Scarto quadratico medio

Lo scarto quadratico medio (o deviazione standard) campionaria è : $\sigma(x) = \sqrt{var(x)}$

Lo scarto quadratico medio (o deviazione standard) empirica è : $\sigma_e(x) = \sqrt{var_e(x)}$

Si introduce in quanto in molte situazioni sarà più utile parlare della radice quadrata della varianza

Tornando agli Esempi: :

Esempi :

Esempio 1 :

$x = (1, 1, 2, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5, 5), \quad n = 14, \quad \bar{x} = 3, \quad var(x) = \frac{22}{13}, \quad var_e(X) = \frac{22}{14}$

$\sigma(x) = \sqrt{\frac{22}{13}} = 1{,}3$

$\sigma_e(x) = \sqrt{\frac{22}{14}} = 1{,}25$

Esempio 2 :

$x = (1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 5, 5, 6, 6), \quad n = 14, \quad \bar{x} = 3, \quad var(x) = \frac{42}{13}, \quad var_e(x) = 3$

$\sigma(x) = \sqrt{\frac{42}{13}} = 1{,}79$

$\sigma_e(x) = \sqrt{3} = 1{,}73$

Esempio 3 :

$x = (1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5), \quad n = 14, \quad \bar{x} = 3, \quad \mathrm{var(x)} = \frac{56}{13}, \quad \mathrm{var_e(x)} = 4$

$\sigma(x) = \sqrt{\frac{56}{13}} = 2{,}07$

$\sigma_e(x) = \sqrt{4} = 2$

PROPOSIZIONE (DISUGUAGLIANZA DI CHEBYSHEV)

Preso un vettore $x = (x_1, \dots, x_n)$, per ogni $d > 0$ si ha che la cardinalità dei dati per cui la distanza dalla media del dato è maggiore di d diviso la numerosità dei dati è minore rispetto alla varianza fratto d al quadrato

$$\frac{\#\{x_i : |x_i - \bar{x}| > d\}}{n} \leq \frac{var_e(x)}{d^2}\left(\leq \frac{var(x)}{d^2}\,\frac{n-1}{n} \quad [per\ la\ varianza\ campionaria]\right)$$

Questa definizione mette in maniera più quantitativa ciò che significa allontanarsi dalla media.

La parte a sinistra conta quanti dati distano di distanza più di d dalla media su tutti i dati.

DIMOSTRAZIONE :

Se vado a contare

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

Questa somma è di certo maggiore uguale di

$$\sum_{\substack{i=1 \\ |x_i - \bar{x}| > d}}^{n} (x_i - \bar{x})^2$$

Quindi dei dati che stanno a distanza maggiore di d

E questo termine è di certo più grande di

$$\sum_{\substack{i=1 \\ |x_i - \bar{x}| > d}}^{n} d^2$$

Quindi:

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \geq \sum_{\substack{i=1 \\ |x_i - \bar{x}| > d}}^{n} (x_i - \bar{x})^2 \geq \sum_{\substack{i=1 \\ |x_i - \bar{x}| > d}}^{n} d^2 = d^2 * \#\{x_i : |x_i - \bar{x}| > d\}$$

La cosa positiva di questa disuguaglianza è che è sempre applicabile in quanto non ci sono altre ipotesi che l'avere dei dati , ma non è un buon risultato per i dati con distribuzione normale.

COROLLARIO CON LO SCARTO QUANDRATICO MEDIO

$$\frac{\#\{x_i : |x_i - \bar{x}| \leq k\sigma(x)\}}{n} \geq 1 - \frac{1}{k^2} \quad \forall\ k \geq 1$$

DIMOSTRAZIONE:

$$\#\{x_i : |x_i - \bar{x}| \leq k\sigma(x)\} = n - \#\{x_i : |x_i - \bar{x}| > k\sigma(x)\}$$

Questa quantità se applico la disuguaglianza

$$\geq n - n\left(\frac{var_e(x)}{k^2\sigma^2(x)}\right)$$

Lo devo mettere in forma di varianza campionaria

$$= n - n\left(\frac{n-1}{n}\frac{var(x)}{k^2\sigma^2(x)}\right)$$

E ora si utilizza il fatto che la varianza campionaria è uguale al quadrato dello scarto quadratico medio campionario

$$= n - (n-1)\left(\frac{\sigma^2(x)}{k^2\sigma^2(x)}\right) = n - \frac{n-1}{k^2}$$

Raccolgo n

$$= n\left(1 - \frac{n-1}{n} * \frac{1}{k^2}\right)$$

E questo termine è maggiore uguale di

$$\geq n\left(1 - \frac{1}{k^2}\right)$$

OSSERVAZIONE

Per la distribuzione normale la percentuale di dati :

- Un intervallo $[\bar{x} - \sigma(x), \bar{x} + \sigma(x)]$ è $di\ circa$ 68%
- Due intervalli $[\bar{x} - \sigma(x), \bar{x} + \sigma(x)]$ è $di\ circa$ 95%
- Tre intervalli $[\bar{x} - \sigma(x), \bar{x} + \sigma(x)]$ è $di\ circa$ 99.7%

(è tipo la probabilità di pescare a caso un dato che è lì dentro)

OSSERVAZIONE

$y_i = bx + a$

$var(y) = b^2 var(x)$
$\sigma(y) = b\sigma(x)$