STATISTICA

CdL in XXX - Prova del xx/xx/xxxx

Cognome _____ Nome _____ Matr _____ Firma _____

ESERCIZIO 1

Si vuole verificare se esiste un legame tra l'età dei volontari e il numero di ore prestate presso gli enti di assistenza. A tal proposito si rilevano i dati relativi ad un gruppo di 8 volontari che hanno prestato il proprio servizio il mese scorso:

	1	2	3	4	5	6	7	8
Numero di ore	7	15	33	20	18	8	12	11
Età	23	25	51	34	29	19	37	22

1) Rappresentare graficamente i due caratteri
2) Confrontare la variabilità del numero di ore di servizio prestato con la variabilità dell'età dei volontari
3) Stabilire se i caratteri sono linearmente legati, e in caso affermativo misurare l'intensità della relazione

ESERCIZIO 2

Nella seguente tabella sono stati riportati i dati relativi alla distribuzione per età e n° di giorni di assenza per malattia dei dipendenti di una azienda:

		Età		
		18 |- 30	**30 |- 45**	**45 |- 60**
Giorni Assenza	**5 -| 10**	5	10	12
	10 -| 20	15	10	6
	20 -| 30	8	2	4

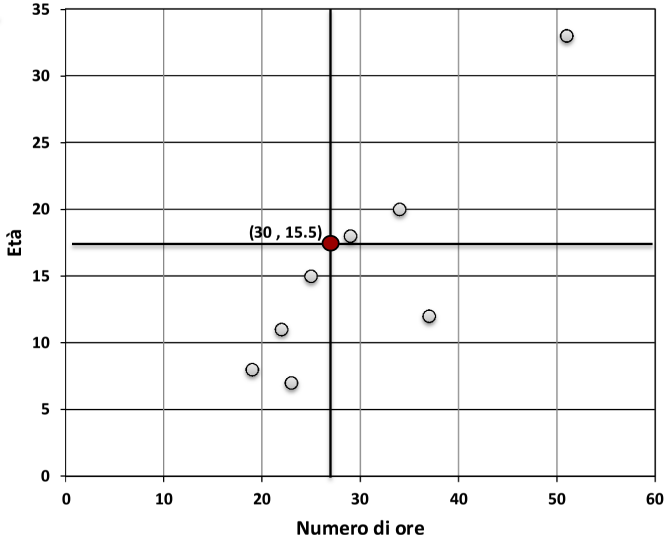
1) Determinare la distribuzione condizionata del n° di assenze per i dipendenti compresi nelle fasce di età 18|-30 e 45|-60 e confrontare il n° medio di giorni di assenza, commentando il risultato
2) Tra tutti coloro che hanno più di 30 anni qual è la percentuale di coloro che si sono assentati dal lavoro meno di 20 giorni?
3) Studiare la connessione tra le due variabili e valutarne in caso affermativo l'intensità

ESERCIZIO 1

	1	2	3	4	5	6	7	8
Numero di ore	7	15	33	20	18	8	12	11
Età	23	25	51	34	29	19	37	22

dall'analisi dello scatterplot possiamo osservare come la nube dei punti sia "stretta": ciò è indice di una relazione lineare tra le varibili, rispetto al collettivo oggetto di studio, con un'alta intensità; il successivo calcolo del baricentro della nube, con coordinate pari alle medie di X e Y (in rosso nel grafico), consente di considerare ammissibile una concordanza tra i caratteri, da verificare in via analitica con il calcolo della covarianza (vedi punto 3)

1



2

$\bar{x} = 15.5$ *(ore)* $\quad \sigma_x^2 = 61.75 \Rightarrow \sigma_x = 7.86$ *(ore)*

$\bar{y} = 30$ *(anni)* $\quad \sigma_y^2 = 95.75 \Rightarrow \sigma_y = 9.79$ *(anni)*

mediamente i volontari hanno prestato servizio per 15.5 ore, con una deviazione standard di ±7.86 ore, e hanno una età di 30 anni, con una deviazione standard di ±9.79 anni; la presenza del volontario n° 3 ha sicuramente una influenza sui valori medi, poiché le modalità ossevate sul numero di ore di servizio e sull'età rappresentano dei valori anomali rispetto alle distribuzioni dei due carateri studiati

$\Rightarrow$ $CV_x = 0.51$ $\quad CV_y = 0.33$

è possibile confrontare i due caratteri utilizzando i coefficienti di variazione, ottenuti dal rapporto tra le deviazioni standard e le medie in valore assoluto; con un coefficiente del 51% e del 33% per il numero di ore e per l'età, rispettivamente, possiamo affermare che l'età dei volontari ha una più bassa variabilità rispetto al numero di ore di servizio prestate: mediamente i valori osservati si discostano maggiormente dal valore medio rispetto all'intensità di quest'ultimo nel primo caso rispetto al secondo

3

$COV(X,Y) = 67.375$

$r(X,Y) = 0.876$

il calcolo della covarianza delle due variabili è diverso da zero, quindi come ipotizzato dall'analisi della rappresentazione grafica è ammissibile una relazione lineare tra nulero di ore di servizio ed età dei volontari; la positività del segno di COV(X,Y) consente poi di affermare che le due variabili sono concordi (a valori del numero di ore superiori alla media si associano valori dell'età superiori alla media, e viceversa): per poter valutare l'intensità di tale legame si ricorre al coefficiente di correlazione, pari in questo caso allo 87.6% della massima relazione lin. positiva e quindi ad un livello alto

ESERCIZIO 2

		Età			
		18 |- 30	30 |- 45	45 |- 60	
Giorni	5 -| 10	5	10	12	27
Assenza	10 -| 20	15	10	6	31
	20 -| 30	8	2	4	14
		28	22	22	72

⟹

		Età			
		18 |- 30	30 |- 45	45 |- 60	
Giorni	5 -| 10	**0.18**	0.45	**0.55**	0.38
Assenza	10 -| 20	**0.54**	0.45	**0.27**	0.43
	20 -| 30	**0.29**	0.09	**0.18**	0.19
		1.00	1.00	**1.00**	1.00

per ottenere le distribuzioni condizionate del n° di assenze (X) rispetto alla classe di età (Y) è necessario dividere le frequenze congiunte di ciascuna colonna per la corrispondente freq. marginale di colonna, cioè calcolare i cosiddetti profili colonna: dall'analisi dei dati così trasformati possiamo dedurre come nella fascia 18 |- 30 (i dipendenti più giovani) più della metà abbia effettuato da 10 a 20 gg di assenze per malattia, mentre nella fascia 45 -| 60 (i dipendenti più anziani) più della metà abbia effettuato meno di 10 gg di assenza; confrontando i gruppi anche rispetto al n° medio di assenze vediamo come sia più basso per questi ultimi di quasi 4 gg rispetto ai primi

1

		Età	
		18 |- 30	45 |- 60
Giorni	5 -| 10	**17.86**	**54.55**
Assenza	10 -| 20	**53.57**	**27.27**
	20 -| 30	**28.57**	**18.18**
		100.00	**100.00**

M(assenze | 18 |- 30) = 16.52 *(giorni)*

M(assenze | 45 |- 60) = 12.73 *(giorni)*

2

Freq (Y>30) = 22+22 = 44
Freq (X<20,Y>30) = 10+10+12+6 = 38

Freq (X<20|Y>30) = 38/44 = 0,86

→ **il numero di dipendenti che ha una età superiore ai 30 anni è dato dalla frequenza marginale corrispondente alla seconda e alla terza colonna della matrice (rispettivamente 22 e 22 unità statistiche, per un totale di 44 dipendenti), mentre il numero di dipendenti che in questo sotto-collettivo hanno effettuato meno di 20 giorni di assenza è leggibile nelle freq. congiunte n_{12} e n_{22} (per i dipendenti tra i 30 e i 45 anni) e n_{13} e n_{23} (per i dipendenti tra i 45 e i 60 anni), per una percentuale dello 86%**

3

frequenze osservate

		Età			
		18 |- 30	30 |- 45	45 |- 60	
Giorni	5 -| 10	5	10	12	27
Assenza	10 -| 20	15	10	6	31
	20 -| 30	8	2	4	14
		28	22	22	72

frequenze teoriche

		Età			
		18 |- 30	30 |- 45	45 |- 60	
Giorni	5 -| 10	10.50	8.25	8.25	27
Assenza	10 -| 20	12.06	9.47	9.47	31
	20 -| 30	5.44	4.28	4.28	14
		28	22	22	72

$$n_{11}^* = \frac{28 \times 27}{72} = 10.50 \neq n_{11} = 5$$

per verificare l'ipotesi di indipendenza in distribuzione, necessaria nello studio sulla connessione tra due variabili, calcoliamo la frequenza teorica relativa alla prima modalità di X e alla prima di Y: poiché la freq. teorica è diversa da quella osservata cade l'ipotesi di indipendenza e si prosegue l'analisi per verificare la connessione tra giorni di assenza ed età dei dipendenti

$\chi^2 = 9.41$

max $\chi^2 = 144$

V = 0.26

→ **una volta calcolate le frequenze teoriche (sotto l'ipotesi di indipendenza tra le variabili) si procede al calcolo del χ^2, in questo caso pari a 9.41: il valore ottenuto è diverso da 0 e quindi c'è connessione tra le variabili; possiamo ora dire qual è l'intensità del legame utilizzando la V di Cramer, che ha al denominatore il massimo valore di connessione che è possibile osservare in questo collettivo: con lo 0.26 (26% della max connessione osservabile), possiamo concludere che c'è un legame debole tra le variabili in esame**