APPROFONDIMENTI – SCHEDA 3

1. La media e le medie

Nelle schede è presentata solo la media aritmetica ed è chiamata semplicemente media, in accordo con la letteratura statistica maggiormente diffusa in ambito scientifico. Anche i software statistici non forniscono direttamente altre medie, quali ad esempio la geometrica e l'armonica.
Questa scelta è motivata dal fatto che gli ambiti di applicazione delle "altre" medie sono estremamente limitati a situazioni particolari. Alleghiamo comunque alcune pagine di Domingo Paola in cui è presentato un percorso didattico per la riflessione sull'uso scorretto della media (aritmetica) in alcuni contesti e sulla necessità di introdurre in queste situazioni altri tipi di medie.
Sottolineamo che – in ogni caso – data una variabile ha senso calcolare di essa una sola delle medie, quella che si applica correttamente al contesto.

2. Alcune dimostrazioni

a) La media minimizza la somma dei quadrati degli scarti: $\sum_{i=1}^{n} (x_i - \bar{x})^2 \le \sum_{i=1}^{n} (x_i - a)^2$ per ogni a reale.

Infatti:

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

Osserviamo che:

$\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \le \sum_{i=1}^{n} x_i^2 - 2na\bar{x} + na^2$ ovvero $0 \le \bar{x}^2 - 2a\bar{x} + a^2$

In quanto il secondo membro della disguaglianza è un quadrato.
Dal punto di vista didattico è importante soffermarsi – dove non già fatto precedentemente – su un punto su cui solitamente i ragazzi incontrano difficoltà: $\sum_{i=1}^{n} k = nk$

b) La mediana minimizza la somma degli scarti assoluti: $\sum_{i=1}^{n} |x_i - Q_2| \le \sum_{i=1}^{n} |x_i - a|$ per ogni a reale.

$$\sum_{i=1}^{n} |x_i - Q_2| = \sum_{x_i < Q_2} Q_2 - x_i + \sum_{x_i \ge Q_2} x_i - Q_2$$

Consideriamo il caso $a < Q_2$. Indichiamo con k il numero di dati minori di a e con h il numero di dati maggiori o uguali ad a e minori di Q2. Si ha: h+k=n/2 se n è pari e h+k=(n-1)/2 se n è dispari.

$$\text{a)}\ \sum_{i=1}^{n} |x_i - Q_2| = \sum_{x_i < Q_2} Q_2 - x_i + \sum_{x_i \ge Q_2} x_i - Q_2 = \sum_{x_i < a} Q_2 - x_i + \sum_{a \le x_i < Q_2} Q_2 - x_i + \sum_{x_i \ge Q_2} x_i - Q_2$$

$$= \sum_{x_i < a} (Q_2 - x_i + a - a) + \sum_{a \le x_i < Q_2} (Q_2 - x_i + a - a) + \sum_{x_i \ge Q_2} (x_i - Q_2 + a - a) =$$

$$= \sum_{x_i < a} (a - x_i) + k(Q_2 - a) + \sum_{a \le x_i < Q_2} (a - x_i) + h(Q_2 - a) + \sum_{x_i \ge Q_2} (x_i - a) + (n - h - k)(a - Q_2) =$$

$$= \sum_{x_i < a} (a - x_i) + \sum_{a \le x_i < Q_2} (a - x_i) + \sum_{x_i \ge Q_2} (x_i - a) + (2h + 2k - n)(Q_2 - a) =$$

b) $\sum_{i=1}^{n} |x_i - a| = \sum_{x_i < a} (a - x_i) + \sum_{x_i \geq a} (x_i - a) = \sum_{x_i < a} (a - x_i) + \sum_{a \leq x_i < Q_2} (x_i - a) + \sum_{x_i \geq Q_2} (x_i - a)$

Confrontando i risultati di a) e b) si ha:

$$\sum_{x_i < a} (a - x_i) + \sum_{a \leq x_i < Q_2} (a - x_i) + \sum_{x_i \geq Q_2} (x_i - a) + (2h + 2k - n)(Q_2 - a) \leq$$

$$\leq \sum_{x_i < a} (a - x_i) + \sum_{a \leq x_i < Q_2} (x_i - a) + \sum_{x_i \geq Q_2} (x_i - a)$$

Infatti la disuguaglianza precedente corrisponde a:

$$(2h + 2k - n)(Q_2 - a) \leq 2 \sum_{a \leq x_i < Q_2} (x_i - a)$$

dove: $x_i - a \geq 0$ per $a \leq x_i < Q_2$

$Q_2 - a > 0$

$2h + 2k - n \leq 0$ si ha l'uguaglianza per n pari e il valore -1 per n dispari

quindi il primo membro è negativo o nullo e il secondo positivo.

La dimostrazione per il caso $a > Q_2$ è del tutto analoga.

c) La somma degli scarti dalla media è 0: $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$

Infatti: $\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = \sum_{i=1}^{n} x_i - n\,\bar{x} = 0$

Da questa proprietà segue che la somma degli scarti dalla media positivi è uguale alla somma degli scarti dalla media negativi.

d) La standard deviation è minore o uguale della metà dell'ampiezza dell'intervallo di variazione: $\sigma \leq \frac{R}{2}$

Questa disuguaglianza è molto importante per valutare "a occhio" lo scarto quadratico medio (e la correttezza dei calcoli).

Indichiamo con Y la variabile aleatoria ottenuta traslando X del centro dell'intervallo di variazione: $Y = X - \frac{x_{(1)} + x_{(n)}}{2}$. Si ha: $\sigma_Y^2 = \sigma_X^2$. I valori di Y stanno fra $-\frac{R}{2}$ e $\frac{R}{2}$; infatti:

$$y_{(1)} = x_{(1)} - \frac{x_{(1)} + x_{(n)}}{2} = \frac{x_{(1)} - x_{(n)}}{2} = -\frac{R}{2} \quad \text{e} \quad y_{(n)} = x_{(n)} - \frac{x_{(1)} + x_{(n)}}{2} = \frac{x_{(n)} - x_{(1)}}{2} = \frac{R}{2}$$

Da cui: $y_i^2 \leq \frac{R^2}{4}$. Quindi: $\sigma_Y^2 = \frac{\sum y_i^2}{n} - \bar{y}^2 \leq \frac{R^2}{4} - \bar{y}^2 \leq \frac{R^2}{4}$

e) Varianza totale e varianze nelle sottopopolazioni: $\sigma_{tot}^2 = \left(f_A\,\sigma_A^2 + f_B\,\sigma_B^2\right) + \left(f_A(\bar{x}_A - \bar{x})^2 + f_B(\bar{x}_B - \bar{x})^2\right)$

Si ha:

$$\sigma_{tot}^2 = \frac{1}{n}\sum_{i=1}^{n} (\bar{x}_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{x_i \in A} (\bar{x}_i - \bar{x})^2 + \sum_{x_i \in B} (\bar{x}_i - \bar{x})^2\right) = \frac{1}{n}\left(\sum_{x_i \in A} (\bar{x}_i - x_A + x_A - \bar{x})^2 + \sum_{x_i \in B} (\bar{x}_i - x_B + x_B - \bar{x})^2\right)$$

Sviluppiamo il primo addendo:

$$\sum_{x_i \in A} (\bar{x}_i - x_A + x_A - \bar{x})^2 = \sum_{x_i \in A} (\bar{x}_i - x_A)^2 + 2\sum_{x_i \in A} (\bar{x}_i - x_A)(x_A - \bar{x}) + \sum_{x_i \in A} (x_A - \bar{x})^2 =$$

$$= \sum_{x_i \in A} (\bar{x}_i - x_A)^2 + 2(x_A - \bar{x})\sum_{x_i \in A} (\bar{x}_i - x_A) + \sum_{x_i \in A} (x_A - \bar{x})^2 = \sum_{x_i \in A} (\bar{x}_i - x_A)^2 + \sum_{x_i \in A} (x_A - \bar{x})^2$$

essendo $\sum_{x_i \in A} (\bar{x}_i - x_A) = 0$.

Quindi:

$$\frac{1}{n}\sum_{x_i \in A}(\bar{x}_i - x_A + x_A - \bar{x})^2 = \frac{n_A}{n}\left(\frac{1}{n_A}\sum_{x_i \in A}(\bar{x}_i - x_A)^2 + \frac{1}{n_A} n_A (x_A - \bar{x})^2\right) = f_A \sigma_A^2 + f_A (x_A - \bar{x})^2$$

3. Perché "n-1"

La varianza e la covarianza sono spesso calcolate usando n-1 al denominatore. Questo deriva da alcuni risultati di statistica inferenziale.

Indichiamo con $X_1, \ldots, X_n$ le variabili aleatorie campionarie di una variabile aleatoria X di cui si osservano le realizzazioni $x_1, \ldots, x_n$. Tali variabili aleatorie sono assunte indipendenti e hanno la stessa distribuzione di X e in particolare la stessa media e la stessa varianza che indichiamo con μ e σ^2.

Indichiamo con $\bar{X}$ la variabile aleatoria media campionaria: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

Il suo valore atteso è: $I\!E(\bar{X}) = I\!E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n} I\!E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} I\!E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$

Si dice che $\bar{X}$ è uno stimatore non distorto di μ, perché la media dei valori campionari $\bar{x}$ calcolati su tutti i possibili campioni il parametro che si vuole stimare.

In modo analogo al caso del valore atteso, si dimostra che la varianza di $\bar{X}$ è $\frac{\sigma^2}{n}$.

Il valore atteso dello stimatore della varianza $\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$ non è σ^2; infatti, se indichiamo tale stimatore con Σ, si ha:

$$I\!E(\Sigma) = \frac{1}{n} I\!E\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right) = I\!E\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - I\!E(\bar{X}^2) = \left(\frac{1}{n}\sum_{i=1}^{n} I\!E(X_i^2)\right) - I\!E(\bar{X}^2) = I\!E(X_i^2) - I\!E(\bar{X}^2)$$

La varianza di una variabile aleatoria si può scrivere come: $\sigma^2 = I\!E(X^2) - I\!E(\bar{X})^2$. Quindi:

$I\!E(X_i^2) - I\!E(\bar{X}_i)^2 = \sigma^2 - \mu^2$ e $I\!E(\bar{X}^2) = \frac{\sigma^2}{n} - \mu^2$ Per cui: $I\!E(\Sigma) = \sigma^2 - \mu^2 - \frac{\sigma^2}{n} + \mu^2 = \frac{n-1}{n}\sigma^2$.

Lo stimatore Σ è quindi uno stimatore distorto (anche se per n grande la sua distorsione è minima).
Dal risultato precedente si può dedurre che uno stimatore non distorto di σ^2 è S^2, definito come:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$