Cos'è la statistica

- Statistiche: al plurale, sinonimo di dati
- Statistica: al singolare, è la disciplina che analizza le statistiche, i dati, cercando di estrarne informazioni utili

L'enorme mole di dati resi oggi disponibili dalla digitalizzazione, rende la statistica indispensabile in qualsiasi ambito.

Data:
venerdì 07.08.2009

Il Sole 24 ORE

Estratto da Pagina:
8

Con la digitalizzazione dei dati, cresce la richiesta da parte delle aziende

Tutti pazzi per gli statistici

Eliana Di Caro

Chi è colto dalla classica incertezza post esami di maturità su cosa fare della propria vita, ha una risposta: iscriversi a statistica e contemplare la partenza per gli Stati Uniti. Dove i laureati in quella facoltà li cercano come il pane e arrivano a guadagnare, al primo anno di lavoro, fino a 125mila dollari.

La spiegazione c'è, semplice e convincente: aumenta la digitalizzazione dei dati, di qualunque tipo, e di pari passo cresce la richiesta di analisti che li leggano e interpretino. Non è un caso che Ibm abbia creato, lo scorso aprile, un pool di oltre 200 matematici e statistici per l'analisi economica e l'ottimizzazione dei servizi. Un progetto che l'azienda proseguirà con il reclutamento di altri 4mila analisti.

Negli Stati Uniti, dove le tendenze si fiutano rapidamente, c'è un altro segnale che la dice lunga: i partecipanti alla conferenza annuale dell'American statistical association, che si è chiusa ieri a Washington, si stima siano stati oltre 6.400, mille in più che nel 2008. Un esercito di uomini e donne, giovani e attempati che per sei giorni si sono distinti in giro per la capitale più di qualunque altro gruppo di turisti.

LA TASK FORCE DELL'IBM

Il colosso informatico ha creato in aprile una squadra di 200 esperti nei suoi laboratori, e progetta di reclutarne altre quattromila

D'altra parte, a vedersela con dati, numeri e parametri di vario tipo non sono solo coloro che tradizionalmente hanno utilizzato queste competenze. L'archeologo e l'esperto di antropologia, il linguista e lo scienziato, tutte le professioni sono obbligate sempre più alle elaborazioni statistiche. Spinte dalla potenza di computer sofisticati e dalla mole di informazioni che il web mette a disposizione. Come ha raccontato al New York Times Carrie Grimes, che lavora per Google proprio nel campo dell'analisi dei dati, «la gente pensa che gli archeologi facciano una vita stile Indiana Jones, e invece gran parte dell'attività è di data analysis». La Grimes lo dice a ragion veduta: laureata ad Harvard in archeologia, è stata in Honduras dove ha mappato i manufatti che indicavano gli insediamenti dei Maya sul territorio. Già allora «computer e modelli matematici» erano parte importante del lavoro. Oggi ha 32 anni ed è uno dei 250 analisti del colosso californiano: il suo compito è accrescere l'efficienza del motore di ricerca, attraverso l'elaborazione statistica di una gigantesca quantità di dati.

Carrie ha fatto un dottorato in statistica a Stanford nel 2003, dunque è anche formalmente qualificata, ma si moltiplicano quelle figure particolari, che vantano conoscenze miste e tagliano trasversalmente diverse discipline (economia, matematica, informatica), reduci dalle esperienze più inusuali. Questo background, unito a tecniche avanzate di lettura dei dati, a volte conta più di una laurea. Anche perché deve affrontare le insidie della rete: il solo volume di dati veicolati dal web può facilmente sopraffare i modelli statistici.

È troppo parlare di *beautiful mind* nell'era di internet, pur senza il genio del matematico John Nash? Sembra di no, a sentire l'entusiasmo di Hal Varian, chief economist di Google, 62 anni e una carriera brillante costruita sui numeri: «Non mi stancherò mai di ripetere che il mestiere più attraente per i prossimi dieci anni sarà quello dello statistico».

eliana.dicaro@ilsole24ore.com

Statistica e Economia

- Descrizione dello stato e dell'andamento nel tempo dei fenomeni economici
- Analisi dei comportamenti degli operatori economici
- Previsioni sulla dinamica degli aggregati economici
- Analisi dei processi e dei risultati produttivi e gestionali
- Valutazione delle condizioni del mercato
- Pianificazione delle strategie di marketing
- Scelta tra portafogli alternativi
- Previsioni sulla dinamica delle misure finanziarie

Statistica: maneggiare con cautela

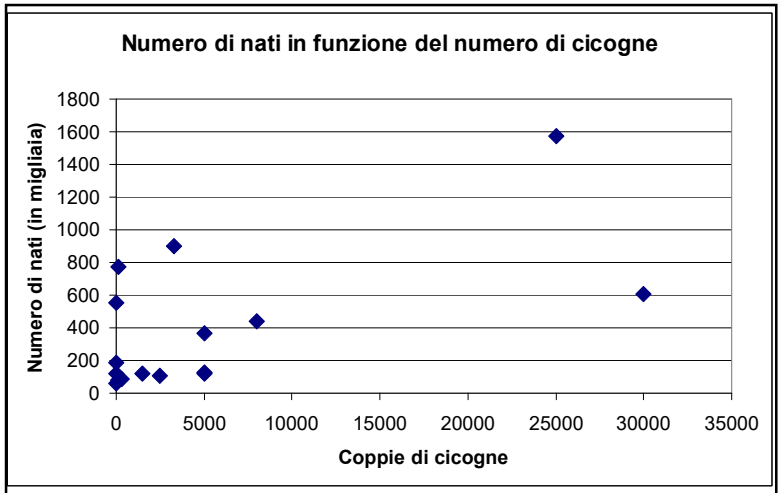
- Mark Twain: «Di solito la gente usa le statistiche come un ubriaco usa un lampione: per appoggiarsi più che come fonte di illuminazione»
- Di seguito alcuni esempi di procedimenti giusti e sbagliati per analizzare i dati.

ESEMPIO 1:

Nascite e cicogne

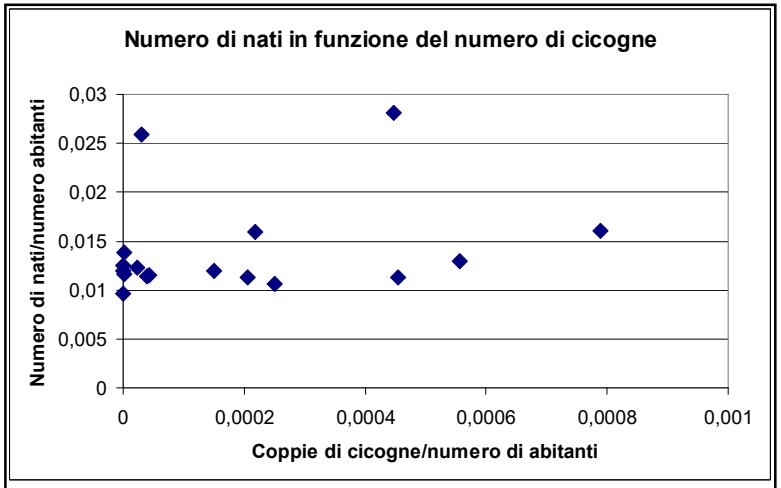
Country	Area (km^2)	Storks (pairs)	Humans (10^6)	Birth rate (10^3/yr)
Albania	28,750	100	3.2	83
Austria	83,860	300	7.6	87
Belgium	30,520	1	9.9	118
Bulgaria	111,000	5000	9.0	117
Denmark	43,100	9	5.1	59
France	544,000	140	56	774
Germany	357,000	3300	78	901
Greece	132,000	2500	10	106
Holland	41,900	4	15	188
Hungary	93,000	5000	11	124
Italy	301,280	5	57	551
Poland	312,680	30,000	38	610
Portugal	92,390	1500	10	120
Romania	237,500	5000	23	367
Spain	504,750	8000	39	439
Switzerland	41,290	150	6.7	82
Turkey	779,450	25,000	56	1576

Table 1. Geographic, human and stork data for 17 European countries



Il grafico mostra chiaramente che i paesi con più cicogne sono quelli con la natalità più elevata, avvalorando la tesi che le cicogne portino i bambini!

6



Un'analisi corretta deve tenere conto della diversa dimensione dei paesi considerati: in questo modo si conclude correttamente che non c'è relazione tra numero di cicogne e di nati.

7

ESEMPIO 2:

- "*Gli automobilisti corretti? Solo l'8%".* (Corriere della Sera del 25/08/2003)
- Dal titolo siamo portati a pensare che quasi tutti gli automobilisti sono scorretti.

8

- Invece, i dati sono i seguenti:
  (Inchiesta Altroconsumo)

	%
guidatori corretti	8
guidatori non completamente corretti	86
guidatori scorretti	6
TOTALE	100

- Un titolo altrettanto parziale, ma un po' più vero di quello scelto:
  "Gli automobilisti scorretti? Solo il 6%".
- E avrebbe trasmesso un'informazione opposta…

ESEMPIO 3:
Abusivismo, la top ten regionale

(Il Messaggero del 17/09/2003)

	Abitazioni abusive costruite nel 2002
Campania	5925
Sicilia	4260
Puglia	3820
Calabria	2919
Lombardia	1901
Lazio	1697
Veneto	1664
Sardegna	1482
Toscana	1327
Abruzzo	1252

- Nei confronti geografici i numeri assoluti sono spesso fuorvianti.
- Tra le dieci regioni della tabella figurano le sette regioni "più grandi" del paese.
- In testa a graduatorie simili non troveremo mai una "piccola" regione.

Come fare un confronto corretto?

	Abitazioni abusive costruite nel 2002	% sul totale delle abitazioni costruite
Campania	5925	28,3
Sicilia	4260	28,9
Puglia	3820	15,8
Calabria	2919	23,2
Lombardia	1901	4,4
Lazio	1697	8,1
Veneto	1664	4,5
Sardegna	1482	15,5
Toscana	1327	8,4
Abruzzo	1252	19,3

- In questo caso, rapportando in ogni regione il numero di abitazioni abusive a quello delle abitazioni costruite.

Ecco come cambia la top ten dell'abusivismo:
(Fonte: La Repubblica del 17 e 21 settembre 2003)

	Abitazioni abusive costruite nel 2002
Campania	5925
Sicilia	4260
Puglia	3820
Calabria	2919
Lombardia	1901
Lazio	1697
Veneto	1664
Sardegna	1482
Toscana	1327
Abruzzo	1252

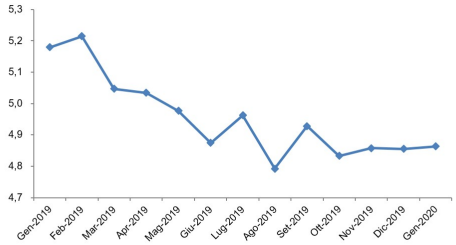
	% sul totale delle abitazioni costruite
Molise	30,7
Sicilia	28,9
Campania	28,3
Calabria	23,2
Abruzzo	19,3
Liguria	18,8
Puglia	15,8
Sardegna	15,5
Basilicata	11,6
Toscana	8,4
ITALIA	11,0

ESEMPIO 4:

- Quando, molti anni fa, la John Hopkins University (USA) iniziò ad accettare anche le donne come studenti, qualcuno pensò di riportare la notizia secondo la quale il 33,3% delle studentesse aveva sposato un insegnante.
- Ma a quell'epoca le donne iscritte erano solo tre ed **una** aveva sposato un professore

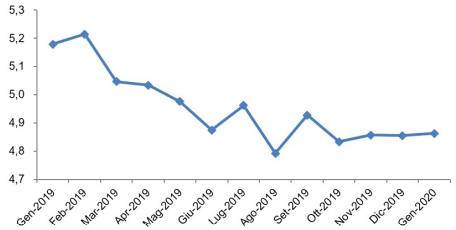
Esempio 5:

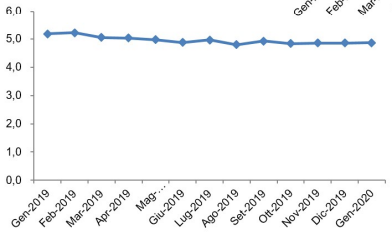
Tasso di disoccupazione, gennaio 2019 - gennaio 2020, dati destagionalizzati, valori percentuali.



... se non lo si osserva attentamente se ne può ricavare una percezione del fenomeno distorta!

I due grafici mostrano **entrambi il calo** del tasso di disoccupazione da gennaio 2019 a gennaio 2020

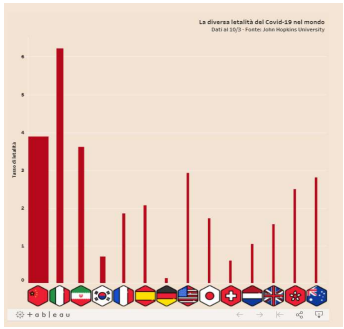




L'unica differenza è nel valore di partenza dell'asse delle ordinate, che provoca una notevole differenza di **scala**: è bastato eliminare la parte tra l'origine (zero) ed il valore 4,7 per dare un'intensità diversa alla stessa rappresentazione…

Esempio 6:

La letalità del Covid 19 in Italia (https://www.infodata.ilsole24ore.com/2020/03/15/coronavirus-piu-letale-italia-dati-spiegati-bene/)



- Le barre rappresentano il tasso di letalità, ovvero la percentuale di decessi sul totale dei contagiati, nei Paesi che, al 10 marzo 2020, avevano registrato almeno 100 casi complessivi ed almeno un decesso.
- La larghezza delle colonne fa riferimento al totale complessivo dei contagiati, dato utilizzato anche per ordinare le nazioni da sinistra verso destra.
- Come si può notare la barra più alta, che corrisponde ad un tasso di letalità maggiore, riguarda proprio l'Italia. **Dove, al 10/03/2020, il 6,2% delle persone cui è stato diagnosticato il coronavirus è morta. In Cina siamo al 3,9%, in Iran al 3,6%, in Corea del Sud addirittura allo 0,7%.** I dati sembrano dire che in Italia si muore di più. La realtà dei fatti, però, potrebbe essere diversa.

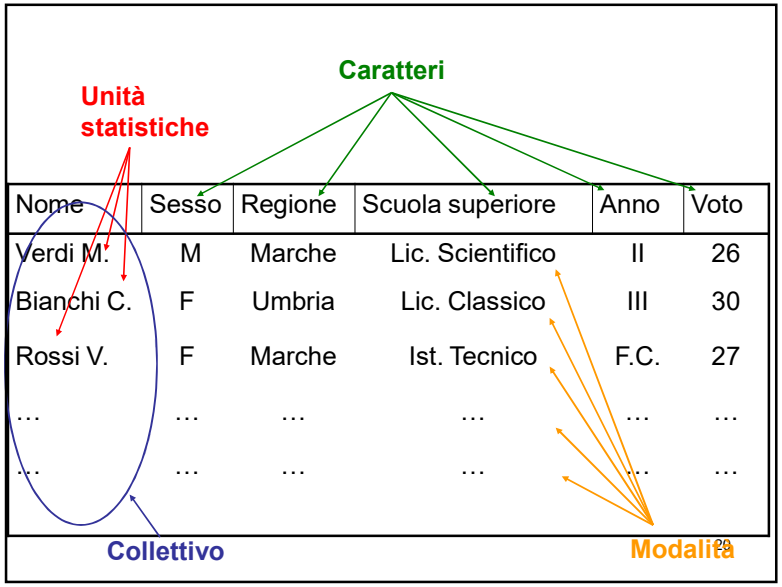
Il lessico della Statistica: qualche definizione

- Statistica: insieme dei principi ai quali dovrebbero ispirarsi la raccolta e l'elaborazione dei dati concernenti fenomeni collettivi
- Statistica descrittiva: si occupa dell'analisi di un fenomeno relativo a un certo gruppo di soggetti (popolazione) sulla base di una rilevazione completa delle informazioni (censimento). Tali informazioni vengono sintetizzate tramite opportuni indici statistici.
- Inferenza statistica: basandosi su un campione estratto dalla popolazione di interesse, trae conclusioni sull'intera popolazione

- Popolazione: insieme di riferimento del fenomeno oggetto di studio
- Unità statistica: singolo caso che compone la popolazione
- Carattere: caratteristica oggetto di rilevazione sulle unità statistiche che formano il collettivo
- Modalità di un carattere: diversi modi con cui il carattere si manifesta nelle unità statistiche

Esempio

- Fenomeno collettivo che si intende studiare: rendimento degli studenti iscritti al corso di laurea EBAM dell'Università di Macerata nell'esame di Statistica, nell'A.A. 2014-2015
- Collettivo statistico: insieme degli studenti iscritti al corso di laurea EBAM dell'Università di Macerata nell'A.A. 2014-2015 e che hanno sostenuto l'esame di Statistica in quell'anno
- Unità statistica: singolo studente iscritto al corso di laurea EBAM dell'Università di Macerata nell'A.A. 2014-2015 e che ha sostenuto l'esame di Statistica in quell'anno
- Caratteri rilevati: sesso, regione di provenienza, tipo di scuola superiore, anno di corso, voto all'esame di statistica
- Matrice di dati: tabella con numero di righe pari al numero di unità statistiche e numero di colonne pari al numero di caratteri rilevati



Classificazione dei caratteri

- A seconda di come sono espresse le sue modalità, un carattere viene classificato in
  - Qualitativo: quando le modalità sono espresse tramite espressioni verbali
  - Quantitativo: quando le modalità sono espresse numericamente
- Un **carattere qualitativo** può essere ulteriormente classificato come:
  - Sconnesso: non esiste un ordine naturale delle modalità
  - Ordinato: esiste un ordine naturale delle modalità

21

- Un **carattere quantitativo** può essere ulteriormente classificato come:
  - Discreto: le modalità possono essere messe in corrispondenza biunivoca con un sottoinsieme dei numeri interi
  - Continuo: si ha una corrispondenza biunivoca con l'insieme dei numeri reali
- Un **carattere quantitativo** si dice **trasferibile** quando la sua intensità può essere trasferita da un'unità all'altra

22

Esercizio

- Si considerino i seguenti caratteri statistici: (a) settore di attività economica prevalente; (b) tipo di contratti stipulati da un'agenzia assicuratrice; (c) giudizio sulla qualità della didattica in un corso di formazione professionale; (d) prezzo al Kg. di un certo prodotto alimentare; (e) giorno della settimana in cui avvengono furti di un certo tipo; (f) numero di stanze nelle abitazioni.
- Per ognuno di essi si indichino: 1. le possibili modalità; 2. la natura del carattere (se qualitativo sconnesso, ordinato ecc.); 3. il collettivo a cui può essere riferito e la corrispondente unità statistica.

23

Esercizio

Si indichi quale carattere può corrispondere ai seguenti gruppi di modalità e completare per ciascun gruppo l'elenco delle possibili ulteriori modalità: (a) nessun mezzo, ferrovia, tram, metro, autobus, mezzo proprio; (b) Piemonte, Valle d'Aosta, Liguria, Lombardia; (c) celibe/nubile, coniugato; (d) laurea, diploma, licenza media inferiore, licenza elementare.

24

Rilevazioni statistiche

- Rilevazione statistica: insieme delle operazioni necessarie per il raccoglimento dei dati necessari ad un'indagine statistica
- A seconda del **metodo**, la rilevazione può essere:
  - Sperimentale: il ricercatore controlla le condizioni sotto le quali si svolge l'osservazione
  - Osservazionale: si osserva la realtà senza intervenire su di essa
- A seconda della **complessità**, la rilevazione può essere:
  - Totale: si osservano tutte le unità della popolazione
  - Parziale: si osserva un campione estratto dalla popolazione
- Tipici **strumenti**, di rilevazione sono:
  - Intervista diretta
  - Intervista telefonica
  - Questionario postale

Esercizio

Il Dipartimento di Economia e Diritto ha necessità di ottenere informazioni riguardanti l'attitudine alla lettura degli studenti iscritti ai propri corsi di laurea e decide, per tale ragione, di realizzare un'indagine campionaria che risponda a queste esigenze.

1. Supponendo di dover partecipare alla progettazione dell'indagine, si decida: (a) qual è il collettivo statistico oggetto di studio e l'unità statistica; (b) quali sono i caratteri che si ritengono più importanti per rispondere alle esigenze informative del Dipartimento e la loro natura, assieme alle corrispondenti modalità.
2. Quale fonte di distorsione si introduce se si procede ad intervistare in modo casuale gli studenti che escono dalle lezioni.

Caso studio 1

I dati (fittizi) riportati nella tabella seguente sono estratti dall'archivio dei clienti di una banca aggiornato al 31/12/2015. Per ciascun cliente sono stati registrati: il sesso, l'età (in anni compiuti), l'ammontare del deposito nel conto corrente (in Euro) ed un giudizio sulla solvibilità (1= buona, 2=sufficiente, 3=scarsa).

I dati grezzi non forniscono informazioni, in quanto scarsamente leggibili. Come estrarre informazioni utili dai dati?

Cliente	Sesso	N.Comp.	Deposito	Giudizio
…	…	…	…	…
1453	M	2	12.500	1
1454	M	3	43.780	2
1455	F	1	9.800	2
1456	M	2	11.345	3
1457	F	6	25.320	1
1458	M	2	12.500	1
1459	M	4	34.780	2
1460	F	1	9.800	2
1461	M	4	11.345	3
1462	F	3	25.320	1
1463	F	3	35.500	3
1464	F	2	12.500	2
1465	M	3	16.590	1
1466	M	2	5.650	2
1467	M	1	14.325	1
1468	F	2	11.080	3
1469	M	4	13.700	1
1470	F	5	19.240	2
1471	F	3	2.500	2
1472	M	2	37.340	3
…	…	…	…	…

Distribuzioni statistiche

- La rilevazione statistica produce come risultato la matrice dei dati

Nome	Sesso	Regione	Scuola superiore	Anno	Voto
Verdi M.	M	Marche	Lic. Scientifico	II	26
Bianchi C.	F	Umbria	Lic. Classico	III	30
Rossi V.	F	Marche	Ist. Tecnico	F.C.	27
Gialli F.	F	Calabria	Ist. Tecnico	II	30
Neri A.	M	Marche	Lic. Scientifico	III	28

- Ogni colonna della matrice costituisce una ***distribuzione disaggregata*** secondo un singolo carattere. Si tratta dell'elencazione delle modalità osservate per ogni una unità

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

- Una distribuzione di questo tipo si chiama ***semplice*** (rispetto ad un solo carattere) ***unitaria*** (unità per unità).

$$x_1, x_2, \ldots, x_i, \ldots, x_n$$

- Se si considerassero più caratteri congiuntamente avremmo una distribuzione ***multipla*** (es. ***doppia*** se si considerassero due caratteri)

- Per *sintetizzare* una distribuzione disaggregata si fa uso di una distribuzione di ***frequenza*** che può essere semplice o multipla

- Una ***distribuzione di frequenza semplice*** viene costruita associando a ognuna delle modalità distinte che sono state osservate,

$$x_1, x_2, \ldots, x_i, \ldots, x_k$$

la corrispondente ***frequenza assoluta*** che è pari al numero di unità statistiche che presentano quella modalità. Per la *i*-esima modalità, la frequenza assoluta viene indicata con n_i.

- Una ***distribuzione di frequenza semplice*** viene rappresentata attraverso una tabella di questo tipo

Modalità (x_i)	Frequenze (n_i)
x_1	n_1
x_2	n_2
…	…
x_i	n_i
…	…
x_k	n_k
Totale	n

Esempio

- Dalla matrice di dati sugli studenti universitari, si possono ricavare 5 distribuzioni semplici secondo i caratteri: sesso, regione, scuola di provenienza, anno di corso e voto in Statistica.

- Per il carattere sesso, le modalità distinte sono M e F con frequenze pari, rispettivamente, a 2 e 3. La corrispondente distribuzione di frequenza è quindi:

Sesso (x_i)	Frequenze (n_i)
M	2
F	3
Totale	5

- Per gli altri 4 caratteri si ha:

Regione (x_i)	Frequenze (n_i)
Marche	3
Umbria	1
Calabria	1
Totale	5

Anno (x_i)	Frequenze (n_i)
II	2
III	2
F.C.	1
Totale	5

Scuola (x_i)	Frequenze (n_i)
Classico	1
Scientifico	2
Tecnico	2
Totale	5

Voto (x_i)	Frequenze (n_i)
26	1
27	1
28	1
30	2
Totale	5

Esercizio

Si considerino i dati del Caso Studio 1.

Per ciascuno dei caratteri presi in esame se ne descriva la natura.

Che tipo di distribuzione è quella presentata in tabella?

Per il carattere Giudizio, si costruisca la distribuzione di frequenze.

Esempio

- Distribuzione delle famiglie per numero di componenti - Regione Marche - Censimento 2001.

NUMERO DI COMPONENTI	Numero di famiglie (in migliaia)
1 persona	124,143
2 persone	149,531
3 persone	124,394
4 persone	107,992
5 persone	31,751
6 o più persone	11,663
Totale	**549,474**

Esempio

- Distribuzione doppia delle famiglie per numero di componenti e per ripartizione geografica - Censimento 2001 (dati in migliaia).

RIPARTIZIONI GEOGRAFICHE	Numero di componenti						
	1 persona	2 persone	3 persone	4 persone	5 persone	6 o più persone	Totale
Italia Nord-Occidentale	1.767,208	1.840,037	1.390,009	966,118	207,367	46,461	6.217,200
Italia Nord-Orientale	1.116,042	1.208,212	962,636	701,273	184,009	59,838	4.232,010
Italia Centrale	1.061,905	1.188,248	941,315	780,561	208,574	61,596	4.242,199
Italia Meridionale	940,888	1.100,449	935,550	1.150,759	474,806	145,822	4.748,274
Italia Insulare	541,578	568,465	476,696	537,495	191,070	55,689	2.370,993
Italia	**5.427,621**	**5.905,411**	**4.706,206**	**4.136,206**	**1.265,826**	**369,406**	**21.810,676**

Esercizio

Si considerino i dati del Caso Studio 1.

Si costruiscano le distribuzione doppie di frequenze secondo i caratteri Sesso e Giudizio e secondo i caratteri Numero di Componenti e Giudizio.

Suddivisione in classi

- Nel caso di un *carattere quantitativo che assume molte modalità* (tipicamente continuo) è conveniente considerare delle ***classi*** al posto delle singole modalità distinte.
- Ogni classe viene identificata da due ***estremi*** (*di sinistra* e *di destra*) che per la *i*-esima classe sono indicati con $c_{i-1} - c_i$.
- Le classi vanno scelte in modo che:
  - il livello di sintesi sia adeguato
  - siano tra loro disgiunte
  - siano esaustive

- Una ***distribuzione di un carattere in classi*** viene rappresentata attraverso una tabella di questo tipo

Classi ($c_{i-1} - c_i$)	Frequenze (n_i)
$c_0 - c_1$	n_1
$c_1 - c_2$	n_2
…	…
$c_{i-1} - c_i$	n_i
…	…
$c_{k-1} - c_k$	n_k
Totale	n

Esempio

- Per il carattere voto in Statistica abbiamo la seguente distribuzione in classi.

Classi di voto	c_{i-1}	c_i	Frequenze (n_i)
18-21	17,5	21,5	0
22-25	21,5	25,5	0
26-28	25,5	28,5	3
29-30	28,5	30,5	2
Totale	-	-	5

- In questo caso, al fine di avere estremi di classe coincidenti, si effettua la ***correzione per continuità***: si sottrae 1/2 all'estremo di sinistra di ogni classe e lo si aggiunge a quello di destra.

Esempio

- Popolazione residente per classi di età - Regione Marche - Censimento 2001

CLASSI DI ETÀ	Popolazione residente (in migliaia)
Meno di 15	189,811
15-24	154,228
25-34	216,417
35-44	218,079
45-54	193,915
55-64	177,476
65-74	168,371
75 e più	152,284
Totale	**1.470,581**

Esercizio

La seguente tabella riporta i dati relativi a 15 aziende agricole umbre (Id. Azienda) che hanno partecipato ad un bando per l'assegnazione di contributi da parte dell'Unione Europea. Si noti che il dato riguardante il Grado di innovazione nei processi dell'azienda (Innovazione) è stato codificato nel modo seguente: 1 = basso, 2 = medio, 3 = alto. Inoltre, il fatturato annuo di ciascuna azienda è espresso in migliaia di Euro.

a) Qual è l'unità statistica? Quali sono i caratteri rilevati? E qual è la loro natura?
b) Quali sono le modalità rilevate del carattere Grado di innovazione? Quali sono le modalità rilevate del carattere Anni di attività?
c) Costruire la distribuzione doppia rispetto al Grado di innovazione e al fatturato (classi 0-5, 5-10, 10-15).

Id. Azienda	1	2	3	4	5	6	7	8
Anni di attività	3	8	13	2	11	18	6	1
Provincia	PG	TR	TR	PG	TR	PG	PG	PG
Innovazione	1	3	2	1	1	2	1	2
Fatturato	5,1	6,8	10,3	14,7	3,5	8,9	11,3	4,5

Id. Azienda	9	10	11	12	13	14	15
Anni di attività	0	12	8	10	15	7	3
Provincia	TR	TR	PG	TR	PG	PG	TR
Innovazione	2	1	3	2	1	3	2
Fatturato	8,3	13,1	7	5,8	8,1	12,6	10,3

Esercizio

Si considerino i dati del Caso Studio 1.

Si costruisca la distribuzione doppia di frequenze secondo i caratteri Ammontare del Deposito e Giudizio. Per il carattere Deposito si considerino le classi: Fino a 15.000, 15.000-30.000 e 30.000-45.000.

Frequenze relative e percentuali

Famiglie per numero di componenti - Italia Settentrionale - Censimento 2001.

NUMERO DI COMPONENTI	Numero di famiglie (x 1000)
1 persona	2.883,250
2 persone	3.048,249
3 persone	2.352,645
4 persone	1.667,391
5 persone	391,376
6 o più persone	106,299
Totale	**10.449,210**

Famiglie per numero di componenti - Italia Meridionale e Isole - Censimento 2001.

NUMERO DI COMPONENTI	Numero di famiglie (x 1000)
1 persona	1.482,466
2 persone	1.668,914
3 persone	1.412,246
4 persone	1.688,254
5 persone	665,876
6 o più persone	201,511
Totale	**7.119,267**

	Italia Settentrionale		Italia Meridionale	
NUMERO DI COMPONENTI	Numero di famiglie	Frequenze relative	Numero di famiglie	Frequenze relative
1 persona	2.883,250	0,276	1.482,466	0,208
2 persone	3.048,249	0,292	1.668,914	0,234
3 persone	2.352,645	0,225	1.412,246	0,198
4 persone	1.667,391	0,160	1.688,254	0,237
5 persone	391,376	0,037	665,876	0,094
6 o più persone	106,299	0,010	201,511	0,028
Totale	**10.449,210**	**1**	**7.119,267**	**1**

	Italia Settentrionale		Italia Meridionale	
NUMERO DI COMPONENTI	Numero di famiglie	Frequenze percentuali	Numero di famiglie	Frequenze percentuali
1 persona	2.883,250	27,593	1.482,466	20,823
2 persone	3.048,249	29,172	1.668,914	23,442
3 persone	2.352,645	22,515	1.412,246	19,837
4 persone	1.667,391	15,957	1.688,254	23,714
5 persone	391,376	3,746	665,876	9,353
6 o più persone	106,299	1,017	201,511	2,831
Totale	**10.449,210**	**100**	**7.119,267**	**100**

- Frequenza relativa: $f_i = \frac{n_i}{n}$
- Frequenza percentuale: $p_i = \frac{n_i}{n} \cdot 100$
- Un'ovvia *proprietà* delle frequenze relative e percentuali è:

$$f_1 + f_2 + \cdots + f_k = \sum_{i=1}^{k} f_i = 1$$

$$p_1 + p_2 + \cdots + p_k = \sum_{i=1}^{k} p_i = 100$$

- Una ***distribuzione di frequenze assolute, relative e percentuali*** viene rappresentata attraverso una tabella di questo tipo

Modalità (x_i)	Frequenze assolute (n_i)	Frequenze relative (f_i)	Frequenze percentuali (p_i)
x_1	n_1	f_1	p_1
x_2	n_2	f_2	p_2
…	…	…	…
x_i	n_i	f_i	p_i
…	…	…	…
x_k	n_k	f_k	p_k
Totale	n	1	100

Esempio

- Per il carattere voto in Statistica abbiamo la seguente distribuzione in classi.

Classi di voto	Frequenze assolute (n_i)	Frequenze relative (f_i)	Frequenze percentuali (p_i)
18-21	0	0	0
22-25	0	0	0
26-28	3	0,6	60
29-30	2	0,4	40
Totale	5	1	100

Esercizio

Una popolazione di 100 individui, di cui 60 donne e 40 uomini, viene intervistata circa le attitudini al fumo ottenendo le seguenti risposte:

Uomini:0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 1 1 0 0 1 0

Donne: 1 0 0 0 0 1 0 0 1 1 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1

dove 1=fumatore e 0=non fumatore.

a) Ricavare una tabella di distribuzione doppia
b) Ricavare la distribuzione di frequenza relativa rispetto al carattere “Attitudine al fumo” separatamente per gli uomini e per le donne.
c) Determinare la percentuale di fumatori tra gli intervistati.

Esercizio

Si considerino i dati del Caso Studio 1.

Si costruisca la distribuzione di frequenze relative e percentuali secondo il carattere Giudizio, separatamente per gli uomini e per le donne.

Frequenze cumulate

Famiglie per numero di componenti - Italia Settentrionale - Censimento 2001.

NUMERO DI COMPONENTI	Numero di famiglie (x 1000)
1 persona	2.883,250
2 persone	3.048,249
3 persone	2.352,645
4 persone	1.667,391
5 persone	391376
6 o più persone	106,299
Totale	**10.449,210**

- Quante sono le famiglie con al massimo due componenti?
- Quante sono le famiglie con al massimo tre componenti?

	Italia Settentrionale				Italia Meridionale			
NUM. COMP.	Num. famiglie	Freq. ass. cum.	Freq. rel. cum.	Freq. perc. cum.	Num. famiglie	Freq. ass. cum.	Freq. rel. cum.	Freq. perc. cum.
1	2.883,250	2.883,250	0,276	27,593	1.482,466	1.482,466	0,208	20,823
2	3.048,249	5.931,499	0,568	56,765	1.668,914	3.151,380	0,443	44,266
3	2.352,645	8.284,144	0,793	79,280	1.412,246	4.563,626	0,641	64,102
4	1.667,391	9.951,535	0,952	95,237	1.688,254	6.251,880	0,878	87,816
5	391,376	10.342,911	0,990	98,983	665,876	6.917,756	0,972	97,169
6 o più	106,299	10.449,210	1	100	201,511	7.119,267	1	100
Totale	**10.449,210**				**7.119,267**			

- Frequenza assoluta cumulata :

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

- Frequenza relativa cumulata:

$$F_i = \frac{N_i}{n} = f_1 + f_2 + \cdots + f_i = F_{i-1} + f_i$$

- Frequenza percentuale cumulata:

$$P_i = \frac{N_i}{n} \cdot 100 = p_1 + p_2 + \cdots + p_i = P_{i-1} + p_i$$

- Una ***distribuzione di frequenze assolute, relative e percentuali cumulate*** viene rappresentata attraverso una tabella di questo tipo

Modalità (x_i)	Frequenze assolute cumulate (N_i)	Frequenze relative cumulate (F_i)	Frequenze percentuali cumulate (P_i)
x_1	N_1	F_1	P_1
x_2	N_2	F_2	P_2
…	…	…	…
x_i	N_i	F_i	P_i
…	…	…	…
x_k	N_k	F_k	P_k

Esercizio

Si considerino i dati del Caso Studio 1.

Qual è la frequenza di clienti che fanno parte di famiglie di al massimo 2 componenti?

Qual è la percentuale di clienti con un Ammontare di Depositi fino a 30.000 euro?

Qual è la percentuale di clienti con Giudizio di Solvibilità almeno sufficiente?

Rappresentazioni grafiche

Permettono di:

- interpretare **più velocemente le informazioni** raccolte sul fenomeno osservato
- coglierne **immediatamente** alcune **caratteristiche**

Ciò **non** significa che i grafici possano **sostituire** i numeri presenti nelle tabelle: essi forniscono un ulteriore utile supporto per l'analisi

Rappresentazione grafica o tabellare?

Alcuni vantaggi che i grafici presentano rispetto alle tabelle che corredano sono:

- **IMMEDIATEZZA**: **visualizzazione immediata dell'andamento** del fenomeno (es.: è in crescita oppure è in diminuzione?) e **della struttura** della distribuzione (es.: sono più i maschi o le femmine?), che consente **una descrizione globale** dei dati
- **SINTESI**: possibilità, in poco spazio, di confrontare più distribuzioni (curve, spezzate, ecc.)
- **SEMPLICITÀ**: il grafico è una **forma più divulgativa** per i dati statistici rispetto a quella tabellare

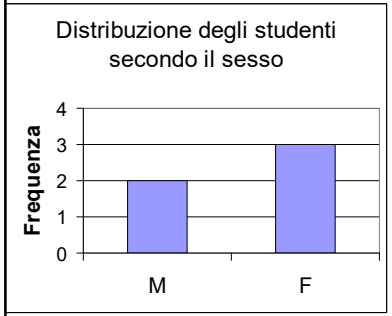
Ad ogni modo è bene ricordare che ...

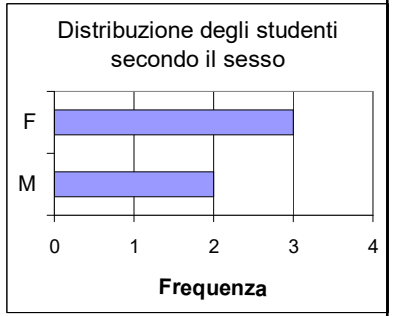
… affinché una rappresentazione grafica sia utile ed efficace deve essere corredata da tutte le informazioni necessarie alla comprensione dei dati in essa rappresentati, ovvero:

- il **titolo**, che deve indicare l'**oggetto**, il **luogo** e l'**epoca** a cui i dati si riferiscono
- il **carattere con le rispettive modalità** (es.: "maschi" e "femmine" per la variabile "sesso"), in funzione delle quali sono classificate le unità statistiche
- l'**unità di misura** impiegata per graduare gli assi
- il **tipo** di valori presentati (assoluti, percentuali,…)
- la **fonte di provenienza** dei dati

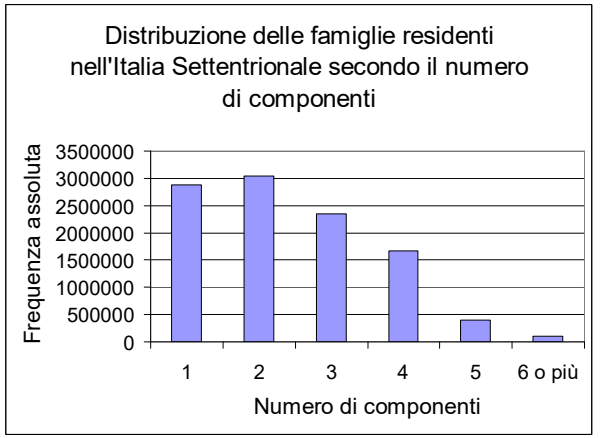
- Per una *distribuzione di frequenza* di un carattere **qualitativo** o **quantitativo discreto**, si utilizza un **grafico a barre** che consiste nel rappresentare, su un piano cartesiano, k barre di altezza $n_1, \ldots, n_k$ in corrispondenza delle ascisse $x_1, \ldots, x_k$.

Esempio

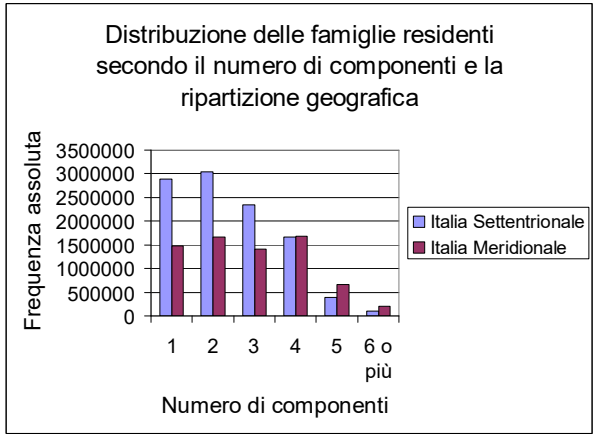




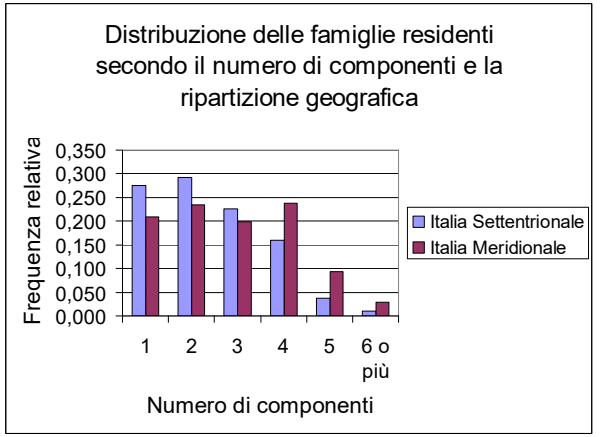
Esempi



62



63



64

Esercizio

Si considerino i dati del Caso Studio 1.

Si rappresenti graficamente la distribuzione di frequenze secondo il carattere Numero di componenti.

Si rappresenti graficamente la distribuzione di frequenze secondo il carattere Giudizio sulla solvibilità, separatamente per Sesso.

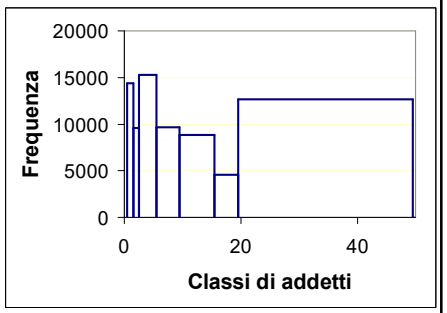
- Nel caso di un **carattere quantitativo in classi**, la distribuzione viene rappresentata tramite un **istogramma di frequenza** costruito tramite una serie di rettangoli corrispondenti alle varie classi. Il rettangolo corrispondente alla *j*-esima classe ha:
  - ***base*** pari all'ampiezza della classe: $a_i = c_i - c_{i-1}$
  - ***altezza*** pari alla densità di frequenza: $h_i = \frac{f_i}{a_i}$
- L'altezza della classe (densità) corrisponde alla frequenza che compete a un sottointervallo di ampiezza unitaria nel caso di uniforme distribuzione delle unità nelle classi. L'istogramma permette quindi di confrontare tra loro classi di diversa ampiezza.
- La caratteristica fondamentale dell'istogramma è che l'area di ogni rettangolo corrisponde alla frequenza della classe a cui si riferisce:

$$a_i \cdot h_i = a_i \cdot \frac{f_i}{a_i} = f_i$$

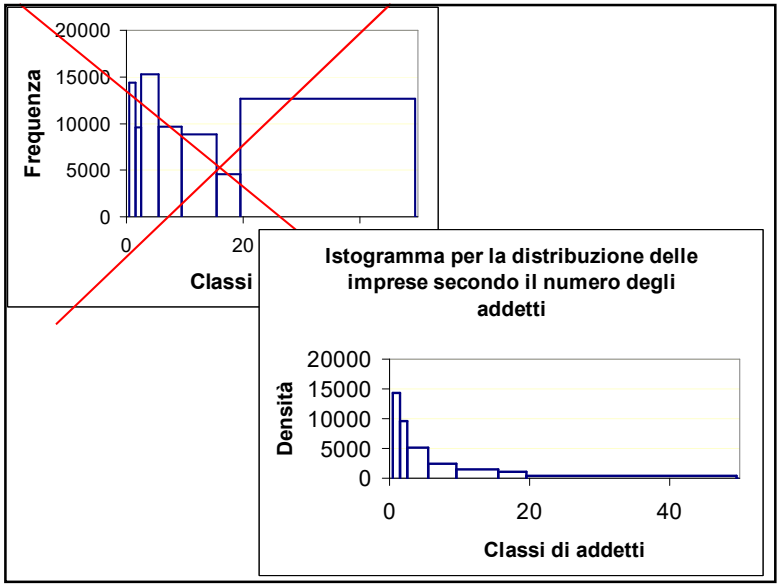
Esempio

•Distribuzione in classi delle imprese della provincia di Macerata con meno di 50 addetti, secondo il numero di addetti – Censimento 2001

Classi di addetti	Frequenze assolute
1	14.349
2	9.588
3--5	15.263
6--9	9.651
10--15	8.837
16--19	4.570
20--49	12.653
Totale:	**74.911**

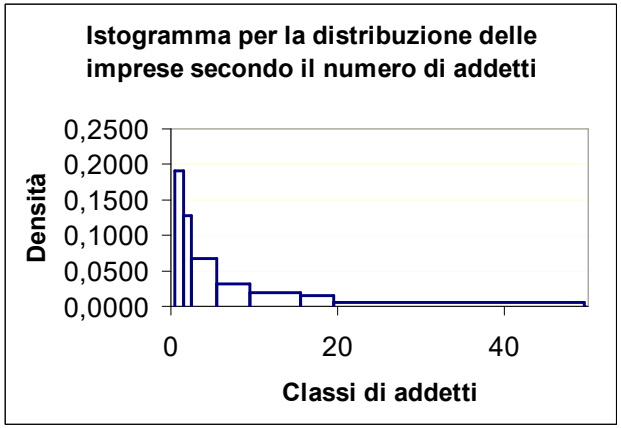


Classi di addetti	Frequenze assolute (n_i)	Ampiezza classi (a_i)	Densità di frequenza (h_i)
1	14.349	1	14.349,0
2	9.588	1	9.588,0
3--5	15.263	3	5.087,7
6--9	9.651	4	2.412,8
10--15	8.837	6	1.472,8
16--19	4.570	4	1.142,5
20--49	12.653	30	421,8
Totale:	**74.911**		



Classi di addetti	Frequenze relative (f_i)	Ampiezza classi (a_i)	Densità di frequenza (h_i)
0,5--1,5	0,1915	1	0,19155
1,5--2,5	0,1280	1	0,12799
2,5--5,5	0,2037	3	0,06792
5,5--9,5	0,1288	4	0,03221
9,5--15,5	0,1180	6	0,01966
15,5--19,5	0,0610	4	0,01525
19,5--49,5	0,1689	30	0,00563
Totale:	**1**		

70



71

Esercizio

Si considerino i dati del Caso Studio 1.

Si rappresenti graficamente la distribuzione di frequenze secondo il carattere Ammontare del Deposito. Si considerino le classi: Fino a 10.000, 10.000-20.000, 20.000-30.000 e 30.000-45.000.

72

Funzione di ripartizione

•La **funzione di ripartizione**, *F(x)* , fornisce la frequenza relativa delle osservazioni che presentano una modalità del carattere non superiore a *x*. Quindi si ha sempre:

$$F(-\infty) = \lim_{x \to -\infty} F(x) = 0 \qquad F(\infty) = \lim_{x \to \infty} F(x) = 1$$

•Per un carattere **qualitativo ordinato** o **quantitativo** non in classi, la funzione di ripartizione è pari a:

$$F(x) = \begin{cases} 0 & \forall\ x < x_1 \\ F_i & x_i \le x < x_{i+1} \\ 1 & \forall\ x > x_k \end{cases}$$

quindi, se *x* è compreso tra la modalità più piccola (x_1) e quella più grande (x_k), *F(x)* è uguale alla frequenza cumulata (F_i) corrispondente alla più grande modalità (x_i) minore o uguale a *x*. Altrimenti, *F(x) = 0* o *F(x) = 1*.

Esempio

NUMERO DI COMPONENTI (x_i)	Numero di famiglie (n_i)	Frequenze relative (f_i)	Frequenze relative cumulate (F_i)
1 persona	2.883,250	0,276	0,276
2 persone	3.048,249	0,292	0,568
3 persone	2.352,645	0,225	0,793
4 persone	1.667,391	0,160	0,952
5 persone	391,376	0,037	0,990
6 persone	106,299	0,010	1
Totale	**10.449,210**	**1**	

•F(2) = 0,568; F(4) = 0,952; F(4,5) = 0,952.

•Nel caso di un **carattere in classi** (tipicamente continuo) ci si basa sull'ipotesi che in ogni classe ci sia uniforme distribuzione: si ha sempre la stessa frequenza in ogni sottointervallo della classe di ampiezza unitaria

•In questo caso la funzione di ripartizione è pari a:

$$F(x) = \begin{cases} 0 & \forall\ x < c_0 \\ F_{i-1} + h_i(x - c_{i-1}) & c_{i-1} \le x < c_i \\ 1 & \forall\ x > c_k \end{cases}$$

e quindi, se *x* è compreso tra l'estremo sinistro della prima classe (c_0) e l'estremo destro dell'ultima classe (c_k), per il calcolo occorre innanzitutto individuare la classe che contiene *x* ($c_{i-1} - c_i$) e poi $F(x) = F_{i-1} + h_i(x - c_{i-1})$; altrimenti, *F(x) = 0* oppure *F(x) = 1*. Si noti che se *x* è uguale all'estremo inferiore di classe (c_{i-1}), si ha $F(x) = F_{i-1}$

Esempio

Imprese attive al 31.12.2010 per classi di fatturato (migliaia di euro) - dati relativi al bilanci anno 2009 - Comune di Padova

Classi di fatturato ($c_{i-1} - c_i$)	Frequenze assolute (n_i)	Frequenze relative (f_i)	Frequenze relative cumulate (F_i)	Densità (h_i)
0--250	2156	0,482219	0,482219	0,001929
250--500	666	0,14896	0,631179	0,000596
500--1.000	517	0,115634	0,746813	0,000231
1.000--2.500	539	0,120555	0,867367	8,04E-05
2.500--5.000	260	0,058153	0,92552	2,33E-05
5.000--10.000	171	0,038246	0,963766	7,65E-06
10.000--25.000	95	0,021248	0,985015	1,42E-06
25.000--50.000	32	0,007157	0,992172	2,86E-07
50.000--100.000	35	0,007828	1	1,57E-07
Totale	**4471**	**1**		

•F(500) = 0,631179; F(450) = 0,482219 + 0,000596(450-250) = 0,601387

Esercizio

Si considerino i dati del Caso Studio 1.

Si consideri la distribuzione di frequenze secondo il carattere Ammontare del Deposito. Si considerino le classi: Fino a 10.000, 10.000-20.000, 20.000-30.000 e 30.000-45.000.

Qual è il valore della funzione di ripartizione in 23.000

a) usando i dati disaggregati

b) usando la distribuzione di frequenza.

Qual è quell'ammontare del deposito al di sotto del quale troviamo il 60% dei clienti

a) usando i dati disaggregati

b) usando la distribuzione di frequenza.

Esercizio

Data la seguente distribuzione delle imprese della provincia di Terni, secondo il numero degli addetti (Fonte: Conoscere l'Umbria, Anno 2009):

Classi di addetti	Frequenze (n_i)
1	38.813
2--5	23.707
6--9	3.914
10--19	2.512
20 e oltre	1.290

1. Si calcolino le frequenze relative, percentuali, relative cumulate e percentuali cumulate.
2. Si rappresenti graficamente l'istogramma di frequenza chiudendo l'ultima classe a 100.
3. Si calcoli, sotto l'ipotesi di uniforme distribuzione nelle classi:
   (a) la frequenza relativa delle imprese con un numero di addetti compreso tra 15 e 30, utilizzando la funzione di ripartizione;
   (b) la frequenza relativa delle imprese con un numero di addetti compreso tra 1 e 7, utilizzando le densità di frequenza e riportando il risultato nell'istogramma;
4. Se facessimo un'unica classe con estremi 1-5, varrebbe l'ipotesi di uniforme distribuzione all'interno della classe?

Rappresentazione grafica della funzione di ripartizione

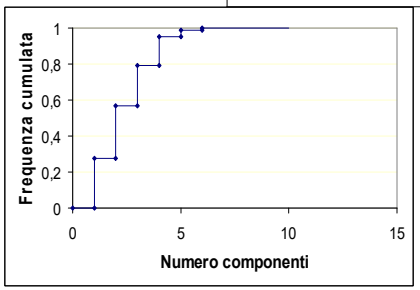
•Rappresentando la funzione di ripartizione si mostra l'andamento delle frequenze cumulate al variare della modalità del carattere.

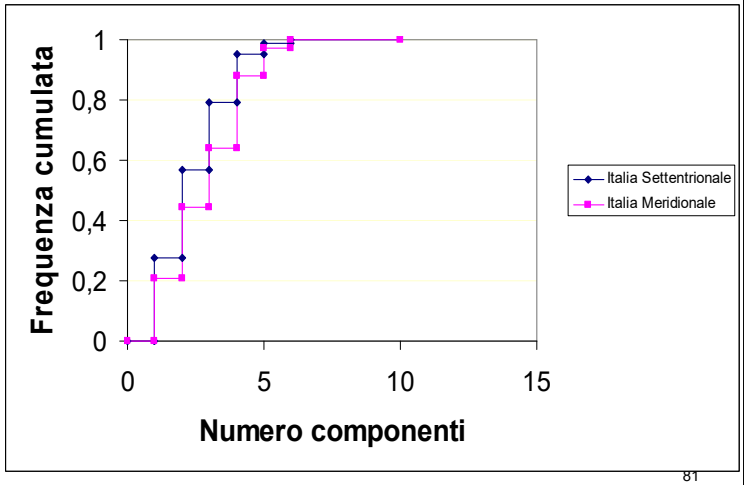
•Per un carattere **qualitativo ordinato** o **quantitativo** la funzione di ripartizione ha una forma "a gradini" ottenuta congiungendo i punti di coordinate (x_i , F_i), per $i = 1,\ldots,k$.

•Nel caso di un carattere in classi si congiungono i punti i coordinate (c_i ,F_i) , $i = 1,\ldots,k$, e il punto (c_0,0) dando origine a una "spezzata".

Esempi

NUMERO DI COMPONENTI (x_i)	Numero di famiglie (n_i)	Frequenze relative (f_i)	Frequenze relative cumulate (F_i)
1 persona	2.883,250	0,276	0,276
2 persone	3.048,249	0,292	0,568
3 persone	2.352,645	0,225	0,793
4 persone	1.667,391	0,160	0,952
5 persone	391,376	0,037	0,990
6 persone	106,299	0,010	1
Totale	**10.449,210**	**1**	





81

Esercizio

Si considerino i dati del Caso Studio 1.

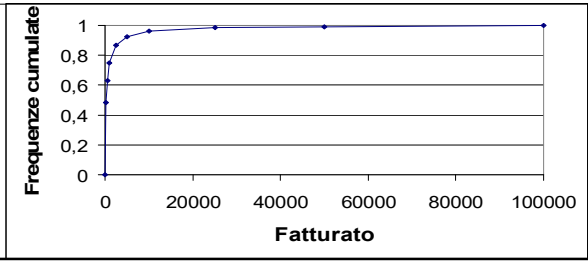
Si consideri la distribuzione di frequenze secondo il carattere Giudizio sulla solvibilità.

Si disegni la funzione di ripartizione e si cerchi di capire che tipo di informazione si può estrarre dal grafico.

Sulla base della funzione di ripartizione, determinare la percentuale di clienti con Giudizio di solvibilità almeno sufficiente.

82

Classi di fatturato ($c_{i-1} - c_i$)	Frequenze assolute (n_i)	Frequenze relative (f_i)	Frequenze relative cumulate (F_i)	Densità (h_i)
0--250	2156	0,482219	0,482219	0,001929
250--500	666	0,14896	0,631179	0,000596
500--1.000	517	0,115634	0,746813	0,000231
1.000--2.500	539	0,120555	0,867367	8,04E-05
2.500--5.000	260	0,058153	0,92552	2,33E-05
5.000--10.000	171	0,038246	0,963766	7,65E-06
10.000--25.000	95	0,021248	0,985015	1,42E-06
25.000--50.000	32	0,007157	0,992172	2,86E-07
50.000--100.000	35	0,007828	1	1,57E-07
Totale	**4471**	**1**		



83

Esercizio

Si considerino i dati del Caso Studio 1.

Si consideri la distribuzione di frequenze secondo il carattere Ammontare del Deposito. Si considerino le classi: Fino a 10.000, 10.000-20.000, 20.000-30.000 e 30.000-45.000.

Si disegnino l'istogramma di frequenze e la funzione di ripartizione e si cerchi di capire che tipo di informazione si può estrarre dai due grafici.

Sulla base della funzione di ripartizione della distribuzione in classi, determinare la percentuale di clienti con un ammontare dei depositi compreso tra 25.000 e 35.000 euro.

84

Esercizio

La seguente distribuzione è tratta dall'indagine ISTAT Struttura e produzione delle aziende agricole - Anno 2007 e considera le aziende agricole esistenti in Umbria secondo la superficie totale (in ettari):

Classi di superficie	Aziende
Meno di 1	5.427
1--2	7.528
2--5	8.900
5--10	6.880
10--20	4.198
20 e oltre	5.272
Totale	**38.205**

a)Assumendo "100" come estremo superiore per l'ultima classe, si rappresentino graficamente l'istogramma di frequenza e la funzione di ripartizione.

Si calcolino inoltre:

b)il valore della funzione di ripartizione nel punto 50;

c)la frequenza relativa delle aziende con una superficie compresa tra 8 e 12 ettari, utilizzando la funzione di ripartizione;

d)la frequenza relativa delle aziende con una superficie tra 15 e 50 ettari, utilizzando le densità di frequenza e riportando il risultato nell'istogramma.

Dove e come studiare

- Libro di testo: S. Borra, A. Di Ciaccio (2014), Cap. 1 e 2
- Svolgere esercitazione 1
- Svolgere i seguenti punti degli esercizi nel file: Esercizi su medie.xls:
  - Foglio 1, punto a) e d)
  - Foglio 2, punto a)
  - Foglio 3, punto a) e c)
  - Foglio 4, punto a) e b)
  - Foglio 5, punto a) e b)
  - Foglio 6, punto a) e b)
  - Foglio 7, punto b)