Laurea in Scienze e Tecnologie Biomolecolari, anno accademico 2014/15
Corso di Matematica e Statistica I

Lezione 14. 27 Ottobre 2014

1 ora *(I dati e la loro descrizione.)*

Un'occhiata ai dati

I dati che provengono dall'esecuzione di un esperimento o dalla misura di una grandezza, o dall'osservazione di un fenomeno, costituiscono il legame tra la nostra rappresentazione della realtà fenomenica e la realtà fenomenica stessa.

Alcuni semplici esempi servono a precisare i differenti contesti in cui abbiamo a che fare con i dati

- L'**altezza degli individui** di una popolazione è una grandezza che può sollecitare il nostro interesse in relazione a varie domande e problemi. La prima cosa da fare è osservarla misurando l'altezza degli individui di un campione significativo della popolazione se non addirittura di tutti gli individui della popolazione stessa. Fissata una unità di misura (ad esempio il centimetro), quello che otteniamo è un insieme di numeri che per comodità metteremo in ordine crescente
$$h_1 \leq h_2 \leq h_3, \quad \dots \quad h_n.$$
notando che alcuni di questi numeri possono essere uguali e che si ripetono nella sequenza che abbiamo indicato. Qui stiamo rilevando i dati di una grandezza che sappiamo variare da individuo a individuo.

- La **carica dell'elettrone** è una grandezza che pensiamo corrisponda ad un valore universale uguale per ogni elettrone. Tuttavia per determinarne l'entità compiamo una serie di misure che ci forniscono risultati differenti perché come succede per ogni operazione di misura intervengono errori che possiamo controllare ed evitare solo parzialmente. Anche in questo caso otteniamo una serie di numeri
$$e_1 \leq e_2 \leq e_3, \quad \dots \quad e_n \ ,$$
in genere diversi.

- Volendo studiare come una certa **sostanza viene assorbita dall'organismo**, possiamo iniettare la sostanza stessa nel sangue di un individuo e, a tempi diversi, misurarne la concentrazione. Il risultato dell'operazione sarà costituito da una sequenza di coppie di numeri che rappresentano rispettivamente il tempo al quale viene misurata la concentrazione e la concentrazione corrispondente.
$$(t_1, c_1), \quad (t_2, c_2) \quad (t_3, c_3) \quad \dots \quad (t_n, c_n) \ ,$$
Immaginiamo che l'assorbimento della sostanza segua una determinata legge (che vogliamo appunto identificare) per cui i dati indicati sopra dovrebbero riferirsi al grafico della funzione che rappresenta la legge.

I dati che possiamo raccogliere in diverse situazioni, diversi contesti e diversi scopi parlano da soli, ma per descriverli in modo significativo occorre adottare criteri sintetici che ne riassumano la struttura.

La media e la varianza

Se il valore medio tra due numeri $\bar{x} = \dfrac{h_1 + h_2}{2}$ è equidistante dai numeri stessi, così , per n numeri $\{h_1, h_2, h_3, \quad \dots \quad h_n\}$ la **media**

$$M[h] = \frac{1}{n} \sum_{i=1}^{i=n} h_i \tag{1}$$

rappresenta un buon compromesso. Se infatti consideriamo la funzione

$$E(x) = \sum_{i=1}^{i=n} (x - h_i)^2, \tag{2}$$

che fornisce la somma delle distanze (quadratiche) del valore x dai valori dei punti in questione, abbiamo

$$E(x) = \sum_{i=1}^{i=n} \left(x^2 + h_i^2 - 2h_i x\right) = nx^2 - 2\left(\sum_{i=1}^{i=n} h_i\right) x + \sum_{i=1}^{i=n} h_i^2$$

e il **minimo** viene assunto proprio per

$$x_m = \frac{1}{n}\sum_{i=1}^{i=n} h_i = M[h]$$

Così la media (1) può essere adottata come rappresentativa dei numeri nel loro complesso, specialmente se unita al valore (minimo) che la funzione $E(x)$ assume in corrispondenza al valore della media. Si chiama infatti **Varianza** o **scarto quadratico medio** la quantità

$$V[h] = \frac{E(M[h])}{n} = \frac{1}{n}\sum_{i=1}^{i=n} (h_i - M[h])^2 \tag{3}$$

che misura la **dispersione dei dati intorno alla loro media**. Osserviamo che, sviluppando il quadrato in (3), si ha

$$V[h] = \frac{1}{n}\sum_{i=1}^{n} h_i^2 - 2\frac{1}{n}\sum_{i=1}^{n} h_i \; M[h] - M^2[h] = M\left[h^2\right] - M^2[h]$$

dove con h^2 indichiamo la sequenza $\{h_1^2,\ h_2^2,\ \ldots\ h_n^2\}$, dei quadrati degli h_i.

Mediana

Un altro modo, per rappresentare i dati utilizza la cosiddetta **mediana** e i **percentili**. In particolare la **mediana** è il valore che divide i dati in due blocchi equi-numerosi. Dunque è definita come

$$\textbf{mediana}[h] = \begin{cases} h_{(n+1)/2} & \text{se } n \text{ è dispari} \\ \dfrac{h_{n/2} + h_{n/2+1}}{2} & \text{se } n \text{ è pari} \end{cases}$$

Per esempio, per i valori $\{1, 2.3, 5, 7.2, 8\}$ la mediana è 5, mentre nel caso $\{1, 5, 7.2, 8\}$ la mediana è 6.1. Osserviamo pure che nel caso in cui nella sequenza ci siano valori ripetuti, questi valori vanno considerati come distinti. Come esempio abbiamo che per i dati $\{1, 1, 1, 2.3, 5, 7.2, 8\}$ la mediana è 2.3. L'operazione di suddivisione dell'insieme dei dati in due può essere ripetuta individuando i **quartili** che nel caso dell'esempio precedente sono 1 e 7.2. Quando dati sono numerosi si può andare avanti fino a ripetere l'operazione cento volte ed ottenere i **percentili**.

Così come la media minimizza l'errore quadratico (2) la mediana minimizza l'errore misurato come

$$E(x) = \sum_{i=1}^{i=n} |x - h_i|, \tag{4}$$

In questo caso abbiamo a che fare con una funzione continua, derivabile *a tratti* che, in ogni intervallo tra due dati $[h_m, h_{m+1}$, si rappresenta come

$$E(x) = (2m - n)x + \sum_{i=m+1}^{i=n} h_i - \sum_{i=1}^{i=m} h_i,$$

e dunque è decrescente quando $2m - n < 0$, crescente quando $2m - n > 0$, costante se $2m - n = 0$. In conclusione vediamo che se n è dispari la funzione $E(x)$ è

- decrescente in ogni intevallo $[h_m, h_{m+1}]$ con $m < \dfrac{n+1}{2}$
- crescente in ogni intevallo $[h_m, h_{m+1}]$ con $m \geq \dfrac{n+1}{2}$

e assume il suo mimimo in $x_m = h_{(n+1)/2}$. Se poi n è pari $E(x)$ è

- decrescente in ogni intevallo $[h_m, h_{m+1}]$ con $m < \frac{n}{2}$
- costante nell'intervallo $[h_{n/2}, h_{n/2+1}]$
- crescente in ogni intevallo $[h_m, h_{m+1}]$ con $m \geq \frac{n}{2} + 1$

e assume il minimo in $x_m = \frac{h_{n/2} + h_{n/2+1}}{2}$.

ESERCIZI

ESERCIZIO 14.1 La media di tre numeri reali può essere maggiore del massimo dei tre numeri? Se sì , fai un esempio. Se no, perché?

ESERCIZIO 14.2 La mediana di quattro numeri reali può essere maggiore del massimo dei quattro numeri?

ESERCIZIO 14.3 Può esistere un insieme di dati con media 0, moda 1 e mediana -1?
Se sì , fai un esempio. Se no, perché?

ESERCIZIO 14.4

(a) Trova un esempio di 10 dati numerici distinti con media uguale alla mediana.

(b) Trova un esempio di 10 dati numerici distinti con media maggiore della mediana.

(c) Trova un esempio di 10 dati numerici distinti con media minore della mediana.

ESERCIZIO 14.5 In una stazione metereologica viene rilevata ogni giorno la temperatura massima. La media delle temperature massime dei primi 6 giorni della settimana è stata di 20° C. Se il settimo giorno la temperatura massima è stata di 18° C, qual è la media delle temperature massime nell'intera settimana?

ESERCIZIO 14.6 Un esperimento ti ha fornito 3 dati x_1, x_2, x_3, tutti compresi fra 12 e 13. La loro deviazione standard può essere uguale a 7?
Se sì , fai un esempio. Se no, perché?

ESERCIZIO 14.7

(a) Trova 10 dati numerici distinti con media 10 e calcolane la deviazione standard.

(b) Trova altri 10 dati numerici distinti con media 10 ma con deviazione standard doppia rispetto ai dati della parte (a).

Questi e altri esercizi, insieme ad appunti, avvisi e istruzioni, si trovano sul sito del corso all'indirizzo
http://www.science.unitn.it/~ iannelli/_corsi/elenco_corsi_2014-2015.html