STATISTICA

CdL in XXX - Prova del xx/xx/xxxx

Cognome _____ Nome _____ Matr _____ Firma _____

ESERCIZIO 1

Nella tabella sono riportati per 250 studenti laureati in Economia Aziendale nella sessione di Luglio il voto di laurea e il tempo trascorso dalla laurea al primo impiego (in mesi):

		Voto di Laurea		
		80 -|90	**90 -|100**	**100 -|110**
Tempo al	**0 -| 6**	20	60	20
I° impiego	**6 -| 12**	30	20	40
(in mesi)	**12 -| 24**	10	30	20

1) Calcolare e commentare le distribuzioni condizionate del tempo al I° impiego e quelle del voto di laurea
2) Studiare la relazione lineare tra i due caratteri valutando l'intensità del loro legame e commentando il risultato

ESERCIZIO 2

Nella tabella è riportata la distribuzione del reddito familiare annuo (in migliaia di €) in Italia, relativo all'anno 2000:

Classi di Reddito (migliaia di €)	Famiglie
0 -| 10	983
10 -| 20	2478
20 -| 30	1878
30 -| 40	1265
oltre 40	1397

(FONTE: Banca d'Italia, Indagine sui bilanci delle famiglie italiane, 2000)

1) Definire il collettivo statistico, il carattere e la natura del carattere
2) Calcolare il reddito familiare mediano, e studiare la variabilità del fenomeno con un indice opportuno, commentando i risultati
3) Studiare la concentrazione del reddito familiare annuo analiticamente

ESERCIZIO 1

		Voto di Laurea			
		80 -|90	90 -|100	100 -|110	Totale
Tempo al I° impiego	0 -| 6	20	60	20	100
	6 -| 12	30	20	40	90
	12 -| 24	10	30	20	60
	Totale	60	110	80	250

1) Per calcolare la distribuzione condizionata del tempo al I° impiego rispetto al voto di laurea (X|Y) è necessario dividere ciascuna frequenza congiunta per il corrispondente totale di colonna (*frequenza marginale di colonna*), ottenendo così i cosiddetti profili colonna:

		Voto di Laurea			
		80 -|90	90 -|100	100 -|110	Totale
Tempo al I° impiego	0 -| 6	33%	55%	25%	40%
	6 -| 12	50%	18%	50%	36%
	12 -| 24	17%	27%	25%	24%
		100%	100%	100%	100%

Dall'analisi della tabella si evince come ad esempio più della metà dei laureati con un voto compreso tra 90 e 100 ha atteso fino a 6 mesi prima di trovare lavoro (55%). Allo stesso modo possiamo vedere come solo un laureato su quattro tra quelli con un voto superiore a 100 ha atteso fino a 6 mesi prima di trovare lavoro (25%), mentre un laureato su tre tra quelli con voto inferiore a 90 ha atteso lo stesso tempo (33%).

Per calcolare la distribuzione condizionata del voto di laurea rispetto al tempo al I° impiego (Y|X) è necessario dividere ciascuna frequenza congiunta per il corrispondente totale di riga (*frequenza marginale di riga*), ottenendo così i cosiddetti profili riga:

		Voto di Laurea			
		80 -|90	90 -|100	100 -|110	
Tempo al I° impiego	0 -| 6	20%	60%	20%	100%
	6 -| 12	33%	22%	44%	100%
	12 -| 24	17%	50%	33%	100%
	Totale	24%	44%	32%	100%

Dall'analisi della seconda tabella si evince come tra i laureati che hanno atteso fino a 6 mesi per trovare lavoro più della metà ha conseguito la laurea con un voto compreso tra 90 e 100 (60%). Analogamente possiamo vedere come tra i laureati che hanno atteso da uno a due anni prima di trovare lavoro, uno su tre ha conseguito la laurea con un voto superiore a 100 (33%).

2) Per studiare la relazione lineare è necessario innanzi tutto verificare se i due caratteri X e Y sono concordi, discordi o incorrelati. A tale scopo è necessario calcolare la covarianza:

$$COV(X,Y)=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(x_i y_j n_{ij}\right)}{n} - M(X)\cdot M(Y)$$

(*per semplificare i calcoli possiamo calcolare le medie con le frequenze marginali relative*)

$$M(X) = (3x0{,}40) + (9x0{,}36) + (18x0{,}24) = 8{,}76 \text{ mesi}$$

$$M(Y) = (85x0,24) + (95x0,44) + (105x0,32) = 95,8 \text{ trentesimi}$$

Mediamente i laureati nella sessione di Luglio hanno conseguito la laurea con un voto medio pari a 95,8 e atteso mediamente 8,76 mesi prima di trovare lavoro.

$$COV(X,Y)=\frac{\sum_{i=1}^{r}\sum_{j=1}^{c}\left(x_i y_j n_{ij}\right)}{n} - M(X)\cdot M(Y) = \frac{3\cdot 85\cdot 20 + ... + 18\cdot 105\cdot 20}{250} - 8,76\cdot 95,8 = 843 - 839,208 = 3,792$$

Avendo ottenuto una COV(X,Y) > 0 possiamo affermare che i due caratteri, nel collettivo esaminato, sono concordi (esiste una *relazione lineare positiva*). Per misurare l'intensità di tale relazione è necessario calcolare il coefficiente di correlazione lineare:

$$r(X,Y) = \frac{COV(X,Y)}{\sqrt{VAR(X)\cdot VAR(Y)}} = \frac{\sum_{i=1}^{r}\sum_{j=1}^{c}[x_i - M(X)][y_j - M(Y)]\cdot n_{ij}}{\sqrt{\sum_{i=1}^{r}[x_i - M(X)]^2 \cdot n_{i.}} \cdot \sqrt{\sum_{j=1}^{c}[y_j - M(Y)]^2 \cdot n_{.j}}}$$

(*per semplificare i calcoli possiamo calcolare le varianze con le frequenze marginali relative*)

$$VAR(X) = (3-8,76)^2x0,40 + (9-8,76)^2x0,36 + (18-8,76)^2x0,24 = 33,7824 \Rightarrow \sigma_x = 5,81 \text{ mesi}$$

$$VAR(Y) = (85-95,8)^2x0,24 + (95-95,8)^2x0,44 + (105-95,8)^2x0,32 = 55,36 \Rightarrow \sigma_y = 7,44 \text{ trentesimi}$$

Mediamente il tempo di attesa per il I° impiego si discosta dal tempo medio di ± 5,81 mesi, mentre il voto di laurea si discosta mediamente dal voto medio di ± 7,44 trentesimi.

$$r(X,Y) = \frac{COV(X,Y)}{\sqrt{VAR(X)\cdot VAR(Y)}} = \frac{3,792}{5,81\cdot 7,44} = 0,0877$$

Abbiamo osservato l'8,77% della massima dipendenza lineare positiva, quindi possiamo affermare che il legame è molto debole (oppure, in alternativa, che il livello di correlazione positiva è molto basso).

ESERCIZIO 2

Classi di Reddito (migliaia di €)	Famiglie
0 -| 10	983
10 -| 20	2478
20 -| 30	1878
30 -| 40	1265
oltre 40	1397
Totale	**8001**

1) Il collettivo statistico è costituito dalle 8001 famiglie, il carattere studiato è il reddito familiare annuo nel 2000. Il carattere ha natura quantitativa continua (**NOTA: *è rappresentato in classi per comodità di lettura, ma di fatto è un carattere continuo perché ogni famiglia può percepire qualsiasi quantità come reddito***).

2) Per calcolare il reddito familiare mediano nel 2000 è necessari innanzi tutto calcolare le frequenze relative cumulate:

Classi di Reddito (migliaia di €)	Famiglie (n_i)	f_i	F_i
0 -| 10	983	0,12	0,12
10 -| 20	2478	0,31	0,43
20 -| 30	1878	0,23	0,67
30 -| 40	1265	0,16	0,83
oltre 40	1397	0,17	1,00
Totale	**8001**	**1,00**	**-**

Dalla lettura delle frequenze relative cumulate individuiamo come classe mediana 20 -| 30 (=> 67%): è la classe che contiene la famiglia con un valore del reddito tale da lasciare a destra e a sinistra il 50% dei valori della distribuzione. A questo punto è necessari individuare il reddito familiare mediano per approssimazione lineare:

$$Me \cong 20 + \frac{0,5 - 0,43}{0,67 - 0,43} \cdot 10 = 20 + 0,287 \cdot 10 = 22,87 \text{ (migliaia di €)}$$

Il reddito familiare mediano nel 2000 ammonta a circa 22870 € annui.

Per quanto riguarda la variabilità del fenomeno, avendo già calcolato la mediana è possibile utilizzare lo scostamento semplice mediano (**NOTA:** ***non è sbagliato usare la varianza, ma dovremmo calcolare prima la media, e ciò comporta ovviamente calcoli aggiuntivi non necessari…***):

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - Me\right| \cdot n_i =$$

$$= \frac{\left|5\text{-}22,87\right| \cdot 983 + \left|15\text{-}22,87\right| \cdot 2478 + \left|25\text{-}22,87\right| \cdot 1878 + \left|35\text{-}22,87\right| \cdot 1265 + \left|45\text{-}22,87\right| \cdot 1387}{8001} =$$

$$= 10,91 \text{ (migliaia di €)}$$

Mediamente si ha uno scostamento dalla mediana di ± 10910 € (**NOTA:** ***poiché abbiamo una distribuzione in classi è necessario utilizzare nella formula i valori centrali al posto delle xi. Dal momento che l'ultima classe è aperta, assumiamo che abbia la stessa ampiezza delle altre e quindi che il valore centrale sia pari a 45***).

3) Per studiare la concentrazione assumiamo all'interno di ciascuna classe di reddito equidistribuzione. in tal modo implicitamente assumiamo che ogni famiglia appartenente ad una data classe abbia come reddito "medio" la quantità espressa dal valore centrale. In tal modo possiamo calcolare l'ammontare complessivo di reddito posseduto dalle famiglie di una classe:

Classi di Reddito (migliaia di €)	Famiglie	v.c.	Reddito Complessivo (migliaia di €)
0 -| 10	983	5	4915
10 -| 20	2478	15	37170
20 -| 30	1878	25	46950
30 -| 40	1265	35	44275
oltre 40	1397	45	62865
Totale	**8001**	**-**	**196175**

Adesso è possibile procedere al calcolo delle frazioni cumulate delle prime i unità statistiche (p_i) e delle frazioni di carattere posseduto dalle prime i unità statistiche (q_i).

Poiché abbiamo una distribuzione in classi possiamo assumere le frequenze relative cumulate F_i come p_i , mentre per calcolare le qi abbiamo bisogno di calcolare prima l'ammontare cumulato di carattere posseduto (A_i) e quindi trasformarlo in frazioni cumulate relative:

Classi di Reddito (migliaia di €)	Famiglie	$F_i = p_i$	A_i	q_i
0 -| 10	983	0,12	4915	0,03
10 -| 20	2478	0,43	42085	0,21
20 -| 30	1878	0,67	89035	0,45
30 -| 40	1265	0,83	133310	0,68
oltre 40	1397	1,00	196175	1,00
Totale	**8001**	-	-	-

Possiamo adesso calcolare il rapporto di concentrazione R per misurare il livello di concentrazione del reddito tra le 8001 famiglie che compongono il nostro collettivo:

$$R = 1 - \frac{\sum_{i=1}^{n-1} q_i}{\sum_{i=1}^{n-1} p_i} = 1 - \frac{1,37}{2,05} = 0,3296 \approx 33\%$$

Abbiamo osservato il 33% della massima concentrazione possibile, quindi concludiamo che il livello di concentrazione del reddito in questo collettivo è medio-basso.