LA STATISTICA

Daniela Tondini

dtondini@unite.it

Facoltà di Bioscienze e Tecnologie agro-alimentari e ambientali e Facoltà di Medicina Veterinaria

C.L. in Biotecnologie

Università degli Studi di Teramo

INDICI STATISTICI

La ***mediana*** o ***valore mediano*** M_e è quell'indice di posizione che, una volta ordinate in senso crescente le osservazioni di un fenomeno, divide la distribuzione in due gruppi di uguale numerosità: al primo gruppo, infatti, appartengono le osservazioni uguali o inferiori alla mediana; al secondo gruppo, invece, quelle superiori o uguali alla mediana. La mediana, dunque, è la modalità dell'unità statistica che occupa il posto centrale nella distribuzione ordinata delle osservazioni. Dato, cioè, un insieme costituito da n intensità (x_1, x_2, …, x_n), la determinazione della mediana è diversa a seconda che n sia pari o dispari, precisamente si ha:

- se n è pari, la mediana è data dalla semisomma delle intensità individuate dalle due posizioni centrali, C_1 e C_2, ovvero dalla loro media aritmetica:

$$C_1 = x_{\frac{n}{2}}, \quad C_2 = x_{\frac{n}{2}+1} \quad \Rightarrow \quad M_e = \frac{C_1 + C_2}{2}$$

- se n è dispari, la mediana è data dal valore che occupa la posizione centrale nella distribuzione dei valori posti in graduatoria:

$$M_e = x_{\frac{n+1}{2}}$$

INDICI STATISTICI

Esempio

La mediana delle seguenti intensità ($n = 7$, dispari):

3; 15; 9; 2; 6; 12; 5

si ottiene ordinando dapprima le intensità in ordine crescente,

$$x_1 = 2;\ x_2 = 3;\ x_3 = 5;\ x_4 = 6;\ x_5 = 9;\ x_6 = 12;\ x_7 = 15$$

e poi considerando l'intensità che occupa il posto centrale, essendo n dispari:

$$M_e = x_4 = 6$$

INDICI STATISTICI

Esempio

La mediana delle seguenti intensità ($n = 8$, pari):

$$7;\ 16;\ 2;\ 3;\ 9;\ 12;\ 15;\ 5$$

si ottiene ordinando dapprima le intensità in ordine crescente,

$$x_1 = 2;\ x_2 = 3;\ x_3 = 5;\ x_4 = 7;\ x_5 = 9;\ x_6 = 12;\ x_7 = 15;\ x_8 = 16$$

e poi considerando le intensità che occupano i due posti centrali, essendo n pari:

$$C_1 = x_{\frac{8}{2}} = x_4 = 7, \quad C_2 = x_{\frac{8}{2}+1} = x_5 = 9 \quad \Rightarrow \quad M_e = \frac{7+9}{2} = \frac{16}{2} = 8$$

INDICI STATISTICI

Se, invece, si ha una distribuzione di frequenze, per calcolare la mediana, occorre determinare le frequenze cumulate: indicando con *n* la somma delle frequenze, se *n* è pari, la mediana è data da

$$\frac{n}{2}$$

Se, invece, *n* è dispari, la mediana è data da:

$$\frac{n+1}{2}$$

INDICI STATISTICI

Esempio

Se si effettua l'indagine su un numero di figli su un campione di famiglie, come riportato nella seguente tabella:

Figli x_i	F.A n_i	F.C.A.
0	3	3
1	8	11
2	7	18
3	4	22
4	1	23
5	1	24
6	1	25
Tot.	25	

essendo n dispari, la mediana è il valore corrispondente a

$$\frac{n+1}{2}=\frac{25+1}{2}=\frac{26}{2}=13$$

ovvero la mediana è 2 poiché 11<13<18.

INDICI STATISTICI

La mediana, pertanto, si può calcolare per tutte quelle variabili le cui modalità possono essere ordinate, ovvero per le variabili qualitative ordinali, e per tutte le variabili quantitative: risulta, infatti, più conveniente usarla qualora si voglia esprimere il valore centrale di distribuzioni di caratteri che non possono essere misurati "esattamente" (ad esempio, i caratteri psicologici graduabili) oppure qualora non si possa far riferimento alla distribuzione normale, proprio grazie alla sua capacità di essere rappresentativa della posizione della distribuzione anche in presenza di valori estremi notevolmente diversi da tutti gli altri.

La mediana, dunque, *minimizza i costi complessivi* ed è soprattutto resistente ai valori estremi: rappresenta, infatti, un indice per decisioni che implicano costi elevati nei casi estremi.

INDICI STATISTICI

La ***moda*** o ***norma*** M_O di una distribuzione di frequenza X, calcolabile per caratteri sia quantitativi sia qualitativi, non risentendo dei valori estremi, rappresenta la modalità, o classe di modalità, caratterizzata dalla massima frequenza (assoluta o relativa) o densità di frequenza, ovvero il valore numerico che, nella distribuzione di frequenza, è maggiormente presente rispetto agli altri. A tal riguardo occorre evidenziare che la moda è una modalità, non una frequenza. Se si rappresenta, pertanto, la distribuzione di frequenza in termini grafici, si può affermare che la moda corrisponde al picco della distribuzione (ad esempio in un grafico a colonne o a nastri, la colonna più alta o il nastro più lungo individua la moda della distribuzione) che, di conseguenza, risulterà *zeromodale* se non ammette alcun valore modale, ovvero nessun picco, *unimodale* se ne ammette uno solo (in tal caso la moda ha significato di sintesi), *bimodale* se ne ammette due, *trimodale* se ne ammette tre, … Per poter determinare, quindi, la classe modale risulta opportuno ricorrere all'istogramma, individuando l'intervallo di altezza massima, ovvero il punto di massimo della curva; la classe con la maggiore densità media, corrispondente proprio all'altezza dell'istogramma, sarà quella modale. La moda, dunque, *minimizza gli scontenti* ed è utilizzata in tutte quelle situazioni ove il consenso ed il numero delle singole unità ha significato per la decisione: la moda, infatti, è un indice utile per individuare la modalità più rappresentativa.

INDICI STATISTICI

Esempio

La moda della seguente successione di termini (n = 13):

$$x_1 = 3;\ x_2 = 5;\ x_3 = 9;\ x_4 = 3;\ x_5 = 5;\ x_6 = 7;\ x_7 = 3;$$
$$x_8 = 2;\ x_9 = 9;\ x_{10} = 3;\ x_{11} = 4;\ x_{12} = 3;\ x_{13} = 6$$

è data dal termine che compare con maggiore frequenza, ovvero è $M_O = 3$ perché compare 5 volte.

Esempio

Data la variabile X = numero di esami sostenuti da sei studenti ed osservati i seguenti valori:

STUDENTI	Nicola	Mary	Eleonora	Beatrice	Davide	Christian
ESAMI	30	19	8	7	27	10

Si può concludere che la variabile X non ha moda, ovvero è *zero modale*, essendo la moda definita come la modalità più frequente: non esiste, infatti, nessuna modalità (numero di esami) ripetuta più delle altre e tutte le modalità hanno la stessa frequenza assoluta pari ad uno studente.

Qual è la modalità più alta? 30

Qual è la modalità più frequente? Nessuna in quanto tutte hanno la stessa frequenza pari ad 1.

Per individuare la moda di una variabile, dunque, bisogna chiedersi in primo luogo qual è la variabile e poi quali sono le modalità e qual è la modalità con la frequenza più alta.

INDICI STATISTICI

Esempi

v.s. discrete

Voti x_i	Numeri di studenti n_i
25	3
26	2
27	8
28	1

v.s. continue di uguale ampiezza

Voti x_i	Numeri di studenti n_i
18---20	3
21---23	5
24---26	10
27---29	4

v.s. continue di diversa ampiezza

Voti x_i	Numeri di studenti n_i	d_i	$H_i = n_i / d_i$
18---21	5	3	5/3 = 1,6
21---23	4	2	4/2 = 2
24---28	6	4	6/4 = 1,5
29---30	3	1	3/1 = 3

INDICI STATISTICI

Si osservi che:

- per *caratteri discreti* la moda si individua facilmente scorrendo lungo la colonna delle frequenze;
- per *caratteri continui*, se le classi di modalità hanno tutte uguale ampiezza, la moda cade nella classe con maggiore frequenza; se le classi di modalità, invece, hanno ampiezza diversa, si divide ogni frequenza per l'ampiezza della rispettiva classe calcolando, così la densità di frequenza; la moda, poi, cade nella classe con maggiore densità di frequenza.

INDICI STATISTICI

I ***quantili*** sono le intensità che dividono, dopo aver ordinato i dati, una distribuzione di frequenza in un certo numero di parti uguali (ad esempio, la mediana è quel valore che divide in due parti uguali l'insieme delle unità ordinate per grandezza, ovvero la distribuzione è divisa, rispetto a tale valore, in due parti ognuna contenente il 50% delle unità). Se si divide la distribuzione in due parti si parla di *terzili* (il primo terzile è quello che lascia alla sua sinistra un terzo delle osservazioni e alla sua destra i rimanenti due terzi; il secondo terzile è quello che lascia alla sua sinistra i due terzi e alla sua destra un terzo rimanente). Se si divide la distribuzione in tre parti si parla di *quartili* (il primo quartile Q_1 lascia alla sua sinistra il 25% dei casi e alla sua destra il rimanente 75%; il secondo quartile Q_2, che coincide con la mediana, lascia alla sua sinistra il 50% dei casi e alla sua destra il rimanente 50%; il terzo quartile Q_3 lascia alla sua sinistra il 75% dei casi e alla sua destra il rimanente 25%). Se si divide la distribuzione in nove parti si parla di *decili*, …, in novantanove parti si parla di *centili*, in cento parti si parla di *percentili*.

INDICI STATISTICI

Se X è un carattere con n modalità ordinate $x_1, x_2, \ldots, x_n$ ($x_1 \leq x_2 \leq \ldots \leq x_n$), per il calcolo dei quartili si procede in maniera analoga a quanto visto in precedenza per la mediana, considerando le posizioni degli elementi:

- se n è pari:

$$Q_1 = \frac{x_{\frac{n}{4}} + x_{\frac{n}{4}+1}}{2}$$

- se n è dispari:

$$Q_1 = x_{\frac{n+1}{4}}$$

I quantili, dunque, si possono calcolare per tutte quelle variabili per le quali risulta possibile ordinarne le modalità, ovvero per variabili qualitative ordinali, oltre che per tutte le variabili quantitative.

INDICI STATISTICI

Esempio

Date le seguenti intensità ($n = 7$, dispari):

$$20;\ 65;\ 2;\ 10;\ 37;\ 15;\ 3$$

il loro quartile Q_1 si ottiene ordinando dapprima le intensità in ordine crescente:

$$x_1 = 2;\ x_2 = 3;\ x_3 = 10;\ x_4 = 15;\ x_5 = 20;\ x_6 = 37;\ x_7 = 65$$

e poi considerando, come primo quartile, l'intensità che occupa il posto:

$$x_{\frac{n+1}{4}} = x_{\frac{7+1}{4}} = x_{\frac{8}{4}} = x_2 = 3 = Q_1$$

Analogamente il terzo quartile Q_3 si ottiene considerando l'intensità che occupa sempre il secondo posto partendo, però, dall'ultima osservazione, ovvero $Q_3 = x_6 = 37$.

INDICI STATISTICI

Esempio

Date le seguenti intensità (n = 8, pari):

20; 65; 83; 10; 37; 15; 3; 2

il loro quartile Q_1 si ottiene ordinando dapprima le intensità in ordine crescente:

$x_1 = 2$; $x_2 = 3$; $x_3 = 10$; $x_4 = 15$; $x_5 = 20$; $x_6 = 37$; $x_7 = 65$; $x_8 = 83$

e poi considerando, come primo quartile, l'intensità che occupa il posto:

$$x_{\frac{n}{4}} = x_{\frac{8}{4}} = x_2 = 3; x_{\frac{n}{4}+1} = x_{\frac{8}{4}+1} = x_{2+1} = x_3 = 10$$

Effettuando, infine, la semisomma tra tali numeri, si ottiene:

$$Q_1 = \frac{3+10}{2} = \frac{13}{2} = 6,5$$

Analogamente il terzo quartile Q_3 si ottiene considerando la semisomma delle intensità che occupano sempre il secondo ed il terzo posto partendo, però, dall'ultima osservazione, ovvero:

$$Q_3 = \frac{37+65}{2} = \frac{102}{2} = 51$$

INDICI DI VARIABILITÀ

Il ***campo di variazione*** o ***range*** R di una sequenza n di numeri $x_1, x_2, \ldots, x_n$ si ottiene effettuando la differenza tra il dato più grande ed il dato più piccolo:

$$R = x_{\max} - x_{\min}$$

Il range, però, pur essendo molto semplice da calcolare, è poco significativo poiché tiene conto solo del valore più piccolo e di quello più grande, trascurando tutti gli altri valori. Può essere utile, ad esempio, in campo meteorologico quando viene indicata l'escursione termica.

Il campo di variazione, pertanto, fornisce informazioni sulla distribuzione dei dati:

- più R è piccolo, più i dati sono concentrati;
- più R è grande, più i dati sono dispersi.

INDICI DI VARIABILITÀ

Lo ***scarto quadratico medio*** o ***deviazione standard*** σ di una sequenza di numeri x_1, x_2, …, x_n rappresenta la media quadratica degli scarti dei dati dalla media aritmetica M_a; in formule è dato da:

$$\sigma = \sqrt{\frac{(x_1 - M_a)^2 + (x_2 - M_a)^2 + \ldots + (x_n - M_a)^2}{n}}$$

La ***varianza*** σ^2 di una sequenza n di numeri x_1, x_2, …, x_n, invece, è il quadrato dello scarto quadratico medio; in formule è data da:

$$\sigma^2 = \frac{(x_1 - M_a)^2 + (x_2 - M_a)^2 + \ldots + (x_n - M_a)^2}{n} = \frac{Dev}{n}$$

essendo *Dev* la ***devianza***, ovvero la somma dei quadrati degli scarti dei numeri dati dalla loro media aritmetica M_a.

INDICI DI VARIABILITÀ

Si osservi, però, che la varianza si può anche ottenere facendo la media dei quadrati meno il quadrato della media, ovvero in formule:

- se i dati sono senza frequenze:

$$\sigma^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - M_a^2$$

- se i dati sono con frequenze:

$$\sigma^2 = \frac{\sum_{i=1}^{s} x_i^2 \cdot n_i}{n = \sum_{i=1}^{s} n_i} - M_a^2$$

INDICI DI VARIABILITÀ

Esempio

Data la seguente tabella:

Valori x_i	F.A. n_i
2	3
4	1
8	2
11	4
Tot.	10

Calcolare scarto quadratico medio e varianza. Si ha:

$$M_a = \frac{2 \cdot 3 + 4 \cdot 1 + 8 \cdot 2 + 11 \cdot 4}{10} = 7$$

INDICI DI VARIABILITÀ

Ne segue che la varianza è data da:

$$\sigma^2 = \frac{\sum_{i=1}^{s} x_i^2 \cdot n_i}{\sum_{i=1}^{s} n_i} - M_a^2 = \frac{2^2 \cdot 3 + 4^2 \cdot 1 + 8^2 \cdot 2 + 11^2 \cdot 4}{10} - 7^2 = 15$$

e lo scarto quadratico medio è dato da:

$$\sigma = \sqrt{\sigma^2} = \sqrt{15} = 3{,}87$$

INDICI DI VARIABILITÀ

Il ***coefficiente di variazione*** *CV* è una misura relativa (le precedenti sono tutte assolute) di dispersione ed è una grandezza adimensionale particolarmente utile quando si devono confrontare le distribuzioni di due gruppi con medie molto diverse o con dati espressi in scale differenti (ad esempio, confronto tra variazione del peso e variazione dell'altezza). In formule, è dato da:

$$CV = \left(\frac{\sigma}{M_a} \cdot 100 \right) \%$$

INDICI DI VARIABILITÀ

Lo ***scostamento semplice medio*** $S(M_a)$ consiste nel calcolare la distanza di tutti i dati dalla media e fare la media aritmetica di tali distanze. In formule, è dato da:

- se i dati sono senza frequenze:

$$S(M_a) = \frac{\sum_{i=1}^{n} |x_i - M_a|}{n}$$

- se i dati sono con frequenze:

$$S(M_a) = \frac{\sum_{i=1}^{s} |x_i - M_a| \cdot n_i}{n = \sum_{i=1}^{s} n_i}$$

INDICI DI VARIABILITÀ

Esempio

Se si considerano le seguenti valutazioni delle tre prove degli esami di stato riportate da quattro studenti:

STUDENTI	Nicola	Mary	Eleonora	Giacomo
PRIMA PROVA	3	5	8	9
SECONDA PROVA	2	7	8	8
TERZA PROVA	6	7	6	6

si ha:

$$(M_a)_1 = (M_a)_2 = (M_a)_3 = 6,25$$

INDICI DI VARIABILITÀ

da cui gli scarti semplici medi delle tre prove sono rispettivamente:

$$\left[S(M_a)\right]_1 = \frac{|3-6,25|+|5-6,25|+|8-6,25|+|9-6,25|}{4} = \frac{3,25+1,25+1,75+2,75}{4} = \frac{9}{4} = 2,25$$

$$\left[S(M_a)\right]_2 = \frac{|2-6,25|+|7-6,25|+|8-6,25|+|8-6,25|}{4} = \frac{4,25+0,75+1,75+1,75}{4} = \frac{8,5}{4} = 2,125$$

$$\left[S(M_a)\right]_3 = \frac{|6-6,25|+|7-6,25|+|6-6,25|+|6-6,25|}{4} = \frac{0,25+0,75+0,25+0,25}{4} = \frac{1,5}{4} = 0,375$$

Si può osservare, quindi, che nella prima prova lo scarto, pari a 2,25 (ovvero i valori della sequenza si discostano mediamente di 2,25 dalla media), è superiore rispetto a quello della terza prova, i dati sono più dispersi ed i risultati più eterogenei; nella terza prova, in cui lo scarto è pari a 0,375, invece, i dati sono più concentrati ed i risultati più omogenei. La distribuzione della prima prova, inoltre, risulta diversa da quella della seconda prova. Dunque, più $S(M_a)$ è piccolo, più i dati sono concentrati, più $S(M_a)$ è grande più i dati sono dispersi. Inoltre, $S(M_a)$ è espresso nella stessa unità di misura dei dati ed $S(M_a)$ tiene conto di tutti i dati della distribuzione.