Università della Calabria

Facoltà di Economia

Statistica

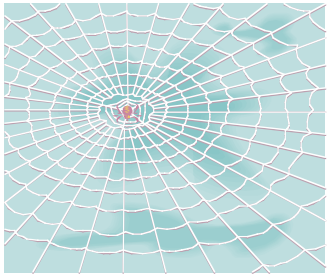
Settore Secs-S/01 - 10 crediti

Prof. Agostino Tarsitano

Dei problemi

Il problema è uno stato di cose di cui non siamo soddisfatti e per il quale siamo incerti sul modo, TRA QUELLI POSSIBILI, per portarlo ad una condizione migliore

Il problema si pone se e solo se c'è la volontà di risolverlo e se le azioni perseguibili sono più di una



La soluzione consiste nella scelta della linea di azione più utile ed efficace per il raggiungimento di un obiettivo.

Una soluzione accettabile si ha anche con la dimostrazione che è indifferente la scelta tra due o più linee di azione.

Modello del problema

In ogni problema ci sono degli aspetti (o fattori, o VARIABILI) che sono per noi:

- **CONTROLLABILI** — **E' noto il loro comportamento e si può predeterminarne i valori**
- **NON CONTROLLABILI** — **E' noto il loro comportamento, ma i loro valori sono -in larga misura- imprevedibili**
- **SCONOSCIUTI** — **Se ne deve postulare l'esistenza, ma non si conosce né il loro comportamento né i loro effetti**

Alcuni fattori sono RILEVANTI cioé hanno un ruolo anche minimo nel problema, altri sono IRRILEVANTI, cioé la loro assenza o presenza non altera la soluzione

Modello del problema/2

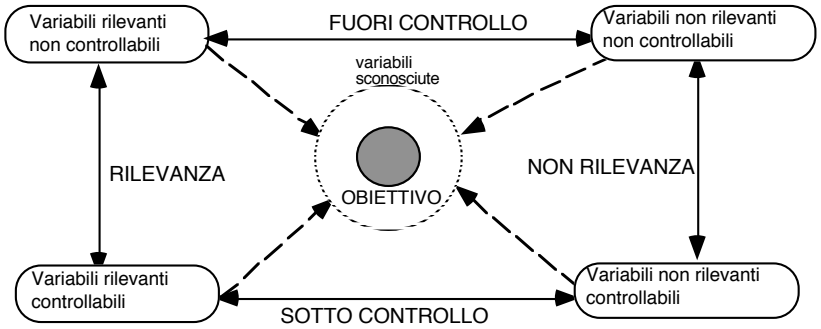
In termini formali possiamo porre la relazione

$$P = f(C; I; S) \quad dove \begin{cases} P = \text{Una misura della prossimità dell'obiettivo} \\ C = \text{Insieme dei fattori sotto controllo} \\ I = \text{Insieme dei fattori incotrollabili} \\ S = \text{Insieme dei fattori sconosciuti} \\ f = \text{schema dell'interazione tra i vari fattori} \end{cases}$$

Il successo o l'insuccesso di una linea di azione dipende da come interagiscono le variabili rilevanti e da come si pongono rispetto all'obiettivo.

Dipende anche dalla circostanza fortunata che i fattori "S" assecondino o non ostacolino quella linea di azione

Modello del problema/3



Per avere un'idea delle incertezze attribuibili ai fattori "S" immaginate questo schema in tre dimensioni e pensate ai nuovi collegamenti che si possono instaurare.

The Art of Problem Solving

Il ruolo della statistica è qui essenziale perché quasi sempre risolvere un problema consiste nell'effettuare una indagine statistica

La statistica:

Ci guida nella PRODUZIONE E RACCOLTA, CLASSIFICAZIONE E SINTESI dei dati.

Risolve le incertezze tra fattori rilevanti e irrilevanti

Fornisce modelli e tecniche per interpretare l'influenza dei fattori sconosciuti

Aiuta nella definizione degli indici di efficienza: COSTI/BENEFICI

Permette il riscontro di efficacia RISULTATI/OBIETTIVI

The Art of Problem Solving/2

L'uso della statistica non garantisce la soluzione del problema

> La combinazione dei dati di qualità ottima unita ad una volontà ferrea di ottenere una risposta non assicura che questa possa essere trovata nemmeno in forma approssimata

D'altra parte, la soluzione potrebbe essere:

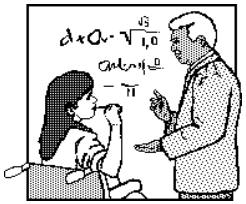
- **INSODDISFACENTE**
- **TARDIVA**
- **NON ACCETTABILE**
- **DISONESTA**
- **GENERATRICE DI ALTRI PROBLEMI PIU' COMPLESSI**

La statistica garantisce solo che i fatti non siano distorti ovvero che la distorsione avvenga in modi trasparenti e ricostruibili (almeno da chi conosce la statistica)

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

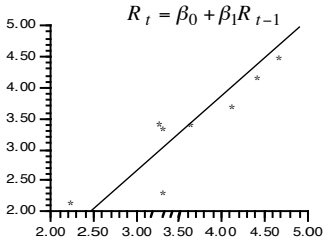
Come il talento forma gli argini rispetto all'avversa fortuna, la statistica riesce a frenare l'erraticità

Uso della statistica

Le situazioni studiate dalla statistica sono reali ovvero sono connesse a fatti concreti

Il rendimento netto di un fondo azionario al tempo "t" indicato con R_t è legato a quello del tempo "t-1".

t	R_{t-1}	R_t
1990	2.10	2.26
1991	2.26	3.34
1992	3.34	3.29
1993	3.29	3.35
1994	3.35	3.65
1995	3.65	4.13
1996	4.13	4.42
1997	4.42	4.69



I valori mostrano un trend raffigurato da una retta. Stimati i parametri si potrà prevedere quale sarà il rendimento del prossimo anno, noto quello dell'anno attuale. Nel caso in esempio, per il 1998, si passerà da 4.69 a 4.63.

USO DELLA STATISTICA/2

EFFICIENZA NELLA PUBBLICA AMMINISTRAZIONE

La legge sulla trasparenza dei procedimenti amministrivi impone tra l'altro che sia predeterminato il tempo entro cui deve concludersi. Occorre perciò stabilire i TEMPI MEDI e MASSIMI di completamento

Un esempio è il controllo del personale e La ripartizione efficiente in base alla produttività

Nell'esempio è aumentato il personale dell'agricoltura ed è diminuito quello dell'industria. Non sembra che equesta sia la tendenza giusta

Dipendenti P.A. e var.%

Ministero	1987	1991	Variaz.Perc
Presid. Consiglio	5287	6230	+17.8%
Affari esteri	6669	7311	+9.6%
Agricoltura	8355	10172	+21.7%
Ambiente	5	279	istituendo
Beni culturali	25707	24749	-3.7%
Bilancio	263	433	64.6%
Commercio estero	538	536	-0.4%
Difesa	294482	307839	4.5%
Finanze	122439	127432	+4.1%
Giustizia	66844	77283	+15.6%
Industria	1462	1434	-1.9%
Interno	121718	146671	+20.5%
Lavori pubblici	4499	4622	+2.7%
Lavoro	15445	15924	+3.1%
Marina Mercantile	1814	2062	+13.7%
Partecip. Statali	137	129	-5.8%
Università	93702	107326	+14.5%

USO DELLA STATISTICA/3

MARKETING

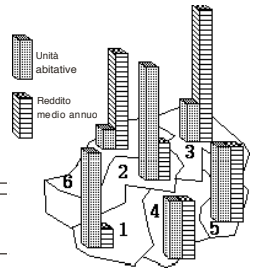
L'obiettivo del marketing è il raccordo della domanda di beni e servizi con la capacità di soddifarle da parte delle imprese.

Lo studio delle esigenze dei clienti, delle abitudini di consumo, della pubblicità, della la strategia rispetto alla concorrenza etc. impongono la conoscenza e la trattazione di informazioni quantitative

ESEMPIO:

La localizzazione di un nuovo punto vendita richiede tra l'altro la zoning di un area secondo il numero di unità abitative ed il reddito

In questa fase la statistica trova largo impiego



Zona	Unità abit.	Disponib.
1	700	9.1
2	1100	18.7
3	270	74.8
4	480	28.5
5	620	39.8
6	110	56.3

Dove non c'è statistica

Certamente la statistica ha poco a che vedere con gli "statisti" e con la "statica", ma ci sono anche altri casi

CASI LIMITE

Monterone (Co) è il comune più piccolo d'Italia (29 ab.) ed è una curiosità per gli studiosi di statistica.

I casi isolati o i casi singoli non interessano la statistica che infatti si presenta come scienza dei collettivi

ACCOSTAMENTI FORZATI

La gazzetta dello sport ha un angolo intitolato: *"Per gli amanti della statistica"* dove riporta dati relativi ai precedenti incontri tra due squadre.

Non si capisce bene quale sia il collegamento se non che i due club hanno lo stesso nome

Dove non c'è statistica/2

GENERALIZZAZIONI SEMPLICISTICHE

La statistica si ritrae quando ci si avventura in estrapolazioni non suffragate da riscontri fattuali

GIORNALISMO NON SPECIALIZZATO

Le statistiche dicono che il 23% dei lettori legge il giornale in meno di 15 minuti.

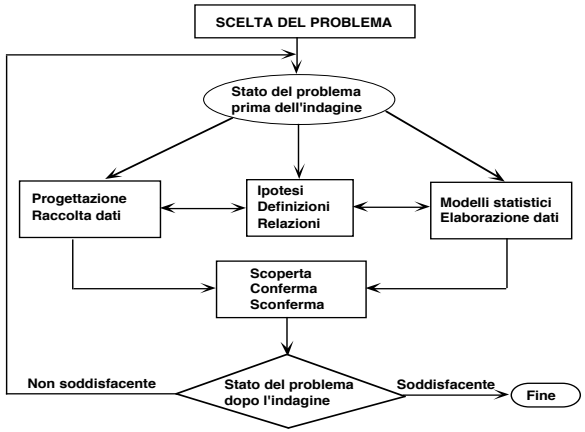
Sembra che l'autore si chiami fuori e che "le statistiche" non siano un supporto essenziale all'articolo.

COMPILAZIONE DI TABELLE

La tabellazione come raccolta organizzata di dati è utile alla statistica, ma non è la statistica.

Lo schema di lavoro

Si effettua un'indagine statistica per dare sostegnoa teorie incerte



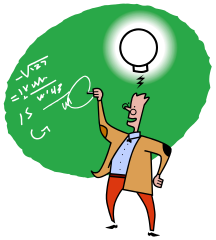
L'insieme delle conoscenze teoriche ed empiriche ed un SANO scetticismo aiutano a spiegare le variazioni tra due stati: prima e dopo l'indagine

L'indagine statistica

Se la trattazione del problema costringe a cercare nuovi dati, questi debbono essere rilevati con uno schema appropriato.

La rilevazione si articola in una sequenza ordinata di casi o repliche che hanno tanti elementi in comune da essere considerati facenti parte di un unico processo: l'indagine statistica.

Ogni indagine ha il suo piano di realizzazione legato alle peculiarità della disciplina in cui il problema è sorto.



Nuove rilevazioni

L'acquisizione di nuovi dati è dovuta al fatto che:

- **La base informativa di un problema non è soddisfacente**
- **E' utile e praticabile realizzarne una nuova o integrare quella esistente**

La rilevazione dei dati consiste nella annotazione sistematica, precisa e impersonale della modalità delle variabili riscontrate sull'unità

Le rilevazioni possono essere classificate in vario modo. Quella più rilevante è la distinzione tra TOTALI e PARZIALI:

TOTALI: coinvolgono tutti gli elementi di una popolazione

PARZIALI: la rilevazione è estesa solo ad una parte, comunque scelta, di popolazione

Le rilevazioni totali

Le RILEVAZIONI TOTALI (O CENSIMENTI) sono quelle in cui sono enumerate o misurate tutte ed indistintamente le unità della popolazione

All'interno delle totali si hanno:

- **RILEVAZIONI GENERALI: riguardano la rilevazione di tutte le unità rispetto alle variabili di interesse (POPOLAZIONE)**

  ***Esempio:** un'indagine sul voto che si rivolga a tutti gli elettori di qualsiasi sesso e regione di residenza*

- **RILEVAZIONI SPECIALI: riguardano la rilevazione delle sole unità rispondenti a certe specifiche (SOTTOPOPOLAZIONE)**

  ***Esempio:** un'indagine sul voto che si rivolga a tutti, ma i soli iscritti alle camere di commercio come "artigiani"*

Le rilevazioni totali/2

- Supponiamo di considerare come unità gli stabilimenti industriali attivi in Calabria per poi circoscrivere l'attenzione a quelli con più di 50 addetti (popolazione TARGET o teorica)

Quello che accomuna le due indagini è che sono enumerate TUTTE le unità che formano la popolazione o la sottopopolazione

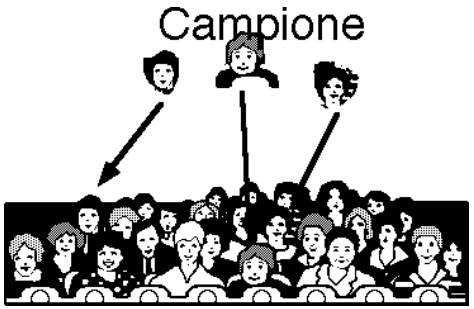
Si tratta perciò di Popolazioni (o sottopolazioni) FINITE E CENSIBILI cioè la cui rilevazione può effetticamente cominciare e finire in tempi e a costi praticabili

Le rilevazioni parziali

Sono limitate solo ad una parte delle unità della popolazione (o sottopopolazione) scelta in base ad opportuni criteri. La parte esaminata si chiama CAMPIONE.

La riduzione delle unità propria del metodo CAMPIONARIO è valida solo se permette il raggiungimento di risultati molto prossimi di quelli ottenibili con la TOTALE.

TOTALE/PARZIALE NON E' UNA COTRAPPOSIZIONE, MA UNA COMPLEMENTARITA'



Esperienze consolidate in molti paesi e in molte discipline dimostrano che si può dare pieno affidamento ai campioni purché scelti con accuratezza.

Tipologia di osservazione

La rilevazione è il rapporto che si instaura tra chi -consapevolemente- osserva ed i soggetti osservati.

Una prima utile distinzione tra i diversi tipi di rilevazione passa per il legame che può intercorrere tra agenti attivi e passivi della rilevazione.

- **ESTERNA** — **si annotano i fatti come i sensi li percepiscono o i sensori li avvertono, senza tentare intervenire sulle unità (tecniche non invasive).**
- **PARTECIPANTE** — **l'osservatore vive nella collettività che studia e interagisce con i soggetti che ne fanno parte modificandone in qualche modo le tendenze naturali (effetto Hawtorne).**

Rilevazione isolata e ripetibile

Un'altro utile distinguo è tra manifestazione

ISOLATA

si effettuano in relazione ad un fatto -volontario o involontario- che non può riaccadere oppure che è costoso o vietato provocare.

Questi di solito non interessano la Statistica

RIPETIBILE

Ogni fatto è unico ed è impossibile replicarlo, ma alcuni elementi essenziali rimangono intatti nelle varie manifestazioni:

tutte le volte che si configura un insieme di circostanze determinate si possono osservare certe conseguenze.

Survey ed esperimenti

All'interno delle rilevazioni ripetibili è importante la differenza tra:

SURVEY

si analizzano eventi che non si possono provocare a volontà, monitorati man mano che si verificano secondo cadenze predeterminate. Fenomeni metereologici, economici, demografici, finanziari

ESPERIMENTO

si creano situazioni di studio artificiali programmate in modo che i risultati possano rispondere a precise domande del tipo causa/effetto.

L'analisi osservativa

Ci interessano rilevazioni ESTERNE su fenomeni che si RIPETONO spontaneamente o che seguono flussi regolari.

L'analisi osservativa o INDAGINE STATISTICA è un insieme di entità elementari dette OSSERVAZIONI.

L'osservazione è composta da DATI: "ciascuno degli elementi di fatto (notizia, comunicazione, messaggio, rilevazione strumentale) rilevante per la soluzione di un problema"

Elementi costitutivi del Dato

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

- **L'UNITA'**
- **LA VARIABILE**
- **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, euro 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**

L'unità statistica

L'unità è il soggetto elementare cui l'indagine si rivolge: una persona fisica oggetto, azienda, o un gruppo di entità che, dal punto di vista dell'indagine, formino un tutt'uno.

Le unità devono essere obiettivamente distinguibili e deve pure essere stabilito quali siano quelle che interessa rilevare e quali debbano invece tralasciarsi.

ESEMPI

a) Interessi maturati su di un conto corrente	**(Il conto corrente)**
b) Tipo di riscaldamento di un appartamento	**(L'appartamento)**
c) Numero di testi consigliati in un corso	**(Il corso)**
d) Emissione di gas tossici da un automobile	**(L'automobile)**
f) Numero di arresti per agente di polizia	**(L'agente)**

Problemi di definizione

INDAGINE SULLE FAMIGLIE

Come si considerano i "single", le coabitazioni, le comunità?

PUBBLICITA' TURISTICA

Non è raro leggere o sentire messaggi promozionali del tipo: 30 giorni di sole nel mese X. Il problema è capire cosa si intende per "giornata di sole": ad esempio nelle ore diurne una sequenza di almeno otto ore di sereno e senza nebbia.

SONDAGGI PRE-ELETTORALI

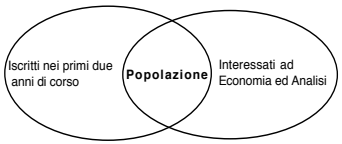
Un intervista telefonica agli abbonati di "La Gazzetta del Sud" può solo indicare come la pensano gli abbonati che hanno risposto alle telefonata.

La popolazione

Dicesi popolazione o UNIVERSO l'insieme di tutte e solo le unità che si è interessati ad osservare in una certa indagine.

ESEMPIO:

Alcuni studenti intendono finanziare le spese di frequenza universitaria avviando un programma di ripetizioni ben fatte ed a basso costo. Quale sarà la popolazione?



E' chiaro che non possono essere tutti gli studenti iscritti. Ci si può limitare agli studenti dei primi due anni.

Occorre poi determinare le materie per cui esistono le competenze: diciamo i corsi fondamentali di economia e matematica.

La delimitazione dell'universo è ora chiara: studenti del biennio che non hanno sostenuto economia e/o analisi.

La popolazione/2

NON è un gruppo di persone che risiedono in una certa zona.

il termine POPOLAZIONE ha una accezione più ampia e più astratta : tutte e solo le unità che hanno in comune una o più proprietà rilevanti per il problema.

La caratteristica unificante deve essere evidente cosicché il riconoscimento avvenga con il minimo di incertezza tenuto conto delle difficoltà create da unità congiunte o sfocate.

Se disponiamo di un elenco del fatturato di 500 imprese edili ciò che studiamo non è la popolazione delle imprese, ma la popolazione dei fatturati.

Tipologia delle popolazioni

La popolazione è un insieme e come tale può essere:

FINITA — **Se include oggetti che possono essere contati ed il conteggio, ad un certo punto si interrompe.**

Esempi: le pagine di un libro, i diplomati di una scuola

ENUMERABILE — **Le unità sono contabili, ma il conteggio non si interrompe mai**

Esempi: i numeri naturali, i lanci di un dado

INFINITA — **Ogni sottoinsieme di popolazione contiene lo stesso numero di entità contenute nella popolazione.**

Esempi: le frazioni tra zero ed uno, le nuances di un colore,

Popolazioni indeterminate ed elusive

Non sempre è nota o determinabile la numerosità della popolazione

INDETERMINATE

L'insieme dei soggeti è finito in quanto esiste un limite fisico non valicabile alla sua crescita, ma le unità sono sparse o rare al punto da rendere impossibile il loro materiale censimento.

Esempi: animali selvatici, tifosi di una squadra, gruppi etnici o religiosi particolari

ELUSIVE

Composte da unità che hanno buone ragioni per non farsi censire. Per queste non si potrà mai essere sicuri che le unità individuate siano tutte o solo una parte perché le altre rimangono nascoste

Esempi: extracomunitari senza documenti, tossicodipendenti, affetti da malattie infettive, affiliati alla onorata società, vincitori di grossi premi alle lotterie, gli idraulici nel mese di agosto.

Popolazioni censibili e virtuali

E' censibile la popolazione le cui unità possono essere esaminate in tempi e costi "ragionevoli"

E' virtuale la popolazione infinita e la enumerabile. Anche la finita lo può essere se la disamina delle unità è costosa, difficile, superflua, impossibile.

Esse esistono solo in via teorica e debbono essere censite comunque in poche unità (campione).

1) Le popolazioni preistoriche possono essere analizzate solo attraverso i pochi resti che gli scavi portano alla luce.

2) Il controllo di qualità non riguarda solo quello che si è già prodotto, ma anche quello che si produrrà, che però non è ancora censibile.

3 I risultati di ogni esperimento sono in realtà solo una parte delle infinite repliche che si potrebbero effettuare.

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

La variabile

E' l'aspetto si intende studiare nel dato.

Può essere una distanza, una numerosità, una forma, un atteggiamento, un grado od anche una composizione di caratteristiche da trattare in modo aggregato.

I simboli più diffusi sono:

$$X,\quad Y,\quad W,\quad Z$$

Che sono la codifica della variabile

La codifica è l'espressione abbreviata con cui le informazioni sulle variabili acquisite dalle unità sono trasferite sui supporti di elaborazione o nei ragionamenti astratti

La variabile/2

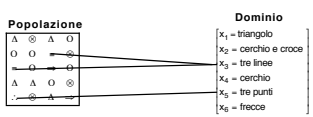
Perché una generica qualità o quantità sia definita "variabile" occorre...

- **ATTINENZA con la realtà di interesse la cui comprensione aumenta (anche di poco) per la disponibilità di dati sulla variabile.**
- **ESSERE SOGGETTA A VARIAZIONI: cioè possa presentarsi con almeno due valori o categorie distinte nell'ambito della popolazione.**
- **ESSERE ACCERTABILE e cioè essere rilevabile strumentalmente senza ambiguità**

Si presuppone inoltre che la variabile possa essere osservata/misurata in modo separato da altre variabili che pure incidono sull'unità.

il dominio della variabile

Individuata la variabile occorre definire l'insieme di tutti e solo i valori o modalità della variabile X (il dominio) riscontrabili nella popolazione:



Ad ogni unità della popolazione sarà associata una ed una sola modalità del dominio.

In questo caso, una delle sei diverse forme presenti. Unità diverse possono presentare la stessa modalità

il dominio della variabile è un insieme di "k" elementi con "k" finito od infinito

$$S=\{X_1, X_2, \ldots, X_k\}$$

L'abbinamento unità/modalità si effettua confrontando ciascuna delle unità è con il dominio "S" ed associandola ad una delle Xi in base ad una regola di classificazione o misurazione.

il dominio della variabile/2

Perché non insorgano ambiguità è necessario che le modalità siano

UNIVOCHE: sia possibile osservarne una sola per ogni unità e sia subito chiaro quale

ESAUSTIVE: non sia possibile osservarne di diverse da quelle già in S

RIPRODUCIBILI: la rilevazione dovrà dar luogo sembre allo stesso schema di attribuzione.

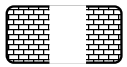
a) Incompatibilità: $X_i \neq X_j \;\; per\ ogni\ i \neq j\ e X_i, X_j \in S$

b) Esaustività: $per\ ogni\ u \in P\ \ X(u) \in S;$

c) Riproducibilità: $X = X_i \;\; se \;\; e \;\; solo \;\; se \;\; X(u) = X_i$

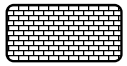
Dominio chiuso o aperto

L'insieme dei valori ammissibili "S" può essere



APERTO **Quando il fenomeno descritto non ha un limite minimo e/o massimo ben definito prima che sia completata la rilevazione**

Esempio: Reddito (che può anche essere negativo)



CHIUSO **Quando le sue modalità sono definite e note in anticipo e non possono cambiare durante la rilevazione**

Esempio: Stato civile

il dominio aperto comporta problemi di elaborazione. Quello chiuso consente dei controlli di validità dei dati

Analisi univariata e multivariata

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

ESEMPIO: dove vanno gli studenti

Studio/Residenza		Nord		Centro		Sud		Totale	
		numero	%	numero	%	numero	%	numero	%
Stessa regione		286995	83.6	178692	90.7	253887	74.7	719.724	81.8
Altre Regioni	Nord ovest	18783	5.5	1635	0.8	8378	2.5	28887	3.3
	Nord Est	27308	8.0	4749	2.4	11312	3.3	43369	4.9
	Centro	9149	2.7	9396	4.8	38800	11.4	57345	6.5
	Sud	929	0.3	2795	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	879906	100.0

La lettura della tabella non è difficile. Lo è la generalizzazione dei risultati

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

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

La definizione operativa

E' l'insieme di regole con cui classificare un concetto, determinarne la misura o, in generale, per aggangiarlo alla realtà osservabile.

cioè impone la definizione operativa solo con variabili di cui sia possibile seguire con facilità il meccanismo di conversione di una proprietà delle unità in una categoria o valore del dominio.

ESEMPIO

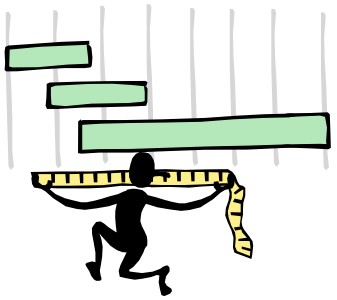
il concetto di interconnessione tra due centri abitati, diciamo "A" e "B" , è misurato con la semisomma degli automezzi che si sono spostati da "A" a "B" e quelli che da "B" sono andatti ad "A".

Classificazione e misurazione

L'acquisizione dei dati può avvenire classificando in categorie distinte la proprietà di cui l'unità è portatrice oppure misurandola in base ad una determinata unità di misura.

con la CLASSIFICAZIONE si identificano l'unità (e le modalità numeriche del dominio sono equivalenti ad ogni altro insieme di simboli);

con la MISURAZIONE si quantifica una proprietà posseduta ed i numeri sono utilizzati in quanto inseriti in un sistema di numerazione.



Classificazione e misurazione/2

La classificazione e la misurazione possono scaturire da due procedure di assegnazione dei valori: enumerazione delle unità rispetto alla proprietà posseduta

oppure comparazione della proprietà studiata rispetto ad un ventaglio di possibilità che, identico per tutte le unità, non dipende dal numero.

ASSEGNAZIONE VALORI

RILEVAZIONE DEL DATO	Enumerazione	Comparazione
Classifizione	Nominazioni	Scala nominale
Misurazione	Graduatorie	Scale ordinali semplici Scale ordinali graduate Scale intervallari Scale proporzionali

Nominazioni e Variabili nominali

Le modalità di queste variabili esprimono categorie, qualità, status: le $\{X_i\}$ in "S" hanno la sola funzione di etichettare le unità per formarne un elenco o per raggrupparle in classi omogenee:.

ESEMPI:

Nominazione: La variabile "Regione" si manifesta con le usuali 20 modalità S={Calabria, Sicilia, ..., Val d'Aosta, Piemonte}.

variabile nominale: Un'impresa può ricadere nel settore {agricoltura, industria, altre attività}.

Le differenze possono essere accertate, ma non ordinate né misurate: si possono scambiar di posto senza che ciò influisca sulla validità della classificazione

Uso dei numeri

la codifica delle modalità porta ad usare dei numeri. Questo però non significa che siano lecite delle operazioni aritmetiche:

i ruoli di una squadra di calcio sono indicati con dei numeri, ma non si può dire che l'ala sinistra ("11") sia maggiore dello stopper ("5") o che l'unità di misura "1" dei calciatori sia il portiere;

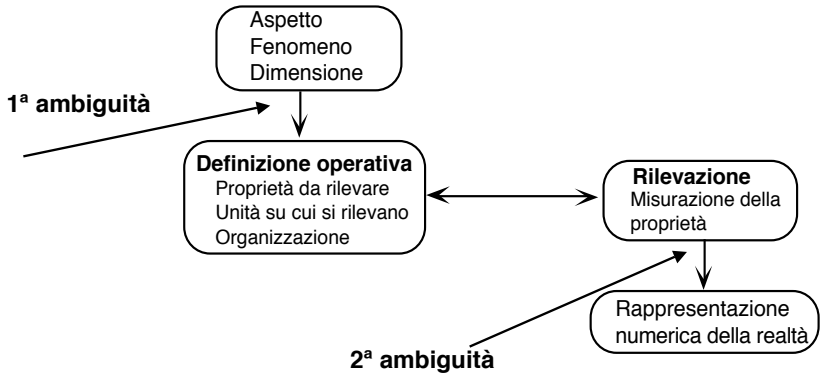
ESEMPI

il numero civico delle abitazioni:

Non ha significato la eventuale progressione delle modalità;

Tecniche di misurazione

il concetto di misurazione è uno dei più controversi tanto che oggi, dopo più di 50 anni, il dibattito è sempre aperto.



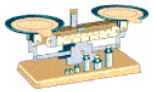
La scala di misurazione

Qui esiste una sovrapponibilità tra una categoria e la successiva che, oltre a contenere quella che la precede, vi aggiunge un quantum di proprietà che la differenzia dalla prima senza cancellarla, anzi inglobandola

ciò che distingue le scale di misurazione è il diverso grado di formalizzazione che si può dare al meccanismo dell'aggiunta del quanto di proprietà.

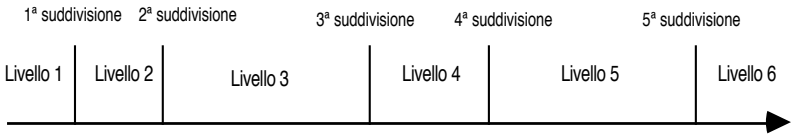
1. **Ordinamento tra valori senza distinguibilità degli scarti;**
2. **Ordinamento tra valori con ordinamento degli scarti;**
3. **Quantificazione dei valori con parità tra scarti: 7-5=3-1;**
4. **Quantificazione dei valori con parità tra rapporti: 8:4=6:3.**

Scala=bilanciamento



Continuo percettivo

L'intensità con cui si avverte una sensazione varia in un continuum di stati. Per misurare il concetto si sovrappone una griglia più o meno regolare



una unità che sia X_i in una rilevazione ed X_j in una successiva con $X_i<X_j$ sarà passata per tutti gli stati intermedi tra X_i ed X_j.

Le suddivisioni non sono però oggettive: osservatori diversi scelgono divisioni diverse ovvero lo stesso punto di separazione ha senso diverso.

N.B. Talvota la proprietà studiata ha natura discontinua: si modifica con una scansione non frazionabile per un numero finito di stati che sono i soli a poter essere osservati.

Ordinamenti

Il termine "scala" ha senso se tra le modalità di "S" sono possibili degli ordinamenti.

1) $X_i < X_j \quad oppure \quad X_i > X_j \quad per \quad ogni \quad i \neq j$

2) $X_i < X_j \Rightarrow X_i \neq X_j$

3) $X_i < X_j \quad e \quad X_j < X_k \Rightarrow X_i < X_k \quad per \quad ogni \quad i < j < k$

Maggiore è il contenuto di "fenomeno" maggiore è la modalità che la rappresenta; esiste perciò una disposizione delle modalità che non può essere alterata senza che ne risulti modificata la rilevazione.

Il dominio si può esprimere con interi consecutivi:

$$S = \{a, a+1,\ a+2,\ \ldots, a+k-1\}$$

Graduatorie

Le modalità "S" sono i ranghi corrispondenti alle posizioni in graduatorie delle unità per i valori possibili sono dati dalla numerosità della rilevazione.

$$S = \{1^{\mathrm{a}},\ 2^{\mathrm{a}},\ \ldots, k^{\mathrm{a}}\}$$

Il processo di misurazione è superficiale, con possibilità elaborative limitate, basate su confronto e sintesi delle posizioni occupate rispetto a variabili diverse.

ESEMPIO

Per stare in testa occorrono buone posizioni su entrambe le graduatorie

Stud.	Grad. Scritto	Grad. Orale	Totale
A	3	1	4
B	2	3	5
C	1	7	8
D	7	2	9
E	5	4	9
F	4	6	10
G	6	5	11

Variabili ordinali

i ranghi sono dei voti che esprimono una valutazione della proprietà posseduta.

Ogni unità è confrontata con una linea (bencjmark) che incasella l'unità in una categoria di valore a prescindere da quello che succede alle altre unità.

ESEMPI:

a) Voti di un giudice: S={0, 1, 2, …, 10};

b) Ammontare di punti da ripartire: {0 -100};

c) Voti grafici: {++, +0, 0+, 00, -0, 0-, - -};

d) Quantificatore verbale: { pianura, collina, montagna}

Invarianza rispetto a trasformazioni monotòne

$$f(X_i) < f(X_j) \quad se \quad X_i < X_j$$

Numero di modalità e Posizione

Non esiste un numero ottimale di livelli: k=7±2 o k=6 sono considerati uno standard nelle ricerche di mercato (Kinnear eTaylor, 1979, p. 30, Malhotra 1996, p. 298).

3 o 4 gradini comportano risultati confusi per l'accorpamento di giudizi eterogenei; d'altra più di sei è utile solo per acquisire variazioni di quantità molto piccole di cui non sempre si ha bisogno.

Anche la disposizione deve essere equilibrata:

ESEMPIO:

Quale delle tre seguenti moltiplicazioni

$P1.\ 9*7*8*6*5*4*3$

$P2.\ 3*4*5*6*7*8*9$

$P3.\ 7*3*8*4*6*5*9$

Effetto posizione

darà il risultato più alto?

Differenziale semantico

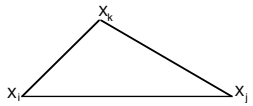
Per attenuare le ambiguità derivanti delle scale ordinali si possono usare delle scale bipolari in cui sono inserite solo le valutazioni più opposte dell'aspetto indagato collocando tra di esse, ad opportune interdistanze, una serie di riquadri

Chi risponde dovrà poi indiviudare il punto più prossimo al suo giudizio ovvero indicare quale descrizione numerica o verbale si adatti al proprio sentire.

ESEMPIO:
"Come giudicate l'operato dei rappresentanti degli studenti nel Senato accademico integrato?"

Scale metriche

Date tre qualsiasi modalità di "S" allora



$d(X_i, X_j) = 0 \ \ se\ e\ solo\ se\ X_i = X_j;$ *Identità*

$d(X_i, X_j) > 0 \ se\ X_i \neq X_j;$ *Positività*

$d(X_i, X_j) = d(X_j, X_i);$ *Simmetria*

$d(X_i, X_k) + d(X_k, X_j) \geq d(X_i, X_j);$ *Disuguaglianza triangolare*

Se il dominio della "X" verifica le quattro condizioni allora su di esso si applicano, sia pure con qualche distinguo, tutte le procedure statistiche.

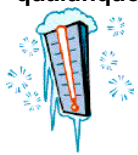
Scale intervallari

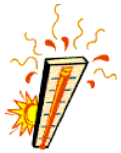
Sivaluta ciò che succede al fenomeno ponendolo in relazione con un movimento lungo un'asta graduata.

Le tacche sono regolari e separate -al livello minimo- da una unità convenzionale che può essere variata senza interferire con ciò che si misura.

L'origine o punto-zero della scala ha un ruolo marginale dato che agisce solo come riferimento e può essere sostituita, senza alcuna conseguenza sull'esito della rilevazione.

Un incremento assoluto tra due misurazioni ha lo stesso significato qualunque sia il livello da cui si calcola l'incremento.





Scale intervallari/2

La differenza tra 40C e 30C gradi è la stessa di quella tra 30C e 20C, ma non si può dire che ad 40C faccia due volte più caldo che a 20C. Se le temperature sono convertite in gradi Fahrenheit si avrà:

$$30\,C \rightarrow \frac{9*30}{5}+32=86F;\quad 40C \rightarrow \frac{9*40}{5}+32=104\,F;\quad 20\,C \rightarrow \frac{9*20}{5}+32=68F;$$

La differenza tra due temperature ha lo stesso significato qualunque sia il livello, ma nessuna asserzione può farsi sul loro rapporto dato che C=0 o F=0 non significa "totale assenza di calore".

Se la "X" è misurata su scala intervallare è lecito -se preferibile- usare in sua vece la variabile ottenuta come trasformazione lineare.

$$Y = a + bx;\quad b > 0$$

Scale proporzionali

Ad un incremento relativo nella misura, corrisponde un incremento relativo di eguale entità in ciò che si misura.

ESEMPIO

La misura di due centimetridi un segmento è -senza incertezze- il doppio di uno con lunghezza pari ad un centimetro;

se la misura aumenta del 50% anche il segmento si allunga di una estensione pari alla sua metà.

E' ammessa ogni trasformazione del tipo:

$$Se\ \ X_j \neq 0\ \ \frac{X_i}{X_j} = c\left[\frac{f(X_i)}{f(X_j)}\right];\quad per\ f(X_j) \neq 0$$

Gerarchia tra i livelli di misurazione

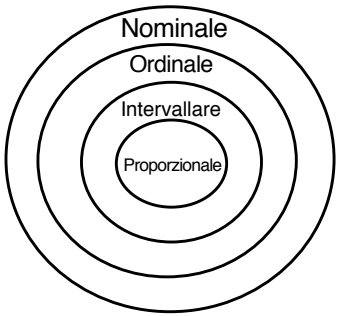
Se una variabile è su scala proporzionale, con un processo di arrotondamenti è possibile riportarla su scala intervallare;

questa a sua volta instaura un ordinamento

che è anche utilizzabile per valutare la similarità delle categorie a quella di riferimento (scala nominale).

Fra le scale esiste perciò una gerarchia:

Il dominio della variabile può anche essere espresso con i livelli esterni, ma non con quelli interni



Variabili discrete

Derivano da un processo di conteggio o di numerazione:

riviste per numero di abbonati
partiti politici per numero di iscritti
catene commerciali per numero di punti vendita affiliati,

Le modalità sono presentate usualmente, ma non sempre, in ordine crescente:

$$X_1 < X_2 < \ldots < X_k$$

Il simbolo "<" ha qui il significato aritmetico di "minore".

La differenza tra due modalità ha significato costante, ma nulla si può dire sul rapporto tra di esse.

In modo alternativo si può dire che le modalità della variabile discreta possono essere contate ovvero poste in corrispondenza biunivoca con l'insieme dei numeri naturali.

Discrete frazionarie e dense

Una variabile può essere discreta, ma espressa con dei numeri decimali

ESEMPIO: lancio di due dadi. Modalità= semisomma dei punti sulle facce superiori. L'uscita di un "6" e di un "5" o di un doppio "6" danno luogo alle modalità

$$\frac{6+5}{2} = 5.5; \quad \frac{6+6}{2} = 6$$

La variabile è discreta perchè tra "5.5" e "6" la variabile non può assumere alcun valore. Le sue modalità sono tutte ISOLATE: E' sempre possibile trovare un intervallo, per quanto piccolo, che contiene una sola modalità.

La variabile DENSA è discrete per natura, ma ha una unità di misura è molto piccola rispetto all'ordine di grandezza con cui si manifesta

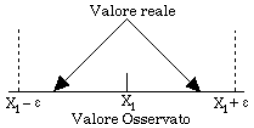
reddito in lire;
circolazione di vetture per numero di auto;
nazioni per numero di abitanti;

La trattazione dei caratteri densi è simile a quella dei caratteri continui

Variabili continue

Non possono essere rilevate puntualmente; il valore assunto è il centro dell'intervallo

$$\left[X_1 - \varepsilon, X_1 + \varepsilon\right]$$



Dire che $X = X_1$ significa dire che $|X - X_1| < \varepsilon$ cioé che si è osservato uno qualsiasi degli infiniti valori compresi in

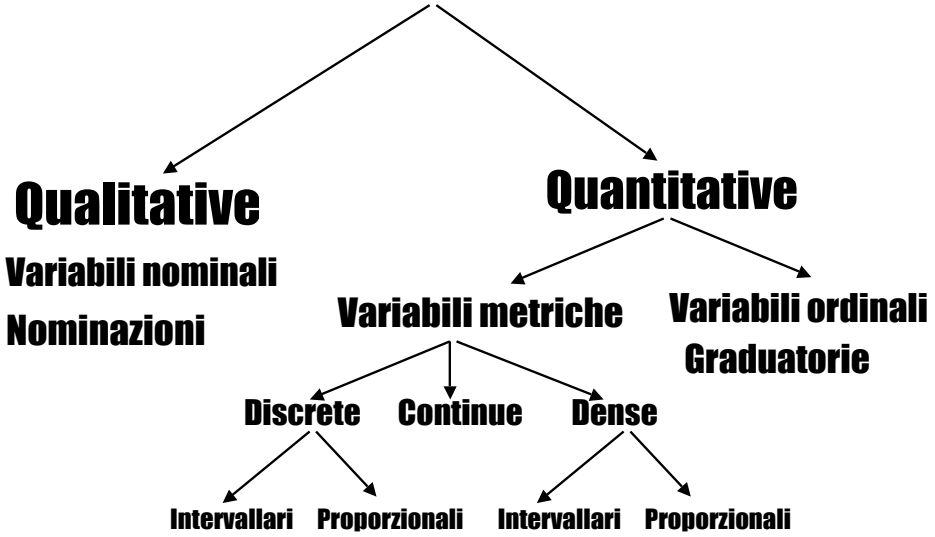
$$\left[X_1 - \varepsilon, X_1 + \varepsilon\right]$$

L'ampiezza del sottointervallo dipende dalla precisione degli strumenti di rilevazione. (Questo è un limite degli strumenti di misurazione non della variabile misurata)

Talvolta le modalità sono presentate come interi. Per distinguerle da quelle di una variabile discreta basta ricordare che:

Tra due modalità discrete non ve ne può essere un'altra, tra due modalità continue se ne trovano infinite

Tipologia delle variabili



Criterio organizzativo

Ogni unità si inserisce in un contesto in cui si distingue e che consente di attribuirle la modalità corretta.

Le tecniche statistiche sono anonime e trascurano la localizzazione dell'unità rispetto alle altre.

In alcune analisi è necessario che l'unità sia ben collocata -nel tempo o nello spazio- ed il suo esame prima o dopo di un altra è rilevante.

Gli ordinamenti possibili sono diversi, ma noi consideriamo solo

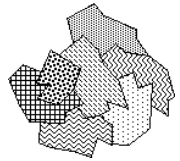
SERIE SPAZIALI (ordinamento geografico)

SERIE STORICHE (ordinamento temporale)

Le serie territoriali

Si ritiene che la modalità o intensità raggiunta dipenda dalla sua posizione topografica. Qui conta il tipo di unità considerata

UNITA' AREALI: rappresentata da una poligolane chiusa
entità fisiche: isola, lago, continente, etc.
entità amministrative: comuni, regioni, nazioni
entità funzionali: distretti sanitari, telefonici, scolastici



Le unità si considerano omogenee al loro interno anche se la rilevazione del carattere si effettua in più punti

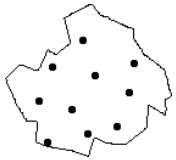
Esempio:
Percentuale di dipendenti pubblici sul totale occupati

Paese	%
Belgio	20.2
Danimarca	29.8
Germania	15.4
Grecia	10.4
Francia	22.8
Spagna	14.3
Regno Unito	19.5
Irlanda	17.9
Italia	17.4
Lussemburgo	11.6
Paesi bassi	15.1
Portogallo	14.1

Altri tipi di unità (territoriali)

UNITA' PUNTUALI: costituiscono i nodi di una maglia più o meno fitta di punti che coprono un dato territorio

misurazioni atmosferiche e idrogeologiche,
censimenti della popolazione
rilevazione della forza lavoro



Le unità puntuali hanno il grande pregio di visualizzare l'ubicazione delle modalità o intensità rivelandone la disseminazione o la concentrazione nel territorio

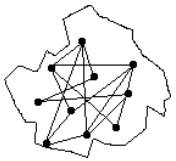
Esempio: Consumi di acqua per uso domestico

Città	Consumo
Bruxelles	108
Amburgo	146
Copenaghen	194
Londra	132
Parigi	147
Roma	220
Lussemburgo	171
Amsterdam	159
Madrid	158

Altri tipi di unità/2

UNITA' RETICOLARI (NETWORK): sono unità che si diramano nel territorio

fiumi, strade, gallerie
direttrici di sviluppo
rotte di navigazione
reti di distribuzione



La rilevazione dei caratteri sui network avviene per punti, in analogia alla osservazione di un flusso che percorre il reticolo (il flusso è spesso la variabile che si intende rilevare)

Esempio:
itinerari turistici calabresi per numero di pro-loco coinvolte

Itinerari	Pro-loco
Catanzaro-S.S.Bruno	5
Catanzaro-Capo Vaticano	7
Catanzaro-Laghi	3
Costa tirrenica	14
Sila e laghi	5
Costa Jonica	14
Aspromonte	5
Costa Viola	13
Magna Grecia	17

Le serie storiche

Spesso si studiano variabili nel tempo e la progressione cronologica dei valori è essenziale per comprendere il comportamento della variabile

VARIABILI DI FLUSSO: procedono in modo continuo. Se il flusso è regolare non importa l'unità di tempo. Se il flusso è erratico mette conto sapere che si rileva per settimane, mesi , trimestri, anni, etc.

Esempio:
Spesa dell'ammnistrazione statale per la cultura
te uto conto che ogni anno non arretra rispetto al periodo precedente

Anno	Spesa
1984	1 055
1985	1 206
1986	1 322
1987	1 954
1988	1 393
1989	1 064
1990	988
1991	947

VARIABILI DI STOCK: si manifestano in un dato istante per poi ripetersi più o meno regolarmente. Ricominciano da zero

Esempio:
Voti validi nelle consultazioni politiche

Anni	Totale
1948	26268912
1953	27092743
1958	29563633
1963	30758031
1968	31803253
1972	33414779
1976	36727273
1979	36671308
1983	36906005
1987	38592383

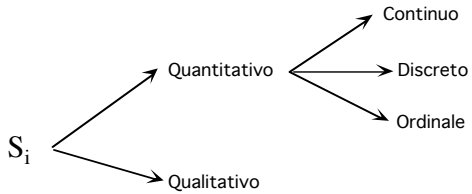
Lo spazio dei dati

Su ogni unità si rilevano m variabili $X_i = (X_{i1}, X_{i2}, \ldots, X_{im}),\ i = 1,2,\ldots,N$

X_i è un blocco o vettore di informazioni detto osservazione i-esima

Ogni variabile ha un suo dominio

$S_1,\ S_2,\ldots,S_m$



La popolazione o collettivo è composta da N unità (N può essere infinito)

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

L'insieme delle osservazioni potenziali su m variabili per le N unità costituisce lo SPAZIO DEI DATI

Modello relazionale dei dati

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 effettivamente riscontrata

Formalmente, d è un elemento del prodotto cartesiano di insiemi

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

cioè dello spazio dei dati

La matrice dei dati

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

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

ESEMPIO

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

SPAZIO DEI DATI

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

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 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

Esempio di data set su pacchetto applicativo. STATISTICA

Caratteristiche di alcune automobili: m=5 variabili per n=22 unità.

CARS

Casename	PRICE	ACCELER	BRAKING	HANDLING	MILAGE
Acura	-0.521	0.477	-0.007	0.382	2.079
Audi	0.866	0.208	0.319	-0.091	-0.677
BMW	0.496	-0.802	0.192	-0.091	-0.154
Buick	-0.614	1.689	0.933	-0.210	-0.154
Corvette	1.235	-1.811	-0.494	0.973	-0.677
Chrysler	-0.614	0.073	0.427	-0.210	-0.154
Dodge	-0.706	-0.196	0.481	0.145	-0.154
Eagle	-0.614	1.218	-4.199	-0.210	-0.677
Ford	-0.706	-1.542	0.987	0.145	-1.724
Honda	-0.429	0.410	-0.007	0.027	0.369
Isuzu	-0.798	0.410	-0.061	-4.230	1.067
Mazda	0.126	0.679	-0.133	0.500	-1.724
Mercedes	1.051	0.006	0.120	-0.091	-0.154
Mitsub.	-0.614	-1.003	0.084	0.382	0.718
Nissan	-0.429	0.073	-0.007	0.263	0.997
Olds	-0.614	-0.734	0.409	0.382	2.114
Pontiac	-0.614	0.679	0.536	0.145	0.195
Porsche	3.454	-2.215	-0.296	0.618	-1.026
Saab	0.588	0.679	0.246	0.263	0.021
Toyota	-0.059	1.218	0.228	0.736	-0.851
VW	-0.706	-0.128	0.102	0.382	0.195
Volvo	0.219	0.612	0.138	-0.210	0.369

Esempio di data in ambiente R

```
> E<-matrix(scan(file="esempio1.txt"),ncol=3,byrow=TRUE)
> E
     [,1] [,2] [,3]
[1,]  197    8  1.8
[2,] 1355   58  1.7
[3,] 2075   81  1.8
```

La statistica che si insegna
è diversa da quella che si usa?

```
> read.table("nazioni.txt")
          PIL Pop Inflazione Area     EU
Austria    NA   8         1.8   84     EC
Francia  1355  58         1.7  544     EC
Germania 2075  81         1.8  358     EC
Svizzera  265   7         1.8   41 non-EU
>
```

I dati mancanti

Derivano da mancata rilevazione o rilevazione manifestamente sbagliata.

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

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

ESEMPIO

Numero di permessi sindacali concessi da ammininistrazioni pubbliche.

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

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

Rilevazione campionaria univariata

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

La codifica

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

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

ESEMPIO:

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

{national, urban, nonagricultural, rural, agricultural}

I meta dati

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

Nessuna unità (etichetta identificativa inclusa) può essere ripetuta

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

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

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

Chiave	X1	X2
A1	10	27
Z3	−1	0.3

Analisi univariata e multivariata

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

ESEMPIO:
dove vanno gli studenti

Stessa regione

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

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

Gli studi multidimensionali sono rinviati. Faremo solo studi univariati.

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

Statistica descrittiva ed inferenziale

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

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

cioè "m" osservazioni su "n" unità.

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

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

Valore massimo per Radiouno; minimo per Isoradio; c'è un gruppo che si addensa intorno a 1000-1500 ascolti.

Le reti pubbliche sono più diffuse di quelle commerciali

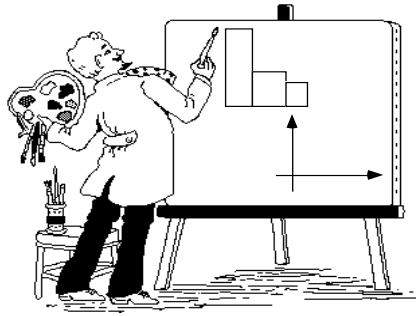
Statistica descrittiva

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

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

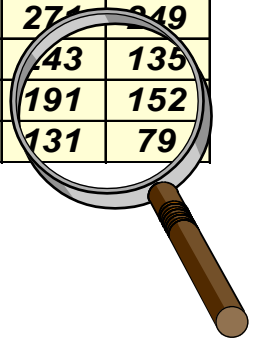
In breve, si configura come una trattazione preliminare indispensabile per affrontare uno studio complesso.

Utilizza tecniche elementari, soprattutto grafiche.



Statistica descrittiva/2

312	281	137	88
145	293	113	239
119	312	175	300
215	196	157	87
200	221	162	367
218	110	327	257
211	126	226	347
211	179	112	79
94	292	271	249
248	-139	243	135
300	248	191	152
253	221	131	79



Calcola indici di

1. Centrallità
2. Variabilità
3. Asimmetria

L'interpretazione dei valori di questi indici è molto legata all'ampiezza del data set

Statistica inferenziale

Inizia laddove il data set è visto come la punta di un iceberg.

I dati sono solo una delle possibili realizzazioni

e riguardano

anche gli ascolti che potevano esserci, ma non ci sono stati nonché gli ascolti che ci saranno o potranno esserci in futuro.

In che modo ed in che misura possiamo estendere agli ascolti potenziali le quantità calcolate sui valori osservati?

Questa è STATISTICA INFERENZIALE

Esempio

Durata in giorni di un premier in Italia.

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

Ci interessa determinare la durata massima del Premier.

Se tali unità fossero le uniche disponibili allora sono tutto ciò che serve per calcolare il massimo. De Gasperi con X_{max}=2808.

Lo stesso vale per la durata media: se tutti i premier fossero durati lo stesso numero di giorni quanto sarebbe durato ciascuno? 814.

Questa è STATISTICA DESCRITTIVA

Scelto a caso un premier qual è la sua durata più probabile? Quale sarà la durata che deve aspettarsi un nuovo premier?

Questa è STATISTICA INFERENZIALE

Descrizione del data set

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

Al momento tralasciamo...

il modo in cui il data set è stato formato

i criteri con cui sono state scelte le variabili

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

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

Presentazione dei dati

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

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

SPOGLIO: Ordinamento dei dati + trattamento dei doppioni

- **Tabella statistica** — **una tabella a due colonne dove si riportano tutte e solo le modalità verificatesi con a fianco il numero di volte che si sono presentate**

ramo

- **Diagramma gambo-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

Lo spoglio dei dati avviene con il computer:

1) **CODIFICA: definizione di una corrispondenza biunivoca tra le cifre numeriche e/o le denominazioni con un insieme di codici che ne facilita l'inputazione e riduce lo spazio che occupano.**

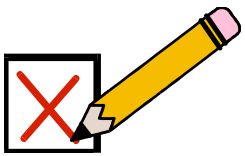
2) **INPUTAZIONE: monotona, ma delicata fase di trasferimento dei dati dal supporto su cui sono già registrati ai programmi di elaborazione.**

3) **ORDINAMENTO: i dati vengono disposti in ordine di grandezza crescente**

4) **RIPETIZIONI: si tiene conto dei doppioni con un segno di spunta: una "X" o una "V".**

Spoglio manuale delle schede elettorali.

Non più compatibile!



Esempio

Reddito procapite delle province italiane

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

FASE_1: ordinamento

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

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

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

Esempio (continua)

Fase_2: eliminazione dei doppioni

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

In questa fase si accorpano i valori ripetuti.

FREQUENZA = NUMERO DI VOLTE CHE SI PRESENTA

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

Il risultato è insoddisfacente. Ci sono troppe informazioni nella tabella

Spoglio e conteggi

Dobbiamo trovare forme semplificate di presentazione dei dati

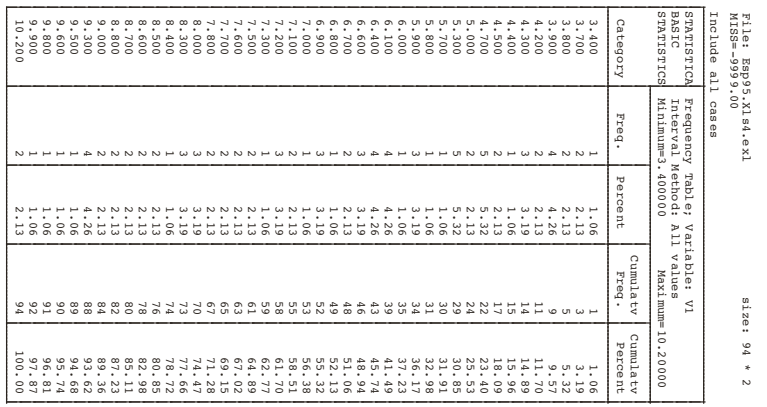
Esempio: Survey di baccelli di Indigofera per numero di semi

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

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

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

Esempio di elaborazione con "Statistica"



Le tabelle

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

Esempio:

Dati in forma narrativa

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

dati in forma tabellare

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

Le stesse informazioni sono molto più intellegibili grazie alla tabella

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

Si interviene anche con accorpamenti e ridefinizioni per semplificare la trattazione

Le tabelle/2

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

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

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

Le tabelle/3

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

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

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

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

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

Esempio

Indagine campionaria di 64 pazienti sottoposti ad una terapia
Dominio: (G, M, TM, NV,TP, P, D)

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

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

La simbologia n_i indica la frequenza della modalità i-esima

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

Nel prospetto di raccolta sapevamo quali pazienti erano guariti.

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

Esempio

Rating del debito estero da parte di Moody's

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

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

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

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

Non c'è garanzia del segreto statistico

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

Diagramma a punti

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

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

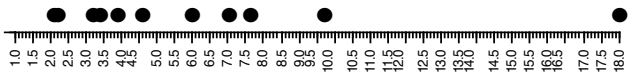
In prossimità del valore più vicino ad ogni modalità si riporta un simbolo (di solito un punto) di dimensione prefissata conforme alla dimensione della linea.

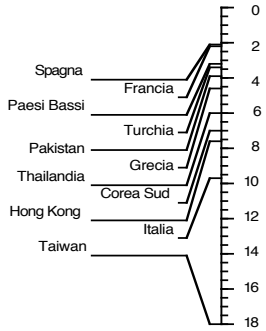
Se più modalità condividono lo stesso punto ovvero sono molto prossime, i punti saranno impilati.

Esempio

Graduatoria delle falsificazioni. Volume di contraffazioni per vari Paesi.

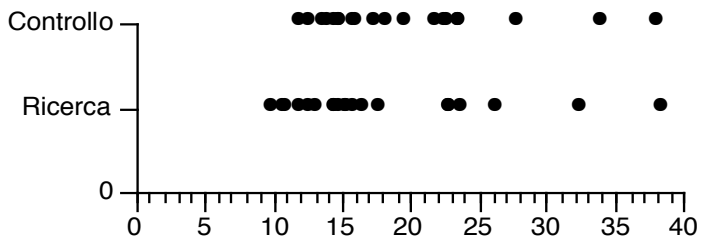
Paese	Volume	Paese	Volume	Paese	Volume	Paese	Volume
Taiwan	18.0	Pakistan	3.9	Francia	2.2	Corea Sud	7.0
Italia	9.7	Paesi Bassi	3.2	Turchia	3.4	Hong Kong	7.6
Thailandia	6.0	Spagna	2.1	Grecia	4.6		





Esempio

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



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

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

Cosa emerge dai dati?

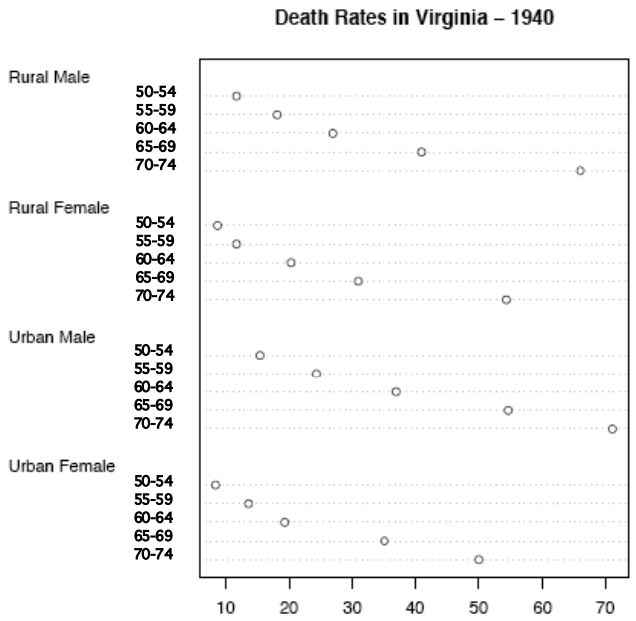


Diagramma ramo-foglia

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

Numero mensile di richiedenti un mutuo fondiario

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

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

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

Le foglie possono sono a loro volta ordinate

28 24

2|84

2|48

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

Esempio di costruzione

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

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

Prospetto di raccolta dati

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

Frequenze

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

Frequenze

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

Diagramma ramo-foglia/2

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

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

Anche il diagramma ramo-foglia disperde delle informazioni:

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

A questo può però ovviare il Digidot

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

Diagramma ramo-foglia/3

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

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

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

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

Esempio: Tasso di variazione in percentuale delle giacenze

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

Formato: ramo: XX.X, foglia: X

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

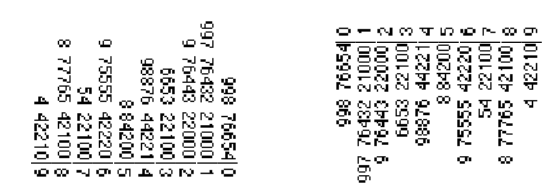
Diagramma ramo-foglia/4

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

Rilevazione campionaria del numero di fabbriche per distretto industriale

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

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



Numero di rami

Esistono vari suggerimenti:

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

[.] parte intera

Se n=50

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

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 sono tutte comprese 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 di libertà.

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

Significa che, note k-1 frequenze relative la mancante si ricava da questo 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.

Esermpio

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

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

da "Il Diavolo sulle Colline" di C. Pavese

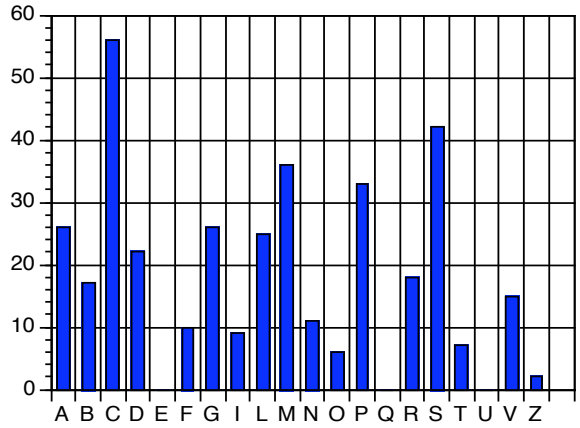
La "a" e la "e" sono dominanti. Le consonanti sono poco presenti alla fine della parola. La "u" è rarissima.

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

La distribuzione dei cognomi

La letttera iniziale dei cognomi ha una distribuzione in cui si notano due picchi in corrispondenza della C e della S. Perché?

Lettera	f.a.	f.r
A	26	7.18%
B	17	4.70%
C	56	15.47%
D	22	6.08%
E	0	0.00%
F	10	2.76%
G	26	7.18%
I	9	2.49%
L	25	6.91%
M	36	9.94%
N	11	3.04%
O	6	1.66%
P	33	9.12%
Q	0	0.00%
R	18	4.97%
S	42	11.60%
T	7	1.93%
U	0	0.00%
V	15	4.14%
Z	2	0.55%
	361	



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 sono presentate con la convenzione

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

Che comporta incertezza nell'assegnare alla classe giusta le modalità limite

Esempio

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

a) Calcolare ampiezze e valori centrali delle classi;

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

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

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

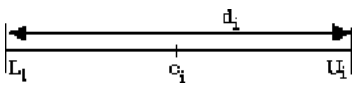
A1) 0.4285, 0.6795, 0.81575, 1.02165
A2) 0.4210, 0.0810, 0.19150, 0.22030

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

Caratterizzazione delle classi

Le classi hanno due elementi importanti:

Ampiezza : $d_i = (U_i - L_i)$



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

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

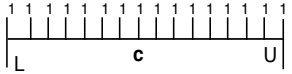
Densità di frequenza

E' una utile caratteristica delle classi $h_i = \frac{f_i}{d_i} = \frac{f_i}{(U_i - L_i)}; \quad i = 1,2,\ldots,k$

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

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

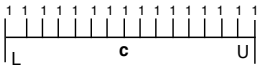
L'indicazione data dalla densità di frequenza è esatta purché la ripartizione delle unità all'interno della classe sia uniforme e cioè del tipo:



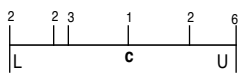
Tipicità del valore centrale

Dipende dalla configurazione con cui si presentano le modalità.

è questionabile nelk caso della uiforme

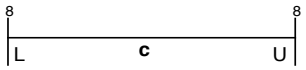


in quanto non c'è ragione di preferire il punto di mezzo.



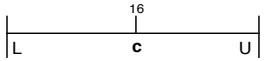
"c" è isolato ed esprime solo se stessa

"c" è poco rappresentattivo



Non rappresenta nessuno

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



Esempio

Consideriamo le seguenti classi:

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

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

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

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

3) **Si sorteggia la classe da aumentare di uno.**

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

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

Limiti delle classi estreme indeterminati

Non sempre le classi estreme hanno limiti espliciti

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

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

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

Ipotesi dell'altezza proporzionale

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

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

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

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

Ne consegue che

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

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

X_{min} ed X_{max} sono le modalità più piccole del dominio (riscontrabili quindi nella popolazione e non necessariamente riscontrate nel data set)

Esempio

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

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

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

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

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

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

Numero delle classi

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

Tassi minimi di sconto commerciale

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

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

Compatte

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

Buone, ma c'è di meglio

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

Sparse

Numero delle classi/2

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

- **Si deve porre un limite minimo per non accorpare troppo valori eterogenei in classi molto vaste**
- **Si deve porre un limite massimo per non vanificare la semplificazione che motiva il raggruppamento**

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

Un'utile regola è quella di Sturges con k arrotondato per difetto o per eccesso secondo la regola del 5 $K = 1 + 3.322 * \mathrm{Log}_{10}(n)$

ESEMPIO:

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

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

Ampiezza delle classi

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

- **Le classi dovrebbero essere della stessa ampiezza per facilitare il confronto tra i diversi livelli raggiunti dalla variabile**
- **Gli estremi dovrebbero essere multipli di 2, 10 e 5 per la loro migliore leggibilità.**
- **La comune ampiezza potrebbe essere ottenuta con la formula**

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

ESEMPIO: per i valori

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

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

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

Ampiezza delle classi/2

L'ampiezza delle classi non deve essere necessariamente costante

- **Se i valori si addensano intorno a valori piccoli o grandi poche classi includerebbero la maggior parte delle modalità.**
- **Le classi debbono essere meno numerose (o più ampie se ci sono poche modalità. Debbono essere più numerose (più sottili) per livelli più densi**

ESEMPIO: rilevazioni dei danni provocati da catastrofi naturali

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

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

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

La regola dell'ampiezza era inadatta

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

Costruzione pratica di una tabella in classi

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

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

Esempio

Indagine campionaria sui tempi di espletamento di un certo compito

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

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

b) $R = 98.15 - 2.42 = 95.73$

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

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

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

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

Se antiestetiche, le due cifre finali possono essere eliminate ponendo

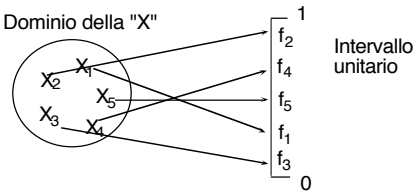
$$U_i = L_{i+1}$$

che però lascerà incertezze sulle modalità estreme

La funzione di distribuzione empirica

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

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



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

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

Grafico della funzione di distribuzione

Molto utile per facilitare il confronto tra distribuzioni di frequenza

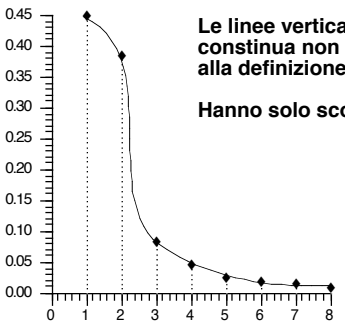
E' costituito dai punti di coordinate $(X_i, f_i), \quad i = 1,2,\ldots,n$

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

A questo fine si usano anche delle linee spezzate o continue

Famiglie abbonate a Cosmopolitan per numero di componenti

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



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

Hanno solo scopo estetico

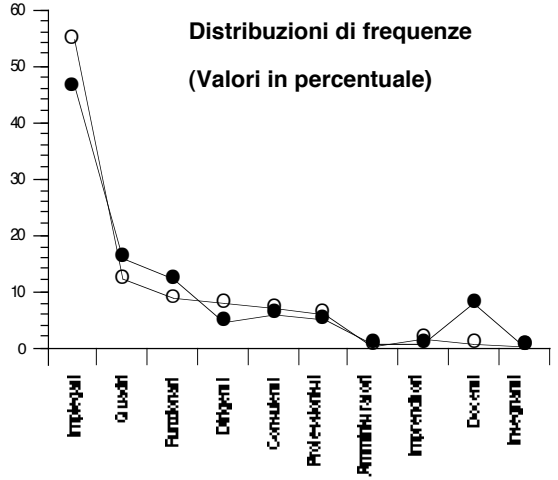
Esempio

Professioni scelte dai laureati di una università del Nord Italia

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

Distribuzioni di frequenze

(Valori in percentuale)



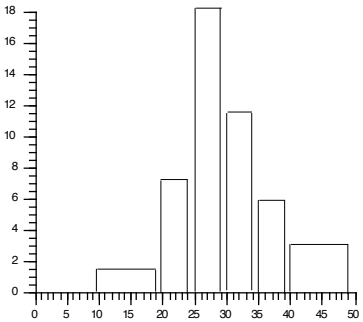
Le differenze maggiori tra ECON.AZ. che diventano impiegati e DES che diventano docenti universitari

Esempio

Le distribuzioni passano attraverso due diverse semplicazioni: tabellare e grafica

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

X=Anni di attività



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

La separazione tra rettangoli è plausibile solo per variabili discrete

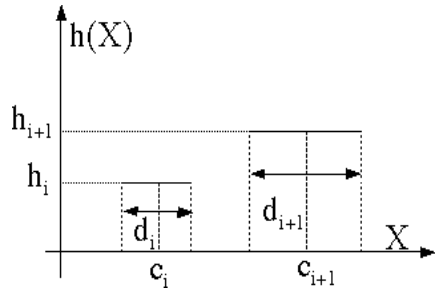
Funzione di distribuzione (per classi)

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

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

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

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



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

L'istogramma delle frequenze

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

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

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

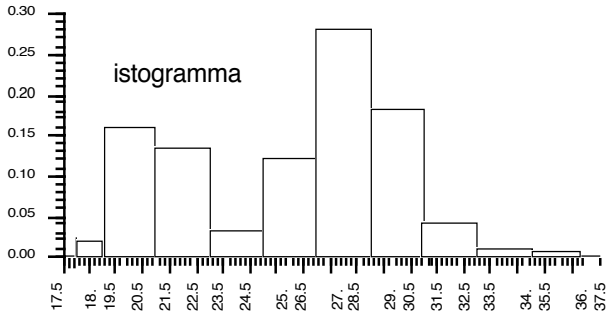
L'area totale dei rettangoli è pari alla costante α *Di solito $\alpha=1$*

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

Esempio

Lunghezza del corpo di un campione di sogliole

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



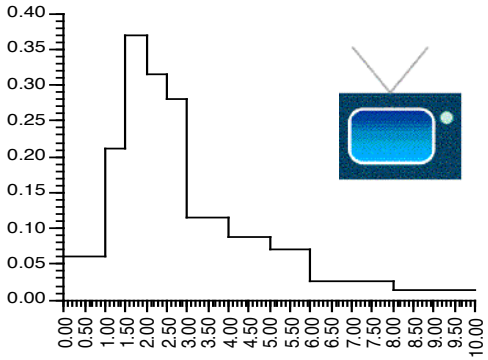
La somma bloccata permette di controllare l'area dei singoli rettangoli e quella complessiva

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

Esempio

Famiglie per il tempo complessivo in cui l'apparecchio rimane acceso.

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



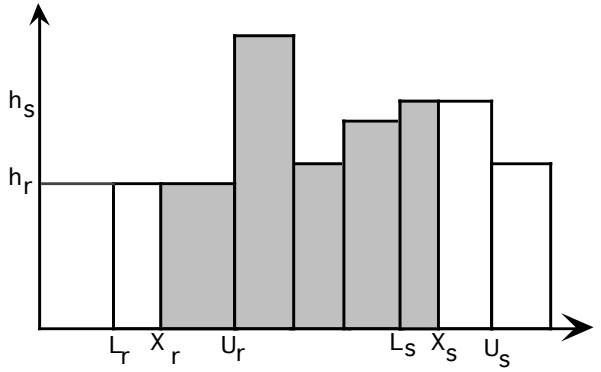
L'ipotesi di equifrequenza nella classe è criticabile perché:

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

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

Additività

Deriva dalla natura di area della frequenza relativa



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

Esempio

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

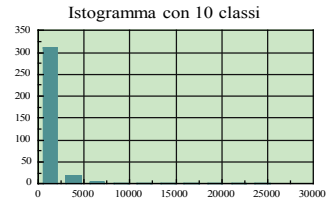
Ignorando i dati di dettaglio, quale percentuale si può presumere ricada tra il 24% e il 35%?

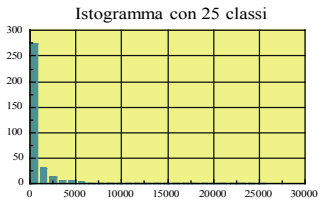
% Spreco	Hard disk	f_i	d_i	h_i
0 - 4	45	0.075	4	0.0188
4 - 8	98	0.163	4	0.0408
8 -15	126	0.210	7	0.0300
15 - 30	200	0.333	15	0.0222
30 - 40	91	0.152	10	0.0152
40 - 50	40	0.067	10	0.0067
	600	1.000		

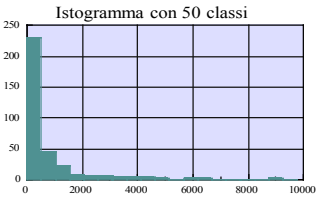
$$A[24,35] = A[24;30] + A[30;35]$$

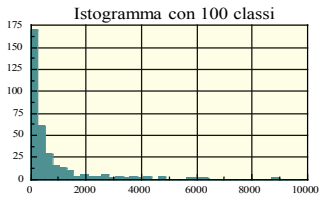
$$= 6 * 0.0222 + 5 * 0.0152 = 0.2092$$

Modifica della forma secondo le classi

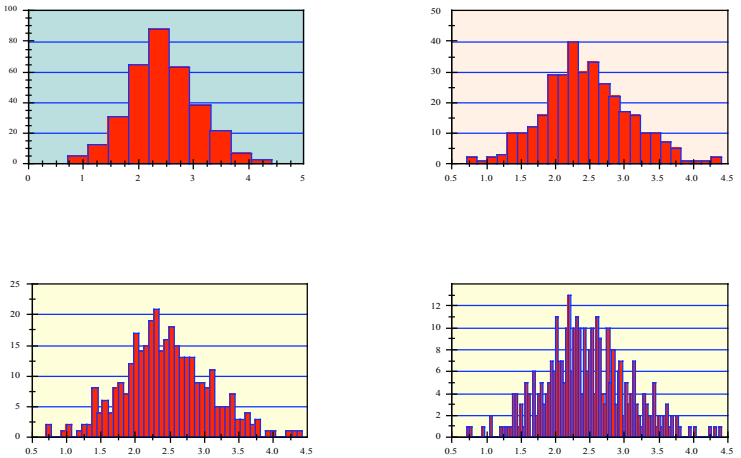








Modalità in scala logaritmica



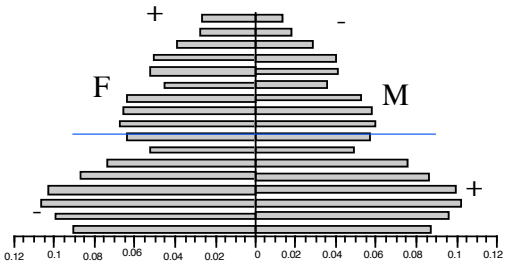
Confronto di distribuzioni

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

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

La piramide della popolazione

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



Classi giovani: + maschi; altre classi: + femmine

Poligono di frequenza

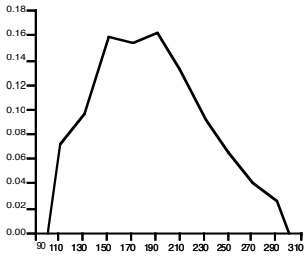
Grafico che discende dall'istogramma ottenuto riportando in un sistema cartesiano i valori centrali delle classi e le frequenze relative

$$(c_i,\ f_i);\quad i=1,2,\ldots,k \qquad \left(c_1-\frac{d_1}{2},\ 0\right);\ \left(c_k+\frac{d_k}{2},\ 0\right)$$

i due punti convenzionali fanno partire e finire il grafico sulle ascisse

Contenuto calorico in alcuni alimenti

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

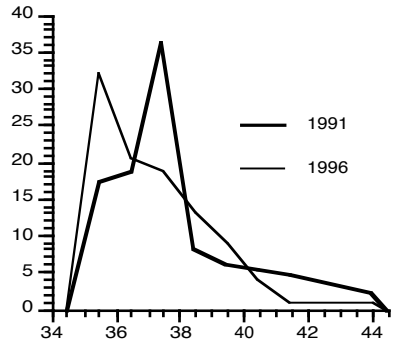


il poligono di frequenza riporta solo il profilo esterno dell'istogramma rendendo Più facile la percezione

Esempio

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

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

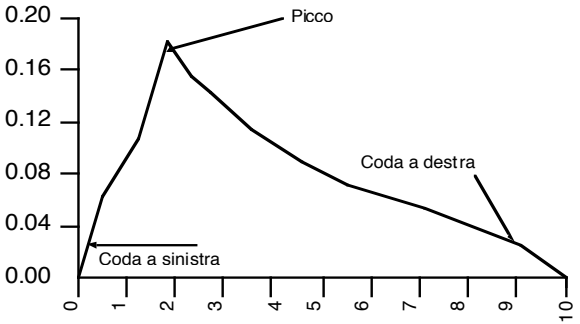


E' evidente a riduzione dell'orario più praticato: da 37 a 35 ore.

Esempio

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

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



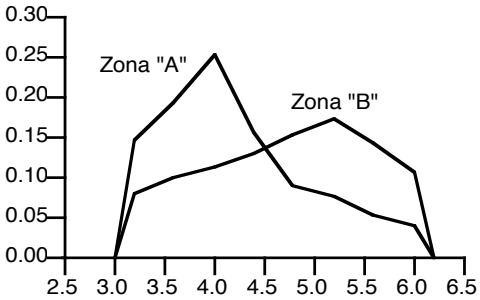
Il picco è il livello in cui la frequenza è massima.

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

Esempio

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

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

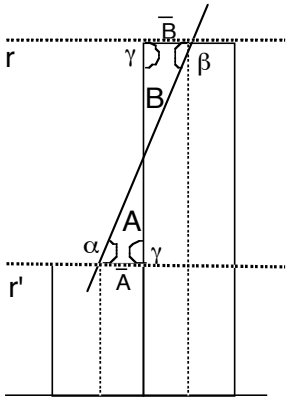


Le differenze sono forti sia al centro che nelle code segno che la concentrazione di sodio segue meccanismi diversi nelle due zone.

Area sottesa

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

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



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

Esempio

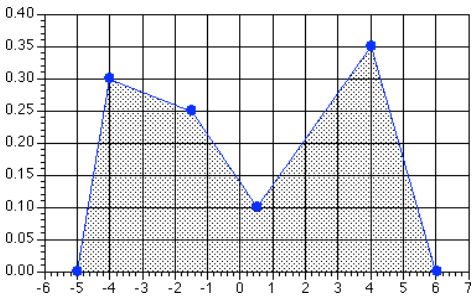
Variazioni di una quotazione azionaria

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

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

Punti convenzionali

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



Frequenze cumulate

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

Frequenza assoluta cumulata

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

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

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

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

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

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

Formule per le frequenze cumulate

Le frequenze cumulate verificano le seguenti relazioni:

$$F_1 = f_1$$

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

$$F_k = 1$$

schema ricorsivo

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

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

x_i		n_i	f_i	N_i	F_i
14	17	6	0.0789	6	0.0789
18	21	11	0.1447	6 + 11 = 17	0.2237
22	25	29	0.3816	6 + 11 + 29 = 17 + 29 = 46	0.6053
26	29	24	0.3158	6 + 11 + 29 + 24 = 46 + 24 = 70	0.9211
30	33	4	0.0526	6 + 11 + 29 + 24 + 4 = 70 + 4 = 74	0.9737
34	38	2	0.0263	6 + 11 + 29 + 24 + 4 + 2 = 74 + 2 = 76	1.0000
		76			

Uso delle frequenze cumulate

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

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

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

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

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

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

Esempio sulle cumulate

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

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

37377 famiglie di 3 componenti pari al 6.96% del totale

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

Il 3.01% ha, almeno, 3 componenti

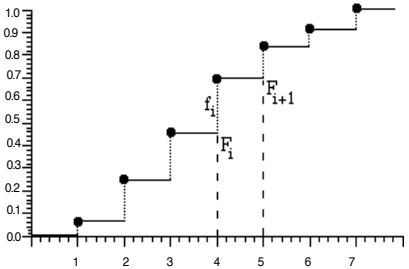
La funzione di ripartizione empirica

E' la sintesi delle frequenze relative cumulate:

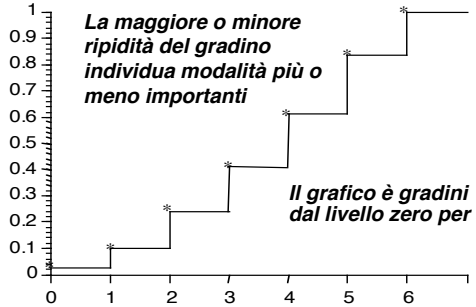
$$F(X)=\begin{cases}0 & \text{se } X<X_{(1)}\\ F_i & \text{se } X_{(i)}\le X\le X_{(i+1)} \text{ per } i=1,2,\ldots,n-1\\ 1 & \text{se } X\ge X_{(n)}\end{cases}$$

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

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



Considerazioni



La maggiore o minore ripidità del gradino individua modalità più o meno importanti

Il grafico è gradini che si alzano dal livello zero per arrivare a 1.

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

La funzione di ripartizione empirica è non decrescente:

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

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

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

La F.R.E. per dati in classi

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

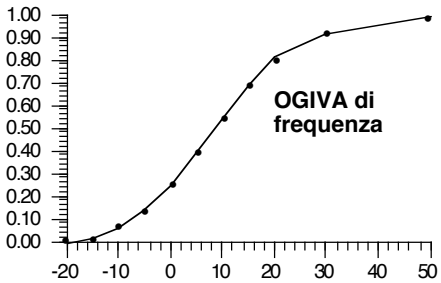
$$h_i=f_i/d_i$$

Si ipotizzano:

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

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

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



Esempio

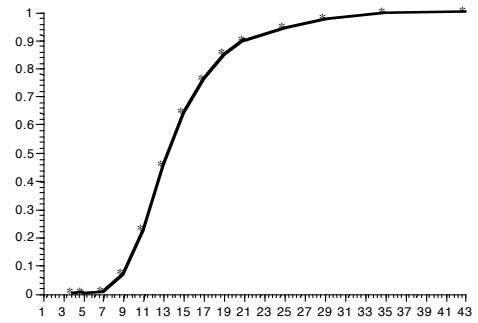
Campione di degenti classificati per tempo trascorso tra ricovero e fase acuta della malattia

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

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

In questo tratto l'ogiva ha la sua massima ripidità

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



La funzione di ripartizione complementare

RICORRE NELLO STUDIO
Della distribuzione dei redditieri per importo posseduto

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

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

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

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

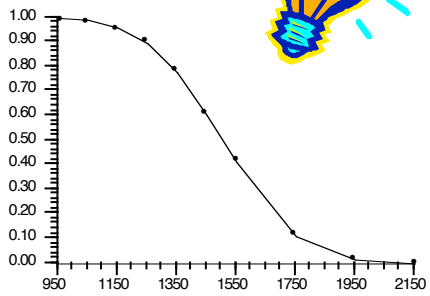
La sua rappresentazione grafica è simile alla curva di ripartizione solo che ora i punti dell'ogiva hanno coordinate (L_i, G_i).

Esempio

Durata (in ore) di un campione di lampadine.

Rappresentazione della funzione di ripartizione complementa

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



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

La funzione quantile empirica

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

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

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

E' evidente che

$$F(X_p) = p$$

Se F è invertibile allora

$$X_p = F^{-1}(p)$$

La funzione quantile/2

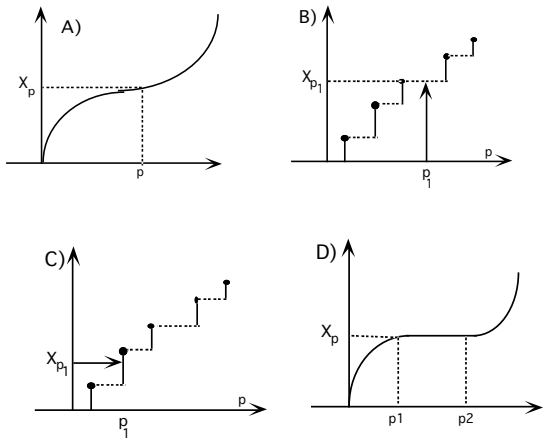
E' positiva e non decrescente ed ha un grafico identico a quello della funzione di ripartizione empirica, ma con gli assi traslati.

A) Situazione ideale: curva continua e relazione biunivoca

B) Non esiste il percentile;

C) Ne esiste più d'uno;

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



La funzione quantile empirica

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

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

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

Esempio:

$n = 17;\; p = 0.65; \quad n*p = 17*0.65 = 11.05,\; i = [n*p] = [11.05] = 11;$
$\gamma = 1; \quad X_{0.65} = X_{(12)}$

La funzione di graduazione/2

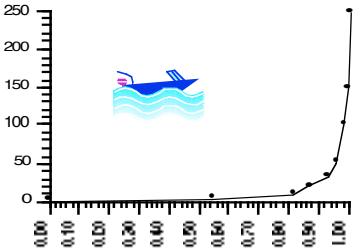
Nel caso di variabili continue, la solita ipotesi di uniformità nelle classi implica:

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

Naviglio a motore per classi di stazza lorda

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

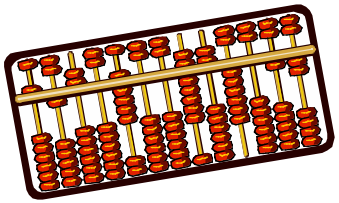
La funzione quanitle molto appiattita rivela una preponderanza delle classi più piccole



Premessa

La distribuzione di frequenze è il risultato della raccolta dati che in essa sono organizzati e semplificati.

Rimangono però ancora tante le informazioni da considerare.



Il nostro obiettivo è riassumerne gli aspetti salienti in pochi valori numerici (parametri statistici)

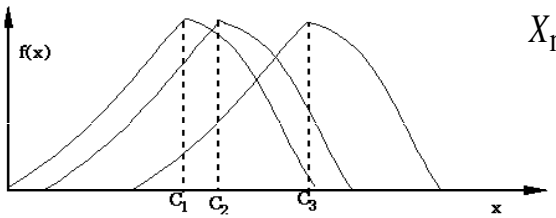
Essi consentono il confronto tra distribuzioni di variabili diverse oppure della stessa variabile in epoche, luoghi ed occasioni diverse.

Concetto di media

Individua il livello di maggiore addensamento delle modalità ovvero la categoria o il valore (espresso nella stessa unità di misura) intorno alla quale sembra ruotare l'intera rilevazione.

Non è necessario che appartenga al dominio della variabile (può essere un valore fittizio).

Bisogna e basta che rispetti la condizione di internalità:



$$X_{\min} \le Media \le X_{\max}$$

Per valori almeno ordinali

Esempi

Variabile qualitativa

Linguaggio	Parlanti (in milioni)
Mandarino	740
Inglese	403
Russo	277
Spagnolo	266
Indostano	264
Arabo	160
Bengalese	155
	2265

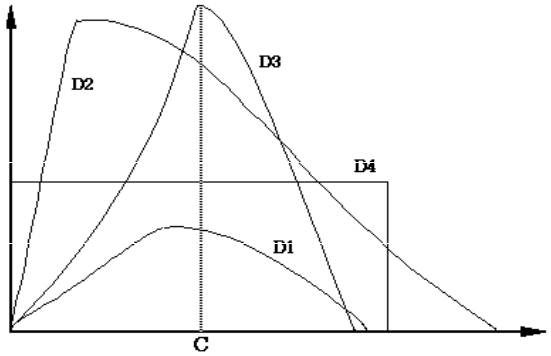
Variabile quantitativa.

Numeri di albini in famiglie di 5 figli	Numero di famiglie
1	25
2	23
3	10
4	1
5	1
	60

Nel primo caso la "media" può essere qualsiasi modalità; nel secondo può essere solo un numero compreso tra 1 e 5

Mancanza di univocità

Il processo di estrema sintesi che porta al collassamento della distribuzione di su di un solo valore è il limite dei parametri di posizione



Distribuzioni molto diverse possono presentare la stessa "media".

Conoscendo questa non è univocamente nota la situazione che l'ha determinata.

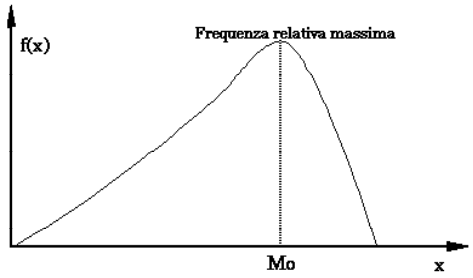
La moda

E' il più facile da calcolare, ma anche quello più grezzo. Si identifica con la modalità corrispondente alla frequenza relativa maggiore:

$$M_o = \left\{ X_m \middle| f_m \geq f_i \quad per \quad i = 1,2,\ldots,k \right\}$$

Classificazione del molare inferiore destro per numero di canali su 1000 soggetti.

Canali x_i	Soggetti n_i	Freq. Rel. f_i
1	2	0.0002
2	914	0.9140
3	76	0.0076
4	8	0.0008
	1000	1.0000



La determinazione della Moda è influenzata solo dai rapporti ordinali tra le frequenze

Esempio

Articoli su riviste per sport prevalente commentato

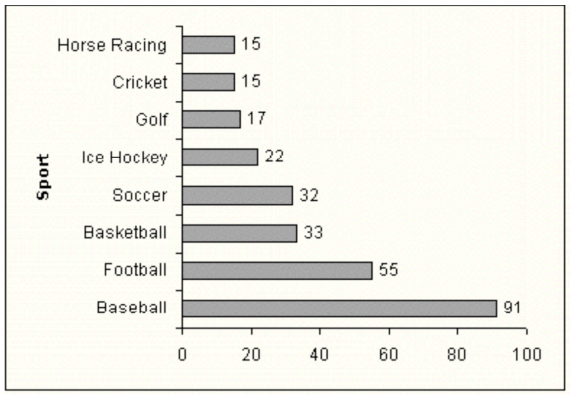
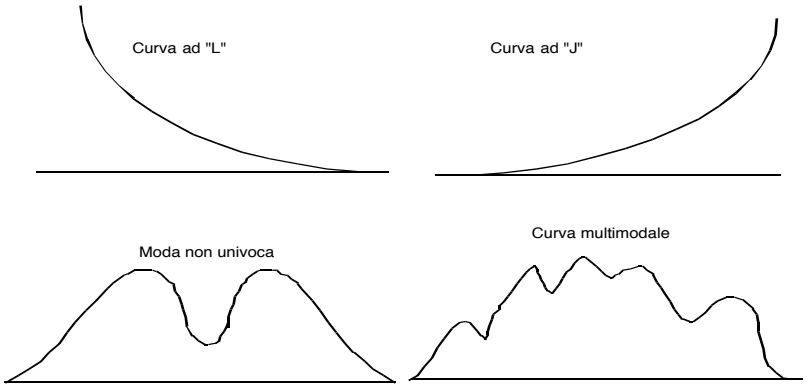


Figure 1. No. Articles in CIS

Qual è la moda?

Amodalità e multimodalità

La moda può essere assente dalla distribuzione.



La frequenza massima può non corrispondere ad una unica modalità.

Inoltre, pur esistendo, la moda può non essere significativa, cioè le frequenze relative dei vari massimi potrebbero differire troppo poco

Difetto della moda

Poiché non usa tutti i dati la moda può dare indicazioni fuorvianti

ESEMPIO:

n=20 uomini arrestati per violenza in famiglia furono sottoposti a vigilanza speciale per 2 anni. Ecco la distribuzion degli arresti alla fine del periodo:

Arresti	Criminali
0	8
1	1
2	1
3	1
4	2
5	4
6	3
	20

La moda è M_o=0 arresti.

La frequenza modale è elevata (doppia della 2ª in ordine di grandezza)

Tuttavia dire che zero arresti è tipico nasconderebbe un gruppo di criminali che ha reiterato il reato ben 5 o 6 volte

Modalità in classi

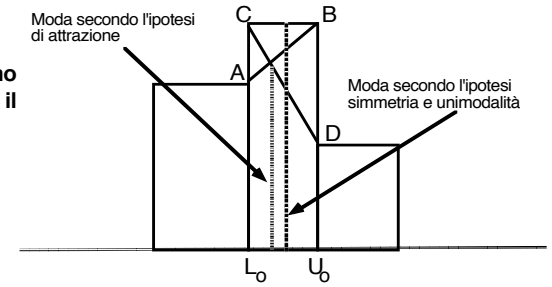
La frequenza relativa più elevata individuerà la classe modale.

Metodo N. 1

I valori della classe modale hanno uguale frequenza. La Moda è il valore centrale della classe:

$$M_o = c_m = \frac{U_m + L_m}{2}$$

$$f_m \geq f_i \quad \forall i$$



Metodo N. 2

La moda è più vicina all'estremo L_m o U_m che confina con la classe con più frequenza (che esercita maggiore "attrazione")

$$M_o = L_m + \frac{(f_m - f_{m-1})}{[(f_m - f_{m-1}) + (f_m - f_{m+1})]} * (U_m - L_m)$$

Esempio

Ordinazioni per importi

Classe modale: (4 - 5.9);

Valore centrale classe modale:

$$\frac{4.0 + 5.9}{2} \cong 5$$

Importi		Ordini	f_i	d_i	h_i
0.0	1.9	102	0.1726	1.9	0.0908
2.0	3.9	175	0.2961	1.9	0.1558
4.0	5.9	208	0.3519	1.9	0.1852
6.0	7.9	76	0.1286	1.9	0.0677
8.0	9.9	23	0.0389	1.9	0.0205
10.0	11.9	7	0.0118	1.9	0.0062
		591	1.0000		

Attrazione: $$4.0 + \left[\frac{0.1852 - 0.1558}{(0.1852 - 0.1558) + (0.1852 - 0.0677)}\right] * 1.9 = 4.38$$

Ampiezze diverse

Se le modalità sono espresse in classi si impone la considerazione della loro eventuale differenza di ampiezza.

L_i	U_i	n_i	d_i	h_i	c_i
3	5	8	2	4	4
5	11	18	6	3	8

In questi casi occorre fare riferimento non più alle frequenze relative, bensì alle densità di frequenza ovvero alle altezze.

$$h_i = \frac{f_i}{d_i}$$

$$M_O = c_m = \frac{L_m + U_m}{2} \qquad \text{dove } (h_m \geq h_i \text{ per } i=1,2,\ldots,k)$$

$$M_O = L_m + \frac{(h_m - h_{m-1})}{[(h_m - h_{m-1}) + (h_m - h_{m+1})]} * (U_m - L_m)$$

La mediana

La Mediana è la modalità che è preceduta e seguita dalle altre con uguale frequenza, si trova cioè in posizione centrale nella graduatoria delle modalità.

ESEMPIO:

Soldati schierati in ordine di altezza: la fila posta al centro è quella mediana

Ricordando che, se non altrimenti indicato, le modalità sono ordinate in senso crescente, la Mediana di "n" osservazioni è data da:

$$M_e = \begin{cases} X_{\left(\frac{n+1}{2}\right)} & se\ "n"\ è\ dispari \\ \dfrac{X_{(n/2)} + X_{(n/2)+1}}{2} & se\ "n"\ è\ pari \end{cases}$$

Esempio

Una radiografia è stata scomposta in pixel e su ciascuno di questi si è misurato il tono di grigio

Classe modale: (31-49);

Valore centrale classe modale:

$$\frac{31+49}{2} = 40$$

Riflettenza		Pixel	f_i	d_i	h_i
0	30	54	0.0113	30.0	0.0004
31	49	613	0.1277	18.0	0.0071
50	98	421	0.0877	48.0	0.0018
99	127	716	0.1492	28.0	0.0053
128	160	432	0.0900	32.0	0.0028
161	191	798	0.1663	30.0	0.0055
192	240	1579	0.3290	48.0	0.0069
241	255	187	0.0390	14.0	0.0028
		4800	1.0000		

Attrazione: $$31 + \left[\frac{0.0071 - 0.0004}{(0.0071 - 0.0004) + (0.0071 - 0.0018)}\right] * 18 = 41.05$$

Esempio

Costi di estrazione dollaro-barile.

Usa/Alaska	7.5	Canada	5.0
Messico	6.0	Venezuela	6.2
Argentina	15.1	Medio Oriente	2.5
Indonesia	10.7	Africa	7.4
Nord Europa	17.5	URSS	6.3

I dati ordinati sono: {2.5,5.0,6.0,6.2,6.3,7.4,7.5,10.7,15.1,17.5}

Poichè n=10 (cioè n è pari)

$$M_e = \frac{X_{\left(\frac{10}{2}\right)} + X_{\left(\frac{10}{2}+1\right)}}{2} = \frac{X_{(5)} + X_{(6)}}{2} = \frac{6.3 + 7.4}{2} = 6.85$$

Se si aggiunge il dato 19.4 allora n=11 è dispari per cui

$$M_e = X_{\left(\frac{10+1}{2}\right)} = X_{(6)} = 7.4$$

Da notare che 7.4 è un dato osservato e 6.85 è un dato fittizio

Proprietà della mediana

La Mediana rende minima la somma dei moduli degli scarti delle modalità. Supponiamo che A>0.

$$Q(A) = \sum_{i=1}^{k} |X_i - A| f_i$$

è minima se A=M_e.

$$\begin{aligned}
Q(A) &= \sum_{X_{(i)}>A} \left(X_{(i)} - A\right) f_{(i)} + \sum_{X_{(i)}\le A} \left(A - X_{(i)}\right) f_{(i)} \\
&= \sum_{X_{(i)}>A} X_{(i)} f_{(i)} - \sum_{X_{(i)}>A} A f_{(i)} + \sum_{X_{(i)}\le A} A f_{(i)} - \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \pm \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \\
&= \sum_{X_{(i)}>A} X_{(i)} f_{(i)} - A\left(\sum_{X_{(i)}>A} f_{(i)}\right) + A\left(\sum_{X_{(i)}\le A} f_{(i)}\right) - \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} + \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} - \sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \\
&= \left[\sum_{X_{(i)}>A} X_{(i)} f_{(i)} + \sum_{X_{(i)}\le A} X_{(i)} f_{(i)}\right] - A\left[1 - F(A)\right] + AF(A) - 2\sum_{X_{(i)}\le A} X_{(i)} f_{(i)} \\
&= C - A\left[1 - 2F(A)\right] - 2\sum_{X_{(i)}\le A} X_{(i)} f_{(i)}
\end{aligned}$$

dove C non dipende da A e l'ultimo termine cresce con A.

Quindi: Q(A) è decrescente per A tale che F(A)<0.5 ed è crescente per F(A)>0.5; il minimo è perciò raggiunto per F(A)=0.5 che corrisponde a A=M_e.

La mediana per dati raggruppati

La mediana corrisponde alla modalità più piccola tra quelle che hanno frequenza relativa cumulata maggiore o uguale a 0.5

$$Min\{x \in S | F(x) \ge 0.5\}$$

ESEMPIO:

Classificazione dei clienti di un punto vendita per numero di acquisti effettuati nel mese

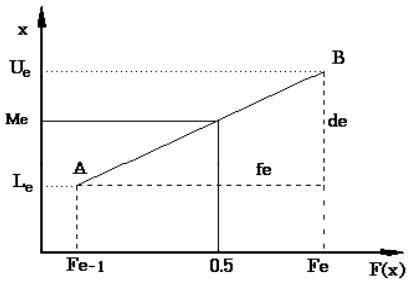
Acquisti	Clienti	f_i	F_i
0	40	0.0964	0.0964
1	69	0.1663	0.2627
2	95	0.2289	0.4916
3	111	0.2675	0.7590
4	74	0.1783	0.9373
5	26	0.0627	1.0000
	415	1.0000	

La Mediana è "3 acquisti"

Modalità in classi

E' possibile individuare solo la classe mediana, ovvero quella cui corrisponde la frequenza relativa cumulata di 0.5.

Per calcolare la mediana si ipotizza che le unità siano uniformi nella classe mediana per cui i segmenti della curva di graduazione sono rette ascendenti.



La retta AB ha equazione

$$X = L_e + \frac{(F - F_{e-1})}{h_e}$$

$$h_e = \frac{f_e}{d_e}$$

$$M_e = L_e + \frac{(0.5 - F_{e-1})}{h_e}; \quad per\ "e"\ tale\ che\ F_e = \underset{1\le j\le k}{Min}\{F | F_j \ge 0.5\}$$

Esempio

Uno studio di consulenza ha classificato le operazioni di auditing per la revisione dei conti annuali secondo la durata in giorni. Calcolo della mediana

Durata		Revisioni	f_i	F_i
5	7	5	0.0595	0.0595
8	10	9	0.1071	0.1667
10	14	14	0.1667	0.3333
15	19	18	0.2143	0.5476
20	24	15	0.1786	0.7262
25	29	12	0.1429	0.8690
30	34	11	0.1310	1.0000
		84	1.0000	

$$M_e = 15 + \frac{(0.5 - 0.3333)}{0.2143/4} = 18.11$$

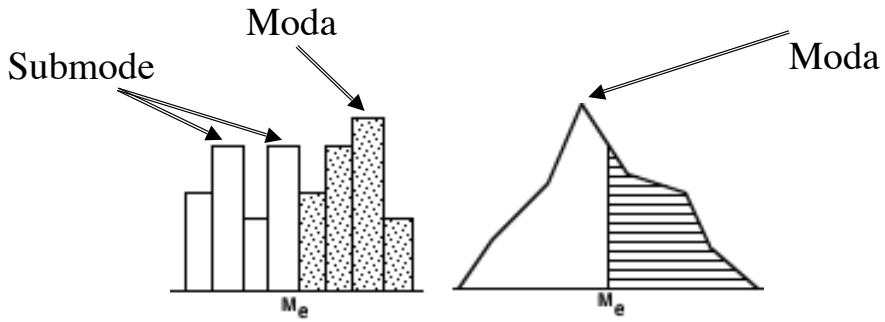
Il calcolo della mediana avviene in due passi:

1) Si individua la classe mediana

2) Si interpola per ottenere il valore puntuale.

Individuazione grafica

La mediana corrisponde alla retta X=M_e che separa due parti uguali dell'istogramma o dell'area sottesa al poligono di frequenza (se le classi hanno uguale ampiezza).



I quantili (o percentili)

L'idea di valori di soglia che suddividano le modalità in particolari gruppi può essere generalizzata definendo il quantile di ordine "p"

Modalità discrete

$$X_p = (1-\gamma)X_{(i)} + \gamma X_{(i+1)},\ 0 < p < 1;\ i \le np < i+1;\ \gamma = \begin{cases} 0.5 & se\ [np] = np \\ 1 & se\ [np] < np \end{cases}$$

Modalità continue non in classi

$$X_p = (1-\gamma)X_{(i)} + \gamma X_{(i+1)} \quad i = [np + 0.5];\ \gamma = np + 0.5 - i$$

Modalità dense o continue in classi

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

il quantile supera il p% delle modalità ed è superato dall' (1-p)%

Esempio_1

Discrete o dense singole

Consideriamo le n=18 rilevazioni degli arrivi di auto ad un punto di imbarco e calcoliamo il quantile del 17%.

612	623	666	744	883	898
964	970	983	1003	1016	1022
1029	1058	1085	1088	1122	1135

$$n * p = 18 * 0.17 = 3.06 \Rightarrow [np] < np \Rightarrow \gamma = 1$$

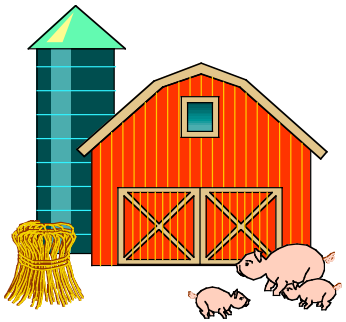
$$X_{0.17} = 0 * X_{(3)} + 1 * X_{(4)} = X_{(4)} = 744$$

Esempio_2

Continue singole

Principali coltivazioni agricole delle Marche. Anno 1998. Valori in ettari. Calcolo del quantile di ordine 0.60

Coltivazione	Superficie
Pomodoro	1'304
Pesca	1'486
Cavolfiore	1'967
Olivo	6'218
Mais ibrido	14'558
Uva da vino	24'272
Grano tenero	36'553
Girasole	38'281
Grano duro	123'049



$$n * p = 9 * 0.60 = 5.4 \Rightarrow i = [5.4 + 0.5] = [5.9] = 5;$$

$$\gamma = 5.9 - 5 = 0.9$$

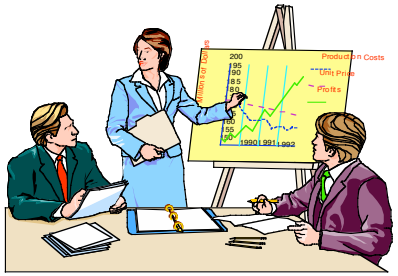
$$X_{0.6} = 0.1 * X_{(5)} + 0.9 * X_{(6)} = 23'300.$$

Esempio_3

Dimensioni delle operazioni di fusione e di acquisizione in Italia per fatturato.

Calcolo di $X_{0.80}$.

Fatturato		Operazioni	f_i	F_i
1	5	30	0.2804	0.2804
5	20	36	0.3364	0.6168
20	40	14	0.1308	0.7477
40	60	10	0.0935	0.8411
60	100	12	0.1121	0.9533
100	150	5	0.0467	1.0000
		107	1.0000	



$$X_{0.80} = 40 + \frac{(0.80 - 0.7477)}{(0.0935/40)} = 51.19$$

Definizione alternativa

I quantili possono essere definiti evitando il riferimento ai valori ordinati.

Il quantile di ordine p con 0<p<1 è dato dalla soluzione del seguente problema di ottimizzazione

$$\underset{A \in R}{Min}\left[\sum_{t \in \{t | X_t \geq A\}} p|X_t - A| + \sum_{t \in \{t | X_t < A\}} (1-p)|X_t - A|\right]$$

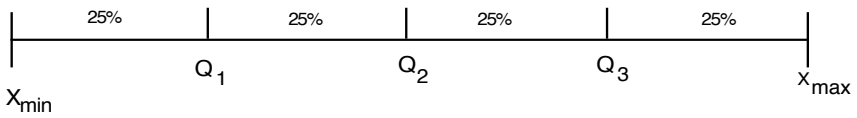
Il caso p=0.5 corrisponde alla mediana.

Questa formulazione può tornare utile per risolvere alcuni problemi di calcolo relativi all'uso dei valori assoluti.

Uso dei quantili

In genere, i quantili si usano come soglie di separazione delle unità in gruppi di numerosità prestabilita.

Fra i più usati sono da annoverare i <u>tre quartili</u> che suddividono le modalità in quattro gruppi ciascuno comprendente il 25% delle modalità:



Q_1 supera il 25% ed è superato del 75%, Q_2 coincide con la mediana ed è superato da tante unità quante ne supera esso stesso; Q_3 supera il 75% delle unità ed è superato dal restante 25%.

Tra due soglie è sempre compreso il 25% di unità.

Esempio

L'esito di una selezione per l'ammissione ad un corso universitario a numero chiuso è riassunto nella tabella.

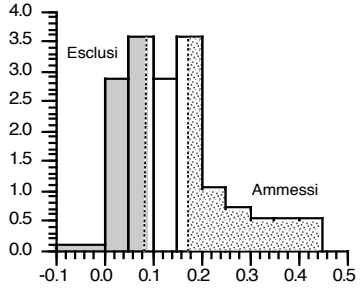
La commissione decide di ammettere il 40% con punteggio più alto, di escludere il 25% inferiore e di sottoporre il restante 35% a dei test suppletivi.

Quali sono le soglie di divisione?

$$X_{0.25} = 0.05 + \left(\frac{0.25 - 0.152}{0.205}\right)(0.10 - 0.05) = 0.0843;$$

$$X_{0.60} = 0.15 + \left(\frac{0.60 - 0.536}{0.143}\right)(0.20 - 0.15) = 0.1724$$

Punteggio		Candidati	f_i	F_i
<0.00		1	0.009	0.009
0.00	0.05	16	0.143	0.152
0.05	0.10	23	0.205	0.357
0.10	0.15	20	0.179	0.536
0.15	0.20	16	0.143	0.679
0.20	0.25	20	0.179	0.857
0.25	0.30	6	0.054	0.911
0.30	0.35	4	0.036	0.946
0.35	0.40	3	0.027	0.973
>0.40		3	0.027	1.000
		112	1.000	



La media aritmetica

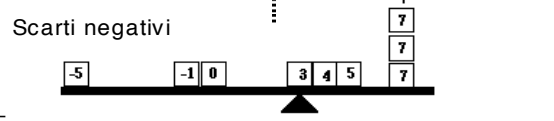
E' la media per antonomasia, quella che si sottintende se non si qualifica ulteriormente il termine di media.

$$\overline{X} = \sum_{i=1}^{k} X_i f_i = X_1 f_1 + X_2 f_2 + \ldots + X_k f_k$$

La media aritmetica è il punto di equilibrio fisico delle modalità spaziate in base al loro valore numerico e pesate con le frequenze relative:

x_i	n_i	f_i
-5	1	1/9
-1	1	1/9
0	1	1/9
3	1	1/9
4	1	1/9
5	1	1/9
7	3	3/9
	9	

$$Ma = -5*\frac{1}{9} - 1*\frac{1}{9} + 3*\frac{1}{9} + 4*\frac{1}{9} + 5*\frac{1}{9} + 7\frac{3}{9} = \frac{27}{9} = 3$$



Esempio

Numero di link visitati per ricerche su Google (max 8)

Siti x_i	Ricerche n_i	Freq. Rel. f_i	Prodotti $x_i f_i$
0	161	0.0042	0.0000
1	152	0.0304	0.0304
2	3957	0.1043	0.2086
3	7603	0.2004	0.6012
4	10263	0.2705	1.0821
5	8498	0.2240	1.1200
6	4984	0.1314	0.7883
7	1055	0.0278	0.1947
8	264	0.0070	0.0557
	37937	1.0000	4.0809

Da notare che si può calcolare moltiplicando le modalità per le frequenze assolute e poi dividendo l'ammontare ottenuto per il totale delle frequenze

Proprietà della media aritmetica

La media aritmetica, se sostituita alle modalità, mantiene inalterato l'ammontare complessivo nella rilevazione.

Il totale delle modalità è infatti: $T = \sum_{i=1}^{k} x_i n_i$

Se al posto di X_i si pone la media aritmetica si ottiene

$$\sum_{i=1}^{n} \overline{X} n_i = \overline{X} \sum_{i=1}^{n} n_i = \overline{X} n = \frac{T}{n} n = T$$

Quindi la media aritmetica è quella quantità che ciascuna unità avrebbe se tutte avessero la stessa parte di variabile.

Internalità

Considerando le modalità in senso ascendente saranno vere le relazioni:

$$\sum_{i=1}^{k} X_{\min} f_i \le \sum_{i=1}^{k} X_i f_i \le \sum_{i=1}^{k} X_{\max} f_i$$

Ogni addendo della prima somma è inferiore o uguale ad ognuno della seconda che a loro volta sono inferiori o uguali a quelli della terza.

Ne consegue che:

$$X_{\min} \sum_{i=1}^{k} f_i \le \sum_{i=1}^{k} X_i f_i \le X_{\max} \sum_{i=1}^{k} f_i \Rightarrow X_{\min} \le \sum_{i=1}^{k} X_i f_i \le X_{\max}$$

Somma nulla degli scarti

La media aritmetica rende nulla la somma degli scarti tra le modalità e la media.

$$\sum_{i=1}^{k}(X_i - \overline{X})f_i = \sum_{i=1}^{k} X_i f_i - \sum_{i=1}^{k} \overline{X} f_i = \sum_{i=1}^{k} X_i f_i - \overline{X} = \sum_{i=1}^{k} X_i f_i - \overline{X} = \overline{X} - \overline{X} = 0$$

ESEMPIO

Numero di dipendenti formati nei comuni e nelle province per settori.

Area	Dipendenti	Scarto
Interventi settoriali	15913	5887
Managerialità	10801	775
Organizzazione	10711	685
Controllo di gestione	9362	-664
Informatizzazione	3938	-6088
Gestione del personale	9431	-595
Totale	60156	0

La media aritmetica è $\overline{X} = 10026$ che, annullando la somma degli scarti, si conferma valore di equilibrio per la distribuzione.

Minimo della somma degli scarti al quadrato

$$\sum_{i=1}^{k}(X_i - A)^2 f_i = \sum_{i=1}^{k}\left[(X_i - \overline{X}) + (\overline{X} - A)\right]^2 f_i$$

$$= \sum_{i=1}^{k}\left[(X_i - \overline{X})^2 + (\overline{X} - A)^2 + 2(\overline{X} - A)(X_i - \overline{X})\right] f_i$$

$$= \sum_{i=1}^{k}(X_i - \overline{X})^2 f_i + \sum_{i=1}^{k}(\overline{X} - A)^2 f_i + 2(\overline{X} - A)\sum_{i=1}^{k}(X_i - \overline{X}) f_i$$

$$= \sum_{i=1}^{k}(X_i - \overline{X})^2 f_i + \sum_{i=1}^{k}(\overline{X} - A)^2 f_i$$

Il terzo termine risulta nullo per la proprietà già dimostrata della media aritmetica di annullare la somma degli scarti semplici.

$$= \sum_{i=1}^{k}(X_i - \overline{X})^2 f_i + (\overline{X} - A)^2$$

il primo addendo non dipende da "A" il 2° è semplicemente un quadrato che ha il minimo nello zero raggiunto per $A = \overline{X}$

Riproducibilità per trasformazioni lineari

Quando la variabile x subisce una trasformazione lineare del tipo y=a + bx lo stesso succede alla media aritmetica.

Infatti: $$\overline{Y} = \sum_{i=1}^{k} Y_i f_i = \sum_{i=1}^{k}(a + bX_i) f_i = \sum_{i=1}^{k} a f_i + b\sum_{i=1}^{k} X_i = a + b\overline{X}$$

ESEMPIO

Bilancio delle principali squadre di calcio di serie A (in milioni). Conversione in migliaia di dollari.

Squadre	M. Lire	m. Dollari
Juventus	1847	947.18
Milan	-27093	-13893.85
Inter	-21442	-10995.90
Roma	504	258.46
Parma	-25418	-13034.87
Lazio	251	128.72
Fiorentina	-10579	-5425.13
Sampdoria	-879	-450.77
Bologna	-8822	-4524.10
	-10181.22	-5221.14

il rapporto di conversione tra milioni di lire e migliaia di dollari è:

$$a = 0.0, \quad b = \frac{1000}{1950} = 0.512821$$

$$\overline{X} = -10181.22, \quad \overline{Y} = 0.512821 * (-10181.22) = -5221.14$$

Proprietà associativa

Supponiamo di individuare "g" gruppi. Le modalità sono individuate con due indici: il primo per il gruppo ed il secondo per le unità del gruppo:

Gruppi	*Valori*			*Unità*
G_1	X_{11}	X_{12}	X_{1k_1}	n_1
G_2	X_{21}	X_{22}	X_{2k_2}	n_2
$\vdots$				
G_g	X_{g1}	X_{g2}	X_{gk_g}	n_g

$$\sum_{i=1}^{g} n_i = n$$

Per ogni gruppo si può calcolare una specifica media aritmetica

$$\mu_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}; \quad i = 1, 2, \ldots, g$$

Proprietà associativa/2

La proprietà associativa consente di ricavare la media aritmetica complessiva:

$$\mu = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i} X_{ij}}{n} = \frac{\sum_{i=1}^{g} n_i \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}}{n} = \frac{\sum_{j=1}^{g} n_i \mu_i}{n} = \sum_{j=1}^{g} \mu_i f_i$$

ESEMPIO

Per tre diverse aree si è considerato il consumo medio annuo di zucchero.

Aree	*Unità*	*Medie*
Zona_A	647	115
Zona_B	173	80
Zona_C	435	75
totale	1255	?

$$\overline{X} = \frac{647}{1255} * 115 + \frac{173}{1255} * 80 + \frac{435}{1255} * 75 = 96.31$$

Modalità raggruppate in classi

E' un caso particolare della proprietà associativa.

Purtroppo le medie aritmetiche di classe non sono note ed occorre stimarle.

La tecnica più usata è quella di adoperare il valore centrale delle classi:

$$\overline{X}_i = \frac{1}{2}L_i + \frac{1}{2}U_i \quad i = 1,2,\ldots,k$$

ESEMPIO

Casi di epatite A in un comune.

Le classi estreme hanno un'ampiezza pari alla metà delle ampiezza delle classi loro continue.

Età		Pazienti	c_i	f_i	Xif_i
≤9		662	6.75	0.5400	3.6448
10	19	420	14.50	0.3426	4.9674
20	29	117	24.50	0.0954	2.3381
30	39	18	34.50	0.0147	0.5065
40	49	5	44.50	0.0041	0.1815
≥50		4	52.25	0.0033	0.1705
		1226		1.0000	11.8087

Le medie di potenze

La media aritmetica rientra in una classe che fornisce una espressione dell'ordine di grandezza del fenomeno a partire da tutti i valori riscontrati:

$$M(X_1,\ldots,X_k;f_1,\ldots,f_k;\alpha) = \left\{\sum_{i=1}^{k} X_i^{\alpha} f_i\right\}^{1/\alpha}$$

Che scaturiscono dal principio di Chisini:

La misura di centralità deve lasciare invariato un particolare aspetto del fenomeno allorché al posto di ogni X_i si sostituisce "M":

Ad esempio, la media aritmetica posta in vece di ciascuna modalità lascia invariato l'ammontare complessivo delle rilevazioni

$$\sum_{i=1}^{k} \overline{X} n_i = \overline{X} \sum_{i=1}^{k} n_i = \frac{\sum_{i=1}^{k} X_i n_i}{n} * n = \sum_{i=1}^{k} X_i n_i = T$$

La media geometrica

E' una media particolarmente adatta come misura di centralità per fenomeni evolutivi che si realizzano proporzionalmente al livello già raggiunto.

Formula classica

$$Mg = \prod_{i=1}^{k} x_i^{fi} = x_1^{f1} * x_2^{f2} * \ldots * x_k^{fk} \quad \text{per } x_i > 0$$

Formula per il calcolo

$$Mg = e^{\left\{\sum_{i=1}^{k} fi\,Ln(xi)\right\}}$$

Esempio.

x_i	n_i	f_i	$Ln(x_i)$	$f_i Ln(x_i)$
1	2	2/10	0.0000	0.0000
3	1	1/10	1.0986	0.1099
4	1	1/10	1.3863	0.1386
5	1	1/10	1.6094	0.1609
6	1	1/10	1.7918	0.1792
7	1	1/10	1.9459	0.1946
9	3	3/10	2.1972	0.6592
	10			1.4424

$$Mg = e^{1.4424} = 4.2308$$

Proprietà della media geometrica

E' meno soggetta a variazioni rispetto alla loro media aritmetica:

X_i	2	4	8	16	32	64	128	256	512	1024		204.6
$Log(X_i)$	0.3010	0.6021	0.9031	1.2041	1.5051	1.8062	2.1072	2.4082	2.7093	3.0103	1.6557	45.2548

G=45.25 supera ed è superata dallo stesso numero di valori rispetto alla media aritmetica: μ=204 che ne supera sette ed è superata solo da tre.

La media armonica

Si tratta di un indice posizione da utilizzare soprattutto per misurare la centralità in situazioni in cui interessa il contributo delle modalità alla composizione di un tutto.

$$H = \frac{1}{\sum_{i=1}^{k} \frac{f_i}{X_i}}; \; per \; X_i \neq 0$$

Figli	Famiglie	f_i	$(1/X_i)f_i$
1	185	0.5362	0.5362
2	78	0.2261	0.1130
3	33	0.0957	0.0319
4	25	0.0725	0.0181
5	13	0.0377	0.0075
6	8	0.0232	0.0039
7	3	0.0087	0.0012
	345	1.0000	0.7119

Per calcolare la media armonica si calcola prima la media aritmetica di reciproci e di questa si calcola il reciproco:

$$H = \frac{1}{0.7119} = 1.4047$$

Valori remoti (outlier)

sono valori (uno o più) che appaiono inusuali senza che li si possa ritenere errati

Ciò dipende dal contesto:

{5, 13, 2, 291, 11, 6}

Non è un outlier né un errore tipografico: numero di clienti in coda ad uno sportello

{3.6, 2.7, 2.8, 3.9, 2.1, 85.8, 3.4, 2.8}

Peso alla nascita di un campione di neonati . E' anomalo se si tratta di persone

Valori remoti /2

Dato un insieme di modalità quantitative, una di esse sarà maggiore delle altre ed un'altra sarà minore.

Se gli estremi sono molto remoti nascerà il sospetto di disfunzioni:

Attenzione! Se un fenomeno può produrre modalità estremamente piccole e/o estremamente grandi è inevitabile che qualcuno si mostri prima o poi.

… E magari proprio nei vostri dati.

Una ASL ha storicamente richiesto il rimborso di un numero mensile di parti cesarei con complicazioni che oscilla tra i 20 ed i 30. Un dato mese, richiede rimborsi per 120.

Questo non necessariamente è un dato anomalo, ma è la spia di un cambiamento nel meccanismo dei rimborsi o nel management della ASL.

Valori anomali e medie

La presenza di valori isolati rispetto al resto della distribuzione ha una incidenza diversa a secondo della media che si usa

A	5	8	9	9	10	11	12	15	20
B	5	8	9	9	10	12	12	15	2013

valore isolato

***μ* passa da 11 a 210**

La media geometrica cresce, ma meno della media aritmetica e più della armonica.

La mediana, non cambia (al massimo si sposta di una posizione)

La moda non cambia (il valore isolato per definizione non può fare moda)

La media aritmetica ponderata

La media aritmetica è un caso speciale della media ponderata:

$$M(w_1, w_2, \ldots, w_k) = \sum_{i=1}^{k} w_i X_i; \;\; con \;\; w_i \geq 0; \;\; \sum_{i=1}^{k} w_i = 1$$

dove w_i è il "peso" con cui la X_i contribuisce alla media. M coincide con $\bar{X}$ se $w_i=f_i$.

Un campione di giovani è stato classificato per numero di domande inviato alle aziende fuori regione

Domande	Disoccupati	f_i	$1/f_i$	w_i	Xw_i
0	25	0.1667	6.0000	0.0239	0.0000
1	30	0.2000	5.0000	0.0199	0.0199
2	26	0.1733	5.7692	0.0230	0.0459
3	20	0.1333	7.5000	0.0298	0.0895
4	14	0.0933	10.7143	0.0426	0.1705
5	11	0.0733	13.6364	0.0543	0.2713
6	8	0.0533	18.7500	0.0746	0.4477
7	7	0.0467	21.4286	0.0853	0.5969
8	4	0.0267	37.5000	0.1492	1.1938
9	3	0.0200	50.0000	0.1990	1.7907
10	2	0.0133	75.0000	0.2984	2.9845
	150	1.0000	251.2985	1.0000	7.6108

La media aritmetica, calcolata in base alle frequenze relative è μ=2.86; nello schema vi è invece il calcolo in base ai pesi:

$$w_i = \frac{1/f_i}{\sum_{j=1}^{k} 1/f_j}$$

che inverte l'importanza delle modalità: quelle meno frequenti diventano più rilevanti e quelle più riscontrate vedono ridotto il loro contributo.

Le medie troncate

La "centralità" può risultare contaminata dalla presenza di valori troppo piccoli o troppo grandi.

la loro influenza può essere controllata escludendo dalle medie una certa frazione di unità.

L'esclusione può avvenire eliminando le unità in una delle due code o in entrambe.

Supponiamo di cancellare i valori inferiori o uguali al quantile $\gamma 2$ e superiori o uguali al quantile $\gamma 1$. La media aritmetica è:

$$M_{\gamma_1,\gamma_2} = \frac{\sum_{x_{\gamma_1} < x < x_{\gamma_1}} X_{(i)}}{n - [n\gamma_1] - [n\gamma_2]}$$

E' nota come media *trimmed* (potata)

Esempio

Un partito politico ha ottenuto in n=29 comuni i seguenti voti:

831	195	781	294	249	241	749	146	286	1445
1367	266	977	1668	1122	563	498	630	1164	1240
620	377	1516	240	724	300	1097	228	2213	

La media aritmetica per l'intera rilevazione è $M_{0,0}$=759.55.

Eliminiamo i valori inferiori al 1° decile ed al 19° ventile, cioè vincoliamo i voti nell'intervallo:

$$X_{0.1} < X < X_{0.95}$$

I quantili sono: $X_{0.1}$=X(3)=228, e $X_{0.95}$=X(28)=1668 e quindi si dovranno escludere dal calcolo X(1)=146,X(2)=195 e X(29)=2213

La media aritmetica troncata è: $M_{0.1,0.05}$=749.

Le medie winsorizzate

Invece di eliminarli, i valori anomali possono essere sostituiti con delle stime: medie winsorizzate.

Paese	Quota	Paese	Quota
Algeria	0.764	Libia	1.409
Gabon	0.293	Nigeria	1.857
Indonesia	1.374	Qatar	0.380
Iran	3.490	Arabia S.	8.395
Iraq	0.500	Emirati	2.260
Kuwait	1.500	Venezuela	2.360

La media aritmetica semplice: $\overline{X} = 2.049$ è ritenuta poco attendibile a causa di modalità abnormi negli estremi. Eliminiamo perciò il valore più piccolo e quello più grande per sostituirli con il primo e l'ultimo quartile rispettivamente: $X_{0.25}=0.5$, $X_{0.75}=2.36$. $M_w=1.563$

La tecnica di sostituzione e molto varia e può risultare arbitraria

Uso dei valori medi_1

I valori medi riescono a dare solo un'idea di massima della distribuzione

Il loro uso non può essere disgiunto dall'apporto informativo di altri indici descrittivi.

Il principio del Chisini è un importante ausilio per la scelta della media. C'è da notare però che le medie potenziate stanno tra di loro in una precisa relazione d'ordine

$$M_h \leq M_g \leq M_a \leq M_q$$

ed è difficile che conclusioni tratte sulla base di una siano poi sconvolte o alterate dall' uso di un'altra.

D'altra parte potrebbero mancare indicazioni su quale funzione dei dati osservati occorra preservare

Uso dei valori medi_2

Le medie troncate e perequate vanno usate con prudenza:

i valori anomali sono tali solo se si ha una conoscenza completa dello spettro dei valori riscontrabili.

Certi suoni a frequenza molto bassa o molto alta sono a noi "remoti"; questo però non significa che non esistono, anzi sono centrali per l'udito di altri esseri viventi.

D'altra parte, è accertata la sensibilità della media aritmetica ai valori grandi ed è perciò inadatta se tali manifestazioni sono marginali:

Uso dei valori medi_3

Un discorso a parte meritano la moda e la mediana (i singoli percentili hanno poco rilievo come indici di posizione).

Innanzitutto la prima può essere usata anche per carattere qualitativi e la seconda per caratteri ordinali. Hanno perciò un raggio d'azione più ampio delle medie potenziate.

La moda ha il difetto di non essere sempre calcolabile o sempre significativa;

La mediana è resistente ai valori anomali, ma come la moda non sfrutta tutte le informazioni rilevate.