2°

Statistica2 – 22/09/2015

I Parametri di dispersione

Oltre alla tendenza centrale come riassumere la dispersione dei dati?

Campo di Variazione

PESO ALLA NASCITA DEI BOVINI

matricola	PESO	SESSO
7	38,00	F
8	38,00	F
1	40,00	F
2	40,00	F
5	40,00	F
10	42,00	F
3	47,00	F
9	47,00	M
4	50,00	M
6	50,00	M
CAMPO DI VARIAZIONE	**50,00**	MAX.
	38,00	min.

Costituisce la misura di dispersione più intuitiva definita anche come “range” o intervallo di dispersione.

Ha lo svantaggio di non essere influenzato dal numero delle osservazioni della serie
Si usano solo i **due valori** più esterni. I valori estremi, essendo i più rari, sono più influenzati dalle oscillazioni accidentali.

Si potrebbe pensare allo scarto di ogni singolo valore dalla media ovvero allo scarto della media da ogni singolo valore.

Scarti dalla media

Σ(x-media)

Contengono tutta l'informazione raccolta

Non si guadagna niente in semplificazione ed in comprensibilità anzi, se si aggiunge anche la media per avere maggiore chiarezza, si aggiunge un ulteriore numero alla serie originale.

PESO ALLA NASCITA DEI BOVINI			
media aritmetica		**43,20**	
matricola	PESO	differenze	SESSO
7	38,00	**-5,20**	F
8	38,00	**-5,20**	F
1	40,00	**-3,20**	F
2	40,00	**-3,20**	F
5	40,00	**-3,20**	F
10	42,00	**-1,20**	F
3	47,00	**3,80**	F
9	47,00	**3,80**	M
4	50,00	**6,80**	M
6	50,00	**6,80**	M

Σ = sommatoria

Analogamente al ragionamento fatto per la media aritmetica si potrebbe pensare alla media degli scarti di ogni singolo valore dalla media (ovvero alla “media degli scarti di ogni singolo valore dalla media ”).

Media = x = Σx/n

La somma degli scarti da sempre come risultato zero (vedi definizione di media).

PESO ALLA NASCITA DEI BOVINI			
matricola	PESO	differenza	SESSO
7	38,00	-5,20	F
8	38,00	-5,20	F
1	40,00	-3,20	F
2	40,00	-3,20	F
5	40,00	-3,20	F
10	42,00	-1,20	F
3	47,00	3,80	F
9	47,00	3,80	M
4	50,00	6,80	M
6	50,00	6,80	M

La media algebrica viene infatti definita anche come quella misura di posizione caratterizzata dal fatto che la somma algebrica degli scarti dei numeri della serie è sempre pari a zero

Si potrebbe pensare allora a non considerare il segno

Deviazione media

Tale dato non viene generalmente **MAI UTILIZZATO**

motivo:

Non permette calcoli successivi.
in termini statistici NON È EFFICACE

PESO ALLA NASCITA DEI BOVINI		
matricola	PESO	deviazione
7	38,00	5,20
8	38,00	5,20
1	40,00	3,20
2	40,00	3,20
5	40,00	3,20
10	42,00	1,20
3	47,00	3,80
9	47,00	3,80
4	50,00	6,80
6	50,00	6,80
Totale	432,00	~~**42,40**~~
media	**43,20**	~~4,24~~

pur:

Non avendo lo svantaggio di "usare" solo i due valori più esterni come per il campo di variazione

Si potrebbe pensare allora a non "perdere" (non considerandolo) il segno ma di sottoporre tutti i dati alla stessa operazione es. elevando gli scarti al quadrato

Somma dei Quadrati degli scarti o SS

Anche semplicemente **Somma dei Quadrati**

SS: Sum of Square

PESO ALLA NASCITA DEI BOVINI			
matricola	PESO	differenze	quadrati
7	38,00	-5,20	27,04
8	38,00	-5,20	27,04
1	40,00	-3,20	10,24
2	40,00	-3,20	10,24
5	40,00	-3,20	10,24
10	42,00	-1,20	1,44
3	47,00	3,80	14,44
9	47,00	3,80	14,44
4	50,00	6,80	46,24
6	50,00	6,80	46,24
Totale	432,00	0,00	**207,60**

Elevando le differenze al quadrato i valori diventano tutti positivi:
"meno per meno è uguale a più!"

A questo punto si potrebbe pensare di dividere la somma dei quadrati per il numero delle osservazioni per ottenere una sorta di media degli scarti **ma…… quante sono le osservazioni vere?**

La nozione:

Gradi di libertà - g.l. o d.f.

g.l.: Gradi di libertà

d.f.: degree of freedom.

Consideriamo la serie

Tutti i valori di una serie di misure sono indipendenti (dipendono solo dalle caratteristiche dell'oggetto misurato e dallo strumento impiegato per la misurazione. NESSUN VALORE PUÒ ESSERE DEDOTTO DALLA CONOSCENZA DEGLI ALTRI

Consideriamo ora gli scarti dalla media

Se $x_s = X - \bar{x}$,

pertanto

$Sx_s = SX - n\bar{x}$;

ma per definizione

$\bar{x} = SX/n$;

e dividendo per n

$Sx_s/n = SX/n - \bar{x}$

$Sx_s/n = \bar{x} - \bar{x} = 0$

Non tutti i valori di una serie di scarti sono indipendenti.
La conoscenza del primo scarto non determina il valore del secondo e così di seguito SOLO FINO AL PENULTIMO.
L'ultimo scarto di una serie non è indipendente:
la somma algebrica degli scarti è uguale a zero!
Passando dalla serie di misure alla serie degli scarti si "perde" un numero libero o grado di libertà!

Se uno scarto è uguale al valore meno la media la somma degli scarti sarà uguale alla somma dei singoli valori meno la media per il numero dei valori; poiché per definizione la media è uguale alla somma degli scarti diviso il numero dei valori la somma degli scarti sarà uguale a zero.

Quanto vale l'ultimo scarto di questa serie?

PESO ALLA NASCITA DEI BOVINI			
matricola	PESO	differenze	quadrati
1	38,00	-5,20	27,04
2	38,00	-5,20	27,04
3	40,00	-3,20	10,24
4	40,00	-3,20	10,24
5	40,00	-3,20	10,24
6	42,00	-1,20	1,44
7	47,00	3,80	14,44
8	47,00	3,80	14,44
9	50,00	6,80	46,24
10	50,00	???	???

Valore dell'ultimo scarto

PESO ALLA NASCITA DEI BOVINI				
matricola	PESO	differenze	quadrati	SESSO
10	**50,00**	**6,80**	**46,24**	M

Il valore del quadrato dell'ultimo scarto si deve trovare elevando al quadrato l'ultimo scarto

La somma dei quadrati degli scarti non è uguale a zero!

A questo punto si potrebbe Ripensare l'idea precedente e **dividere la somma dei quadrati per** il numero degli scarti **realmente indipendenti** cioè per **i gradi di libertà**!

n.b. **somma dei quadrati sta per somma dei quadrati degli scarti**

perchè c'è la necessità di dividere la somma dei quadrati per le osservazioni indipendenti?

Analogia con media

Da più fiducia un parametro stabilito a partire da una serie di 50 rilevazioni (= 49 gradi gradi di libertà degli scarti) oppure un parametro stabilito a partire da una serie di 3 rilevazioni (= 2 gradi di libertà degli scarti)?

3°

GRADI DI LIBERTÀ: (gl o df) esprimono il numero dei dati effettivamente disponibili per valutare la quantità di informazione contenuta nel parametro stesso

È evidente che dia più fiducia un parametro stabilito a partire da una serie di 50 rilevazioni rispetto ad uno stabilito a partire da una serie di 3 rilevazioni.

Non è però il numero totale dei dati che è importante ma, come abbiamo visto, il numero di dati indipendenti!

Cioè non **n** ma **n-1** = 50-1 = **49** e 3-1 = **2**

La nozione:

Varianza Var o MS

MS: Mean Square

VARIANZA: la media così calcolata dei quadrati degli scarti.
Somma dei quadrati degli scarti diviso i gradi di libertà.

$S(X-\bar{x})^2 /(n-1)$

$MS = \frac{SS\text{: Sum Square}}{df}$

La varianza non ha lo stesso "ordine di grandezza della media:
La **media** deriva da una **somma semplice** la **varianza** da una **somma di valori** (gli scarti) elevati **al quadrato**!

Si potrebbe pensare quindi di **uniformare gli "ordini di grandezza"** riportando la Varianza allo stesso "ordine di grandezza" della media aritmetica per renderla più "comprensibile".

la radice quadrata della varianza si chiama:

SD: Standard Deviation

DEVIAZIONE STANDARD: DS = alla radice quadrata della varianza.

Se un indice di dispersione è pari a 1,50 (per es.. la d.s.) su una media di 2,00 oppure su una media di 200,0 la variazione è evidentemente molto “diversa pur essendo uguale”, o meglio, la variazione non ha evidentemente lo stesso significato. Si potrebbe pensare pertanto di rapportare gli indici di variazione alla media:

COEFFICIENTE DI VARIAZIONE: CV = la deviazione standard diviso la media. Ovvero l’incidenza della variazione dei dati rapportata al valore della media degli stessi.

ERRORE STANDARD (DELLA MEDIA) = SE = la deviazione standard diviso la radice quadrata del numero delle osservazioni (non dei gradi di libertà in questo caso).

SE: Standard Error

Parametri di dispersione VALIDI

2°

- **VARIANZA:**
- **DEVIAZIONE STANDARD:**
- **COEFFICIENTE DI VARIAZIONE:**
- **ERRORE STANDARD:**

Dal valore di uno qualsiasi di questi parametri di dispersione si possono calcolare i valori di tutti gli altri. Vedi esercizi

Da R. Fisher:

statistico!

"La grande fiducia che i biologi hanno nel valore di questi parametri è probabilmente basata assai più sulla eccellenza dei risultati da loro ottenuti dopo anni di lavoro usandoli che non sulla dimostrazione formale (matematica)" della loro correttezza e validità"

La dimostrazione matematica non è facile da assimilare e può non essere indispensabile per dei "non statistici" o dei "non matematici".

in pratica per guidare la macchina non è necessario saperla progettare in toto. Bisogna però sapere che il freno è necessario!

I parametri di dispersione, così come la media, però non "**contengono più**" la dimensione del campione.

È necessario quindi riportare anche la dimensione del campione

cioè

3°

Il numero dei dati o delle osservazioni – n oppure i gradi di libertà

La **media** + un parametro di **dispersione** delle osservazioni + il **n**umero delle osservazioni (o i gl = df)

Tutti insieme contengono tutta l'informazione contenuta nei dati raccolti = <u>descrivono completamente le osservazioni</u>

questi tre parametri devono essere sempre riportati per la valutazione scientifica delle osservazioni biologiche.

Calcola tutti i parametri di dispersione validi per la seguente serie di dati:

PESO ALLA NASCITA DEI BOVINI	
matricola	PESO
1	40,00
2	40,00
3	47,00
4	50,00
5	40,00
6	50,00
7	38,00
8	38,00
9	47,00
10	42,00
Totale	432,00

PESO ALLA NASCITA DEI BOVINI

matricola	PESO	differenze	quadrati	SESSO
1	40,00	-3,20	10,24	F
2	40,00	-3,20	10,24	F
3	47,00	3,80	14,44	F
4	50,00	6,80	46,24	M
5	40,00	-3,20	10,24	F
6	50,00	6,80	46,24	M
7	38,00	-5,20	27,04	F
8	38,00	-5,20	27,04	F
9	47,00	3,80	14,44	M
10	42,00	-1,20	1,44	F
Totale	432,00	0,00	207,60	
media	**43,20**		**20,76**	
n.	**10**	10	10	
g.l.		9	9	
VARIANZA	**23,06666667**		23,066667	
DEVIAZIONE STANDARD	**4,802776974**		4,802777	
COEFFICIENTE DI VARIAZIONE	**0,111175393**		0,1111754	
ERRORE STANDARD	**1,518771433**		1,5187714	

gl=df	numero osservazioni - 1	(n-1)
x	scarto dalla media (aritmetico)	$\mathbf{X-\bar{X}}$
$\mathbf{\bar{X}}$	Media	**SX/n**
SS =	Somma quadrati scarti	$\mathbf{sx^2}$
s^2 = **Var** = **MS** = Varianza		$\mathbf{sx^2/(n-1)}$
s = **DS** = **SD** = Deviazione standard		$\mathbf{\sqrt{s^2}}$
cv	Coefficiente di variazione	$\mathbf{s/\bar{X}}$
sm =**es**=**SE** = Errore standard della media		$\mathbf{s/\sqrt{n}}$

Un parametro statistico è tanto più efficace:
1 quanto più riassume i dati iniziali con la minore perdita di informazioni
2 - quanto meglio si presta a calcoli ed a test successivi

Per
Descrivere completamente una serie di misure
È necessario riportare sempre

1- Un parametro di **tendenza centrale**

+

2- Un parametro di **dispersione**

+

3- La **quantità delle misurazioni** fatte

esempio

1° Media aritmetica

+

2° Deviazione Standard

+

3° Gradi di libertà

Per descrivere una serie di misure compiutamente è necessario riportare SEMPRE in una tabella:

1° **Numero delle osservazioni = n**

oppure gl =df

2° **Un Parametro di dispersione megio d.s. Oppure es. o var**

3° **Un Parametro di tendenza centrale es. $\bar{X}$**

La nostra serie di misure viene descritta da:

n = 10

$\bar{x}$ = media (aritmetica) = 43,2

VARIANZA o MS= 23,066