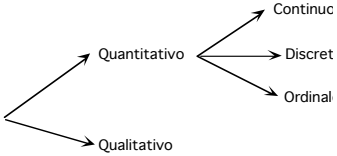
Lo spazio dei dati

Su ogni unità si rilevano "m" variabili $X_1,\ X_2,\ \ldots\ ,\ X_m$

Ogni variabile ha un suo dominio $S_1,\ S_2,\ldots,S_m$



Si possono analizzare in tutto "N" unità (ma N può essere infinito)

$$P = \{U_1, U_2, \ldots, U_N\}$$

P è la popolazione (o universo) formata da tutte e solo le unità di interesse di Una ricerca

Su ogni unità è possibile rilevare un insieme di "m" informazioni detto vettore della osservazione

$$X_i = (X_{i1}, X_{i2}, \ldots, X_{im}),\ i = 1,2,\ldots,N$$

Modello relazionale dei dati

Deriva dal concetto matematico di RELAZIONE

Noti gli insiemi $S_1,\ S_2,\ldots,S_m$ **coincidenti ognuno con un dominio**

"d" è una RELAZIONE se si configura come una "m-tupla" ordinata di valori

$$d = (d_1, d_2, \ldots, d_m)$$

tali che $d_1 \in S_1,\ d_2 \in S_2,\ \ldots,\ d_m \in S_m$

E' evidente che "d" coincide con una osservazione

"d" è un elemento del prodotto cartesiano di insiemi

$$D = S_1 \otimes S_2 \otimes \ldots \otimes S_m$$

Che costituisce lo SPAZIO DEI DATI

La matrice dei dati

Una rilevazione di dati consiste nella osservazione delle variabili sulle unità confluite del campione o della popolazione

Le osservazioni campionarie sono i vettori

$$X_i,\quad i = 1,2,\ldots,n$$

I cui valori disposti in ordine di acquisizione formano la MATRICE DEI DATI

ESEMPIO

Lo staff tecnico di una organizzazione è composto da 6 persone: Donne o uomini, laureate o no, residenti, vicini, fuori sede.

SPAZIO DEI DATI

(D, L, R, u1)	(D, L, V, u1)	(D, L, F, u1)	(D, L, R, u2)	(D, L, V, u2)	(D, L, F, u2)	(D, L, R, u3)	(D, L, V, u3)	(D, L, F, u3)
(D, N, R, u1)	(D, N, V, u1)	(D, N, F, u1)	(D, N, R, u2)	(D, N, V, u2)	(D, N, F, u2)	(D, N, R, u3)	(D, N, V, u3)	(D, N, F, u3)
(U, L, R, u1)	(U, L, V, u1)	(U, L, F, u1)	(U, L, R, u2)	(U, L, V, u2)	(U, L, F, u2)	(U, L, R, u3)	(U, L, V, u3)	(U, L, F, u3)
(U, N, R, u1)	(U, N, V, u1)	(U, N, F, u1)	(U, N, R, u2)	(U, N, V, u2)	(U, N, F, u2)	(U, N, R, u3)	(U, N, V, u3)	(U, N, F, u3)
(D, L, R, u4)	(D, L, V, u4)	(D, L, F, u4)	(D, L, R, u5)	(D, L, V, u5)	(D, L, F, u5)	(D, L, R, u6)	(D, L, V, u6)	(D, L, F, u6)
(D, N, R, u4)	(D, N, V, u4)	(D, N, F, u4)	(D, N, R, u5)	(D, N, V, u5)	(D, N, F, u5)	(D, N, R, u6)	(D, N, V, u6)	(D, N, F, u6)
(U, L, R, u4)	(U, L, V, u4)	(U, L, F, u4)	(U, L, R, u5)	(U, L, V, u5)	(U, L, F, u5)	(U, L, R, u6)	(U, L, V, u6)	(U, L, F, u6)
(U, N, R, u4)	(U, N, V, u4)	(U, N, F, u4)	(U, N, R, u5)	(U, N, V, u5)	(U, N, F, u5)	(U, N, R, u6)	(U, N, V, u6)	(U, N, F, u6)

Ciò che era possibile osservare

MATRICE DEI DATI

Persona	Sesso	Titolo	Residenza
u_1	D	L	F
u_2	D	L	V
u_3	D	L	V
u_4	D	L	R
u_5	D	N	R
u_6	U	N	V

Ciò che si è effettivamente osservato

Le dimensioni della matrice dei dati

La matrice dei dati ha dimensioni $(n\ x\ m)$

n è il numero di righe dove ogni riga (*record*) corrisponde ad una unità
m è il numero di colonne dove ognuna corrispondente ad una variabile

Matrice dei dati = *data set*

Insieme strutturato di informazioni

n=20
m=5

indagine sul *self-service*

meta-dato

Nome	N.libri	Tempo	Posiz.	Corso	Giudizio
A.C.	6	6	Coll.		Medio
A.R.	10	6	4	DES	Medio
A.G.	6	11	Doc		Pessimo
A.T	5	1	FC	EA	Medio
D.I.	6	5	Dlp		Pessimo
D.S.	7	8	FC	SSA	Medio
F.D.	11	5	Doc		Ottimo
G.A.	1	4	2	DUS	Ottimo
G.G.	10	1	3	DES	Buono
G.L.	2	1	Est.		Medio
G.P.	8	6	4	SSA	Pessimo
G.S.	4	12	Imp		Cattivo
L.F.	2	7	1	EA	Cattivo
M.B.	8	8	Doc		Pessimo
M.P.	8	3	3	DEAI	Ottimo
P.A.	5	5	4	SSA	Medio
P.C.	8	2	FC		Medio
R.B.	6	4	2	DES	Cattivo
R.T.	1	4	2	EA	Buono
S.B.	5	2	Doc		Ottimo

Esempio di data set su foglio elettronico

Variabili e dati sul Piano integrato Territoriale (PIT) "Serre vibonesi"
N=24, m=13

Codice	NOME	SUP	POPRES	DENS99	VEC98	DIP98	LUADIP	TANALF	VPR9981	TIM	TIN	IMPRLA	TIMPR	DENSOC
101	Acquaro	2532	3018	119.2	104.1	63.7	13.4	14.6	-8.4	-14.4	2.2	47.2	29.4	38.2
106	Arena	3235	1983	61.3	102.1	62.5	15.0	14.3	-15.2	-4.6	0.2	47.9	24.8	30.6
114	Brognaturo	2450	801	32.7	82.5	66.2	16.5	4.9	-0.2	-10.2	3.8	42.0	25.2	57.9
116	Capistrano	2094	1244	59.4	118.9	61.8	8.0	15.7	-4.2	-6.4	0.6	42.1	22.0	26.6
141	Dasa'	619	1378	222.6	164.8	61.1	5.5	11.4	-14.0	-5.7	-2.9	45.4	50.2	71.2
144	Dinami	4406	3222	73.1	68.7	60.0	7.6	12.3	-0.9	-8.3	6.5	59.9	33.5	47.0
146	Fabrizia	3878	2776	71.6	95.8	63.7	6.0	15.6	-17.0	-15.0	2.4	55.5	39.5	58.4
149	Filadelfia	3048	6742	221.2	109.2	57.3	11.1	12.2	-20.6	-24.0	3.5	47.8	35.5	57.9
151	Filogaso	2369	1390	58.7	58.2	53.3	9.5	10.0	18.2	-4.7	6.0	72.1	39.9	97.2
153	Francavill	2825	2670	94.5	95.0	56.4	7.8	9.1	-12.4	-17.3	2.6	33.4	16.3	26.6
157	Gerocarne	4493	2633	58.6	78.6	58.8	7.9	14.1	-12.9	-23.5	5.5	44.8	29.9	38.2
179	Mongiana	2070	848	41.0	86.8	63.8	10.4	10.8	-14.2	-19.1	2.8	44.8	26.8	51.4
182	Monterosso	1816	2063	113.6	147.3	58.5	18.7	9.5	-11.2	-6.9	-2.7	53.2	47.4	64.3
184	Nardodipac	3278	1532	46.7	97.0	63.7	4.7	14.8	-25.8	-17.2	3.2	26.7	15.0	23.4
198	Pizzoni	2323	1440	62.0	128.8	63.3	10.5	16.8	-19.8	-14.9	-2.5	51.8	25.2	36.4
200	Polia	3178	1290	40.6	153.0	78.5	14.6	15.7	-16.9	-16.9	-2.4	48.8	28.2	53.6
212	San Nicola	1932	1727	89.4	164.7	76.0	18.0	16.8	-11.0	-6.4	-3.8	46.1	27.1	35.4
228	Serra San	3958	6894	174.2	106.4	52.8	17.8	9.3	8.2	-2.1	5.3	54.0	44.5	72.6
232	Simbario	1925	1139	59.2	130.4	72.3	20.1	9.6	-20.5	-8.1	-0.6	57.6	28.8	36.2
235	Sorianello	972	1682	173.0	62.0	55.9	10.2	14.7	-0.6	-8.7	8.5	71.3	31.3	57.2
236	Soriano Ca	1517	3154	207.9	77.7	52.7	15.9	8.7	1.6	-9.2	5.9	103.3	65.8	110.9
240	Spadola	958	818	85.4	116.8	54.8	20.3	7.6	6.1	-0.5	-0.7	45.0	69.7	99.3
252	Vallelonga	1753	852	48.6	138.5	66.9	15.0	15.2	1.5	-1.2	-2.8	40.6	30.5	36.3
253	Vazzano	1985	1283	64.6	131.3	56.7	14.1	9.6	4.4	-0.8	-2.8	56.4	31.8	44.2

Le non risposte

In ogni indagine è presente una certa percentuale di unità che non collabora o che non è posta in grado di collaborare.

Sfuggono all'indagine per ragioni loro proprie (unità

Gli incaricati non riescono ad individuarle

Entrano in contatto tra di loro e concordano le rispo

Regolano le loro risposte in base a chi le contatta

La risposta dipende dalla forma di contatto: telefono, posta, rete, intervista, etc.

Per correggere queste anomalie bisogna eliminare delle unità dalla frame oppure -se possibile- ripetere l'acquisizione

I dati mancanti

I cosiddetti *missing values* sono quelli dovuti a non risposte o non coverages insanabili.

Derivano anche da mancata rilevazione o rilevazione manifestamente sbagliata.

L'elaborazione automatica dei dati non consente di lasciare dei vuoti nelle celle della matrice dei dati.

Per quelli che mancano si adotta un codice convenzionale che non può essere confuso con i dati rilevabili nella particolare indagine

ESEMPIO

Numero di permessi sindacali concessi da ammininistrazioni pubbliche.

Le sedi che non hanno risposto sono indicate con "-99"

E' anche interessante capire il perché dei "missing values"

Rilevazione campionaria univariata

133	197	165	214	188	237	188	115	128	213	120
204	**-99**	232	230	236	149	153	112	68	117	153
94	72	222	220	139	219	144	137	98	80	**-99**
209	93	181	249	200	128	82	**-99**	103	182	156
71	182	199	126	127	187	185	87	177	94	92
145	115	**-99**	203	233	64	227	88	67	243	240
204	156	118	**-99**	91	115	243	74	192	74	**-99**
197	245	235	88	141	116	168	204	62	**-99**	128
242	67	130	158	184	114	232	122	70	122	72

La codifica

Le denominazioni delle modalità sono talvolta lunghe o espresse con termini scomodi che complicano il ragionamento.

Si stabiliscono abbreviazioni (codifica) per facilitarne la trattazione informatica e saranno poi queste a comparire nella matrice dei dati.

ESEMPIO:

In una indagine internazionale sulla distribuzione dei redditi, il grado di copertura della popolazione di cui si sono considerate le entrata venne rilevata con il dominio S={NL, URB, NAG, RRL, AG} che sono abbreviazioni di *{national, urban, nonagricultural, rural, agricultural}*

La codifica è utile per sveltire le operazioni di trasferimento dei dati dai moduli con cui sono acquisite (questionari, schede di richiesta, fogli di controllo, etc.) e per limitare le sviste nella trascrizione.

I meta dati

Nel modello relazionale ogni data set è un insieme e in quanto tale

Nessuna unità (etichetta identificativa inclusa) può essere ripetuta

L'ordine con cui i dati sono inseriti nella relazione deve essere specificato con un attributo (key) collegato alla frame.

La chiave di accesso alla singola unità cioè il codice o l'insieme di codici che consentono di identificare il singolo dato sono dei meta dati cioè dati su dati.

I meta dati sono essenziali per accedere alle informazioni già raccolte in fonti ufficiali e disponili su supporto informatico

Chiave	X1	X2
A1	10	27
Z3	−1	0.3

Analisi univariata e multivariata

Ogni problema è una ragnatela: se si tocca un filo tutti gli altri vibrano. Lo stesso sucede per le variabili.

Lo studio univariato ha solo scopo didattico. Nella pratica i dati sono sempre multivariati

ESEMPIO: dove vanno gli studenti

		Stessa regione							
		Nord		Centro		Sud		Totale	
		numero	%	numero	%	numero	%	numero	%
		286555	83.6	178692	90.7	253887	74.7	719.124	81.8
	Nord Ovest	18783	5.5	1526	0.8	8378	2.5	28687	3.3
	Nord Est	27308	8.0	4749	2.4	11312	3.3	43369	4.9
Altre regioni	Centro	9149	2.7	9396	4.8	38800	11.4	57345	6.5
	Sud	929	0.3	2756	1.4	27296	8.0	30981	3.5
	Totale	56169	16.4	18427	9.3	85786	25.3	160382	18.2
	Italia	342724	100.0	197119	100.0	339663	100.0	879506	100.0

La lettura di una tabella a più variabili non è difficile. Lo è la generalizzazione dei risultati

Gli studi multidimensionali sono al momento rinviati. Faremo solo studi univariati.

Col presupposto che si possa avere l'idea di un concetto multilaterale studiando separatamente le sue componenti

Statistica descrittiva ed inferenziale

L'escussione delle unità rispetto alle variabili produce il data set

$$C = \{X_1,\ X_2,\ \ldots,\ X_n\}$$

cioè "m" osservazioni su di "n" unità

Si parla di STATISTICA DESCRITTIVA se il data set è analizzato per quello che è senza uno sfondo su cui proiettare i dati

Emittenti	Ascolti	Emittenti	Ascolti	Emittenti	Ascolti	Emittenti	Ascolti
Radiouno	7616	Radioverderai	791	R.D.S.	2671	Lattemiele	1145
Radiodue	6137	Isoradio	594	Rete 105	2607	Radio cuore	1135
Radiotre	1458	Radio deejay	3687	RTL 102.5	2112	Radio Maria	1105
Stereorai	1282	Radio italia SMI	3178	Radio Radicale	1541	Italia Network	1056
CNR	1468	Radio Montecarlo	1460	Radio Kiss Kiss	1393	Kiss Kiss Italia	972
105 Classic	786						

massimo per Radiouno; minimo per Isoradio; c'è un gruppo che si addensa intorno a 1000-1500 ascolti; Le reti pubbliche sono più diffuse di quelle commerciali

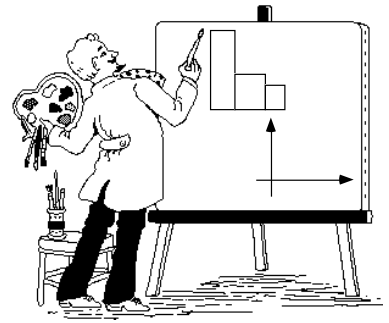
Statistica descrittiva ed inferenziale/2

La STATISTICA DESCRITTIVA mira alla organizzazione, all'analisi tabellare e grafica, al calcolo di grandezze sintetiche di ciò che si è rinvenuto nella rilevazione

E' anche nota come analisi esplorativa *(Exploratory Data Analisys)* proposta soprattutto da J.W. Tukey nel 1977

In breve si ocnfigura come una trattazione preliminare indispensabile per affrontare uno studio complesso.

Utilizza tecniche elementari, soprattutto grafiche, di grande efficacia nell'aiutare a comprendere l'esito della rilevazione



Statistica descrittiva ed inferenziale/3

Un data set è importante solo in quanto mezzo per studiare la popolazione. E' quello, ma poteva essere un altro

Teoricamente, ogni unità può ruotare in una particolare posizione del data set per cui è variabile ciò che osserviamo in ogni posizione

Inoltre, il valore presentato dall'unità dipende da un modello comportamentale che produce potenzialmente un valore diverso in ogni rilevazione con qualcuno più favorito di altri.

Ne consegue che ogni rilevazione (totale o parziale) dà una certa immagine delle variabili rilevate.

Quale corrispondenza ci si può aspettare con ciò che si osserverebbe se la rilevazione fosse ripetuta o effettuata con strumenti perfetti ed infallibili

Quale corrispondenza se fossero considerate tutte le possibili unità?

Statistica descrittiva ed inferenziale/4

A queste domande risponde la STATISTICA INFERENZIALE

Ci interessa determinare la durata massima del Premier.

Se tali unità fossero le uniche disponibili allora sono tutto ciò che serve per calcolare il massimo. Li ordiniamo in senso ascendente per ottenere: De Gasperi con Xmax=2808.

Lo stesso vale per la durata media:
Se tutti i premier fossero durati lo stesso numero di giorni quanto sarebbe durato ciascuno?
Un semplice problema algebrico risolto sommando le durate e dividendo per il numero degli addendi: 814.

Questa è STATISTICA DESCRITTIVA

Durata in giorni di un premier in Italia.

Premier	giorni	Premier	giorni
Andreotti	2669	Amato	300
Colombo	560	Berlusconi	226
Craxi	1351	Ciampi	353
De Gasperi	2808	Cossiga	441
De Mita	465	Dini	351
Fanfani	1660	Forlani	253
Moro	2277	Goria	260
Prodi	509	Leone	337
Rumor	1098	Pella	154
Scelba	511	Tambroni	123
Segni	1087	Zoli	408
Spadolini	521		

La statistica inferenziale inizia laddove il data set è visto come la punta di un iceberg cioè i dati sono solo una delle possibili realizzazioni e riguardano anche i premier che verranno.

In che modo ed in che misura possiamo estendere al futuro le misure calcolate sui valori passati?

Descrizione del data set

Fase essenziale di ogni ricerca statitstica è l'acquisizione di dati:

Al momento tralasciamo...

il modo in cui il data set è stato formato

i criteri con cui si sono scelte le variabili

L'attezione è limitata alla descrizione del *data set o collettivo statistico*

- **Sintesi tabellare e grafica**
- **Parametri rilevanti**

Presentazione dei dati

Dalla raccolta dei dati si esce con il PROSPETTO DI RACCOLTA: una disposizione righe per colonne di dati non ordinati.

Dal prospetto di raccolta occorre passare a modi di presentazione più semplici e comprensibili per mezzo delle operazioni di

SPOGLIO: Ordinamento dei dati + trattamento dei doppioni

- **Tabella statistica** — **una tabella a due colonne dove si riportano tutte e solo le modalità verificatesi con a fianco il numero di volte che si sono presentate**
- **Diagramma ramo -e-foglia** — **Si ordinano in modo crescente i dati e si trascrivono le cifre più grandi. Di riportano poi le cifre più piccole per un numero pari alla frequenza del dato**

Lo spoglio

SPOGLIO AUTOMATICO Se i dati sono molto numerosi lo spoglio dei dati avviene con il computer:

1) **CODIFICA:** definizione di una corrispondenza biunivoca tra le cifre numeriche e/o le denominazioni con un insieme di codici che ne facilita l'inputazione e riduce lo spazio che occupano.
2) **INPUTAZIONE:** monotona, ma delicata fase di trasferimento dei dati dal supporto su cui sono già registrati ai programmi di elaborazione.

SPOGLIO MANUALE Se i dati sono pochi o non si sa o non si vuole usare il computer si può procedere come segue:

a) **ORDINAMENTO:** i dati vengono disposti in ordine di grandezza crescente
b) **RIPETIZIONI:** si tiene conto dei doppioni con un segno di spunta: una "X" o una "V".

Esempio

Reddito procapite delle province italiane

8.6	8.0	7.8	7.0	8.8	9.9	9.3	10.2	8.4	8.0	8.3	7.7	8.6	7.1	6.1	8.3
9.0	7.3	8.0	8.8	6.7	6.4	10.2	7.6	5.9	8.3	9.3	6.4	6.7	6.9	6.9	6.9
6.6	7.5	7.1	9.0	7.2	8.5	7.8	9.3	9.6	8.7	9.5	8.7	6.1	6.4	6.8	6.6
3.8	5.3	3.4	7.2	5.0	4.7	3.8	3.9	4.3	4.3	4.3	3.9	3.7	5.3	6.0	5.9
3.9	5.9	4.5	4.2	5.0	7.6	4.5	4.2	6.6	9.3	7.5	5.3	4.7	5.8	6.4	
4.7	7.2	5.3	5.7	4.4	3.9	4.7	3.7	4.7	6.1	6.1	7.7	5.3	8.5	9.8	

FASE_1: ordinamento

3.4	3.9	4.3	4.7	5.3	5.8	6.1	6.4	6.8	7.1	7.3	7.8	8.3	8.7	9.3	9.8
3.7	3.9	4.3	4.7	5.3	5.9	6.1	6.6	6.9	7.2	7.6	8.0	8.4	8.7	9.3	9.9
3.7	3.9	4.4	4.7	5.3	5.9	6.1	6.6	6.9	7.2	7.6	8.0	8.5	8.8	9.3	10.2
3.8	4.2	4.5	4.7	5.3	5.9	6.4	6.6	6.9	7.2	7.7	8.0	8.5	8.8	9.3	10.2
3.8	4.2	4.5	5.0	5.3	6.1	6.4	6.7	7.0	7.3	7.7	8.3	8.6	9.0	9.3	
3.9	4.3	4.7	5.0	5.7	6.1	6.4	6.7	7.1	7.3	7.8	8.3	8.6	9.0	9.6	

Si individuano subito la modalità più piccola e la modalità più grande

$$X_{(1)} = 3.4; \quad X_{(95)} = 10.2$$

Esempio (continua)

Fase_2: eliminazione dei doppioni

X	cont.	Freq	X	Cont.	Freq	X	Cont.	Freq	X	Cont.	Freq
3.4	x	1	3.7	xx	2	3.8	xx	2	8.7	xx	2
3.9	xxxx	4	4.2	xx	2	4.3	xxx	3	9.3	xxxx x	5
4.4	x	1	4.5	xx	2	4.7	xxxx x	5	10.2	xx	2
5	xx	2	5.3	xxxx	4	5.7	x	1	8.7	xx	2
5.8	x	1	5.8	x	1	5.9	xxx	3	9.6	x	1
6.1	xxxx x	5	6.4	xxxx	4	6.6	xxx	3	9	xx	2
6.7	xx	2	6.8	x	1	6.9	xxx	3	9.9	x	1
7	x	1	7.1	xx	2	7.2	xxx	3	8.4	x	1
7.3	xxx	3	7.6	xx	2	7.7	xx	2	8.5	xx	2
7.8	xx	2	8	xxx	3	8.3	xxx	3	8.6	xx	2

In questa fase si accorpano i valori ripetuti che costituiscono la frequenza di una data modalità all'interno della serie dei valori campionari.

FREQUENZA = NUMERO DI VOLTE CHE SI PRESENTA

Le più presenti sono: 6.1, 4.7, e 9.3 che compaiono 5 volte

Il risultato è insoddisfacente. Ci sono troppe informazioni nella tabella

Spoglio e conteggi

Dobbiamo trovare forme semplificate di presentazione dei dati

Esempio: Survey di baccelli di Indigofera per numero di semi

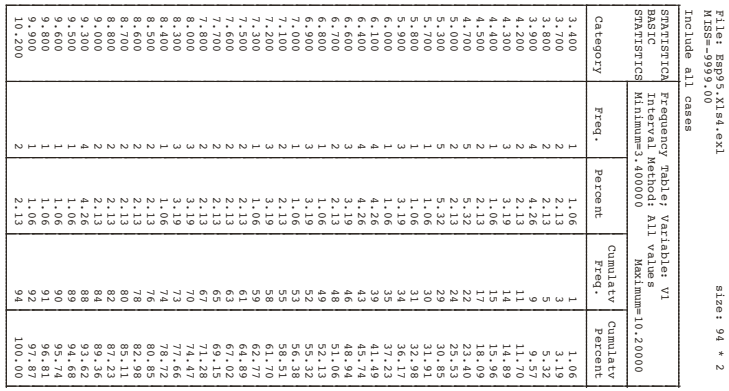
3	6	5	9	10	12	5	3	9
8	7	10	7	8	9	7	6	5
9	8	7	6	9	9	9	6	12
11	8	7	9	11	8	10	9	8
4	7	6	5	9	8	7	9	8
8	8	7	8	9	6	7	8	9
9	5	10	9	7	6	9	9	8
8	9	8	6	6	9	9	9	6
5	10	4	3	7	11	9	7	8
9	4	8	7	8	9	9	8	6
9	6							

X_i	Conteggi	Freq.
3	XXX	3
4	XXX	3
5	XXXX XX	6
6	XXXX XXXX XXXX	12
7	XXXX XXXX XXXX X	13
8	XXXX XXXX XXXX XXXX XXX	19
9	XXXX XXXX XXXX XXXX XXXX XXXX XX	26
10	XXXX X	5
11	XXX	3
12	XX	2

Le annotazioni di conteggio pur utili non sono necessarie alla comprensione dell'indagine campionaria degli '89 baccelli

E' sufficiente un combinato Modalità/ Frequenze

Esempio di elaborazione con “Statistica”



File: Esp95_Xl84.exl
MISS=-9999.00
size: 94 * 2
Include all cases
STATISTICA Frequency Tables; Variable: V1
BASIC Interval Method: All values
STATISTICS Minimum=3,400000 Maximum=10,20000

Category	Freq.	Percent	Cumulativ Freq.	Cumulativ Percent
3-400	1	1.06	1	1.06
3-700	2	2.13	3	3.19
3-800	2	2.13	5	5.32
3-900	4	4.26	9	9.57
4-200	2	2.13	11	11.70
4-300	3	3.19	14	14.89
4-400	1	1.06	15	15.96
4-700	5	5.32	20	21.28
5-000	2	2.13	22	23.40
5-300	5	5.32	27	28.72
5-700	1	1.06	28	29.79
5-800	1	1.06	29	30.85
5-900	3	3.19	32	34.04
6-000	1	1.06	33	35.11
6-100	4	4.26	37	39.36
6-400	4	4.26	41	43.62
6-600	3	3.19	44	46.81
6-700	2	2.13	46	48.94
6-800	1	1.06	47	50.00
6-900	3	3.19	50	53.19
7-000	1	1.06	51	54.26
7-100	2	2.13	53	56.38
7-200	3	3.19	56	59.57
7-500	1	1.06	57	60.64
7-600	2	2.13	59	62.77
7-700	2	2.13	61	64.89
7-800	2	2.13	63	67.02
8-000	3	3.19	66	70.21
8-300	3	3.19	69	73.40
8-400	1	1.06	70	74.47
8-500	2	2.13	72	76.60
8-600	2	2.13	74	78.72
8-700	2	2.13	76	80.85
9-000	2	2.13	78	82.98
9-300	4	4.26	82	87.23
9-500	2	2.13	84	89.36
9-600	1	1.06	85	90.43
9-800	1	1.06	86	91.49
9-900	1	1.06	87	92.55
10-200	2	2.13	89	94.68
			91	96.81
			92	97.87
			94	100.00

Le tabelle

C'è bisogno di una organizzazione e presentazione dei dati più efficiente

Esempio:

Dati in forma narrativa

Persone che non si recano al lavoro per motivi di salute. Il 14.4% dei dirigenti si assentano da uno a tre giorni; il 3.3% da quattro a sette giorni; il 3.2% da 8 a 14 giorni e per più di 14 giorni si assenta il 2.9%. Tra gli impiegati il 60.2% non si assenta mai, il 10.8% si assenta da uno a tre giorni; il 9.9% da quattro a sette giorni; il 4.4% da 8 a 14 giorni ed il 6.0% per almeno 15 giorni. Il 52.6% dei capi operai non restano a casa per motivi di salute. Si assenta da uno a tre giorni l'11.1% e da quattro a sette giorni il 16.1%. Più di 7 giorni, ma meno di 15 si assenta il 2.8% e per più di 14 giorni resta a casa il 9.4%.

dati in forma tabellare

	A casa per motivi di salute				
Professione	mai	1-3gg	4-7gg	6-14gg	>di 14gg
Dirigente	65.3	24.4	5.3	3.2	2.9
Impiegato	60.2	21.8	9.9	4.4	3.7
Capo Operaio	52.6	19.1	16.1	2.8	9.4

Le stesse informazioni sono molto più intellegibili grazie alla tabella

Nelle tabelle statistiche si effettua la prima sgrezzatura dei dati che vengono disposti in ordine logico dopo aver eliminto le ripetizioni

Si interviene anche con accorpamenti e ridefinizioni per semplificare la trattazione

Le tabelle/2

I numeri scritti per esteso non sono comprensibili, ma la loro lettura deve essere aiutata con accorgimenti migliorativi

	A casa per motivi di salute				
Professione	mai	1-3gg	4-7gg	6-14gg	>di 14gg
Dirigente	65.3	24.4	5.3	3.2	2.9
Impiegato	60.2	21.8	9.9	4.4	3.7
Capo Operaio	52.6	19.1	16.1	2.8	9.4

- **Linee di separazione della testata**
- **Linee di contorno**
- **Spaziatura comoda e regolare delle colonne**
- **Uso di una font (helvetica) senza "grazie" che risulta molto efficace per la redazione e lettura delle tabelle**

Le tabelle/3

La riduzione del numero di cifre (eliminando quelle non essenziale al confronto per ordine di grandezza) si migliora la comprensibilità dei dati

	A casa per motivi di salute				
Professione	mai	1-3gg	4-7gg	6-14gg	>di 14gg
Dirigente	65	24	5	3	3
Impiegato	60	22	10	4	4
Capo Operaio	53	19	16	3	9

il dettaglio dei valori con molte cifre è rassicurante per l'impressione di precisione che sembra comunicare.

Non si capisce perché si debbano considerare cifre decimali se i confronti si fanno con cifre intere o quasi:

I valori 89.93 e 45.39 sono precisi, ma 90 e 45 sono più chiari: il primo è il doppio del secondo

Esempio

Indagine campionaria di 64 pazienti sottoposti ad una terapia

Dominio: (G, M, TM, NV,TP, P, D)

TM	D	TP	NV	TP	M	TM	P	G	NV	TP	P	G	NV	G	M
P	G	M	TM	M	NV	M	TM	M	P	P	G	TM	M	M	G
G	TP	M	D	P	TM	NV	TM	TP	NV	G	G	NV	G	TP	TM
NV	P	TM	M	M	TP	M	M	M	TP	TM	G	M	G	TM	M

X_i	n_i
D	2
P	7
TP	8
NV	11
TM	16
M	12
G	8
	64

La simbologia n_i indica la frequenza della modalità i-esima

Raggruppare dei valori in tabelle ha l'inconveniente di disperdere i dettagli

Nel prospetto di raccolta sapevamo quali pazienti erano guariti.

Nella tabella ciò è incerto perchè sono raggruppati: $n_1 = 2$

Esempio

Rating del debito estero da parte di Moody's

Argentina	B1	Corea S.	AA1	India	BAA1	Singapore	AA3
Australia	AA2	Danimarca	AA1	Irlanda	AA3	Slovenia	A1
Austria	AAA	Egitto	BAA3	Italia	AA3	Spagna	AA2
Belgio	AA1	Filippine	AA3	Messico	BA2	Stati UNiti	AAA
Bolivia	AA1	Finlandia	BA1	Norvegia	AA1	Svezia	AA2
Brasile	B2	Francia	AAA	Nuova Zelanda	AA3	Svizzera	AAA
Canada	AAA	Germania	AAA	Paraguay	BAA3	Turchia	BAA3
Cil	B2	Giappone	AAA	Portogallo	A1	Ungheria	BA1
Cina	BAA1	Grecia	BAA1	Regno Unito	AAA	Venezuela	AA2

L'esempio illustra due elementi da non trascurare nella costruzione di tabelle statistiche

Rating	36
A1	2
AA1	5
AA2	4
AA3	5
AAA	8
B1	1
B2	2
BA1	2
BA2	1
BAA1	3
BAA3	3

La presenza di modalità con frequenza unitaria non sempre è opportuna dato che consente la identificazione dell'unità: B1=>Argentina

Non c'è garanzia del segreto statistico

Riportare il totale in testa è una idea che appare efficace, anche se c'è il rischio che il totale sia risommato

Diagramma a punti

Preso un foglio, si traccia (in verticale o in orizzontale) una linea delimitata in modo che il valore più piccolo possibile X_{min} e quello più grande X_{max} siano chiaramente evidenziati.

La linea deve essere graduata con tacche equispaziate corrispondenti a dei valori interi (o comunque di facile lettura nel constesto dell'applicazione).

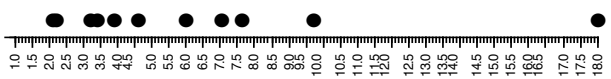
In prossimità del valore più vicino ad ogni modalità si riporta un simbolo (di solito un punto) di dimensione prefissata conforme alla dimensione della linea.

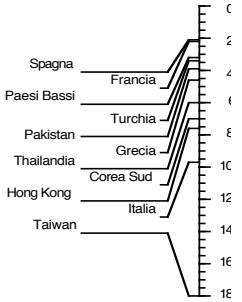
Se più modalità condividono lo stesso punto ovvero sono molto prossime, i punti saranno impilati.

Esempio

Graduatoria delle falsificazioni. Volume di contraffazioni per vari Paesi.

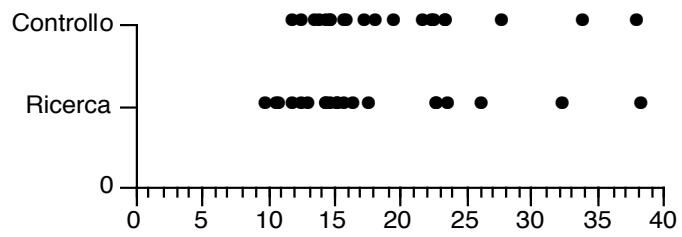
Paese	Volume	Paese	Volume	Paese	Volume	Paese	Volume
Taiwan	18.0	Pakistan	3.9	Francia	2.2	Corea Sud	7.0
Italia	9.7	Paesi Bassi	3.2	Turchia	3.4	Hong Kong	7.6
Thailandia	6.0	Spagna	2.1	Grecia	4.6		





Esempio

Tempi di scioglimento del 75% di un analgesico ottenuti nel laboratorio di ricerca e nel centro controllo produzione.



Nell'esempio si nota che il tempo di dissoluzione trovato dal centro di controllo è tendenzialmente superiore a quello proposto dal centro di ricerca.

Che poi lo scarto sia o meno compatibile con una "sostanziale equivalenza" tra i risultati è un problema che affronteremo nella statistica inferenziale.

Diagramma ramo-foglia

E' un modo diverso e più informativo di presentare i dati (di una variabile discreta o con valori aventi poche cifre).

Numero mensile di richiedenti un mutuo fondiario

45	46	55	52	51	65	48	67	65	66	54	37	70	60	68	58	58	48
67	65	66	54	53	48	59	53	51	60	60	48	61	56	48	59	60	51
47	70	66	55	61	51	71	70	48	70	61	53	46	38	71	46	48	52
66	39	45	68	67	54	70	68	65	45	46	58	72	39	48	71	58	55
28	82	24	80	27	35	81	33	88	85	85	47	29	59	58	89	73	75

Si ordinano in senso crescente i dati e si trascrivono verticalmente i valori che costituiranno i rami.

A fianco di ciascuna si riportano i valori più piccoli (le foglie)

Le foglie possono sono a loro volta ordinate

28 24

2 | 84

2 | 48

Rami (1^a cifra)	Foglie (Cifra finale)						n_j
2	4789						4
3	3578	99					6
4	5556	6667	7788	8888	8		17
5	1111	2233	3444	5558	8888	999	23
6	0000	1115	5556	6667	7788	8	21
7	0000	0111	2235				12
8	0125	589					7

Esempio di costruzione

Una analista contabile vuole capire l'andamento dei saldi crediti al consumo attivi in un supermarket

Non intende perdere tempo esaminandoli tutti. Ne sceglie un campione di 40

Prospetto di raccolta dati

71	**58**	**66**	**119**	**55**	**46**	**22**	**69**	**84**	**72**
45	**61**	**45**	**84**	**68**	**107**	**96**	**58**	**47**	**61**
91	**47**	**102**	**76**	**63**	**55**	**52**	**69**	**75**	**10**
85	**32**	**63**	**55**	**55**	**65**	**66**	**35**	**70**	**78**

Frequenze

1	0	
2	2	
3	25	
4	6557	
5	8585255	
6	6918139356	
7	12650	
8	44758	
9	61	
10	72	
11	9	

Frequenze

1	0	1
2	2	1
3	25	2
4	5567	4
5	2555588	7
6	1133566899	10
7	01256	5
8	44578	5
9	16	2
10	27	2
11	9	1

Diagramma ramo-foglia/2

il diagramma ramo-foglia dà le stesse informazione della tabella, ma aggiunge una dimensione visiva interessante

Ruotando opportunamente il diagramma si ottiene una visione d'assieme molto utile dei valori riscontrati nel campione

Anche il diagramma ramo-foglia disperde delle informazioni:

E' persa la sequenza di acquisizione ed i valori non sono riconducibili alle unità su cui sono stati rilevati

A questo può però ovviare il Digidot

Rami (1^acifra)	Foglie (Cifra finale)						n_j
2	4789						4
3	3578	99					6
4	5556	6667	7788	8888	8		17
5	1111	2233	3444	5558	8888	999	23
6	0000	1115	5556	6667	7788	8	21
7	0000	0111	2235				12
8	0125	589					7

Diagramma ramo-foglia/3

Può essere usato per modalità frazionarie. Basterà adottare delle formule di arrotondamento

Se i valori hanno 4 cifre frazionarie e si ne vogliono usare solo le 2 più significative

$$X_i^* = \frac{[X_i * 100 + 0.5]}{100}$$

$$1.567 \Rightarrow \frac{[1.567*100+0.5]}{100} = \frac{[156.7+0.5]}{100} = \frac{[157.2]}{100} = \frac{157}{100} = 1.57$$

Esempio: Tasso di variazione in percentuale delle giacenze

-1.01	-2.39	-0.01	-0.72	-1.05	-0.05	-0.02	-1.04	-2.35
0.15	-1.54	-1.05	0.13	0.11	-0.43	-0.72	-1.36	-0.41
-1.50	0.14	-1.36	-0.07	-1.56	-0.43	0.12	-1.54	-0.71
-0.01	0.17	-0.01	0.19	-0.41	0.00	-1.07	-0.45	-1.36
-2.50	-2.33	0.13	0.35	-0.44	0.33	-1.51	-0.78	0.36
0.37	-1.58	-1.58	-0.44	0.39	0.31	-0.04	-1.34	-0.75
-0.05	0.35	0.16	-2.36	-0.74	-2.58	-0.02	-0.76	-2.31

Formato: ramo: XX.X, foglia: X

-2.5	80
-2.3	96531
-1.5	88644 10
-1.3	6664
-1.0	73541
-0.7	86542 21
-0.4	54433 11
0.0	75542 22111 0
0.1	12334 5679
0.3	13356 79

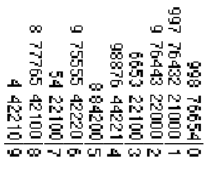
Diagramma ramo-foglia/4

Lo sviluppo delle foglie non deve per forza essere da sinistra verso destra e l'orientamento può anche essere verticale

Rilevazione campionaria del numero di fabbriche per distretto industriale

42	86	85	88	46	48	29	10	75	20	62	94	71
65	8	31	48	20	10	81	52	14	65	24	62	50
19	80	42	80	44	41	22	69	62	70	35	36	65
30	13	27	58	54	12	47	92	16	60	20	19	44
91	58	94	12	65	9	87	87	6	72	87	6	84
11	70	49	82	33	24	74	17	7	92	90	72	50
67	4	23	26	32	5	10	36	9	64	30	17	32

998 76654	0
997 76432 21000	1
9 76443 22000	2
6653 22100	3
98876 44221	4
8 84200	5
9 75555 42220	6
54 22100	7
8 77765 42100	8
4 42210	9



Numero di rami

Esistono vari suggerimenti:

- EMERSON-HOAGLIN: [10Log(n)]
- Proporzione radice: [1.5√n]

[.] parte intera

Se n=50

$$E-H \Rightarrow \left[10\,Log(50)\right] = 16, \qquad PR = \left[1.5\sqrt{50}\right] = 10$$

Spezzatura dei rami

Se le cifre iniziali sono poche e i rami molto lunghi conviene dividerli

Si spezza il ramo ripetendo la sua cifra seguita da due diversi segni per separare i valori inferiori o uguali alla metà e quelli superiori

0	⇒	0 *
		0 +
1		1 *
		1 +
2		2 *
		2 +
3		3 *
		3 +

11	26	32	44	51
13	26	33	45	51
14	26	34	46	51
15	27	35	47	53
16	27	36	47	54
16	28	37	47	55
16	29	37	47	57
18	30	38	48	58
22	30	38	49	58
22	30	38	49	58
23	31	42	50	59
24	31	42	50	59

1*	13455
1+	6668
2*	2234
2+	6667789
3*	000112345
3+	677888
4*	2245
4+	67777899
5*	00111345
5+	788899

ESEMPIO: società per numero di licenze software

Le frequenze relative

Le tabelle riassumono il modo in cui le unità si ripartiscono fra le varie modalità.

x_i **Modalità i-esima**
n_i **Frequenza assoluta (numero di presenze di x_i)**

$n = \sum_{i=1}^{n_i} n_i$ **Totale delle rilevazioni**

$f_i = \frac{n_i}{n}$ **frequenza relativa (peso di X_i nella rilevazione)**

Le frequenze relative, per costruzione, verificano le seguenti relazioni

a) $0 \le f_i \le 1 \quad (i=1,2, \ldots ,k)$

b) $\sum_{i=1}^{k} f_i = 1$

k è il numero di modalità distinte che è possibile rilevare nell'indagine

Frequenze relative/2

Le frequenze relative sono confrontabili tra loro ed in rilevazioni diverse dato che hanno perso l'ordine di dimensionalità (sono tutte tra zero ed uno)

$f_i = 0$ **Significa che la modalità i-esima era osservabile nella Popolazione (spazio dei dati), ma non è stata osservata nella rilevazione**

$f_i = 1$ **Significa che, sebbene nella popolazione era possibile osservare più di una modalità, le unità incluse nella rilevazione hanno presentato modalità costante X_i**

La semplificazione ottenuta non è senza costo.

Il passaggio dalle frequenze assolute alle relative comporta la perdita di un grado diLibertà. Infatti, il vincolo

$$1 = \sum_{i=1}^{k} f_i$$

Significa che, note k-1 frequenze relative qualsiasi quella mancante si ricava dal vincolo.

ESEMPIO

In una area di sviluppo si sono censiti gli addetti nelle piccole imprese (meno di 10 addetti).

7	5	9	2	2	4	7	8	4	7	2	2
9	5	4	5	6	7	8	2	9	8	2	9
2	9	4	7	8	5	6	7	2	5	8	8
5	7	5	6	5	2	9	6	6	3	2	5
3	5	6	4	8	5	4	2	6	3	7	4
4	3	7	9	2	8	3	3	4	4	5	8

X_i	n_i	f_i
2	12	0.1667
3	6	0.0833
4	10	0.1389
5	12	0.1667
6	7	0.0972
7	9	0.1250
8	9	0.1250
9	7	0.0972
	72	1.0000

Da un esame rapido emerge che gli addetti sono ripartiti in modo abbastanza uniforme presso le piccole aziende.

Lo scarto massimo da 6 a 12 presenze non appare enorme alla luce delle 72 unità rilevate.

ESEMPIO

Redigiamo la distribuzione di frequenze delle parole classificate per vocale finale nel seguente brano (separare le parole apostrofate s'era=si era).

La casa di Oreste era un terrazzo rosso e scabro e dominava nella gran luce un mare di valli e burroni che faceva male agli occhi. Ero corso per tutto il mattino nella pianura che conoscevo e dal finestrino avevo intravisto le rogge alberate della mia infanzia: specchi d'acqua, branchi di oche, praterie. Ci pensavo ancora quando il treno s'era messo per ripe scoscese e dove bisognava guardare in su per vedere il cielo. Dopo una stretta galleria s'era fermato. Nell'afa e nella polvere mi ritrovai sulla piazzetta della stazione, gli occhi pieni di coste calcinate. Un carrettiere grasso mi mostrò la strada; dove salire salire, il paese era in alto. Gettai la valigetta sul carro e al passo lento dei buoi salimmo insieme [...

da "Il Diavolo sulle Colline" di C. Pavese

La "a" e la "e" sono dominanti. Le consonanti sono poco presenti alla fine della parola. La "u" è rarissima.

X_i	n_i	f_i
A	30	0.2344
E	33	0.2578
I	23	0.1797
O	25	0.1953
U	1	0.0078
Cons.	16	0.1250
	128	1.0000

Modalità in classi

il raggruppamento delle modalità del dominio è utile in varie occasioni

- **Variabili continue o dense**
- **Presenza di modalità con frequenze piccole**
- **Per fenomeni di cui interessa la gradualità più che l'intensità**
- **Rilevazioni puntuali incerte o di affidabilità limitata**
- **Semplificazione della presentazione dei dati raccolti**

L'uso del raggruppamento in classi NON è applicato per la elaborazione poiché provoca la perdita di informazioni di dettaglio

Se però siamo eredi di dati raccolti da altri e presentati in classi dobbiamo saperli trattare

Modalità in classi/2

$$X_i: (L_i, U_i), \quad i = 1,2,\ldots,k \quad \text{con} \quad L_i \le U_i$$

Gli estremi possono essere inclusi oppure esclusi (uno o entrambi)

$$\begin{aligned} X_i: &\ \{X_i | L_i \le X \le U_i\} && \text{Chiuse} \\ X_i: &\ \{X_i | L_i < X < U_i\} && \text{Aperta} \\ X_i: &\ \{X_i | L_i < X \le U_i\} && \text{Aperta a sinistra} \\ X_i: &\ \{X_i | L_i \le X < U_i\} && \text{Aperta a destra} \end{aligned}$$

La distinzione è importante dato che talvolta le classi di variabili continue o dense vengono presentate con la convenzione

$$L_i = U_{i-1}, \quad i = 2,3,\ldots,k$$

ciò potrebbe comportare incertezza nell'assegnare alla classe giusta le modalità limite

ESEMPIO

S supponga che in una tabella si abbiano le classi riportate a destra:

a) Calcolare ampiezze e valori centrali delle classi;

b) in quali classi ricadono le frequenze: 0.6395, 0.7189, 0.9114?

	X_i	
1)	0.218	0.639
2)	0.639	0.720
3)	0.720	0.9115
4)	0.9115	1.1318

N.B. Le classi sono aperte a sinistra e chiuse a detra

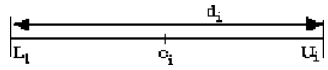
A1) 0.4285, 0.6795, 0.81575, 1.02165
A2) 0.4210, 0.0810, 0.19150, 0.22030

B1) "2", "2", "3"

Caratterizzazione delle classi

Le classi hanno due elementi importanti:

Ampiezza : $d_i = (U_i - L_i)$



Valore centrale $c_i = \dfrac{U_i + L_i}{2}$

Ai fini del calcolo dei valori centrali e delle ampiezze. NON rileva che gli estremi siano inclusi o no

Classe		Ampiezza	Valore Centrale	
-4	-2	-2 -(-4)=2	$\frac{-2+(-4)}{2}$	= -3
-2	-1	-1 -(-2)=1	$\frac{-1+(-2)}{2}$	= -1.5
-1	2	2-(-1) = 3	$\frac{2+(-1)}{2}$	= 0.5
2	6	6 -2 = 4	$\frac{6+2}{2}$	= 4

Densità di frequenza

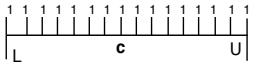
E' una utile caratteristica delle classi

$$h_i = \frac{f_i}{d_i} = \frac{f_i}{(U_i - L_i)}; \quad i = 1,2,\ldots,k$$

che misura quanta parte della frequenza relativa spetterebbe ad una sotto classe del denominatore se a ciascuna ne toccasse in parti uguali.

Mesi	Maschere	d_i	f_i	h_i	Mesi	Maschere	d_i	f_i	h_i
0 - 6	28	6	0.0142	0.0024					
6 - 12	92	6	0.0467	0.0078	24 - 30	649	6	0.3297	0.0550
12 - 18	270	6	0.1371	0.0229	30 - 36	134	6	0.0680	0.0113
18 - 24	702	6	0.3567	0.0595	36 - 52	93	16	0.0472	0.0030
						1968		1.0000	

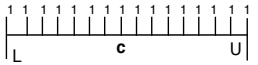
L'indicazione data dalla densità di frequenza è esatta purché la ripartizione delle unità all'interno della classe sia uniforme e cioè del tipo:



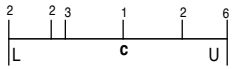
Tipicità del valore centrale

Dipende dalla configurazione con cui si presentano le modalità.

Nel caso della uniforme

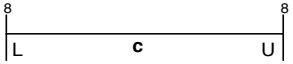


è questionabile in quanto non c'è ragione di preferire il punto di mezzo.



"c" è isolato ed esprime solo se stessa

"c" è poco rappresentattivo



Non rappresenta nessuno

L'uso del valore centrale è corretto in caso di classe degenere:

ESEMPIO

Consideriamo le seguenti classi:

(A) 10'000 - 20'000
(B) 20'000 - 30'000

Se succede che X=20'000 ci sono varie possibilità:

1) Si aumenta di uno la classe (A) nel presupposto che 20'000 sia il massimo di tale classe.

2) Si aumenta di uno la classe (B) nel presupposto che 20'000 sia il minimo di tale classe.

3) Si sorteggia la classe da aumentare di uno.

4) Si aumentano alternativamente di uno le classi (A) e (B) cominciando da una scelta casualmente

In generale, se non altrimenti indicato si intenderà la classe come chiusa a sinistra ed aperta a destra

Classi per dati arrotondati

Per variabili continue oppure dense si arrotonda di solito con la regola del 5

Se minore di 5 si arrotonda all'unità più piccola
Se maggiore o uguale a 5 all'unità più grande

La differenza minima osservabile tra due valori è l'unità di arrotondamento.

Tra limiti reali e riportati vale la relazione:

$$\text{Limite reale } L_i = \text{Limite riportato } L_i - \frac{\text{unità di arrotondamento}}{2};$$

$$\text{Limite reale } U_i = \text{Limite riportato } U_i + \frac{\text{unità di arrotondamento}}{2}$$

il dato arrotondato ricade all'interno degli estremi (che sono più larghi), ma il valore reale potrebbe invece sconfinare di classe

Esempio

Nel prospetto è classificato un campione di comuni secondo il rapporto di composizione: (suoli agricoli/superficie totale)*100.

i	Reali	Riportati	ni
1	0.0 - 1.5	0 - 1	35
2	1.5 - 3.5	2 - 3	62
3	3.5 - 6.5	4 - 6	46
4	6.5 - 9.5	7 - 9	23
5	9.5 - 13.5	10 - 13	19
6	13.5 - 20.5	14 - 21	5
			190

Se per un comune il rapporto vale 3.4 si incrementerà di uno la frequenza della seconda classe riportata (2 - 3) corrispondente alla classe reale (1.5 - 3.5).

I valori dovranno confluire in tale classe fino a 3.5 e solo a questo punto l'incremento di frequenza per un nuovo dato scatterà per la terza classe (4 - 6).

Limiti delle classi estreme indeterminati

Non sempre le classi estreme hanno limiti espliciti

meno di U_1, PIù di L_k, al più U_1, almeno L_k

- **Perché non si conoscono**
- **Perché non interessano**
  - **Perché "remoti" rispetto al blocco centrale dei dati**
  - **Perché non esiste un limite preciso**

In questi casi è difficile stabilire valori centrali ed ampiezze delle classi. Occorre fare delle ipotesi soggettive (arbitrarie)

Ipotesi dell'altezza proporzionale

Un'ipotesi che risolve incertezze su altezze e valori centrali è la seguente:

$$\frac{h_1}{h_2}=\frac{h_2}{h_3};\qquad \frac{h_{k-1}}{h_{k-2}}=\frac{h_k}{h_{k-1}}$$

il rapporto tra altezze di classe estrema e classe contigua è pari al rapporto delle due classi immediatamente precedenti (superiore) o seguenti (inferiore)

$$h_i=\frac{f_i}{U_i-L_i}$$

Ne consegue che

$$c_1=\text{Max}\left\{\frac{X_{min}+U_1}{2};\ U_1-\frac{h_3}{h_2^2}*\frac{f_1}{2}\right\};\qquad L_1=2*c_1-U_1$$

$$c_k=\text{Min}\left\{\frac{X_{max}+L_k}{2};\ L_k+\frac{h_{k-2}}{h_{k-1}^2}*\frac{f_k}{2}\right\};\qquad U_k=2*c_k-L_k$$

Xmin ed Xmax sono le modalità più piccole del dominio (riscontrabili quindi nella popolazione e non necessariamente riscontrate nel campione)

Esempio

Distribuzioni di frequenza relative alle macchine eoliche installate nel mondo. Per questa distribuzione sembra ragionevole proporre:

Potenza in kW	Macchine	f	d	h	i
< 5	15000	0.8214	-	-	1
6 - 25	200	0.0110	20	0.0005476	2
26 -100	2500	0.1369	75	0.0018254	3
101-300	550	0.0301	200	0.0001506	4
301-700	9	0.0005	400	0.0000012	5
> 700	2	0.0001	-	-	6
	18261				

$$X_{min}=1\ \text{kW};\quad X_{max}=1000\ \text{kW}$$

Sono limiti fisici suggeriti dall'indagine, ma altri limiti sarebbero plausibili

$$c_1=\text{Max}\left\{\frac{1+5}{3};\ 5-\frac{0.0018254}{(0.0005476)^2}*\frac{0.8214}{2}\right\}=\text{Max}\{3;\ -2495\}=3;\qquad L_1=2*6-5=1$$

$$c_6=\text{Min}\left\{\frac{1000+700}{2};700+\frac{0.0001506}{(0.0000012)^2}*\frac{0.0001}{2}\right\}=\text{Min}\{850;6116\}=850;\ U_k=2*850-700=1$$

Numero delle classi

Il numero e le ampiezze delle classi dovranno scaturire da un compromesso tra esigenze contrastanti: l'accuratezza della presentazione, la semplicità della presentazione.

Tassi minimi di sconto commerciale

6.56	7.31	7.31
6.75	11.25	6.62
4.62	15.37	7.31
6.37	12.43	6.87
5.62	8.00	7.25
7.12	8.68	6.93
4.75	8.62	6.87
6.18	7.68	6.81
5.62	6.93	8.18
4.81	6.62	10.31
4.81	10.43	12.75
5.25	12.06	9.68

X_i		n_i
4.6	8.2	26
8.2	11.8	16
11.8	15.4	4
		36

Compatte

X_i		n_i
4.6	6.4	9
6.4	8.2	17
8.2	8.2	6
10.0	10	3
13.6	15.4	1
		36

Buone, ma c'è di meglio

X_i		n_i
4.6	5.5	5
5.5	6.4	4
6.4	7.3	11
7.3	8.2	6
8.2	9.1	2
9.1	10.9	3
10.9	11.8	1
11.8	12.7	2
12.7	13.6	1
13.6	14.5	0
14.5	15.4	1
		36

Sparse

Numero delle classi/2

Non esistono regole granitiche, ma suggerimenti empirici più o meno validi

Si deve porre un limite minimo per non accorpare troppo valori eterogenei in classi molto vaste

Si deve porre un limite massimo per non vanificare la semplificazione che motiva il raggruppamento

Di solito si pone $5 \le k \le 25$

Un'utile regola è quella di Sturges con "K" arrotondato per difetto o per eccesso secondo la regola del 5

$$K = 1 + 3.322 * \mathrm{Log}_{10}(n)$$

ESEMPIO:

Un campione di n=179 valori dovrebbe essere raggruppato in k=8 classi

$$k = 1 + 3.322 * \mathrm{Log}_{10}(179) = 1 + 3.322 * 2.2528 = 8.4838 \cong 8$$

Ampiezza delle classi

Anche qui suggerimenti pratici, ma la cui applicabilità deve essere valutata di volta in volta

Le classi dovrebbero essere della stessa ampiezza per facilitare il confronto tra i diversi livelli raggiunti dalla variabile

Gli estremi dovrebbero essere multipli di 2, 10 e 5 per la loro migliore leggibilità.

La comune ampiezza potrebbe essere ottenuta con la formula

$$d = \frac{X_{(n)} - X_{(1)}}{1.5\sqrt[3]{n}}$$

ESEMPIO: per i valori

Cosenza	240	Acri	700	Rossano	270
Corigliano Calabro	219	Rende	481	Catanzaro	343
San Giovanni in Fiore	1008	Reggio Calabria	29	Vibo Valentia	556
Crotone	43	Lametia Terme	210	Paola	94
Cutro	221	Cassano allo Jonio	250	Siderno	5
Gioia Tauro	23	Palmi	250		
Taurianova	208	Castrovillari	350		

$$d = \frac{1008 - 5}{1.5\sqrt[3]{19}} = \frac{1003}{4} \cong 252$$

0	**250**
250	**500**
500	**750**
750	**1008**

Ampiezza delle classi/2

L'ampiezza delle classi non deve essere necessariamente costante

Se i valori si addensano intorno a valori piccoli o grandi poche classi includerebbero la maggior parte delle modalità e resterebbero classi praticamente vuote

Le classi debbono essere meno numerose (o più ampie) laddove si riscontrino poche modalità. Debbono essere più numerose (più sottili) per livelli più affollati

ESEMPIO: rilevazioni dei danni provocati da catastrofi naturali

5	5	5	5	7	7	7	7	8	9	9
9	11	11	11	15	15	15	15	15	16	17
20	20	20	20	20	20	20	20	20	21	21
22	23	23	23	24	24	24	26	26	27	27
27	27	29	31	33	33	36	38	40	40	40
40	40	40	40	40	43	44	45	45	45	
45	45	45	46	46	47	47	48	48	48	49
50	50	50	50	50	50	50	50	50	50	50
50	50	50	58	63	70	70	70	70	75	75

In questo caso si è privilegiata la uniformità delle frequenze e si sono stabilite le classi in modo da includere lo stesso numero di casi.

X	n_i
5 - 8	9
9 - 15	11
16 - 20	11
21 - 38	10
40 - 44	11
45 - 48	13
49 - 50	15
58 - 75	8
	88

La regola dell'ampiezza era inadatta

$$d = \frac{75 - 5}{1.5\sqrt[3]{99}} \cong 10$$

Costruzione pratica di una tabella in classi

- **Si ordinano i dati in senso crescente e si trovano** $X_{(1)}$ ed $X_{(n)}$
- **Si calcola il campo di variazione** $R = X_{(n)} - X_{(1)}$
- **Si sceglie "k" con la regola di Sturges**
- **La comune ampiezza delle classi è data da** $M = \left(\frac{R}{k}\right)$

  con "r" cifre decimali dove "r" è il numero di cifre con cui sono riportati i dati
- **Si sceglie un conveniente estremo inferiore:** $L_1 \le X_{(1)}$
- **Si pone:** $L_i = L_1 + (i-1)*d \quad \text{per } i = 2,3,\ldots,k$
- **Si pone:** $U_i = L_{i+1} - \delta \quad \text{per } i = 1,3,\ldots,k-1$ **con** $\delta = (0.1)^r$
- **Si sceglie un conveniente estremo superiore:** $U_k \ge X_{(n)}$

Esempio

Indagine campionaria sui tempi di espletamento di un certo compito

11.24	14.23	73.56	7.23	29.52	64.71	22.14	38.19	19.66	34.45	23.56	12.71	94.82	42.44	55.37
11.36	2.42	15.35	44.14	95.61	19.73	89.55	17.64	21.69	56.28	12.81	26.40	57.57	61.00	23.22
98.15	72.30	16.41	3.87	5.23	13.37	10.31	36.16	66.17	23.89	28.00	69.43	15.70	12.76	94.72
39.91	16.84	13.81	17.29	46.38	51.17	24.29	33.91	49.82	21.73	26.15	55.52	34.23	26.50	57.49

a) $X_{(1)} = 2.42, \quad X_{(60)} = 98.15;$

b) $R = 98.15 - 2.42 = 95.73$

c) $k = 1 + 3.322 * \text{Log}_{10}(60) = 6.9 \cong 7;$

d) $d = \frac{R}{k} = \frac{95.73}{7} = 13.67 \cong 14;$ e) $L_1 = 2;$ f) $L_i = 2 + (i-1)*14 \Rightarrow (2,16,30,44,58,72,86)$

g) $\delta = (0.1)^2 = 0.01;$ h) $U_i = L_{i+1} - 0.01 \Rightarrow (15.99,29.99,43.99,57.99,71.99,85.99);$ i) $U_7 = 99$

X_i			n_i
2.00-	15.99	xxxxx xxxxx xxxxx	15
16.00-	29.99	xxxxx xxxxx xxxxx xxx	18
30.00-	33.99	xxxxx xx	7
44.00-	57.99	xxxxx xxxx	9
58.00-	71.99	xxxx	4
72.00-	85.99	xx	2
86.00-	99.00	xxxx	4
			60

Se antiestetiche, le due cifre finali possono essere eliminate ponendo

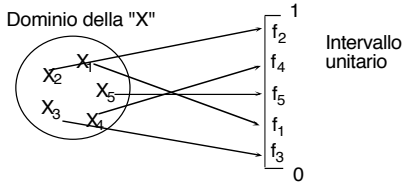
$$U_i = L_{i+1}$$

che però lascerà incertezze sulle modalità estreme

La funzione di distribuzione empirica

Le informazioni dello schema riassuntivo possono essere ulteriormente sintetizzate nella FUNZIONE DI DISTRIBUZIONE EMPIRICA:

$$f(X) = \begin{cases} f_i & \text{se } X = X_i \\ 0 & \text{altrimenti} \end{cases} \qquad i = 1,2,\ldots,k$$



La funzione di distribuzione è un meccanismo che associa ad ogni modalità la frequenza relativa con cui si è presentata.

Si aggiunge l'aggettivo "empirica" in quanto basata su valori osservati

Grafico della funzione di distribuzione

Molto utile per facilitare il confronto tra distribuzioni di frequenza

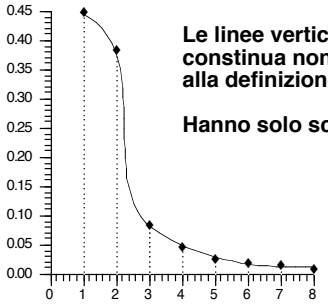
E' costituito dai punti di coordinate $(X_i, f_i), \quad i = 1,2,\ldots,n$

Tali punti sono talvolta collegati a mezzo di aste per migliorarne la leggibilità.

A questo fine si usano anche delle linee spezzate o continue

Famiglie abbonate a Cosmopolitan per numero di componenti

X_i	n_i	f_i
1	2226	0.4452
2	1908	0.3816
3	405	0.0810
4	206	0.0412
5	103	0.0206
6	71	0.0142
7	50	0.0100
8	31	0.0062
	5000	1.0000



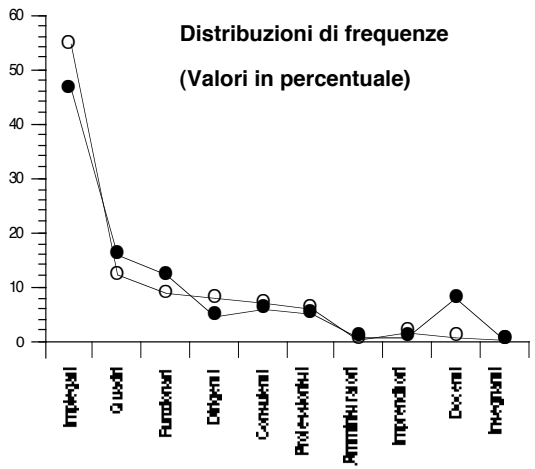
Le linee verticali e la linea constinua non appartengono alla definizione.

Hanno solo scopo estetico

Esempio

Professioni scelte dai laureati di una università del Nord Italia

	CL. EC. AZ.	CL. D.E.S.
Impiegati	54.4	46.5
Quadri Privati	12.1	16
Funzionari pubblici	9	12.3
Dirigenti	8	4.7
Consulenti	7.2	5.8
Professionisti	6.1	5.2
Amministratori	0.5	0.8
Imprenditori	1.6	0.8
Docenti	0.9	7.8
Insegnanti	0.2	0.1
	100	100



Distribuzioni di frequenze

(Valori in percentuale)

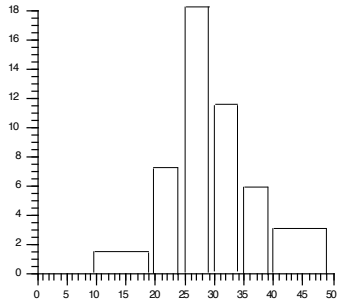
Le differenze maggiori tra ECON.AZ. che diventano impiegati e DEs che diventano docenti universitari

Esempio

Le distribuzioni passano attraverso due diverse semplicazioni: tabellare e grafica

X_i		n_i
10	19	3
20	24	7
25	29	18
30	34	11
35	39	6
40	49	5
		50

X=Anni di attività



Le informazioni e le impressioni che si ricavano sono diverse, ma hanno un fine comune: capire che cosa i dati vogliono dire

La separazione tra rettangoli è plausibile solo per variabili discrete

Da notare che in questo caso il fattore di propozionalità è a=n il che implica un asse delle ordinate che riporta le frequenze assolute

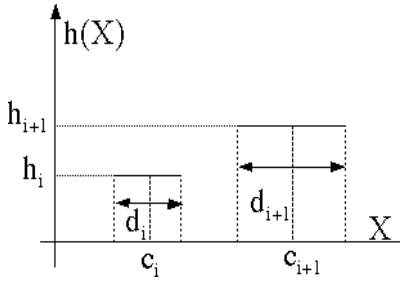
Funzione di distribuzione (per classi)

E' simile a quella per singole modalità, ma i singoli punti non sono più rappresentabili

$$f(X)=\begin{cases} f_i & \text{se } X\in(L_i,U_i) \\ 0 & \text{altrimenti} \end{cases} \qquad i=1,2,\ldots,k$$

Per riportare correttamente in grafico la funzione di distribuzione per dati raggruppati occorrono le altezze (o densità di frequenza)

$$h_i=\frac{f_i}{d_i}$$ **per i=1,2,…,k**



Le altezze esprimono quante osservazioni sono di pertinenza di un sottointervallo unitario della classe (supponendo che le frequenze siano equiripartite).

L'istogramma delle frequenze

E' il grafico della funzione di distribuzione per dati in classi

In un sistema di assi cartesiano si pongono le modalità sulle ascisse e si costruiscono dei rettangoli di area proporzionale alla frequenze relative

$$A(L_i,U_i)=\alpha*(d_i*h_i) \quad dove \begin{cases} \alpha & \text{fattore di proporzionalità} \\ d_i=(U_i-L_i) \\ h_i=\frac{f_i}{d_i} \end{cases}$$

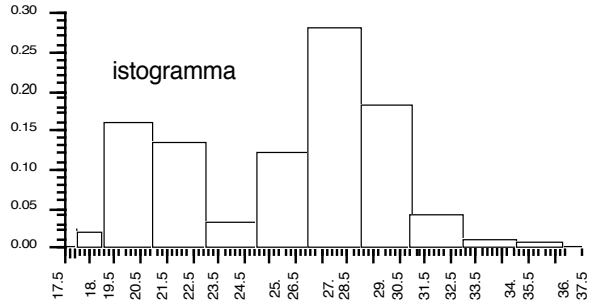
L'area totale dei rettangoli è pari alla costante α *Di solito $\alpha=1$*

$$\sum_{i=1}^{k}A(L_i,U_i)=\sum_{i=1}^{k}\alpha(d_i*h_i)=\sum_{i=1}^{k}\alpha f_i=\alpha\sum_{i=1}^{k}f_i=\alpha$$

Esempio

Lunghezza del corpo di un campione di sogliole

X_i		n_i
18.	19.	25
20.	21.	180
22.	23.	150
24.	25.	36
26.	27.	136
28.	29.	322
30.	31.	203
32.	33.	49
34.	35.	11
36.	37.	4
		1116



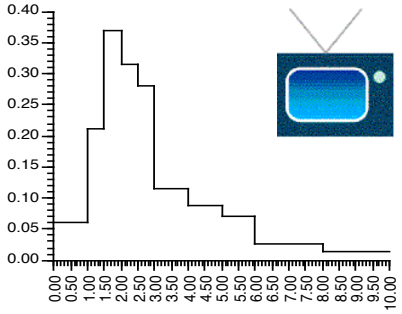
La somma bloccata permette di controllare l'area dei singoli rettangoli e quella complessiva

La doppia gobba indica la presenza di due razze diverse oppure di due diverse fasi di sviluppo

Esempio

C di famiglie per il tempo complessivo in cui l'apparecchio rimane acceso.

X_i		n_i	f_i	h_i
0.0	1.0	7	0.0614	0.0614
1.0	1.5	12	0.1053	0.2105
1.5	2.0	21	0.1842	0.3684
2.0	2.5	18	0.1579	0.3158
2.5	3.0	16	0.1404	0.2807
3.0	4.0	13	0.1140	0.1140
4.0	5.0	10	0.0877	0.0877
5.0	6.0	8	0.0702	0.0702
6.0	8.0	6	0.0526	0.0263
8.0	10.0	3	0.0263	0.0132
		114	1.0000	



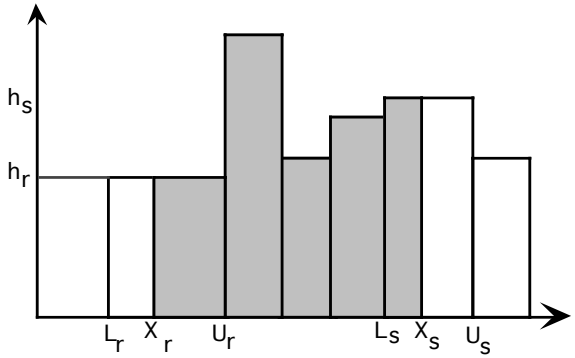
L'ipotesi di equifreqeunza nella classe è criticabile perché:

Introduce distorsioni se la variabile è discreta o densa in quanto assegna ordinate anche ad ascisse inesistenti

E' una forzatura se la variabile è continua in quanto la condizione di equifrequenza è raramente giustificata.

Additività

Deriva dalla natura di area della frequenza relativa



$$A(X_r, X_s) = A(X_r, U_r) + \sum_{i=r+1}^{s-1} A(L_i, U_i) + A(L_s, X_s)$$

Esempio

Un costruttore di *hard disk* ha fatto rilevare lo spazio non utilizzato sulla memoria di massa di n=600 utenti.

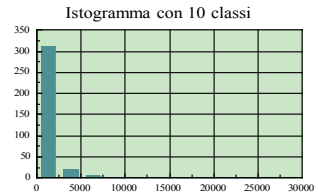
Ignorando i dati di dettaglio, quale percentuale si può presumere ricada tra il 24% e il 35%?

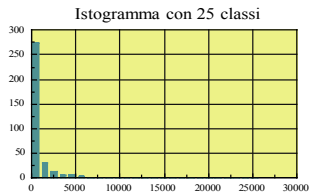
% Spreco	Hard disk	f_i	d_i	h_i
0 - 4	45	0.075	4	0.0188
4 - 8	98	0.163	4	0.0408
8 -15	126	0.210	7	0.0300
15 - 30	200	0.333	15	0.0222
30 - 40	91	0.152	10	0.0152
40 - 50	40	0.067	10	0.0067
	600	1.000		

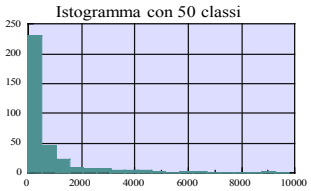
$$A[24,35] = A[24;30] + A[30;35]$$

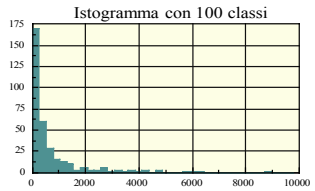
$$= 6 * 0.0222 + 5 * 0.0152 = 0.2092$$

Modifica della forma secondo le classi

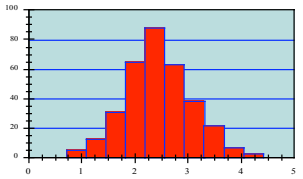


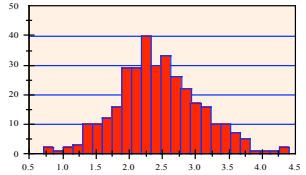


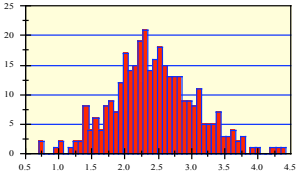


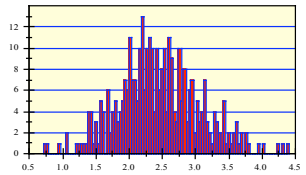


Modalità in scala logaritmica









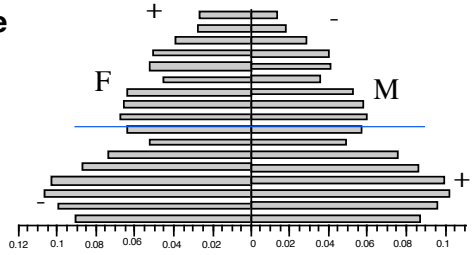
Confronto di distribuzioni

L'istogramma delle frequenze è utile l'analisi congiunta di due variabili rilevate -con le stesse classi- in due occasioni diverse.

Per chiarire il confronto I due istogrammi sono presentati in a forma una piramide.

La piramide della popolazione

Età	Femm.	Maschi	Età	Femm.	Maschi
meno di 4	0.0809	0.0871			
5 _ 9	0.0889	0.0957	45 _49	0.0587	0.0575
10_14	0.0955	0.1025	50 _54	0.0572	0.0527
15 _ 19	0.0921	0.0997	55 _59	0.0409	0.0359
20 _ 24	0.0783	0.0863	60 _64	0.0472	0.0414
25 _ 29	0.0664	0.0759	65 _69	0.0453	0.0403
30 _ 34	0.0473	0.0489	70 _74	0.0349	0.0285
35 _39	0.0571	0.0569	75 _79	0.0252	0.0179
40 _44	0.0603	0.0592	più di 80	0.0238	0.0134
				1.0000	1.0000



Classi giovani: + maschi; altre classi: + femmine

Poligono di frequenza

Grafico che discende dall'istogramma ottenuto riportando in un sistema cartesiano i valori centrali delle classi e le frequenze relative

$$(c_i,\ f_i);\quad i = 1,2,\ldots,k$$

a questi si aggiungono i due punti convenzionali:

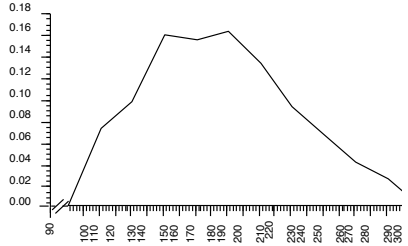
$$\left(c_1 - \frac{d_1}{2},\ 0\right);\ \left(c_k + \frac{d_k}{2},\ 0\right)$$

così il grafico parte e finisce sull'asse delle ascisse

il poligono di frequenza riporta solo il profilo esterno dell'istogramma rendendo Più facile la percezione

Contenuto calorico in alcuni alimenti

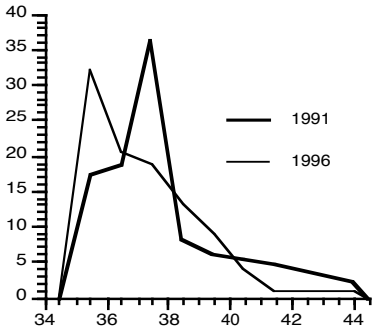
x_i		n_i	c_i	f_i
100	120	14	110	0.0714
120	140	19	130	0.0969
140	160	31	150	0.1582
160	180	30	170	0.1531
180	200	32	190	0.1633
200	220	26	210	0.1327
220	240	18	230	0.0918
240	260	13	250	0.0663
260	280	8	270	0.0408
280	300	5	290	0.0255
		196		1.0000



Esempio

Confronto della distribuzione di frequenza degli stabilimenti tedeschi per numero di ore settimanali lavorate.

Ore	1991	1996
35 _ 35.9	17.5	32.2
36 _ 36.9	19.0	20.6
37 _ 37.9	36.4	18.8
38 _ 38.9	8.4	13.2
39 _ 39.9	6.3	9.1
40 _ 40.9	5.5	4.0
41 _ 41.9	4.6	1.2
42 _ 45.9	2.3	0.9
	100.0	100.0

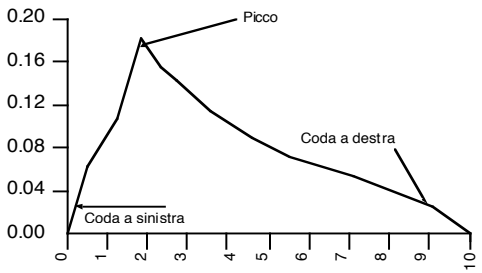


E' evidente a riduzione dell'orario più praticato: da 37 a 35 ore.

Esempio

Famiglie per tempo (in ore) complessivo in cui il televisore rimane acceso.

X_i		f_i
0.0	1.0	0.0614
1.0	1.5	0.1053
1.5	2.0	0.1842
2.0	2.5	0.1579
2.5	3.0	0.1404
3.0	4.0	0.1140
4.0	5.0	0.0877
5.0	6.0	0.0702
6.0	8.0	0.0526
8.0	10.0	0.0263
		1.0000



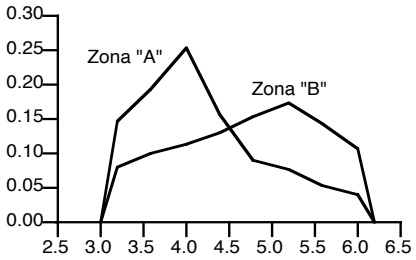
Il picco è il livello in cui la frequenza è massima.

Le code sono gli allungamenti che si riscontrano in corrispondenza dei valori più bassi e più alti della distribuzione

Esempio

Esito di una analisi comparativa rispetto alla concentrazione di sodio delle acque di due zone residenziali.

Concentr.		Zona "A"	Zona "B"
3.0	3.4	36	24
3.4	3.8	48	30
3.8	4.2	63	34
4.2	4.6	39	39
4.6	5.0	22	46
5.0	5.4	19	52
5.4	5.8	13	43
5.8	6.2	10	32
		250	300

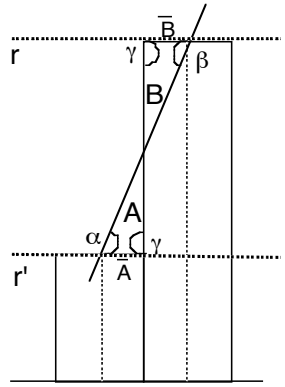


Le differenze sono forti sia al centro che nelle code segno che la concentrazione di sodio segue meccanismi diversi nelle due zone.

Area sottesa

Quando le ampiezze delle classi sono uguali, l'area sottesa al poligono è pari ad "1" (oppure α)

a) due rette parallele: r e r' formano con una trasversale coppie di angoli alterni uguali: α e β;
b) Gli angoli indicati con "γ" sono uguali perché entrambi retti.
c) i segmenti $\overline{A}, \overline{B}$ sono uguali perché le classi hanno ampiezze uguali
d) I triangoli A e B sono uguali perché hanno in comune un lato e i due angoli ad esso adiacenti.



Ciò che dell'istogramma è escluso è pari a ciò che di esterno è incluso

Esempio

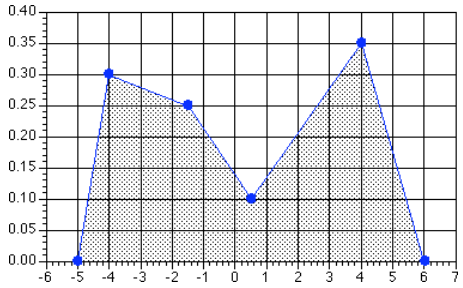
Variazioni di una quotazione azionaria

Xi	ni	fi	ci
da -5 a -3	30	0.30	-4.0
da -3 a 0	25	0.25	-1.5
da 0 a 1	10	0.10	0.5
da 2 a 6	35	0.35	4.0
	100		

L'area sottesa NON è pari ad uno perché le classi hanno ampiezza diversa

Punti convenzionali

ascisse	Ordinate
-5.0	0.00
-4.0	0.30
-1.5	0.25
0.5	0.10
4.0	0.35
6.0	0.00



Frequenze cumulate

Hanno senso solo per variabili almeno su scala ordinale $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ **e di fatto ordinate**

Frequenza assoluta cumulata

$$N_i = n_1 + n_2 + \ldots + n_i = \sum_{j=1}^{i} n_j \qquad (i=1,2,\ldots,k) \qquad \text{con} \quad N_k = n$$

indica il numero complessivo di unità che presentano modalità minore ("precedente") o uguale alla Xi.

Frequenza relativa cumulata $F_i = \frac{N_i}{N_k} = \frac{N_i}{n}$ con $F_k = 1$

indica la frazione di unità che presentano modalità minore ("precedente") o uguale alla Xi.

In caso di modalità tutte distinte, le frequenze cumulate sono date dalla formula:

$$F_i = \frac{i}{n}; \quad i = 1,2,\ldots,n$$

Formule per le frequenze cumulate

Le frequenze cumulate verificano le seguenti relazioni:

$$F_1 = f_1$$

$$F_i = F_{i-1} + f_i \qquad i = 2,3,\ldots,k-1$$

$$F_k = 1$$

schema ricorsivo

Per comodità si pone convenzionalmente: $F_0 = f_0 = 0$

Un campione di donne è stato classificato secondo l'età (in anni) al primo matrimonio. Calcolo delle frequenze cumulate

X_i		n_i	f_i	N_i	F_i
14	17	6	0.0789	6	0.0789
18	21	11	0.1447	6+11=17	0.2237
22	25	29	0.3816	6+11+29=17+29=46	0.6053
26	29	24	0.3158	6+11+29+24=46+24=70	0.9211
30	33	4	0.0526	6+11+29+24+4=70+4=74	0.9737
34	33	2	0.0263	6+11+29+24+4+2=74+2=76	1.0000
		76			

Uso delle frequenze cumulate

Servono a calcolare la frazione di unità compresa tra due qualsiasi modalità X_r<X_s :

$$A(X_r, X_s) = F_s - F_r = \sum_{j=1}^{s} f_j - \sum_{j=1}^{r} f_j = f_r + f_{r+1} + \ldots + f_s$$ ***Estremi inclusi***

$$A(X_r, X_s) = F_{s-1} - F_{r-1} = \sum_{j=1}^{s-1} f_j - \sum_{j=1}^{r-1} f_j = f_{r+1} + f_{r+2} + \ldots + f_{s-1}$$ ***Estremi esclusi***

Rileva l'inclusione o l'esclusione degli estremi solo se la "X" è discreta

Servono a calcolare le frequenze retrocumulate: se $X_s = X_{max}$ allora:

$$F_s - F_i = 1 - F_i = 1 - \sum_{j=1}^{i} f_i = f_k + f_{k-1} + \ldots + f_{i+1}$$

Esempio sulle cumulate

In uno studio demografico sulla regione Lombardia si era interessati alle famiglie in cui nessun componente svolgeva lavoro retribuito. In particolare interessava la classificazione per numero di componenti.

Modalità x_i	Frequenze Assolute n_i	Frequenze Relative f_i	Frequenze R. Cumul. F_i	Frequenze R. R. Cumul. G_i
1	261449	0.4866	0.4866	0.5134
2	222323	0.4138	0.9004	0.0996
3	37377	0.0696	0.9699	0.0301
4	10778	0.0201	0.9900	0.0100
5	3684	0.0069	0.9968	0.0032
6	1073	0.0020	0.9988	0.0012
7	376	0.0007	0.9995	0.0005
8 e più	246	0.0005	1.0000	0.0000
	537306	1.0000		

37377 famiglie di 3 componenti pari al 6.96% del totale

Il 96.99% delle famiglie ha, al più, 3 componenti

Il 3.01% ha, almeno, 3 componenti

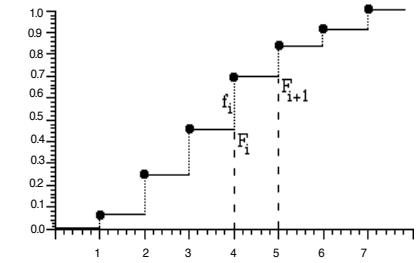
La funzione di ripartizione empirica

E' la sintesi delle frequenze relative cumulate:

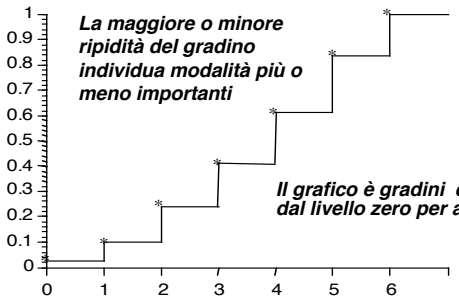
$$F(X)=\begin{cases}0 & \text{se } X<X_{(1)}\\ F_i & \text{se } X_{(i)}\le X\le X_{(i+1)} \text{ per } i=1,2,\ldots,n-1\\ 1 & \text{se } X\ge X_{(n)}\end{cases}$$

Associa ad ogni X la frazione di unità che, complessivamente, ha presentato una modalità inferiore o uguale ad X.

Essa ha un grafico a scala con gradini che si alzano dal livello zero per arrivare al livello 1. Quindi la F(x) è una funzione non decrescente:



Considerazioni



X_i	n_i	f_i	F_i
0	2	0.0250	0.0250
1	6	0.0750	0.1000
2	11	0.1375	0.2375
3	14	0.1750	0.4125
4	16	0.2000	0.6125
5	18	0.2250	0.8375
6	13	0.1625	1.0000
	80	1.0000	

La funzione di ripartizione empirica è non decrescente:

$$\text{Se } X_2 > X_1 \Rightarrow F(X_2)\ge F(X_1)$$

E' inoltre continua solo a destra in quanto a sinistra si verifica un salto

$$F(X_i-\varepsilon)=F_{i-1};\ F(X_i+\varepsilon)=F_i$$

La F.R.E. per dati in classi

$$F(X)=\begin{cases}0 & \text{se } X<L_1\\ F_{i-1}+h_i\left[X-L_i\right] & \text{se } L_i\le X<U_i \text{ per } i=1,2,\ldots,k-1\\ 1 & \text{se } X\ge U_k\end{cases}$$

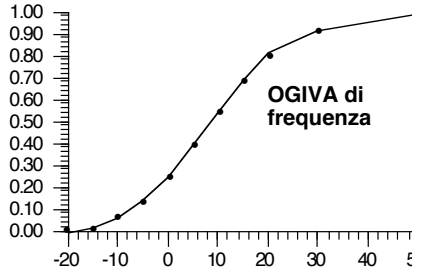
$$h_i=f_i/d_i$$

Si ipotizzano:

- **Classi chiuse a sinistra e aperte a destra**
- **Unità sono collocate uniformemente nella classe**

X_i		n_i	f_i	F_i
-20	-15	7	0.0159	0.0159
-15	-10	21	0.0478	0.0638
-10	-5	33	0.0752	0.1390
-5	0	49	0.1116	0.2506
0	5	62	0.1412	0.3918
5	10	64	0.1458	0.5376
10	15	70	0.1595	0.6970
15	20	52	0.1185	0.8155
20	30	45	0.1025	0.9180
30	50	36	0.0820	1.0000
		439	1.0000	

Per le discrete o dense tale grafico non è corretto, ma usato



Esempio

Campione di degenti classificati per tempo trascorso tra ricovero e fase acuta della malattia

X_i		n_i	f_i	F_i	X_i		n_i	f_i	F_i
4	5	2	0.0024	0.0024	18	19	73	0.0884	0.8498
6	7	13	0.0157	0.0081	20	21	40	0.0484	0.8982
8	9	40	0.0484	0.0665	22	25	37	0.0448	0.9430
10	11	131	0.1586	0.2251	26	29	27	0.0327	0.9757
12	13	192	0.2324	0.4575	30	35	16	0.0194	0.9952
14	15	152	0.1840	0.6415	36	43	4	0.0048	1.0000
16	17	99	0.1199	0.7614			826	1.0000	

Il "blocco'" centrale delle unità si colloca tra i 13 ed i 19 giorni:

In questo tratto l'ogiva ha la sua massima ripidità

La funzione di ripartizione è anche definita per valori inferiori a 4 (ha però sempre valore zero)

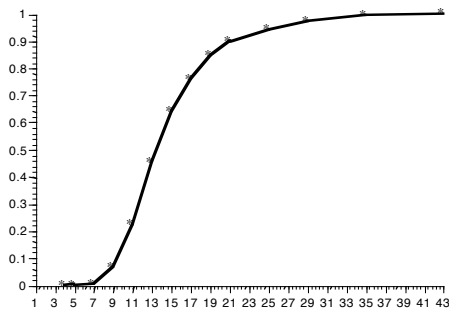


Grafico quantile (Quantile Plot)

E' una evoluzione dell'ogiva di frequenza per variabili continue o dense rilevate con modalità non raggruppate.

In ascissa si riportano le modalità osservate ed in ordinata si inserisce una trasformata lineare delle frequenze cumulate detta punto grafico (plotting position).

$$p_i = \frac{i - a}{n + b}$$

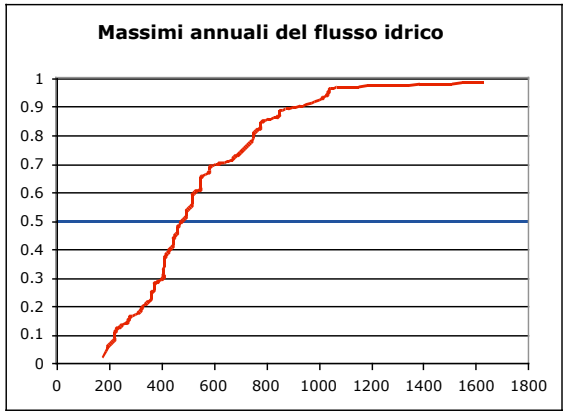
Per costruzione i punti grafico sono compresi ta zero ed uno.

Si ipotizza che la rilevazione sia accurata al punto da evitare valori ripetuti

Riferimento	a	b	p_i
Weibull	0	1	$\frac{i}{n+1}$
Blom	0.375	0.25	$\frac{i-0.375}{n+0.25}$
Cunnane	0.4	0.2	$\frac{i-0.4}{n+0.2}$
Gringorten	0.44	0.12	$\frac{i-0.44}{n+0.12}$
Hazen	0.5	0.0	$\frac{i-0.5}{n}$
Landwehr	0.35	0.0	$\frac{i-0.35}{n}$
Barnett	0.3	0.4	$\frac{i-0.3}{n+0.4}$
Tukey	0.333	0.333	$\frac{i-1/3}{n+1/3}$

i	Xi	pi			
1	183.00	0.020	26	500.00	0.510
2	193.00	0.039	27	503.00	0.529
3	206.00	0.059	28	524.00	0.549
4	227.00	0.078	29	524.01	0.569
5	229.00	0.098	30	527.00	0.588
6	242.00	0.118	31	557.00	0.608
7	276.00	0.137	32	559.00	0.627
8	286.00	0.157	33	559.01	0.647
9	324.00	0.176	34	589.00	0.667
10	336.00	0.196	35	594.00	0.686
11	364.60	0.216	36	672.00	0.706
12	366.00	0.235	37	695.00	0.725
13	379.00	0.255	38	717.00	0.745
14	381.00	0.275	39	742.00	0.765
15	412.00	0.294	40	755.00	0.784
16	412.01	0.314	41	761.00	0.804
17	415.00	0.333	42	785.00	0.824
18	419.00	0.353	43	789.00	0.843
19	420.00	0.373	44	858.00	0.863
20	437.00	0.392	45	860.00	0.882
21	452.00	0.412	46	959.00	0.902
22	452.01	0.431	47	1018.00	0.922
23	466.80	0.451	48	1043.00	0.941
24	469.00	0.471	49	1054.00	0.961
25	481.00	0.490	50	1629.00	0.980

Esempio



E' evidente il forte allungamento del grafico per i valori grandi.

Da notare che il grafico si mantiene lineare fino al punto grafico 0.7 corrispondente al livello di 672 m^3/s

La funzione di ripartizione complementare

RICORRE NELLO STUDIO
Della distribuzione dei redditieri per importo posseduto

Dell'andamento di unità sopravviventi dopo un certo decorso del tempo sperimentale

E' definita come il complemento ad uno della funzione di ripartizione:

$$G(X) = 1 - F(X)$$

La funzione G(x) è costruita con le frequenze retrocumulate ed esprime perciò la frazione di unità che ha presentato un valore almeno uguale ad "X".

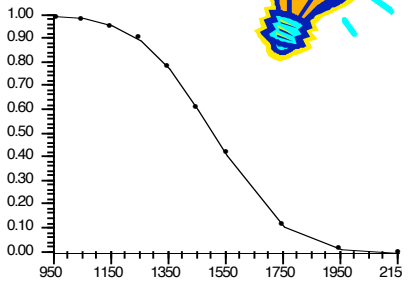
La sua rappresentazione grafica è simile alla curva di ripartizione solo che ora i punti dell'ogiva hanno coordinate (L_i, G_i).

Esempio

Durata (in ore) di un campione di lampadine.

Rappresentazione della funzione di ripartizione complementare.

X_i	n_i	f_i	F_i	G_i
950 - 1050	4	0.0133	0.0133	0.9867
1050 - 1150	9	0.0300	0.0433	0.9567
1150 - 1250	19	0.0633	0.1067	0.8933
1250 - 1350	36	0.1200	0.2267	0.7733
1350 - 1450	51	0.1700	0.3967	0.6033
1450 - 1550	58	0.1933	0.5900	0.4100
1550 - 1750	90	0.3000	0.8900	0.1100
1750 - 1950	29	0.0967	0.9867	0.0133
1950 - 2150	4	0.0133	1.0000	0.0000
	300	1.0000		



Rispetto alle funzione di ripartizione l'ogiva ha solo cambiato inclinazione

La funzione di graduazione empirica

Un aspetto rilevante delle frequenze cumulate è la funzione di ripartizione inversa o funzione di graduazione

$$X_p = \begin{cases} X_{(1)} & \text{se } p = 0 \\ \text{Minimo}\{X|F(x) \geq p\} & \text{se } 0 < p \leq 1 \end{cases}$$

Lega la frazione "p" con X_p detta percentile di ordine "p", a cui è associata la frequenza cumulata più piccola fra tutte quelle che hanno una frequenza relativa cumulata maggiore o uguale a p.

E' evidente che $F(X_p) = p$ **Se F è invertibile allora** $X_p = F^{-1}(p)$

La funzione di graduazione empirica/2

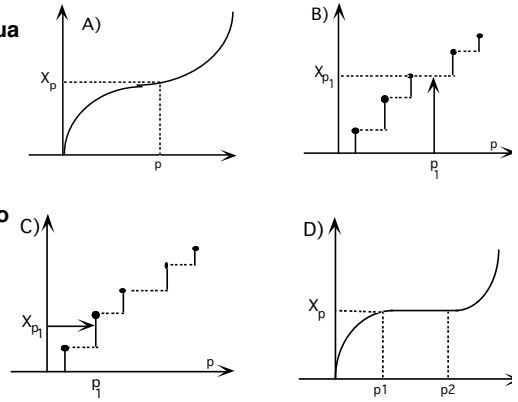
E' positiva e non decrescente ed ha un grafico identico a quello della funzione di ripartizione empirica, ma con gli assi traslati.

A) Situazione ideale: curva continua e relazione biunivoca

B) Non esiste il percentile;

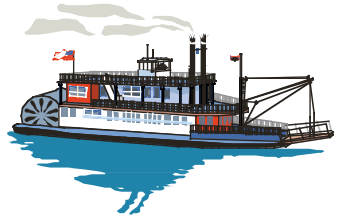
C) Ne esiste più d'uno;

D) Lo stesso percentile è associato a più di una frazione



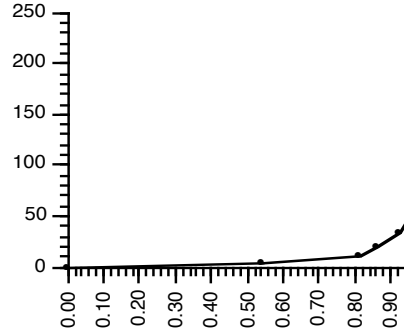
Esempio

Naviglio a motore per classi di stazza lorda



X_i		n_i	f_i	F_i
	<4	12234	0.5412	0.5412
4	10	6192	0.2739	0.8151
10	20	1061	0.0469	0.8620
20	35	1392	0.0616	0.9236
35	50	761	0.0337	0.9573
50	100	577	0.0255	0.9828
100	150	288	0.0127	0.9955
	>150	99	0.0045	1.0000
		22 604	1.0000	

La curva di graduazione molto appiattita rivela una preponderanza delle classi più piccole



La funzione di graduazione empirica

Un aspetto rilevante delle frequenze cumulate è la funzione di ripartizione inversa o funzione di graduazione

$$X_p = (1-\gamma)X_{(i)} + \gamma X_{(i+1)} \;;\; 0 < p \le 1$$

dove i=[n*p] è l'intero più grande minore di g=n*p.

Se la variabile è discreta allora: $\gamma = \begin{cases} 1 & se \;\; i < g \\ 0 & se \;\; i = g \end{cases}$

Esempio:

$n = 17;\; p = 0.65; \quad n*p = 17*0.65 = 11.05,\; i = [n*p] = [11.05] = 11;$

$\gamma = 1; \quad X_{0.65} = X_{(12)}$

La funzione di graduazione/2

Nel caso di variabili continue, la solita ipotesi di uniformità nelle classi implica:

$$X_p = (1-\gamma)L_i + \gamma U_i; \qquad F_{i-1} \le p < F_i; \;\; \gamma = \frac{p - F_{i-1}}{F_i - F_{i-1}}$$

E' positiva e non decrescente ed ha un grafico identico a quello della funzione di ripartizione empirica, ma con gli assi traslati:

X_i		n_i	f_i	F_i
	<4	12234	0.5412	0.5412
4	10	6192	0.2739	0.8151
10	20	1061	0.0469	0.8620
20	35	1392	0.0616	0.9236
35	50	761	0.0337	0.9573
50	100	577	0.0255	0.9828
100	150	288	0.0127	0.9955
	>150	99	0.0045	1.0000
		22 604	1.0000