Distribuzioni statistiche

L'operazione di determinazione delle modalità del carattere per ciascuno degli elementi del collettivo origina una *distribuzione del collettivo secondo il carattere considerato*.
Il nome "distribuzione" deriva dal fatto che mediante essa si indica come la modalità del carattere considerato si distribuiscono nelle unità del collettivo.

- La distribuzione secondo un carattere origina una <u>distribuzione semplice</u>;
- La distribuzione secondo due caratteri origina una <u>distribuzione doppia</u>;
- In generale, la distribuzione secondo più caratteri origina una <u>distribuzione multipla</u>

Distribuzioni unitarie

$$
\begin{array}{lccccccc}
\text{popolazione } \Omega & = & \{\omega_1, & \omega_2 & \omega_3, & \cdots, & \omega_N\} \\
 & & \downarrow & \downarrow & \downarrow & & \downarrow \\
\text{carattere } X & = & \{a_1, & a_2, & a_3, & \cdots, & a_N\}
\end{array}
$$

- La successione $a_1, \ldots, a_N$ costituisce una *distribuzione unitaria* di Ω secondo il carattere o *variabile statistica* X;
- a_i è la *modalità* con cui il carattere X si presenta nella unità statistica ω_i.
- Ovviamente nell'insieme $\{a_1, a_2, \ldots, a_N\}$ possono essere presenti più gruppi di elementi uguali, cioè vi possono essere più unità statistiche per cui il carattere X assume la stessa modalità.
- Esempio: redditi, elezioni politiche, residenza.

Distribuzioni di frequenze

Raggruppando fra loro gli elementi uguali di una distribuzione unitaria di N elementi si origina una *distribuzione di frequenze* (*assolute* o *relative*).

Nella popolazione Ω vi saranno:

n_1 elementi in cui il carattere X assume modalità x_1,

n_2 elementi in cui il carattere X assume modalità x_2,

...,

n_k elementi in cui il carattere X assume modalità x_k.

Si ottiene la *distribuzione di frequenze assolute* di X:

$$X = \begin{cases} x_1 \quad x_2 \quad \cdots \quad x_k & \leftarrow \quad \textit{modalità di } X \\ n_1 \quad n_2 \quad \cdots \quad n_k & \leftarrow \quad \textit{frequenze assolute} \end{cases}$$

Ovviamente risulta: $n_1 + n_2 + \cdots + n_k = N$.

Distribuzioni di frequenze relative

Sia assegnata una popolazione Ω e consideriamo la distribuzione di frequenze assolute di X:

$$X = \begin{cases} x_1 & x_2 & \cdots & x_k \\ n_1 & n_2 & \cdots & n_k \end{cases}$$

Si definisce *frequenza relativa* della modalità x_i

$$f_i = \frac{n_i}{N} \qquad \text{o, in percentuale} \qquad f_i\% = \frac{n_i}{N} \cdot 100\%$$

Si ottiene pertanto una *distribuzione di frequenze relative*:

$$X = \begin{cases} x_1 & x_2 & \cdots & x_k \\ f_1 & f_2 & \cdots & f_k \end{cases}$$

Per ciascuna modalità x_i, questa distribuzione fornisce la frazione f_i dei casi in cui è stata osservata tale modalità.

Ovviamente risulta: $f_1 + f_2 + \cdots + f_k = 1$.

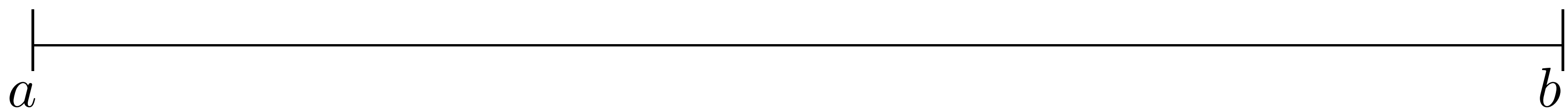
Esempio

Consideriamo la seguente distribuzione delle famiglie italiane per numero di componenti al Censimento 1991:

Componenti (x_i)	n_i	$f_i\%$
1	4.099.970	20,6
2	4.920.050	24,7
3	4.410.961	22,2
4	4.228.722	21,2
5	1.576.409	7,9
6	474.343	2,4
7 e più	198.548	1,0
Totale	19.909.003	100,0

Distribuzioni per classi di valori

Sia X un carattere quantitativo continuo in base al quale studiare una popolazione Ω. Poichè il carattere è continuo, il carattere può assumere un qualunque valore in un certo intervallo $[a, b] \subseteq \mathbb{R}$.



Allora si divide l'intervallo $[a, b]$ in più sottointervalli:

ed associamo ad ogni unità statistica ω_i il sottointervallo in cui ricade il valore assunto dal carattere X in ω_i.
Si origina così una *distribuzione per classi*:

$$X = \begin{cases} [a_0, a_1[& [a_1, a_2[& \cdots & [a_{k-1}, a_k] \\ n_1 & n_2 & \cdots & n_k \end{cases}$$

Esempio

La tabella seguente riporta la popolazione residente in Italia per classi di età (Censimento 1991):

Classi di età	$f_i\%$
fino a 5 anni	4,9
5 - 14	11,0
15 -19	7,6
20 - 39	30,1
40 - 59	25,3
60 - 74	14,4
75 e oltre	6,7

Distribuzioni cumulate

Un carattere quantitativo può essere assegnato mediante una *distribuzione cumulata di frequenze* che fa corrispondere a ciascuna modalità x_i il numero $n_1 + \cdots n_i$ di unità statistiche per cui il carattere X assume un valore al più uguale a X:

$$X = \begin{cases} x_1 & x_2 & x_3 & \cdots & x_k \\ n_1 & n_1 + n_2 & n_1 + n_2 + n_3 & \cdots & N \end{cases}$$

o, in termini di frequenze relative

$$X = \begin{cases} x_1 & x_2 & x_3 & \cdots & x_k \\ f_1 & f_1 + f_2 & f_1 + f_2 + f_3 & \cdots & 1 \end{cases}$$

Nel seguito porremo:

$$N_i = n_1 + n_2 + \cdots + n_i \; = \; N_{i-1} + n_i$$

$$F_i = f_1 + f_2 + \cdots + f_i \; = \; F_{i-1} + f_i$$

Esempio

Consideriamo la seguente distribuzione delle famiglie italiane per numero di componenti al Censimento 1991:

Componenti (x_i)	n_i	N_i	f_i	F_i
1	4.099.970	4.099.970	0,21	0,21
2	4.920.050	9.020.020	0,25	0,45
3	4.410.961	13.430.981	0,22	0,67
4	4.228.722	17.659.703	0,21	0,89
5	1.576.409	19.236.112	0,08	0,97
6	474.343	19.710.455	0,02	0,99
7 e più	198.548	19.909.003	0,01	1,00
Totale	19.909.003		1,00	

Esempio

Distribuzione di 890 utenti internet nel 1996 secondo l'età:

Età	Numero di utenti n_i	Freq. relativa f_i (in %)	Distr. cumulativa F_i (in %)
meno di 20	30	3.4	3.4
20 - 24	107	12.0	15.4
25 - 29	213	23.9	39.3
30 - 34	217	24.4	63.7
35 - 39	139	15.6	79.3
40 - 44	95	10.7	90.0
45 - 49	50	5.6	95.6
50 - 54	28	3.2	98.8
55 e più	11	1.2	100.0
Totale	890	100.0	

Esempio

Consideriamo la seguente distribuzione di $N = 290$ aziende parmensi per numero di postazioni Internet:

numero postazioni	n_i	f_i
0	56	19,31%
1	103	35,52%
2-5	60	20,69%
6-10	41	14,41 %
> 10	28	9,66%
mancata risposta	2	0,69%
Totale	290	100,00%

Densità di frequenza 1/2

Nel caso di distribuzioni per classi di valori, una quantità importante è la *densità di frequenza* o *frequenza specifica della classe i-esima* d_i data dal rapporto fra la frequenza relativa della classe e la sua ampiezza:

$$d_i := \frac{f_i}{\Delta_i} \ .$$

Densità di frequenza 2/2

La seguente tabella riporta la distribuzione dei redditi, in milioni di lire, dichiarati ai fini IRPEF, dei lavoratori dipendenti nel 1995:

Classi di età	Δ_i	n_i	N_i	f_i	F_i	d_i
0-9	9	1.787	1.787	0,12	0,12	0,0138
9-15	6	1.310	3.097	0,09	0,22	0,0152
15-19	4	972	4.069	0,07	0,28	0,0169
19-25	6	2.753	6.822	0,19	0,48	0,0320
25-35	10	4.227	11.049	0,29	0,77	0,0295
35-50	15	2.174	13.223	0,15	0,92	0,0101
50-100	50	920	14.143	0,06	0,99	0,0013
100-150	50	138	14.281	0,01	1,00	0,0002
150-300	150	54	14.335	0,00	1,00	0,0000
300 e oltre	150	9	14.344	0,00	1,00	0,0000
Totale		14.344		1,00		

Funzione di ripartizione o cumulata

Una variabile statistica X con distribuzione $\{f_i\}_i$ può essere caratterizzata anche in termini della cosidetta *funzione di distribuzione cumulata di frequenza* o *funzione di ripartizione*.
Sia X una v.s. discreta con distribuzione $\{(x_1, f_1), \ldots, (x_k, f_k)\}$ definita su Ω, si definisce *funzione cumulativa di frequenza* la funzione reale di variabile reale $F : \mathbb{R} \to [0, 1]$ definita per ogni $x \in \mathbb{R}$ come segue:

$$F(x) := \sum_{x_j \leq x} f_j \ .$$

Funzione di ripartizione - Proprietà

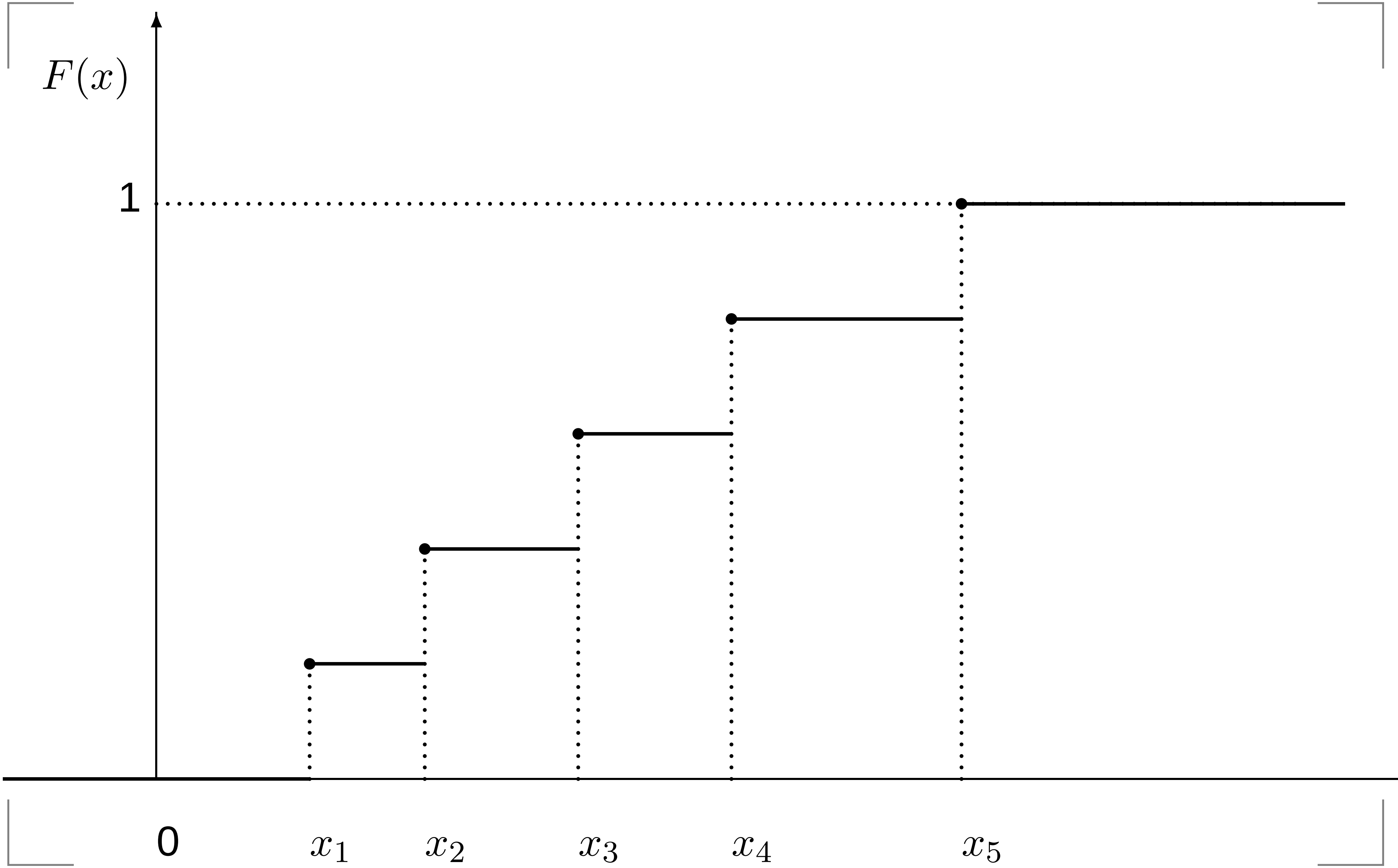
Indicate con $x_1, \ldots, x_k$ le modalità di una variabile statistica X, con $x_1 < x_2 < \cdots < x_k$, la funzione di ripartizione di X presenta tre proprietà:

i. $F(x) = 0$ per $x < x_1$,

ii. $F(x) = 1$ per $x > x_k$,

iii. $F(x)$ è non decrescente in $\mathbb{R}$.

sempio di diagramma di funzione cumulat



Rappresentazioni grafiche

Hanno lo scopo di fornire <u>immediatamente</u> a chiunque le caratteristiche essenziali del fenomeno oggetto dell'indagine.

- Tali rappresentazioni sono basate essenzialmente su una proporzionalità fra le frequenze (assolute o relative) e delle grandezze geometriche (aree o lunghezze) che vengono utilizzate per rappresentare il fenomeno.
- Non esistono regole fisse generali per la scelta della rappresentazione grafica con cui sintetizzare una distribuzione. L'importante è che venga assicurata l'immediata percezione del fenomeno in esame.
- Le rappresentazioni grafiche possono essere d'aiuto per scoprire relazioni fra le caratteristiche di distribuzioni.

Rapp. grafiche di caratteri qualitativi

- Diagramma a settori circolari o a torta in cui, a ciascuna modalità x_i si associa un settore circolare avente area proporzionale alla sua frequenza f_i.
- Diagramma a canne d’organo in cui a ciascuna modalità x_i si associa un rettangolo avente base costante ed un’altezza proporzionale alla frequenza f_i.
- Diagrammi figurativi in cui si utilizzano delle figure per rappresentare la distribuzione in esame: ciascuna figura rappresenta una modalità e la sua dimensione è proporzionale alla sua frequenza. (maneggiare con cura!)

Rapp. grafiche di caratteri quantitativi 1/2

- Istogramma (caratteri continui): plurirettangolo avente basi proporzionale all'ampiezza delle classi e aree proporzionali alla frequenza.
  Nota: Poichè nell'istogramma le aree dei singoli rettangoli sono proporzionali alle frequenze delle rispettive classi, l'altezza h_i del rettangolo della classe i-esima deve essere proporzionale al rapporto fra la frequenza della classe e la corrispondente ampiezza:

$$h_i \quad \propto \quad \frac{f_i}{x_{i+1} - x_i}$$

  La quantità h_i viene anche chiamata *frequenza specifica della classe i-esima*.

Rapp. grafiche di caratteri quantitativi 2/2

- Poligono di frequenza è una poligonale che unisce i punti centrali delle basi superiori dei rettangoli dell'istogramma.
- Diagrammi a segmenti (caratteri discreti): grafico cartesiano in cui, in corrispondenza di ciascuna modalità x_i, si riporta un segmento di altezza proporzionale alla corrispondente frequenza relativa (f_i) oppure frequenza assoluta (n_i)