

Università di Cassino

Esercitazione di Statistica 1 del 30 ottobre 2006

Dott.ssa Simona Balzano

Considerando il DATASET STUDENTI:

- 1) Costruire le distribuzioni di frequenza per i caratteri SESSO, N.SCARPA e CORSO LAUREA e calcolare per tali distribuzioni gli indici di posizione e di variabilità.
- 2) Verificare la proprietà associativa della media aritmetica del carattere PESO rispetto al carattere CORSO LAUREA.
- 3) L'altezza degli studenti iscritti al corso di laurea in fisica è uguale a 170 cm più il 5% dell'altezza degli iscritti in informatica. Sapendo gli informatici hanno un'altezza media pari a 174,16667 cm, con una varianza di 41,8, qual è l'altezza media dei fisici?

Soluzioni

1)

A. Carattere SCELTA

La distribuzione di frequenza del carattere è:

SCELTA	n_i	f_i
M	12	0,60
F	8	0,40
Totale	20	1

Essendo il carattere di tipo qualitativo nominale (dicotomico) possiamo calcolare la moda e gli indici di mutabilità.

Moda

$$M_o = M$$

Indice di Gini

$$G = 1 - \sum_{i=1}^K f_i^2 = 1 - (0,6^2 + 0,4^2) = 1 - (0,36 + 0,16) = 0,48$$

$$G_{max} = 1 - \frac{1}{K} = 1 - \frac{1}{2} = 0,5$$

Indice di eterogeneità normalizzato di Gini ($0 \leq G^* \leq 1$):

$$G^* = \frac{G}{G_{max}} = \frac{0,48}{0,5} = \mathbf{0,96}$$

Indice di entropia di Shannon:

$$\begin{aligned} H &= -\sum_{i=1}^2 f_i \cdot \log(f_i) = \\ &= -[0,6 \times \log(0,6) + 0,4 \times \log(0,4)] = \\ &= -[0,6 \times (-0,222) + 0,4 \times (-0,398)] = -[-0,133 - 0,159] = 0,2923 \\ H_{max} &= \log(K) = \log 2 = 0,3 \end{aligned}$$

Indice normalizzato di Shannon ($0 \leq H^* \leq 1$):

$$H^* = \frac{H}{H_{max}} = \frac{0,2923}{0,3} = \mathbf{0,971}$$

Conclusioni:

G^* ed H^* molto vicini a 1 \rightarrow la distribuzione è molto eterogenea: entrambe le modalità sono presenti e con frequenze equilibrate tra loro.

B. Carattere NUMERO SCARPA

La distribuzione di frequenza è la seguente:

N. scarpa	n_i	f_i	F_i
36	1	0,05	0,05
37	4	0,2	0,25
38	2	0,1	0,35
39	1	0,05	0,4
40	1	0,05	0,45
41	2	0,1	0,55
42	3	0,15	0,7
43	4	0,2	0,9
45	2	0,1	1
Totale	20	1	

Il carattere è di tipo quantitativo discreto, quindi è possibile calcolare tutti gli indici di posizione e di variabilità.

Media aritmetica

$$\mu = \frac{\sum_{i=1}^n x_i}{n} = \frac{809}{20} = \mathbf{40,45}$$

Moda

$$Mo_1 = 37; \quad Mo_2 = 43$$

Mediana

Consideriamo la successione ordinata di valori del carattere:

36 37 37 37 37 38 38 39 40 41 | 41 42 42 42 43 43 43 43 45 45

Essendo n pari la mediana è ottenuta come:

$$\text{Me} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{x_{10} + x_{11}}{2} = \mathbf{41}$$

Quartili

Osservando le frequenze cumulate:

$$Q_1 = \mathbf{37}$$

$$Q_3 = \mathbf{43}$$

Range

$$\text{Range} = x_{\max} - x_{\min} = 45 - 36 = 9$$

Differenza interquartile

$$Q_3 - Q_1 = 43 - 37 = 6$$

Varianza

Metodo 1: A partire dalla distribuzione semplice:

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} = \\ &= \frac{(36 - 40,45)^2 + (37 - 40,45)^2 + (37 - 40,45)^2 + \dots + (45 - 40,45)^2}{20} = \\ &= \frac{156,95}{20} = \mathbf{7,8475}\end{aligned}$$

Metodo 2: A partire dalla distribuzione di frequenza:

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^k (x_i - \mu)^2 n_i}{N} = \\ &= \frac{(36 - 40,45)^2 \times 1 + (37 - 40,45)^2 \times 2 + \dots + (45 - 40,45)^2 \times 2}{20} = \\ &= \frac{156,95}{20} = \mathbf{7,8475}\end{aligned}$$

Metodo 3: Formula alternativa:

$$\begin{aligned}\sigma^2 &= \frac{\sum_{i=1}^n x_i^2}{n} - \mu^2 = \frac{\sum_{i=1}^k x_i^2 n_i}{n} - \mu^2 = \\ &= \frac{36^2 + 37^2 + \dots + 45^2}{20} - 40,45^2 = \frac{32881}{20} - 1636,2 = \mathbf{7,8475}\end{aligned}$$

Scarto quadratico medio

$$\sigma = \sqrt{\sigma^2} = \sqrt{7,8475} = \mathbf{2,8}$$

Coefficiente di variazione

$$CV = \frac{\sigma}{\mu} = \frac{2,8}{40,45} = \mathbf{0,07}$$

C. Carattere CORSO LAUREA

Il carattere è di tipo qualitativo nominale e segue la seguente distribuzione di frequenza:

Negozi	n_i	f_i
Matematica	7	0,35
Informatica	12	0,6
Biologia	1	0,05
Totale	20	1

Moda

Mo = **Informatica**

Indice di Gini

$$G = 1 - \sum_{i=1}^4 f_i^2 = 1 - (0,35^2 + 0,6^2 + 0,05^2) = 1 - (0,1225 + 0,36 + 0,0025) = \mathbf{0,485}$$

$$G_{max} = 1 - \frac{1}{K} = 1 - \frac{1}{3} = 0,67$$

Indice di eterogeneità normalizzato di Gini ($0 \leq G^* \leq 1$):

$$G^* = \frac{G}{G_{max}} = \frac{0,485}{0,67} = \mathbf{0,73}$$

Indice di entropia di Shannon:

$$H = -\sum_{i=1}^K f_i \cdot \log(f_i) =$$
$$= -[0,35 \times \log(0,35) + 0,6 \times \log(0,6) + 0,05 \times \log(0,05)] =$$
$$= -[0,35 \times (-0,15958) + 0,6 \times (-0,13311) + 0,05 \times (-0,06505)] = \mathbf{0,357}$$

$$H_{max} = \log(K) = \log(3) = 0,477$$

Indice normalizzato di Shannon ($0 \leq H^* \leq 1$):

$$H^* = \frac{H}{H_{max}} = \frac{0,357}{0,477} = \mathbf{0,75}$$

Conclusioni:

G^* ed H^* abbastanza elevati \rightarrow la distribuzione è abbastanza eterogenea: tutte le modalità sono presenti ma con frequenze non molto simili tra loro.

2)

Secondo il raggruppamento indotto dalle modalità del carattere CORSO LAUREA, verifichiamo che la media generale del PESO è uguale alla media delle medie di gruppo.

$$n = 20$$

$$\mu = \frac{\sum_{i=1}^n x_i}{n} = \frac{1281}{20} = \mathbf{64,05}$$

Le medie dei $k = 3$ gruppi (le 3 modalità di CORSO LAUREA) sono:

B	I	I	I	I	I	I	I	I	I	I	I	I	M	M	M	M	M	M	M
49	65	62	50	49	68	85	56	72	65	75	70	80	47	48	56	59	73	86	66

$$\mu_B = 49$$

$$\mu_I = \frac{\sum_{i=1}^{n_I} x_{I_i}}{n_I} = \frac{797}{12} = 66,41667$$

$$\mu_M = \frac{\sum_{i=1}^{n_M} x_{M_i}}{n_M} = \frac{435}{7} = 62,14286$$

La proprietà associativa è verificata dalla seguente uguaglianza:

$$\frac{\sum_{i=1}^3 \mu_i n_i}{n} = \frac{(49 \times 1) + (66,41667 \times 12) + (62,14286 \times 7)}{20} = \frac{1280}{20} = \mathbf{64,05 = \mu}$$

3)

L'altezza dei fisici è deducibile a partire da quella degli informatici, cui è legata dalla relazione lineare:

$$\text{ALTEZZA}_F = 170 + 0,05 \text{ALTEZZA}_I$$

Per conoscere la media dell'altezza dei fisici non è necessario calcolare le altezze delle singole unità, perché la media aritmetica gode della proprietà della linearità, per cui può essere ottenuta come:

$$\begin{aligned} \mu_F &= a + b\mu_I = \\ &= 170 + 0,05 \mu_I = 170 + 0,05 \times 174,16 = 178,7083 \end{aligned}$$