

STATISTICA DESCRITTIVA

Introduzione

La statistica è la disciplina che studia la raccolta, l'organizzazione, l'elaborazione e l'interpretazione dei dati. *Lo scopo della statistica è lo studio di un fenomeno collettivo, cioè di situazioni singole ripetibili con caratteristiche comuni, riscontrando le regolarità che sono alla base del fenomeno collettivo.*

Comunemente si distingue tra **statistica descrittiva**, che tratta la raccolta, l'organizzazione e la descrizione sintetica dei dati, e **statistica inferenziale** (o deduttiva o induttiva o matematica: sono tutti sinonimi), che trae conclusioni probabilistiche dai dati usando concetti e metodi del calcolo delle probabilità.

Nel raccogliere i dati riguardo alle caratteristiche di un gruppo di individui o di oggetti, è spesso impossibile o poco pratico osservare l'intero gruppo, specialmente se è grande. Invece di analizzare l'intero gruppo, chiamato **popolazione** o universo, si esamina una piccola parte della popolazione, chiamata **campione**. Esempi di popolazione sono: gli alunni di una classe; gli abitanti di una città in una determinata data; i bulloni prodotti in un giorno in una giornata precisa.

Le caratteristiche della popolazione (e del campione) che vengono prese in considerazione prendono il nome di variabili. Le **variabili** possono essere: - quantitative (quando sono espresse da un numero, es. il numero dei figli, l'età, il peso) - qualitative (quando non possono essere espresse da un numero, es. il colore degli occhi, lo sport praticato). Se una variabile quantitativa può assumere qualunque valore fra due dati valori è detta **variabile continua** (es. l'altezza di una persona), altrimenti è detta **variabile discreta** (es. il risultato del lancio di un dado: 1,2,3,4,5,6).

Il momento dell'elaborazione dei dati è veramente importante, in essa i dati acquisiti verranno catalogati, trascritti in apposite tabelle, per poi infine rappresentarli graficamente nei più svariati modi per poterli, così, **interpretare**. A seconda dei dati statistici raccolti, possiamo creare vari tipi di grafici e tabelle. Vari tipi di grafici: lineare (FIG.1a), istogramma (o diagramma a rettangoli, FIG.1b), diagramma circolare (o a torta, FIG.1c).

Distribuzione di frequenze

Quando si vogliono riassumere grandi quantità di dati grezzi, è spesso utile distribuire i dati stessi in classi e determinare il numero di individui appartenenti a ciascuna classe (**frequenza della classe**). Un ordinamento tabulare dei dati secondo le classi e secondo le corrispondenti frequenze delle classi è detto **distribuzione di frequenze**. Più sotto è riportata una tabella che è una distribuzione di frequenza dei pesi (arrotondati al kg più prossimo) di 100 studentesse di una scuola (TAB.1). La prima classe comprende i pesi da 45 kg a 47 kg ed è indicata con il simbolo 45-47. Poiché 5 studentesse hanno pesi appartenenti a questa classe, la frequenza della classe è 5. La dicitura 45-47 è detta intervallo della classe ed i numeri 45 e 47 sono detti limiti della classe (limite inferiore e superiore). I dati riassunti ed ordinati come in tabella sono detti dati raggruppati.

Anche se il procedimento di raggruppamento distrugge molte informazioni sui dati originali, si ha il vantaggio di ottenere una maggiore sintesi delle relazioni tra i dati stessi e quindi una migliore interpretazione.

Data una distribuzione di frequenze, la sua rappresentazione grafica è l'**istogramma** o il poligono (vedi TAB.2 e FIG.2). La **frequenza relativa** di una classe è la frequenza divisa per il totale delle frequenze di tutte le classi e viene spesso espressa in percentuale. La somma di tutte le frequenze relative è 1, cioè il

100%. La tabella risultante e' detta distribuzione di frequenze relative che sara' rappresentata da istogrammi percentuali e poligoni di frequenze relative. La frequenza totale di tutti i valori inferiori al confine superiore di una data classe e' detta **frequenza cumulata**. La tabella che presenti le frequenze cumulate e' detta **distribuzione di frequenze cumulate** o **distribuzione comulata (o cumulativa)**. Il grafico che la rappresenta e' il poligono di frequenze cumulate che ha come ultimo valore il numero totale dei membri del campione (vedi TAB.3 e FIG.3). Spesso e' volentieri si usa la **distribuzione di frequenze cumulate relative (o percentuali)**, rappresentata dal poligono di frequenze cumulate relative (o percentuali) che ha come ultimo valore il valore 1 (o 100%), detto anche **ogiva**. Vedi Tabella e Figura.

Supponiamo di avere un campione vastissimo: si possono scegliere delle classi molto piccole. Ne consegue che il poligono di frequenze diventa la **curva di frequenza**, che sono poi spesso studiate con un approccio matematico. Le curve di frequenza possono avere varie forme (vedi FIG.4).

Indici di posizione

Dato un insieme di dati (X_1, X_2, \dots, X_N), per studiarlo si usano dei parametri che possano essere rappresentativi dell'intero insieme (e' piu' semplice confrontare i valori di alcuni parametri piuttosto che tabelle o curve di frequenza).

Il parametro primario e' un parametro per indicare la posizione del valore "tipico". Vari sono gli indicatori di posizione usati:

MEDIA (ARITMETICA):

$$\langle X \rangle = \bar{X} = \frac{X_1 + X_2 + \dots + X_N}{N} = \frac{\sum_{i=1}^N X_i}{N}. \quad (1)$$

MEDIA PESATA, cioe' dati i dati e i pesi (W_1, W_2, \dots, W_N): $\tilde{X} = \frac{\sum_{i=1}^N W_i X_i}{\sum_{i=1}^N W_i}$; un caso particolare e' fare la media usando la distribuzione di frequenze: supponiamo di avere N dati X distribuiti in n valori (Y_1, Y_2, \dots, Y_n) con frequenze (f_1, f_2, \dots, f_n), invece di fare la media degli N dati X , si puo' fare: $\langle X \rangle = \frac{\sum_{i=1}^n f_i Y_i}{\sum_{i=1}^n f_i}$.

MEDIANA: si ordinano i dati, il valore centrale (o la media aritmetica dei due valori centrali) e' il valore mediano. Data una curva di frequenza, la mediana divide l'area sotto la curva a meta'. I quartili, decili o percentili sono i valori che dividono l'area sotto la curva in 4, 10, 100 parti.

MODA: e' il valore piu' frequente, piu' comune. Data una curva di frequenza il valore di picco e' la moda.

Se la curva e' simmetrica (come la gaussiana), media, mediana e moda coincidono, in FIG.5 vi e' un caso in cui differiscono.

Indici di dispersione

Un secondo parametro importante e' la misura di quanto i dati si dispongano vicini o lontani dal valore medio. Si chiama scarto il valore $V_i = X_i - \bar{X}$.

CAMPO DI VARIAZIONE: e' la differenza fra il valore minimo e massimo dei dati.

SCARTO QUADRATICO MEDIO:

$$s = \sqrt{\left(\sum_{i=1}^N (X_i - \bar{X})^2\right)/(N-1)} = \sqrt{\left(\sum_{i=1}^N (V_i)^2\right)/(N-1)}; \quad (2)$$

si noti che gli scarti sono sommati in quadratura in modo da costituire una quantita' positiva (la somma degli scarti avrebbe dato zero!). In tutta la popolazione si sarebbe definito $\sigma = \sqrt{(\sum_{i=1}^N (X_i - \langle X \rangle)^2)/(N)}$. Il valore s del campione cosi' definito, cioe' dividendo per $N-1$, e' la stima migliore dello scarto quadratico medio σ della popolazione. Questo risultato fa parte della "inferenza statistica" che qui non trattiamo e noi useremo il simbolo σ , detto anche **dispersione**, anche per il campione (si noti che per N grandi le definizioni coincidono). La **varianza** e' pari a s^2 (o σ^2). Piu' σ e' grande, piu' i dati sono discosti dal valore medio!

Indici successivi

Si definisce momento di ordine r della variabile X : $\overline{X^r} = (\sum_{i=1}^N X_i^r)/N$ e momento di ordine r dalla media aritmetica \bar{X} :

$$m_r = \left(\sum_{i=1}^N (X_i - \bar{X})^r\right)/N = \overline{(X - \bar{X})^r}. \quad (3)$$

Se $r = 1$, $m_1 = 0$. Se $r = 2$, $m_2 = \sigma^2$.

Il momento di ordine 3 e' legato alla simmetria della curva e il momento di ordine 4 fornisce informazioni su quanto e' piccata la curva (curtosi, si veda FIG.6). L'asimmetria e' $= m_3/\sqrt{m_2^3}$ e la curtosi e' $= m_4/m_2^2$.

Vari tipi di grafici

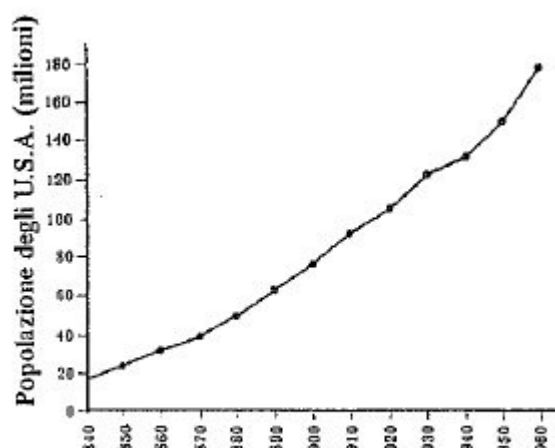


FIG.1a

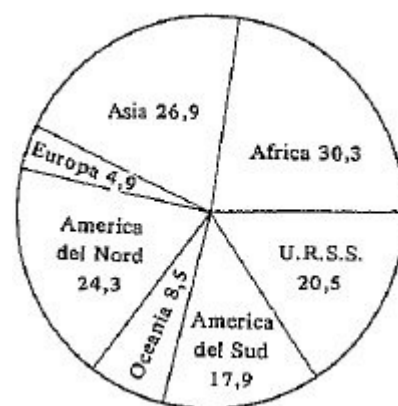
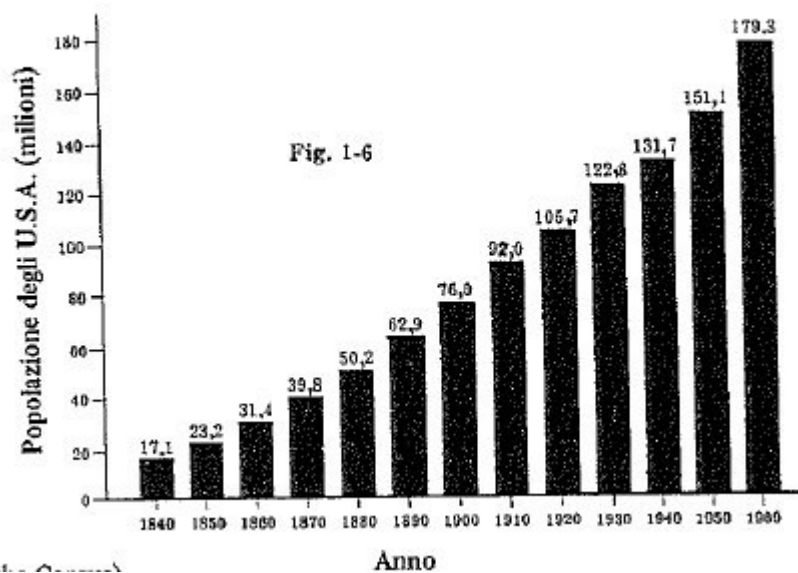


FIG.1b

FIG.1c

Tabella TAB.1 a cui si riferisce descrizione nel testo.

Peso in KG	Numero di studente sse
45 – 47	5
48 – 50	18
51 – 53	42
54 – 56	27
57 – 59	8
	100

Tabella 2 e Istogramma delle frequenze (TAB.2 e FIG.2)

Massa (in kilogrammi)	Numero di studenti
60 – 62	5
63 – 65	18
66 – 68	42
69 – 71	27
72 – 74	8
Totale 100	

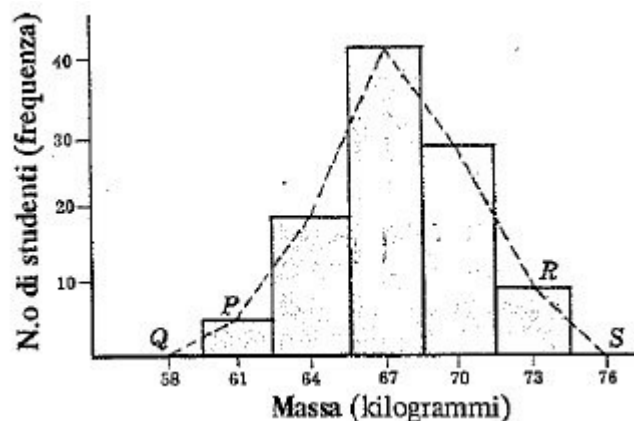


Tabella 3 e Grafico delle frequenze cumulate (TAB.3 e FIG.3)

Tabella 2.2

Massa (kilogrammi)	Numero di studenti
fino a 59,5	0
fino a 62,5	5
fino a 65,5	23
fino a 68,5	65
fino a 71,5	92
fino a 74,5	100

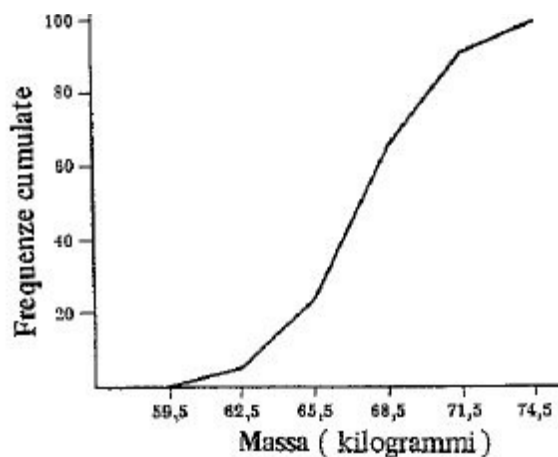
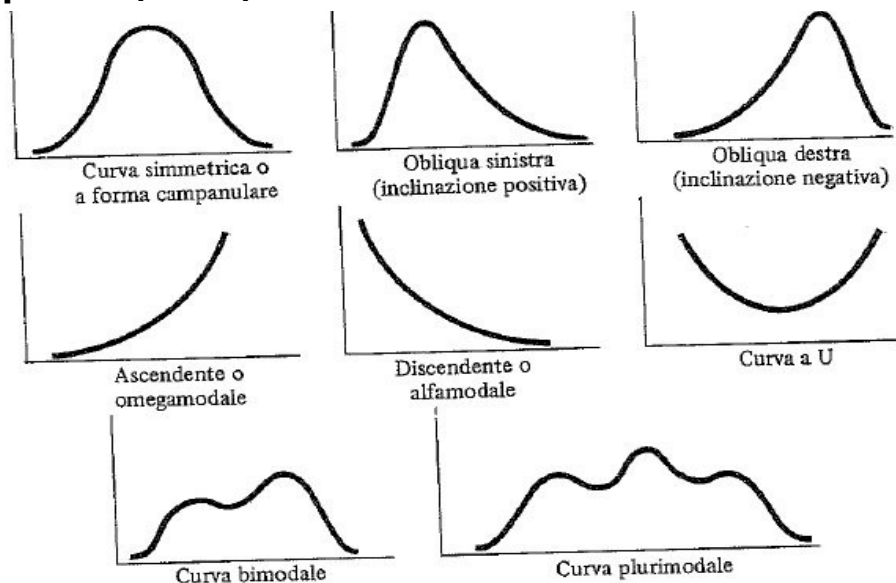
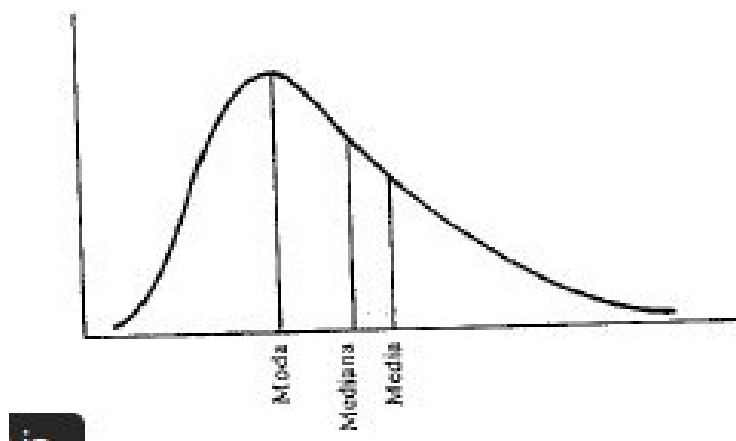


Fig. 2-2

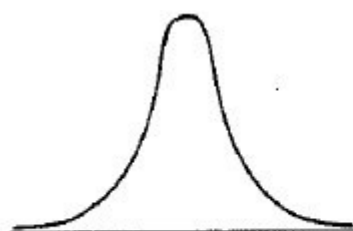
curve di frequenza (FIG.4)



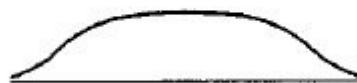
Cfr. Media, moda, mediana (FIG.5)



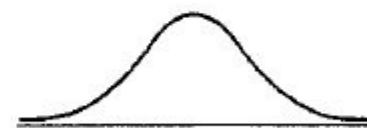
curve con curtosi differenti (FIG.6)



(a) Leptocurtosi



(b) Platicurtosi



(c) Mesocurtosi