

MORE ESERCISES ON POSITION INDICES

Esercizio 4

I seguenti dati forniscono il fatturato annuo in milioni di vecchie lire di **171** industrie tessili.

Fatturato	Freq. ass.
[300,500]	20
(500,800]	45
(800,1500]	56
(1500,2000]	50
TOTALE	171

- Qual è la **percentuale di industrie** con fatturato annuo **superiore a 500 milioni e non superiore a 1.5 miliardi?**
- Quale è la **classe modale** del fatturato?
- Quale è il **fatturato medio?**

Classe	Freq. ass.	Freq. rel.	Ampiezza della classe
[300,500]	20	20/171	200
(500,800]	45	45/171	300
(800,1500]	56	56/171	700
(1500,2000]	50	50/171	500
TOTALE	171	1	

- Quale è la percentuale di industrie con fatturato annuo **superiore a 500 milioni e non superiore a 1.5 miliardi?**

È la somma delle frequenze relative delle classi (500, 800] e (800, 1500] moltiplicata per 100.

$$\left(\frac{45}{171} + \frac{56}{171} \right) \cdot 100 = \frac{101}{171} \cdot 100 = 59.06\%$$

Classe	Freq. ass.	Freq. rel.	Ampiezza della classe
[300,500]	20	20/171	200
(500,800]	45	45/171	300
(800,1500]	56	56/171	700
(1500,2000]	50	50/171	500
TOTALE	171	1	

- Quale è la **classe modale** del fatturato?

È la classe con la frequenza più elevata (ossia quella corrispondente alla **moda**), quindi è **(800, 1.500]**.

- Quale è il **fatturato medio**?

Il calcolo esatto non è possibile ma è sensato supporre che i dati di ogni classe siano ben approssimati dal **valore centrale della classe**

$$\bar{x} = \frac{1}{171} (400 \cdot 20 + 650 \cdot 45 + 1.150 \cdot 56 + 1.750 \cdot 50) = \mathbf{1.106,14}$$

Esercizio 5.

Si considerino i dati

2 0 3 2 4 **x**

per quali dei seguenti valori di **x** la mediana vale 2?

☐ 0

☐ 3

☐ 4

☐ 3.5

La taglia del campione è 6, numero pari. Quindi la mediana è la media aritmetica tra il terzo ed il quarto valore $m = \frac{x_3 + x_4}{2}$.

Se **x** = 0 disponendo i dati in ordine crescente si ha

0 0 2 2 3 4

e quindi la mediana è 2.

Esercizio 6.

Un lavoratore neoassunto per decidere se usare la metropolitana o l'automobile per recarsi al lavoro fa la seguente indagine:
va a lavoro per i primi 12 giorni in automobile e per i successivi 12 in metropolitana e registra il tempo impiegato in minuti.

Giorno	1	2	3	4	5	6	7	8	9	10	11	12
Auto	23	32	44	21	36	30	28	33	45	34	29	31
Metro	22	24	22	33	26	31	24	28	32	31	37	24

$\bar{x}_A$ **media campionaria auto:**

$$\bar{x}_A = \frac{1}{12} (23+32+44+21+36+30+28+33+45+34+29+31) = 32,167$$

x_M **media campionaria metro:**

$$\bar{x}_M = \frac{1}{12} (22+24+22+33+26+31+24+28+32+31+37+24) = 27,833$$

INDICI DI DISPERSIONE

Consideriamo le seguenti due variabili x e y entrambe di taglia 5:

$$\begin{array}{rcccccc} x : & 14 & 15 & 16 & 16 & 19 \\ y : & 0 & 1 & 16 & 17 & 46 \end{array}$$

Queste variabili hanno la stessa mediana $m = 16$ e la stessa media:

$$\begin{aligned}\bar{x} &= \frac{14 + 15 + 2 \cdot 16 + 19}{5} = \frac{80}{5} = 16 \\ \bar{y} &= \frac{1 + 16 + 17 + 46}{5} = \frac{80}{5} = 16\end{aligned}$$

Ma i dati sono molto diversi: i valori di x sono vicini mentre quelli di y sono sparsi. Si rende necessario introdurre un **indice di dispersione**.

VARIANZA CAMPIONARIA

Data una variabile x di taglia n e di valori $x_1, \dots, x_n$, i seguenti numeri

$$x_i - \bar{x}, \quad i = 1, 2, \dots, n$$

si chiamano **scarti dalla media**. La somma degli scarti è nulla:

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = x_1 + x_2 + \dots + x_n - n\bar{x} = 0$$

Per avere un indice di dispersione significativo si introduce la **varianza campionaria**

$$\begin{aligned} s^2 &= \frac{1}{n-1} \left[(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2 \right] \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \end{aligned}$$

Calcoliamo le varianze dei dati dell'esempio considerato prima:

$$\begin{array}{rcccccc} x : & 14 & 15 & 16 & 16 & 19 \\ y : & 0 & 1 & 16 & 17 & 46 \end{array}$$

Avevamo calcolato $\bar{x} = 16 = \bar{y}$.

$$\begin{aligned} s_x^2 &= \frac{1}{4} \left[(14 - 16)^2 + (15 - 16)^2 + 2 \cdot (16 - 16)^2 + (19 - 16)^2 \right] \\ &= \frac{1}{4} [4 + 1 + 9] = \frac{14}{4} = 3,5 \end{aligned}$$

$$\begin{aligned} s_y^2 &= \frac{1}{4} \left[(0 - 16)^2 + (1 - 16)^2 + (16 - 16)^2 + (17 - 16)^2 + (46 - 16)^2 \right] \\ &= \frac{1}{4} [256 + 225 + 1 + 900] = \frac{1382}{4} = 345,5 \end{aligned}$$

La varianza 345,5 è un numero grande come ci aspettavamo, inoltre è un numero molto maggiore di 3,5 che è la varianza di x .

UNA PROPRIETA' DELLA VARIANZA CAMPIONARIA

Vale la seguente identità:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$$

La dimostrazione è una semplice verifica, infatti:

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

dove si è usata la definizione di media:

$$\sum_{i=1}^n x_i = n\bar{x}$$

SCARTO QUADRATICO MEDIO

Si definisce scarto quadratico medio o **deviazione standard** la **radice quadrata della varianza**

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)}$$

Nota che lo scarto quadratico medio è misurato nelle stesse unità delle variabili x_i e della loro media $\bar{x}$.

Esercizio

Calcolare la varianza campionaria s^2 del seguente campione di taglia $n = 5$:

1 2 5 6 6

La **media campionaria** è : $\bar{x} = \frac{1}{5}(1 + 2 + 5 + 2 \cdot 6) = \frac{20}{5} = 4$

Usando la formula $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ otteniamo:

$$s^2 = \frac{1}{4} [(-3)^2 + (-2)^2 + 1 + 2 \cdot 2^2] = \frac{1}{4} (9 + 4 + 1 + 2 \cdot 4) = \frac{22}{4}$$

Usando la formula $s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$

$$s^2 = \frac{1}{4} [1 + 2^2 + 5^2 + 2 \cdot 6^2 - 5 \cdot 4^2] = \frac{1}{4} (1 + 4 + 25 + 2 \cdot 36 - 80) = \frac{22}{4}$$

CALCOLO DELLA VARIANZA CON LE FREQUENZE

Esempio. Abbiamo i seguenti dati disposti in una tabella delle frequenze assolute

Valore	Freq. assoluta
4	1
6	4
7	2
TOTALE	7

Quindi l'insieme dei dati originali è composto da 7 valori che disposti in modo crescente sono:

4 6 6 6 6 7 7.

Abbiamo visto che la media campionaria è $\bar{x} = 6$. La varianza campionaria è

$$s^2 = \frac{1 \cdot (4 - 6)^2 + 4 \cdot (6 - 6)^2 + 2 \cdot (7 - 6)^2}{7 - 1} = 6$$

In generale sia x una variabile numerica di taglia n con k valori distinti ordinati $x_1 < x_2 < \dots < x_k$ con frequenze assolute $n_1, n_2, n_3, \dots, n_k$ ($n_1 + n_2 + \dots + n_k = n$). La varianza campionaria è data da

$$s^2 = \frac{n_1 \cdot (x_1 - \bar{x})^2 + n_2 \cdot (x_2 - \bar{x})^2 + \dots + n_k \cdot (x_k - \bar{x})^2}{n - 1}$$

which, given that $f_i = \frac{n_i}{N}$ can be eventually written as

$$s^2 = \frac{1}{n-1} \sum_{j=1}^k f_j (x_j - \bar{x})^2.$$

MORE ABOUT VARIANCE AND STANDARD DEVIATION

Let us define the vector ξ whose n components are the $x_i - \bar{x}$, i.e. the vector

$$\xi = (\xi_1, \xi_2, \dots, \xi_n) = (x_1 - \bar{x}, x_2 - \bar{x}, \dots, x_n - \bar{x})$$

We may write the **sample variance** as

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{|\xi|^2}{n-1}$$

and the **standard deviation** as

$$s = \sqrt{s^2} = \frac{|\xi|}{\sqrt{n-1}}$$

Note that the standard deviation is measured in the same units as the variables x_i and their mean $\bar{x}$.

UN ALTRA PROPRIETA' DELLA VARIANZA CAMPIONARIA

Se si somma una costante c a tutti i dati la varianza non cambia.

In formule: data una variabile x di taglia n con valori $x_1, \dots, x_n$ si definisce la variabile y con valori $y_i = x_i + c$, $i = 1, \dots, n$.

Sappiamo che $\bar{y} = \bar{x} + c$ quindi

$$\begin{aligned}s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i + c - \bar{x} - c)^2 \\ &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = s_x^2\end{aligned}$$

Utile per fare i calcoli.

ANCORA UNA PROPRIETA' DELLA VARIANZA CAMPIONARIA

Si moltiplicano tutti i valori per un numero $c \neq 0$ e si ottiene la variabile numerica y con valori

$$y_i = cx_i, \quad i = 1, 2, \dots, n$$

La media campionaria $\bar{y} = c\bar{x}$ e la varianza

$$\begin{aligned} s_y^2 &= \frac{1}{n-1} \left(\sum_{i=1}^n (cx_i)^2 - n(c\bar{x})^2 \right) \\ &= \frac{1}{n-1} \left(\sum_{i=1}^n c^2 x_i^2 - n c^2 \bar{x}^2 \right) \\ &= c^2 \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n \bar{x}^2 \right) = c^2 s_x^2 \end{aligned}$$

RIEPILOGO

Data una variabile numerica x di taglia n con valori

$$x_1, x_2, \dots, x_n$$

e due numeri non nulli a e b , la variabile numerica y di taglia n con valori

$$y_1 = ax_1 + b, y_2 = ax_2 + b, \dots, y_n = ax_n + b$$

valgono le seguenti relazioni:

$$\bar{y} = a\bar{x} + b, \quad s_y^2 = a^2 s_x^2$$

che possono essere direttamente verificate come **Esercizio**.

La seconda relazione può anche essere espressa in termini di scarto quadratico come $s_y = |a|s_x$ dove $|a| = \sqrt{a^2}$ è il modulo (o valore assoluto) di a .

Esercizio di ricapitolazione 1

Si consideri il seguente campione di taglia $n = 9$:

1 6 5 6 2 5 2 6 3

Si calcoli la **mediana**, la **moda**, la **media campionaria**, la **varianza** e lo **scarto quadratico**.

La **media campionaria** è :

$$\bar{x} = \frac{1}{9}(1 + 6 + 5 + 6 + 2 + 5 + 2 + 6 + 3) = \frac{36}{9} = 4$$

Ordiniamo i dati in ordine crescente

1 2 2 3 5 5 6 6 6

la **mediana** corrisponde al valore centrale per cui $m = x_5 = 5$ e la **moda** corrisponde al valore della classe più numerosa per cui è uguale a 6.

Calcoliamo ora la **varianza campionaria s^2** :

Usando la formula $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ otteniamo:

$$\mathbf{s^2} = \frac{1}{8} (9 + 4 + 4 + 1 + 1 + 1 + 4 + 4 + 4) = \frac{32}{8} = \mathbf{4}$$

Usando la formula $s^2 = \frac{1}{n-1} \left(\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right)$:

$$\mathbf{s^2} = \frac{1}{8} (1 + 4 + 4 + 9 + 25 + 25 + 36 + 36 + 36 - 144) = \mathbf{4}$$

Lo **scarto quadratico** è semplicemente **s** = 2.

Esercizio di ricapitolazione 2

Le temperature alle ore 6 del primo gennaio sono state rilevate all'Aquila negli ultimi 100 anni. In questo esercizio le temperature sono le variabili x_i con $i = 1, 2, \dots, n$ e $n = 100$.

Dai dati risulta che la temperatura media, calcolata in gradi Celsius, è stata $\bar{x} = -2,4$ con scarto quadratico $s_x = 4,4$ (si noti che a volte si usa dire che la temperatura è stata di $-2,4 \pm 4,4$).

Quale è stata la temperatura media e lo scarto quadratico in gradi Fahrenheit?

Le temperature y_i in gradi Fahrenheit si ottengono dalle temperature x_i in gradi Celsius dalla relazione $y_i = ax_i + c$ con $a = 1,8$ e $c = 32$ (alla pressione di una atmosfera l'acqua congela a 32 gradi e bolle a 212 gradi Fahrenheit).

Si avrà $\bar{y} = a\bar{x} + c = -2,4 \cdot 1,8 + 32 = 27,68$ e $s_y = as_x = 4,4 \cdot 1,8 = 7,92$.

Esercizio di ricapitolazione 3

Assume you made the following observations:

x_j	f_j
1	8
2	5
3	2
4	3
5	4
6	2
7	1
8	2
9	1
10	2
k=10	n=30

where the x_i are the different observed values and the f_i are the absolute frequencies.

The mean is

$$\bar{x} = \frac{1}{n} \sum_j^k f_j x_j = 4,$$

which allows us to write

x_j	ξ_j	ξ_j^2	f_j
1	-3	9	8
2	-2	4	5
3	-1	1	2
4	0	0	3
5	1	1	4
6	2	4	2
7	3	9	1
8	4	16	2
9	5	25	1
10	6	36	2

where $\xi_j = x_j - \bar{x}$.

The mode equals 1 since the most populated class is that corresponding to $x_1 = 1$. The median is $m = \frac{x_{15} + x_{16}}{2} = \frac{3+4}{2} = 3.5$.

Moreover, the variance is

$$s^2 = \frac{1}{n-1} \sum_{j=1}^k f_j \xi_j^2 = 8.72$$

and the standard deviation is $s \simeq \sqrt{8.72} \simeq 2.95$.

We can resume this result by writing

$$x_i = \bar{x} \pm s = 4 \pm 3$$

which in a condensate form tells us that the values are distributed around 4, mostly in a region ranging from 1 to 7.