Università di Cassino

Esercitazione di Statistica 1 del 6 novembre 2006

Dott.ssa Simona Balzano

Considerando il DATASET STUDENTI risolvere i seguenti quesiti:

1) 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 che la varianza dell'altezza degli informatici è pari a 41,8, qual è la varianza dell'altezza dei fisici?
2) Calcolare per il carattere PESO lo scostamento semplice medio dalla media e lo scostamento semplice medio dalla mediana.
3) Per il carattere quantitativo ISCRITTI osservato nei 3 corsi di laurea (n = 3) misurare la mutua variabilità attraverso la differenza semplice media e quadratica.
4) Verificare la proprietà di scomponibilità della devianza per la distribuzione del carattere PESO rispetto al CORSO LAUREA.
5) Calcolare il grado di concentrazione del carattere ISCRITTI a partire dalla segunte distribuzione:

osservazione	x_i
B	540
I	2473
M	480
F	516
C	642

e costruire il diagramma di Lorenz.

Soluzione

1)
L'altezza dei fisici è legata a quella degli informatici dalla relazione lineare:

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

La varianza di una variabile Y = a + bX può essere ottenuta a partire dalla varianza di X come:

$$\sigma_y^2 = b^2\sigma_x^2$$

Quindi:

$$\sigma_F^2 = b^2\sigma_I^2 =$$

$$= 0{,}05^2 \times 41{,}8 = 0{,}104$$

2)

$$S(\mu) = \frac{\sum_{i=1}^{N} |x_i - \mu|}{N} \; ; \qquad S(Me) = \frac{\sum_{i=1}^{N} |x_i - Me|}{N}$$

La media del carattere PESO è 64,05.
La mediana è 65.

A. Scostamento semplice medio dalla media

$$S(\mu) = \frac{|65-64{,}05| + |62-64{,}05| + \cdots + |49-64{,}05|}{20} = \frac{200{,}9}{20} = \mathbf{10{,}045}$$

B. Scostamento semplice medio dalla mediana

$$S(Me) = \frac{|47-65| + |48-65| + \cdots + |86-65|}{20} = \frac{199}{20} = \mathbf{9{,}95}$$

3)
Il carattere quantitativo ISCRITTI osservato rispetto alle 3 modalità del carattere qualitativo corso di laurea da origine alla seguente distribuzione semplice di n = 3 unità:

osservazione	x_i
B	1
I	7
M	12

la cui media aritmetica è data da:

$$\mu = \frac{\sum_{i=1}^{3} x_i}{3} = \frac{1+12+7}{3} = 6{,}67$$

Per il calcolo della **differenza media semplice**:

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|}{n(n-1)} = \frac{2\sum_{i>j} |x_i - x_j|}{n(n-1)}$$

può essere di aiuto costruire la seguente tabella delle differenze in valore assoluto tra tutte le coppie (x_i, x_j):

	1	7	12	
1	0	6	11	
7	6	0	5	
12	11	5	0	
				44

la cui somma è pari a 44 e da cui si ricava:

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|}{n(n-1)} = \frac{44}{3(3-1)} = \frac{44}{6} = 7,33$$

$$\Delta_{MAX} = 2\mu = 2\times 6,67 = 13,34$$

Differenza media semplice normalizzata $0 \le R \le 1$:

$$R = \frac{\Delta}{\Delta_{max}} = \frac{7,33}{13,34} = 0,55$$

Conclusione: R ha un valore intermedio tra 0 ed 1 $\rightarrow$ le osservazioni si differenziano l'una dall'altra ma non in misura molto elevata.

La differenza media quadratica è data da:

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2}{n(n-1)} = \frac{2\sum_{i>j}\left(x_i - x_j\right)^2}{n(n-1)}$$

La tabella dei quadrati delle diferenze è:

	1	7	12
1	0	36	121
7	36	0	25
12	121	25	0

Quindi:

$$D = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(x_i - x_j\right)^2}{n(n-1)} = \frac{364}{3(3-1)} = 60,67$$

4)

La verifica della proprietà della scomponibilità può essere effettuata indifferentemente rispetto alla devianza o alla varianza.
Di seguito sono riportati entrambi i procedimenti affiancati, per i quali si riportano in rosso i risultati relativi alla devianza ed in blu quelli relativi alla varianza.

Le modalità del carattere CORSO DI LAUREA definiscono 3 gruppi, di seguito riportati con i rispettivi valori del carattere PESO:

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

Calcolo di <u>devianza e varianza totale</u> indicate con **dev(X)** e $\boldsymbol{\sigma_X^2}$:

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

Devianza totale	Varianza totale
$\text{dev}(X) = \sum_{i=1}^{20}(x_i - \mu)^2 = \mathbf{2852{,}95}$	$\sigma_X^2 = \frac{\sum_{i=1}^{20}(x_i-\mu)^2}{20} = \frac{2852{,}95}{20} = \mathbf{142{,}6475}$

Calcolo delle medie dei gruppi:

$$\mu_{BIO} = 49$$

$$\mu_{INF} = \frac{\sum_{i=1}^{n_{INF}} x_{INF_i}}{n_{INF}} = \frac{797}{12} = 66{,}41667$$

$$\mu_{MAT} = \frac{\sum_{i=1}^{n_{MAT}} x_{MAT_i}}{n_{MAT}} = \frac{435}{7} = 62{,}14286$$

Calcolo di devianza e varianza <u>tra i gruppi</u> o <u>esterna</u> indicate con $\mathbf{dev_{EXT}}$ o $\mathbf{dev_B}$ (between) e σ_{EXT}^2:

Devianza esterna	Varianza esterna
$\text{dev}_{EXT} = \sum_{i=1}^{G}(\mu_i - \mu)^2 n_i =$ $(49-64{,}05)^2 \times 1 +$ $+(66{,}41-64{,}05)^2 \times 12 +$ $+(62{,}143-64{,}05)^2 \times 7 = \mathbf{319{,}17619}$	$\sigma_{EXT}^2 = \frac{\sum_{i=1}^{G}(\mu_i-\mu)^2 n_i}{n} =$ $\frac{1}{20}\Big[(49-64{,}05)^2 \times 1 + +(66{,}41-64{,}05)^2 \times 12 +$ $+(62{,}143-64{,}05)^2 \times 7\Big] = \frac{319{,}17619}{20} = \mathbf{15{,}95881}$

Calcolo di devianza e varianza entro i gruppi o interna, indicate con $\mathbf{dev_{INT}}$ o $\mathbf{dev_{W}}$ (within) e σ^2_{INT}:

Devianza interna	Varianza interna
$dev_{INT} = \sum_{i=1}^{G}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2 =$ $= \left[(49-49)^2\right] +$ $+\left[(65-66,41667)^2+\cdots+(80-66,41667)^2\right]+$ $+\left[(47-62,14286)^2+\cdots+(66-62,14286)^2\right]=$ $=$ **2533,77381**	$\sigma^2_{INT} = \dfrac{\sum_{i=1}^{G}\sum_{j=1}^{n_i}(x_{ij}-\mu_i)^2}{20} =$ $= \dfrac{1}{20}\left\{\left[(49-49)^2\right]+\right.$ $+\left[(65-66,41667)^2+\cdots+(80-66,41667)^2\right]+$ $\left.+\left[(47-62,14286)^2+\cdots+(66-62,14286)^2\right]\right\}=$ $= \dfrac{2533,77381}{20} =$ **126,68869**

✓ La devianza interna può essere calcolata come **somma** delle devianze entro ciascun gruppo, ossia:

$$dev(BIO) = 0$$
$$dev(INF) = \sum_{i=1}^{12}\left(x_{INF_i}-\mu_{INF}\right)^2 = 1354,917$$
$$dev(MAT) = \sum_{i=1}^{7}\left(x_{MAT_i}-\mu_{MAT}\right)^2 = 1178,857$$

la cui somma è proprio **2533,77381**

✓ La varianza interna può essere calcolata come **media** delle devianze entro ciascun gruppo, ossia:

$$\sigma^2_{INT} = \frac{\sum_{i=1}^{G} dev_i}{n} = \frac{dev(BIO)+dev(INF)+dev(MAT)}{20} = \mathbf{126,68869}$$

Verifica della proprietà della scomponibilità:

Per la devianza:

$$dev_{EXT} + dev_{INT} = dev(X)$$

$$\mathbf{319{,}17619 + 2533{,}77381 = 2852{,}95}$$

Per la varianza:

$$\sigma^2_{EXT} + \sigma^2_{INT} = \sigma^2_X$$

$$\mathbf{15{,}95881 + 126{,}68869 = 142{,}64750}$$

5)

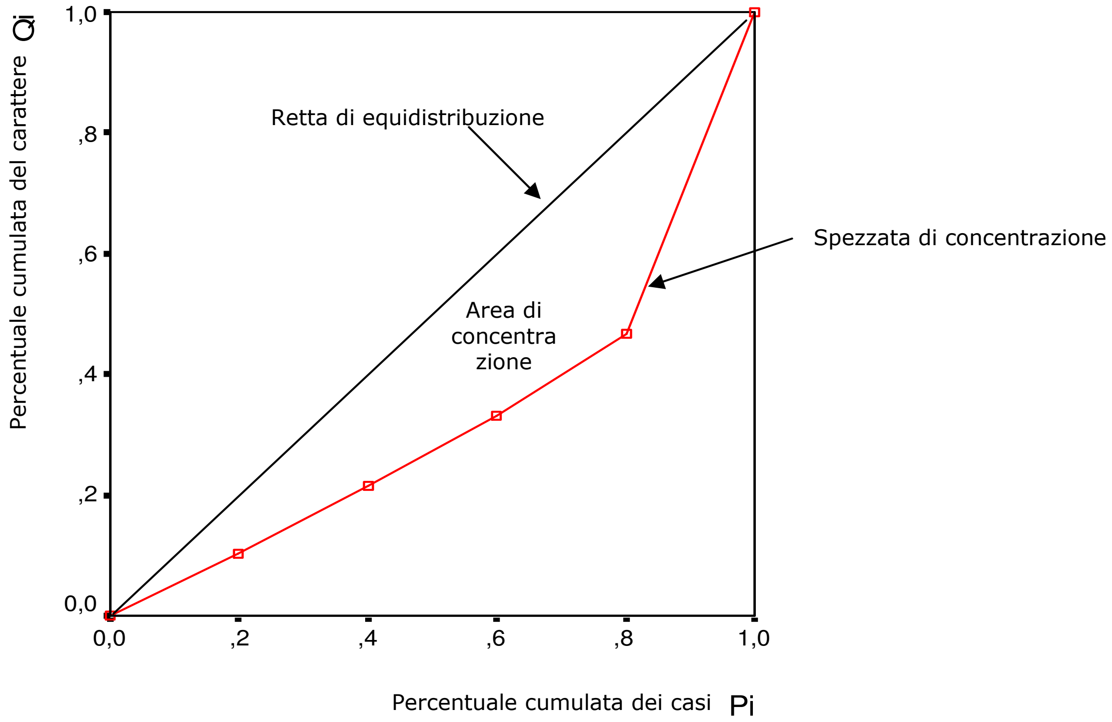
Per il calcolo del rapporto di concentrazione di Gini sono necessarie le seguenti quantità:

osservazione	x_i	p_i	A_i	q_i	p_i-q_i
M	480	0,2	480	0,103	0,097
F	516	0,4	996	0,214	0,186
B	540	0,6	1536	0,330	0,270
C	642	0,8	2178	0,468	0,332
I	2473	-	4651		
	4651	**2**			**0,884**

Da cui si ricava:

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{0,884}{2} = \mathbf{0,44}$$

La rappresentazione grafica del rapporto di Gini è la curva di Lorentz, dove l'area compresa tra la spezzata di concentrazione e la retta di equidistribuzione è pari ad R


Percentuale cumulata del carattere Qi
1,0
,8
,6
,4
,2
0,0
Retta di equidistribuzione
Spezzata di concentrazione
Area di concentra zione
0,0
,2
,4
,6
,8
1,0
Percentuale cumulata dei casi Pi