Lezione 3: Distribuzioni di frequenza e rappresentazioni grafiche

Corso di Statistica
Facoltà di Economia
Università della Basilicata

Prof. Massimo Aria

- aria@unina.it

Distribuzione unitaria

Una **distribuzione statistica** consiste nell'insieme delle risposte assunte da un carattere statistico osservato su un dato collettivo.

Si immagini di aver osservato o rilevato sulle N unità statistiche della popolazione una variabile X le cui modalità sono risultate essere (x_1, x_2,…, x_l,…, x_N).
Tale insieme di dati prende il nome di **distribuzione unitaria** della variabile X.

Il pedice "l" individua l'unità statistica sul quale è stato rilevata la variabile X. Ne deriva che con x_l si indica la modalità assunta dalla variabile X per l'elle-sima unità statistica (con l=1,2,….,N).

Una distribuzione unitaria, pur essendo un'informazione esaustiva sul fenomeno, non consente una immediata individuazione delle caratteristiche salienti dello stesso: massimo, minimo, modalità più frequente, ecc.
In altre parole non è utile per fornire informazioni di sintesi.

Esempi di distribuzioni unitarie

Unità	Reddito
1	20345,54
2	15982,26
3	98456,20
4	10639,20
5	12983,52
6	9876,34
7	98121,00
8	12934,20
...	...
...	...
...	...
...	...
N	x_N

Unità	Residenza
1	Napoli
2	Milano
3	Milano
4	Firenze
5	Napoli
6	Roma
7	Napoli
8	Roma
...	...
...	...
...	...
...	...
N	x_N

Esempi di distribuzioni unitarie

Distribuzione di quantità

La **distribuzione di quantità** è una organizzazione dei dati in forma tabellare tale che per ogni modalità della variabile X si fa corrispondere la quantità totale misurata/rilevata sulle N unità della popolazione.

Essa esplicita, quindi, come l'ammontare complessivo del fenomeno si distribuisce tra le modalità del carattere X.

Il pedice "i" indica la generica modalità del carattere (con i=1,2,…,k).

Modalità	Quantità
x_1	A_1
x_2	A_2
...	
x_i	A_i
...	
x_k	A_k

Distribuzione di quantità

Esempio di distribuzione di quantità

Fabbricazione di elementi da costruzione in metallo	Quantità
Ponti ed elementi di ponti di ferro o acciaio	59.434
Torri e piloni di ferro o acciaio	784.948
Materiale per impalcature e simili di ferro o acciaio	548.739
Dighe, chiuse, porte di carico di ferro o acciaio	10.836
Pannelli costituiti da due lamiere di acciaio profilate con un'anima isolante	265.119
Altre costruzioni unicamente o principalmente di lamiera	670.376
Altre strutture di ferro o acciaio	2.386.128
Costruzioni e parti di costruzione di alluminio	272.532
Porte, finestre e loro intelaiature, stipiti e soglie di ferro o di acciaio	3.727.455
Porte, finestre e loro intelaiature, stipiti e soglie di alluminio	7.698.480

Distribuzione delle quantità prodotte (in numero di pezzi) nel settore degli elementi da costruzione in metallo (fonte ISTAT, anno 2007)

Distribuzione di frequenza

La **distribuzione di frequenza** è una organizzazione dei dati in forma tabellare tale che ad ogni modalità della variabile X si fa corrispondere la rispettiva frequenza.

In altre parole, la distribuzione di frequenza esplicita quante volte una determinata modalità si presenta nel collettivo oggetto di studio.

Essa è un modo sintetico per rappresentare le unità statistiche che assumono uguale modalità indicandone unicamente la frequenza di risposta.

Modalità	Frequenza assoluta
x_1	n_1
x_2	n_2
…	
x_i	n_i
…	
x_k	n_k

Distribuzione di frequenza

Distribuzione di frequenza per variabili discrete

Si immagini una popolazione composta da N unità su cui è osservata una variabile X che assume *k* distinte modalità (variabile discreta).

E' possibile rappresentare le osservazioni in una distribuzione di frequenza in cui:

• x_i consiste nella i-esima modalità della X

• n_i consiste nella i-esima **frequenza assoluta**
(il numero di volte che la modalità i si presenta nel collettivo osservato)

• $f_i = n_i/N$ consiste nella i-esima **frequenza relativa**
(la proporzione con cui la modalità i si presenta nel collettivo osservato)

Modalità	Frequenza assoluta	Frequenza relativa
x_1	n_1	$f_1 = n_1/N$
x_2	n_2	$f_2 = n_2/N$
…		
x_i	n_i	$f_i = n_i/N$
…		
x_k	n_k	$f_k = n_k/N$

Distribuzione di frequenza per variabili discrete

Frequenze assolute n_i

$$\sum_{i=1}^{k} n_i = N; \quad 0 \le n_i \le N$$

Frequenze relative $f_i = n_i/N$

$$\sum_{i=1}^{k} f_i = 1; \quad 0 \le f_i \le 1$$

Proprietà delle frequenze assolute e relative

Rappresentazioni grafiche per variabili discrete

Una prima sintesi grafica della distribuzione di una variabile discreta può essere ottenuta attraverso un **diagramma a barre** (o diagramma cartesiano).

Esso è costruito ponendo sull'asse delle ascisse le modalità della variabile X e sulle ordinate le frequenze (assolute o relative) corrispondenti ad ogni modalità.

Si ottiene così una rappresentazione detta "*a barre verticali*".

Alcune osservazioni:
• L'impiego delle frequenze assolute o relative non cambia la forma della distribuzione.
• Il ricorso alle frequenze relative è necessario se si vogliono confrontare due diverse distribuzioni
(es. distribuzione del numero di figli in Italia e in Francia).

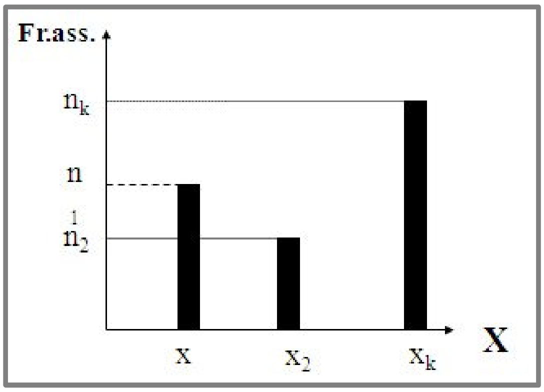
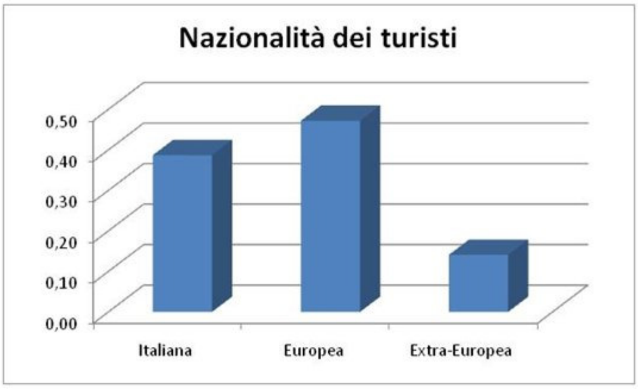


Diagramma a barre

Esempi di distribuzioni per variabili discrete

Nazionalità	Frequenze assolute	Frequenze relative
Italiana	1286	0,39
Europea	1568	0,47
Extra-Europea	469	0,14
Totale	3323	1,00



Rappresentazione tabellare e grafica della distribuzione della nazionalità
(Fonte: Risorsa Turismo, 2008)

Distribuzione di frequenza per variabili continue

Nel caso di una variabile continua non è possibile far corrispondere ad ogni modalità la rispettiva frequenza, in quanto il carattere potrebbe assumere infinite distinte modalità (ognuna delle quali avrebbe frequenza assoluta pari a 1)

Per fornire una rappresentazione tabellare di una variabile continua si ricorre quindi ad una **suddivisione in classi** delle modalità di risposta.

Ciò consente di determinare le frequenze assolute e relative delle classi di risposta in luogo delle singole modalità.

La suddivisione in classi

Si definisce una **generica classe** come:

$[x_{i-1}, x_i]$

"*in essa sono incluse tutte le modalità di X maggiori di x_{i-1} e minori o uguali a x_i*"

Si definisce **ampiezza di una classe** $[x_{i-1}, x_i]$,
la differenza tra l'estremo superiore e l'estremo inferiore della stessa:

$a_i = x_i - x_{i-1}$

I **criteri di suddivisione** delle modalità in classi sono i seguenti:

• *sulla base della conoscenza del fenomeno*
Con l'ausilio di esperti si determina la suddivisione che si ritiene maggiormente esplicativa dell'oggetto di studio

• *sulla base di regole empiriche*

- In classi di pari ampiezza (*equi-ampie*)
- In classi di pari frequenza (*equi-frequenti*)

Classi di modalità	Frequenza assoluta	Frequenza relativa	Ampiezza della classe
$x_0 \le x \le x_1$	n_1	$f_1 = n_1/N$	$x_1 - x_0$
$x_1 < x \le x_2$	n_2	$f_2 = n_2/N$	$x_2 - x_1$
…	…	…	…
$x_{i-1} < x \le x_i$	n_i	$f_i = n_i/N$	$x_i - x_{i-1}$
…	…	…	…
$x_{k-1} < x \le x_k$	n_k	$f_k = n_k/N$	$x_k - x_{k-1}$

Distribuzione di frequenza in classi

Costruzione di distribuzioni in classi equi-ampie

Classi equi ampie

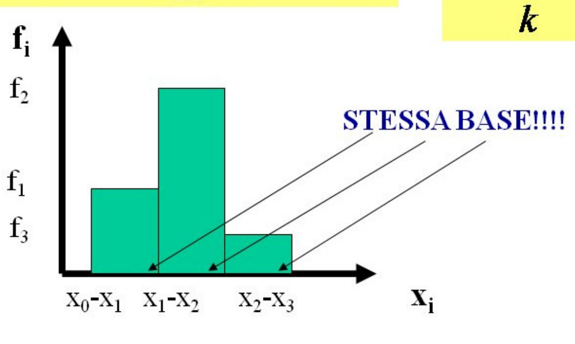
$$\frac{x_{\max} - x_{\min}}{k} \Rightarrow \text{ampiezza di ogni classe}$$

$k = \#$ classi (è scelto a priori!)

Esempio:

$x_{\max} = 100 \qquad x_{\min} = 20 \qquad k = 4$

$$\frac{x_{\max} - x_{\min}}{k} = \frac{100 - 20}{4} = 20$$



Ogni classe è di ampiezza 20!

Si tiene conto delle x_i ma NON delle frequenze!!!

Costruzione di distribuzioni in classi equi-frequenti

Classi equi frequenti

L'ampiezza di ogni classe

deve essere tale da contenere $\frac{N}{k}$ frequenze

k =# classi (è scelto a priori!)

Esempio:

$N = 600 \quad k = 6$

$$\frac{N}{k} = \frac{600}{6} = 100$$

Si tiene conto delle frequenze e NON delle x_i!!!

Ogni classe è di frequenza 100!

X	n_i
x_0—x_1	100
x_1—x_2	100
x_2—x_3	100
x_3—x_4	100
x_4—x_5	100
x_5—x_6	100

Densità di frequenza

Per la rappresentazione grafica di una distribuzione in classi, non è possibile utilizzare il grafico a barre in quanto le classi potrebbero avere diversa ampiezza e le frequenze non sarebbero quindi confrontabili.

In luogo delle frequenze si introduce il concetto di **densità di frequenza**.

Essa si definisce come il rapporto tra la frequenza di una classe e la rispettiva ampiezza:

$d_i = n_i / (x_i - x_{i-1})$

Classi di modalità	Frequenza assoluta	Ampiezza della classe	Densità di frequenza
$x_0 \le x \le x_1$	n_1	$a_1 = x_1 - x_0$	$d_1 = n_1/a_1$
$x_1 < x \le x_2$	n_2	$a_2 = x_2 - x_1$	$d_2 = n_2/a_2$
…	…	…	…
$x_{i-1} < x \le x_i$	n_i	$a_i = x_i - x_{i-1}$	$d_i = n_i/a_i$
…	…	…	…
$x_{k-1} < x \le x_k$	n_k	$a_k = x_k - x_{k-1}$	$d_k = n_k/a_k$

Densità di frequenza

Rappresentazioni grafiche per variabili continue

La rappresentazione grafica di una variabile continua avviene attraverso l'impiego dell'**istogramma**.

Esso fa corrispondere ad ogni classe un rettangolo la cui ampiezza è pari all'ampiezza della classe e la cui altezza è pari alla rispettiva densità di frequenza.

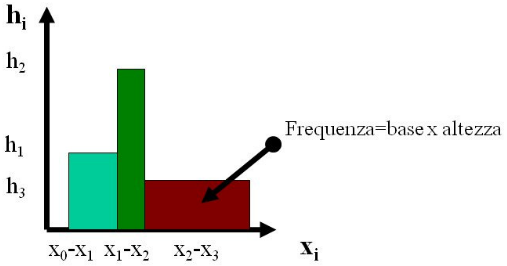
Le barre così ottenute hanno un'area pari alla frequenza assoluta delle corrispondenti classi e forniscono, quindi, una informazione non distorta sulla forma della distribuzione.

Alcune osservazioni:

- nei diagrammi a barre l'ampiezza della barra è convenzionale, mentre nell'istogramma indica l'ampiezza della classe.
- nei diagrammi a barre la frequenza è pari all'altezza della barra, mentre nell'istogramma essa è pari all'area della barra.
- Nel caso di una distribuzione in classi equi-ampie di ampiezza pari a 1, l'istogramma equivale ad un diagramma a barre.

$$a_i = x_i - x_{i-1} \Rightarrow \text{ampiezza (base)}$$

$$d_i = \frac{n_i}{a_i} \Rightarrow \text{densità (altezza)}$$



Densità di frequenza

Esempio di distribuzione di frequenza in classi

Km percorsi giornalmente	Frequenza assoluta n_i	Ampiezza a_i	Densità assoluta d_i
(2; 20]	15	18	0,833
(20; 30]	2	10	0,200
(30; 42]	3	12	0,250
Totale	*20*		

Esempi di distribuzione di frequenza in classi

Funzione di ripartizione empirica

La **funzione di ripartizione empirica** è una funzione che associa ad ogni valore reale x_i la proporzione di unità statistiche che assumono valori uguali o inferiori a x_i.

In pratica, la funzione di ripartizione empirica è ottenuta cumulando progressivamente le frequenze relative al crescere di X.

Il concetto di funzione di ripartizione trova applicazione nel contesto delle variabili quantitative.

Definizione

F_i = frequenza relativa cumulata

$$F_i = \sum_{h=1}^{i} f_h \qquad F(x_i) = \frac{\#(X \le x_i)}{N}$$

Proprietà

$0 \le F(x) \le 1$

$F(x)$ è non decrescente

$F(-\infty) = 0 \quad F(+\infty) = 1$

$F(x)$ è continua da destra

Definizione e proprietà della funzione di ripartizione empirica

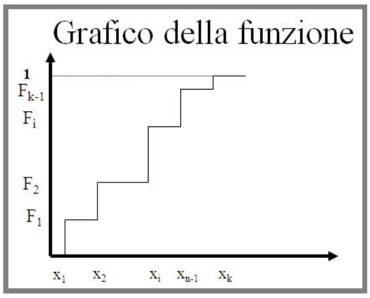


Grafico della funzione di ripartizione empirica

Nella prossima lezione

Nella prossima lezione si affronteranno i seguenti argomenti:

- concetto di centro di una distribuzione
- indici statistici di posizione