Esercitazioni di Statistica Matematica

Allievi Meccanici II Anno

1 semestre 2001 - 2002

Maurizio Dini

Argomenti odierni:

STATISTICA DESCRITTIVA

Analisi univariata

1) distribuzioni, frequenze assolute, relative e specifiche, frequenze cumulate e retrocumulate

2) indici di posizione (media, mediana, moda)

3) indici di dispersione (varianza, dev. standard, quantile, quartile, IQR)

4) indici di forma (skewness)

Analisi multivariata

5) frequenze marginali e congiunte

6) coefficienti di correlazione

ESERCIZIO 1

Supponiamo di dover dimensionare l'organico di una biglietteria in modo da ridurre al minimo le code allo sportello e le conseguenti attese dei clienti in coda. Per semplicità trascuriamo soluzioni come flessibilità, turni, rotazioni, part-time, ecc.

Si tratterà quindi di rilevare una serie di dati da utilizzare come campione per le nostre analisi.

L'obiettivo verrà raggiunto individuando il massimo numero di clienti in coda nel periodo di campionamento.

La seguente tavola riporta le rilevazioni, effettuate per 50 giorni feriali in 5 orari fissati, della lunghezza della coda davanti allo sportello.

Rappresentiamo con tabella e grafici:

1) le frequenze cumulate e retrocumulate

2) l'ampiezza delle classi

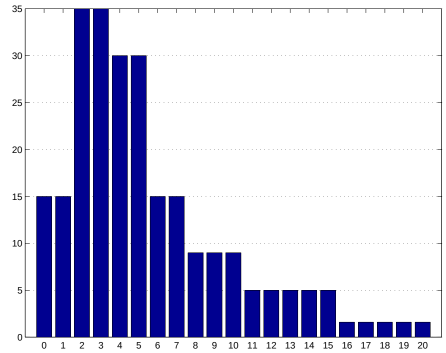
3) le frequenze specifiche

Classi x	Osservazioni n_x	FC $N(x)$	FRC $\overline{N}(x)$	Amp. classe	FS f_s
0 - 1	30				
2 - 3	70				
4 - 5	60				
6 - 7	30				
8 - 10	27				
11 - 15	25				
16 - 20	8				
Totale	250	/	/	/	/

SOLUZIONE

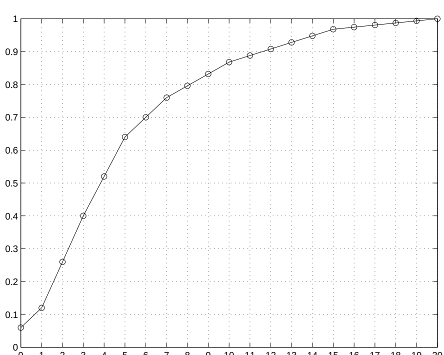
Classi x	Osservazioni n_x	FC $N(x)$	FRC $\overline{N}(x)$	Amp. classe	FS f_s
0 - 1	30	30	250	2	15
2 - 3	70	100	220	2	35
4 - 5	60	160	150	2	30
6 - 7	30	190	90	2	15
8 - 10	27	217	60	3	9
11 - 15	25	242	33	5	5
16 - 20	8	250	8	5	1.6
Totale	250	/	/	/	/

Grafico a barre della distribuzione della lunghezza delle code



Naturalmente, per garantire assenza di code dovranno essere sempre presenti 20 addetti (pari al massimo numero di clienti in coda rilevato durante il campionamento)

Plot della frequenza specifica relativa cumulata



Questo grafico ci consente, ad esempio, di affermare che con 12 addetti il 90% dei clienti non farà coda allo sportello.

Da un campione di 40 studenti Fisica di sesso maschile sono stati rilevati i segenti dati: l'altezza, il peso, il voto ottenuto all'esame di Informatica. I dati vengono raggruppati in classi, e si ottengono le seguenti tabelle per le frequenze assolute congiunte:

peso (Kg) altezza (m)	50-60	60-70	70-80	80-90
1.50-1.60	3	2	1	0
1.60-1.70	2	4	3	1
1.70-1.80	1	3	6	4
1.80-1.90	0	2	4	4

peso voto	50-60	60-70	70-80	80-90
19-21	1	1	2	1
22-24	1	2	3	2
25-27	3	4	5	3
28-30	1	4	4	3

1. Ricavare le frequenze marginali assolute.
2. Dalle informazioni fornite è possibile ricavare la tabella delle frequenze congiunte delle variabili *altezza-voto*?
3. Calcolare i coefficienti di correlazione ρ_{ap} e ρ_{vp}. Commentare i risultati ottenuti. È possibile calcolare ρ_{av}?
4. Calcolare la frequenza congiunta definita da $f(i,j) = f_v(i) f_p(j)/n$ dove f_p è la frequenza marginale dei pesi, f_v è la frequenza marginale dei voti e n è il numero di studenti ($n = 40$). Confrontare $f(i,j)$ con $f_{vp}(i,j)$.

Soluzione. Chiamiamo f_{ap} la frequenza congiunta *altezza-peso* corrispondente alla prima tabella, e f_{vp} la frequenza congiunta *voto-peso* corrispondente alla seconda tabella.

1.

$$f_a(i) = \sum_{j=1}^{4} f_{ap}(i,j) = [6\ 10\ 14\ 10]$$

$$f_p(j) = \sum_{i=1}^{4} f_{ap}(i,j) = [6\ 11\ 14\ 9]$$

$$f_v(i) = \sum_{j=1}^{4} f_{vp}(i,j) = [5\ 8\ 15\ 12]$$

2. Le informazioni non sono sufficienti per ricavare le frequenze congiunte delle variabili *altezza-voto*.

3. Definiamo i vettori

$$a = [1.55\ 1.65\ 1.75\ 1.85]$$

$$p = [55\ 65\ 75\ 85] \quad v = [20\ 23\ 26\ 29]$$

Che corrispondono ai punti medi degli intervalli delle classi rispettivamente delle variabili altezza, peso e voto.

$$\overline{a} = \frac{1}{4}\sum_{i=1}^{4} f_a(i)a(i) = 1.72m \quad \overline{p} = \frac{1}{4}\sum_{i=1}^{4} f_p(i)p(i) = 71.5Kg$$

$$\overline{v} = \frac{1}{4}\sum_{i=1}^{4} f_v(i)v(i) = 25.5$$

$$s_a = \sqrt{\frac{1}{3}\sum_{i=1}^{4} f_a(i)(a(i)-\overline{a})^2} = 10.18cm$$

$$s_p = \sqrt{\frac{1}{3}\sum_{i=1}^{4} f_p(i)(p(i)-\overline{p})^2} = 10.0128Kg$$

$$s_v = \sqrt{\frac{1}{3}\sum_{i=1}^{4} f_v(i)(v(i)-\overline{v})^2} = 3.0038$$

$$s_{ap} = \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{3} f_{ap}(i,j)(a(i)-\overline{a})(p(j)-\overline{p}) = 0.533$$

$$s_{vp} = \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{3} f_{vp}(i,j)(v(i)-\overline{v})(p(j)-\overline{p}) = 0.6923$$

$$\rho_{ap} = \frac{s_{ap}}{s_a s_p} \approx 0.5233 \quad \rho_{vp} = \frac{s_{vp}}{s_v s_p} \approx 0.023$$

Chiaramente le variabili *voto* e *peso* sono incorrelate, mentre vi è una correlazione positiva tra le variabili *altezza* e *peso*: persone più alte tendono a pesare di più.

Non è possibile calcolare il coefficiente di correlazione tra le variabili *altezza* e *voto* perché la frequenza congiunta $f_{av}(i,j)$ non è nota.

4. la frequenza congiunta definita da $f(i,j) = f_v(i)f_p(j)/n$ è data dalla seguente tabella:

peso / voto	50-60	60-70	70-80	80-90
19-21	.75	1.375	1.75	1.125
22-24	1.2	2.2	2.8	1.8
25-27	2.25	4.125	5.25	3.375
28-30	1.8	3.3	4.2	2.7

Rappresentiamo nella seguente tabella il rapporto $f(i,j)/f_v(i,j)$:

peso voto	50-60	60-70	70-80	80-90
19-21	0.75	1.375	0.875	1.125
22-24	1.2	1.1	0.9333	0.9
25-27	0.75	1.0312	1.05	1.125
28-30	1.8	0.825	1.05	0.9

Constatiamo che i valori sono molto prossimi a 1: infatti se le variabili sono incorrelate la distribuzione di frequenza congiunta è bene approssimata dal prodotto delle frequenze marginali diviso il numero dei dati.