MEDIA, VARIANZA, MEDIANA

1. La media

1.1. Definizione e proprietà. Siano dati tre numeri reali a, b e c. Si dice media di questi tre numeri il numero seguente
$$Av = \frac{a+b+c}{3}.$$
Supponiamo che $a < b < c$.
Si verifica facilmente che
$$a < Av < c.$$
Adesso rimuoviamo l'ipotesi che $a < b < c$, ma supponiamo che a, b e c siano distinti. In questo caso segue
$$\min\{a, b, c\} < Av < \max\{a, b, c\}.$$
Infine siano a, b e c tre numeri reali qualsiasi.
Verificare che si ha
$$\min\{a, b, c\} \leq Av \leq \max\{a, b, c\}. \tag{1.1}$$
È facile dimostrare che in (1.1), vale l'eguaglianza se e solo se $a = b = c = Av$.

Sia ora n un intero positivo. Supponiamo che siano assegnati n numeri reali $a_1, a_2 \ldots, a_n$. Possiamo scrivere in modo più compatto l'n-pla precedente con il simbolo a. La media dell'n-pla è definita da
$$Av(a) = \frac{1}{n}\sum_{i=1}^{n} a_i$$
Sia $m = \min_i\{a_i\}$ e $M = \max_i\{a_i\}$. Allora si ha
$$m \leq Av \leq M \tag{1.2}$$
che, per ragioni che saranno spiegate nel seguito, chiameremo teorema della media per un numero finito di elementi. Inoltre in (1.2) valgono entrambe le uguaglianze se e solo se $a_i = Av$, per ogni $i = 1, \ldots n$. Questo risultato si dimostra facilmente, infatti risulta evidente che
$$m \leq a_i \leq M, \text{ per ogni } i = 1, \ldots, n. \tag{1.3}$$
Sommando le precedenti disuguaglianze otteniamo che
$$nm \leq \sum_{i=1}^{n} a_i \leq nM$$
e quindi dividendo per n otteniamo proprio la (1.2).

Supponiamo ora che in una delle disuguaglianze (1.3) valga un segno stretto, allora sommando le disuguaglianze tale segno stretto varrebbe anche nella (1.2). Questo significa che se vogliamo che valga $m = Av(a) = M$, dovrà necessariamente valere $m = M = a_i$ per ogni $i = 1\ldots,n$ e quindi $Av(a) = m = M = a_i$ per ogni $i = 1,\ldots n$.

Per apprezzare il significato geometrico convinciamoci che la media rappresenta una "altezza media". A tale proposito assumiamo che a_i siano tutti numeri positivi e disegnamo un grafico dove alle prime n ogni unità sull'asse delle x associamo i valori a_i, $i = 1\ldots,n$, (vedi Figura 1). Indichiamo con $\mathcal{A}$ la figura data dall'unione dei vari rettangoli. Poiché la base di ogni rettangolo misura 1 la somma delle aree dei vari rettangoli, quindi di $\mathcal{A}$ corrisponde a $nAv(a) = \sum_{i=1}^{n} a_i$ (area in grigio nella figura). Se confrontiamo la figura $\mathcal{A}$ con il rettangolo che ha come base la stessa base e come altezza il minimo m dei dati a_i osserviamo subito che $\mathcal{A}$ contiene tale rettangolo. Se invece consideriamo il rettangolo che ha sempre la stessa base e come altezza il massimo M dei dati a_i allora $\mathcal{A}$ sarà contenuta in tale rettangolo (vedi Figure 2,3). Evidentemente se vogliamo un rettangolo con quella base e che abbia esattamente l'area uguale ad $\mathcal{A}$ dobbiamo scegliere un'altezza intermedia, più precisamente $Av(a)$ (vedi Figura 4)

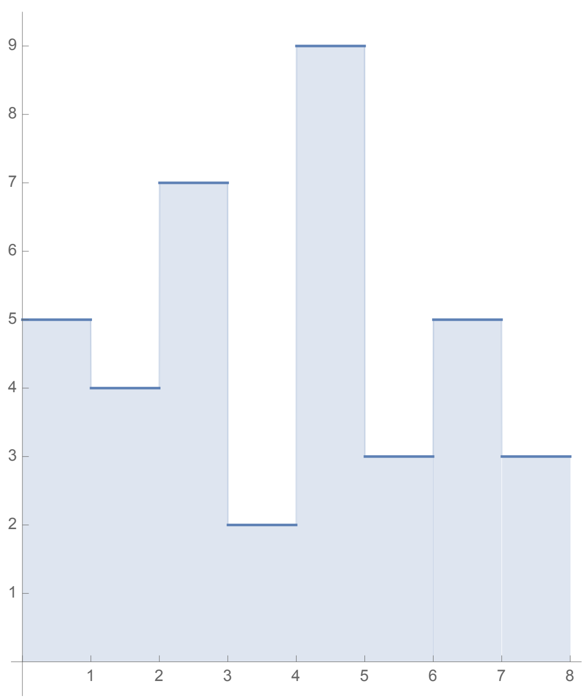


Figura 1. Caso $n = 7$: $a_1 = 5$, $a_2 = 4$, $a_3 = 2$, $a_4 = 9$, $a_5 = 3$, $a_6 = 6$, $a_7 = 3$

Vediamo ora alcune proprietà della media. Una prima osservazione, ovvia, che segue dalla definizione, in particolare dalla proprietà commutativa della somma, è che possiamo riordinare i numeri a_i come vogliamo ma la loro media rimane la stessa. In particolare, dati n numeri possiamo sempre riordinarli in modo che valga $a_1 \leq a_2 \leq \cdots \leq a_i \leq a_{i+1} \leq \cdots \leq a_n$.

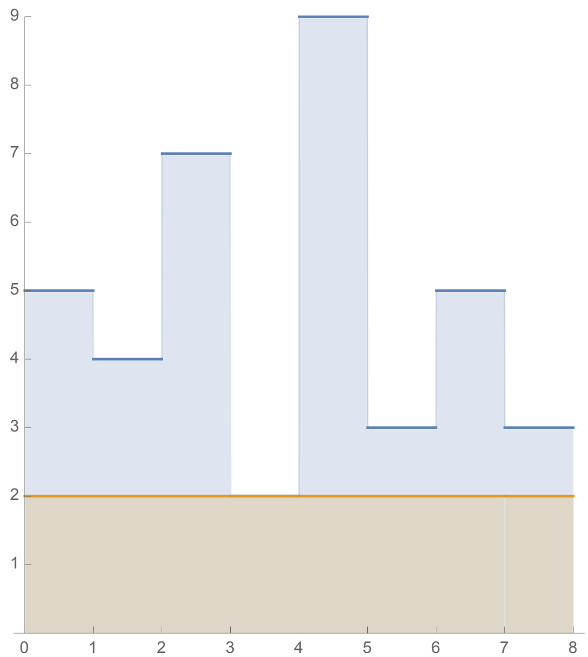


FIGURA 2. Caso $n = 7$: $a_1 = 5$, $a_2 = 4$, $a_3 = 2$, $a_4 = 9$, $a_5 = 3$, $a_6 = 6$, $a_7 = 3$

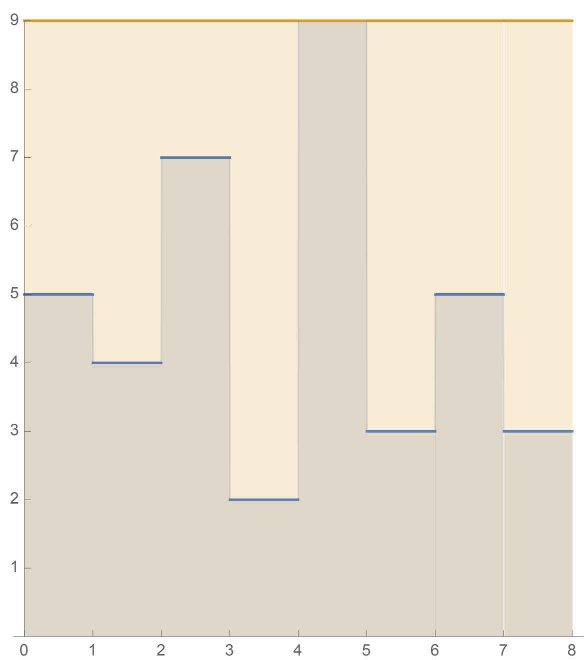


FIGURA 3. Caso $n = 7$: $a_1 = 5$, $a_2 = 4$, $a_3 = 2$, $a_4 = 9$, $a_5 = 3$, $a_6 = 6$, $a_7 = 3$

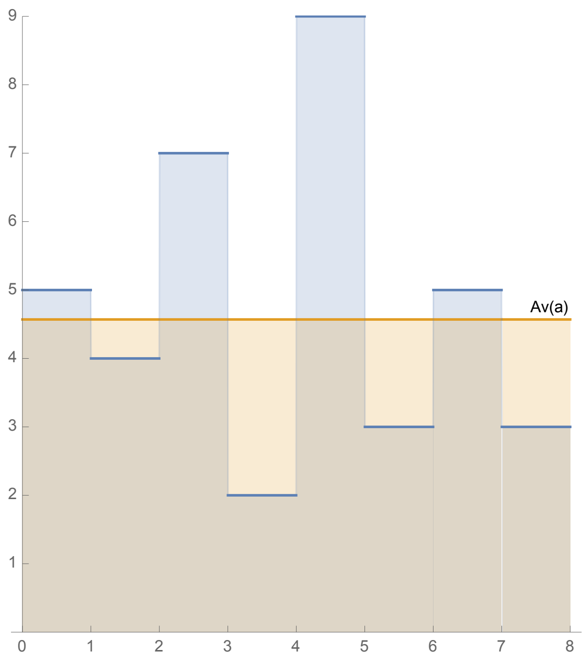


Figura 4. Caso $n = 7$: $a_1 = 5$, $a_2 = 4$, $a_3 = 2$, $a_4 = 9$, $a_5 = 3$, $a_6 = 6$, $a_7 = 3$. Se consideriamo un rettangolo con altezza Av(a) uguale circa a 4.57, tale rettangolo avrà la stessa area di $\mathcal{A}$

Una seconda osservazione riguarda il fatto che se aggiungiamo ad ogni elemento a_i una costante fissata c, ovvero consideriamo la nuova n-pla di numeri $\overline{a}_i = c + a_i, i = 1, \dots n$, che indichiamo brevemente con $\overline{a}$. Si ottiene subito $Av(\overline{a}) = c + A(v)$. Perciò se tutti i dati vengono traslati di uno stesso valore anche la media viene traslata dello stesso valore. Stesso discorso se moltiplichiamo ogni dato per una stessa costante fissata α e quindi consideriamo l'n-pla αa formata dai numeri $\alpha a_1, \dots \alpha a_n$, avremo che $Av(\alpha a) = \alpha Av(a)$. Più in generale si può verificare che la media è un "operatore lineare" ovvero date due costanti qualsiasi α_1 e α_2 e prese due n-ple di dati a e b avremo che $Av(\alpha_1 a + \alpha_2 b) = \alpha_1 Av(a) + \alpha_2 Av(b)$.

1.2. **Varianza discreta.**
Dati i tre numeri a_1, a_2 e a_3 consideriamo la media dei loro quadrati ovvero

$$\frac{a_1^2 + a_2^2 + a_3^2}{3}.$$

Possiamo scrivere $Av(a_1^2, a_2^2, a_3^2)$ oppure, per brevità $A(a^2)$ identificando i tre numeri $\{a_1^2, a_2^2, a_3^2\}$ con il simbolo a^2.

Consideriamo una popolazione di 100 persone con età media 50 anni. Possiamo dire molto poco su quante persone hanno circa 50 anni. Ad esempio potrebbe darsi che esattamente la metà abbia 45 anni e metà 55 anni. Un modo per capire quanto sono distanti le età delle singole persone dalla media è di calcolare la varianza.

Cominciamo a domandarci qual è il caso in cui le persone hanno quasi tutti l'età media vicina a 50 anni. Se tutte le persone avessero 50 anni la media coinciderebbe con l'età di ognuno dei componenti. In qualunque altro caso le cose vanno diversamente. Ad esempio, tornando all'esempio in cui la metà della persone ha 55 anni e l'altra metà 45. Per tutti quelli che hanno 55 anni, lo scarto dalla media è 5 (in eccesso) e per tutti quelli che hanno 45 anni lo scarto è ancora 5 anni, ma questa volte in difetto. Per tenere conto di questi "errori" si possono fare cose molto diverse ma, in ogni caso, sia che sia in difetto che in eccesso, sempre di "errore" si tratta. Per ora valuteremo questo errore per le persona i di età a_i con la quantità $(a_i - Av(a))^2$. L'errore medio quindi sarà dato dalla formula

$$Var(a) = \frac{1}{n}\sum_{i=1}^{n}(a_i - Av(a))^2. \tag{1.4}$$

Nell'esempio precedente ogni termine $(a_i - Av(a))$ vale 5 e quindi il suo quadrato vale 25 e si hanno 100 termini, ognuno che vale 25

$$Var(a) = \frac{1}{100}\sum_{i=1}^{100} 25 = 25. \tag{1.5}$$

In generale data una n-pla di dati a definiamo tramite la (3.2) la sua varianza e la indichiamo con il simbolo $Var(a)$.

Ci si può domandare se la varianza possa essere introdotta attraverso un' ottimizzazione. La risposta è positiva. Per poter verificare questo fatto cominciamo ad osservare alcune proprietà della varianza. Come detto la varianza nasce come strumento per valutare l'errore che si commette prendendo la media di n dati rispetto ai dati stessi. Tra le proprietà sulla media abbiamo osservato che la media di una traslazione dei dati per un vettore costante c corrisponde alla traslazione della media, ovvero $Av(a + c) = Av(a) + c$. Nel caso della varianza cosa succede dopo una traslazione? Intuitivamente non dovrebbe cambiare nulla, ovvero $Var(a + c) = Var(a)$, la varianza non dipende dalla traslazione. Questo si verifica facilmente usando la sua definizione algebrica e la formula di sopra per la media della traslazione. Infatti

$$Var(a+c) = \frac{1}{n}\sum_{i=1}^{n}(a_i+c-Av(a+c))^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i+c-Av(a)-c))^2 = \frac{1}{n}\sum_{i=1}^{n}(a_i-Av(a))^2 = Var(a)$$

Un'altra cosa utile da sapere è che la varianza può essere scritta in un modo equivalente.

Consideriamo prima per semplicità il caso $n = 3$ Dati $\{a_1, a_2, a_3\}$ e posto $Av(a) = \frac{1}{3}(a_1 + a_2 + a_3)$ e $Av(a^2) = \frac{1}{3}(a_1^2 + a_2^2 + a_3^2)$, dimostrate, svolgendo la somma ed i quadrati, che

$$\frac{1}{3}\sum_{i=1}^{3}(a_i - Av(a))^2 = Av(a^2) - (Av(a))^2.$$

La formula è valida in generale:

$$Var(a) = \frac{1}{n}\sum_{i=1}^{n}(a_i - Av(a))^2 = Av(a^2) - (Av(a))^2. \tag{1.6}$$

La formula è banale quando $Av(a) = 0$. Nel caso generale, usando semplicemente la formula del quadrato della differenza $(x-y)^2 = x^2 + y^2 - 2xy$, otteniamo

$$(a_i - Av(a))^2 = a_i^2 + Av(a)^2 - 2a_iAv(a) \quad \text{per ogni } i = 1, \dots n,$$

per cui

$$Var(a) = \frac{1}{n}\sum_{i=1}^{n}(a_i - Av(a))^2 = \frac{1}{n}\sum_{i=1}^{n} a_i^2 + Av(a)^2 - 2a_iAv(a) =$$

$$\frac{1}{n}\sum_{i=1}^{n} a_i^2 + \frac{nAv(a)^2}{n} - 2Av(a)\frac{1}{n}\sum_{i=1}^{n} a_i = Av(a^2) - Av(a)^2.$$

Ci chiediamo ora, cosa possiamo dire della varianza quando moltiplichiamo tutti i dati per una costante fissata λ? Anche in questo caso si tratta solo di fare un semplice calcolo. Ricordando che $Av(\lambda a) = \lambda Av(a)$ otteniamo

$$Var(\lambda a) = \frac{1}{n}\sum_{i=1}^{n}(\lambda a_i - \lambda Av(a))^2 = \lambda^2\frac{1}{n}\sum_{i=1}^{n}(a_i - Av(a))^2 = \lambda^2 Var(a).$$

La varianza si comporta come un operatore quadratico, nel senso che amplifica al quadrato la dilatazione dei dati. Anche in virtù di questa caratteristica è sensato definire lo scarto quadratico medio, o deviazione standard di una sequenza di dati come la radice quadrata della varianza. Più precisamente

$$\sigma(a) = \sqrt{Var(a)}.$$

Ovviamente per quanto detto sopra avremo che $\sigma(\lambda a) = |\lambda|\sigma(a)$ per ogni $\lambda \in \mathbb{R}$.
Per le proprietà della media e della deviazione standard, possiamo sempre ricondurci attraverso appropriate traslazioni e dilatazioni al caso in cui la media è nulla e la deviazione uguale ad 1.

Come vedremo nel prossimo paragrafo, la conoscenza della media e della varianza può essere un dato estremamente indicativo in quanto capita spesso che dati sperimentali seguano un comportamento di tipo "gaussiano". In tal caso i 2 indicatori fondamentali: media e varianza, forniscono in modo sperimentalmente rilevante "la distribuzione dei dati" . Questa avrà la forma di una campana rovesciata centrata rispetto al valore della media e di larghezza legata alla deviazione standard (vedi paragrafo 2). Per cui, da quanto osservato, se sappiamo che i nostri dati seguono una legge gaussiana e conosciamo i valori numerici di una distribuzione gaussiana di media nulla e deviazione 1 possiamo poi riportarli al nostro caso con semplici operazioni algebriche.

Supponiamo ora di voler capire quanto i singoli dati di un insieme $\{a_1, \dots a_n\}$ i sia diverso dalla media. Ad esempio, quando $n = 3$, possiamo considerare una funzione f

definita come $f(x) = \frac{1}{3}((a_1 - x)^2 + (a_2 - x)^2 + (a_3 - x)^2)$ e domandarci quale sia la scelta migliore di x. Migliore rispetto al fatto di rendere f piccola, o almeno più piccola possibile.

Un calcolo elementare mostra che

$$f(x) = x^2 - \frac{2}{3}x(a_1 + a_2 + a_3) + \frac{1}{3}(a_1^2 + a_2^2 + a_3^2).$$

Il grafico di f dato da una parabola e la geometria analitica insegna che il valore minimo di f è assunto quando $x = x_0$ con $x_0 = \frac{1}{3}(a_1 + a_2 + a_3) = Av(a)$. Sostituendo questa espressione di trova che

$$f(x) \geq f(x_0) = Av(a^2) - (Av(a))^2 = Var(a).$$

Quindi abbiamo due notevoli risultati. Il migliore possibile x è la media di a ed il valore più basso possibile è la varianza di a.

Questo risultato di ottimizzazione si può generalizzare ad numero qualsiasi di dati, ovvero presa la funzione

$$F(x) = \frac{1}{n}\sum_{i=1}^{n}(a_i - x)^2$$

questa assumerà il valore minimo proprio nel punto $Av(a)$, tale minimo sarà quindi $F(Av(a)) = Var(a)$

2. La frequenza

Un concetto interessante nell'analisi di un numero finito di dati è la frequenza. Supponiamo di avere una sequenza di dati del tipo

$$3, 4, 7, 4, 4, 6, 8, 9, 8, 5, 4, 3, 5, 5, 4, 8, 3, 9.$$

In questo caso la sequenza di dati non è ordinata, ma questo non ha importanza, già sappiamo che per quanto riguarda la media il loro ordinamento non ha alcun peso. Quello che salta all'occhio è che alcuni numeri si ripetono più volte. Per tener conto di questo aspetto riportiamo su una tabella i numeri diversi nella sequenza di dati, che per semplicità ordiniamo mettendo a fianco il numero di volte in cui si ripetono

Numeri	ripetizioni
3	3
4	5
5	3
6	1
7	1
8	3
9	2

Facciamo delle osservazioni. La prima cosa che salta all'occhio è che il 4 è il numero che capita più di frequente. La seconda è che bisogna fare attenzione a distinguere il numero di dati considerati e i diversi numeri che compaiono nella sequenza di dati. In questo caso i diversi numeri che abbiamo sono $\{3, 4, 5, 6, 7, 8, 9\}$, invece il numero complessivo di

dati considerati si ottiene sommando la riga delle ripetizioni, ovvero in questo caso 18. Utilizziamo il termine frequenza per indicare il numero di volte in cui un determinato elemento si ripete. Ad esempio il numero 4 ha frequenza uguale a 5, il numero 9 ha frequenza uguale a 2 e così via. Quindi la somma delle frequenze coincide con il numero di dati presi in esame. Chiameremo, invece, frequenza relativa di un elemento, il rapporto tra la frequenza di quell'elemento e il numero complessivo di dati. Tornando all'esempio precedente la frequenza relativa di 3 sarà $\frac{3}{18} = \frac{1}{6}$, di 4 sarà $\frac{5}{18}$.....
In generale, data una sequenza di dati, il cui numero totale indichiamo con N, consideriamo gli elementi diversi della sequenza, il cui numero totale indichiamo con K, ovviamente $K \leq N$. Ordiniamo questi K valori distinti, se indichiamo con l'indice i l'isimo elemento, questo avrà una frequenza associata che possiamo chiamare F_i, corrispondente al numero di volte in cui tale elemento si ripete, denotiamo, invece con f_i la frequenza relativa, data appunto dal rapporto tra F_i e il numero complessivo di dati N. Ci chiediamo che rapporto esiste tra il calcolo della media, la frequenza e la frequenza relativa?

Un modo semplice per calcolare la media è distinguere gli elementi distinti e della successione di dati e moltiplicarli per la loro frequenza. Più precisamente se la sequenza di dati è data da K elementi distinti $\{a_1, \ldots, a_K\}$, avremo

$$Av(a) = \frac{1}{N}\sum_{j=1}^{K} a_i F_i = \sum_{j=1}^{K} a_i \frac{F_i}{N} = \sum_{j=1}^{K} a_i f_i.$$

Per evitare di essere impressionati da tutti questi indici torniamo all'esempio precedente. In quel caso la sequenza di dati è formata da 18 dati. Il numero di elementi distinti è 7, come abbiamo visto tali elementi corrispondono all'insieme $\{3, 4, 5, 6, 7, 8, 9\}$. Aggiorniamo la tabella mettendo le frequenze relative in luogo delle ripetizioni

Numeri	frequenza relativa f_i
3	$\frac{3}{18}$
4	$\frac{5}{18}$
5	$\frac{3}{18}$
6	$\frac{1}{18}$
7	$\frac{1}{18}$
8	$\frac{3}{18}$
9	$\frac{2}{18}$

la media della sequenza dei dati è data da

$$Av(a) = 3\frac{3}{18} + 4\frac{5}{18} + 5\frac{3}{18} + 6\frac{1}{18} + 7\frac{1}{18} + 8\frac{3}{18} + 9\frac{2}{18} = \frac{99}{18} = 5.5.$$

Pensiamo al caso illustrato nel precedente paragrafo di un campione di 100 persone tali che la metà ha 45 anni e l'altra metà 55. La media come abbiamo visto corrisponde a 50. Quanti sono gli elementi differenti che compaiono in questa serie di dati? Sono solamente 2, ovvero 45 e 55. La loro frequenza relativa? Ovviamente risulterà uguale ad $\frac{1}{2}$, infatti per entrambi gli elementi abbiamo 50 ripetizioni su 100 dati.

Sicuramente il concetto di media è legato al fatto che i nostri dati siano dei numeri. D'altra parte come è facile immaginare ha senso raccogliere dei dati che non siano necessariamente valori numerici. Si possono fare diversi esempi. Se chiediamo ad un campione di persone, in che mese sono nati, in quale città sono nati oppure qual è il loro gruppo sanguinio ecc., raccoglieremo dei dati non numerici ma non per questo meno interessanti ai fini statistici. Ovviamente per tali dati non avrà senso parlare di media. Avrà però senso considerare la frequenza in cui i dati si ripetono. In tale contesto i concetti di frequenza e frequenza relativa saranno sicuramente ancora definiti. In tutti questi casi sarà possibile rappresentare graficamente i risultati relativi alla frequenza attraverso dei grafici a barre o tramite istogrammi. In generale nell'asse delle x mettiamo delle classi di oggetti distinti, non necessariamente numeri, mentre nell'asse delle y mettiamo la frequenza o frequenza relativa di quella classe. Quando non pensiamo a volori numerici la loro posizione sull'asse delle x sarà assolutamente arbitraria, sarà comunque sensato utilizzare intervalli contigui di uguale lunghezza.
Ad esempio chiediamo a 100 persone di dirci quale è il gradimento per un determinato film. Decidiamo di fare 4 classi: insoddisfatto, abbastanza soddisfatto, soddisfatto, molto soddisfatto. Dopo aver raccolto i dati otteniamo la seguente tabella

Giudizi	frequenza
insoddisfatto	15
poco soddisfatto	20
soddisfatto	40
molto soddisfatto	25

che possiamo esprimere graficamente attraverso il seguente diagramma a barre. Se avessimo usato invece della frequenza la frequenza relativa poco sarebbe cambiato per la rappresentazione. Sarebbe stato come cambiare l'unità di misura sull'asse delle y.

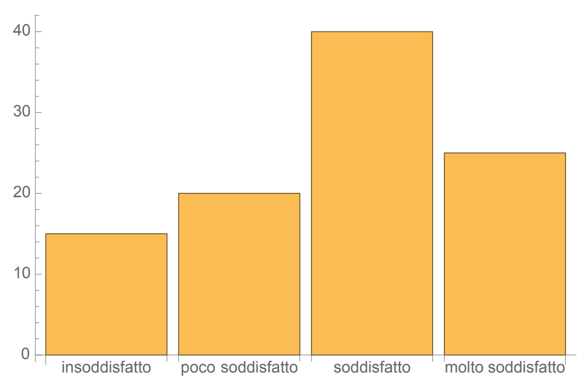


Figura 5

Torniamo ora al caso in cui i dati raccolti siano di carattere numerico. Nuovamente possiamo usare una rappresentazione come quella sopra dove però sull'asse delle x ha senso considerare realmente dei numeri. In tale situazione possiamo associare ad ogni dato la sua frequenza oppure la sua frequenza relativa. Se facessimo il grafico otterremmo un numero finito di punti nel piano. Questo avrebbe sicuramente molto senso quando sappiamo che i dati appartengono ad una classe ben definita, ad esempio i numeri interi o meglio ancora un sottoinsieme fissato di numeri interi. Ad esempio se raccogliamo il dato di un uscita del lotto sappiamo che questo è un numero compreso tra 1 e 90, se raccogliamo il dato di un lancio di un dado questo è un numero compreso tra 1 e 6 ecc. In tutte le situazioni in cui sappiamo che i dati assumono prefissati valori numerici (non necessariamente interi) può essere utile usare la seguente rappresentazione. Supponiamo che ci siano k valori distinti che i dati possono assumere, indichiamo questi con x_i $i = 1\dots,k$ supponendo che siano ordinati in modo che $x_1 < x_2 < \cdots < x_k$. Ora consideriamo degli intervalli chiusi a sinistra e aperti a destra del tipo $[x_i, x_{i+1})$ se $i = 1,\dots k-1$. Consideriamo anche le semirette $(-\infty, x_1)$ e $[x_k, +\infty)$. Possiamo definire la seguente funzione a scalino

$$F(x) = \begin{cases} 0 & \text{se } x < x_1 \\ f_i & \text{se } x_i \le x < x_{i+1}, \quad i = 1\dots k-1 \\ f_n & \text{se } x \ge x_k \end{cases}$$

dove f_i rappresenta la frequenza relativa del valore x_i. Nel caso dell'esempio della sequenza di dati $\{3, 4, 7, 4, 4, 6, 8, 9, 8, 5, 4, 3, 5, 5, 4, 8, 3, 9\}$, gli x_i distinti apparterranno all'insieme $\{3, 4, 5, 6, 7, 8, 9\}$ usando la tabella di sopra per le frequenze avremo il seguente grafico

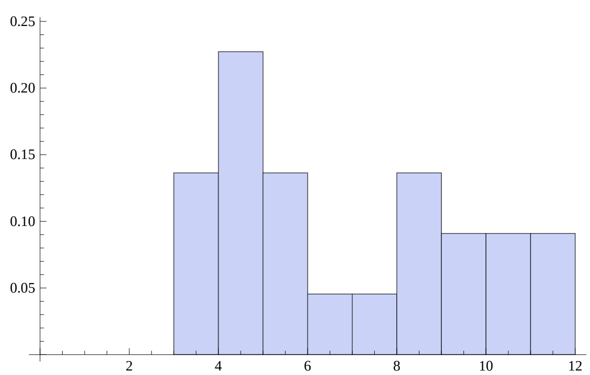


Figura 6

In qualche modo possiamo pensare che il valore della frequenza relativa sia strettamente legato alla probabilità che un determinato valore sia assunto rispetto agli altri. Più un valore è frequente e più appare probabile rispetto agli altri. Questo ovviamente non è completamente vero soprattutto se il numero della sequenza dei dati N è relativamente basso. L'idea di sovrapporre il concetto di frequenza relativa con il concetto di probabilità diventa interessante quando N è molto grande ma soprattutto quando i dati ottenuti sono legati essi stessi ad una legge probabilistica. Pensiamo al lancio di un dado, il numero ottenuto sarà sempre compreso tra 1 e 6. Se lanciamo N volte il dado possiamo costruirci la funzione a gradino. Se N non è molto grande allora potremmo ottenere un grafico casuale.

Se N diventa molto grande allora ci aspettiamo che i valori delle frequenze relative tendano ad essere sempre più vicini all'effettiva probabilità che esca un numero rispetto ad un altro. Sappiamo bene che se il dado non è truccato tale probabilità sarà uguale per tutti e 6 i numeri e quindi uguale ad $\frac{1}{6}$. Quindi per N grande tutti gli f_i dovranno necessariamente avvicinarsi al valore $\frac{1}{6}$.

Un'altra funzione interessante da introdurre in questo contesto è la seguente

$$P(x) = \sum_{i=1}^{j} f_i \quad \text{dove} \quad \text{dove } j \text{ è il più grande indice tale che } x \geq x_j,$$

pensando all'interpretazione probabilistica questa funzione dovrebbe rappresentare la probabilità che un certo dato sia inferiore ad x.

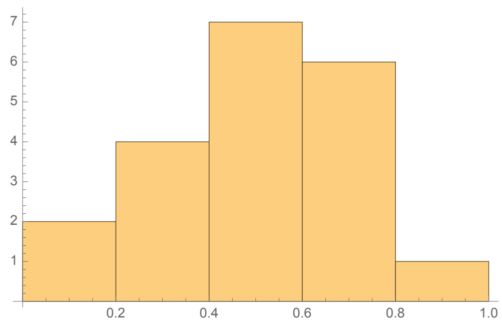
La funzione a scalino introdotta sopra ci è utile per avere una rappresentazione grafica delle frequenze, d'altra parte ha il grosso svantaggio di essere sensata solo quando sappiamo che a priori i dati assumono solo valori predeterminati. Inoltre poiché in generale le ampiezze degli intervalli $[a_j, a_{j+1})$ dipendono da j (nel nostro esempio erano tutte uguali ad 1) l'area del plurirettangolo associato non ha molto a che fare con l'effettiva frequenza ("probabilità") che un insieme di dati sia assunto. A tale proposito rimane invece significativa la funzione P. Supponiamo di voler sapere la frequenza relativa ("probabilità") di un insieme di dati $(x, y]$. Evidentemente nell'intervallo $(x, y]$ cadranno solo un numero finito di valori x_j. Non è difficile convincersi che possiamo scrivere la frequenza relativa di tutti questi valori attraverso la semplice formula $P(y) - P(x)$.

In realtà, spesso, nella raccolta di dati si ha a che fare con dati che provengono da misurazioni e nascono per forza di cose come approssimazioni di quantità reali. Poiché nelle approssimazioni è sottinteso che ci sia un errore, se vogliamo dar ancora senso alla frequenza dobbiamo decidere che se peschiamo 2 dati diversi nello stesso intervallo fissato questi incidano non sulla frequenza del singolo dato quanto piuttosto sulla frequenza di un intervallo di dati. Ovviamente per far questo dobbiamo decidere a priori la lunghezza degli intervalli. A priori si potrebbero prendere intervalli di ampiezza diversa. Per semplicità noi supporremo che gli intervalli abbiano tutti la stessa lunghezza, più precisamente: partendo dall'origine, dividiamo la retta in infiniti intervalli tutti di ampiezza uguale ad $a > 0$. Poiché vogliamo che tutti questi intervalli siano disgiunti e tali che l'unione sia tutto $\mathbb{R}$, li prendiamo per convenzione chiusi a sinistra e aperti a destra. Quindi avremo intervalli del tipo $[0, a)$, $[a, 2a)$, $[-a, 0)$,.... A questo punto se consideriamo una sequenza di dati reali e fissiamo il passo a possiamo disegnare un istogramma di tale sequenza associando ad ogni intervallo la sua frequenza. Costruiamo quindi una funzione a gradino, alle cui basi ci sono gli intervalli fissati e le cui altezze corrispondono alle frequenze (oppure alle frequenze relative) della sequenza di dati.

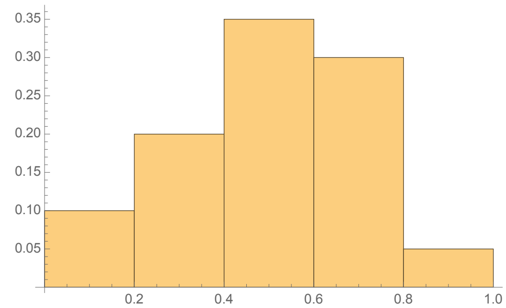
Ad esempio prendiamo la sequenza

$$\{0.3, 0.35, 0.51, 0.13, 0.43, 0.41, 0.4, 0.45, 0.7, 0.78, 0.9, 0.63, 0.68, 0.73, 0.46, 0.23, 0.25, 0.1, 0.54, 0.64\}$$

si tratta di 20 dati. Scegliamo come ampiezza dell'intervallo 0.2, tracciamo l'istogramma associato. Dalla Figura 7 (a) riconosciamo che la frequenza più alta corrisponde ai

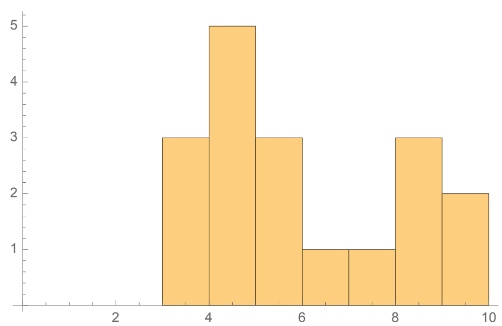


(A) Istogramma passo 0.2 delle frequenze

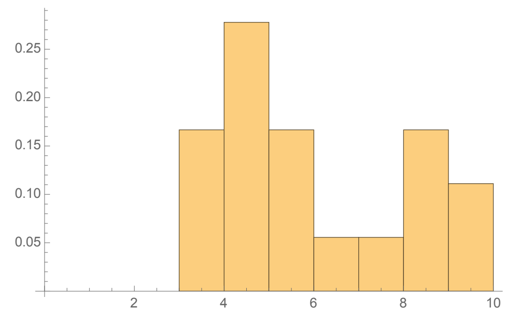


(B) Istogramma passo 0.2 delle frequenze relative

FIGURA 7



(A) Istogramma passo 1 delle frequenze



(B) Istogramma passo 1 delle frequenze relative

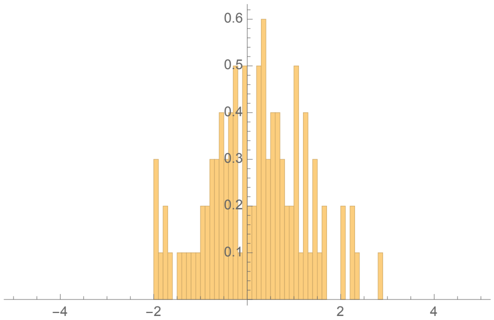
FIGURA 8

dati compresi nell'intervallo $[0.4, 0, 6)$. Possiamo anche tracciare l'istogramma usando le frequenze relative (Figura 7 (b)).

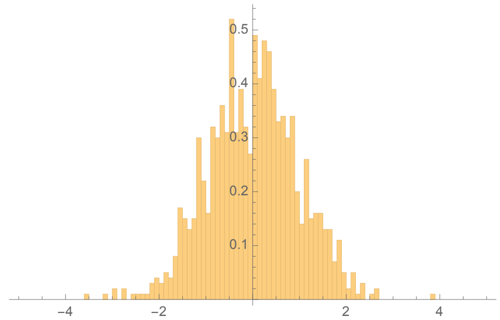
Tornando invece all'esempio fornito all'inizio di questo paragrafo legato alla sequenza di dati $\{3, 4, 7, 4, 4, 6, 8, 9, 8, 5, 4, 3, 5, 5, 4, 8, 3, 9\}$, poiché si tratta di dati interi conviene scegliere il passo $a = 1$ in modo tale che la frequenza dell'intervallo corrisponda alla reale frequenza del numero intero. In questo caso otteniamo l'istogramma delle frequenze assolute e relative vedi Figura 8.

Ovviamente l'interesse per queste cose nasce quando i dati sono in numero elevato. Questi esempi servono solo per dare un'idea di come si tracciano gli istogrammi.

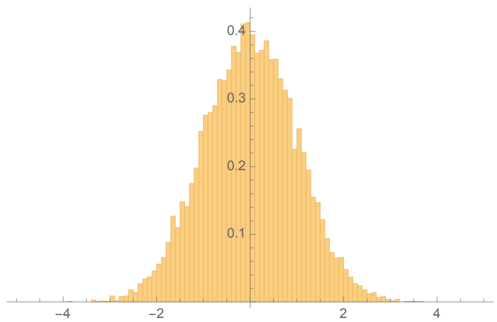
Concludiamo questo paragrafo osservando che il metodo usato nel primo paragrafo per rappresentare i dati, è molto diverso da quello usato in questo paragrafo. Si pensi alla Figura 1. In quel caso il valore del dato era legato alle ordinate e nelle ascisse venivano considerati degli intervalli unitari quello aveva il vantaggio di darci un legame tra la media dei dati e l'area del plurirettangolo (unione di rettangoli) ottenuto nel grafico. In questo caso le ordinate rappresentano le frequenze dei dati mentre nelle ascisse ci sono i dati stessi. Questa rappresentazione, ci dice in qualche modo quali sono gli intervalli di



(A) Istogramma rinormalizzato di 100 dati che seguono una legge gaussiana



(B) Istogramma rinormalizzato di 1000 dati che seguono una legge gaussiana



(C) Istogramma rinormalizzato di 10000 dati che seguono una legge gaussiana

FIGURA 9

dati più probabili rispetto agli altri, ovvero quelli che hanno frequenza relativa maggiore. Potremmo pensare, ad esempio, che il grafico delle frequenze relative ci dica tramite l'ordinata associata ad un determinato intervallo quale sia la probabilità che tale intervallo di valori sia preso. Questo ha senso soprattutto nel caso in cui la sequenza di dati segua una legge probabilistica. Osserviamo che l'area del rettangolo di base $[na, (n+1)a)$ e altezza la frequenza relativa f_n ha area uguale a af_n. In tale prospettiva conviene effettuare un riscalamento del grafico delle frequenze relative dividendo il valore delle ordinate per il passo a. In questo modo se ci chiediamo quale è la probabilità che un dato sia compreso nell'intervallo (x, y) questa sarà in buona approssimazione (a abbastanza grande) data dall'area della regione che si trova sotto il grafico della funzione a gradino limitatamente all'intervallo (x, y). Più in particolare questo tipo di grafico ci dirà in modo empirico la legge di probabilità seguita dai dati, più sarà elevato il numero dei dati e migliore sarà l'approssimazione. La distribuzione di probabilità più nota e che si presenta frequentemente nelle applicazioni è la distribuzione gaussiana. Se consideriamo dei dati casuali che seguano tale distribuzione un suo istogramma rinormalizzato è dato in Figura 9 dove sono considerati rispettivamente 100, 1000 e 10000 dati casuali che seguono una legge gaussiana. Come si può notare all'aumentare dei dati la forma del grafico assomiglia sempre di più ad una campana centrata in 0. Questa è la forma della distribuzione gaussiana di media nulla e varianza 1. La media nulla fa si che la campana sia centrata in 0 e la varianza è legata alla larghezza della campana. I nostri dati casuali seguono esattamente la legge gaussiana di media nulla e varianza uguale ad 1.

Esercizio, esperimento Prendete 2 dadi e lanciateli 30 volte. Ogni volta raccogliete il dato ottenuto sommando le facce esposte dei 2 dadi. I dati saranno degli interi compresi tra 2 e 12. Calcolate la media, quindi disegnate l'istogramma delle frequenze e l'istogramma delle frequenze relative ottenute. Poiché si tratta di interi si suggerisce di prendere come passo per l'istogramma il valore 1.

3. La mediana

Ora introduciamo un altro concetto un pochino diverso dalla media ma che ha anch'esso rilevanza statistica. Supponiamo di avere un insieme ordinato di dati, ad esempio raccogliamo i voti presi in una determinata prova scritta da una classe di studenti. Supponiamo che ci siano 25 studenti ed associamo ad ognuno di questi un voto compreso tra 0 e 10. Possiamo trascrivere su una tabella questi voti, indicando semplicemente con una lettera diversa dell'alfabeto i diversi studenti

$$\begin{bmatrix} A & B & C & D & E & F & G & H & I & J & K & L & M & N & O & P & Q & R & S & T & U & V & W & X & Y \\ 7 & 5 & 4 & 9 & 7 & 6 & 7 & 7 & 9 & 7 & 6 & 5 & 7 & 8 & 6 & 6 & 7 & 5 & 5 & 7 & 7 & 6 & 5 & 4 & 8 \end{bmatrix}$$

Abbiamo 25 dati corrispondenti ai diversi studenti, questi dati assumono un valore intero compreso tra 0 e 10. In realtà i nostri studenti non sono mai andati sotto il 4. Ordiniamo ora i dati, attenzione non ci interessa più associarli alle diverse persone (lettere). La lunga lista sarà:

$$4,4,5,5,5,5,5,6,6,6,6,6,7,7,7,7,7,7,7,7,7,8,8,9,9. \tag{3.1}$$

Scriviamo la tabella delle frequenze

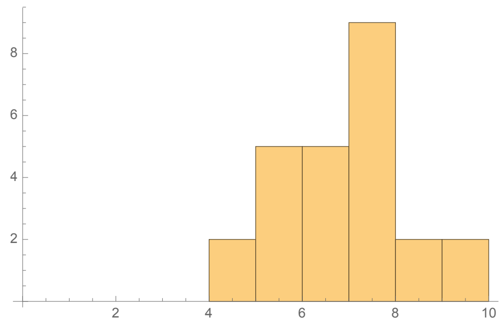
Voti	frequenza	frequenza relativa
4	2	$\frac{2}{25}$
5	5	$\frac{5}{25}$
6	5	$\frac{5}{25}$
7	9	$\frac{9}{25}$
8	2	$\frac{2}{25}$
9	2	$\frac{2}{25}$

Calcoliamo la media, facendo uso della tabella delle frequenze

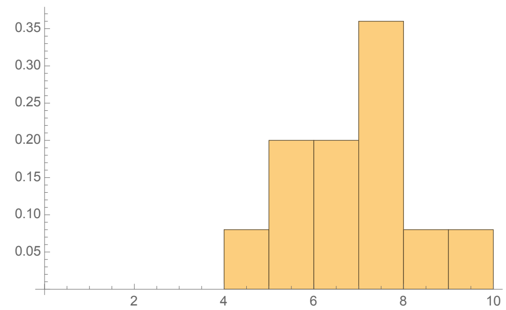
$$Av = \frac{4\cdot 2 + 5\cdot 5 + 6\cdot 5 + 7\cdot 9 + 8\cdot 2 + 9\cdot 2}{25} = \frac{160}{25} = 6.4.$$

Quindi la classe in media ha preso poco più della sufficienza. Già che ci siamo disegniamo gli istogrammi associati, vedi Figura 9.

Ci possiamo però chiedere un altro indice, anch'esso rilevante. Ovvero quale è quel voto che si trova in mezzo, ovvero tale che ci sono tot studenti che hanno preso una valutazione



(A) Istogramma passo 1 delle frequenze



(B) Istogramma frequenze relative

FIGURA 10

uguale o minore di quel voto e lo stesso numero che ha preso un voto maggiore o uguale di quel voto. In questo caso è semplice verificare che ci sono 12 persone che hanno preso un voto inferiore o uguale a 7 e 12 persone che hanno preso un voto maggiore o uguale a 7. Il numero 7 corrisponde alla **mediana** dei dati che abbiamo considerato. Non è stato difficile determinarla, il dato era composto da un numero dispari di elementi (25) e quindi non abbiamo fatto altro che prendere quello "in mezzo" ovvero il tredicesimo, quindi 7.

Da una semplice osservazione vediamo che, variando di poco i dati, ad esempio supponendo che lo studente A abbia preso 6 invece di 7 allora la media varia di poco ovvero avremo $Av = \frac{159}{25} = 6.36$. Invece la mediana cambia in modo più evidente ovvero diventa 6 invece di 7. D'altra parte, è vero che, se lo studente A avesse preso 10 invece di 7, la mediana sarebbe rimasta invariata uguale a 7, mentre la media sarebbe cambiata. Questo significa che la mediana risente in modo molto particolare delle variazioni dei singoli dati. La mediana in certi casi riveste una certa importanza, infatti ci fornisce un valore attraverso il quale sappiamo se ci troviamo sopra o sotto la "metà" di un campione di dati. In certi casi si sceglie proprio il valore della mediana per dividere un gruppo di elementi in due sottoclassi una "negativa" e una "positiva". Attenzione, per forza di cose, essendo n dispari le due classi non saranno composte dallo stesso numero di elementi. Tornando al nostro esempio, se decidiamo di mettere nella classe "positiva" tutti quelli che hanno preso un voto maggiore o uguale alla mediana avremo in quella classe 13 elementi e nell'altra classe 12 elementi. Invece se decidiamo di mettere nella classe "positiva" tutti quelli che hanno preso un voto strettamente più alto della mediana allora questa risulterà composta da appena 4 elementi. La mediana diventa un valore significativo soprattutto se il numero di dati da esaminare è molto alto e si decide di mettere una soglia a posteriori una volta raccolti un certo numero di dati e che possa servire su base statistica per nuovi dati raccolti.

Rimanendo per ora nel caso n dispari diamo la seguente definizione

Definizione 3.1. *Sia a una sequenza di dati ordinati: $a_1 \leq a_2 \leq \cdots \leq a_{n-1} \leq a_n$, dove n è un numero dispari. Si definisce mediana di a e si indica con $m(a)$ il punto di "mezzo" di tali dati, ovvero*

$$m(a) = a_{\frac{n+1}{2}}$$

Fate attenzione per il calcolo della mediana è importante non confondersi. Se si considerano solo gli elementi distinti tra i dati si commette un errore. Nella definizione di mediana gli elementi di dati che si ripetono pesano esattamente nello stesso modo degli elementi distinti. Quindi un istogramma o una tabella delle frequenze non ci aiutano a visualizzare la mediana ad occhio. Ci aiuta, invece, se il numero di dati non è eccessivo, ordnare i dati come in (3.1).

Esempio Consideriamo la lista di dati $\{1.2, 2, 3.2, 1.3, 5, 3.2, 1.2.1, 2, 3, 3, 1.2\}$. Si tratta di un numero dispari di dati, esattamente 11 dati, ordiniamoli

$$1.2 \leq 1.2 \leq 1.2 \leq 1.3 \leq 2 \leq 2.1 \leq 3 \leq 3 \leq 3.2 \leq 3.2 \leq 5.$$

La mediana corrisponderà al sesto tra questi ovvero 2.1.

Ora torniamo ad una problematica già affrontata quando abbiamo introdotto la varianza. Come possiamo valutare la distanza tra i nostri dati e la media o la mediana. La varianza è una delle strade, in quel caso si valuta la media dei dati $(a_i - Av(a))^2$. Più precisamente

$$Var(a) = \frac{1}{n}\sum_{i=1}^{n}(a_i - Av(a))^2. \tag{3.2}$$

Come visto nel paragrafo 1 la varianza risulta un ottimo indicatore per "stimare" l'errore dei dati rispetto alla loro media, in quanto verifica una proporietà di ottimizzazione.
D'altra parte il termine $(a_i - Av(a))^2 = |a_i - Av(a)|^2$. Sappiamo che il modulo della differenza non è altro che la misura della distanza tra due numeri sulla retta reale $\mathbb{R}$. A questo punto la questione che ci si può porre, in modo naturale, è la seguente: perché non prendere direttamente la media delle distanze tra i dati e la loro media invece che prendere la media dei quadrati delle distanze tra i dati e la loro media ?

Questo ci porta ad introdurre il seguente termine

$$D(a) = \frac{1}{n}\sum_{i=1}^{n}|a_i - Av(a)|$$

dove $D(a)$ misura la media della distanza tra i dati e la loro media. La prima cosa che ci chiediamo, per valutare se esso è un indicatore valido come la varianza, è se anche in questo caso vale una proprietà di ottimizzazione rispetto alla media.
Poiché il coefficiente $\frac{1}{n}$ non gioca nessun ruolo nel problema del minimo, possiamo ometterlo e considerare direttamente la funzione

$$G(x) = \sum_{i=1}^{n}|a_i - x|.$$

Ci chiediamo per quali valori di x la funzione G assume il minimo?
Consideriamo prima il caso $n = 3$. Abbiamo assunto che i dati siano ordinati per cui si ha $a_1 \leq a_2 \leq a_3$. Per vedere dove tale funzione assume il minimo dovremmo considerare i diversi casi e togliere il modulo a seconda della posizione del punto x. Ma a volte, si fatica di meno se si usa l'ingegno. È inutile fare troppe discussioni e considerare i 4 casi possibili. Sappiamo già che in tutti i casi verrà fuori una retta a volte con coefficiente angolare negativo altre volte positivo. In ogni caso il minimo sarà sempre raggiunto in uno

degli estremi e in nessun altro punto. La vera questione è essere sicuri che tale minimo sia effettivamente raggiunto. Ma questo possiamo supporlo, proponendoci poi di dimostrarlo quando avremo a disposizione i potenti mezzi dell'analisi reale. Preso per buono il fatto che il minimo ci sia questo dovrà necessariamente essere assunto in uno dei 3 punti a_1, a_2 e a_3. Abbiamo

$$G(a_1) = |a_2-a_1|+|a_3-a_1|,\ G(a_2) = |a_2-a_1|+|a_3-a_2| = |a_3-a_1|,\ G(a_3) = |a_3-a_2|+|a_3-a_1|,$$

da cui segue subito che il minimo è assunto nel punto a_2.

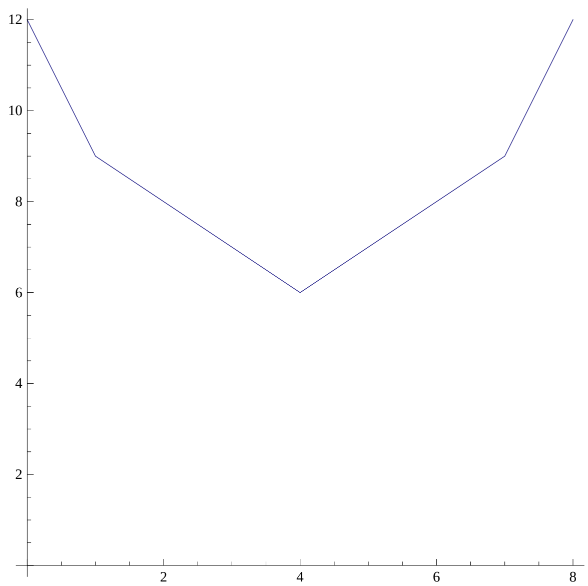


FIGURA 11. Grafico della funzione $G(x)$ nel caso $n = 3$ con $a_1 = 1$, $a_2 = 4$, $a_3 = 7$

A questo punto osserviamo che a_2 è proprio la mediana della sequenza di dati $a_1 \leq a_2 \leq a_3$.
Consideriamo il caso generale in cui n è dispari. Ci è utile scriverlo nel seguente modo $n = 2k + 1$ per qualche $k \in \mathbb{N}$. Dimostriamo che il minimo di G sarà assunto nel punto intermedio a_{k+1} corrispondente alla mediana dei dati. Utilizziamo il metodo dell'induzione. Quindi vogliamo far vedere che per ogni $n = 2k + 1$, k intero maggiore o uguale ad 1, il minimo di G è effettivamente assunto quando $x = a_{k+1} = m(a)$. Se $k = 1$ ovvero $n = 3$ l'abbiamo mostrato sopra. Supponiamo vero il caso $n = 2k + 1$ e mostriamo che esso è vero per $n = 2(k + 1) + 1 = 2k + 3$. Consideriamo la funzione

$$G(x) = \sum_{i=1}^{2n+3} |a_i - x| = |a_1 - x| + \sum_{i=2}^{2n+2} |a_i - x| + |a_{2n+3} - x|.$$

Ragionando come nel caso di 3 soli elementi, non è necessario analizzare gli innumerevoli casi per studiare il modulo, sappiamo già, dato per scontato che la funzione ammetta il minimo, che esso è sicuramente assunto in uno dei punti a_j, $j = 1, \dots 2n+3$. Se noi calcoliamo $G(x)$ in un qualsiasi punto a_j abbiamo

$$G(a_j) = |a_j - a_1| + |a_{2n+1} - a_j| + \sum_{i=2}^{2n+2} |a_i - a_j| = |a_{2n+3} - a_1| + \sum_{i=2}^{2n+2} |a_i - a_j|.$$

Infatti la somma dei primi 2 termini è indipendente dalla scelta di a_j. Quindi il minimo è assunto nel punto in cui l'assume il resto della sommatoria $\sum_{i=2}^{2n+2} |a_i - a_j|$. Possiamo quindi usare l'ipotesi induttiva in quanto siamo nel caso in cui ci sono $2k+1$ termini. Quindi il minimo è assunto nel punto intermedio a questi punti ovvero a_{k+2} che coincide con il punto intermedio dei valori $a_1, \dots a_{2k+3}$, da cui la tesi.

Analizziamo ora il caso n pari e cerchiamo di capire come definire la mediana. Riprendiamo l'esempio dei voti considerato all'inizio di questo paragrafo. Aggiungiamo un 26-esimo alunno che chiamiamo Z. Aggiorniamo quindi la tabella supponendo che abbia preso esattamente 7

$$\begin{bmatrix} A & B & C & D & E & F & G & H & I & J & K & L & M & N & O & P & Q & R & S & T & U & V & W & X & Y & Z \\ 7 & 5 & 4 & 9 & 7 & 6 & 7 & 7 & 9 & 7 & 6 & 5 & 7 & 8 & 6 & 6 & 7 & 5 & 10 & 7 & 7 & 6 & 5 & 4 & 8 & 7 \end{bmatrix}.$$

Non ci interessa discutere la media, visto che per come è definita non c'è differenza tra n pari o dispari. Vogliamo capire, invece, in che modo è sensato definire la mediana. Ordinando i dati avremo

$$4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 9, 9.$$

Sono 26 elementi. Concentriamoci su quelli in mezzo. Essendo n pari ne abbiamo 2 a metà ovvero il tredicesimo e il quattordicesimo. In questo caso essi coincidono e sono entrambi uguali a 7 per cui è sensato dire che la mediana è uguale a 7 anche in questa situazione. Supponiamo invece che Z invece di prendere 7 avesse preso 5. In tale situazione i dati ordinati diventano

$$4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 7, 7, 7, 7, 7, 7, 8, 8, 9, 9,$$

analizzando quelli in mezzo ci accorgiamo che il tredicesimo corrisponde a 6 mentre il quattordicesimo a 7. Come possiamo definire la mediana? Avendo in mente il concetto che la mediana deve essere un numero tale che sia maggiore o uguale alla metà dei dati e nello stesso tempo minore o uguale all'altra metà ci accorgiamo che tale numero non è più unico come nei casi precedenti. Infatti, qualsiasi valore compreso tra 6 e 7 può giocare tale ruolo, d'altra parte la scelta di tale numero tra 6 e 7 non è così significativa se vogliamo solo che esso sia un valore soglia rispetto ai dati che abbiamo. Per questo motivo una possibile scelta è prendere la media tra i 2 valori che si trovano "in mezzo". Nel nostro ultimo esempio $\frac{6+7}{2} = 6.5$. Nell'esempio predente invece poiché i valori centrali coincidevano la media corrispondeva proprio al valore comune 7. Riassumendo per n pari la mediana è data da $\frac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2}$. Alla luce di queste considerazioni diamo la seguente definizione di mediana per n pari

Definizione 3.2. *Sia a una sequenza di dati ordinati: $a_1 \leq a_2 \leq \cdots \leq a_{n-1} \leq a_n$, dove n è un numero pari. Si definisce mediana di a e si indica con $m(a)$ il punto*

$$m(a) = \frac{a_{\frac{n}{2}} + a_{\frac{n}{2}+1}}{2}.$$

Vediamo cosa succede al problema dell'ottimizzazione della funzione G nel caso n pari. In particolare analizziamo il caso $n = 4$. Dobbiamo studiare la funzione

$$G(x) = |a_1 - x| + |a_2 - x| + |a_3 - x| + |a_4 - x|.$$

Di nuovo se volessimo discutere il modulo, dovremmo considerare addirittura 6 casi. In 5 di essi verrà sempre una funzione lineare con coefficiente angolare positivo o negativo a seconda delle situazione e quindi il possibile minimo dovrà essere assunto in uno dei valori a_1, a_2, a_3, a_4. C'è un unico caso in cui vale la pena soffermarsi ovvero quello in cui x si trovi nell'intervallo $[a_2, a_3]$. Per ognuno di questi x si avrà

$$G(x) = x - a_1 + x - a_2 + a_3 - x + a_4 - x = a_3 + a_4 - a_1 - a_2$$

quindi la funzione risulterà costante. A questo punto, preso per buono che il minimo esiste, questo andrà cercato tra questo valore costante e i valori assunti dalla funzione in a_1 e a_4. Abbiamo

$$G(a_1) = a_2 - a_1 + a_3 - a_1 + a_4 - a_1 \geq a_4 - a_1 + a_3 - a_2$$

analogamente

$$G(a_4) = a_4 - a_1 + a_4 - a_3 + a_4 - a_2 \geq a_4 - a_1 + a_3 - a_2$$

Quindi il minimo della funzione G corrisponde a $a_3 + a_4 - a_1 - a_2$ ed è assunto per ogni x nell'intervallo $[a_2, a_3]$. Questo risultato si sovrappone al concetto di mediana per n pari. Ovvero tutti i valori che si pongono "in mezzo". In particolare ritroviamo il valore che abbiamo scelto come mediana ovvero $\frac{a_2+a_3}{2}$.

Con una dimostrazione simile a quella fatta nel caso $n = 4$ e usando il principio di induzione come nel caso n dispari (provate!) si dimostra che anche per n pari, $n = 2k$, la funzione G assume il minimo per tutti i valori in mezzo, ovvero contenuti nell'intervallo $[a_k, a_{k+1}]$, in particolare nel punto che abbiamo scelto per definire la mediana, ovvero $m(a) = \frac{a_k+a_{k+1}}{2}$.

In conclusione è proprio la mediana un ottimizzatore della funzione $G(x)$ e non è perciò interessante analizzare l'errore $D(a)$ introdotto sopra bensì l'errore

$$H(a) = \frac{1}{n} \sum_{i=1}^{n} |a_i - m(a)|.$$

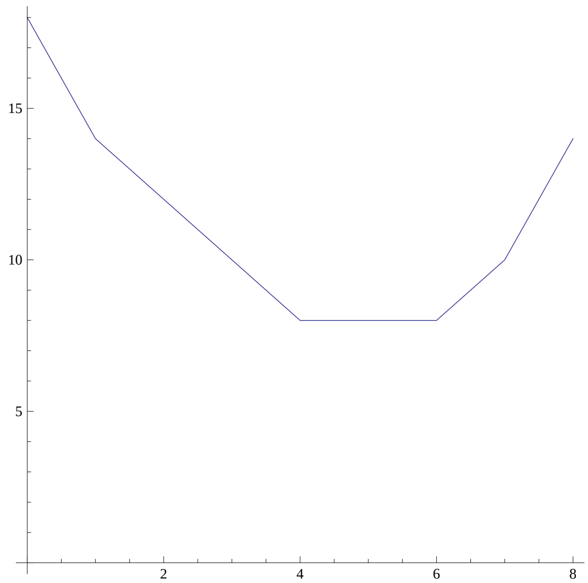


FIGURA 12. Grafico della funzione $G(x)$ nel caso $n = 4$ con $a_1 = 1$, $a_2 = 4$, $a_3 = 6$, $a_4 = 7$