STATISTICA

corso reiterato per D.E.S./V.=O

(prof. A. TARSITANO)

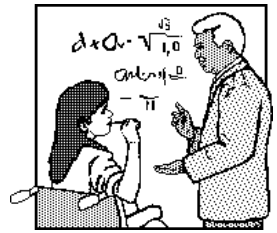
Le informazioni sul corso sono reperibili nel sito

http://www.ecostat.unical.it/Tarsitano/Coreista.htm

La statistica

La statistica è una scienza che raccoglie tutti i metodi e le tecniche che hanno come obiettivo

- **LA SCOPERTA**
- **LA NEGAZIONE**
- **L'ESTRAZIONE**



Del contenuto *informativo* di un insieme di dati

Elementi costitutivi del Dato

La statistica è centrata sul dato che studiamo nei suoi elementi costitutivi:

- **L'UNITA' SU CUI E' RILEVATO**
- **LA VARIABILE STUDIATA**
- **LA SCALA DI MISURAZIONE**
- **IL CRITERIO ORGANIZZATIVO**

ESEMPIO

Nell'idea che i disavanzi delle aziende pubbliche si concentrino in particolari regioni a fianco c'è la tabella che li riporta, in milioni, per alcune regioni.

La caratterizzazione dei dati è ora: {Regione, Disavanzo, Milioni di lire, Ordinamento alfabetico};

Regioni	Disavanzo
Abruzzi	110558
Calabria	49991
Campania	2189901
Emilia R.	478704
Lazio	2739464
Liguria	378193
Lombardia	1111113
Marche	83445
Piemonte	342798
Puglia	360113
Toscana	562888
Umbria	143723
Veneto	600062
Totale	**9150955**

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

Lo spazio dei dati

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

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

- Quantitativo
  - Continuo
  - Discreto
  - Ordinale
- Qualitativo

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

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

La matrice dei dati

Una rilevazione consiste nella osservazione delle variabili sulle unità

Le osservazioni sono i vettori $X_i,\quad i = 1,2,\ldots,n$

I cui valori 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

indagine sul *self-service* di una biblioteca

meta-dato

n=20
m=5

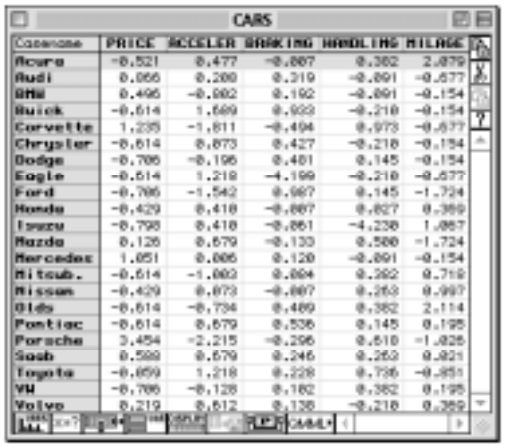
Nome	N.libri	Tempo	Posiz.	Corso	Giudizio
A.C.	6	6	Coll.		Medio
A.R.	10	6	4	DES	Medio
A.G.	6	11	Doc		Pessimo
A.T	5	1	FC	EA	Medio
D.I.	6	5	Dlp		Pessimo
D.S.	7	8	FC	SSA	Medio
F.D.	11	5	Doc		Ottimo
G.A.	1	4	2	DUS	Ottimo
G.G.	10	1	3	DES	Buono
G.L.	2	1	Est.		Medio
G.P.	8	6	4	SSA	Pessimo
G.S.	4	12	Imp		Cattivo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

STATISTICA

Caratteristiche di alcune auto:

m=5 variabili per n=22 unità.



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

Microdati e macrodati

L'unità per cui si cercano i dati (unità di rilevazione) non sempre coincide con quella oggetto di studio (unità di indagine)

Esempio:
La rilevazione delle scuole materne può essere effettuata per comuni, ma essere poi elaborata per provincie

I microdati sono i valori riferiti all'unità elementare che non può essere ulteriormente scomposta.

I macrodati sono i valori ottenuti o direttamente o dalla aggregazione di più dati elementari.

I microdati sono un sistema di rilevazione comodo quando non si è sicuri della scala di aggregazione che poi potrà servire

I metadati

Sono codici che identificano in modo sintetico e senza ambiguità le unità

Esempi:
Se si tratta di persone il record include nome e cognome e altre informazioni età, sesso, professione

Nel caso di imprese: settore produttivo, forma societaria, dipendenti, sede degli stabilimenti.

Per daati territoriali è inserito il riferimento geografico delle unità.

I metadati sono dei dati per accedere ad altri dati. Sono il mezzo di contatto tra riclevazioni diverse sulle stesse unità

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 dati mancanti

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

Derivano anche da mancata rilevazione o rilevazione manife- stamente sbagliata.

L'elaborazione dei dati non consente vuoti nelle celle. Se mancano i dati si adotta un codice convenzionale

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 dei dati

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

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	Centro	9149	2.7	9396	4.8	38800	11.4	57345	6.5
regioni	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

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'attenzione è limitata alla descrizione del *data set*

- **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 faciliti l'inputazione e riduca 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 (in milioni di lire) 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 ascendente dei valori.

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	

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

Esempio di elaborazione con “Statistica”

File: Esp95_Xls4.exl
MISS=-9999.00
Include all cases
size: 94 * 2

STATISTICA BASIC STATISTICS
Frequency Table; Variable: V1
Interval Method: All values
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
4.500	2	2.13	17	18.09
4.700	5	5.32	22	23.40
5.000	2	2.13	24	25.53
5.300	5	5.32	29	30.85
5.700	1	1.06	30	31.91
5.800	1	1.06	31	32.98
5.900	3	3.19	34	36.17
6.000	1	1.06	35	37.23
6.100	4	4.26	39	41.49
6.400	4	4.26	43	45.74
6.600	3	3.19	46	48.94
6.700	2	2.13	48	51.06
6.800	1	1.06	49	52.13
6.900	3	3.19	52	55.32
7.000	1	1.06	53	56.38
7.100	3	3.19	56	59.57
7.200	1	1.06	57	60.64
7.300	2	2.13	59	62.77
7.500	2	2.13	61	64.89
7.600	2	2.13	63	67.02
7.700	2	2.13	65	69.15
7.800	2	2.13	67	71.28
8.000	3	3.19	70	74.47
8.300	3	3.19	73	77.66
8.400	1	1.06	74	78.72
8.500	2	2.13	76	80.85
8.600	2	2.13	78	82.98
8.700	2	2.13	80	85.11
8.800	2	2.13	82	87.23
9.000	4	4.26	88	89.36
9.300	1	1.06	89	94.68
9.500	1	1.06	90	95.74
9.600	1	1.06	91	96.81
9.800	1	1.06	92	97.87
9.900	1	1.06	94	100.00
10.200	2	2.13		

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

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).

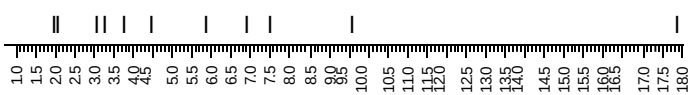
In prossimità del valore più vicino ad ogni modalità si riporta un simbolo (di solito un punto) di dimensione prefissata conforme alla dimensione della linea.

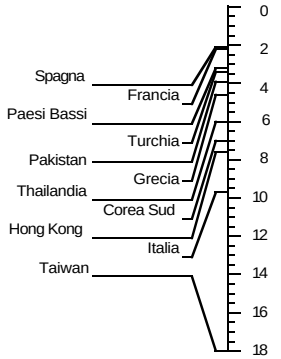
Se più modalità condividono lo stesso punto ovvero sono molto prossime, i punti saranno impilati.

ESEMPIO

Graduatoria delle falsificazioni. Volume di contraffazioni per vari Paesi.

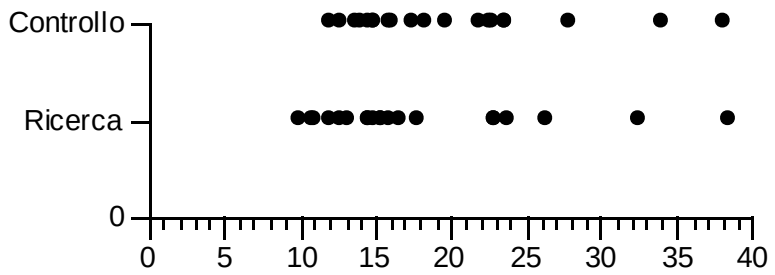
Paese	Volume	Paese	Volume	Paese	Volume	Paese	Volume
Taiwan	18.0	Pakistan	3.9	Francia	2.2	Corea Sud	7.0
Italia	9.7	Paesi Bassi	3.2	Turchia	3.4	Hong Kong	7.6
Thailandia	6.0	Spagna	2.1	Grecia	4.6		





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_i
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
0	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_i
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

Numero di rami

Esistono vari suggerimenti:

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

[.] parte intera

Se n=50

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

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 berroni che faceva agli occhi. Ero corso per tutto il mattino nella pianura che conoscevo e dalla finestra avevo intravisto le rogge alberate della mia infanzia, specchi d'acqua, branchi di oche, praterie pei. Ci savo ancora quando il treno s'era messo per ripe scosceso e dove bisognava guardare insu per vedere il cielo. Dopo una stretta galleri s'era fermato. Nell'afa nella polvere mi ritrovai sulla piazzetta della stazione, gli occhi pieni di coste calcinate, un carretti e grassi mi mostr la strada, dove salire e salire, il paese era in alto. Gettai la valigetta sul sedile posteriore dei la cimmo insieme...

da "Il Diavolo tes Colline" di C. Pavese

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

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

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)

$$X_i: \{X_i | L_i \le X \le U_i\} \quad \text{Chiuse}$$
$$X_i: \{X_i | L_i < X < U_i\} \quad \text{Aperta}$$
$$X_i: \{X_i | L_i < X \le U_i\} \quad \text{Aperta a sinistra}$$
$$X_i: \{X_i | L_i \le X < U_i\} \quad \text{Aperta a destra}$$

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

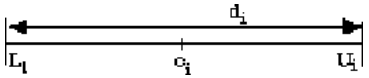
$$L_i = U_{i-1}, \quad i = 2,3,\ldots,k$$

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

Caratterizzazione delle classi

Le classi hanno due elementi importanti:

Ampiezza : $d_i = (U_i - L_i)$



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

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

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**

Densità di frequenza

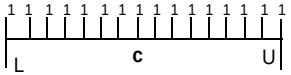
E' una utile caratteristica delle classi

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

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:



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:

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

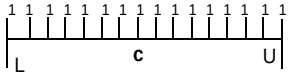
$$\text{Limite reale } U_i = \text{Limite riportato } U_i + \frac{\text{unità di arrotondamento}}{2}$$

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

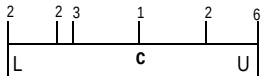
Tipicità del valore centrale

Dipende dalla configurazione con cui si presentano le modalità.

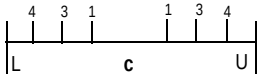
Nel caso della uniforme



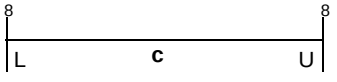
è questionabile in quanto non c'è ragione di preferire il punto di mezzo.



"c" è isolato ed esprime solo se stessa

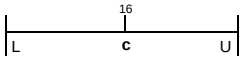


"c" è poco rappresentattivo



Non rappresenta nessuno

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



Esempio

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).

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: $L_i = L_1 + (i-1)*\delta$ per $i = 2,3,\ldots,k$
- Si pone: $U_i = L_{i+1} - \delta$ 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) $k = 1 + 3.322 * \mathrm{Log}_{10}(60) = 6.9 \cong 7;$

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

Frequenze cumulate

Hanno senso per variabili almeno su scala ordinale: $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$

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:

Schema della ricorsività →

$$F_1 = f_1$$
$$F_i = F_{i-1} + f_i \qquad i = 2,3,\ldots,k-1$$
$$F_k = 1$$

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

ESEMPIO

Donne per (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			

ESEMPIO

Nella tabella è riportata la distribuzione delle unità "giorni di apertura di un distributore" per numero di litri di liquido antigelo-antiossidante venduto.

X_i		n_i	f_i	N_i	F_i	X_i		n_i	f_i	N_i	F_i
0	8	13	0.2500	13	0.2500	33	40	5	0.0962	41	0.7885
9	16	10	0.1923	23	0.4423	41	48	4	0.0769	45	0.8654
17	24	8	0.1538	31	0.5962	49	56	4	0.0769	49	0.9423
25	32	5	0.0962	36	0.6923	57	64	3	0.0577	52	1.0000
								52	1.0000		

Nella 2ª e 3ª colonna ci sono le frequenze assolute e relative, nella 4ª e 5ª le analoghe quantità cumulate.

E' facile controllare che le relazioni indicate in precedenza sono tutte rispettate..

Frequenze retrocumulate

Con un ragionamento analogo possiamo calcolare la frazione di unità con modalità strettamente superiore a $X_{(i)}$ con la formula:

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

indica il numero complessivo di unità che presentano modalità maggiore ("successiva") alla X_i.

Esempio:

a) Richieste mensili di sostituzioni di pezzi per un elettrodomestico

Richieste	Mesi	f_i	F_i	G_i
0	24	0.333	0.33	0.67
1	19	0.264	0.60	0.40
2	10	0.139	0.74	0.26
3	8	0.111	0.85	0.15
4	7	0.097	0.94	0.06
5	3	0.042	0.99	0.01
6	1	0.014	1.00	0.00
	72	1.00		

La richiesta di almeno un pezzo è avvenuta nel 67% delle volte cioè la frequenza relativa retrocumulata associata alla modalità X=0;

Due o più richieste sono avvenute nel 40% dei mesi cioè la frequenza relativa retrocumulata associata a X=1.

ESEMPIO

Tempo in ore prima che un dispositivo elettronico mostri segni di usura.

X		n	f	F	G	r
300	499	2	0.02	0.02	0.98	0.02
500	699	8	0.08	0.10	0.90	0.08
700	899	13	0.13	0.23	0.77	0.14
900	1099	17	0.17	0.40	0.60	0.22
1100	1299	21	0.21	0.61	0.39	0.35
1300	1499	14	0.14	0.75	0.25	0.36
1500	1699	11	0.11	0.86	0.14	0.44
1700	1899	7	0.07	0.93	0.07	0.50
1900	2099	4	0.04	0.97	0.03	0.57
2100	2299	2	0.02	0.99	0.01	0.67
2300	2499	1	0.01	1.00	0.00	1.00
		100				

Le frequenze relative retrocumulate per modalità articolate in classi sono ottenute avviando direttamente l'accumulo dall'ultima classe

Nessun dispositivo ha mostrato un tempo di durata regolare di 2500 ore o più; Il 98% ha avuto una durata almeno pari a 500 ore.

Nell'ultima colonna è calcolato il rapporto $r_i = f_i/G_{i-1}$ che misura il rischio di disfunzioni per dispositivi che hanno già raggiunto l'età X_i.

L'istogramma delle frequenze

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

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

L'area totale dei rettangoli è pari ad α

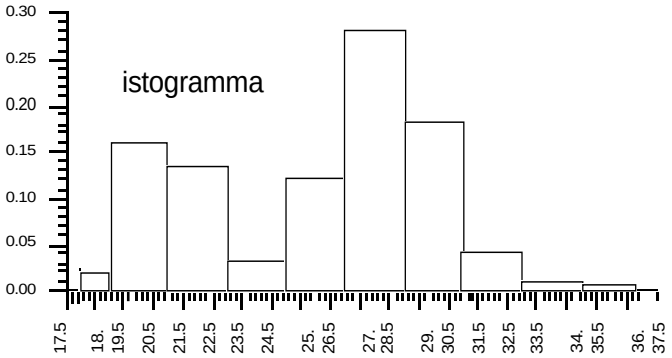
Di solito $\alpha=1$

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

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

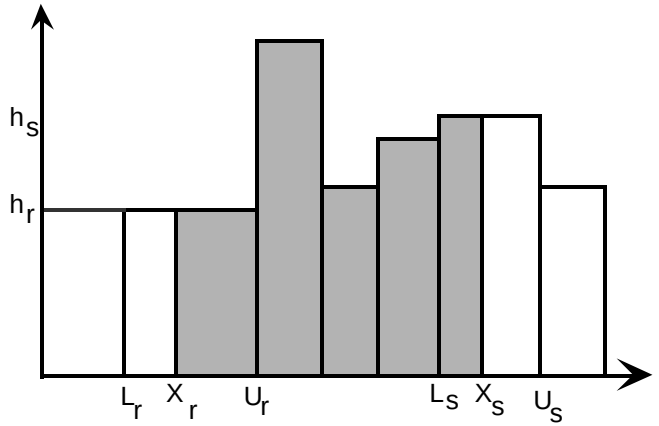


La somma bloccata permette di controllare l'area dei singoli rettangoli e quella complessiva

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

Additività

Deriva dalla natura di area della frequenza relativa



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

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		

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

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

Poligono di frequenza

Grafico che discende dall'istogramma è 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:

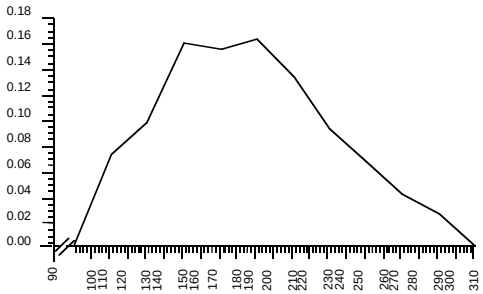
$$\left(c_1 - \frac{d_1}{2},\ 0\right);\ \left(c_k + \frac{d_k}{2},\ 0\right)$$

così il grafico parte e finisce sull'asse delle ascisse

il poligono di frequenza riporta solo il profilo esterno dell'istogramma facilitandone la percezione

Contenuto calorico

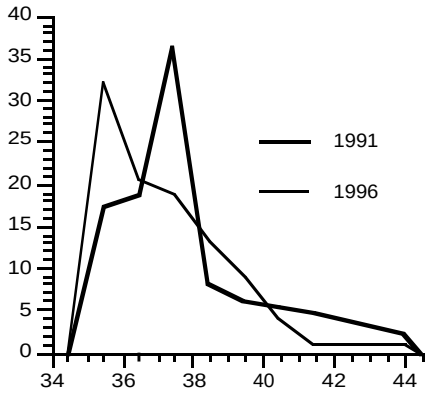
X_i		n_i	c_i	f_i
100	120	14	110	0.0714
120	140	19	130	0.0969
140	160	31	150	0.1582
160	180	30	170	0.1531
180	200	32	190	0.1633
200	220	26	210	0.1327
220	240	18	230	0.0918
240	260	13	250	0.0663
260	280	8	270	0.0408
280	300	5	290	0.0255
		196		1.0000



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

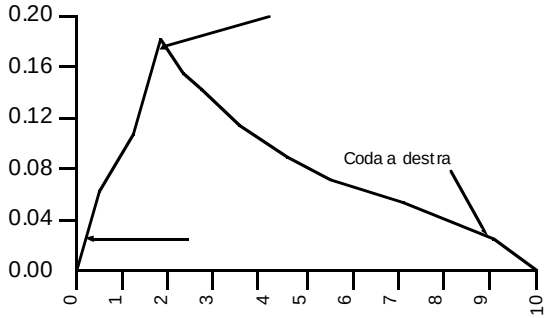


E' evidente a riduzione dell'orario più praticato: da 37 a 35 ore.

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



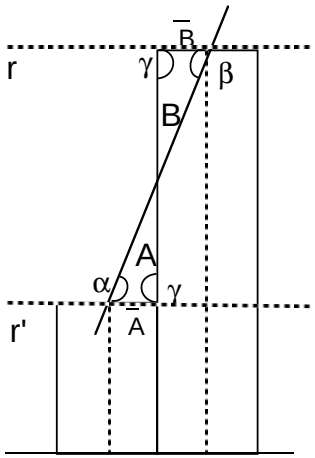
Il picco è il livello in cui la frequenza è massima.

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

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.

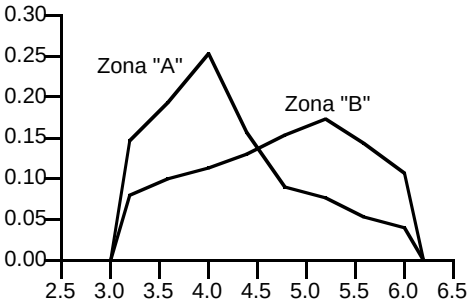


Ciò che dell'istogramma è escluso è pari a ciò che di esterno è incluso

ESEMPIO

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.

L'ogiva delle frequenze

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

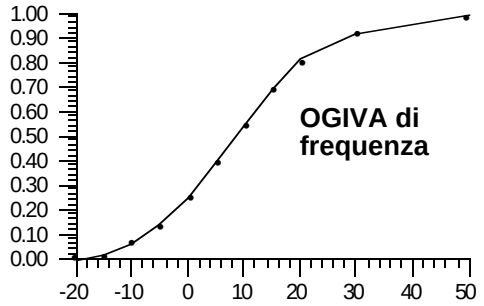
$$h_i = f_i / d_i$$

Ipotesi:

- **Le unità sono collocate uniformemente nella classe**
- **La variabile è continua**

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

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	



ESEMPIO

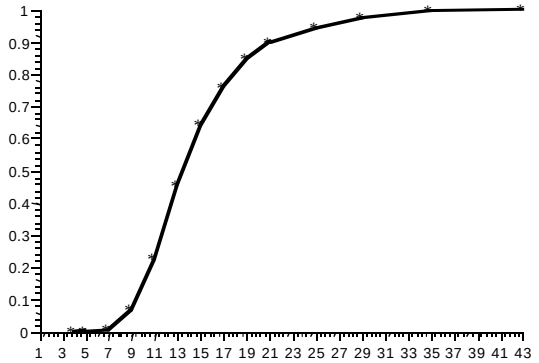
Campione di degenti classificati per tempo trascorso tra ricovero e fase acuta della malattia

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, ma qui assume valore zero



L'ogiva complementare

RICORRE NELLO STUDIO
Della distribuzione dei redditieri per importo posseduto

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

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

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".

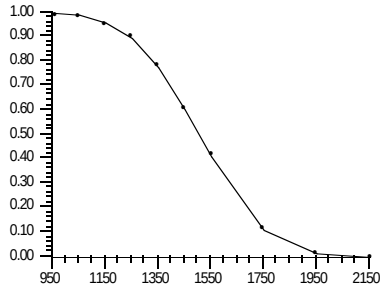
La sua rappresentazione grafica è simile alla ogiva delle frequenze solo che ora i punti hanno coordinate (L_i, G_i).

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

Si parla di funzione di sopravvivenza o di ripartizione complementare