## 3 Le medie

La costruzione di una distribuzione di frequenza consente di disporre di una rappresentazione più compatta e informativa rispetto alla serie dei dati osservati.

Alle distribuzioni di frequenza vanno affiancate le rappresentazioni grafiche che, sebbene non consentano di evidenziare eventuali sfumature del fenomeno oggetto di studio, tuttavia ne danno una visione immediata, interpretabile non solo da un esperto di Statistica.

Si è detto che uno dei compiti fondamentali della Statistica è quello di riassumere, in alcune costanti di sintesi, caratteristiche particolari del fenomeno.

Esistono diverse categorie di costanti sintetiche, ciascuna descrive un aspetto di una distribuzione. In particolare, i valori medi, se i dati sono quantitativi, ne pongono in evidenza la "dimensione" o "intensità", ossia il loro ordine di grandezza. La scelta del tipo di media da utilizzare dipende dalla tipologia dei dati a disposizione e dagli scopi che ci si propone in una ricerca.

Le *medie che discendono dalla definizione di Chisini* sono grandezze che derivano o dipendono dai valori dati e che sostituite ad essi li sintetizzano senza alterare la visione d'insieme del fenomeno considerato.

Le *medie di posizione* suddividono la serie osservata in un numero prefissato di parti uguali; tali medie trovano giustificazione nella definizione di Cauchy: "*è valore medio di una serie di dati qualsiasi valore compreso tra il più piccolo e il più grande di essi*".

Le *medie decisionali* derivano dalla minimizzazione di una funzione di perdita dell'informazione. I dati osservati, infatti, se risultano dalla misura ripetuta di uno stesso oggetto o soggetto, sono affetti inevitabilmente, da errori accidentali.

### 3.1 Medie secondo il Chisini

Le medie di *Chisini* si applicano su dati rilevati su oggetti/soggetti diversi, omogenei, ossia rilevati con la stessa unità di misura, e per ipotesi non affetti da errori.

Fissata una funzione *f*, si chiama media quel valore costante *M* che, sostituito ad ogni singolo valore, lascia inalterata la seguente uguaglianza:

$$f(M, M, \ldots, M) = f(x_1, x_2, \ldots, x_n).$$

Se il fenomeno è additivo, la funzione *f* è la funzione somma, e la media *M* che si ricava dall'uguaglianza suddetta è la media aritmetica:

$$\sum_{i=1}^{n} M = \sum_{i=1}^{n} x_i \qquad nM = \sum_{i=1}^{n} x_i \qquad M = \frac{\sum_{i=1}^{n} x_i}{n}$$

Se il fenomeno è moltiplicativo, ovvero se si evolve in modo più che proporzionale rispetto all'unità di misura considerata, la funzione *f* è la funzione prodotto e la media *M* che si ricava dalla precedente uguaglianza è la media geometrica:

$$\prod_{i=1}^{n} M = \prod_{i=1}^{n} x_i \qquad M^n = \prod_{i=1}^{n} x_i \qquad M = \sqrt[n]{\prod_{i=1}^{n} x_i} \quad .$$

Se le $x_i$ sono funzioni di altre variabili: $x_i = f(y_i)$, per esempio $x_i = y_i^m$, l'uguaglianza diviene:

$$f(M^m, M^m, \ldots, M^m) = f(y_1^m, y_2^m, \ldots, y_n^m)$$

da cui, se *f* è la funzione somma, si ricava la *media potenziata di ordine m*:

$$\sum_{i=1}^{n} M^m = \sum_{i=1}^{n} y_i^m \qquad nM^m = \sum_{i=1}^{n} y_i^m \qquad M^m = \frac{\sum_{i=1}^{n} y_i^m}{n}$$

$$M = \sqrt[m]{\frac{\sum_{i=1}^{n} y_i^m}{n}} = \left( \frac{\sum_{i=1}^{n} y_i^m}{n} \right)^{1/m}$$

Per una distribuzione di frequenze è:

$$M = \left( \frac{\sum_{i=1}^{k} y_i^m n_i}{n} \right)^{1/m} \quad .$$

Al variare di *m*, si ricavano le seguenti medie:

*m=-1* *media armonica*

$m \to 0$ *media geometrica*

*m=1* *media aritmetica*

*m=2* *media quadratica*

*m=3* *media cubica*

tra le quali vale la relazione

$M_{-1} \le M_0 \le M_1 \le M_2 \le M_3$,

avendosi l'uguaglianza solo nel caso in cui le $y_i$ siano costanti.

Se $f$ è la funzione prodotto, si ottiene la media geometrica:

$$\prod_{i=1}^{n} M^m = \prod_{i=1}^{n} y_i^m \qquad M^{mn} = \prod_{i=1}^{n} y_i^m$$

$$M = \left( \prod_{i=1}^{n} y_i^m \right)^{\frac{1}{mn}} = \sqrt[n]{\prod_{i=1}^{n} y_i} \quad .$$

Considerandone il logaritmo si ha:

$$\log M = \frac{1}{n} \sum_{i=1}^{n} \log y_i \ .$$

La media geometrica deve il suo nome al fatto che rappresenta il termine centrale di una progressione geometrica, con un numero di termini dispari.

*Proprietà della media geometrica*

1) La m.g. di una serie di valori moltiplicati per una costante è uguale alla costante per la m.g. dei valori;
2) La m.g. di una serie di rapporti di valori è uguale al rapporto tra le m.g. delle due serie di valori;
3) La m.g. del reciproco di una serie di valori è uguale al reciproco della m.g.

**Esempi sulle medie potenziate di ordine m**

*Media quadratica*

Si abbiano quattro piastrine d'oro quadrate di uguale spessore, ma di lati rispettivamente uguali a 2, 4, 10, 8 *cm*. Si vogliano fondere e forgiare in 4 piastrine quadrate di lato uguale. Il lato medio sarà:

$$M_2 = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} = \sqrt{\frac{2^2 + 4^2 + 10^2 + 8^2}{4}} = \sqrt{\frac{184}{4}} = \sqrt{46} = 6{,}7823$$

*Media cubica*

Si abbiano 4 cubetti d'oro di diverso volume. Si vogliano fondere e forgiare in 4 cubetti di uguale volume. Se i lati dei cubetti misurano rispettivamente *mm* 2, 4, 10, 8, il lato medio sarà:

$$M_3 = \left( \frac{\sum_{i=1}^{n} x_i^3}{n} \right)^{1/3} = \left( \frac{2^3 + 4^3 + 10^3 + 8^3}{4} \right)^{1/3} = \left( \frac{1584}{4} \right)^{1/3} = \sqrt[3]{396} = 7{,}34\,.$$

*Media geometrica*

*Esempio 1*

Un bene dal costo iniziale C subisce:

- il 1° anno un aumento del 9%;
- il 2° anno un aumento del 14% sul costo del 1° anno;

- il 3° anno un aumento del 12% sul costo del 2° anno;
- il 4° anno un aumento del 10% sul costo del 3° anno.

Determinare l'aumento percentuale medio.

$r_1=0,09 \qquad r_2=0,14 \qquad r_3=0,12 \qquad r_4=0,10$

$C_1=C+Cr_1=C(1+r_1)$

$C_2=C_1+C_1r_2=C_1(1+r_2)= C(1+r_1)\,(1+r_2)$

$C_3=C_2+C_2r_3=C_2(1+r_3)= C(1+r_1)\,(1+r_2)\,(1+r_3)$

$C_4=C_3+C_3r_4=C_3(1+r_4)= C(1+r_1)\,(1+r_2)\,(1+r_3)\,(1+r_4)$

$$C(1+r_1)\,(1+r_2)\,(1+r_3)\,(1+r_4)= C(1+r_M)^4$$

$$\sqrt[4]{1,09 \cdot 1,14 \cdot 1,12 \cdot 1,10} = 1 + r_M$$

$$1,1123-1= r_M \qquad \Rightarrow \qquad r_M=0,1123$$

Dunque il tasso di aumento medio durante i 4 anni è dell'11,23%.

*Esempio 2*

Il numero di microrganismi in una certa coltura è aumentato da 2000 a 9000 in tre giorni.

Qual è stato l'incremento medio giornaliero?

Il n. dei microrganismi dopo un giorno sarà:

$n_1=2000+2000r=2000(1+r)$

Dopo 2 giorni:

$n_2=n_1+n_1r=n_1(1+r)=2000(1+r)^2$

Dopo 3 giorni:

$n_3=n_2+n_2r=n_2(1+r)=2000(1+r)^3$

Poiché il n. dei microrganismi alla fine dei 3 giorni è uguale a 9000, si ha:

$n_3=9000=2000(1+r)^3$

da cui, risolvendo rispetto ad r si ottiene:

$4{,}5=(1+r)^3 \quad \Rightarrow \quad \sqrt[3]{4{,}5}=1+r \quad \Rightarrow \quad \sqrt[3]{4{,}5}-1=r \quad \Rightarrow \quad r=0{,}6509$

Il tasso di crescita medio è stato dunque del 65,1%.

*Media armonica*

Viene utilizzata quando si hanno quantità tra cui esiste una relazione inversa (es. durata e consumi, velocità e tempo, ecc…)

*Esempio 1*

In 4 prove di velocità sul km lanciato, un corridore in bicicletta ha realizzato, rispettivamente, le velocità di 62, 64, 65, 68 *km* all'ora.

I reciproci di queste velocità forniscono il tempo ($v=s/t$), in frazioni di ora, impiegato in ciascuna delle 4 prove, per percorrere un *km*: 1/62, 1/64, 1/65, 1/68.

Determinare quella velocità media che lasci invariato il tempo totale cronometrato nelle 4 prove:

$$\frac{1}{62}+\frac{1}{64}+\frac{1}{65}+\frac{1}{68}=4\cdot\frac{1}{x}$$

da cui

$$x=\frac{4}{\frac{1}{62}+\frac{1}{64}+\frac{1}{65}+\frac{1}{68}}=64{,}68$$

*Esempio 2*

Nelle analisi di mercato spesso è interessante conoscere il consumo medio annuo di un determinato prodotto. Supponiamo si voglia indagare sul consumo medio annuo di lamette da barba; viene dunque intervistato un campione di consumatori:

| persone | durata media in giorni di una lametta | consumo annuo di lamette |
|---|---|---|
| 1 | 10 | 365:10=36,5 |
| 2 | 6 | 365:6=60,8 |
| 3 | 30 | 365:30=12,2 |
| 4 | 5 | 365:5=73 |
| 5 | 14 | 365:14=26,1 |
| totale | 65 | 208,6 |

consumo pro-capite: $\frac{208,6}{5} = 41,7$ lamette

durata media di ogni lametta: $\frac{365}{41,7} = 8,8$ giorni.

Più semplicemente:

$$M_{-1} = \frac{5}{\frac{1}{10}+\frac{1}{6}+\frac{1}{30}+\frac{1}{5}+\frac{1}{14}} = 8,8\,.$$

*Esempio 3*

Un individuo spende per il riscaldamento di 3 anni consecutivi sempre la stessa cifra di 1500 ∈ all'anno, acquistando il combustibile a:

- 0,30 ∈ il 1° anno;
- 0,40 ∈ il 2° anno;
- 0,50 ∈ il 3° anno.

Determinare il costo medio di 1 l di combustibile per l'intero periodo.

Sono stati acquistati:

- il 1° anno $\frac{1500}{0,30} = 5000$ l di combustibile;
- il 2° anno $\frac{1500}{0,40} = 3750$ l di combustibile;
- il 3° anno $\frac{1500}{0,50} = 3000$ l di combustibile.

Il costo medio al l per l'intero periodo è:

$$\frac{\text{COSTO TOTALE}}{\text{TOTALE LITRI}} = \frac{3 \cdot 1500}{5000 + 3750 + 3000} = 0{,}38 \in .$$

Più rapidamente, basta calcolare la media armonica del costo al l:

$$M_{-1} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{3}{\frac{1}{0{,}30} + \frac{1}{0{,}40} + \frac{1}{0{,}50}} = 0{,}38 \in .$$

### 3.2 Medie di posizione

Le medie di posizione trovano applicazione nel contesto di una serie di modalità/valori ordinati in successione non decrescente:

$$x_{(1)} , x_{(2)} , x_{(3)} , \ldots\ldots , x_{(n)}$$

Definiamo "QUANTILI" quei valori che ripartiscono la serie osservata in $(q+1)$ parti di uguale numerosità; ovviamente è $q \leq n-1$. Al variare di $q$, si ottengono i seguenti quantili:

*q=1 mediana*

*q=2 terzili*

*q=3 quartili*

*q=5 sestili*

*q=9 decili*

*q=99 centili*.

Nella stessa serie il 2° quartile, così come il 3° sestile, coinciderà con la mediana:

$$M_e = \begin{cases} x_{\frac{n+1}{2}} & \text{se } n \text{ è dispari} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{se } n \text{ è pari} \end{cases}$$

il pedice indica la posizione che il valore $x$ occupa nella serie.

Ad esempio, supponiamo di aver rilevato il peso in kg di 13 uomini:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 78 | 75 | 73 | 90 | 88 | 87 | 83 | 76 | 88 | 78 | 80 | 83 | 81 |

Volendo calcolare la mediana, dobbiamo innanzitutto ordinare la serie:

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 73 | 75 | 76 | 78 | 78 | 80 | 81 | 83 | 83 | 87 | 88 | 88 | 90 |

Poiché il numero delle osservazioni $n$=13 è dispari, la mediana è:

$$M_e = x_{\frac{n+1}{2}} = 81.$$

Se non avessimo osservato l'ultimo valore $x_{(13)} = 90$, il numero delle osservazioni $n$=12 sarebbe stato pari. In tal caso,

$$M_e = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} = \frac{80+81}{2} = 80{,}5\,.$$

Se la variabile in esame è quantitativa continua, i quantili possono essere calcolati nel seguente modo:

$$\overline{x}_{\frac{i}{q+1}} = \begin{cases} x_{\left[\frac{ni}{q+1}\right]+1} & \text{se} \left[\frac{ni}{q+1}\right] \neq \frac{ni}{q+1} \\ \left( x_{\left[\frac{ni}{q+1}\right]} + x_{\left[\frac{ni}{q+1}\right]+1} \right) : 2 & \text{se} \left[\frac{ni}{q+1}\right] = \frac{ni}{q+1} \end{cases}$$

dove *i=1, 2,…, q*.

Supponiamo, ad esempio, di aver rilevato il peso in *kg* di n=8 donne. Si riporta la serie già ordinata:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 52 | 54 | 58 | 59 | 60 | 60 | 63 | 65 |

<u>Mediana</u>

$$\overline{x}_{\frac{1}{2}} = \overline{x}_{0,5} = \frac{x_4 + x_5}{2} = \frac{59+60}{2} = 59{,}5$$

Terzili

$$\bar{x}_{\frac{1}{3}} = \bar{x}_{0,3\bar{3}} = x_3 = 58$$

$$\bar{x}_{\frac{2}{3}} = \bar{x}_{0,6\bar{6}} = x_6 = 60$$

Quartili

$$\bar{x}_{\frac{1}{4}} = \bar{x}_{0,25} = \frac{x_2 + x_3}{2} = \frac{54 + 58}{2} = 56$$

$$\bar{x}_{\frac{2}{4}} = \bar{x}_{0,5} = \frac{x_4 + x_5}{2} = 59{,}5$$

$$\bar{x}_{\frac{3}{4}} = \bar{x}_{0,75} = \frac{x_6 + x_7}{2} = \frac{60 + 63}{2} = 61{,}5$$

Sestili

$$\bar{x}_{\frac{1}{6}} = \bar{x}_{0,1\bar{6}} = x_2 = 54$$

$$\bar{x}_{\frac{2}{6}} = \bar{x}_{0,3\bar{3}} = x_3 = 58$$

$$\bar{x}_{\frac{3}{6}} = \bar{x}_{0,5} = \frac{x_4 + x_5}{2} = 59{,}5$$

$$\bar{x}_{\frac{4}{6}} = \bar{x}_{0,6\bar{6}} = x_6 = 60$$

$$\bar{x}_{\frac{5}{6}} = \bar{x}_{0,8\bar{3}} = x_7 = 63$$

Vediamo adesso come calcolare i quantili su una distribuzione di frequenze.

Consideriamo la seguente distribuzione:

| TITOLO DI STUDIO | $n_i$ |
|---|---|
| **licenza elementare, nessun titolo** | 20442469 |
| **licenza media** | 16403989 |
| **qualifica professionale** | 2554109 |
| **maturita'** | 11254538 |
| **dottorato, laurea, diploma universitario** | 3267219 |
| TOTALE | 53922324 |

Popolazione residente in Italia nell'anno 1999 secondo il titolo di studio

Fonte: ISTAT, Annuario statistico italiano 1999

Per determinare i quantili occorre calcolare le frequenze cumulate:

| $N_i = \sum_{h=1}^{i} n_h$ | $f_i$ | $F_i = \sum_{h=1}^{i} f_h$ | $f_i*100$ | $F_i*100$ |
|---|---|---|---|---|
| 20442469 | 0,379 | 0,379 | 37,911 | 37,911 |
| 36846458 | 0,304 | 0,683 | 30,422 | 68,332 |
| 39400567 | 0,047 | 0,731 | 4,737 | 73,069 |
| 50655105 | 0,209 | 0,939 | 20,872 | 93,941 |
| 53922324 | 0,061 | 1 | 6,059 | 100 |
| | 1 | | 100 | |

Poiché $N$=53922324 è pari, la mediana occuperà una posizione compresa tra $\frac{N}{2} = 26961162$ e $\frac{N}{2}+1 = 26961163$. Tali posizioni sono contenute nella seconda frequenza cumulata $N_2$=36846458, cui è associata la modalità "licenza media". D'altra parte, guardando le frequenze relative o percentuali cumulate, si evince subito che il 50% delle osservazioni è contenuto proprio in $F_2$.

Consideriamo adesso la distribuzione di frequenze del numero di carburatori osservati su 32 automobili di marca diversa:

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ | $x_i n_i$ |
|---|---|---|---|---|---|
| 1 | 7 | 0,219 | 7 | 0,219 | 7 |
| 2 | 10 | 0,313 | 17 | 0,531 | 20 |
| 3 | 3 | 0,094 | 20 | 0,625 | 9 |
| 4 | 10 | 0,313 | 30 | 0,938 | 40 |
| 5 | 0 | 0 | 30 | 0,938 | 0 |
| 6 | 1 | 0,031 | 31 | 0,969 | 6 |
| 7 | 0 | 0 | 31 | 0,969 | 0 |
| 8 | 1 | 0,031 | 32 | 1 | 8 |
| *totale* | 32 | 1 | | | 90 |

Calcoliamo la mediana e la media aritmetica:

$$M_e = \frac{x_{\frac{N}{2}} + x_{\frac{N}{2}+1}}{2} = \frac{x_{16} + x_{17}}{2} = 2$$

$$M = \frac{\sum_{i=1}^{8} x_i n_i}{N} = \frac{90}{32} = 2{,}8125$$

Notiamo che la media aritmetica, essendo espressa da un numero decimale, non può rappresentare il numero di carburatori di un automobile! Per variabili di conteggio, dunque, la media aritmetica assume valore "indicativo-formale", mentre i valori medi di posizione assumono pienezza di significato.

Consideriamo la distribuzione del numero di prodotti difettosi di un certo processo produttivo:

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 0 | 3 | 0,06 | 3 | 0,06 |
| 1 | 9 | 0,18 | 12 | 0,24 |
| 2 | 13 | 0,26 | 25 | 0,50 |
| 3 | 11 | 0,22 | 36 | 0,72 |
| 4 | 8 | 0,16 | 44 | 0,88 |
| 5 | 4 | 0,08 | 48 | 0,96 |
| 6 | 2 | 0,04 | 50 | 1,00 |
| totale | 50 | 1,00 | | |

Poiché *N=50* è pari, la mediana occuperà una posizione compresa tra $\frac{N}{2} = 25$ e $\frac{N}{2} + 1 = 26$. Osserviamo però che la 25° osservazione è compresa nella terza frequenza cumulata $N_3$ =25, cui è associato il valore 2, mentre la 26° osservazione è compresa nella quarta frequenza cumulata $N_4$ =36, cui è associato il valore 3.

Per convenzione si considera, allora, la semisomma di tali valori:

$$M_e = Q_2 = \frac{2+3}{2} = 2{,}5 .$$

Volendo calcolare gli altri due quartili, $Q_1$ e $Q_3$, basta osservare le frequenze relative cumulate; quella che contiene il 25% delle osservazioni è $F_3$, mentre quella che contiene il 75% delle osservazioni è $F_5$, dunque $Q_1$ =2 e $Q_3$ =4.

Consideriamo la distribuzione di un gruppo di famiglie agricole secondo il numero dei figli:

| $x_i$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 0 | 4 | 0,009 | 4 | 0,009 |
| 1 | 9 | 0,021 | 13 | 0,030 |
| 2 | 34 | 0,078 | 47 | 0,107 |
| 3 | 77 | 0,176 | 124 | 0,284 |
| 4 | 94 | 0,215 | 218 | 0,499 |
| 5 | 88 | 0,201 | 306 | 0,700 |
| 6 | 65 | 0,149 | 371 | 0,849 |
| 7 | 40 | 0,092 | 411 | 0,940 |
| 8 | 15 | 0,034 | 426 | 0,975 |
| 9 | 4 | 0,009 | 430 | 0,984 |
| 10 | 5 | 0,011 | 435 | 0,995 |
| 11 | 2 | 0,005 | 437 | 1,000 |
| totale | 437 | | | |

$Q_1$ =3

$Q_2 = M_e = x_{\frac{N+1}{2}} = x_{219} = 5$

$Q_3$ =6.

Supponiamo adesso di voler calcolare i quartili su una distribuzione di frequenze per classi. Si considerino le temperature (in gradi) minime giornaliere di 25 mesi di luglio in una zona delle alpi orientali:

| $x_i$ -$x_{i+1}$ | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 6-7 | 1 | 0,001 | 1 | 0,001 |
| 7-8 | 1 | 0,001 | 2 | 0,002 |
| 8-9 | 4 | 0,005 | 6 | 0,007 |
| 9-10 | 15 | 0,019 | 21 | 0,027 |
| 10-11 | 52 | 0,067 | 73 | 0,094 |
| 11-12 | 84 | 0,108 | 157 | 0,202 |
| 12-13 | 131 | 0,169 | 288 | 0,371 |
| 13-14 | 121 | 0,156 | 409 | 0,527 |
| 14-15 | 108 | 0,139 | 517 | 0,667 |
| 15-16 | 114 | 0,147 | 631 | 0,814 |
| 16-17 | 75 | 0,097 | 706 | 0,911 |
| 17-18 | 45 | 0,058 | 751 | 0,969 |
| 18-19 | 13 | 0,017 | 764 | 0,986 |
| 19-20 | 9 | 0,012 | 773 | 0,997 |
| 20-21 | 2 | 0,003 | 775 | 1,000 |
| | 775 | 1 | | |

Poiché *N=775* è dispari, la mediana occuperà la posizione $\frac{N+1}{2}=388$. Tale posizione è compresa nell'8° frequenza cumulata, cui è associata la classe 13-14. La mediana, pertanto, sarà un valore all'interno di tale classe. Per convenzione, si sceglie il valore centrale della classe $M_e=\frac{13+14}{2}=13{,}5$ o, meglio, si utilizza il metodo dell'interpolazione. A tale proposito, ricordiamo che, l'equazione di una retta passante per due punti A e B è:

$$\frac{y-y_A}{y_B-y_A}=\frac{x-x_A}{x_B-x_A}.$$

Considerato un punto P(*x*, *y*) appartenente alla retta, di cui è nota l'ordinata *y*, è semplice, quindi, determinarne l'ascissa *x*:

$$x=\frac{y-y_A}{y_B-y_A}\left(x_B-x_A\right)+x_A.$$

Se ipotizziamo che A e B abbiano coordinate A($x_i$, $N_{i-1}$), B($x_{i+1}$, $N_i$), il punto $P(M_e, \frac{N+1}{2})$ è interno al segmento A-B, per cui la mediana si determina facilmente:

$$\frac{\frac{N+1}{2}-N_{i-1}}{N_i-N_{i-1}}=\frac{M_e-x_i}{x_{i+1}-x_i}$$

e poiché $N_i - N_{i-1}=n_i$:

$$M_e=\frac{\frac{N+1}{2}-N_{i-1}}{n_i}\left(x_{i+1}-x_i\right)+x_i=\frac{388-288}{121}\left(14-13\right)+13=13{,}83.$$

Ovviamente, se $\frac{N+1}{2}=N_i$, allora $M_e= x_{i+1}$.

Consideriamo un altro esempio, in cui *N* è pari:

| $x_i$ -$x_{i+1}$ | $n_i$ | $N_i$ |
|---|---|---|
| 50-100 | 110 | 110 |
| 100-200 | 400 | 510 |
| 200-300 | 90 | 600 |
| *totale* | 600 | |

*N=600* $\frac{N}{2} = 300$ $\frac{N}{2} + 1 = 301$

$$\frac{\frac{\left(\frac{N}{2} + \frac{N}{2} + 1\right)}{2} - N_{i-1}}{N_i - N_{i-1}} = \frac{M_e - x_i}{x_{i+1} - x_i}$$

In luogo di $\frac{\left(\frac{N}{2} + \frac{N}{2} + 1\right)}{2}$ si può considerare semplicemente $\frac{N}{2}$:

$$M_e = \frac{\frac{N}{2} - N_{i-1}}{n_i}\left(x_{i+1} - x_i\right) + x_i = \frac{300 - 110}{400}\left(200 - 100\right) + 100 = 147{,}5\,.$$

In modo analogo si possono determinare gli altri quartili.

**3.3 Medie decisionali**

Questi valori medi rivestono un'importanza particolare dal punto di vista scientifico, perché presentano un valore informativo molto elevato.

Trovano collocazione nell'ambito di tutti quei fenomeni del reale ripetibili, per i quali cioè è possibile ripetere più volte, e nelle stesse condizioni, la misura di una grandezza incognita *X*.

Questi dati risultano affetti da errori accidentali, dovuti alla presenza di infiniti fattori di disturbo, che non consentono di determinare con esattezza la misura *X* della grandezza cui siamo interessati.

La Statistica è chiamata a trovare il modo più opportuno di combinare le osservazioni, al fine di ottenere la migliore valutazione del valore investigato *X*.

Se assumiamo che la relazione fra il vero valore $X$ e l'errore casuale sia di tipo additivo

$$x_i = X + \varepsilon_i \qquad\qquad i = 1, 2, \ldots, n$$

ciò che interessa è minimizzare l'errore $\forall x_i$:

$$\varepsilon_i = x_i - X,$$

per cui la migliore combinazione $\bar{x}$ delle osservazioni $x_i$ si ottiene minimizzando la *funzione di perdita globale* dell'informazione contenuta nei dati:

$$\Sigma |\varepsilon_i|^p = \Sigma |x_i - \bar{x}|^p,$$

$p \in (0, \infty)$ è un parametro che dipende dalla natura probabilistica dell'errore $\varepsilon_i$.

La metodologia statistica si fonda in buona parte sull'assunzione che gli errori seguano una distribuzione di probabilità normale.

In realtà, gli errori seguono una distribuzione di tipo simmetrico ed unimodale, che varia, al variare di $p$, da forme cuspidate a forme più appiattite (*famiglia di curve normali di ordine p*).

In particolare si dimostra che:

- per $p=1$, $\varepsilon_i \sim$ LAPLACE $\Rightarrow$ $\bar{x}$ : mediana
- per $p=2$, $\varepsilon_i \sim$ NORMALE $\Rightarrow$ $\bar{x}$ : media aritmetica
- per $p \to \infty$, $\varepsilon_i \sim$ UNIFORME $\Rightarrow$ $\bar{x}$ : semisomma dei valori estremi

Questa media $\bar{x}$, che indicheremo con $M_{p-1}$, per sottolineare che dipende da $p$, rappresenta il vero valore della grandezza investigata $X$, cioè il valore che avremmo misurato se non ci fosse stata la presenza dell'errore.

La media $M_{p-1}$, detta "media di norma p", si ottiene, come si è detto, minimizzando la funzione di perdita globale, ossia risolvendo l'equazione:

$$-p\Sigma |x_i - M_{p-1}|^{p-1} segno(x_i - M_{p-1}) = 0,$$

che ha soluzione esplicita solo per *p=2*:

$$\sum_{i=1}^{n}\left(x_i - M_1\right)^2 = \text{minimo}$$

$$-2\sum_{i=1}^{n}\left(x_i - M_1\right) = 0 \qquad \Rightarrow \qquad \sum_{i=1}^{n} x_i - nM_1 = 0 \qquad \Rightarrow \qquad M_1 = \frac{\sum_{i=1}^{n} x_i}{n}.$$

*Esempi di medie decisionali*

La seguente tabella riporta 150 misure sperimentali rilevate in un punto di un circuito elettronico con un voltmetro digitale; i valori (misure della tensione) sono espressi in volt:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 5,145 | 5,120 | 5,146 | 5,114 | 5,134 | 5,148 | 5,146 | 5,143 | 5,145 | 5,156 |
| 5,132 | 5,138 | 5,139 | 5,140 | 5,139 | 5,132 | 5,128 | 5,142 | 5,132 | 5,138 |
| 5,143 | 5,159 | 5,123 | 5,148 | 5,131 | 5,143 | 5,146 | 5,129 | 5,141 | 5,135 |
| 5,145 | 5,139 | 5,136 | 5,161 | 5,118 | 5,141 | 5,138 | 5,152 | 5,146 | 5,138 |
| 5,131 | 5,160 | 5,169 | 5,142 | 5,129 | 5,131 | 5,128 | 5,140 | 5,150 | 5,130 |
| 5,124 | 5,150 | 5,140 | 5,136 | 5,150 | 5,158 | 5,144 | 5,132 | 5,145 | 5,142 |
| 5,133 | 5,137 | 5,131 | 5,137 | 5,154 | 5,155 | 5,126 | 5,126 | 5,133 | 5,149 |
| 5,128 | 5,125 | 5,133 | 5,134 | 5,144 | 5,133 | 5,157 | 5,134 | 5,138 | 5,142 |
| 5,143 | 5,166 | 5,154 | 5,134 | 5,124 | 5,129 | 5,155 | 5,153 | 5,146 | 5,154 |
| 5,158 | 5,148 | 5,140 | 5,133 | 5,134 | 5,133 | 5,152 | 5,155 | 5,132 | 5,135 |
| 5,136 | 5,148 | 5,153 | 5,150 | 5,147 | 5,162 | 5,129 | 5,148 | 5,151 | 5,157 |
| 5,151 | 5,137 | 5,128 | 5,140 | 5,143 | 5,140 | 5,130 | 5,153 | 5,142 | 5,151 |
| 5,146 | 5,148 | 5,137 | 5,157 | 5,158 | 5,157 | 5,153 | 5,131 | 5,164 | 5,159 |
| 5,134 | 5,148 | 5,144 | 5,143 | 5,156 | 5,147 | 5,145 | 5,123 | 5,140 | 5,162 |
| 5,139 | 5,152 | 5,132 | 5,154 | 5,128 | 5,140 | 5,151 | 5,138 | 5,139 | 5,142 |

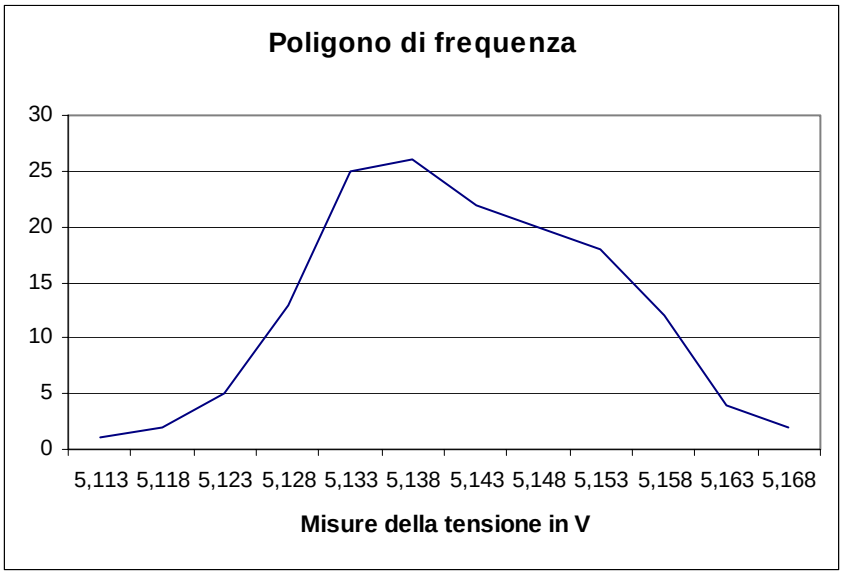
Si tratta di misure ripetute della stessa grandezza, affette da errori accidentali, dunque i valori medi più idonei a rappresentare tale grandezza sono le medie decisionali.

Si è scelto di raggruppare i dati in 12 classi di ampiezza pari a 0,005 V. Anche se i dati assumono valori compresi tra 5,114 e 5,169, si è ampliato l'intervallo di variazione e scelto come estremo inferiore 5,110 e come estremo superiore 5,170. Le classi sono chiuse a destra:

| $x_i$ | $x_{i+1}$ | $n_i$ |
|---|---|---|
| 5,110 | 5,115 | 1 |
| 5,115 | 5,120 | 2 |
| 5,120 | 5,125 | 5 |
| 5,125 | 5,130 | 13 |
| 5,130 | 5,135 | 25 |
| 5,135 | 5,140 | 26 |
| 5,140 | 5,145 | 22 |
| 5,145 | 5,150 | 20 |
| 5,150 | 5,155 | 18 |
| 5,155 | 5,160 | 12 |
| 5,160 | 5,165 | 4 |
| 5,165 | 5,170 | 2 |
| *totale* | | 150 |

Dalla rappresentazione grafica dei dati, si possono avere informazioni, sebbene molto grossolane, sulla distribuzione degli errori. I dati, infatti, differiscono dagli errori per una costante:

$$x_i = X + \varepsilon_i.$$




Se si suppone che i dati provengano da una distribuzione normale ($p=2$), la media più appropriata è la media aritmetica:

| $_cx_i$ | $n_i$ | $_cx_i \cdot n_i$ |
|---|---|---|
| 5,113 | 1 | 5,113 |
| 5,118 | 2 | 10,235 |
| 5,123 | 5 | 25,613 |
| 5,128 | 13 | 66,658 |
| 5,133 | 25 | 128,313 |
| 5,138 | 26 | 133,575 |
| 5,143 | 22 | 113,135 |
| 5,148 | 20 | 102,950 |
| 5,153 | 18 | 92,745 |
| 5,158 | 12 | 61,890 |
| 5,163 | 4 | 20,650 |
| 5,168 | 2 | 10,335 |
| | 150 | 771,210 |

$$M_1 = \frac{\sum_{i=1}^{n} {}_c x_i n_i}{n} = \frac{771{,}21}{150} = 5{,}1414$$

Se si suppone che i dati provengano da una distribuzione di Laplace (*p=1*); la media più appropriata è la mediana:

| $x_i$ | $x_{i+1}$ | $n_i$ | $N_i$ |
|---|---|---|---|
| 5,110 | 5,115 | 1 | 1 |
| 5,115 | 5,120 | 2 | 3 |
| 5,120 | 5,125 | 5 | 8 |
| 5,125 | 5,130 | 13 | 21 |
| 5,130 | 5,135 | 25 | 46 |
| 5,135 | 5,140 | 26 | 72 |
| 5,140 | 5,145 | 22 | 94 |
| 5,145 | 5,150 | 20 | 114 |
| 5,150 | 5,155 | 18 | 132 |
| 5,155 | 5,160 | 12 | 144 |
| 5,160 | 5,165 | 4 | 148 |
| 5,165 | 5,170 | 2 | 150 |
| *totale* | | 150 | |

$$M_0 = \frac{\frac{N}{2} - N_{i-1}}{n_i}\left(x_{i+1} - x_i\right) + x_i = \frac{\frac{150}{2} - 72}{22}\left(5{,}145 - 5{,}140\right) + 5{,}140 = 5{,}1407\,.$$

### 3.4 Proprieta' della media aritmetica

La media aritmetica ha una capacità informativa notevole (a meno che non sia calcolata per variabili di tipo enumerazione o conteggio; nel qual caso assume

valore puramente indicativo), sia se ricavata dalla definizione di Chisini, sia come media decisionale.

A prescindere dalla definizione da cui deriva, la media aritmetica gode di due importanti proprietà:

1) la somma degli scarti dei valori osservati dalla propria media aritmetica è sempre nulla:

$$\Sigma(x_i-M)=0$$

*dimostrazione*:

$\Sigma(x_i-M)= \Sigma x_i-nM=\Sigma x_i-\Sigma x_i=0$

2) la somma dei quadrati degli scarti dei valori dalla propria media aritmetica è un minimo rispetto alla somma dei quadrati degli scarti degli stessi valori da qualsiasi altra media:

$$\Sigma(x_i-M)^2=minimo$$

*dimostrazione*:

$\Sigma(x_i-k)^2= \Sigma[(x_i-M)+(M-k)]^2=\Sigma[(x_i-M)^2+(M-k)^2+2(x_i-M)(M-k)]=$

$=\Sigma(x_i-M)^2+n(M-k)^2+2(M-k)\Sigma(x_i-M)$

Poichè:

- $n(M-k)^2\geq 0$, essendo $(M-k)^2$ un quadrato ed $n$ una quantità positiva;
- $2(M-k)\Sigma(x_i-M)=0$, essendo $\Sigma(x_i-M)=0$ per la 1° proprietà;

allora $\Sigma(x_i-M)^2\leq\Sigma(x_i-k)^2$, dove l'uguaglianza si ha per $k=M$.

<u>esempi</u>:

| $x_{(1)}$ | $x_{(2)}$ | $x_{(3)}$ | $x_{(4)}$ | $x_{(5)}$ |
|---|---|---|---|---|
| 160 | 162 | 164 | 166 | 168 |

$M$=164 poiché è il termine centrale di una serie aritmetica con un numero di termini dispari. Infatti:

$$M = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{160+162+164+166+168}{5} = \frac{820}{5} = 164$$

*1° proprietà*

$\Sigma(x_i - M) = (160-164) + (162-164) + (164-164) + (166-164) + (168-164) = -4-2+2+4=0$

*2° proprietà*

$\Sigma(x_i - M)^2 = 16+4+4+16=40$

$k = 162 < 164$

$\Sigma(x_i - k)^2 = (160-162)^2 + (162-162)^2 + (164-162)^2 + (166-162)^2 +$

$+(168-162)^2 = 4+4+16+36=60 \quad \Rightarrow \quad 40<60$

$k = 166 > 164$

$\Sigma(x_i - k)^2 = (160-166)^2 + (162-166)^2 + (164-166)^2 + (166-166)^2 +$

$+(168-166)^2 = 36+16+4+4=60 \quad \Rightarrow \quad 40<60$

In forma tabellare:

| $x_i$ | $x_i - M$ | $(x_i - M)^2$ | $x_i - 162$ | $(x_i - 162)^2$ | $x_i - 166$ | $(x_i - 166)^2$ |
|---|---|---|---|---|---|---|
| 160 | -4 | 16 | -2 | 4 | -6 | 36 |
| 162 | -2 | 4 | 0 | 0 | -4 | 16 |
| 164 | 0 | 0 | 2 | 4 | -2 | 4 |
| 166 | 2 | 4 | 4 | 16 | 0 | 0 |
| 168 | 4 | 16 | 6 | 36 | 2 | 4 |
| totale | 0 | 40 | | 60 | | 60 |

La media aritmetica gode di altre proprietà.

Sia $X$ una variabile statistica con media $M_X$ .

Sia $Y$ una trasformazione lineare di $X$: $Y = \alpha + \beta X$; dimostriamo che:

$$M_Y = \alpha + \beta M_X$$

*dimostrazione*

$$M_Y = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (\alpha + \beta x_i)}{n} = \frac{n\alpha}{n} + \beta \frac{\sum_{i=1}^{n} x_i}{n} = \alpha + \beta M_X$$

Se $\beta=1$, $Y=\alpha+X$ ed è:

$$M_Y = \alpha + M_X$$

Ovvero, se la variabile $X$ subisce una traslazione, la media subisce la stessa trasformazione della variabile.

Se $\alpha=0$, $Y=\beta X$ , ovvero $X$ subisce solo un cambiamento di scala ed è:

$$M_Y = \beta M_X$$

Supponiamo, ad esempio, di aver rilevato la statura, in *m*, su 5 soggetti:

$X$: 1,50 1,60 1,70 1,80 1,90 $M_X$=1,70 *m*

Volendo disporre della media in *cm*, trasformiamo i dati da *m* in *cm*:

$Y$: 150 160 170 180 190 $M_Y$=170 *cm*

oppure possiamo trasformare direttamente $M_X$:

$M_Y = \beta M_X$ =100·1,70=170 *cm*.

Si consideri adesso una variabile statistica $X$ con media $M_X$ e una variabile statistica $Y$ con media $M_Y$. Sia $Z=X+Y$. Dimostriamo che:

$M_Z=M_X+M_Y$

*dimostrazione*

$$M_Z = \frac{\sum_{i=1}^{n} z_i}{n} = \frac{\sum_{i=1}^{n} (x_i + y_i)}{n} = \frac{\sum_{i=1}^{n} x_i}{n} + \frac{\sum_{i=1}^{n} y_i}{n} = M_X + M_Y$$