frequenza, indice di tendenza centrale e confronto della variabilità

Un laureando in sociologia ha intervistato nel mese di Febbraio 30 persone a cui ha chiesto "quante volte si sono recate al cinema nell' ultimo mese". Il giovane ha preso nota delle risposte ottenute, che risultano essere le seguenti:

1, 0, 4, 2, 2, 4, 5, 0, 1, 1, 2, 3, 4, 3, 2, 2, 2, 3, 0, 1, 3, 4, 0, 0, 3, 3, 5, 6, 4, 2.

a) Aiutiamolo a sistematizzare i dati, costruendo una tabella di distribuzione delle frequenze (assolute, percentuali e cumulate) relative ai dati rilevati con le sue interviste.

c) Determinare media, moda, mediana e varianza relative alla frequentazione delle sale cinematografiche.

d) Non soddisfatto dei dati raccolti per la propria tesi, decide di condurre altre interviste, nel mese di Giugno. Le risposte fornite da altre 30 persone hanno dato un valor medio uguale a 2,5 ed una varianza uguale a 3. Confrontare con i dati rilevati in inverno e in primavera, commentando le differenze tra la voglia di andare al cinema in periodi dell'anno differenti e la variabilità delle risposte ottenute dagli intervistati.

Soluzione

N. volte in cui è andati al cinema	n (assolute)	f (relative)	%	% cumulate
0	5	0,17	17%	17%
1	4	0,13	13%	30%
2	7	0,23	23%	53%
3	6	0,20	20%	73%
4	5	0,17	17%	90%
5	2	0,07	7%	97%
6	1	0,03	3%	100%
Tot.	30	1,00	100%	/

Moda= 2

Mediana = 2

Media = 2,4

s^2 = 2.64 (varianza)

s = 1.62 (deviazione standard)

CV_i (coeff. Variazione inverno) = 0.67

CV_p (coeff. Variazione primavera) = 0.69

Mediamente la frequenza al cinema è simile in entrambe le stagioni, e anche la variabilità relativa delle risposte ottenute è praticamente identica.

Esercizio 7: scelta dell'indice di posizione

La popolazione delle prime 10 città statunitensi in milioni è la seguente.

New York (New York): 9,21

Los Angeles (California): 4,05

Chicago (Illinois): 2,83

Houston (Texas): 2,01

Phoenix (Arizona): 1,55

Filadelfia: 1,45

Dallas (Texas): 1,31

San Diego (California): 1,30

San Antonio (Texas): 1,24

San Jose (California): 0,94

Calcolare la popolazione media e la popolazione mediana. Quale indice di tendenza centrale è il più adatto?

Soluzione

La popolazione totale è 25.80 milioni.

Quindi la media è 2.58 milioni.

La mediana è la semisomma tra 1.45 e 1.55, cioè 1.5 milioni di abitanti.

È meglio la mediana perché non risente troppo dei valori anomali (valore estremo come quello della città di New York).

Esercizio 8: calcolo dei quartili

Determinare la mediana ed i quartili della seguente distribuzione di frequenze per singoli valori.

VALORI x_i	FFREQUENZE ASSOLUTE n_i
3	25
5	30
8	40
10	52
15	45
18	38
20	27
24	21
	278

Soluzione

Xi	ni	Ni	% cumul
3	25	25	9%
5	30	55	20%
8	40	95	34%
10	52	147	53%
15	45	192	69%
18	38	230	83%
20	27	257	92%
24	21	278	100%
	N= 278		

Essendo 278 la numerosità totale, la mediana è il valore che occupa il posto n/2 -esimo (cioè il 139°) che corrisponde al valore 10 (dove si trova il 50% delle frequenze percentuali cumulate); quindi Me = 10.

È superfluo determinare l'(n/2 +1) -esimo valore (cioè il 140°) e farne la media col precedente, perché è anch'esso uguale a 10.

Per determinare il primo quartile si può:

- dividere 278 per 4 e poiché il quoziente non è intero (69,5) si verifica che il primo quartile si identifica con la media tra il 69° e il 70° valore (che sono però entrambi uguali a 8)
- osservare dove si trova il 25% delle frequenze cumulate. In entrambi i casi si ha Q1 = 8.

Ragionamento simile si effettua per il terzo quartile, Q3, che corrisponde a 18.

Esercizio 9: rapporto di concentrazione di Gini

Nell'a.a. 1988-89, gli iscritti all'Università in Italia per Facoltà risultano:

Facoltà	***Studenti in corso*** *(in migliaia)*
Scientifiche	*146*
Mediche	*100*
Ingegneria	*193*
Economiche-Giuridiche-Sociali	*520*
Letterarie	*239*

a) È possibile svolgere l'analisi della concentrazione degli studenti italiani iscritti all'Università nell'anno accademico 1988-99?
b) In caso affermativo si eseguano gli opportuni calcoli individuando un opportuno indice di concentrazione del fenomeno.

Soluzione

a) L'analisi della concentrazione è possibile quando ci si trova davanti ad un *carattere trasferibile*. Idealmente, potremmo considerare tutti gli studenti universitari e ipotizzare che si trasferiscano da una Facoltà all'altra.
b) Per condurre l'analisi è necessario ordinare le modalità del carattere "numero di studenti in corso" in senso non decrescente. Si procede poi calcolando la cumulata dell'intensità assoluta (ci), la cumulata dell'intensità relativa (qi) e la cumulata di frequenza relativa (pi), come risulta nella seguente tabella:

x_i	Intensità cumulate c_i	Cumulate intensità relative q_i	p_i
100	100	0,083	0,2
146	246	0,21	0,4
193	439	0,36	0,6
239	678	0,56	0,8
520	1198	1	1

L'indice che sintetizza la concentrazione del fenomeno è il "rapporto di concentrazione di Gini", di cui di seguito si riporta la formula come promemoria.

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

Dalla precedente tabella risulta che:

Cumulate intensità relative q_i	p_i	$p_i - q_i$
0,083	0,2	0,117
0,21	0,4	0,19
0,36	0,6	0,24
0,56	0,8	0,24
$\sum_{i=1}^{n-1}$	2	0,787

Il rapporto di concentrazione di Gini (che varia da 0 a 1, assumendo valore 0 in caso di equi-distribuzione del fenomeno e 1 in caso di massima concentrazione) risulta essere il seguente: R= 0,787/2= 0,3935.

Esercizio 10: tabella a doppia entrata (tavola di contingenza)

In un campione di studenti universitari al secondo anno di corso è stato rilevato il voto riportato all'esame di Statistica e quello riportato all'esame di Storia Contemporanea:

Studente	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Voto a Statistica (X)	*28*	*22*	*18*	*18*	*20*	*30*	*20*	*23*	*23*	*27*
Voto a Storia Contemporanea (Y)	*30*	*28*	*27*	*18*	*28*	*28*	*28*	*27*	*27*	*18*

a) Costruire la tabella a doppia entrata (tavola di contingenza) X, Y, calcolando le frequenze marginali di riga e di colonna
b) Calcolare il voto mediano dell'esame di Statistica
c) Costruire la tabella con per frequenze marginali di riga e quella con le frequenze marginali di colonna.

Soluzione:

a) La tavola di contingenza è una tabella a doppia entrata, che registra quante volte (cioè la frequenza assoluta) una coppia di modalità (xi, yj) si presenta contemporaneamente per le unità statistiche. È una distribuzione bivariata perché prende in considerazione due variabili.

Voto a statistica (X)	Voto a storia contemporanea (Y)				
	18	27	28	30	*Tot di riga*
18	1	1	0	0	2
20	0	0	2	0	2
22	0	0	1	0	1
23	0	2	0	0	2
27	1	0	0	0	1
28	0	0	0	1	1
30	0	0	1	0	1
Tot di colonna	2	3	4	1	10

I totali per riga e per colonna sono le frequenze corrispondenti alle variabili X e Y e sono definite frequenze assolute marginali (cioè distribuzioni di frequenza marginali univariate di X e Y).

b) Per prima cosa occorre ordinare i voti riportati
18, 18, 20, 20, 22, 23, 23, 27, 28, 30

Il numero di unità statistiche è di N=10, quindi pari; allora si devono considerare i voti riportati dalle unità statistiche che occupano le posizioni N/2=5° e (N/2)+1=6° e fare la media. La mediana è dunque:

Me = (22+23)/2 = 22,5.

c) Percentuali di riga:

	Voto di storia contemporanea (Y)				
Voto di statistica (X)	18	27	28	30	%
18	50	50	0	0	100
20	0	0	100	0	100
22	0	0	50	0	100
23	0	100	0	0	100
27	100	0	0	0	100
28	0	0	0	100	100
30	0	0	100	0	100

Percentuali di colonna:

	Voto di storia contemporanea (Y)			
Voto di statistica (X)	18	27	28	30
18	50	33	0	0
20	0	0	50	0
22	0	0	25	0
23	0	67	0	0
27	50	0	0	0
28	0	0	0	100
30	0	0	25	0
%	100	100	100	100

Esercizio 11: tavola di contingenza e Chi-quadro

Negli Stati Uniti, sembra che la razza abbia influenza sul fatto che un omicida sia condannato a morte o meno. La Tabella seguente riporta 326 casi in cui l'imputato è stato accusato di omicidio.

a) Costruire la tavola delle frequenze attese (*in caso di indipendenza*) e valutare se le due variabili sono tra loro indipendenti.
b) Valutare l'eventuale associazione con un indice appropriato.
c)

	Condanna a morte		
Razza imputato	sì	no	Tot.
Bianca	19	141	160
Nera	17	149	166
Tot.	36	290	326

Soluzione

a) Tabella delle frequenze attese (frequenze teoriche)

	Condanna a morte		
Razza imputato	sì	no	Tot.
Bianca	17,7	142,3	160
Nera	18,3	147,7	166
Tot.	36	290	326

Siccome le due tabelle divergono (anche se di poco), le due variabili NON sono indipendenti.

b) Un indice appropriato è l'indice chi-quadro che risulta infatti molto basso (chi^2= 0,22), per cui tra le due variabili esiste un'associazione, ma di lievissima intensità.

Esercizio 12: tabella a doppia entrata, indipendenza tra variabili e Chi quadro

La tabella seguente riporta le abitudini nei confronti del fumo di un gruppo di studenti e dei loro genitori.

Lo studente fuma?	*I genitori fumano?*		
	nessuno	uno solo	Entrambi
No	1168	1823	1380
Sì	188	416	400

a) Quanti studenti vengono descritti in questa tabella? Quale percentuale di studenti fuma?
b) Se consideriamo la variabile "fumo genitori" come indipendente, quale tra le percentuali di riga o quelle di colonna dovrei calcolare? Effettuare il calcolo commentando i risultati.
c) Costruire la tabella delle frequenze attese (teoriche), spiegando che cosa significano. C'è una relazione tra le abitudini nei confronti del fumo dei genitori e quelle dei figli oppure sono indipendenti?
d) A quali indici posso ricorrere per valutare la dipendenza/indipendenza tra le due variabili?

Soluzione

a) Nella tabella vengono descritti 5375 studenti. La percentuale di studenti che fuma è:

100*(188+416+400)/5375 = 18,68%

b) Per calcolare l'influenza della variabile posta in riga sulla variabile posta in colonna è opportuno calcolare le percentuali di colonna.
TABELLA
c) Per frequenze attese o frequenze teoriche si intendono quelle che otterrei nel caso in cui ci fosse indipendenza tra le due variabili. Le due variabili non sono quindi indipendenti, in quanto la tabella delle frequenze attese è diversa da quella delle frequenze osservate.
TABELLA
d) Per valutare il grado di associazione si calcola un indice di associazione, per esempio il chi-quadrato o l'indice V di Cramer.

Esercizio 13: calcolo dell'indice Chi quadro e V di Cramer

La tabella seguente riporta la produzione di energia elettrica suddivisa per fonti, nelle 3 grandi ripartizioni italiane:

RIPARTIZIONI GEOGRAFICHE	FONTI DI ENERGIA			TOTALI
	IDROELETTRICA	TERMOELETTRICA	GEOTERMOELETTRICA	
NORD	38.005	78.507	-	116.512
CENTRO	3.470	33.265	3.254	39.989
SUD	3.748	54.184	-	57.932
TOTALE	45.223	165.956	3.254	214.433

Calcolare l'indice del Chi quadro e l'indice medio di contingenza V di Cramer.

Soluzione

Si costruisce la tabella delle frequenze attese (teoriche) in caso di indipendenza.

RIPARTIZIONI GEOGRAFICHE	FONTI DI ENERGIA			TOTALI
	IDROELETTRICA	TERMOELETTRICA	GEOTERMOELETTRICA	
NORD	24.571,88	90.172,06	1.768,06	116.512
CENTRO	8.433,51	30.948,66	606,83	39.989
SUD	12.217,61	44.835,28	879,11	57.932
TOTALI	45.223	165.956	3.254	214.433

Il Chi quadro risulta di: 33.963

V di Cramer = 33.963/ 214.433x2 = 0,08

Ricordiamo che la V di Cramer è utilizzata per misurare il grado di associazione tra due variabili nominali, il risultato è un valore reale nell'intervallo tra 0 e 1. Orientativamente se il valore ottenuto è compreso tra 0 e 0,3 si ha una bassa connessione, da 0,3 a 0,6 si ha una buona connessione, da 0,6 a 1 si ha un'ottima connessione.

Esercizio 14: costruzione della retta di regressione lineare

La seguente tabella a doppia entrata raccoglie i valori relativi alla variabile X (numero di macchinari su cui si lavora) e Y (ore di assenza nell'ultimo mese) rilevati su un campione di 106 dipendenti di un'azienda di termo-induzione dell'acciaio.

x_i	y_h 3	7	10	18	Frequenze Marginali per riga
3	15	11	6	4	36
6	8	10	13	3	34
11	6	7	10	13	36
Frequenze marginali Per colonna	29	28	29	20	106

Supponendo che sussista un rapporto di dipendenza lineare della Y dalla X, siano dati i seguenti valori:

Covarianza (X, Y) = 609,89

Dev Standard (X) = 1174,03

Dev Standard (Y) = 2804,36

Si determinino:

- la retta di regressione di Y su X
- il coefficiente di correlazione lineare.

Soluzione

Per il calcolo della retta di regressione, si calcolano le medie (poiché le deviazioni standard e la covarianza sono date).

Media x = 708/106= 6,68

Media y = 933/106 = 8,80

b = 609,89/1174,03 = 0,52

a = 8,80 - 0,52 x 6,68= 5,33

Per cui la retta di regressione risulta: y = 5,33 + 0,52 x

Ricordando la formula del coefficiente di regressione:

$$\rho XY = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

p = 609,892/1174,03 x 2804,36 = 0,113

Esercizio 15: costruzione di uno scatter-plot (diagramma di dispersione)

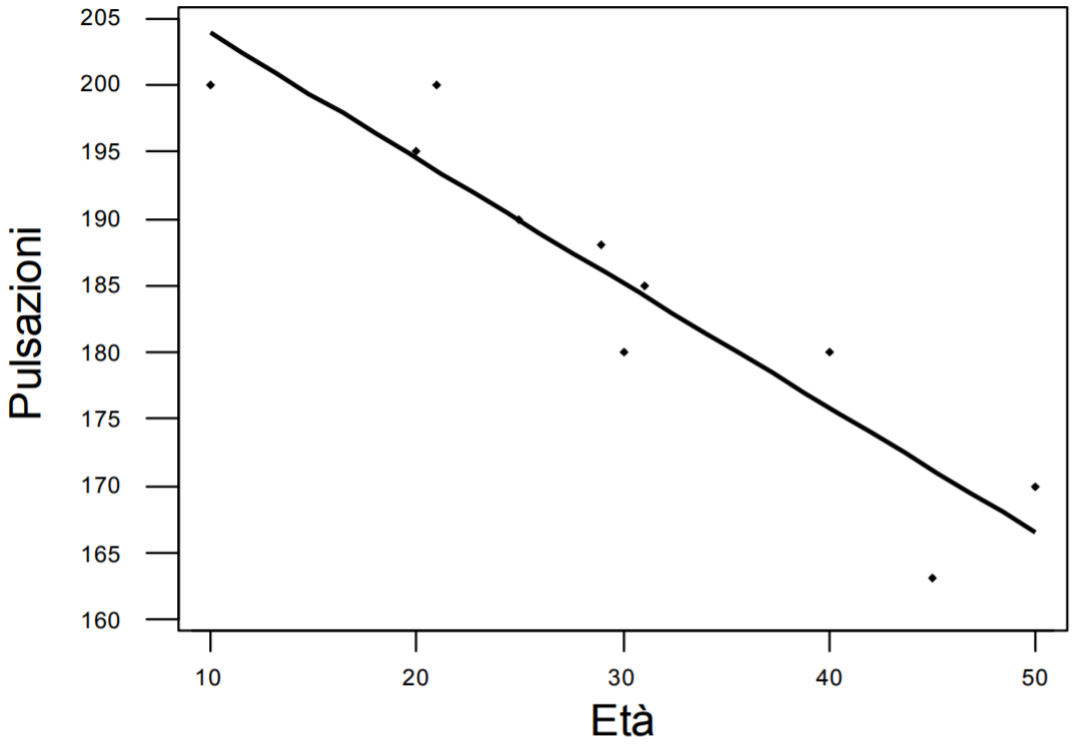
In un esperimento diretto allo studio della relazione tra il numero di pulsazioni sotto sforzo (per minuto) e l'età (in anni) sono stati rilevati i seguenti dati su 10 soggetti di sesso maschile:

Pulsazioni	200	195	200	190	188	180	185	180	163	170
Età	10	20	21	25	29	30	31	40	45	50

1. Costruire lo scatter-plot (diagramma di dispersione)
2. Commentare il grafico individuando il tipo e la forma della relazione (esplicitando se diretta o inversa) tra le due variabili.

Soluzione

1.



2. Graficamente è possibile interpretare una relazione lineare (e quindi effettuare i calcoli per la regressione lineare semplice) di tipo inverso: al crescere dell'età diminuiscono le pulsazioni al minuto (correlazione negativa).