

Un po' di statistica

Christian Ferrari

Laboratorio di Matematica

1 Introduzione

La **statistica** è una parte della matematica applicata che si occupa della raccolta, dell'analisi e dell'interpretazione di dati osservati o misurati. Essa è fondamentale per tutti quegli ambiti che di basano su delle osservazioni, ad esempio la fisica o l'economia.

In statistica per descrivere, misurare o osservare dei fenomeni si introducono delle *variabili*, in questo senso la variabile è alla base dei metodi della statistica. Per esempio studiando il comportamento di un gas, le variabili possono essere: X la temperatura, Y la pressione. Un altro esempio potrebbe essere lo studio delle note conseguite da tutti gli allievi di I liceo nelle materie scientifiche, allora le variabili sono X nota di matematica, Y nota di fisica, W nota di chimica e Z nota di biologia.

Quando si osserva o si misura la variabile X un numero n di volte notiamo

$$x_1, x_2, \dots, x_n$$

questi *valori*. Se questi valori sono rappresentativi della variabile X si parla di *campione*, se invece n coincide con l'insieme di tutti i possibili individui si parla di *popolazione*. Ogni possibile valore è chiamato *modalità*.

In quello che segue supponiamo che lo studio del fenomeno si possa fare osservando interamente la collettività di individui (popolazione): si parla di **statistica descrittiva**; essa consiste in un insieme di metodi e tecniche per sintetizzare l'informazione contenuta nei dati. Gli strumenti di sintesi sono essenzialmente di tre tipi:

- tabelle;
- rappresentazioni grafiche;
- indici sintetici.

Quando sintetizziamo l'informazione contenuta nei dati, ne perdiamo una parte. Gli strumenti di sintesi devono essere scelti in modo tale da preservare, per quanto possibile, l'informazione rilevante per il problema analizzato ed eliminare l'informazione non necessaria.

2 Tabelle

Nella tabella sono riportate:

- le modalità della variabile;
- le frequenze associate a ciascuna modalità.

La **frequenza assoluta** n_i misura quante volte una certa modalità x_i è stata osservata nella popolazione studiata.

La **frequenza relativa** rappresenta la proporzione (talvolta in percentuale) di osservazioni che presentano una certa modalità della variabile analizzata

$$p_i = \frac{n_i}{n} (\times 100) \quad (1)$$

La **frequenza cumulata** assoluta N_i (relativa P_i) associata ad una modalità della variabile indica il numero (la proporzione) di osservazioni che presentano un valore minore o uguale rispetto a quello della modalità data.

Se siamo disposti a rinunciare ad una certa parte di informazione, la distribuzione di frequenza può essere costruita anche per variabili continue (ossia che assumono valori in un insieme continuo $I \subseteq \mathbb{R}$). Generalmente si opera nel modo seguente:

- si suddivide l'insieme dei valori che la variabile può assumere in intervalli, detti *classi*;
- si determina il numero di osservazioni che cadono all'interno di ciascuna classe.

Invece dei valori x_1, x_2, \dots, x_n avremo un insieme di classi, che noteremo in modo analogo. Come per i valori, in modo analogo, si rappresentano i dati in forma di tabelle per le classi. La costruzione delle classi vale anche se si vuole “compattare” i risultati di una variabile discreta.

Non esistono regole assolute per la costruzione delle classi. In generale:

- evitare di costruire classi con frequenze molto basse;
- modulare l'ampiezza delle classi in funzione della disponibilità di informazione “locale”;
- se possibile, non variare l'ampiezza di classe (semplifica l'interpretazione).

3 Rappresentazioni grafiche

Le rappresentazioni grafiche sono strumenti molto utili per visualizzare le caratteristiche di una variabile. Ne esistono di svariati tipi, a seconda delle esigenze di analisi. Alcuni riproducono le stesse informazioni di una distribuzione di frequenza, altri riassumono caratteristiche difficilmente rappresentabili mediante tabelle.

Come rappresentare la distribuzione di frequenza di una variabile continua o una variabile discreta “compattata”? Se le classi sono di ampiezza diversa, le frequenze non sono direttamente confrontabili. Per costruire un grafico che rappresenti in modo adeguato l'informazione è necessario eliminare l'effetto dell'ampiezza di classe. Il rapporto tra la frequenza e l'ampiezza (indicata con Δ_i) di una classe è detto **densità di frequenza**

$$d_i = \frac{p_i}{\Delta_i} \quad \text{oppure} \quad D_i = \frac{P_i}{\Delta_i} . \quad (2)$$

Le densità di frequenza sono fra loro confrontabili.

In un istogramma di frequenza ad ogni classe è associato un rettangolo, per l'ordinata dell'istogramma ci sono tre possibilità:

- la frequenza assoluta;
- la frequenza relativa;
- la densità di frequenza.

Da un istogramma è possibile determinare alcune rilevanti caratteristiche del fenomeno, per esempio:

- tendenza centrale;
- dispersione;
- grado di simmetria della distribuzione.

4 Indici sintetici

Le caratteristiche più rilevanti di una distribuzione, per esempio la tendenza centrale del fenomeno e il grado di dispersione possono essere rappresentate mediante numeri, detti *indici sintetici*.

Gli indici di posizione servono per individuare la *tendenza centrale* del fenomeno studiato. I più utilizzati sono:

- media aritmetica;
- moda;
- mediana.

La *dispersione* indica in quale misura i valori osservati differiscono da un valore di riferimento, ad esempio la media aritmetica. Gli indici di dispersione più utilizzati sono:

- la varianza;
- lo scarto quadratico medio o deviazione standard.

4.1 Media aritmetica

La *media aritmetica* \bar{x} è il più importante indice di posizione. La formula per il calcolo della media aritmetica è

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i \quad (3)$$

ossia la media è la somma dei valori osservati divisa per il numero di osservazioni. Ecco alcune proprietà della media aritmetica:

- la media aritmetica è sempre compresa tra il minimo ed il massimo dei valori osservati

$$x_{\min} \leq \bar{x} \leq x_{\max}$$

- la somma degli scarti dalla media è sempre pari a zero

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

4.2 Moda

La **moda** di una distribuzione è la modalità più frequente.

4.3 Mediana

La **mediana** \tilde{x} è il valore che occupa la posizione centrale nella distribuzione, tale che:

- metà delle osservazioni sono uguali o minori;
- metà delle osservazioni sono uguali o maggiori.

La mediana divide in due parti di ugual numero l'insieme dei valori osservati.

Per calcolare la mediana bisogna: (1) ordinare i valori osservati in ordine crescente, (2) prendere il valore centrale nella graduatoria ordinata. Il modo di procedere per il secondo punto varia a seconda del numero di osservazioni

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & \text{se } n \text{ è dispari} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n+2}{2}}) & \text{se } n \text{ è pari} \end{cases}$$

Infatti:

- se n è *dispari*, esiste un unico valore che divide esattamente in due la distribuzione. Il valore centrale occupa la posizione $\frac{n+1}{2}$;
- se n è *pari*, si considerano valori centrali quelli che occupano le posizioni $\frac{n}{2}$ e $\frac{n}{2} + 1$, si usa come mediana la semisomma dei valori centrali.

4.4 Varianza

Il grado di dispersione delle singole osservazioni è misurato dagli scarti

$$x_i - \bar{x}$$

un buon indice di dispersione deve essere una sintesi di queste quantità.

La **varianza** è definita da

$$v = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2 \quad (4)$$

Osserviamo che elevando al quadrato, trascuriamo il segno degli scarti.

4.5 Scarto quadratico medio o deviazione standard

Lo **scarto quadratico medio** o **deviazione standard** è la radice quadrata della varianza

$$s = \sqrt{v} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i^2 - \bar{x}^2)} \quad (5)$$

È l'indice più frequentemente utilizzato perché è espresso nella stessa unità di misura della variabile d'interesse.

L'interpretazione della deviazione standard si basa sul fatto empirico seguente: spesso, circa 2/3 dei dati si trova a una distanza minore di una deviazione standard dalla media; circa il 95% di essi si trova a meno di due deviazioni standard dalla media.

5 Esercizi

Esercizio 1

Le precipitazioni annuali a Ginevra Y tra il 1826 e il 1843 sono state le seguenti (in mm):

583 890 777 958 875 926 524 756 619
730 688 528 901 884 969 1258 850 939

1. Formando classi di larghezza $\Delta = 100$ mm, determina le frequenze assoluta e relativa, rappresenta poi i dati in forma tabellare.
2. Determina la densità di frequenza relativa e rappresenta i dati in un istogramma.
3. Determina la media aritmetica, la mediana e la deviazione standard.

Esercizio 2

La tabella seguente riporta il numero di aziende agricole secondo la superficie agricola totale, secondo le classi indicate.

Superficie	n_i	p_i	N_i	P_i	Δ_i	d_i	D_i
0 + 1	2406						
1 + 2	3404						
2 + 3	2857						
3 + 5	4415						
5 + 10	6856						
10 + 20	5708						
20 + 30	1365						
30 + 50	751						
50 + 100	410						
100+	238						

1. Completa la tabella.
2. Rappresenta i dati in un istogramma in modo tale che:
 - la base del rettangolo è pari all'ampiezza di classe;
 - l'altezza del rettangolo è pari alla densità di frequenza;
 - l'area del rettangolo è per costruzione la frequenza (assoluta o relativa) associata alla classe.

Esercizio 3

Su 100 soggetti è stato rilevato il gruppo sanguigno. I risultati sono stati riportati nella tabella seguente riuniti per gruppo

Gruppo	n_i	p_i
$x_1 = A$	20	
$x_2 = B$	5	
$x_3 = AB$	2	
$x_4 = 0$	23	
Totale		

Completa la tabella e determina la moda.

Esercizio 4

Le temperature medie in gradi °C a Ginevra X durante i mesi di gennaio dal 1950 al 1975 sono state le seguenti

0,7	2,7	0,6	-0,8	0,0	2,9	3,8	-0,2	1,5
2,4	1,6	2,6	3,8	-3,1	-1,0	1,5	0,5	1,5
0,3	2,1	0,4	-1,4	1,2	0,3	3,4	3,5	

1. Formando classi di larghezza $\Delta = 1^\circ\text{C}$, determina la frequenza relativa e rappresenta i dati in forma tabellare.
2. Determina la densità di frequenza e rappresenta i dati in un istogramma.
3. Determina la media aritmetica, la mediana e la deviazione standard.

Esercizio 5

La tabella seguente si riferisce all'altezza X rilevata su 20 soggetti.

1,82	1,84	1,71	1,75	1,81	1,83	1,74	1,72	1,77	1,81
1,72	1,82	1,68	1,75	1,66	1,68	1,73	1,61	1,75	1,65

1. Rappresenta i dati in una tabella, indicando la frequenza relativa.
2. Determina gli indici della tendenza centrale.
3. Determina gli indici della dispersione.

Esercizio 6

I dati seguenti rappresentano il carico massimo (in tonnellate) che possono sopportare dei cavi prodotti da una certa fabbrica

10,1	12,2	9,3	12,4	13,7	10,8	11,6	10,1	11,2	11,3
12,2	12,6	11,5	9,2	14,6	11,1	13,3	11,8	7,1	10,5

1. Rappresenta in un istogramma i dati, scegliendo delle classi appropriate.
2. Determina gli indici sintetici.
3. Rappresenta su di un grafico a dispersione xy la densità di frequenza cumulata.