FRANCESCO DELVECCHIO

STATISTICA PER LA RICERCA SOCIALE

CACUCCI EDITORE

FRANCESCO DELVECCHIO

STATISTICA per la ricerca sociale

CACUCCI EDITORE

INDICE

CONCETTI GENERALI INTRODUTTIVI

CAPITOLO I — **Le fasi della ricerca**

CAPITOLO II — **Fase di rilevazione**

CAPITOLO III — **Rappresentazione grafica**

CAPITOLO IV — **Elaborazione dei dati**

CAPITOLO V — **Cenni di calcolo delle probabilità**

CAPITOLO VI — **Alcune distribuzioni notevoli**

CAPITOLO VII — **Stime puntuali. Grado di fiducia delle stime**

CAPITOLO VIII — **La stima intervallare**

CAPITOLO XVI — **La misura delle relazioni fra più variabili**

CAPITOLO XVII — **Cenni sui modelli log-lineari**

CAPITOLO XVIII — **Dalla qualità alla quantità**

APPENDICE

PREFAZIONE

Il presente volume trae origine dal desiderio di fornire i principali elementi di Metodologia statistica a chi intenda intraprendere ricerche in campo sociale ed abbia una preparazione di Matematica solo a livello di scuola media superiore. Esso, perciò, può essere di ausilio non solo agli studenti universitari dei Corsi di laurea in Sociologia, Psicologia, Scienze Politiche e di altri Corsi ove sia contemplato l'insegnamento della Statistica e non quello della Matematica, ma anche agli operatori sociali che non abbiano una adeguata preparazione in Matematica.

Ovviamente, anche se nel testo non sono esposti i passaggi matematici per pervenire alla formulazione delle teorie statistiche, si è cercato di rendere ben chiari la logica posta alla base di detta formulazione, i requisiti a cui ogni test deve soddisfare per poter essere usato ed, infine, l'interpretazione dei risultati.

Poiché lo studioso dei fatti sociali molto spesso è chiamato ad operare con campionamenti stratificati, a due stadi ed a grappoli, si è ritenuto opportuno porre in luce come effettuare sia tali tipi di campionamento, sia le stime di alcune grandezze caratteristiche di grande uso nel campo sociale valutandone altresì il grado di precisione.

Naturalmente, grande importanza è stata data all'inferenza statistica, segnatamente ai piccoli campioni ed ai test non parametrici, a causa del frequente utilizzo di tali tecniche, specie nel campo delle scienze del comportamento.

Sempre per venire incontro alle esigenze dei ricercatori, il volume, nelle successive edizioni, è stato arricchito di argomenti che spesso in letteratura non sono trattati con applicazioni al campo sociale ma solo a livello strettamente metodologico: sono stati inseriti, così, la quantificazione delle mutabili, la tecnica della risposta casualizzata mediante domanda incorrelata, la distinzione tra intervallo di confidenza ottimale ed intervallo di confidenza centrato e le relative determinazioni, i test multinomiale ed ipergeometrico multivariato (per la prova d'ipotesi che valori osservati siano uguali a valori attesi nel campo di variabili discrete e di piccoli campioni), l'autocorrelazione dei residui, le *dummy variables*, i confronti ortogonali, la somiglianza di distribuzioni statistiche e di unità statistiche, ecc.. È stato anche riscritto il capitolo "La misura delle relazioni fra più variabili" e ciò non solo per una maggiore comprensione dell'argomento (ad esempio, la correlazione parziale

fra due variabili X_1 e X_2 è stata studiata come correlazione lineare fra i residui della regressione di X_1 da un insieme di variabili esplicative ed i residui della regressione di X_2 dallo stesso insieme di variabili esplicative), ma anche per precisare alcuni metodi di scelta delle variabili (*backward elimination, forward selection, stepwise regression*) che trovano frequenti applicazioni in campo sociale, anche per merito dei numerosi pacchetti di software reperibili sul mercato. In detto capitolo sono stati aggiunti, inoltre, rispetto alle edizioni precedenti, altri tre argomenti (l'individuazione di dati anomali, l'erronea specificazione del modello ed il modello moltiplicativo) per consentire allo studioso di fatti sociali la migliore formulazione delle ipotesi di relazioni fra variabili.

Infine, in questa edizione è stato soppresso il capitolo dell'analisi della covarianza, perché esposto in maniera più semplice (con l'ausilio, però, dell'algebra delle matrici) nel nostro volume *Analisi statistica di dati multidimensionali*, Cacucci Editore, Bari, 1992, a cui si rimanda il lettore desideroso di maggiori approfondimenti. Sono state invece inserite le nozioni di base relative ai modelli log-lineari, perché molto utili allo studioso di fatti sociali che voglia porre in luce relazioni fra caratteri qualitativi.

Forse la materia contenuta nel volume è troppo vasta per essere trattata in un corso annuale di lezioni: il docente, però, può non svolgere alcuni argomenti che ritenga meno utili per lo studente in un primo approccio con la disciplina, anche perché molti di essi risultano tra loro indipendenti. Gli argomenti trascurati in un primo momento possono costituire, però, oggetto di studio o in eventuali corsi successivi, o in discussioni seminariali o nella stesura di tesi di laurea.

Oltre tutto, a nostro parere, il volume può essere utilizzato soprattutto in ricerche inerenti al lavoro che lo studioso di fatti sociali deve svolgere. È noto, infatti, che non tutto ciò che serve all'esercizio della professione in campo sociale può essere svolto nei corsi universitari e che, una volta immesso in un'attività lavorativa, il neolaureato si trova a dover approfondire metodologie e tecniche per l'analisi dei fenomeni oggetto di studio: é proprio in quell'occasione che può far comodo avere sottomano un manuale ove siano riportate le principali tecniche, in modo che l'ulteriore approfondimento sia un prosieguo di ciò che ha imparato sui banchi universitari.

Per concludere, ringrazio tutti i colleghi che, in questa e nelle altre edizioni, mi hanno gentilmente fornito consigli e suggerimenti.

PREFAZIONE ALLA SETTIMA EDIZIONE

Le edizioni di questo volume, che si sono succedute nel tempo, hanno subito degli aggiornamenti allo scopo di rispondere sempre meglio alle esigenze della didattica e della ricerca in campo sociale. In particolare, in codesta edizione sono stati precisati e approfonditi taluni concetti, come, ad esempio, quelli concernenti il rapporto di durata, l'efficienza di uno stimatore e l'uguaglianza di due o più leggi di distribuzione.

Alcuni argomenti sono stati, invece, completamente riscritti. Fra questi: il metodo dei "confronti", quando è respinta l'ipotesi di uguaglianza di più di due medie, quelli di uguaglianza e somiglianza di distribuzioni e quello sull'intervallo di confidenza ottimale, introducendo, per quest'ultimo, un nuovo metodo che rende più attendibili i risultati. Sono stati, inoltre, generalizzati il "test esatto di Fisher" estendendolo a tabelle r×s, e quelli di Wilcoxon e Friedman, estendendoli al caso di "pareggi", che è quello che si presenta più frequentemente nella ricerca sociale.

CONCETTI GENERALI INTRODUTTIVI

1. - Fenomeni tipici e atipici. Fenomeni collettivi

Fenomeno (o fatto) è tutto ciò che può essere direttamente o indirettamente osservato.

Un fenomeno che può essere osservato direttamente è, ad es., una nascita, un decesso, una nevicata, uno sciopero, una denuncia all'Autorità Giudiziaria, il saper leggere, ecc..

Un fatto che può essere osservato indirettamente è, ad es., l'intelligenza di un individuo (questa, invero, appare dall'osservazione di certi aspetti del comportamento dell'individuo come la capacità di risolvere problemi, quiz, ecc.), oppure il tenore di vita di una popolazione.

Una prima distinzione è quella che differenzia i fenomeni in tipici e atipici.

Un *fenomeno tipico* è quello che si presenta sempre nello stesso modo, per cui l'osservazione di un solo caso fornisce le caratteristiche degli altri nelle medesime circostanze, ovvero l'effettuazione di un solo esperimento consente di conoscere i risultati di altri nelle stesse condizioni. Ad es., è sufficiente osservare una sola volta che l'acqua distillata in riva al mare bolle a 100 gradi C, per poter asserire che sotto le stesse condizioni ciò accade sempre.

Un *fenomeno atipico* è, invece, quello per cui una certa caratteristica è rilevata più o meno frequentemente, ma non sempre. Ad es., le professioni degli individui variano, in genere, da persona a persona; la durata di degenza ospedaliera varia, in genere, da paziente a paziente; i decessi che avvengono giornalmente nella città di Roma variano da giorno a giorno; ecc..

Un'altra classificazione è quella che distingue i *fenomeni in individuali e collettivi.*

Così, ad es., la nascita di un bambino è un fatto individuale, mentre la natalità di una popolazione è un fenomeno collettivo.

Quanto alle *cause* che agiscono sui fenomeni, queste possono essere semplici o complesse. In genere sono semplici le cause che agiscono sui fatti individuali: ad es., quelle che intervengono sulla nascita di un bambino; mentre sono *complesse* quelle che agiscono sui fenomeni collettivi: ad es., quelle che influenzano la natalità di una popolazione.

Se, poi, è nota l'azione di tutte le cause, allora è perfettamente noto anche il fenomeno: in tal caso basta, invero, una sola osservazione o un solo esperimento per poterlo descrivere: si tratta, quindi, di un fenomeno tipico. Ad es., sullo spazio s percorso nel vuoto da un grave in caduta libera influisce solo il quadrato del tempo t di caduta e l'accelerazione di gravità g: allora, nello stesso arco di tempo e nello stesso luogo (cioé con la stessa accelerazione g), tutti i corpi in caduta libera nel vuoto percorrono lo stesso spazio fornito dall'espressione

$$s = \frac{1}{2} g t^2.$$

Da quanto detto in precedenza appare evidente che i fenomeni oggetto di studio sono di natura diversa e investono campi diversi. In particolare i *fenomeni sociali* sono quelli che traggono origine (o sono avvertiti) dall'organismo sociale, cioé sono quelli relativi alla vita di una collettività in quanto influiscono sulle azioni degli individui di quella collettività.

La morte non è, ad es., un fatto sociale, ma un fatto individuale; mentre la mortalità è un fenomeno sociale in quanto, esprimendo la relazione esistente fra i decessi avvenuti in un dato periodo in una certa collettività e l'ammontare degli individui di quella collettività, influisce, ad es., sulla composizione per età della popolazione e, quindi, sulle forze di lavoro, sulla produzione, ecc.. Ed ancora, un individuo che si sposta non costituisce un fenomeno sociale, ma lo è, invece, l'emigrazione perché questa comporta la soluzione di problemi che riguardano tutti gli individui di quella collettività.

Comunque, di qualsiasi natura siano i fenomeni, essi vanno studiati scientificamente.

Studiare scientificamente un fenomeno significa ricondurlo a schemi in modo da poterlo misurare, descrivere per ricavarne, infine, delle regolarità e poterne anche prevedere lo andamento futuro. Naturalmente lo studio non si esaurisce solo nell'applicazione delle tecniche all'analisi dei fenomeni, ma comprende altresì l'interpretazione dei risultati ai quali si è pervenuti attraverso detta applicazione.

I primi fenomeni studiati scientificamente sono stati quelli fisici, però, con l'affinarsi della metodologia statistica, anche i fenomeni sociali possono essere studiati scientificamente: oggi, invero, si fa sempre più sentire l'esigenza di conoscere meglio la realtà sociale per far sì che si possano anche formulare razionalmente delle previsioni sullo sviluppo di talune manifestazioni della vita sociale.

2. - Sul concetto di Statistica

Studiando i fenomeni atipici su un gran numero di elementi (o effettuando un gran numero di esperimenti) ci si accorge che, in genere, essi manifestano delle "regolarità": considerando un grandissimo numero di dati si osserva, ad es., che all'incirca nascono 105 maschi ogni 100 femmine.

Da tutto ciò si deduce, allora, che i fenomeni atipici osservati su un gran numero di casi (o anche con un gran numero di esperimenti) possono essere considerati *collettivamente tipici* se presentano delle regolarità.

La *Statistica* è la disciplina che studia i fenomeni collettivamente tipici allo scopo di mettere in luce *regolarità* nascoste o dal modo con cui si sono manifestati i singoli casi, o dai risultati apparsi nei vari esperimenti.

Facciamo notare che la parola *"statistica"* deriva dal latino "status" perché all'inizio stava ad indicare la scienza che si occupa degli avvenimenti notevoli dello Stato. Oggi, invece, detta parola, usata al plurale (cioè "le statistiche"), sta ad indicare un insieme di dati numerici relativi a gruppi di persone, di animali, di cose o di fatti in senso lato; mentre, adoperata al singolare (cioè "la Statistica"), sta a significare l'insieme delle teorie e dei metodi (da cui anche il nome "Sta-

tistica metodologica") che permettono di studiare fenomeni collettivamente tipici.

Applicando, poi, i metodi statistici allo studio dei fenomeni dei vari campi, nascono le *statistiche applicate.*

La *Statistica sociale* è la scienza che studia i fenomeni sociali sotto l'aspetto quantitativo. Essa, attraverso il metodo induttivo [1], cerca di evidenziare le regolarità e tutte le possibili relazioni esistenti fra detti fenomeni. Ovviamente, tali regolarità hanno un valore limitato al tempo e al luogo a cui si riferiscono: i fenomeni sociali, invero, non solo sono diversi da luogo a luogo (ad es., in Italia la natalità è diversa per le varie regioni), bensì si evolvono e si trasformano nel tempo dando origine ad un processo di rinnovamento di tutta la società (si pensi, ad es., all'aumento della vita media alla nascita, alla diminuzione della mortalità e all'aumento della scolarizzazione che hanno agito sulla modificazione delle forze di lavoro, sulla disoccupazione, ecc.; oppure si pensi all'esodo dall'agricoltura verso l'industria e le attività terziarie).

L'esigenza di conoscere quantitativamente alcuni fenomeni sociali fu sentita, invero, fin dall'antichità, proprio perché la conoscenza di tali fenomeni serve per poter organizzare meglio l'amministrazione della società. Così, ad es., Mosè volle conoscere l'ammontare della popolazione israelita, Servio Tullio istituì il "Census" per accertare il numero di cittadini e l'ammontare dei loro beni, Carlo Magno istituì rilevazioni a carattere finanziario e amministrativo, ecc..

I fenomeni sociali che interessano l'amministrazione della società sono, però, innumerevoli; da ciò l'utilità di studiare autonomamente gruppi di essi: sono sorte così altre apposite discipline:

— la *Demografia* studia le popolazioni umane per conoscere, ad es., sia le caratteristiche ad un dato istante e sia gli aspetti di rinnovamento dovuti alle nascite, ai decessi, ai movimenti migratori, all'urbanizzazione, ecc.;

[1] Il *metodo induttivo* è un procedimento a posteriori, in quanto dallo esame dei fatti si risale alle leggi che quei fatti regolano. *Il metodo deduttivo*, invece, è un procedimento a priori, perché dai princìpi generali si passa ai fatti reali.

— *la Statistica sanitaria* studia i fenomeni cosiddetti sanitari (morbosità, mortalità, attrezzature e servizi sanitari, ecc.);
— la *Biometria* studia gli aspetti biologici degli organismi viventi;
— la *Statistica giudiziaria* studia i fenomeni che derivano dall'attività della magistratura (denunciati all'A.G., condannati, assolti, delitti, contravvenzioni, procedimenti esauriti, protesti, divorzi, ecc.);
— la *Statistica economica* studia i fatti cosiddetti economici (prezzi, consumi, produzione, risparmio, importazioni, esportazioni, reddito nazionale, ecc.);

ed altre ancora.

E' evidente, inoltre, che un fenomeno collettivo può riguardare anche più campi di studi, non sempre indipendenti gli uni dagli altri: i decessi, ad es., interessano sia la Demografia che la Statistica sanitaria, mentre l'utilizzo delle strutture ospedaliere concerne sia la Statistica sanitaria che la Statistica economica.

3. - Definizioni generali

Abbiamo già detto che la Statistica studia i fenomeni collettivi. Questi possono essere intesi come fenomeni che riguardano un collettivo statistico: occorre, perciò, definire cosa si intende per collettivo statistico.

Per *collettivo statistico* si intende un gruppo di elementi di natura qualsiasi, in modo che si possa stabilire se un elemento appartiene o no al gruppo. Ad es., è un collettivo statistico il gruppo di persone giudicate nel Distretto di Corte d'Appello di Bari nel 1985, sicché, se una persona è stata giudicata in epoca diversa o in un Distretto di Corte d'Appello diverso non fa parte del collettivo statistico definito. Un collettivo statistico è anche il gruppo di procedimenti di cognizione avvenuti in Italia nel 1985. Altri esempi di collettivi statistici possono essere forniti dall'insieme degli ospedali di una regione in una certa epoca, dai dimessi da un ospedale in un certo periodo di tempo, dai disoccupati nella regione

Puglia nel mese di luglio 1985, dai decessi avvenuti nella città di Roma in un certo giorno, ecc..

Il collettivo statistico si chiama anche *massa* o *popolazione* o *universo*.

Gli elementi singoli di cui è formato un collettivo statistico, diconsi *unità statistiche*.

Le unità statistiche possono essere *semplici* o *composte*.

Esempi di unità statistiche semplici sono: il singolo ammalato di un ospedale, la singola denuncia all'A.G., la singola abitazione, il singolo occupato in una azienda, ecc..

Una unità statistica composta è l'insieme di più unità semplici omogenee riguardo all'aspetto considerato. Un reparto di ospedale è un esempio di unità statistica composta: invero, tale reparto è costituito da più unità statistiche semplici (i degenti, i medici che vi lavorano, ecc.). Una famiglia di censimento è una unità statistica composta perché è costituita da più persone (che sono le unità semplici).

I collettivi statistici possono essere *finiti* o *infiniti*. Sono finiti quando è possibile togliere da essi uno dopo l'altro tutti gli elementi (ad es., il gruppo degli studenti di una scuola in un certo periodo). I collettivi che non sono finiti si dicono infiniti (ad es., l'universo delle prove di un lancio di moneta, oppure quello dei risultati ottenuti dalla ripetizione di un certo esperimento).

Su ogni unità statistica vengono rilevati diversi aspetti ciascuno dei quali è chiamato *carattere*.

Il *carattere* è dunque un qualunque attributo posseduto da una unità statistica. Esso può manifestarsi in varie maniere ognuna delle quali è detta *modalità*.

Nella cartella clinica di un ricoverato, ad es., sono rilevati vari caratteri: il sesso (carattere) con due modalità (maschio e femmina); il titolo di studio (carattere) con cinque modalità (privo di titolo, licenza elementare, licenza media, maturità, laurea); ecc..

I caratteri si distinguono in *qualitativi* e *quantitativi*.

I primi sono quelli per i quali le modalità sono esprimibili soltanto con forme verbali, mentre i secondi sono quelli per i quali le modalità sono esprimibili con dei numeri. Sono caratteri qualitativi il sesso, il titolo di studio, la religione,

il tipo di malattia, lo stato civile, la regione di residenza, ecc.. Sono caratteri quantitativi la statura, il peso, il perimetro toracico, la pressione arteriosa (massima o minima), il reddito prodotto, il numero di componenti una famiglia, la durata di degenza dei dimessi in un certo periodo da un ospedale, ecc..

I caratteri qualitativi possono, a loro volta, essere distinti in caratteri qualitativi *non ordinabili* e caratteri qualitativi *ordinabili:* i primi non ammettono un ordine naturale di successione, i secondi lo ammettono.

Caratteri qualitativi non ordinabili sono il sesso, la professione, la religione, il tipo di malattia, ecc..

Quando sulle unità statistiche si rilevano caratteri qualitativi non ordinabili, diremo che detti caratteri sono misurabili a *livello di scale nominali*.

Caratteri qualitativi ordinabili sono il titolo di studio (le cui modalità si possono ordinare in ordine crescente considerando come prima modalità "privo di titolo" e come ultima modalità "laurea"), le qualifiche professionali dei medici di una divisione ospedaliera (le modalità sono: assistente, aiuto, primario), ecc..

Quando sulle unità statistiche si rilevano caratteri qualitativi ordinabili, diremo che i medesimi sono *misurati a livello di scale ordinali*.

Il numero d'ordine che compete ad una certa modalità, cioé il posto che ad essa compete in una certa graduatoria, si chiama *rango*. Ad es., nella graduatoria crescente delle qualifiche attribuite al personale medico di un ospedale, il primario ha rango 1, l'aiuto il rango 2 e l'assistente il rango 3.

I caratteri quantitativi possono essere distinti in *discreti* e *continui*.

Sono *discreti* quando le modalità possono assumere solo alcuni valori (generalmente interi). Ad es., il numero dei posti letto di un ospedale, il numero di giornate di degenza dei dimessi da un ospedale in un certo periodo, il numero di componenti una famiglia, il numero di vani delle abitazioni, il numero di gravidanze di una donna, il numero di malattie avute nell'ultimo anno da un individuo, ecc..

I caratteri quantitativi sono *continui* quando le modalità possono assumere qualsiasi valore reale compreso fra il più

piccolo e il più grande. Ad es., sono caratteri continui la statura di un individuo, il peso, l'età, la temperatura corporea, ecc..

Dalle definizioni date e dagli esempi riportati appare evidente che, se vogliamo confrontare due modalità di uno stesso carattere qualitativo non ordinabile osservate su due unità statistiche, tale confronto può solo mettere in luce se esse sono uguali o diverse, ma non se una è maggiore o minore dell'altra. Ad es., se di due soggetti si rilevano le professioni, si può solo dire se essi esercitano la stessa professione oppure no, ma non si può dire se una professione è maggiore o minore di un'altra (invero, possiamo solo dire che la professione di avvocato è diversa da quella di professore di matematica, ma non quale delle due è maggiore).

Se vogliamo confrontare due modalità di uno stesso carattere qualitativo ordinabile, questa volta non solo possiamo dire se sono uguali o diverse, ma altresì se una è maggiore o minore dell'altra, anche se non si è in grado di calcolare di quanto. Ad es., se di due medici ospedalieri si rilevano le qualifiche, si può subito vedere se sono uguali e, nel caso siano diverse, si può stabilire anche chi ha la maggiore (infatti, l'assistente ha una qualifica minore dell'aiuto, mentre il primario ne ha una superiore), ma non si può dire di quanto sia la loro differenza. Da quando detto si deduce, inoltre, che se A, B, C sono tre modalità di uno stesso carattere qualitativo ordinabile, allora, se A è minore di B e B è minore di C, anche A è minore di C. Cioè, per i caratteri anzidetti è valida la proprietà transitiva della disuguaglianza.

Se vogliamo, invece, confrontare due modalità di uno stesso carattere quantitativo, questa volta si può dire sia se sono uguali, sia se una è maggiore dell'altra e sia di quanto. Ad es., se un individuo A ha una statura di 170 cm e un individuo B ha una statura di 180 cm, non solo possiamo dire che la statura di B è più elevata di quella di A, ma anche che B è più alto di A di cm 10.

I caratteri quantitativi, inoltre, godono della proprietà che se la differenza fra la grandezza A e la grandezza B è

uguale alla differenza fra la grandezza C e la grandezza D, anche mis(A) — mis(B) = mis(C) — mis(D). Questi caratteri si dicono misurati a *livello di scala ad intervallo*.

Ovviamente, per poter misurare con tale scala, è necessario fissare, per la misura delle grandezze, una unità di misura che possa essere utilizzabile in modo da ottenere sempre gli stessi risultati; inoltre alle differenze fra misure si possono applicare le regole dell'aritmetica.

Nelle scienze sociali le misure a livello di scale ad intervalli sono quelle migliori. Vi sono, però, livelli di misurazione superiori.

Ad es., supponiamo di misurare la temperatura in gradi centigradi (C°). Se la grandezza A ha una temperatura di 20 C° e la grandezza B ha una temperatura di 10 C°, non si può dire che la temperatura di A è doppia di quella di B: infatti, basta cambiare l'unità di misura (ad es., basta prendere come unità di misura il grado Fahrenheit) per rendersi conto che non è vero che la temperatura di A è doppia di quella di B. Se, invece, A ha un peso di 20 Kg e B un peso di 10 Kg, allora A ha un peso doppio di B in quanto, comunque si fissi l'unità di misura (ad es., grammi, once, ecc.) A ha sempre un peso doppio di B.

Se un livello di misurazione è tale che, oltre a soddisfare le proprietà delle scale ad intervallo, soddisfa anche la proprietà che, se A è k volte B, anche mis(A) = k mis(B), allora diremo che il livello di misurazione è di *scala di rapporti*. Solo se le grandezze si misurano a livello di scala di rapporti, si applicano alle misure le stesse regole dell'aritmetica e le conclusioni che si traggono sono valide anche in termini delle vere grandezze.

Si osservi che tali scale, oltre ad avere una unità di misura, hanno anche uno zero assoluto (sicché dire, ad es., che il peso di un corpo ha una misura di zero unità significa che ha massa zero).

Concludendo, si può affermare che i caratteri quantitativi sono quelli che forniscono più informazioni, mentre quelli qualitativi non ordinabili sono quelli che ne danno di meno.

4. - Le rilevazioni statistiche

Per lo studio dei fenomeni sociali occorre conoscere le caratteristiche delle unità statistiche di cui è costituito il collettivo. La conoscenza delle unità statistiche avviene per mezzo delle rilevazioni statistiche.

Le rilevazioni statistiche possono essere *totali* e *parziali*.

Sono *totali* quelle rilevazioni che riguardano tutte le unità del collettivo statistico. Sono *parziali* quelle rilevazioni che comprendono solo una parte di dette unità. Le rilevazioni parziali vengono dette anche *campionarie* e l'insieme parziale tratto dal collettivo viene anche chiamato *campione statistico*.

Le rilevazioni campionarie assumono notevole importanza perché consentono di avere (sotto certe condizioni che stabiliremo in seguito) informazioni attendibili sull'universo, con una riduzione di costi e di tempi occorrenti per la raccolta dei dati statistici: ad es., la spesa media annua per spettacoli di una famiglia di una città si può dedurre da un campione di un conveniente numero di famiglie di quella città; ed ancora, il professore assegna il voto ad un alunno formulando solo alcune domande che rappresentano un campione di tutte le domande che si possono formulare.

Definiremo *rappresentativo* un campione estratto con un criterio oggettivo: la rappresentatività è, perciò, una proprietà del modo con cui si formano i campioni e non dei gruppi di unità che si estraggono.

In genere è molto difficile stabilire a priori le caratteristiche delle unità statistiche da rilevare ai fini dell'indagine. Per questo motivo, per far sì che in sede di interpretazione dei risultati non manchi qualche informazione, tenuto conto anche del fatto che ci possono essere informazioni inutili ai fini dello studio del fenomeno (per cui la ricerca è appesantita dal tempo e dal costo maggiore per rilevare queste informazioni), occorre che detta ricerca sia adeguatamente pianificata.

E' questo l'oggetto del capitolo successivo.

CAPITOLO I

LE FASI DELLA RICERCA

Per evitare che la ricerca si estrinsechi in improvvisazione, è opportuno fissare, a priori, le varie tappe che si possono seguire. Si capisce che ciò che diremo è solo una guida per lo studente e per chi intenda effettuare una ricerca sociale.

La ricerca, di solito, comprende cinque fasi: una fase preliminare, una di programmazione, una di rilevazione, una di elaborazione ed una fase di interpretazione dei risultati.

1. - Fase preliminare

In questa fase della ricerca si seguono certi schemi e si realizzano certe tappe:

a) si cerca, innanzitutto, di definire il tema della ricerca ed i sottotemi.

Ciò è indispensabile in quanto non solo spesse volte accade che di mano in mano che si discute sul tema della ricerca occorre cambiare l'obiettivo originario, ma non è raro che si inizino ricerche che non vanno a buon fine perchè, in fase avanzata, ci si accorge che o non è definito bene l'oggetto della ricerca, per cui le risposte date dagli intervistati non sono confrontabili, oppure la massa di informazioni è talmente eterogenea che non si sa da dove iniziare. E' ovvio che occorre anche definire con chiarezza il "collettivo statistico" da rilevare o da stimare: per es., occorre decidere quali categorie di persone rilevare (cioè di quale città o di quale quartiere, di quale classe di età, di quale classe sociale, di quale settore di lavoro, ecc.); oppure, nel caso, ad es., si decida di fare una rilevazione sugli anziani, bisogna definire esattamente cosa si intende per anziano; se, invece, si vuole fare una rilevazione sugli enti assistenziali bisogna prima definire cosa s'intende per ente e, in seguito, fra tutti quelli che soddisfano alla definizione andare a vedere quali hanno le caratteristiche assistenziali; e così via.

Per dare un esempio di come si fissano i sottotemi, riportiamo ciò che hanno esposto, in una esercitazione, gli studenti del 2° anno della Scuola di Statistica dell'Università di Bari nell'a.a. 1975 - 76, a proposito di uno studio su "Gli studenti della Facoltà di Economia e Commercio dell'Università di Bari sotto il profilo socio-economico" (tema), il cui questionario è riportato in seguito.

I *sottotemi* che sono stati sviluppati in detta indagine sono:

— notizie riguardanti lo studente;
— notizie socio-economiche sullo studente e sulla famiglia;
— motivi che hanno spinto lo studente a scegliere la Facoltà di Economia e Commercio;
— alcune notizie di vita universitaria;
— lavoro a cui lo studente aspira dopo la laurea.

b) Si cerca di procurarsi materiale bibliografico sia in relazione a ricerche teoriche che a ricerche empiriche già effettuate ed inerenti al problema in esame e ciò per avere utili indicazioni sull'argomento e per poter stendere un primo documento generico sul tema.

c) Si cominciano a precisare le domande relative al tema e ai sottotemi. Se non si hanno sufficienti idee per esprimere queste domande, per definire il problema e per circoscrivere il tema, si possono seguire due strade:

— *ricognizione dell'ambiente*, nel senso che esperti dell'équipe stabiliscono contatti personali con alcuni elementi oggetto della rilevazione e ciò per ottenere nuove informazioni;
— *indagine pilota*, nel senso di effettuare una indagine su pochi elementi del collettivo con un questionario a domande aperte (a domande, cioè, nelle quali non si prevede una risposta fra quelle indicate nel questionario).

Una domanda aperta si può fare, ad es., chiedendo all'intervistato quale aspetto dell'argomento in oggetto piace di più e quale di meno. Dalle risposte ricevute si possono trarre utili informazioni per ridurre la gamma delle risposte ad una determinata domanda e poter, quindi, formulare il questionario finale da una parte senza il timore di aver trascurato aspetti essenziali per la conoscenza dell'oggetto e dall'altra senza avere

miriadi di informazioni che fanno perdere di vista il problema essenziale.

Sempre riferendoci alla esercitazione anzidetta, diamo un esempio di domande da porre in relazione al sottotema "*Notizie riguardanti lo studente*":

— stato civile;
— anno di corso;
— distanza del luogo di residenza da Bari;
— titolo di studio presentato per l'immatricolazione;
— frequenta abitualmente le lezioni;
— non frequenta le lezioni;
— viaggia;
— quante volte alla settimana viaggia;
— dimora abitualmente a Bari.

d) Si formulano le ipotesi che si vogliono verificare con la ricerca. Queste ipotesi non sono altro che le varie risposte che si possono dare ad ognuno dei quesiti posti in relazione ai temi e sottotemi della ricerca.

Ad esempio le ipotesi che si vogliono verificare relative al quesito "Perchè non frequenta le lezioni?" possono essere:

— perchè non lo ritiene opportuno;
— perchè lavora;
— perchè vive fuori sede e le condizioni economiche della famiglia non lo consentono;
— perchè il paese dove abita non è ben collegato con Bari.

E' evidente che uno degli scopi della ricerca è proprio quello di conoscere la composizione percentuale di queste risposte.

2. - Fase di programmazione

In questa fase l'équipe dei ricercatori dovrà stabilire:

a) *le modalità di rilevazione delle unità statistiche.*

Queste modalità sono diverse. Noi citeremo le più comuni:

— *Esame dei documenti.*

Questi documenti possono essere dati da atti di nascita, di matrimonio, di battesimo o di morte, come nel caso delle ri-

cerche di demografia storica, oppure, possono essere dati anche dalle statistiche ufficiali e non ufficiali. Nel caso, ad es., si debba fare una rilevazione sugli handicappati è evidente che bisogna rilevare le informazioni dalla scheda di ciascun handicappato, esistente presso l'Istituto assistenziale. Ovviamente questi documenti vanno vagliati con spirito critico per accertare preliminarmente fino a che punto siano attendibili.

— *Interviste.*

In verità questo punto andrebbe trattato dopo aver parlato del questionario, però è evidente che la formulazione del questionario viene ad essere condizionata anche dal modo di rilevazione delle modalità delle unità statistiche (se, ad es., l'intervista è telefonica, il questionario deve essere molto breve altrimenti si corre il rischio che l'intervistato interrompa la telefonata).

Le interviste possono essere *dirette, postali* e *telefoniche.* Ovviamente ognuno di questi tipi di intervista ha i suoi pregi e difetti.

La scelta di uno dei tre tipi di intervista dipende, s'intende, dalle caratteristiche principali a cui devono soddisfare le indagini, e cioè: il costo della rilevazione, il tempo massimo da impiegare (se l'inchiesta dura molto, rischia, quando è finita, di non essere più attuale) e il grado di precisione della stima che è, evidentemente, legato al problema dell'ampiezza del campione quando la rilevazione non è totale.

L'intervista diretta è quella che si fa con l'ausilio di un intervistatore. E' il tipo di intervista più usato in quanto dà più garanzie, essendo nota l'identità dell'intervistato ed essendo bassa la percentuale dei rifiuti. Ha l'inconveniente, però, di essere più costosa delle altre in quanto bisogna pagare l'intervistatore.

Essa, in linea di massima, può essere di due tipi: *libera* e *con questionario.*

L'intervista libera può essere a sua volta *focalizzata*, quando l'intervistatore (pur dialogando in modo libero) riesce ad ottenere dall'intervistato risposte precise a domande esplicite, oppure *globale* o *biografica* quando l'intervistatore chiede la opinione dell'intervistato su una determinata questione che interessa la ricerca e sollecita il racconto delle sue esperienze.

Essa consente di conoscere i moventi e persino il subcosciente dell'intervistato, per cui le risposte sono più attendibili.

L'*intervista con questionario* consiste nel fatto che l'intervistato risponde ai quesiti letti dall'intervistatore sul questionario.

L'intervista con questionario ha il pregio che le risposte si possono facilmente tabulare, mentre per tabulare le risposte in una intervista libera si corre il rischio di una cattiva interpretazione da parte di chi deve tradurre in numero le risposte date.

L'intervista postale ha due pregi essenziali: l'uno di avere un costo relativamente basso e l'altro che l'intervistato può rispondere con comodo a domande anche imbarazzanti conservando l'anonimato. Ha, però, dei difetti: il primo consiste nel fatto che son pochi quelli che rispediscono il questionario compilato, specie se questo questionario è molto lungo ed elaborato.

Naturalmente si sono approntate anche tecniche per avere dei "ritorni" più numerosi come, ad es., il sollecito oppure un premio per chi risponde. Ciò, però, ha l'inconveniente di aumentare il costo della rilevazione e di dover conoscere l'identità di chi risponde (per mandargli il premio o per sollecitare chi non risponde). E ciò, ovviamente, ridimensiona uno dei pregi fondamentali di questo tipo di intervista e cioè l'anonimato. C'è chi propone di aumentare l'ampiezza iniziale del campione per ottenere, alla fine (dopo, cioè, i mancati ritorni), un campione rappresentativo. Però in tal modo il campione sarebbe ugualmente *distorto*, nel senso che è stato dimostrato che a rispondere per prima sono gli appartenenti alle categorie sociali più istruite: sicchè nel campione così formato sarebbero poco rappresentati (o addirittura non rappresentati) gli appartenenti agli strati sociali meno colti.

Il secondo difetto potrebbe essere quello che il questionario sia compilato da altre persone: in molti di questi casi, però, le risposte rispecchiano ugualmente il livello sociale dell'intervistato; si aggiunga, inoltre, che il pregio dell'anonimato è di gran lunga superiore al difetto menzionato.

La difficoltà principale, invece, è quella di disporre degli indirizzi di tutto l'universo per poi, eventualmente, estrarre il campione. Qualcuno si serve delle liste elettorali, ma in tali liste non tutti sono iscritti, ad es., mancano i minori di 18 anni e quelli esclusi dalle liste per una qualsiasi ragione prevista

dalla legge elettorale (ad es., i condannati ad oltre 5 anni di pena). Ci sono anche le anagrafi comunali, però i dati dell'anagrafe sono riservati ed occorre un permesso speciale per rilevarli. Inoltre c'è da aggiungere che entrambe le liste sono formate dalle persone residenti: sicchè non sono compresi gli immigrati che conservano la residenza nel paese d'origine, mentre sono compresi gli emigrati che risultano ancora residenti.

L'intervista telefonica. Essa, come già detto, presuppone che il questionario sia molto breve ed ha il difetto di non rispettare molto la casualità della scelta degli elementi: infatti, non tutte le persone hanno il telefono. Se l'indagine è fatta nell'ambito cittadino è la più economica, ma diventa, invece, onerosa se le telefonate sono in teleselezione.

Le interviste per telefono sono usate dalle emittenti televisive per conoscere le opinioni nei riguardi di determinati problemi.

— Il metodo del sopralluogo.

Consiste nel mandare un osservatore nei luoghi oggetto di indagine per rilevare il comportamento dei soggetti.

— Il metodo del libretto o "panel".

Questo metodo si usa generalmente per conoscere giorno per giorno i tipi di consumo e le spese effettuate dalle famiglie a cui è stato consegnato il libretto. Questa indagine è di tipo continuativo e per questo vi è difficoltà a trovare famiglie disposte a collaborare. Si cerca di invogliare le famiglie offrendo dei premi a quelle che registrano e inviano libretti con tempestività: però, in tal modo, il campione è "distorto" in quanto si consegnano i libretti solo a quelle famiglie che accettano e non a quelle scelte casualmente.

b) *Tecniche di rilevazione delle unità statistiche.*

Qui bisogna stabilire se occorre rilevare tutte le unità di un collettivo oppure solo una parte (*campione*). Se si decide di fare una indagine campionaria bisogna, poi, fissare l'ampiezza del campione e il tipo di campionamento da usare.

Di questi problemi, però, ci occuperemo in apposito paragrafo. Tuttavia, anche se questi concetti saranno approfonditi in seguito dal punto di vista metodologico, è qui opportuno precisare come può avvenire la scelta delle unità da rilevare da parte dell'intervistatore.

La scelta delle unità può essere già fissata, in modo casuale, dal piano d'inchiesta: in tal caso saranno forniti all'intervistatore i nomi e gli indirizzi delle persone da intervistare, nomi che sono stati precedentemente estratti a sorte dagli elenchi contenenti tutti i nominativi (coi rispettivi indirizzi) del collettivo statistico. Ovviamente, in questo caso, se l'intervistatore non trova la persona cercata non può intervistare un'altra in sua vece.

E' possibile, però, che il piano di ricerca fissi, in modo casuale, il numero civico dello stabile (è il caso di quando non si hanno gli elenchi da cui estrarre i nominativi) e preveda anche che l'intervistatore interroghi, ad es., un componente della famiglia che abiti sul pianerottolo di sinistra del primo piano, uno che abiti sul pianerottolo di destra del secondo piano e così via. In questo caso è opportuno che l'intervistatore esegua alla lettera le istruzioni altrimenti il campione può risultare *distorto* (cioè gli elementi che entrano a far parte del campione non hanno tutti la stessa probabilità di essere scelti). A tal proposito Huguette Dautriat[2] osserva che in una indagine la maggior parte degli intervistati era formata da portinai ed artigiani proprio perchè gli intervistatori, per far prima e per evitare di salire le scale e forse anche a causa delle assenze, consegnavano il questionario agli abitanti del pian terreno.

Il campione può essere per *quota*: in tal caso si fissa una certa percentuale di persone fra gruppi aventi prefissate caratteristiche (ad es., in una indagine sul tipo di lettura è indispensabile conoscere i gusti di tutti gli strati sociali della popolazione, per cui, in un campione di 100 persone, si possono, ad es., intervistare 20 insegnanti, 20 impiegati, 5 ingegneri, 5 medici, 10 avvocati, 10 operai, 10 commercianti, 5 possidenti e 15 casalinghe). L'intervistatore è, quindi, libero di scegliersi i suoi soggetti purchè rispetti le quote. Va osservato, però, che anche in questo caso bisogna fare in modo che gli intervistati rappresentino tutti i quartieri della città e non solo, per comodità, il quartiere dove abita l'intervistatore.

---

[2] H. DAUTRIAT, *Il Questionario*, Franco Angeli Editore, Milano, 1966.

Questo tipo di campionamento si usa, generalmente, nelle indagini pilota per ottenere risposte a certi quesiti o per conoscere meglio l'ambiente su cui, poi, si deve fare la ricerca.

c) *Stesura del questionario.*

Per stendere il questionario in modo definitivo (o quasi definitivo) bisogna tener conto di tutto ciò che si è detto finora. Ovviamente anche per questo punto ci sono dei suggerimenti da dare, ma data l'importanza dell'argomento rimandiamo ad altro paragrafo.

d) *La verifica del questionario.*

Bisogna collaudare il questionario compilato per verificare la sua idoneità. Esso lo si distribuisce ad un campione di elementi, molto più piccolo di quello di cui bisogna poi servirsi per la ricerca, però, anche se questo campione non è rappresentativo come numerosità, deve rappresentare tutti gli strati del collettivo statistico oggetto di rilevazione (cioè, è un campione per quota).

La verifica del questionario, che avviene, in genere, tramite intervista, è importante perchè oltre a porre in luce le domande poco chiare, è utile ai fini della trasformazione di una domanda aperta in una strutturata ed anche ha lo scopo di ottenere una migliore sequenza logica delle domande stesse.

Solo dopo che questa verifica è stata positiva o solo dopo che si sono apportate le modifiche suggerite dall'indagine pilota, si potrà dire che il questionario è definitivo.

e) *Le norme per il rilevatore.*

Queste hanno lo scopo di chiarire le informazioni che devono essere raccolte sulle unità statistiche da rilevare e il modo come il rilevatore si deve comportare.

Anche questo argomento sarà trattato in apposito paragrafo.

f) *Le norme per informare l'intervistato.*

A questo punto bisogna chiarire, anche, in che modo informare l'intervistato che le risposte date sono riservate e l'indagine viene svolta per fini scientifici, e ciò per evitare che l'intervistato si rifiuti di rispondere o che risponda in modo non veritiero: usualmente si sceglie di inviare delle lettere alle

persone da intervistare, che illustrino lo scopo della ricerca e preannuncino la visita dell'intervistatore, oppure può l'intervistatore stesso presentarsi all'intervistato per illustrare brevemente le finalità della ricerca e l'ente che la organizza.

Naturalmente esistono anche altri metodi d'informazione quali, ad es., manifesti murali, giornali locali, radio o televisione (come accade, ad es., in occasione del censimento della popolazione).

g) *Previsione dei costi e dei tempi.*

Questo punto assume particolare rilievo perchè spesse volte le ricerche non si possono completare per l'esaurimento dei fondi, non avendo previsto in tempo il costo complessivo della indagine.

La previsione del costo dell'indagine deve riguardare: la spesa per il calcolatore che deve elaborare i risultati, la spesa per la tipografia, la spesa per gli intervistatori in relazione al numero medio giornaliero di intervistati, il costo dei mezzi di trasporto, il costo per preparare l'intervista, ecc..

3. - Il questionario

La difficoltà della preparazione del questionario è nella formulazione delle domande. Invero, dal modo come si pongono le domande dipende l'attendibilità delle risposte.

Per preparare il questionario il ricercatore, dopo aver consultato la bibliografia, deve servirsi innanzitutto della collaborazione dei responsabili di ciascuno stadio dell'inchiesta ed inoltre deve consultare gli schedari dell'impresa o dell'ente per conto del quale fa la ricerca: ad es., per fare un'indagine sulla provenienza sociale degli handicappati di un certo Centro assistenziale, bisogna che il ricercatore si consigli, soprattutto, con gli addetti al servizio sociale e medico di quel Centro e, prima di formulare le domande, occorre che consulti i documenti esistenti nell'archivio del Centro.

Nella formulazione delle domande si deve evitare che l'intervistato faccia sforzi di memoria. Ad es., in una indagine sull'alimentazione invece di chiedere quanta carne si è consumata

in un anno, è più opportuno chiedere quanta carne, in media, si consuma in una settimana.

Quando non si tratti di indagine in cui si voglia conservare l'anonimato, bisogna evitare di formulare domande imbarazzanti. Ad esempio, in un'indagine religiosa è stato chiesto: « Siete stato domenica scorsa a messa? ». Ha risposto di sì il 60% degli intervistati. E' stato ripetuto il sondaggio con gli stessi criteri ma cambiando domanda: « Indicate esattamente cosa avete fatto domenica scorsa », e solo il 40% degli intervistati ha detto di essere stato a messa[3].

E' opportuno che le domande siano formulate in modo semplice e con linguaggio elementare se si vogliono ottenere risultati precisi: ad es., se si fa un'inchiesta sulla limitazione delle nascite è sconsigliabile chiedere « lei usa contraccettivi? » perchè le donne con basso grado di istruzione difficilmente capiranno la domanda.

Analogamente se in un'indagine sociale sull'impiego del tempo libero si chiede « come passi la ricreazione? » si rischia di avere risposte disparate se non si precisa bene cosa vogliamo intendere per ricreazione, in quanto ognuno può avere un concetto diverso di ricreazione. Per il Grazia Resi[4], ad es., ricreazione è una attività (cioè è escluso l'ozio) non remunerativa, scelta spontaneamente con lo scopo di trarne un'intima soddisfazione.

Inoltre il numero delle domande influisce sulla preparazione del questionario. Infatti detto numero dipende sia dal luogo ove avviene l'intervista, sia dall'argomento dell'inchiesta e sia dal tipo di persone intervistate. Invero, se l'intervista avviene in luogo pubblico, il numero delle domande deve essere necessariamente esiguo. Ancora esiguo deve essere tale numero se l'argomento è difficile perchè, in tal caso, l'intervistato si stanca più facilmente. E' evidente che il numero delle domande dipende anche dalla classe sociale delle persone intervistate: infatti, se si intervista un professionista e l'argomento dell'inchiesta è interessante, il numero delle domande può essere anche consistente.

[3] G. TAGLIACARNE, *Tecnica e pratica delle ricerche di mercato*, Giuffrè Editore, Milano, 1960, pag. 64.

[4] B. GRAZIA RESI, *Considerazioni sulla Ricreazione Sociale*, Rivista internazionale di Scienze sociali, Vol. 26°, fascicolo VI, 1955.

Ovviamente se da una parte il questionario deve essere formato da poche domande, dall'altra c'è la necessità che tali domande siano sufficienti ad avere una visione esauriente del fenomeno sociale che si vuol studiare.

Nel formulare le domande bisogna tener conto anche dell'ordine. E cioè, si formulano prima le domande

— *di informazione e di opinione.*

A tal proposito si fa osservare che bisogna evitare, per quanto è possibile, che due domande che abbiano fra di loro un nesso causale siano troppo vicine: è opportuno, perciò, che l'intervistato, senza leggere prima il questionario, risponda alle domande secondo l'ordine con cui si susseguono;

— poi le *domande difficili.*

Si consiglia, per queste domande, di porle in modo da non urtare la suscettibilità dell'interessato e che l'intervistato non venga messo in soggezione. Ad es., non bisogna chiedere « Cosa pensi di.....? », ma « Alcuni pensano che....., altri il contrario; secondo lei chi ha ragione? ».

Spesso per avere notizie sullo stato economico-sociale dell'intervistato si vuol conoscere il reddito familiare. Ovviamente, l'intervistato non sempre è disposto a dichiarare il reddito per paura di essere colpito dal fisco. Si cerca perciò di avere una risposta per via indiretta. Il Luzzatto-Fegiz[5] per accertare il reddito, fra le altre domande ha posto le seguenti: « Secondo voi quanto dovrebbe spendere in media una famiglia composta come la vostra e della medesima condizione sociale per vivere in questa città (o paese) senza lussi, ma senza privarsi del necessario? ». E dopo aver posto altre domande, ne ha formulato altre due: « Dal lato puramente economico (guadagno, benessere materiale della vostra famiglia) siete contento o malcontento? ». « Se siete malcontento, sapreste dire quanto al mese vi occorrerebbe in aggiunta a ciò che guadagnate, per vivere

[5] P. LUZZATTO-FEGIZ, *Il volto sconosciuto dell'Italia*, Giuffré, Milano, 1956.

senza lussi, ma senza privarsi del necessario? »; ovviamente il reddito presunto, che si potrebbe tirar fuori dalle tre domande, è sempre approssimato, ma indubbiamente indicativo. Si consiglia, comunque, di alternare le domande difficili con quelle di informazione e di opinione;

— poi quelle *che riguardano la persona dell'intervistato.*

E' ovvio che queste domande si pongono alla fine perchè non è semplice, per l'intervistatore, cominciare il discorso con « che età ha », « quale professione esercita », ecc..

Si capisce che se per il seguito dell'indagine queste domande sono importanti, si possono anche mettere in principio;

— ed, infine, le *domande supplementari* che, anche se non hanno una utilità immediata, servono a suscitare interesse ed a far procedere agevolmente il colloquio.

Qualche volta, nel questionario, si inseriscono anche le *domande di controllo* per vedere l'attendibilità delle risposte date. Ad es., in un'indagine sulla lettura tra l'altro si può chiedere: « Quali di questi settimanali legge abitualmente? » e si fa seguire un elenco di settimanali A, B, C, D, ... Se l'intervistato risponde A, M, Z e uno di questi settimanali o non è mai esistito o non si pubblica più, è evidente che l'intervista non è attendibile.

E' d'uopo, ora, ricordare che esistono anche *domande filtro*, le quali servono a selezionare gli intervistati e a far, quindi, guadagnare tempo. Ad es., se si desidera fare un'indagine sulla lettura dei settimanali, si può porre come prima domanda: « Lei dedica parte del suo tempo libero alla lettura? ». Se la risposta è affermativa si prosegue, altrimenti l'intervista è conclusa. Come domanda successiva si può chiedere: « Preferisce la lettura di un quotidiano, di un settimanale o di romanzi? ». Se la risposta è « settimanali » si prosegue, altrimenti l'intervista è conclusa. La 3ª domanda, ora, può essere: « Che tipo di settimanale legge? ». Ecc..

Ed ora un'ultima avvertenza: è preferibile, quando è possibile, formulare *domande chiuse (strutturate)* anzichè domande *aperte*, perchè spesso non è facile codificare le risposte date alla domanda aperta. Tuttavia è da notare che, a volte, sono necessarie le domande aperte proprio perchè quelle chiuse potrebbero falsare la verità. E' il caso delle indagini di opinioni: invero, l'intervistato, di fronte ad un insieme di risposte relative ad una domanda, potrebbe segnarne una, attribuendosi, così, in buona fede, un'opinione, anche se in effetti prima dell'intervista non ne aveva alcuna.

Inoltre bisogna ricordarsi, quando si formulano domande strutturate, di inserire, fra le altre, la risposta « diversa dalle precedenti », in quanto è possibile che la risposta che vuol dare l'intervistato sia diversa da quelle formulate nel questionario.

Per concludere questi brevi cenni sulla preparazione del questionario diciamo che alla fine possono anche scriversi le annotazioni degli intervistatori specie in relazione al comportamento ed all'atteggiamento dell'intervistato in modo che si possa anche dare un giudizio sul grado di attendibilità delle risposte avute.

4. - Esempio di questionario

GLI STUDENTI DELLA FACOLTA' DI ECONOMIA E COMMERCIO DELL'UNIVERSITA' DI BARI

SOTTO IL PROFILO SOCIO - ECONOMICO

La presente indagine ha lo scopo di studiare i problemi socio-economici degli studenti universitari iscritti alla Facoltà di Economia e Commercio dell'Università di Bari.

La preghiamo, perciò, di collaborare, compilando questo modulo.

Si fa presente che le notizie fornite saranno elaborate statisticamente in forma anonima.

Sbarrare il quadratino ☐ in corrispondenza della risposta che si desidera dare, o scrivere sulla riga punteggiata.

Sezione A

1. - Cognome e nome .....

	giorno	mese	anno
2. - Data di nascita	.....	.....	19.....

3. - Sesso M ☐ F ☐

4. - Stato civile: celibe o nubile ☐, coniugato ☐, vedovo ☐, separato legalmente ☐, divorziato ☐.

5. - Anno di corso: I ☐, II ☐, III ☐, IV ☐, I fuori corso ☐, oltre il I fuori corso ☐.

6. - Luogo di residenza .....

7. - Distanza del luogo di residenza da Bari: Km. .....

8. - Anno accademico di immatricolazione presso la Facoltà: 19.....

9. - Titolo di studio presentato per l'immatricolazione: .....
.....

10. - Ha conseguito già altra laurea o diploma universitario: Sì ☐ No ☐.
Se sì, indicare quale .....

11. - Se proviene da altra Facoltà, indicare quale: .....
.....

12. - Frequenta abitualmente le lezioni? Se

Sì ☐

- Viaggia? Sì ☐ No ☐
- Se segue le lezioni viaggiando, con quali mezzi?
.....
- Quanto tempo al giorno impiega per venire a Bari e tornare a casa? .....
.....
- Quante volte alla settimana viaggia? .....
- Dimora abitualmente a Bari? Sì ☐ No ☐
Se sì, dove?
- presso parenti ☐
- in pensione ☐
- presso la Casa dello Studente ☐
- in appartamento libero ☐

Se dimora a Bari, quanto spende, in media, mensilmente per il soggiorno?
.....

No ☐

Se non segue le lezioni, perchè?

Perché non lo ritiene opportuno ☐
Perché lavora ☐
Perché vive fuori sede e le condizioni economiche della famiglia non lo consentono ☐
Perché il paese dove abita non è ben collegato con Bari ☐
Per altri motivi ☐
Quali altri motivi? .....
.....
.....
.....
.....
.....
.....
.....
.....

Sezione B

13. - Attualmente lavora, anche saltuariamente? Sì ☐ No ☐

14. - Se lavora, qual è la sua posizione in questo lavoro?
— Datore di lavoro (imprenditore, commerciante al minuto e all'ingrosso, proprietario fondiario non conduttore, esercente di bar, ecc.) ☐
— Libero professionista (medico, avvocato, farmacista, ecc.) ☐
— Dirigente ed alto funzionario (compresi gli ufficiali di grado superiore) ☐
— Impiegato (compresi gli ufficiali fino a capitano, i sottufficiali, gli insegnanti, ecc.) ☐
— Lavoratore in proprio ☐
— Lavoratore dipendente ☐
— Lavoro con i familiari ☐

15. - Se lavora, qual è il ramo di attività economica in cui lavora?
agricoltura ☐, industria ☐, commercio ☐, Pubblica Amministrazione ☐, altri (indicare quale) .....

16. - Secondo lei, quanto dovrebbe guadagnare una famiglia della stessa condizione sociale della sua e con lo stesso numero di componenti, per vivere nel suo paese senza lussi, ma senza privarsi del necessario? .....

17. - Quanti sono i componenti del suo nucleo familiare?
Padre ☐, Madre ☐, Figli ☐, (indicare quanti), Altri ☐ (indicare quanti)

18. - Suo padre è pensionato? Sì ☐ No ☐
Suo padre è disoccupato? Sì ☐ No ☐

19. - Sua madre è pensionata? Sì ☐ No ☐
Sua madre è disoccupata? Sì ☐ No ☐

20. - Qual è la posizione dei suoi genitori nel lavoro che svolgono? (Se pensionati o disoccupati indicare la posizione del lavoro che svolgevano prima)

	Padre	Madre
— Datore di lavoro	☐	☐
— Libero professionista	☐	☐

— Dirigente ed alto funzionario ☐ ☐
— Impiegato ☐ ☐
— Lavoratore in proprio ☐ ☐
— Lavoratore dipendente ☐ ☐
— Senza professione (casalinghe, detenuti, mendicanti, ecc.) ☐ ☐

21. - In quale ramo di attività economica lavorano i suoi genitori?

	Padre	Madre
— Agricoltura	☐	☐
— Industria	☐	☐
— Commercio	☐	☐
— Pubblica Amministrazione	☐	☐
— Altri	☐	☐

22. - I suoi fratelli e sorelle lavorano? Sì ☐ No ☐
In quanti? .....

23. - Dal lato puramente economico è contento del tenore di vita che la sua famiglia può condurre? Sì ☐ No ☐

24. - Se non è contento sa dire quanto al mese servirebbe alla sua famiglia, in aggiunta a ciò che guadagna, per vivere senza lussi ma senza privarsi del necessario? .....

Sezione C

25. - Per quale motivo si è iscritto alla Facoltà di Economia e Commercio?
— Per attitudine agli studi economici ☐
— Perché spera in maggiori possibilità di lavoro ☐
— Perché già avviato in attività economiche ☐
— Perché è stata consigliata questa Facoltà in quanto più rispondente alle esigenze dell'attuale società ☐
— Per il tipo di studi effettuati nella scuola media superiore ☐
— Perchè suo padre è laureato in Econom. e Commercio ☐
— Perché in famiglia c'è un'attività economica ove c'è bisogno di un laureato in Economia e Commercio ☐
— Perchè ritiene la laurea in Economia e Commercio più facile delle altre lauree ☐
— Per altro motivo ☐ quale? .....

26. - Ritiene di essere stato influenzato nella scelta della Facoltà?

Sì ☐
Se fosse stato libero quale facoltà avrebbe scelto? .....
.....
Perché? .....
.....

No ☐
In base all'esperienza vissuta in Facoltà, ritiene di essere contento della scelta fatta?
Sì ☐ No ☐

Sezione D

27. - Usufruisce del presalario? Sì ☐ No ☐
28. - Ne ha fatto domanda? Sì ☐ No ☐
29. - L'anno scorso ha usufruito di presalario? Sì ☐ No ☐
30. - Nel precedente anno accademico, quali corsi ha frequentato di più? .....
31. - Ha incontrato difficoltà per frequentare regolarmente le lezioni? Sì ☐ No ☐
Se sì, quali? .....
.....
32. - Ha trovato difficoltà nei programmi di studi? Sì ☐ No ☐
Se sì, quali? .....
.....

Sezione E

33. - Crede che la laurea in Economia e Commercio dia oggi ampia possibilità di inserimento nel mondo del lavoro?
Sì ☐ No ☐
34. - A quale lavoro aspira dopo la laurea?
Insegnamento ☐ Impiego in amministrazioni pubbliche ☐ Impiego in aziende industriali e commerciali ☐
Libera professione ☐ Condurre la propria azienda ☐
Condurre l'azienda familiare ☐
A quale altro? .....
35. - Ritiene utile, dopo la laurea, iscriversi ad un corso di specializzazione? Sì ☐ No ☐
Se sì, quale? .....
.....

5. - Norme per gli intervistatori

Diamo alcune norme di carattere generale per gli intervistatori. E' evidente che per ogni indagine si impartiranno, poi, anche delle istruzioni particolari per condurre bene le interviste.

L'intervistatore deve esattamente interrogare le persone indicate sul piano di lavoro (a meno che non si tratti di campionamento per quote) e, se non riesce ad intervistarne qualcuna, deve effettuare almeno tre visite in ore e giorni diversi prima di desistere. Naturalmente non può rimpiazzare chi non ha interrogato, ma deve ritornare il questionario non compilato, a meno che non gli sia stato fornito, dal piano di lavoro, un elenco complementare da cui estrarre un altro nominativo.

Prima di iniziare ad intervistare, l'intervistatore deve conoscere *molto bene* il contenuto del questionario per poter essere in grado di chiarire tutti i dubbi prima dell'inizio dell'intervista.

L'intervistatore, però, *non deve* assolutamente influenzare le risposte dell'intervistato e deve evitare qualsiasi commento; al più, se qualche domanda non è stata capita dall'intervistato, deve rileggerla esattamente nell'ordine con cui è scritta nel modulo d'intervista.

Se l'intervistato, incuriosito, chiede informazioni supplementari, l'intervistatore deve dare chiarimenti solo alla fine dell'intervista e nell'ambito del rispetto del segreto professionale.

Se l'intervista è libera, l'intervistatore deve riportare esattamente i commenti dell'intervistato.

Inoltre, quando l'intervista è libera, l'intervistatore deve evitare di interrogare individui che sono in attesa, ad es., di un autobus perchè si rischia di non portare a termine l'intervista.

Se l'intervista deve essere effettuata in una Fiera o in un locale pubblico, l'intervistatore non deve intervistare il personale o i vigili, bensì le persone che escono dalla Fiera o dal locale pubblico: si presume, invero, che solo codeste persone abbiano visto tutto e siano in grado di dare risposte veritiere.

Non bisogna intervistare una massaia prima delle ore 9,30 o nell'ora di pranzo, né bisogna disturbare un commerciante nell'ora di punta bensì la mattina all'apertura dell'esercizio quando i clienti sono ancora pochi.

Non bisogna mai intervistare persone che siano state presenti ad interviste fatte ad altri.

L'intervistatore, quando la sua visita non sia stata preannunciata, deve avvicinare le persone nel modo più semplice per non destare diffidenza e brevemente deve rassicurare lo intervistato sulla serietà e sullo scopo della ricerca. Per es., può presentarsi così: « Buon giorno. La signora Rossi è in casa? ». Se chi ha aperto la porta dice che la signora Rossi è lei, allora l'intervistatore può aggiungere: « Sono un intervistatore dello Istituto di ricerche sociali di ..... Non ho niente da venderle e posso anche evitare di entrare. Desidero solo alcune informazioni su ..... Le sarei grato se volesse, a tal proposito, esprimere le sue idee e le prometto di non farle perdere molto tempo. Se lei permette la prima domanda è questa ..... ». Nel caso la signora Rossi non possa riceverlo, l'intervistatore deve chiedere quando può tornare. E' importante, però, che intervisti solo la signora Rossi e non altre persone della famiglia.

A volte delle signore rispondono di non avere esperienza e invitano l'intervistatore a ritornare quando c'è il marito. In tal caso l'intervistatore deve far osservare alla signora che, ai fini della ricerca, interessa conoscere proprio il suo punto di vista.

Se poi l'intervistato si rifiuta di rispondere l'intervistatore deve fargli capire che, rifiutandosi, non gli consente di svolgere bene il suo lavoro.

E' evidente che altri suggerimenti che si danno sulle norme particolari per gli intervistatori sono tratti dall'esperienza fatta nell'indagine pilota e nella ricognizione sull'ambiente.

Alla fine dell'intervista bisogna ricordarsi di ringraziare lo intervistato per l'aiuto dato, lasciando in lui la sensazione che ha compiuto una cosa importante e che le sue risposte sono state interessanti.

Da quanto detto appare evidente che l'intervistatore deve avere diversi requisiti: una buona cultura e modi distinti che destino fiducia e simpatia nell'intervistato, scioltezza nel lin-

guaggio e spiccato spirito di osservazione ma, soprattutto, onestà ed imparzialità.

L'intervistatore, poi, deve anche vestire con eleganza perchè la prima impressione che desta è data proprio dall'aspetto esteriore. Molti sono d'accordo nel ritenere che le donne sono molto indicate per questo tipo di lavoro, specie se insegnanti o studentesse universitarie.

Diamo, qui di seguito, come esempio, le *norme per l'intervistatore* fissate dall'Istat per la rilevazione delle forze di lavoro in Italia.

« Nel presente modello l'intervistatore deve indicare le notizie riguardanti *le famiglie che non hanno pututo intervistare.*

I casi di impossibilità ad eseguire l'intervista possono essere i seguenti:

1. - *RIFIUTO.* - Nel caso che la persona indicata nell'elenco o un suo familiare, si rifiuti di rispondere e di fornire le notizie da trascrivere sul modello (ISTAT/P/50), l'intervistatore farà opera di persuasione assicurandola della riservatezza delle notizie raccolte e sottolineando quanto in proposito è detto nel tesserino di riconoscimento rilasciato dal Comune.

Se la persona persiste nel rifiuto, prenderà nota della mancata intervista nel presente modello.

2. - *IRREPERIBILITA'.* - Può darsi che l'intervistatore non trovi la persona indicata sull'elenco e gli venga detto che è sconosciuta.

In tal caso deve assicurarsi che non si tratti di alterazione di cognome dovuto ad errore di trascrizione (es., Tonti invece di Tenti) chiedendo il cognome della famiglia che è nell'abitazione o, se necessario, eseguendo il controllo del nominativo presso l'ufficio del Comune.

Può anche darsi, se il capo famiglia è donna coniugata o vedova, che nell'elenco sia indicato il cognome da nubile, spesso sconosciuto al portiere o agli stessi vicini di casa; basterà chiedere il cognome da coniugata per chiarire l'equivoco. Talvolta può trattarsi di alterazione del nome della via o piazza (Piazza Cavour invece di via Cavour, situata in altra zona della città) o di numero errato.

L'intervistatore, ove non riesca ad individuare il nominativo errato, annoterà la mancata intervista nel presente modello.

3. - *MORTE.* - Nel caso che la persona indicata nell'elenco sia l'unico componente della famiglia e risulti deceduto in epoca precedente il giorno di riferimento, l'intervistatore ne prenderà nota nel presente modello.

4. - *MOMENTANEA ASSENZA.* - Se la persona che si cerca è in viaggio per affari o per diporto o è fuori casa per commissioni, si intervisterà la *persona di famiglia* che ne fa le veci o quella che sia in grado di rispondere alle domande che le saranno rivolte. Nel caso che la famiglia sia completamente assente chiederà notizie alle famiglie vicine o al portiere. Occorrendo lascerà, nella cassetta delle lettere o sotto la porta, una cartolina avviso con l'indicazione dell'ora e del giorno in cui sarà ripetuta la visita.

Nel caso di assenza che si verifichi per tutta la settimana di rilevazione, l'intervistatore prenderà nota della mancata intervista nel presente modello.

5. - *DOMICILIATA DI FATTO IN ALTRO COMUNE.* - Nel caso di famiglia che risulti domiciliata in altro Comune dove ha preso dimora abituale, l'intervistatore prenderà nota della mancata intervista nel presente modello.

6. - *EMIGRATA ALL'ESTERO.* - Anche nel caso di famiglia completamente emigrata all'estero l'intervistatore prenderà nota della mancata intervista nel presente modello.

L'intervistatore dovrà tenere presente anche i seguenti casi e accorgimenti:

FAMIGLIA ABITANTE IN ALTRA CASA. - Dovrà recarsi al nuovo indirizzo se l'abitazione è nella sua area di rilevazione; in caso diverso, dovrà prenderne nota nel presente modello e riferirne all'Ufficio Comunale *che vi provvederà con l'intervistatore dell'altra area di rilevazione* .

NOMINATIVO INDICATO NELL'ELENCO CHE NON FA' PIU' PARTE DELLA FAMIGLIA. - Considererà la famiglia nella *composizione in cui si trova all'indirizzo segnato* e interrogherà la *persona di famiglia* che ha assunto le funzioni di capo di famiglia o altra che sia in grado di rispondere alle domande.

ACCORGIMENTI DA SEGUIRE NELL' INTERVISTA. - Quando si presenta al domicilio della famiglia da intervistare,

avrà cura di declinare subito la sua identità e la sua qualifica di incaricato del Comune. Quindi, dopo essersi assicurato di trovarsi nel domicilio della famiglia indicata nell'elenco e che la persona con la quale parla è il capo famiglia (in caso diverso chiederà di lui, o, in sua assenza, di un altro membro della famiglia capace di rispondere alle domande), esporrà in modo succinto lo scopo della visita. Se la famiglia è stata intervistata altre volte, ricorderà che i buoni risultati della rilevazione hanno messo in evidenza l'utilità di rinnovare l'indagine per accertare le variazioni complessive che sono nel frattempo intervenute.

Durante l'intervista avrà cura:

a) di porre le domande con la massima cortesia e di formularle in modo da non esercitare alcuna influenza sulla risposta;

b) di non intervistare mai in presenza di estranei;

c) di non mostrare all'intervistato i questionari già riempiti riguardanti altre famiglie;

d) di non mostrare sorpresa alle risposte date e, meno ancora, fare commenti su particolari condizioni di qualche componente la famiglia o su altra circostanza che risultasse dall'intervista;

e) di non suscitare o alimentare conversazioni o discussioni di carattere politico ».

6. - Tecniche di campionamento

Abbiamo già visto, nel paragrafo 2, che una rilevazione campionaria può essere effettuata in vari modi a seconda di come si estraggono gli elementi dalla popolazione. Si tratta, qui, di chiarire meglio in che cosa consistono queste tecniche di campionamento.

a) Campione casuale semplice (o bernoulliano).

Per formare questi campioni è necessario conoscere gli elenchi (*base* o *frame*, secondo la denominazione anglosassone) di tutte le unità statistiche dell'universo.

Per determinare un campione casuale bisogna far in mo-

do che nell'estrazione delle unità dalla popolazione, le medesime abbiano tutte la stessa probabilità di far parte del campione.

Scelta casuale non significa, però, scelta senza regola, affidata alla libera iniziativa dell'intervistatore: in tal caso, invero, le unità da rilevare non avrebbero tutte la stessa probabilità di far parte del campione, visto che l'intervistatore potrebbe essere portato (anche in buona fede) a scegliere alcune unità e non altre solo perché più facilmente rilevabili.

Per formare un *campione casuale semplice* (o *bernoulliano* o *con ripetizione*) di *n* elementi basta numerare tutti gli N elementi dell'universo ed estrarre da un'urna, dove sono stati imbussolati tutti i numeri corrispondenti a quegli elementi, n numeri con l'accortezza, però, di rimettere di volta in volta — nell'urna — il numero estratto.

Il numero n di elementi che fanno parte del campione si chiama *ampiezza* o *dimensione del campione*, mentre il rapporto n/N prende il nome di *frazione di campionamento*.

L'insieme di tutti i possibili campioni di ampiezza n che si possono estrarre da una popolazione costituisce *l'universo dei campioni di dimensione n.*

Ad es., data la popolazione di N = 4 elementi (a), (b), (c), (d),

l'universo dei campioni bernoulliani di ampiezza n = 1, è costituito dagli stessi elementi (il numero di campioni di tale universo è, quindi, N^1). L'universo dei campioni di ampiezza n = 2 è formato dalle coppie di elementi che si ottengono scrivendo a fianco di ciascun campione di ampiezza 1 gli elementi stessi: infatti, se nella prima estrazione è venuto fuori l'elemento a, nella seconda estrazione può venire fuori ancora l'elemento a (visto che tale elemento è stato reinserito nell'urna), oppure uno qualsiasi degli altri elementi. L'universo dei campioni bernoulliani di ampiezza n = 2 è, allora,

(aa), (ab), (ac), (ad), (ba), (bb), (bc), (bd), (ca), (cb), (cc), (cd), (da), (db), (dc), (dd).

Tale universo è costituito, dunque, da N^2 campioni.

L'universo dei campioni bernoulliani di ampiezza n = 3 si ottiene scrivendo a fianco di ciascun campione di dimensione

2, gli elementi stessi: ad es., dalla coppia campionaria (aa) si ottengono i campioni

(aaa), (aab), (aac), (aad);

analogamente, dalla coppia (ab) si ottengono i campioni

(aba), (abb), (abc), (abd);

e così via.

Tutti i campioni d'ampiezza $n = 3$ sono, dunque, N^3.

In generale l'universo dei campioni, d'ampiezza n, estratti bernoullianamente è costituito da N^n campioni.

A questo punto è opportuno precisare che i risultati ricavati da una rilevazione parziale non coincidono, in genere, con quelli ricavati da una rilevazione totale: per tale motivo diciamo che una caratteristica ricavata con i dati di un campione è una *stima* più o meno precisa della stessa caratteristica ricavata con i dati della popolazione.

In seguito chiariremo il concetto di precisione di una stima.

Si dimostra, inoltre, che quando il campione è scelto in modo casuale la precisione della stima aumenta con l'aumentare della dimensione del campione.

Il ricorso al campionamento per avere informazioni su una caratteristica della popolazione comporta, quindi, in genere, un errore detto *errore di campionamento* per distinguerlo dagli *errori di osservazione e di rilevazione* (presenti anche nelle rilevazioni totali) e dall'*errore sistematico* che conduce a stime che si discostano dalla vera caratteristica dell'universo sempre nello stesso senso (si pensi, ad es., ad una bilancia non ben tarata le cui misure sono sempre sovrastime o sottostime del peso reale di qualsiasi corpo).

E' logico che l'errore totale di una rilevazione campionaria è la somma dell'errore di campionamento e di tutti gli altri tipi di errori.

b) Campione casuale senza ripetizione e campione casuale in blocco

Qualche volta, però, la scelta con ripetizione risulta non agevole o addirittura impossibile (è il caso, ad es., di quan-

do si vuol controllare, tramite un campione, la durata di una popolazione di lampadine). Molto spesso si ricorre, perciò, ad un *campione casuale senza ripetizione* (o *senza reinserimento*).

In questo tipo di campionamento, se un elemento è stato estratto per far parte del campione, tale elemento viene escluso dalle successive estrazioni.

Facciamo osservare che se N è finito, nello schema bernoulliano può essere anche $n > N$, mentre nello schema senza ripetizione è sempre $n < N$.

Vediamo, ora, com'è costituito l'universo dei campioni nel caso di estrazione senza ripetizione, riferendoci all'esempio precedente.

L'universo dei campioni senza ripetizione di ampiezza 1 è formato, ovviamente, dagli elementi stessi (quindi il numero di campioni è N).

L'universo dei campioni senza ripetizione, di ampiezza $n = 2$, si ottiene scrivendo a fianco di ciascun campione di ampiezza 1, gli elementi rimasti: infatti, se nella prima estrazione è stato estratto, ad es., l'elemento a, questo non può essere estratto nella seconda estrazione: per cui le coppie che si possono formare sono

(ab), (ac), (ad).

L'universo dei campioni senza ripetizione, d'ampiezza $n = 2$, è, perciò,

(ab), (ac), (ad), (ba), (bc), (bd), (ca), (cb), (cd), (da), (db), (dc).

Tale universo è costituito dunque da $N(N-1)$ campioni.

Se si desidera scrivere l'universo dei campioni senza ripetizione, ognuno di dimensioni $n = 3$, basta scrivere a fianco di ciascun campione d'ampiezza 2 gli elementi rimasti: infatti, se nella prima e nella seconda estrazione sono stati estratti, ad es., gli elementi (bc), questi non possono uscire nella terza estrazione: per cui le terne che si possono formare sono (bca), (bcd). Il numero di campioni, di dimensione $n = 3$, estratti senza ripetizione è $N(N-1)(N-2)$. Ecc..

E' evidente che i campioni estratti senza ripetizione sono campioni ordinati nel senso che, ad es., il campione (ab) è

diverso dal campione (ba): invero, per il campione (ab) è stato estratto prima a e poi b, mentre per il campione (ba) è stato estratto prima b e poi a.

Se gli elementi si estraggono, invece, in *blocco* non esiste l'ordine di estrazione.

In tal caso, ordinati gli elementi, per avere l'universo dei campioni in blocco di dimensione 2 basta scrivere a fianco di ciascun elemento gli elementi che seguono. Cioé

(ab), (ac), (ad), (bc), (bd), (cd).

Per formare l'universo dei campioni in blocco di dimensione 3, basta scrivere a fianco di ciascun campione di dimensione 2 gli elementi che, nell'ordine, seguono l'ultimo elemento della coppia. Si ha, allora,

(abc), (abd), (acd), (bcd).

Analogamente si procede per $n > 3$.

c) *Tavole di numeri aleatori*

E' evidente che se la popolazione è formata da molti elementi il procedimento dell'estrazione dall'urna è molto scomodo, perciò si ricorre ad altri metodi.

Un altro metodo per scegliere le unità in modo casuale consiste nel servirsi delle tavole aleatorie. Queste tavole sono formate da tante colonne di numeri fra 0 e 9, scelti a caso: scelti, ad es., estraendoli da un'urna contenente i 10 numeri 0, 1, 2, ..., 9.

Una volta numerate tutte le unità statistiche dell'universo, vediamo come estrarre le unità che devono far parte del campione. Supponiamo, ad es., che di tutte le 584 famiglie iscritte nello schedario anagrafico delle famiglie di un certo Comune se ne vogliono estrarre 50. Poiché 584 è un numero di tre cifre, prendiamo 3 colonne nelle tavole di numeri aleatori: partendo da un numero qualsiasi (che, essendo 3 le colonne, è formato da 3 cifre), scegliamo i primi 50 numeri non superiori a 584 anche se non distinti (*campione bernoulliano*); se, invece, scegliamo i primi 50 numeri non superiori a 584 e diversi fra loro, otteniamo un *campione casuale senza ripetizione.*

Aggiungiamo che ora sono in commercio calcolatrici, anche tascabili, che forniscono numeri aleatori dell'ordine che interessa.

d) *Campione a più stadi*

Quando la popolazione è molto numerosa, per limitare i costi (a parità di rappresentatività del campione), si usano i campioni a più stadi.

Ad es., se vogliamo estrarre un campione di famiglie italiane conviene prima estrarre a caso un campione di provincie (1° stadio) e le provincie saranno chiamate *unità primarie;* nell'ambito di ciascuna provincia estratta campioneremo poi un certo numero di comuni (2° stadio) e i comuni saranno chiamati *unità di secondo stadio*; ed infine, nell'ambito di ciascun comune sceglieremo a caso un certo numero di famiglie (3° stadio) e le famiglie si chiameranno *unità finali*. Queste unità finali possono anche essere diverse dalle *unità elementari* che sono le unità a cui ci si riferisce nell'indagine.

Si aggiunga, inoltre, che conviene usare questo tipo di campionamento anche quando non si ha un unico elenco degli elementi.

e) *Campione stratificato*

Quando la popolazione è molto numerosa (come è il caso, ad es., della rilevazione delle forze di lavoro in Italia) non solo è molto difficile numerare tutti gli elementi della popolazione, ma si corre anche il rischio di inserire nel campione elementi atipici: si ricorre, perciò, al *campione stratificato.*

Il metodo consiste nel classificare le unità statistiche in un certo numero di classi (*strati*) da cui poi estrarre (in mo-

do casuale) gli elementi del campione. Il raggruppamento deve essere fatto in modo tale che ogni elemento deve comparire in un solo strato; inoltre gli elementi che fanno parte di un medesimo strato devono essere simili tra loro (cioé, le differenze esistenti fra di loro devono essere piccole).

Il campionamento stratificato si usa non solo quando la numerosità dell'universo è elevata, ma soprattutto quando gli elementi non sono posti in un'unica lista. Ad es., quando si vogliono rilevare le famiglie della Puglia, bisogna tener conto che esse sono incluse nelle liste comunali: non esiste, cioé, una lista unica dove sono elencate tutte le famiglie pugliesi.

Diciamo subito che raggruppando la popolazione in strati omogenei (ad es., i comuni possono essere raggruppati secondo l'ampiezza demografica, la popolazione geografica, la zona altimetrica, ecc.) *si migliora* — rispetto al campionamento casuale semplice — *la precisione* con cui si stima la grandezza che si vuol rilevare. Invero è evidente che se vogliamo, ad es., stimare il reddito medio della popolazione barese tramite un campione di mille individui, tale reddito è stimato meglio se stratifichiamo la popolazione in classi omogenee di reddito (ad es., in classi professionali); infatti, operando con un campione bernoulliano, si corre il rischio di estrarre, ad es., i mille individui tutti dalla classe inferiore (visto che tale classe è la più numerosa) ed in tal caso il reddito medio sarebbe, ovviamente, sottostimato.

Ci sono vari modí, poi, per stabilire l'ampiezza del campione di ogni strato. Quello generalmente usato è il *campione stratificato proporzionale alle unità della popolazione di ogni strato:* supposto che N_1, N_2, ..., N_s siano il numero delle unità della popolazione contenute negli strati (con $N_1 + N_2 + \ldots + N_s = N$) e supposto che $f = n/N$ sia la frazione di campionamento, le unità da campionare in ogni strato sono date da:

$$n_1 = f N_1, \qquad n_2 = f N_2, \qquad \ldots \qquad n_s = f N_s.$$

E' evidente che questo tipo di campionamento si può effettuare quando si conoscono le unità della popolazione contenute nello strato.

f) Campione sistematico

In questo tipo di campione, numerati gli N elementi dell'universo e posto $\frac{N}{n} = k$ (*passo di estrazione*), si include nel campione un elemento ogni k partendo da un numero $s \leqslant k$ scelto a caso. Ad es., dovendo scegliere 50 elementi da una lista numerata di 1.000 elementi, poiché $k = \frac{1.000}{50} = 20$, si procede così

a) si sceglie a caso (ad es., da un'urna) un numero s compreso fra 1 e 20: ad es., s = 15;

b) si includono, nel campione, gli elementi numero 15, 35, 55, 75, ecc..

Si dimostra che se l'elenco degli elementi di tutta la popolazione è fatto in modo casuale, anche il campione sistematico è di tipo casuale.

Nelle indagini sociali si ritengono come casuali anche quei campioni estratti da elenchi alfabetici perché, mentre da una parte è possibile saltare gruppi di famiglie con lo stesso cognome, dall'altra ciò non è rilevante rispetto alle altre variabili studiate.

Vi sono casi, però, in cui questo tipo di campionamento non è applicabile perché si può ottenere un campione distorto.

Ad es., se abbiamo un elenco di 10.000 contribuenti elencati secondo l'ordine decrescente del reddito posseduto e si vuol conoscere il loro reddito medio tramite un'indagine campionaria di n = 100 elementi, poiché $k = \frac{10.000}{100} = 100$, evidentemente non è la stessa cosa porre s = 1 oppure s = 100: invero, nel 1° caso si ha un valore medio molto elevato, mentre nel secondo caso detto valore medio sarà ovviamente sottostimato.

A cagione di questi motivi, è opportuno che il ricercatore esamini attentamente gli elenchi da cui trarre il campione

sistematico per stabilire se esiste o meno il requisito della casualità.

Un esempio di campionamento sistematico è dato dalla estrazione delle famiglie dalle anagrafi comunali, da parte dell'Istat, per rilevare le forze di lavoro.

g) Campione a grappoli

Un campionamento a *grappoli* si realizza quando si estraggono contemporaneamente tutti gli elementi di un insieme di unità contigue dell'universo (insieme detto, appunto, grappolo).

Ad esempio, se si vuol fare un'indagine sulle caratteristiche degli studenti di scuola media inferiore della provincia di Bologna, si può formare la lista di tutte le scuole medie della provincia (ognuna delle quali è un grappolo), si estraggono alcune scuole e poi si intervistano tutti gli studenti di quelle scuole estratte.

E' evidente, allora, che in questo modo si ha un'ulteriore riduzione di costi di rilevazione: si pensi, ad es., se l'indagine è nazionale, alla difficoltà di intervistare i soggetti dislocati in tutta l'area nazionale e al costo dovuto, anche, alla trasferta da pagare agli intervistatori: è molto più economico, invero, mandare un intervistatore in un posto e farlo rimanere per qualche giorno che non farlo viaggiare ogni giorno (se mai anche per rintracciare e intervistare un individuo al giorno).

Le caratteristiche rilevate con un campionamento a grappoli sono, però, tanto meno precise di quelle ottenute con un campionamento bernoulliano, quanto più omogenei sono gli elementi appartenenti allo stesso grappolo e quanto più grandi sono le differenze esistenti fra i vari grappoli. Ad es., se di 100 quartieri (di una città) se ne estraggono 10, e se le caratteristiche delle unità statistiche all'interno di ogni quartiere sono molto simili, si rischia di includere nel campione quartieri molto atipici e poiché il grappolo di unità statistiche contenute all'interno di ogni quartiere è omogeneo, grande può essere l'errore di campionamento: si pensi, ad es., se i 10 quartieri estratti sono tutti nella zone più ricca della città e se tutte le famiglie all'interno di quei quartieri sono molto

facoltose, in tal caso il reddito familiare della città sarebbe sovrastimato e quindi avremmo un errore di campionamento tanto elevato da non rendere attendibile la stima effettuata.

h) Campione areolare

Nel campionamento *areolare* l'unità di rilevazione è una area geografica, perciò la base è costituita da una carta geografica o da una mappa molto precisa.

Per realizzare un campionamento areolare si può, ad es., dividere la città in gruppi di isolati (aree), si estraggono casualmente alcune di queste aree ed infine si scelgono tutte le unità statistiche in esse contenute (famiglie, persone, aziende, ecc.) oppure soltanto una parte di esse.

Questo tipo di campionamento è molto utile quando non si hanno a disposizione liste aggiornate e complete delle unità statistiche da rilevare.

Stabilito ad es., (nel solito modo) il numero n_i di famiglie (o persone) da rilevare nell'area i, il metodo risulta di facile applicazione in quanto al rilevatore, una volta assegnata l'area, si può dire di intervistare un componente della famiglia di un certo numero civico di un certo stabile, abitante, ad es., sul pianerottolo di sinistra del primo piano, uno che abiti sul pianerottolo di destra del secondo piano, e così via. Il metodo, cioé, non prevede un elenco aggiornato da cui estrarre gli elementi del campione e ciò costituisce un grande vantaggio: molte volte, infatti, gli indirizzi forniti agli intervistatori non sono esatti o perché sbagliati o perché ci sono stati dei trasferimenti non ancora denunciati all'anagrafe.

Prima di concludere facciamo osservare che se rileviamo tutte le unità statistiche contenute nelle aree estratte, il campionamento areolare coincide con quello a grappoli: ogni area, infatti, può considerarsi un grappolo.

Per tale motivo anche il campionamento areale va usato con cautela in quanto ad esso si muovono le critiche già esposte per il campionamento a grappoli.

7. - Errori di rilevazione nelle indagini campionarie

La teoria dei campioni è in grado di tenere sotto controllo gli errori campionari esprimendo un giudizio sull'ordine di grandezza di tali errori. Lo statistico deve però anche cercare di ridurre gli altri tipi di errori, perché è inutile aumentare l'ampiezza campionaria per ridurre l'errore campionario se poi è ugualmente elevato l'errore complessivo a causa degli errori di rilevazione. Poiché, in genere, è molto difficile stimare l'ordine di grandezza di questi errori, è opportuno conoscere quali sono i tipi più comuni allo scopo di prevenirli.

Elenchiamo, perciò, i tipi più comuni di errori di rilevazione che si riscontrano nelle ricerche sociali:

a) errori dovuti alla poca chiarezza delle domande formulate nei questionari (per questo è opportuno effettuare indagini pilota per la verifica del questionario);
b) errori dovuti alla imprecisa definizione dell'universo;
c) errori dovuti alle mancate risposte al questionario (infatti non si sa come avrebbero risposto coloro che non avessero compilato il questionario);
d) errori nelle risposte date dagli intervistati, causati da suggerimenti e interpretazioni errate dell'intervistatore (per questo occorre istruire bene gli intervistatori);
e) errori commessi durante l'operazione di spoglio dei questionari (ad es., mancata registrazione di qualche risposta, oppure errata interpretazione di qualche risposta se il questionario è a domande aperte);
f) errori dovuti alle volontarie risposte errate dell'intervistato sia nei riguardi di domande delicate (non formulate garantendo sufficientemente l'anonimato: ad es., domande sull'uso di droga, sull'esercizio della prostituzione, ecc.) e sia per i pregiudizi non rimossi (come il timore che le risposte date siano usate contro l'intervistato: è il caso, ad es., delle evasioni fiscali);
g) errori di trascrizioni (ad es., 77 invece di 67).

Come si vede, i possibili errori di rilevazione sono molti. E' compito dei ricercatori cercare di eliminarne il maggior numero possibile attraverso un accurato controllo di tutte le fasi dell'indagine.

CAPITOLO II

FASE DI RILEVAZIONE

La rilevazione dei dati è quel complesso di operazioni che consente di pervenire alla conoscenza delle caratteristiche delle singole unità statistiche. Essa può dividersi in tre sottofasi: *raccolta, classificazione* e *rappresentazione grafica* dei dati.

1. - La raccolta dei dati

La raccolta dei dati è la prima importante operazione che lo statistico deve attuare. Tale operazione, nel caso la ricerca sia promossa dallo Stato e da Enti, può essere eseguita da appositi uffici: ad es., l'ISTAT raccoglie, nelle sue pubblicazioni ufficiali, i dati per conto dello Stato.

Quando, invece, la ricerca è effettuata da una équipe di studiosi privati, la raccolta sarà a cura dei medesimi: è il caso, ad es., degli studi realizzati da operatori sociali per capire meglio l'ambiente in cui lavorano ed i problemi che affrontano o devono affrontare.

In sostanza, quindi, la raccolta dei dati consiste nella pura e semplice registrazione di questi su schede, su questionari, su cartelle cliniche, ecc..

Ovviamente, man mano che si raccolgono i moduli (o i questionari) bisogna revisionarli per accertare se la loro compilazione è soddisfacente e ciò per evitare una delle numerose fonti di errori che si possono commettere nelle indagini.

La raccolta dei dati può essere *occasionale* (come, ad es., nelle indagini svolte per particolari aspetti del campo clinico di un ospedale); *periodica* (ad es., nei censimenti demografici) e *continua* (ad es., nelle rilevazioni anagrafiche).

2. - La classificazione dei dati e la tabulazione dei risultati

Una volta raccolti tutti i moduli (o, ad es., tutti i questionari) che servono per l'indagine, si può passare ad esaminare ed organizzare le informazioni contenute in essi: si procede, cioé, allo spoglio dei moduli.

Lo spoglio o classificazione è quella operazione che consiste nel raggruppare tutti i dati secondo le modalità dello stesso carattere. Per effettuarlo più facilmente si fa corrispondere ad ogni modalità una casella (in tal modo si formano tante caselle quante sono le modalità del carattere) e, in quest'ultima, si pone un'asta ogni qualvolta l'unità statistica possiede la modalità corrispondente a quella casella. Generalmente, per poterle più facilmente contare, si raggruppano le aste a cinque a cinque (ogni qualvolta si raggiunge la quinta unità, l'asta si scrive trasversalmente sopra le precedenti, cioé il simbolo 𝍸 sta a significare cinque unità).

Supponiamo, ad es., di voler classificare 25 lavoratori secondo il settore di attività economica: agricoltura (A), industria (I), commercio (C), altre attività (AA)

A, I, I, C, C, AA, C, A, A, C, C, I, AA, A, A, C, C, A, AA, I, A, I, C, I, AA.

Poiché le modalità sono quattro, si formano quattro caselle e si trascrive, con un'asta, ogni soggetto nella casella corrispondente al suo settore di attività economica. Si ottiene, così, il prospetto seguente.

A	𝍸 II
I	𝍸 I
C	𝍸 III
AA	IIII

Si osservi che il primo soggetto è riportato nella prima casella del prospetto, il 2° e 3° soggetto nella seconda casella, il 4° e 5° soggetto nella terza, ecc..

Concluso lo spoglio, si tabulano i risultati. La *tabulazione* consiste nel mettere i dati sotto forma di tabella. La tabella statistica o *distribuzione statistica* semplice è un prospetto di due colonne: nella prima sono elencate le modalità (possibilmente ordinate) del carattere, mentre nella seconda sono riportati i numeri che esprimono le unità statistiche che posseggono tali modalità. Questi numeri si chiamano frequenze assolute.

La *frequenza assoluta* di una generica modalità x_i è, quindi, il numero di volte n_i con cui si presenta quella modalità. Nel caso dell'esempio dei lavoratori classificati secondo il settore di attività, si ha la Tav. II/1.

Tav. II/1. - Distribuzione di 25 lavoratori secondo il settore di attività economica.

Settore di attività economica	N. di lavoratori
Agricoltura	7
Industria	6
Commercio	8
Altre attività	4
Totale	25

In detta tavola al "Commercio" spetta la frequenza 8 (cioé, vi sono 8 soggetti che lavorano nel settore del Commercio), all'Agricoltura spetta la frequenza 7 (cioé, vi sono 7 persone che lavorano in Agricoltura), ecc..

Si noti che per non creare confusione, le tabelle devono riportare una didascalia che illustri il carattere rilevato, ed anche le modalità e le rispettive frequenze.

Le distribuzioni statistiche proprio perché associano ad ogni modalità x_i una frequenza n_i, vengono anche chiamate *distribuzioni di frequenze.*

Ogni qualvolta i dati esprimono una misura (ad es., misura del peso di un neonato in relazione ai giorni di crescita), oppure l'ammontare di un carattere (ad es., l'ammontare della popolazione presente nelle varie regioni italiane al censimento 1981) si parla di *distribuzione d'intensità.*

Le distribuzioni statistiche il cui carattere è qualitativo si chiamano anche *serie* o *mutabili statistiche* (indicate sinteticamente con "m.s.").

A seconda che il carattere qualitativo sia non ordinabile o ordinabile, la m.s. si dirà non ordinabile o ordinabile.

La Tav. II/1 mostra un esempio di m.s. non ordinabile, la Tav. II/2 e la Tav. II/3 rappresentano esempi di m.s. ordinabili.

Fra le m.s. ordinabili assumono notevole importanza le m.s. *cicliche*: esse sono quelle mutabili le cui modalità presentano un ordine definito di successione senza che si possa dire, se non facendo una convenzione, quale è la prima e quale è la ultima (ad es., giorni della settimana) (Tav. II/3).

Se il carattere è quantitativo, le distribuzioni statistiche si chiamano *seriazioni*.

Tav. II/2 - Distribuzione di 50 ufficiali secondo il grado.

Gradi militari	n_i
sottotenente	9
tenente	20
capitano	8
maggiore	6
tenente colonnello	4
colonnello	2
generale	1

Tav. II/3 - Personale assente, nella 1ª settimana del mese di maggio 1986, in un complesso alberghiero.

Giorni	Personale assente
lunedì	7
martedì	9
mercoledì	8
giovedì	4
venerdì	11
sabato	7
domenica	2

Tav. II/4 - Morti in Italia per malattie dell'apparato respiratorio, 1974 - 83.

Anni	Numero di morti
1974	39.238
1975	47.465
1976	41.630
1977	42.246
1978	37.388
1979	36.053
1980	39.397
1981	39.456
1982	33.475
1983	39.243

Tav. II/5 - Famiglie per numero di componenti. Marche. Censimento 1981.

Numero di componenti	Numero di famiglie
1	66.002
2	105.399
3	104.162
4	104.076
5	45.964
6 e più	25.518
Totale	451.121

Poiché in una distribuzione statistica le modalità variano, noi useremo chiamare le seriazioni con la dizione *variabili statistiche* (indicate sinteticamente con v.s.).

Sicché diremo che una v.s. è una variabile che assume le modalità $x_1, x_2, \ldots, x_s$ con frequenze $n_1, n_2, \ldots, n_s$ e la indicheremo con

$$\begin{pmatrix} x_1 & x_2 & \ldots & x_s \\ n_1 & n_2 & \ldots & n_s \end{pmatrix}$$

oppure con

x_i	n_i
x_1	n_1
x_2	n_2
.	.
.	.
.	.
x_s	n_s

Analoga definizione è valida per la mutabile statistica.

Secondo che il carattere quantitativo sia discreto o continuo, la v.s. si dirà discreta o continua.

Nella Tav. II/5 è riportato un esempio di v.s. discreta, e nella Tav. II/7 uno di v.s. continua.

Fra le serie, particolare importanza assumono le *serie storiche*. In esse le modalità x_i si riferiscono al tempo, mentre le n_i esprimono l'ammontare del fenomeno: ad es., i morti in Italia per malattia dell'apparato respiratorio negli anni 1974 - 1983 (Tav. II/4).

Le *serie territoriali*, invece, sono distribuzioni in cui le x_i sono unità territoriali, mentre le n_i rappresentano l'ammontare del fenomeno relativo al territorio x_i: ad es., gli indici di diffusione della televisione (n_i) secondo le regioni (x_i) italiane (Tav. II/6), ove con n_i abbiamo indicato il numero di televisori per 100 famiglie della regione x_i.

Dalla Tav. II/7 appare evidente che quando il carattere è continuo, le modalità vanno raggruppate in classi: si usa par-

Tav. II/6 - Indici di diffusione della televisione per regioni. 1974.

Regioni	Indici di diffusione
Piemonte	71,91
Valle d'Aosta	64,99
Lombardia	80,21
Trentino Alto Adige	73,67
Veneto	81,32
Friuli Venezia Giulia	77,03
Liguria	74,91
Emilia Romagna	76,85
Toscana	80,59
Umbria	78,13
Marche	78,35
Lazio	74,66
Abruzzi e Molise	68,32
Campania	56,89
Puglia	67,66
Basilicata	57,70
Calabria	50,15
Sicilia	50,39
Sardegna	65,70
ITALIA	71,32

Fonte: *M. Lo Presti*, Aspetti statistici del fenomeno televisivo in Puglia, Rassegna Economica, n. 5, 1975.

Tav. II/7 - Ricoverati in una divisione di pediatria secondo gruppi di età.

Classi di età	Numero di ricoverati
Meno di 1 anno	20
Da 1 a 2 anni compiuti	15
Da 3 a 5 anni compiuti	12
Da 6 a 9 anni compiuti	10
Da 10 a 11 anni compiuti	3
Totale	60

lare, perciò, anche di v.s. divise in intervalli.

Ad es., nella Tav. II/7, nella prima classe (indicata con "Meno di 1 anno") sono inclusi bambini che hanno un'età compresa fra la nascita ed 1 anno meno 1 giorno alla data in cui si effettua l'indagine; nella seconda classe (indicata con "Da 1 anno a 2 anni compiuti") sono inclusi bambini che hanno un'età compresa fra 1 anno e 3 anni meno 1 giorno alla data di rilevazione; ecc..

La distribuzione della Tav. II/7 può anche scriversi come la Tav. II/8 e la Tav. II/9. Si ponga attenzione che le classi sono sempre le stesse, ma indicate in maniera diversa. La simbologia della Tav. II/8 è quella che attualmente usa l'Istat.

Tav. II/8 - Ricoverati in una divisione pediatrica secondo gruppi di età.

Classi di età (in anni compiuti)	Numero di ricoverati
Meno di 1	20
1 - 2	15
3 - 5	12
6 - 9	10
10 - 11	3
Totale	60

Tav. II/9 - Ricoverati in una divisione pediatrica secondo gruppi di età.

Classi di età (in anni compiuti)	Numero di ricoverati
Meno di 1	20
1 ⊢ 3	15
3 ⊢ 6	12
6 ⊢ 10	10
10 ⊢ 12	3
Totale	60

Nel caso della simbologia della Tav. II/9, si conviene considerare l'estremo inferiore della classe come appartenente ad essa e l'estremo superiore come appartenente alla classe successiva: in tal caso si dice che gli intervalli sono chiusi a sinistra e aperti a destra e si scrive $x_{i-1} \vdash x_i$.

La differenza fra i due estremi (superiore ed inferiore) della classe si chiama *ampiezza* o *modulo* della classe: in ge-

nere, se gli estremi della classe sono x_{i-1} ed x_i, l'ampiezza della classe è

$$d_i = x_i - x_{i-1}.$$

Ad es., dalle Tavv. II/7, 8, 9 si deduce che la prima classe ha ampiezza 1 (= 1 – 0), mentre la seconda ha ampiezza 2 (= 3 – 1), la terza ha ampiezza 3 (= 6 – 3), ecc..

Dicesi *valore centrale* di una classe la semisomma degli estremi: cioé, $\frac{x_{i-1} + x_i}{2}$ è il valore centrale della classe $x_{i-1} \vdash\!\!- x_i$, mentre $\frac{0 + 1}{2} = 0{,}5$ è il valore centrale della classe "Meno di 1", $\frac{1 + 3}{2} = 2$ è il valore centrale della classe $1 \vdash\!\!- 3$.

La determinazione di detto valore è utile per gli ulteriori calcoli che si devono eseguire nella fase di elaborazione dei risultati.

Per concludere l'argomento della tabulazione dei dati diciamo che, se il numero di moduli (di cui bisogna fare lo spoglio) è elevato, tale operazione va eseguita con calcolatori elettronici.

3. - Il raggruppamento dei dati in classi

Alle volte riportare tutti i dati osservati di una v.s. continua è fonte di confusione perché è difficile sintetizzare gli aspetti più significativi del fenomeno: si ricorre, allora, al raggruppamento in classi. Questi gruppi, pur facendo perdere qualche informazione, consentono di avere una visione chiara e semplice di come sono distribuiti i dati.

Riguardo questi ultimi, nel paragrafo precedente abbiamo solo detto che, quando il carattere è continuo, essi possono essere raggruppati in classi. Tuttavia non sono stati forniti dei criteri per il raggruppamento: le classi fin qui studiate

erano, cioé, gia determinate a priori sicché, rilevate le modalità di un carattere continuo, non bisognava far altro che contare il numero di casi appartenenti ad una certa classe.

Non esistono, purtroppo, delle regole fisse per raggruppare i dati: i metodi usati sono legati, invero, al tipo di problema che si sta esaminando. I suggerimenti da dare per l'individuazione delle classi riguardano, comunque, tre aspetti:

a) le proprietà cui esse devono soddisfare;

b) l'ampiezza di ciascuna;

c) il loro numero.

In relazione al punto a), occorre che le classi soddisfino due proprietà:

— esse *devono essere esaustive*, cioé, tutti gli elementi dello universo devono poter trovare la loro collocazione;

— *non devono essere sovrapposte*, e ciò per evitare di trovarsi in presenza di unità statistiche che possono essere collocate in più classi.

Quanto al secondo aspetto, un criterio molto usato è quello di dividere l'intervallo compreso fra il più piccolo e il più grande dei valori osservati (*campo di variazione*) in un numero di classi di uguale ampiezza: in tal modo possono essere comparate le frequenze corrispondenti ad ognuna di esse. Si osservi, però, che tale metodo non è raccomandabile sia quando qualche classe ha frequenza zero, sia quando a qualcuna compete il maggior numero dei casi: in quest'ultima circostanza, infatti, si ha una perdita notevole di informazioni rispetto alla distribuzione originale.

Un altro metodo è quello di dividere il campo di variazione in intervalli disuguali in modo, però, che ad ogni classe competa all'incirca la stessa frequenza.

Quando, poi, si osservano valori estremi che si discostano notevolmente dagli altri dati, alle volte si usano *classi non limitate*: classi, cioé, per le quali non è indicato o l'estremo inferiore o l'estremo superiore. A tale riguardo facciamo notare

che quando non viene precisato l'estremo inferiore della 1ª classe, essa sarà indicata con "Meno di ...."; se x_s è l'estremo inferiore dell'ultima classe, quando non ne viene dato l'estremo superiore, tale classe sarà indicata con "x_s e più". Nelle indagini sociali è opportuno, però, evitare che le classi siano non limitate, e ciò per non complicare le elaborazioni successive specie quando è necessario determinare il valore centrale della classe.

Quanto al problema del numero di classi, conviene, in prima approssimazione, suddividere il campo di variazione in un numero sufficientemente elevato di classi uguali per poi, eventualmente, raggrupparle in classi più ampie (uguali o disuguali): se usassimo, invero, poche classi (ma più ampie) non potremmo poi più dividerle ove fosse necessario.

Esistono numerosi altri metodi (ad es., quello della *Clusters Analysis*), molto più precisi e meno soggettivi, l'uso dei quali richiede, però, nozioni di Matematica che vanno al di là dei programmi della scuola media superiore.

Facciamo un esempio.

Intervistati 80 individui, si considerino i guadagni (in migliaia di lire) che i medesimi hanno dichiarato di aver percepito nei 6 giorni precedenti l'intervista

900, 200, 195, 185, 150, 160, 163, 170, 197, 185, 200, 30; 130, 147, 150, 160, 170, 190, 174, 160, 190, 158, 50, 160, 170, 187, 180, 165, 170, 170, 180, 170, 160, 55, 160, 176, 190, 170, 178, 165, 170, 160, 185, 165, 90, 170, 192, 195, 180, 170, 165, 197, 160, 180, 350, 100, 160, 170, 198, 160, 160, 164, 170, 185, 196, 120, 160, 178, 180, 170, 500, 155, 160, 170, 190, 170, 160, 122, 150, 150.

Possiamo, ad es., chiederci:

— quanti individui, in quei 6 giorni, hanno guadagnato meno di 100.000 lire?

— quanti individui hanno guadagnato 200.000 lire e più?

Per rispondere più facilmente a queste domande è opportuno raggruppare i dati in classi.
Tenuto presente che il più piccolo valore è 30 ed il più alto è 900, il campo di variazione (*range* in inglese) è

$$x_s - x_1 = 900 - 30 = 870.$$

Poiché i valori estremi si discostano notevolmente dagli altri dati, come prima classe considereremo la classe non limitata inferiormente "meno di 100" e come ultima la classe non limitata superiormente "200 e più".

Visto che tutti gli altri valori sono compresi fra 100 e 199 possiamo pensare, ad es., di raggruppare i dati in 10 classi tutte di ampiezza uguale a 10. Si ha così il seguente prospetto.

Classi (in migliaia di lire)	Frequenze	Classi (in migliaia di lire)	Frequenze
Meno di 100	𝍸	150 - 159	𝍸 |
100 - 109	|	160 - 169	𝍸 𝍸 𝍸 𝍸
110 - 119		170 - 179	𝍸 𝍸 𝍸 ||||
120 - 129	||	180 - 189	𝍸 𝍸
130 - 139	|	190 - 199	𝍸 𝍸 |
140 - 149	|	200 e più	𝍸

Appare evidente, ora, che non solo non c'è alcun individuo il cui guadagno è compreso nella classe (110 - 119) (per cui tale classe è inutile), ma anche che le frequenze di talune classi sono troppo esigue (per cui le classi stesse si possono ulteriormente raggruppare). In definitiva, quindi, ove gli scopi dell'indagine non consigliano diversamente, suggeriamo la distribuzione della Tav. II/10.

Tav. II/10 - Guadagni (in migliaia di lire) di 80 individui nei giorni precedenti l'intervista.

Classi (in migliaia di lire)	Frequenze
Meno di 100	4
100 - 149	5
150 - 159	6
160 - 169	20
170 - 179	19
180 - 189	10
190 - 199	11
200 e più	5
Totale	80

4. - Frequenze relative - Frequenze cumulate - Frequenze relative cumulate

Indichiamo con n_i sia la frequenza (assoluta) con la quale si presenta la generica modalità x_i quando la v.s. è discreta, sia la frequenza dei casi che cadono nella classe $x_{i-1} \vdash\!\!- x_i$ quando la v.s. è continua.

Il numero $$n = n_1 + n_2 + \ldots + n_s = \sum_{i=1}^{s} n_i$$

rappresenta la totalità delle frequenze [6].

Il numero $$f_i = \frac{n_i}{n}$$

si chiama *frequenza relativa* della modalità x_i o della i.ma classe: la frequenza relativa è dunque il rapporto fra la frequenza assoluta e il totale delle frequenze. La 3ª colonna della Tav. II/11 è stata ottenuta ponendo

$f_1 = 2/20 = 0{,}10$; $f_2 = 3/20 = 0{,}15$; $f_3 = 5/20 = 0{,}25$; ecc..

Tav. II/11 - Bambini di una collettività. Esempio di frequenze assolute, relative, cumulate e frequenze relative cumulate.

Classi d'età (in anni)	n_i	f_i	N_i	F_i
Meno di 1	2	0,10	2	0,10
1 ⊢ 3	3	0,15	5	0,25
3 ⊢ 6	5	0,25	10	0,50
6 ⊢ 10	10	0,50	20	1

E' evidente che la somma delle frequenze relative risulta uguale all'unità.

---

[6] Dati più elementi dipendenti da un indice
$$n_1, n_2, \ldots, n_r, \ldots, n_s, \ldots$$
il simbolo $\sum_{i=r}^{s} n_i$, che si legge "sommatoria di n_i, per i che varia da r ad s", sta ad indicare la somma degli elementi che vanno dal posto r al posto s: nel simbolo è specificato, quindi, il primo e l'ultimo indice dei termini della somma. Sicché:

$$\sum_{i=3}^{6} n_i = n_3 + n_4 + n_5 + n_6; \quad \sum_{i=2}^{4} n_i = n_2 + n_3 + n_4;$$

$$\sum_{i=1}^{s} n_i = n_1 + n_2 + \ldots + n_s.$$

Il simbolo $\sum$ (senza indici) indica, invece, la somma di tutti i termini.

Le frequenze relative moltiplicate per 100 si chiamano *percentuali*, per cui $f_1 \cdot 100 = 10$ sta a significare che il 10% del totale dei bambini ha meno di 1 anno di età; $f_2 \cdot 100 = 15$ sta a significare che il 15% dei bambini di quella collettività ha un'età fra 1 anno e 3 anni meno un giorno; ecc..

I numeri

$$N_1 = n_1$$
$$N_2 = n_1 + n_2$$
$$N_i = n_1 + n_2 + \ldots\ldots + n_i$$
$$N_s = n_1 + n_2 + \ldots\ldots + n_s = n,$$

si chiamano *frequenze cumulate o frequenze sommatorie.*

Dalla 4ª colonna della Tav. II/11, si ricava che $N_1 = n_1 = 2$ è la frequenza dei bambini di età inferiore ad 1 anno; $N_2 = n_1 + n_2 = 2 + 3 = 5$ è la frequenza dei bambini di età inferiore a 3 anni; $N_3 = n_1 + n_2 + n_3 = 2 + 3 + 5 = 10$ è la frequenza dei bambini di età inferiore a 6 anni; ecc..

La distribuzione

$$\begin{pmatrix} x_1 & x_2 \ldots\ldots x_s \\ N_1 & N_2 \ldots\ldots N_s \end{pmatrix}$$

si chiama *distribuzione cumulativa di frequenze.*

I numeri

$$F_1 = f_1$$
$$F_2 = f_1 + f_2$$
$$\ldots\ldots$$
$$F_s = f_1 + f_2 + \ldots\ldots + f_s = 1$$

si chiamano *frequenze relative cumulate.*

Nella 5ª colonna della Tav. II/11, $F_1 = f_1 = 0{,}10$ fornisce la frequenza relativa dei bambini d'età inferiore ad 1 anno, $F_2 = f_1 + f_2 = 0{,}10 + 0{,}15 = 0{,}25$ è la frequenza relativa dei bambini d'età inferiore a 3 anni, $F_3 = f_1 + f_2 + f_3 = 0{,}10 + 0{,}15 + 0{,}25 = 0{,}50$ è la frequenza relativa dei bambini con un'età minore di 6 anni; ecc..

La frequenza relativa di tutti i valori (che può assumere la X) minori della modalità x_i, cioè

$$F(x_i) = Fr(X < x_i),$$

si chiama *funzione di distribuzione* (o di *ripartizione*).

Se la v. s. è discreta, la funzione di distribuzione è definita, perciò, da

$$F(x_i) = \begin{cases} 0 & \text{per } x < x_1 \\ F_i & \text{per } x_i \leqslant x < x_{i+1} \\ 1 & \text{per } x \geqslant x_s . \end{cases}$$

Nei punti di discontinuità risulta, allora, che i valori della funzione di distribuzione coincidono con le frequenze relative cumulate, cioè

$F(x_1) = 0$, $F(x_2) = F_1$, $F(x_3) = F_2$, ..., $F(x_s) = 1$.

5. - Altre definizioni

Quando il carattere è ordinabile e le modalità x_i sono disposte in ordine non decrescente o in ordine non crescente, la v. s. si dice *graduata*. Il numero d'ordine che compete ad una certa modalità (cioè il posto che detta modalità occupa in una certa graduatoria) si chiama *rango*.

Assegnate due v. s.

$$\begin{pmatrix} x_1 & x_2 \ldots\ldots x_s \\ n_1 & n_2 \ldots\ldots n_s \end{pmatrix} \quad \text{e} \quad \begin{pmatrix} x'_1 & x'_2 \ldots\ldots x'_s \\ n'_1 & n'_2 \ldots\ldots n'_s \end{pmatrix}, \qquad [1]$$

la modalità x_i si dice *cograduata* alla modalità x'_j quando entrambe le modalità occupano lo stesso posto nella graduatoria non decrescente delle modalità stesse; invece, x_i si dice *contrograduata* alla modalità x'_j quando entrambe occupano lo stesso posto l'una nella graduatoria non decrescente, mentre l'altra nella graduatoria non crescente. E' opportuno far notare che le modalità x_i e x'_j possono essere cograduate o contrograduate anche se la X e la X' non si riferiscono allo stesso fenomeno, cioè non sono omogenee.

Quando le modalità di due v. s. riguardano la stessa unità statistica allora si dicono *associate*. Ad esempio, si considerino le votazioni conseguite da 4 studenti A, B, C, D agli esami di Statistica (X) e Sociologia (X')

$$X \equiv \begin{pmatrix} A & B & C & D \\ 27 & 25 & 24 & 25 \end{pmatrix} \quad X' \equiv \begin{pmatrix} A & B & C & D \\ 30 & 24 & 24 & 26 \end{pmatrix}.$$

Le votazioni di 27 in Statistica e di 30 in Sociologia sono associate in quanto riguardano lo stesso individuo A. E' ovvio che, in questo caso, $n = n'$.

Due v.s. sono *uguali* quando si riferiscono allo stesso carattere, sono rilevati nella stessa unità ed hanno i valori associati uguali.

Due distribuzioni si dicono *somiglianti* quando hanno la stessa funzione di distribuzione.

Si noti che se due distribuzioni sono uguali sono anche somiglianti, ma non è vero il contrario.

6. - Le tabelle a doppia entrata

Fin qui abbiamo classificato i dati secondo le modalità di un solo carattere.

Supponiamo, ora, che per ogni unità statistica di una determinata popolazione si rilevino due caratteri. Ad esempio, supponiamo di rilevare il numero di vani ed il numero di persone che abitano in 20 appartamenti. Ad ogni appartamento (unità statistica) si può associare, quindi, una coppia ordinata di numeri reali (x, y): il numero x rappresenta il numero di vani di quell'appartamento ed il numero y il numero di persone che vi abitano. I 20 appartamenti rilevati saranno rappresentati, perciò, da 20 coppie ordinate di numeri reali, ad esempio da:

(2,2), (2,2), (2,3), (2,3), (3,3), (3,2), (2,4), (3,4), (4,2), (4,3),
(5,2), (5,4), (5,4), (5,4), (5,4), (5,3), (5,3), (5,3), (5,2), (5,4).

In questo caso il *carattere* X è dato dal numero di vani di ogni appartamento ed il *carattere* Y è dato dal numero di persone per appartamento.

Poichè, nel nostro caso, esistono appartamenti di 2, 3, 4, 5 vani, 2, 3, 4 e 5 si dicono le *modalità* del carattere X.

Analogamente, poiché esistono appartamenti abitati da 2, 3, 4 persone, si dice che 2, 3, 4 sono le modalità del carattere Y.

In generale, quindi, si può dire che il carattere X si manifesta con modalità $x_1, x_2, \ldots\ldots, x_s$, ed il carattere Y con le modalità $y_1, y_2, \ldots\ldots y_t$.

Specie quando le unità raccolte sono molte, per avere una visione globale del fenomeno, si classificano i dati in una tabella a doppia entrata.

Vediamo in che modo si forma una tabella a doppia entrata considerando le 20 coppie di numeri precedentemente scritti. Si comincia col disegnare un rettangolo. Sul lato orizzontale

superiore si scrivono le modalità della X ordinate e prese una sola volta (nel nostro caso, quindi, 2, 3, 4, 5), mentre a fianco al lato verticale di sinistra si scrivono le modalità della Y ordinate e prese una sola volta (nel nostro caso 2, 3 e 4). Si mandano delle linee orizzontali che dividono il rettangolo in tante strisce quante sono le modalità della Y (nel nostro caso, quindi, in tre strisce) in modo tale, però, che in ogni striscia vi capiti una sola modalità della Y. Analogamente si mandano delle linee verticali che dividono il rettangolo in tante strisce in modo tale che in ogni striscia vi capiti una sola modalità della X. In tal modo, se s sono le modalità della X e t le modalità della Y, il rettangolo è decomposto in $s \cdot t$ caselle (nel nostro caso è stato decomposto in $4 \cdot 3$ caselle). A questo punto ogni unità statistica (formata, come si è detto, da una coppia di numeri reali) viene scritta con un'asta in quella casella che ha in alto il primo numero della coppia ed a sinistra il secondo numero della stessa coppia. Ad esempio, le prime due coppie sono state riportate nella I^a casella in alto a sinistra della Tav. II/12/a; la 3^a e la 4^a coppia sono state riportate nella casella sotto la precedente (nella stessa Tav. II/12/a). Operando nello stesso modo con tutte le unità statistiche considerate (nel nostro caso, 20 appartamenti), si ottiene la Tav. II/12/b.

Tav. II/12/a

Y	X			
	2	3	4	5
2	||			
3	||			
4				

Tav. II/12/b

Y	X			
	2	3	4	5
2	||	|	|	||
3	||	|	|	|||
4	|	|		𝍸

Si osservi che, per facilitare il conteggio, quando si sono già scritte quattro unità nella stessa casella, la quinta unità si scrive trasversalmente sopra le precedenti cioè 𝍸 (vedasi ad esempio l'ultima casella in basso a destra della Tav. II/12/b). A questo punto si contano le unità di ciascuna casella e si portano in una tabella del tipo della Tav. II/13.

Tav. II/13 - Appartamenti secondo il numero di vani ed il numero di persone che vi abitano.

Persone	Vani				n_{0h}
	2	3	4	5	
2	2	1	1	2	6
3	2	1	1	3	7
4	1	1	–	5	7
n_{i0}	5	3	2	10	20

Tav. II/14 - Esempio teorico di tabella a doppia entrata.

Y	X						Freq. marg.
	x_1	x_2	· · ·	x_i	· · ·	x_s	
y_1	n_{11}	n_{21}	· · ·	n_{i1}	· · ·	n_{s1}	n_{01}
y_2	n_{12}	n_{22}	· · ·	n_{i2}	· · ·	n_{s2}	n_{02}
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
y_h	n_{1h}	n_{2h}	· · ·	n_{ih}	· · ·	n_{sh}	n_{0h}
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮
y_t	n_{1t}	n_{2t}	· · ·	n_{it}	· · ·	n_{st}	n_{0t}
Freq. marg.	n_{10}	n_{20}	· · ·	n_{i0}	· · ·	n_{s0}	n

Ovviamente, nel caso n sia molto grande, il conteggio non sarà fatto a mano bensì con l'ausilio di un calcolatore.

In generale si può, dunque, scrivere la Tav. II/14, ove i numeri che stanno dentro il rettangolo limitato dalla doppia linea sono frequenze.

Ad es., n_{11} sta ad indicare quante volte si presenta la coppia (x_1, y_1), n_{ih} sta ad indicare quante volte si presenta la coppia (x_i, y_h). Inoltre si è posto $n_{i0} = n_{i1} + n_{i2} + \ldots + n_{it}$. Dunque, n_{i0} indica quante volte si presenta la sola modalità x_i a prescindere da come è accoppiata con le varie modalità della Y. Ad es., nella Tav. II/13, $n_{10} = 5$ sta ad indicare quanti sono gli appartamenti con 2 vani (qualsiasi le persone che l'occupano).

Analogamente $n_{20} = n_{21} + n_{22} + \ldots + n_{2t}$ indica quante volte si presenta x_2. Così $n_{20} = 1 + 1 + 1 = 3$ indica quanti appartamenti hanno tre vani. E così via.

Invece $n_{01} = n_{11} + n_{21} + \ldots + n_{s1}$ indica quante volte si presenta la modalità y_1; $n_{02} = n_{12} + n_{22} + \ldots + n_{s2}$ indica quante volte si presenta la modalità y_2; e così via.

Ad es., nella Tav. II/13, $n_{02} = 7$ indica quanti sono gli appartamenti abitati da 3 persone qualsiasi siano i vani degli appartamenti stessi. Le frequenze $n_{01}, n_{02}, \ldots, n_{0h}, \ldots, n_{0t}$ e le analoghe $n_{10}, n_{20}, \ldots, n_{i0}, \ldots, n_{s0}$ si chiamano *frequenze marginali*. La somma n delle frequenze marginali poste nell'ultima colonna o di quelle poste nell'ultima riga della Tav. II/10 cioè

$n = \sum_{i=1}^{s} n_{i0} = \sum_{h=1}^{t} n_{0h}$, fornisce (nell'esempio portato) il totale degli appartamenti .

Un esempio di tabella a doppia entrata in cui un carattere è qualitativo può essere dato dalla distribuzione di un gruppo di laureati in Statistica secondo l'età e il tipo di maturità (Tav. II/11).

La tabella a doppia entrata viene anche chiamata *matrice dei dati* (in analogia alle tabelle di elementi che in Matematica si chiamano matrici).

E' evidente che, ad es., il 20 che nella Tav. II/15 è segnato con la linea sta ad indicare che esistono 20 laureati in Statistica che hanno un'età di 24 anni compiuti e sono dotati di maturità scientifica.

Per concludere diciamo che a seconda che i due caratteri siano entrambi quantitativi, entrambi qualitativi, oppure uno quantitativo e l'altro qualitativo, la tabella a doppia entrata rappresenta una v.s. doppia, una m.s. doppia, oppure una distribuzione doppia mista. Una m.s. doppia viene denominata anche *tabella di contingenza*.

Tav. II/15 - Laureati in Statistica secondo l'età e il tipo di maturità.

Età in anni compiuti	Tipo di maturità						Totale n_{0h}
	classica	scientifica	magistrale	tecnica industr.	tecnica comm.	altre	
22	10	8	1	—	1	—	20
23	20	18	2	—	2	—	42
24	10	20	2	2	5	—	39
25	5	13	2	4	10	1	35
26	5	7	4	6	3	1	26
27	3	5	2	4	3	—	17
28 e più	2	5	2	5	7	—	21
Totale n_{i0}	55	76	15	21	31	2	200

7. - Distribuzioni condizionate - Distribuzioni marginali

Una distribuzione statistica semplice relativa ad una qualunque colonna o riga della tabella a doppia entrata è denominata *distribuzione condizionata* del carattere Y sotto la condizione $x = x_i$, o del carattere X sotto la condizione $y = y_i$.

Ad es., la distribuzione della Tav. II/16, estratta dalla Tav. II/13 associando alle modalità del carattere "Persone" le frequenze riguardanti la prima colonna della tabella a dop-

Tav. II/16

Persone (Y)	frequenze
2	2
3	2
4	1
Totale	5

Tav. II/17

Vani (X)	frequenze
2	2
3	1
4	1
5	3
Totale	7

pia entrata, è la distribuzione del carattere "Persone" condizionata al fatto che gli appartamenti hanno tutti due vani.

Analogamente, la distribuzione della Tav. II/17, ottenuta associando alle modalità del carattere "Vani" la seconda riga della Tav. II/13, è la distribuzione dei vani condizionata al fatto che tutti gli appartamenti sono abitati da 3 persone.

Le due distribuzioni statistiche semplici che si ottengono associando alle modalità del fattore "Persone", oppure alle modalità del fattore "Vani", le rispettive frequenze marginali, cioé

Y	n_{0h}
2	6
3	7
4	7
Totale	20

X	n_{i0}
2	5
3	3
4	2
5	10
Totale	20

si chiamano, invece, *distribuzioni marginali* rispettivamente del carattere "Persone" e del carattere "Vani".

In generale, quindi, le distribuzioni marginali dei due caratteri di una tabella a doppia entrata sono quelle che associano alle modalità dei due caratteri le loro frequenze marginali. Cioé, in simboli,

Y	Freq. marginali
y_1	n_{01}
y_2	n_{02}
.	.
.	.
.	.
y_t	n_{0t}
Totale	n

X	Freq. marginali
x_1	n_{10}
x_2	n_{20}
.	.
.	.
.	.
x_s	n_{s0}
Totale	n

In una tabella a doppia entrata, le *frequenze relative* sono date dai rapporti fra le frequenze di ogni casella e il totale n (ad es., riferendoci alla Tav. II/15, la frequenza relativa dei laureati in Statistica di 25 anni compiuti ed in possesso di Maturità scientifica è 13/200); le *frequenze marginali relative* sono fornite dai rapporti fra le frequenze marginali e la frequenza totale (ad es., riferendoci alla Tav. II/13 la frequenza relativa di tutti gli appartamenti che hanno due vani è 5/20); le *frequenze condizionate relative* sono date dai rapporti fra le frequenze di ciascuna casella e le frequenze marginali della colonna e della riga a cui la casella appartiene (ad es., dalla Tav. II/15 si ricava che la frequenza relativa dei laureati in Statistica di 25 anni compiuti condizionati ad avere maturità scientifica è 13/76, mentre la frequenza relativa dei laureati in Statistica che posseggono maturità scientifica condizionati ad avere 25 anni compiuti è 13/35: si deduce, allora, che per ogni casella esistono due frequenze condizionate relative).

Prima di concludere questa sottofase è opportuno far rilevare che se la raccolta è effettuata dalla Pubblica Amministrazione, i dati vengono posti negli *archivi di informazione* che costituiscono, poi, le fonti d'informazione ufficiali e non.

ESERCIZI DA SVOLGERE

1) Su 100 domande 90 studenti hanno dato le seguenti risposte giuste

74	35	70	84	48	90	82	47	56	68
68	81	73	61	66	96	79	86	65	83
75	88	83	75	82	89	69	67	73	72
73	84	78	65	78	72	62	80	78	67
93	75	71	75	72	60	77	74	74	75
62	90	71	76	76	65	60	69	95	71
88	78	59	76	76	65	74	79	75	87
90	93	77	62	54	86	84	91	93	90
60	68	94	73	75	88	77	72	83	83

a) Costruire la distribuzione di frequenze;
b) calcolare le frequenze cumulate e le frequenze relative cumulate;

c) determinare la percentuale di studenti che hanno ottenuto un numero di risposte giuste inferiore a 70;
d) calcolare la percentuale di studenti che hanno ottenuto un numero di risposte giuste superiore a 90.

2) Dopo aver estratto dall'Annuario Statistico Italiano l'ammontare della popolazione residente in Italia al censimento demografico 1981, in età di 6 anni in poi, secondo il grado di istruzione:

a) individuare il tipo di distribuzione;
b) calcolare le frequenze percentuali;
c) calcolare le frequenze cumulate.

3) Dopo aver rilevato dall'Annuario Statistico Italiano, il numero di famiglie residenti in Italia, al censimento demografico 1981, per numero di componenti:

a) individuare il tipo di distribuzione;
b) calcolare le frequenze percentuali;
c) calcolare le frequenze cumulate.

4) Dall'Annuario Statistico Italiano si estragga la serie storica dei procedimenti di scioglimento di matrimonio in Italia esauriti in fase ordinaria.

5) Dall'Annuario Statistico Italiano ricavare il numero delle abitazioni, censite il 25 ottobre 1981, occupate in Puglia secondo il titolo di godimento (proprietà, affitto, altro titolo). Calcolare:

a) le frequenze cumulate;
b) le frequenze relative;
c) le frequenze relative cumulate.

6) Allo scopo di accertare se la professione del padre influisce sul rendimento scolastico degli alunni, è stata condotta un'indagine su un campione di 60 alunni scelti a caso fra gli esaminati alla licenza di scuola media in un Istituto di una città.

Le professioni dei padri sono state classificate come segue:

— imprenditore e libero professionista (x_1),
— dirigente e impiegato (x_2),
— lavoratore in proprio (x_3),
— operaio e assimilato (x_4).

I gradi di rendimento scolastico sono i seguenti:

— insufficiente (y_1),
— sufficiente (y_2),
— discreto (y_3),
— buono (y_4),
— ottimo (y_5).

I risultati sono riportati nella seguente tabella.

Alunno	Professione paterna X	Grado di rendimento Y	Alunno	Professione paterna X	Grado di rendimento Y	Alunno	Professione paterna X	Grado di rendimento Y
1	x_1	y_2	21	x_4	y_2	41	x_3	y_2
2	x_2	y_2	22	x_4	y_4	42	x_3	y_2
3	x_3	y_4	23	x_3	y_3	43	x_3	y_4
4	x_2	y_5	24	x_3	y_3	44	x_4	y_5
5	x_4	y_3	25	x_3	y_3	45	x_2	y_4
6	x_3	y_3	26	x_4	y_3	46	x_1	y_5
7	x_2	y_2	27	x_3	y_2	47	x_1	y_4
8	x_2	y_5	28	x_3	y_2	48	x_2	y_4
9	x_4	y_2	29	x_2	y_2	49	x_2	y_4
10	x_3	y_4	30	x_4	y_1	50	x_2	y_3
11	x_4	y_4	31	x_3	y_2	51	x_3	y_3
12	x_1	y_4	32	x_3	y_3	52	x_4	y_3
13	x_1	y_5	33	x_2	y_3	53	x_4	y_3
14	x_1	y_3	34	x_4	y_3	54	x_3	y_4
15	x_2	y_3	35	x_2	y_3	55	x_4	y_3
16	x_4	y_1	36	x_3	y_3	56	x_3	y_3
17	x_2	y_4	37	x_3	y_3	57	x_3	y_3
18	x_4	y_2	38	x_4	y_3	58	x_3	y_2
19	x_4	y_2	39	x_1	y_3	59	x_3	y_2
20	x_4	y_2	40	x_1	y_3	60	x_3	y_2

Costruire la tabella a doppia entrata e calcolare, poi,

a) il numero di coloro che non hanno superato l'esame;
b) il numero degli alunni che hanno conseguito *"ottimo"*;
c) il numero dei figli dei liberi professionisti;
d) il numero dei figli di operai che hanno conseguito "buono".

7) Quaranta persone sono classificate secondo:

— il sesso: Maschio (M), Femmina (F);
— lo stato civile: Celibe o Nubile (N), Coniugato-a (C), Vedovo-a (V), Separato-a o Divorziato-a (S);

— l'istruzione: privo di titolo di studio o con licenza elementare (A), licenza scuola media inferiore (M_i), diploma di scuola media superiore (D), laurea (L).

I dati sono riportati nella seguente tabella.

Persone	Sesso X	Stato civile Y	Istruzione Z	Persone	Sesso X	Stato civile Y	Istruzione Z
1	M	N	A	21	F	S	A
2	M	S	M_i	22	M	V	M_i
3	M	V	D	23	F	N	A
4	F	C	L	24	F	N	A
5	F	N	A	25	M	N	A
6	M	S	A	26	M	C	A
7	F	N	A	27	M	C	M_i
8	F	N	M_i	28	M	N	M_i
9	F	N	M_i	29	M	C	M_i
10	M	V	A	30	F	N	A
11	M	C	M_i	31	M	N	M_i
12	M	C	M_i	32	F	N	A
13	F	C	A	33	F	N	M_i
14	M	N	M_i	34	M	C	D
15	F	N	D	35	M	C	L
16	M	C	D	36	M	V	A
17	M	C	D	37	M	S	M_i
18	M	C	L	38	F	S	M_i
19	M	N	L	39	F	N	M_i
20	M	N	D	40	F	N	A

a) Formare le tabelle a doppia entrata: (sesso - stato civile), (sesso - istruzione), (stato civile - istruzione).

b) Costruire la tabella stato civile - istruzione, secondo il sesso.

c) Quanti sono i maschi?

d) Quante sono le femmine laureate?

e) Quanti sono i maschi divorziati?

f) Quanti sono i divorziati (M e F) che sono forniti di laurea?

g) Quanti sono i maschi celibi forniti di diploma di scuola media superiore?

8) In un campione di 30 persone appartenenti alle forze di lavoro si considerino, nell'ordine, i seguenti caratteri:

— età in anni compiuti (si classifichino i dati nelle classi di età 14-18, 18-25, 25-65);

— condizione professionale: Occupato (O), parzialmente occupato (S), disoccupato (NO);

— settore di attività economica: Agricoltura (A), Industria (I), Altre attività (T);

— posizione nella professione: Imprenditore e Libero professionista (L), Dirigente e Impiegato (D), Lavoratore in proprio (P), Operaio e assimilato (Op).

I dati sono riportati nella seguente tabella.

Persone	Età (in anni compiuti)	Condizione professionale	Settore attività economica	Posizione nella professione	Persone	Età (in anni compiuti)	Condizione professionale	Settore attività economica	Posizione nella professione
1	16	O	A	Op	16	26	O	I	D
2	40	O	I	Op	17	53	O	I	D
3	22	S	A	Op	18	25	NO	I	Op
4	50	O	I	L	19	61	O	T	L
5	60	O	I	D	20	20	S	A	Op
6	30	NO	I	Op	21	49	O	I	Op
7	17	S	A	Op	22	15	S	A	Op
8	63	O	A	L	23	36	O	A	D
9	55	O	A	D	24	24	O	A	Op
10	45	O	T	L	25	21	O	A	Op
11	39	O	A	D	26	23	O	A	D
12	27	NO	I	Op	27	20	O	A	Op
13	16	S	T	Op	28	21	O	I	Op
14	48	O	I	D	29	51	O	I	L
15	45	O	I	Op	30	23	S	T	Op

Costruire, con i dati precedenti, tutte le possibili tabelle a doppia entrata.

CAPITOLO III

RAPPRESENTAZIONE GRAFICA

Quando si osservano tabelle di dati, non è sempre facile avere una visione globale del fenomeno.

La rappresentazione grafica dei dati consente, appunto, di cogliere con immediatezza le principali caratteristiche della distribuzione statistica.

E' ovvio che, perchè un grafico sia comprensibile, deve essere corredato di un titolo (che illustri brevemente quale fenomeno si sta rappresentando) e di una scala di misura necessaria per la lettura del grafico. Sugli assi, inoltre, bisogna indicare il nome delle variabili.

A seconda del modo come sono espressi i dati si avranno rappresentazioni di tipo diverso.

Nelle scienze sociali le principali rappresentazioni grafiche sono:

1. - Il diagramma a settori circolari

Esso si usa per rappresentare le mutabili statistiche non ordinabili e, a volte, anche le serie territoriali.

Consiste nel dividere il cerchio in settori proporzionali all'intensità del fenomeno. Per far ciò si trova l'angolo al centro di ogni settore dalla proporzione $\alpha_i : n_i = 360 : n$. Ove, come al solito, n_i è la frequenza (o intensità) che corrisponde alla modalità x_i ed $n = \sum_{i=1}^{s} n_i$. Nell'esempio della Fig. III/1 è stato riportato il diagramma a settori circolari relativo alle percentuali degli occupati in Italia nel 1976, secondo l'attività economica.

2. - Il cartogramma

Questo tipo di grafico è molto usato nel caso delle serie territoriali. Ad es., per rappresentare gli indici di diffusione

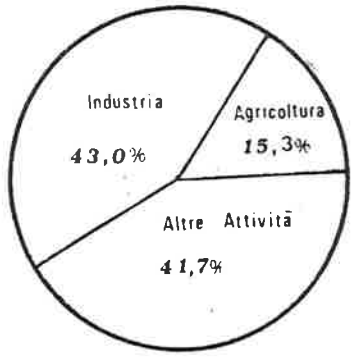


Fig. III/1 - Occupati in Italia nel 1976, secondo l'attività economica.

della televisione (abbonamenti su 100 famiglie) nelle varie regioni italiane, riportati nella Tav. II/6, si colorano (o si tratteggiano) le regioni secondo l'intensità della diffusione.

E' logico che, per la corretta interpretazione, bisogna riportare a fianco l'intensità che corrisponde ad una prefissata colorazione (o tratteggio) (Fig. III/2).

3. - L'ortogramma

E' un tipo di rappresentazione tramite rettangoli, che si usa spesso per rappresentare m.s. non ordinabili, o, anche, per comparare fenomeni relativi a serie territoriali. I rettangoli hanno tutti la stessa base, ma hanno un'altezza proporzionale all'intensità (o frequenza) del fenomeno. Un esempio di ortogramma è riportato nella Fig. III/3. In detta figura sono rappresentati i consumi di proteine animali e di proteine totali in grammi/abitante/giorno, per grandi regioni o gruppi di paesi (Anno 1970).

4. - Il diagramma a scala naturale

E' un tipo di grafico che si adatta molto bene alla rappresentazione di v. s. continue e di serie storiche.

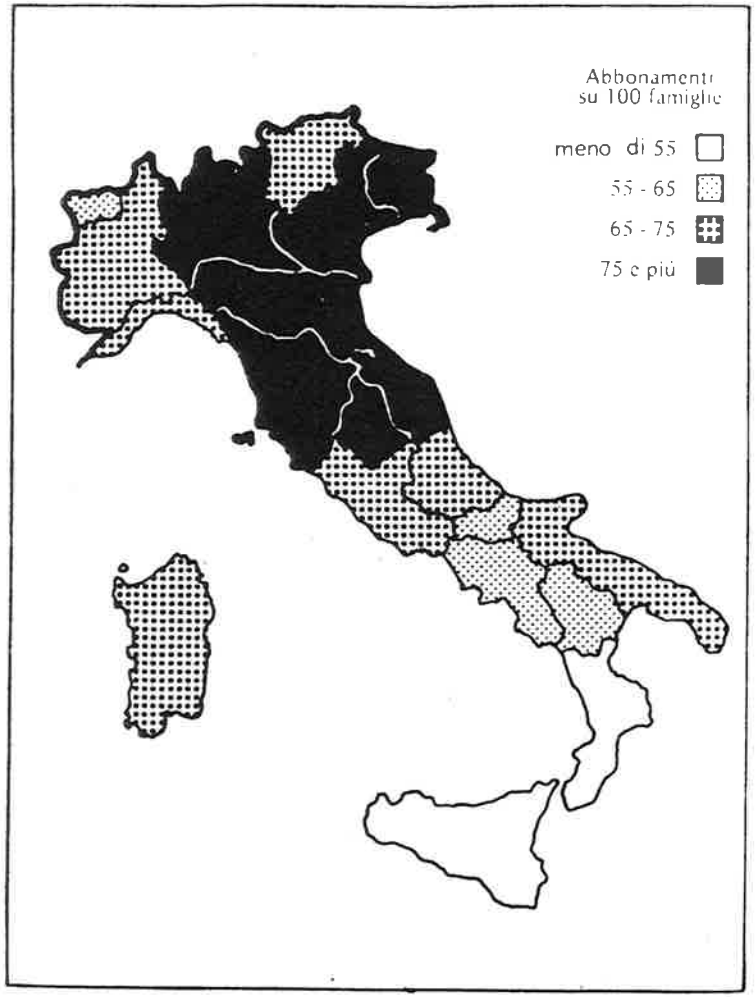


Fig. III/2 - Indici di diffusione della televisione nelle regioni italiane al 1974. Fonte: M. Lo Presti, *Aspetti statistici del fenomeno televisivo in Puglia*, Rassegna Economica, n. 5, 1975.

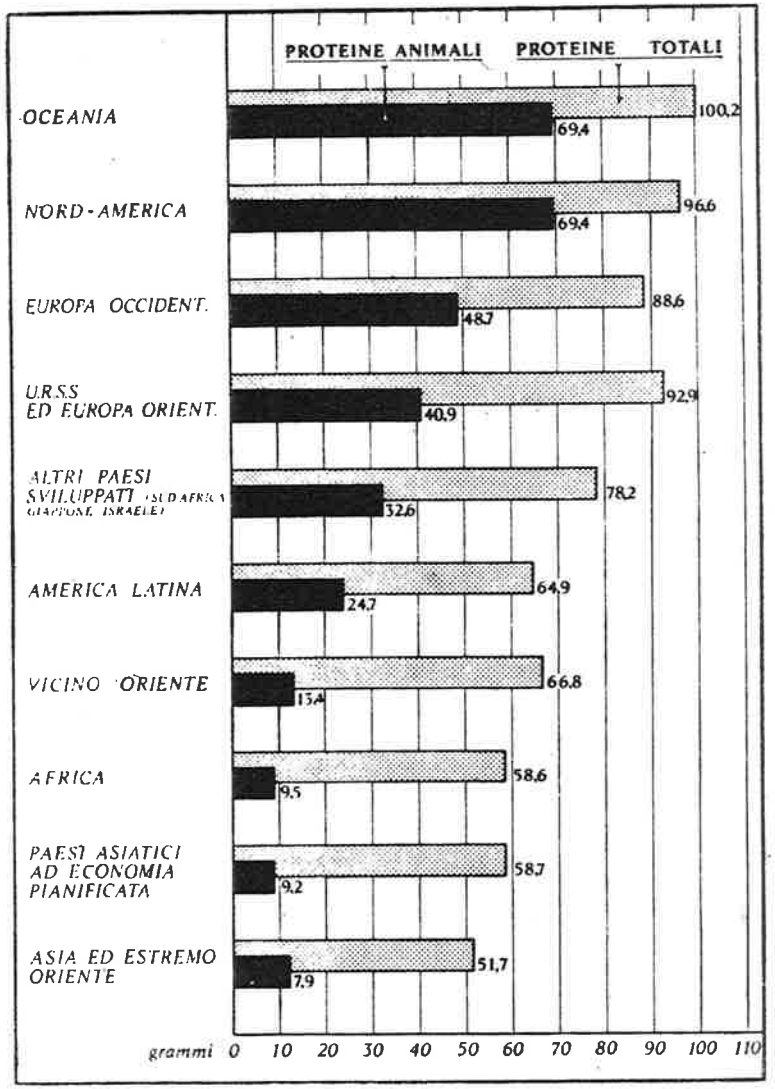


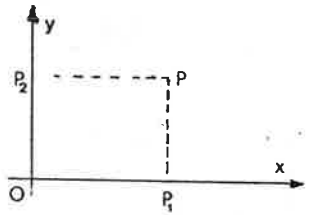
Fig. III/3 - Consumi di proteine animali e di proteine totali in grammi/abitante/giorno, per grandi regioni o gruppi di paesi (Anno 1970). Fonte: G. Galeotti, *Popolazione e ambiente*, Cacucci editore, Bari, 1974.

Per rappresentare i diagrammi cartesiani si utilizza un *sistema di riferimento cartesiano ortogonale.*

Detto sistema è costituito da due assi orientati e perpendicolari (quello orizzontale orientato da sinistra a destra, quello verticale dal basso verso l'alto) e da due unità di misura (la prima per misurare le modalità e la seconda per misurare le frequenze o le intensità).

L'incontro dei due assi è l'origine O del sistema.

Vediamo in che modo si individua la posizione di un punto P nel piano.



Fissato il punto P nel piano, siano P_1 e P_2 le sue proiezioni sugli assi.

Dicesi *ascissa* x di P la misura del segmento orientato $\overline{OP_1} = \overline{P_2\,P}$ (tale ascissa è positiva se $\overline{OP_1}$ è nel verso della freccia, è negativa se per andare da O verso P_1 si procede nel verso contrario a quello della freccia); dicesi *ordinata* y di P la misura del segmento $\overline{OP_2} = \overline{P_1\,P}$ (tale ordinata è positiva o negativa se per andare da O verso P_2 si procede nel verso della freccia o in quello contrario): sicché dato P ad esso sono associati due numeri reali, l'ascissa e l'ordinata di P. Viceversa, data una coppia (x,y) di numeri reali si può individuare il punto P nel piano: si riporta l'ascissa x sull'asse orizzontale (che perciò vien detto asse x o delle ascisse) e dal punto P_1 così determinato si manda la perpendicolare all'asse x, si riporta l'ordinata y sull'asse verticale (che perciò viene denominato asse y o delle ordinate) e dal punto P_2 così individuato si manda la parallela all'asse x: il punto P è dato dalla intersezione delle due rette anzidette. In questo modo si è stabilita una corrispondenza biunivoca fra punti del piano e coppie ordinate di numeri reali.

L'ascissa e l'ordinata si chiamano *coordinate* del punto P.

Per dire che questo è individuato dalle coordinate (x,y), si scrive P (x,y).

Tutto ciò premesso il *diagramma a scala naturale* si ottiene congiungendo, in un sistema di riferimento cartesiano ortogonale, i punti P_i (x_i,n_i) che si ottengono al variare di i, ove x_i è la generica modalità della v.s. assegnata ed n_i è la frequenza (o intensità) che compete ad x_i.

Tale tipo di rappresentazione presuppone che la rilevazione ha senso in ogni istante e che nell'intervallo fra due ascisse il fenomeno varî in maniera uniforme.

Ad es., le temperature (in gradi centigradi) osservate nelle varie ore del giorno 5 luglio 1978 all'osservatorio dell'Aeronautica di Palese (BA), sono fornite dalla seguente distribuzione.

$$\begin{pmatrix} \text{ore } 8 & \text{ore } 12 & \text{ore } 16 & \text{ore } 20 \\ 21^0 & 28^0 & 30^0 & 24^0 \end{pmatrix}.$$

Poichè la temperatura varia con continuità, la rappresentazione grafica del fenomeno è data dal diagramma a scala naturale della Fig. III/4.

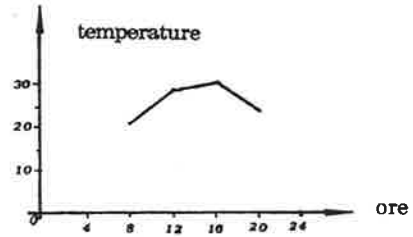


Fig. III/4 - Andamento della temperatura durante le ore del giorno.

Questi tipi di grafici si usano in medicina per vedere come varia la temperatura corporea degli ammalati.

Esistono, poi, anche strumenti che riescono automaticamente a tracciare tali grafici: si pensi, ad es., agli strumenti che misurano l'andamento delle maree, che tracciano i cardiogrammi, gli encefalogrammi, ecc..

5. - Il diagramma a segmenti

Quando la variabile non è continua, allora la rappresentazione grafica è come quella della Fig.III/5: in detta figura è rappresentata la v.s. della Tav. III/1.

Tav. III/1 - Appartamenti secondo il numero di stanze.

Stanze (x_i)	N. di appartamenti (n_i)
2	7
3	15
4	12
5	8
6	3

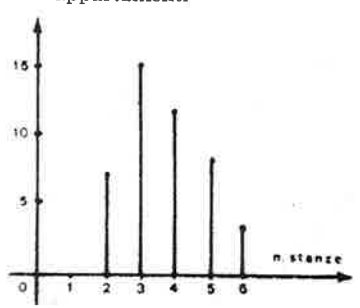


Fig. III/5 - Appartamenti secondo il numero di stanze.

6. - L'istogramma

E' una rappresentazione tramite rettangoli e si usa quando vogliamo rappresentare caratteri continui (come il tempo, l'età, la statura, ecc.) o semicontinui (come il reddito, il salario, ecc.), le cui modalità sono raggruppate in classi aventi determinata ampiezza.

Le frequenze o le intensità delle varie modalità sono rispettivamente uguali alle *aree* dei rettangoli aventi per base i segmenti che indicano le ampiezze delle classi corrispondenti. Il rapporto fra la frequenza e l'ampiezza della base dà ciò che si chiama la *frequenza unitaria* o *densità di frequenza:* la densità di frequenza è, quindi, uguale all'altezza del rettangolo. Cioè,

$$h_i = \frac{n_i}{d_i} .$$

Esempio:

La distribuzione degli operai di una data azienda secondo l'età, è data dalle prime due colonne della Tav. III/2.

Nella terza colonna sono riportate le ampiezze delle classi, mentre nella quarta colonna sono riportate le densità di frequenza che risultano, rispettivamente:

60 : 4 = 15, 116 : 4 = 29, 200 : 8 = 25, ecc..

L'istogramma è quello della Fig. III/6.

Tav. III/2 - Operai di una azienda secondo l'età.

Età (x_i) in anni compiuti	Numero di operai (n_i)	d_i	$h_i = \frac{n_i}{d_i}$	N_i
14 - 17	60	4	15	60
18 - 21	116	4	29	176
22 - 29	200	8	25	376
30 - 39	230	10	23	606
40 - 49	150	10	15	756
50 - 64	150	15	10	906

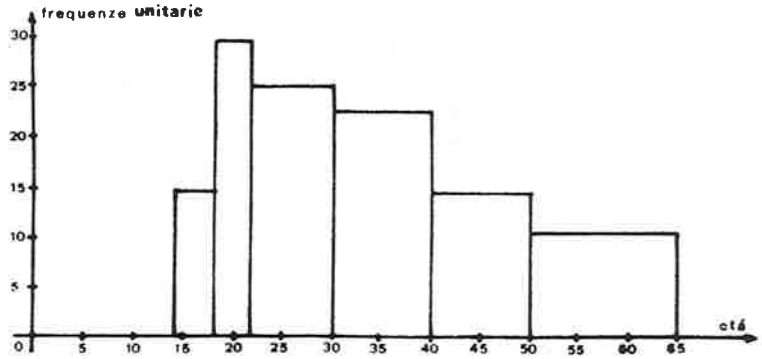


Fig. III/6 - Operai di una azienda secondo l'età.

7. - Il diagramma polare

I *diagrammi a coordinate polari* si usano per rappresentare le mutabili cicliche.

Si fissa nel piano un punto O che chiamiamo *polo* e una semiretta uscente da O che chiamiamo *asse polare.* Si assume come unità di misura delle intensità o frequenze un determinato segmento e come unità di misura degli angoli il grado sessagesimale (1°) e si fissa il verso positivo (verso antiorario) delle rotazioni dell'asse polare intorno ad O. Ogni punto del piano è individuato da due coordinate: il *raggio vettore,* (che rappresenta la distanza del punto P da O) e l'*anomalia* o *argomento* (che è l'angolo formato dall'asse polare col raggio vettore). Generalmente si indica il raggio vettore con ρ (leggi: *ro*), e l'anomalia con φ (leggi: *fi*).

Per rappresentare la mutabile ciclica si divide (con delle semirette) l'angolo giro in tante parti quante sono le modalità della mutabile statistica, e su ciascuna semiretta si riporta un segmento $\rho_i = n_i$ (uguale cioè alla frequenza o intensità del fenomeno). Congiungendo i vari punti si ottiene il *diagramma polare.*

Il diagramma polare della Fig. III/7 rappresenta la mutabile statistica della Tav. II/3. In esso l'angolo che ogni semiretta forma con la semiretta consecutiva è

$$\varphi = \frac{360^0}{7} \simeq 51^0 25'$$

Qualche volta, con centro in O, si traccia una circonferenza con raggio uguale alla media aritmetica μ del fenomeno: in tal modo si può vedere quali sono i giorni in cui l'intensità o la frequenza del fenomeno eccede oppure è inferiore alla media μ.

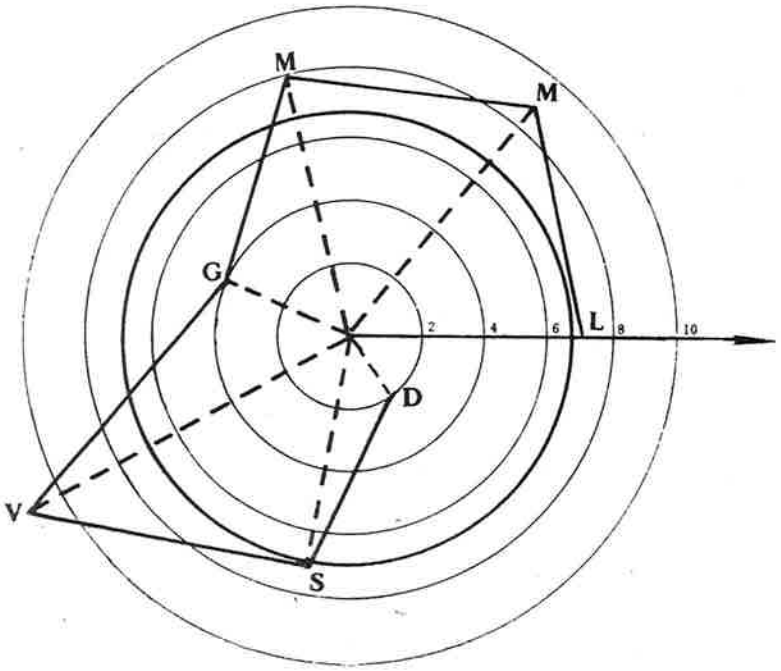


Fig. III/7 - Personale assente, nella prima settimana del mese di maggio del 1986, in un complesso alberghiero

8. - Confronto di grafici

Per rendere agevole il confronto di grafici bisogna cercare di utilizzare le stesse unità di misura, ma tratteggio (o colore) diverso.

In particolare, i diagrammi si disegnano nello stesso sistema di assi di riferimento, con tratteggio (o colore) diverso: in tal caso a fianco di ogni diagramma bisogna indicare il fenomeno al quale la curva si riferisce, oppure occorre riportare la legenda per ogni tratteggio o colore usato.

Naturalmente se la rappresentazione non dovesse risultare chiara perchè i fenomeni (che si stanno riportando in grafico) si accavallano e si aggrovigliano, allora è consigliabile usare assi di riferimento diversi. Un esempio è riportato in Fig. III/8.

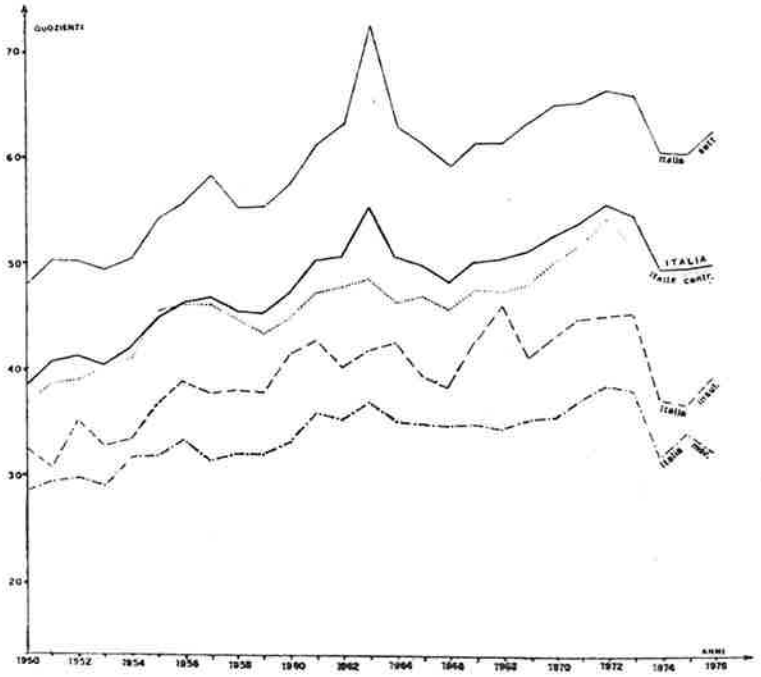


Fig. III/8 - Morti per accidenti in Italia per 100.000 persone residenti secondo le circoscrizioni geografiche. (Fonte: *F. Delvecchio - A. Piccinno*, La mortalità per accidente in Italia, "Studi di Demografia", Quaderno n. 19, Cacucci Editore, Bari, 1979).

9. - Diagrammi integrali

Il *diagramma integrale* è quello che si ottiene rappresentando graficamente la funzione di distribuzione.

Se il carattere è continuo e le modalità sono raggruppate in classi e se x_0, x_1, ..., x_{i-1}, x_i, ..., x_s sono i punti di divisione delle classi, tale diagramma si ottiene congiungendo i punti di coordinate $P_i\ (x_i\ ,\ F_i)$, punti, cioè, che hanno come ascissa l'estremo superiore delle classi e come ordinata le frequenze relative cumulate. E' evidente che così facendo si ipotizza che all'interno di ciascuna classe le frequenze si distribuiscono uniformemente. Si noti che ad x_0 corrisponde un'ordinata pari a 0, visto che la frequenza dei valori minori di x_0 è nulla.

Ad es., il diagramma integrale della distribuzione riportata nella Tav. III/2 è quello riportato nella Fig. III/9.

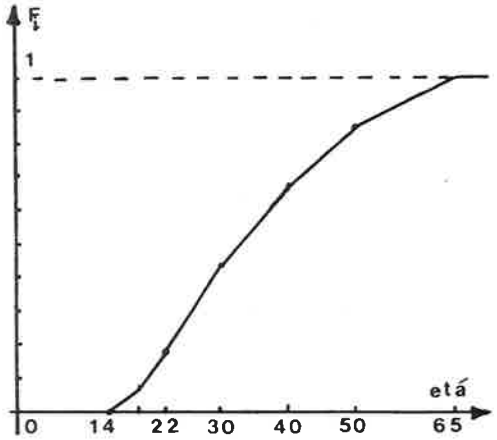


Fig. III/9 - Diagramma integrale degli operai di un'azienda secondo l'età.

Fig. III/10 - Diagramma integrale degli appartamenti secondo il numero di stanze.

Se, invece, il carattere è discreto, dal punto $P_i\ (x_i\ ,\ F_i)$ si manda un segmento parallelo all'asse x di lunghezza pari alla ampiezza dell'intervallo $(x_i\ ,\ x_{i+1})$.

Ad es., il diagramma integrale della distribuzione riportata nella Tav. III/1 è quello riportato in Fig. III/10.

10. - Poligono di frequenze. Curva di frequenze

Nel caso di v.s. continue, congiungendo fra loro i punti centrali delle basi superiori dei rettangoli di un istogramma si ottiene una spezzata che viene denominata *poligono di frequenze*.

Rispetto all'istogramma il poligono di frequenze ha il vantaggio di poter rappresentare più facilmente, nello stesso grafico, più di un fenomeno.

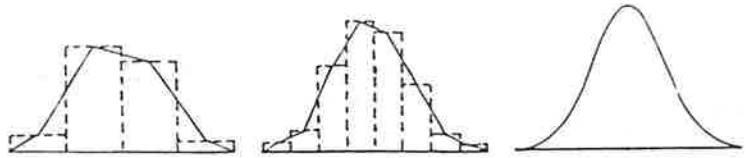
Va osservato, però, che è possibile confrontare più poligoni di frequenze quando le distribuzioni hanno le stesse classi e lo stesso numero di casi. Se, invece, hanno le stesse classi ma un numero di casi diverso, per il confronto occorre riferirsi a frequenze relative.

Dalla figura appare evidente, poi, che se rimpiccioliamo sempre più le classi ed aumentiamo sempre più il numero di osservazioni, e ciò perché conservi significato la frequenza dei casi che cadono in ciascuna classe (naturalmente ci riferiamo a frequenze relative), il poligono di frequenze tende a trasformarsi in una curva continua detta *curva di frequenze*.

Questo risultato è importante ai fini di trovare la curva teorica che può essere considerata l'andamento della distribuzione del carattere nell'universo, del quale le osservazioni riportate nella distribuzione empirica rappresentano un campione.

Per concludere diciamo che, se ci riferiamo a frequenze relative, anche l'area sottesa al poligono di frequenze o alla curva di frequenze è uguale ad 1.

Va chiarito, inoltre, che l'ordinata di un punto della spezzata (o della curva) di frequenze non sta ad indicare la frequenza che spetta alla modalità del fenomeno corrispondente all'ascissa di quel punto, in quanto in questo tipo di grafico le frequenze sono rappresentate da aree: perciò la frequenza dei casi che cadono in una certa classe è data dall'area sottesa al poligono (o alla curva) di frequenze in corrispondenza di quella classe.



ESERCIZI DA SVOLGERE

Si desumano i dati relativi agli esercizi di questo capitolo dall'Annuario Statistico Italiano o dal Compendio Statistico Italiano.

1) Si costruisca il grafico relativo alla popolazione residente in Italia, di 6 anni e oltre, per grado di istruzione (tale popolazione è stata già rilevata all'esercizio 2 del capitolo precedente).

2) Si costruisca il grafico relativo alle famiglie residenti in Italia per numero di componenti (tale distribuzione è stata già rilevata all'esercizio 3 del capitolo precedente).

3) Si costruisca il grafico relativo alle percentuali delle abitazioni occupate in Puglia, secondo il titolo di godimento (tale distribuzione è stata già rilevata all'esercizio 5 del capitolo precedente).

4) Si costruisca il grafico della serie storica della popolazione residente in Italia a fine anno a partire dal 1971.

5) Si costruisca il grafico della popolazione scolastica italiana secondo il tipo di scuola (materna, elementare, scuola media, scuola secondaria superiore).

6) Si costruisca il grafico delle percentuali della spesa del pubblico per spettacoli in Italia, secondo il tipo di spettacolo (teatro, cinematografo, trattenimenti vari, manifestazioni sportive).

7) Si costruisca il grafico relativo ai detenuti condannati in Italia, secondo la pena inflitta.

8) Si costruisca il grafico relativo agli entrati negli istituti per minori, secondo il grado di istruzione.

9) Costruire il cartogramma della serie territoriale, relativa alle regioni italiane, dei procedimenti di scioglimento di matrimonio, esauriti in fase ordinaria, ogni 1000 famiglie (ottenuta, cioè, rapportando il numero dei procedimenti di scioglimento di matrimonio esauriti in fase ordinaria in un anno al numero delle famiglie [dello stesso anno e della medesima regione] e moltiplicando il risultato per 1000).

10) Costruire il cartogramma della serie territoriale relativa alle regioni italiane, dei matrimoni per 100.000 abitanti.

11) Costruire il diagramma polare della serie mensile dei suicidi in Italia.

12) Costruire il grafico delle giornate di degenza per posto letto.

13) Costruire l'istogramma, per maschi e per femmine, della popolazione italiana censita il 24 ottobre 1971, per le classi di età: meno di 6, 6-14, 14-21, 21-25, 25-45, 45-65, 65 e più.

14) Costruire il diagramma polare degli incidenti stradali per mese.

CAPITOLO IV

ELABORAZIONE DEI DATI

L'elaborazione dei dati consiste in un insieme di tecniche che permettono l'utilizzazione delle informazioni.

Noi, di queste tecniche, tratteremo solo quelle di più largo uso nelle Scienze sociali.

1. - Valori medi

I valori medi hanno il vantaggio di sintetizzare le modalità delle distribuzioni considerate. Fra questi i più usati nelle ricerche sociali sono:

a) *La media aritmetica*

La media aritmetica, che d'ora in poi chiameremo *media* e che indicheremo con μ, è quel valore che sintetizza il carattere di un collettivo statistico in modo che, sostituito ai singoli termini, rimanga invariato l'ammontare del carattere stesso: cioè, se ad ogni termine si sostituisce μ, l'ammontare del carattere rimane lo stesso. Se il carattere è X, la media si indica anche con $\bar{x}$ oppure con $\mu(X)$; mentre, se il carattere è Y, la media si indica con $\bar{y}$ oppure con $\mu(Y)$; ecc..

La denominazione di media aritmetica deriva dal fatto che μ è il termine centrale di un numero dispari di elementi in progressione aritmetica (tali, cioè, che la differenza fra ogni termine ed il precedente sia costante).

Ad es., tre individui hanno percepito, nella prima settimana di maggio del 1973, i seguenti salari settimanali (in lire)

60.000, 65.000, 70.000.

La *media* μ deve essere tale che

$$60.000 + 65.000 + 70.000 = \mu + \mu + \mu$$

da cui si ricava che

$$\mu = \frac{60.000 + 65.000 + 70.000}{3} = 65.000.$$

Sicchè, in generale, sempre nel caso di singoli valori,

$$\mu = \bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_s}{s} = \frac{\sum_{i=1}^{s} x_i}{s}. \qquad [1]$$

Nel caso non si abbiano valori singoli di salari (in lire), ma, ad es., si abbia la distribuzione di frequenze

$$\begin{pmatrix} 60.000 & 65.000 & 70.000 \\ 3 & 4 & 3 \end{pmatrix},$$

il salario complessivo percepito da quei dieci individui è dato da

$$60.000 \cdot 3 + 65.000 \cdot 4 + 70.000 \cdot 3,$$

per cui il salario medio deve essere tale che

$$60.000 \cdot 3 + 65.000 \cdot 4 + 70.000 \cdot 3 = 10\ \mu$$

e, quindi,

$$\mu = \frac{60.000 \cdot 3 + 65.000 \cdot 4 + 70.000 \cdot 3}{3 + 4 + 3} = \frac{650.000}{10} = 65.000.$$

In generale, allora, per una distribuzione di frequenze, la media è data da

$$\mu = \bar{x} = \frac{x_1 n_1 + x_2 n_2 + \ldots\ldots + x_s n_s}{n_1 + n_2 + \ldots\ldots + n_s} = \frac{\sum_{i=1}^{s} x_i n_i}{\sum_{i=1}^{s} n_i}. \qquad [2]$$

Si noti che questa formula equivale alla precedente nel caso che gli n_i siano uguali ad 1. Diciamo, ancora, che la media aritmetica si usa quando le modalità del carattere sono quantitative.

Prima di procedere oltre è opportuno far osservare che la media aritmetica è espressa nella stessa unità di misura adoperata per le x_i: se si considera, quindi, una distribuzione di individui secondo le stature espresse in cm, anche la media è espressa in cm; se invece si tratta di una distribuzione di individui secondo le età espresse in anni, anche la media di tale distribuzione è espressa in anni.

Diciamo *scarto* o *scostamento* dalla media aritmetica la differenza fra il valore x_i e la media μ, cioè

$$e_i = x_i - \mu \quad .$$

La media aritmetica gode di alcune proprietà, di cui le più importanti sono:

1) *La somma algebrica degli scarti* $x_i - \mu$ *è uguale a zero.*

Infatti, dalla [2] si ricava

$$x_1 n_1 + x_2 n_2 + \ldots\ldots + x_s n_s = \mu\ (n_1 + n_2 + \ldots\ldots + n_s).$$

Da cui, portando tutto al primo membro, si ottiene

$$(x_1 - \mu)\, n_1 + (x_2 - \mu) n_2 + \ldots + (x_s - \mu) n_s = \sum_{i=1}^{s} (x_i - \mu)\ n_i = 0.$$

Questa proprietà è caratteristica della media aritmetica nel senso che, se μ è la media aritmetica, $\Sigma\ (x_i - \mu) n_i = 0$; viceversa se $\Sigma\ (x_i - k) n_i = 0$, allora $k = \mu$.

Verifichiamo la 1ª proprietà considerando il caso della distribuzione di frequenze precedente (cioè quella riguardante il salario di dieci individui). Si ha, appunto,

$$(60.000 - 65.000)3 + (65.000 - 65.000)4 + (70.000 - 65.000)3 = 0.$$

2) *La somma dei quadrati degli scarti* $x_i - \mu$ *è un minimo,* ossia è minore della somma dei quadrati degli scarti fra le x_i ed un valore k diverso da μ. Infatti,

$$\Sigma\ (x_i - k)^2 n_i = \Sigma\ (x_i - \mu + \mu - k)^2 n_i = \Sigma[(x_i - \mu)^2 + 2\ (\mu - k)\ (x_i - \mu) + (\mu - k)^2]\ n_i = \Sigma\ (x_i - \mu)^2\ n_i + 2\ (\mu - k)\ \Sigma\ (x_i - \mu) n_i + (\mu - k)^2\ \Sigma\ n_i = \Sigma\ (x_i - \mu)^2\ n_i + n\ (\mu - k)^2,$$

in quanto, per la 1ª proprietà, $\Sigma\ (x_i - \mu) n_i = 0$ ed, inoltre, $\Sigma\ (\mu - k)^2 n_i = (\mu - k)^2\ \Sigma\ n_i$.

La relazione scritta dimostra, appunto, la 2ª proprietà della media. Infatti,

$$\Sigma\ (x_i - k)^2\ n_i > \Sigma\ (x_i - \mu)^2\ n_i$$

perché $n\ (\mu - k)^2$ è una quantità positiva.

Anche questa proprietà è caratteristica della media aritmetica nel senso che, se μ è la media aritmetica, $\Sigma\ (x_i - \mu)^2\ n_i$ = minimo; viceversa, se $\Sigma\ (x_i - k)^2\ n_i$ = minimo, allora $k = \mu$.

Verifichiamo la 2ª proprietà considerando la seguente distribuzione di stature (in cm)

$$\begin{pmatrix} 170 & 172 & 174 \\ 2 & 4 & 2 \end{pmatrix}.$$

La media è

$$\mu = \frac{170 \cdot 2 + 172 \cdot 4 + 174 \cdot 2}{8} = 172.$$

Se supponiamo, ad es., k = 173, si ha

$$\sum_{i=1}^{3} (x_i - \mu)^2 n_i = (170-172)^2 \cdot 2 + (172-172)^2 \cdot 4 + (174-172)^2 \cdot 2 = 16,$$

$$\sum_{i=1}^{3} (x_i - k)^2 n_i = (170-173)^2 \cdot 2 + (172-173)^2 \cdot 4 + (174-173)^2 \cdot 2 = 24,$$

e ciò prova la proprietà anzidetta.

3) *La media aritmetica è associativa*, nel senso che se la v.s. viene divisa in k gruppi e se indichiamo con $\mu_1, \mu_2, \ldots, \mu_k$ le medie parziali dei singoli gruppi e con $n_1, n_2, \ldots, n_k$ le frequenze, la media generale μ può ottenersi come media ponderata delle medie parziali. Cioè

$$\mu = \frac{\mu_1 n_1 + \mu_2 n_2 + \ldots + \mu_k n_k}{n_1 + n_2 + \ldots + n_k}.$$

Questa formula è molto utile perché consente di ricavare la media generale quando si conoscano le medie parziali e non le singole distribuzioni parziali.

Così, ad es., se un gruppo di 12 maschi ha avuto un voto medio di 6 decimi in Storia e un gruppo di 8 femmine ha avuto un voto medio di 7,5 decimi nella stessa disciplina, il voto medio complessivo è

$$\mu = \frac{6 \cdot 12 + 7{,}5 \cdot 8}{20} = 6{,}6.$$

Se la v.s. è divisa in intervalli, per calcolare la media aritmetica μ bisogna distinguere due casi:

a) *Per ogni classe si conosce la frequenza n_i e l'ammontare complessivo A_i del carattere.*

Allora, se indichiamo con μ_i il valore medio (incognito) di quella classe, si ha $n_i \mu_i = A_i$, da cui $\mu_i = A_i/n_i$.

Trovato il valore di μ_i, per calcolare μ si può applicare la proprietà associativa della media, cioè

$$\mu = (\Sigma \mu_i n_i)/n = (\Sigma A_i)/n.$$

ossia la media μ si ottiene dividendo l'ammontare complessivo del carattere $A = \Sigma A_i$ per la frequenza totale $n = \Sigma n_i$.

Si calcoli, ad es., l'ampiezza demografica media dei comuni della provincia di Brindisi al censimento 1981.

Comuni della provincia di Brindisi secondo l'ampiezza demografica, censimento 1981.

Ampiezza demografica (numero di abitanti)	Numero dei comuni (n_i)	Popolazione residente (A_i)
Fino a 5.000	1	4.781
5.001 - 10.000	5	35.706
10.001 - 20.000	8	110.272
20.001 - 50.000	5	150.519
oltre 50.000	1	89.786
Totale	20	391.064

Applicando la formula si ha

$$\mu = (\Sigma A_i)/n = A/n = 391.064/20 \simeq 19.553.$$

Cioè, in media, i comuni della provincia di Brindisi hanno una popolazione residente di 19.553 abitanti.

b) *Per ogni classe si conosce solo la frequenza n_i dei casi che cadono nella classe.*

Questa volta, per giungere alla determinazione di μ, si ipotizza che in ogni intervallo le intensità del carattere siano concentrate nel valore centrale della classe stessa (valore centrale che, com'è noto, è dato dalla semisomma dei valori estremi della classe).

Questa ipotesi equivale ad ipotizzare che all'interno di ogni classe le unità statistiche siano distribuite uniformemente in modo che possano essere considerate addensate nel punto di mezzo della classe.

Diciamo subito che il valore trovato è un valore approssimato; comunque l'approssimazione è tanto più attendibile quanto più piccole sono le ampiezze delle classi: invero classi molto piccole fanno perdere meno informazioni sui dati nella forma originale (cioè nella forma prima del raggruppamento dei dati in classi).

Per calcolare la media aritmetica nel caso di v.s. divise in classi è necessario, dunque, calcolare il valore centrale delle singole classi: occorre, perciò, che siano noti i limiti (inferiore e superiore) di ciascuna classe. Qualche volta, però, le classi estreme non sono limitate (ad es.,

qualche volta la 1ª classe è definita da « meno di x_1 », mentre l'ultima classe è definita da « x_s e più »): in tal caso, risalendo ai dati originari, bisogna limitare le anzidette classi (cioè occorre fissare il limite inferiore per la prima e il limite superiore per l'ultima). Ove ciò non sia possibile, si fissa intuitivamente il valore centrale della classe, possibilmente tenendo conto della struttura del fenomeno.

Per questo motivo, quando si raggruppano i dati in classi, è sempre conveniente formare *classi limitate*.

Si calcoli, ad es., la media aritmetica dei redditi mensili riportati nella tavola seguente.

Numero di redditieri secondo classi di reddito in lire 1983.

Redditi mensili (migliaia di lire 1983)	Redditieri (n_i)	Valori centrali (x_i)	Stima dei redditi delle classi $x_i\, n_i$
Meno di 1.000	50	600	30.000
1.000 |– 1.500	30	1.250	37.500
1.500 |– 2.000	15	1.750	26.250
2.000 e più	5	3.000	15.000
Totale	100		108.750

Abbiamo stimato il valore centrale della prima classe pari a lire 600.000 e il valore centrale dell'ultima classe pari a L. 3.000.000. Ciò posto, risulta

$$\mu = \frac{\Sigma x_i n_i}{\Sigma n_i} = \text{L. } \frac{108.750.000}{100} = \text{L. } 1.087.500.$$

Nel caso si tratti di tabella a doppia entrata si possono calcolare, poi, varie medie. Ad es., riferendoci alla Tav. II/15 del Cap. II, ci si può chiedere: "In media qual è l'età dei laureati in Statistica forniti di maturità classica?" In tal caso si calcola la media degli anni associata alla prima modalità del tipo di maturità, cioè

$$\bar{y}_1 = \frac{22 \cdot 10 + 23 \cdot 20 + 24 \cdot 10 + 25 \cdot 5 + 26 \cdot 5 + 27 \cdot 3 + 30 \cdot 2}{55} = 23{,}927$$

ove invece di "28 anni e più", abbiamo posto "30 anni".

In simboli si ha

$$\bar{y}_1 = \frac{\sum_{h=1}^{t} y_h n_{1h}}{n_{10}}$$

Analogamente, si può chiedere l'età media $\bar{y}_2$ dei laureati in Statistica forniti di maturità scientifica. Si ha

$$\bar{y}_2 = \frac{22 \cdot 8 + 23 \cdot 18 + 24 \cdot 20 + 25 \cdot 13 + 26 \cdot 7 + 27 \cdot 5 + 30 \cdot 5}{76} = 24{,}5,$$

e in simboli

$$\bar{y}_2 = \frac{\sum_{h=1}^{t} y_h n_{2h}}{n_{20}}$$

Analogamente, $\bar{y}_3 = 25{,}6$ $\bar{y}_4 = 26{,}762$, $\bar{y}_5 = 26{,}032$, $\bar{y}_6 = 25{,}5$. Inoltre si può calcolare la media generale per conoscere l'età media dei laureati in Statistica qualsiasi sia la maturità posseduta. Essa si calcola ponderando le età y_h con le frequenze marginali n_{0h} dell'ultima colonna. Si ottiene

$$\bar{y} = \frac{22 \cdot 20 + 23 \cdot 42 + 24 \cdot 39 + 25 \cdot 35 + 26 \cdot 26 + 27 \cdot 17 + 30 \cdot 21}{200} = 24{,}91,$$

ossia

$$\bar{y} = \frac{\sum_{h=1}^{t} y_h n_{0h}}{n}.$$

Come facilmente si verifica la media generale si può anche ottenere ponderando le medie parziali con le frequenze marginali n_{i0} dell'ultima riga della tavola anzidetta. Cioè

$$\bar{y} = \frac{\bar{y}_1 \cdot n_{10} + \bar{y}_2 \cdot n_{20} + \ldots\ldots + \bar{y}_s n_{s0}}{n} = \frac{23{,}927 \cdot 55 + 24{,}5 \cdot 76 + 25{,}6 \cdot 15}{200} +$$

$$+ \frac{26{,}762 \cdot 21 + 26{,}032 \cdot 31 + 25{,}5 \cdot 2}{200} = 24{,}91.$$

In simboli, quindi,

$$\bar{y} = \frac{\sum_{h=1}^{t} y_h n_{0h}}{n} = \frac{\sum_{i=1}^{s} \bar{y}_i n_{i0}}{n}.$$

Analoghe formule valgono per il carattere X se è quantitativo.

b) *La media geometrica*

La media geometrica Mg è quel valore che sintetizza il carattere di un collettivo statistico di s elementi in modo che, sostituito ai singoli termini, rimanga invariato il prodotto delle modalità attribuite ad essi. Cioè Mg deve essere tale che

$$x_1 \cdot x_2 \cdot \ldots \cdot x_s = Mg \cdot Mg \cdot \ldots \cdot Mg = \overset{s}{Mg}$$

da cui

$$Mg = \sqrt[s]{x_1 \cdot x_2 \cdot \ldots \cdot x_s} = \sqrt[s]{\prod_{i=1}^{s} x_i} \qquad [3]$$

(ove il simbolo $\prod_{i=1}^{s}$ sta a significare che bisogna fare il prodotto degli elementi x_i quando i varia da 1 a s).

La denominazione di media geometrica è dovuta al fatto che Mg è il termine centrale di un numero dispari di elementi in progressione geometrica (tali, cioè, che il rapporto di ogni termine al precedente sia costante).

Naturalmente, nel caso si tratti di una distribuzione di frequenze, si ha

$$Mg = \sqrt[\sum_{i=1}^{s} n_i]{x_1^{n_1} \cdot x_2^{n_2} \cdot \ldots \cdot x_s^{n_s}} = \sqrt[\sum_{i=1}^{s} n_i]{\prod_{i=1}^{s} x_i^{n_i}}. \qquad [4]$$

Si può facilmente verificare che, se applichiamo la media aritmetica agli stessi elementi, $Mg \leqslant \mu$ (l'uguaglianza, ovviamente, si ottiene quando gli elementi son tutti uguali tra loro).

Ad es.,

$$Mg\,(1,3,9) = \sqrt[3]{1 \cdot 3 \cdot 9} = 3, \quad \mu\,(1,3,9) = \frac{1+3+9}{3} \simeq 4{,}3.$$

E' evidente che per calcolare il valore di Mg bisogna far uso di calcolatrici, oppure bisogna ricorrere all'uso dei logaritmi [7]. In tal caso, si ha

$$\lg_{10} Mg = \frac{\lg_{10} x_1 + \lg_{10} x_2 + \ldots + \lg_{10} x_s}{s} = \frac{\sum_{i=1}^{s} \lg_{10} x_i}{s}.$$

Trovato il valore dell'espressione precedente, basta calcolare l'antilogaritmo per ottenere il valore di Mg.

Esempio

Se P_0 e P_1 indicano l'ammontare di una popolazione all'anno $t = 0$ e all'anno $t = 1$, dicesi *tasso annuo di sviluppo della popolazione* considerata il numero

$$r = \frac{P_1 - P_0}{P_0}. \qquad [5]$$

Ciò posto, se si indicano con $r_1, r_2, \ldots, r_t$ i tassi di sviluppo della popolazione rispettivamente nel 1°, nel 2°, ..., nel t^0 anno, dalla [5] si deduce

$$P_1 = P_0 + P_0 r_1 = P_0 (1 + r_1)$$

$$P_2 = P_1 + P_1 r_2 = P_1 (1 + r_2) = P_0 (1 + r_1)(1 + r_2)$$

ed, in generale,

$$P_t = P_0 (1 + r_1)(1 + r_2) \ldots (1 + r_t). \qquad [6]$$

Se si vuol trovare il tasso medio annuo r con cui si deve sviluppare la popolazione in modo che dopo t anni abbia l'ammontare espresso dalla [6], deve essere

$$P_0 (1 + r)^t = P_0 (1 + r_1)(1 + r_2) \ldots (1 + r_t), \qquad [7]$$

da cui

$$1 + r = \sqrt[t]{(1 + r_1)(1 + r_2) \ldots (1 + r_t)}. \qquad [8]$$

Il valore di r si ottiene dalla [8], in cui il secondo membro è la media geometrica dei valori $1 + r_1$, $1 + r_2$, ..., $1 + r_t$.

In pratica, si presenta il problema di calcolare l'ammontare della popolazione negli intervalli fra due censimenti conoscendo la popolazione ai censimenti. Occorre, allora, calcolare prima r.

[7] In appendice sono riportati alcuni cenni sui logaritmi.

Se s è l'intervallo tra due censimenti, ricavato il radicando della formula [8] con la [6], si può scrivere

$$1 + r = \sqrt[s]{\frac{P_s}{P_0}} \quad , \qquad [9]$$

da cui si può ricavare r.

Applichiamo quanto detto al calcolo del tasso medio annuo di sviluppo della popolazione residente in Italia fra i due censimenti del 24 ottobre 1971 e del 25 ottobre 1981. Poichè P_{1971} = 54.136.547 abitanti, P_{1981} = 56.556.911 abitanti ed s = 10 anni, si deduce

$$r = \sqrt[s]{\frac{P_s}{P_0}} - 1 = \sqrt[10]{\frac{56.556.911}{54.136.547}} - 1 = 0{,}00438.$$

La popolazione italiana fra il 1971 ed il 1981 si è incrementata di circa il 4,4 per mille abitanti in media all'anno.

Supposto, poi, che dopo il 1981 la popolazione italiana continui a svilupparsi con lo stesso tasso medio annuo, l'ammontare della popolazione al 25/X/1988 si otterrà con la formula

$$P_t = P_0 (1 + r)^t \qquad [10]$$

(ricavata esplicitando P_t dalla [9] e sostituendo *s* con *t*), ove si ponga

$$P_t = P_{1988},\ P_0 = P_{1981},\ t = 7 \text{ ed } r = 0{,}00438.$$

Si ha

$$P_{1988} = 56.556.911\ (1 + 0{,}00438)^7 \text{ abitanti} = 58.313.898 \text{ abitanti.}$$

In tal modo si è effettuata una previsione dell'ammontare della popolazione italiana al 25/X/1988.

c) *Scelta della media. La media armonica*

Quanto detto pone in luce che non sempre si usa la media aritmetica per sintetizzare un fenomeno: la scelta del tipo di media va fatta in relazione al tipo di problema che si sta studiando.

Vediamo qualche altro esempio.

Si calcoli per l'Italia e per il 1981 il consumo finale per residente conoscendo i consumi finali per abitante nelle singole ripartizioni geografiche e la popolazione residente (Tav. IV/1).

Tav. IV/1 - Consumo pro-capite (in lire) e popolazione residente delle **ripartizioni geografiche, 1981.**
(Fonte: Annuario di Contabilità nazionale, vol. XII, Tomo II, 1983).

Ripartizioni geografiche	Consumo finale per abitante (in lire)	Popolazione residente (al 30 giugno)
Italia Settentrionale	6.429.150	25.695.900
Italia Centrale	6.160.716	10.788.892
Italia Meridionale	4.458.452	13.527.432
Italia Insulare	4.813.713	6.489.897

Si indichino con C_i, c_i, P_i, il consumo finale globale, il consumo finale per abitante e la popolazione residente della ripartizione i.ma e con C, c, P le analoghe grandezze per l'Italia.

E' noto inoltre che il consumo finale pro-capite di un territorio è dato dal rapporto fra il consumo finale globale e lo ammontare della popolazione residente di quel territorio, cioè

$$c_i = \frac{C_i}{P_i} \quad , \qquad c = \frac{C}{P} \quad . \qquad [11]$$

Si ha, quindi,

$$c = \frac{C}{P} = \frac{C_1+C_2+C_3+C_4}{P_1+P_2+P_3+P_4} = \frac{c_1P_1+c_2P_2+c_3P_3+c_4P_4}{P_1+P_2+P_3+P_4} = \frac{\Sigma\, c_i\, P_i}{\Sigma\, P_i}$$

(ove, non essendo noti i consumi globali C_i, si è posto C_i = $= c_i P_i$ ricavato dalla [11]).

Il nostro problema si risolve cancolando, dunque, la media aritmetica dei consumi finali pro-capite delle singole ripartizioni stesse:

$$c = \text{L.}\frac{6.429.150 \cdot 25.695.900 + 6.160.716 \cdot 10.788.892}{56.502.121} + \frac{4.458.452 \cdot 13.527.432 + 4.813.713 \cdot 6.489.897}{56.502.121} = \text{L. } 5.720.529$$

cioé, un consumo finale per abitante di L. 5.720.529.

Si calcoli, ora, per l'Italia e per il 1981, il consumo finale

pro-capite conoscendo il consumo finale pro-capite delle singole ripartizioni geografiche e il consumo totale delle stesse ripartizioni (Tav. IV/2).

Tav. IV/2 - Consumo finale totale (in miliardi di lire) e consumo finale per abitante (in lire), 1981.

Ripartizioni geografiche	Consumo per abitante (lire)	Consumo totale (miliardi di lire)
Italia Settentrionale	6.429.150	165.202,8
Italia Centrale	6.160.716	66.467,3
Italia Meridionale	4.458.452	60.311,4
Italia Insulare	4.813.713	31.240,5

Si ha

$$c = \frac{C}{P} = \frac{C_1+C_2+C_3+C_4}{P_1+P_2+P_3+P_4} = \frac{C_1+C_2+C_3+C_4}{\frac{C_1}{c_1}+\frac{C_2}{c_2}+\frac{C_3}{c_3}+\frac{C_4}{c_4}} = \frac{\Sigma C_i}{\Sigma \frac{C_i}{c_i}}, \quad [12]$$

(ove, non essendo note le popolazioni P_i, si è posto $P_i = \frac{C_i}{c_i}$):

cioé, il nostro problema è risolto calcolando la media armonica dei consumi pro-capite delle singole ripartizioni geografiche (c_i) con pesi uguali ai consumi globali delle ripartizioni stesse (C_i). Siamo giunti, così, al concetto di un nuovo tipo di media.

In generale, assegnata la v. s.

$$\begin{pmatrix} x_1 & x_2 & \dots & x_s \\ n_1 & n_2 & \dots & n_s \end{pmatrix},$$

dicesi *media armonica* [8] quel valore M_a che sostituito ai singoli

[8] M_a è detta *media armonica* perchè è il termine centrale dei numeri 6, 4, 3, secondo i rapporti dei quali devono stare le lunghezze delle corde musicali per formare una certa armonica.

termini della distribuzione lascia invariata la somma dei reciproci dei termini stessi. Cioè,

$$\frac{n_1}{x_1} + \frac{n_2}{x_2} + \dots + \frac{n_s}{x_s} = \frac{n_1}{M_a} + \frac{n_2}{M_a} + \dots + \frac{n_s}{M_a},$$

da cui $\Sigma \frac{n_i}{x_i} = \frac{\Sigma n_i}{M_a}$ e, quindi,

$$M_a = \frac{\Sigma n_i}{\Sigma \frac{n_i}{x_i}},$$

che è analoga alla [12].

Nel caso dell'esempio si ha, dunque,

$$c = \text{L.}\frac{165.202,8 + 66.467,3 + 60.311,4 + 31.240,5}{\frac{165.202,8}{6.429.150} + \frac{66.467,3}{6.160.716} + \frac{60.311,4}{4.458.452} + \frac{31.240,5}{4.813.713}} = \text{L. } 5.720.529$$

Si può concludere allora dicendo che il tipo di media da usare è in relazione a una certa proprietà delle v.s. che vogliamo sintetizzare.

d) *La mediana*

La mediana (M_e), dopo aver ordinato i dati in modo non decrescente, è quel valore che lascia alla sua sinistra ed alla sua destra un ugual numero di termini. Essa, nelle Scienze sociali, si usa generalmente quando il carattere è qualitativo ordinabile.

Così, ad es., se si hanno 51 operai distribuiti secondo la graduatoria crescente di salari, il salario mediano è quello posseduto dal 26° operaio. Se, invece, gli operai sono in numero pari il salario mediano si ottiene dalla media aritmetica dei due salari che stanno al centro. Ad es., se gli operai sono 50

il salario mediano è la media aritmetica dei salari posseduti dal 25° e dal 26° operaio.

Si consideri, ora, la distribuzione $\begin{pmatrix} 2 & 3 & 4 & 5 \\ 2 & 4 & 3 & 2 \end{pmatrix}$. Essa può scriversi 2,2,3,3,3,3,4,4,4,5,5, da cui si vede che la modalità mediana è 3: appare evidente, allora, che la mediana M_e di una v.s. è quella modalità x_i a cui corrisponde la 1ª frequenza cumulata $N_i > n/2$; se $N_i = n/2$ allora $M_e = (x_i + x_{i+1})/2$.

Se la v.s. è divisa in intervalli, la classe mediana $[x_{m-1}, x_m[$ è quella per cui $N_m \geqslant n/2$. Ad es., la classe mediana della v.s. di cui alla Tav. III/2 è [30,40[anni, in quanto la 1ª frequenza cumulata maggiore di $n/2 = 453$ appartiene a quella classe. Per calcolare, poi, la mediana M_e di tale v.s., si ipotizza che all'interno della classe mediana il fenomeno si distribuisca uniformemente: si ipotizza, cioè, che le differenze fra il valore mediano M_e e i punti estremi x_{m-1} e x_m dell'intervallo mediano siano proporzionali alle differenze fra le frequenze cumulate corrispondenti a quei punti. Si perviene, allora, alla proporzione

$$(M_e - x_{m-1}) : (x_m - x_{m-1}) = \left(\frac{n}{2} - N_{m-1}\right) : (N_m - N_{m-1})$$

da cui

$$M_e = x_{m-1} + \frac{\left(\frac{n}{2} - N_{m-1}\right)\ (x_m - x_{m-1})}{N_m - N_{m-1}}\ .$$

Nel caso dell'esempio precedente, poichè $x_{m-1} = 30$, $x_m = 40$, $\frac{n}{2} = 453$, $N_{m-1} = 376$, $N_m = 606$, il valore mediano è

$$M_e = \left[30 + \frac{(453 - 376)\ \ (40 - 30)}{606 - 376}\right] \text{ anni } \simeq 33{,}34 \text{ anni, cioè,}$$

è circa 33 anni e 4 mesi.

Se si considerano, ora, gli scarti $x_i - M_e$ dei valori x_i dalla mediana, è facile verificare la seguente proprietà: *la somma dei valori assoluti dei suddetti scarti è un minimo,* nel senso che se K è un valore diverso dalla mediana si ha

$$\Sigma\ |x_i - M_e| < \Sigma\ |x_i - K|\ .$$

e) *La moda*

La moda (M_0) è quel valore al quale corrisponde la massima frequenza. Si usa, generalmente, quando si tratta di mutabili statistiche non ordinabili. Le distribuzioni che hanno un solo massimo si chiamano unimodali, quelle che hanno due massimi bimodali. Se la v. s. è divisa in intervalli, la classe modale è quella a cui corrisponde la massima densità di frequenza.

Nell'esempio della Tav. III/2, la classe modale è quella di 18 - 22 anni perchè ad essa compete la massima densità di frequenza

$$h_2 = \frac{n_2}{d_2} = \frac{116}{4} = 29.$$

f) *Quantili*

Anche se poco usati nella ricerca sociale, è opportuno che lo studente impari a calcolare i quantili, ovverosia quei valori che ripartiscono i dati (disposti in ordine crescente) in parti uguali. In particolare:

— i *quartili* suddividono i dati in quattro parti uguali, sicché il primo quartile Q_1 è preceduto da un quarto dei dati ed è seguito da tre quarti (il 2° quartile è, dunque, uguale alla mediana). Ad esempio, il primo quartile dell'insieme numerico

2, 2, 4, 4, 5, 6, 6, 7, 8, 9, 9, 9, 9, 10, 10, 10

è $Q_1 = \frac{4+5}{2} = 4{,}5$ in quanto è preceduto da 1/4 dei dati ed è seguito da 3/4 dei dati;

— i *decili* suddividono i dati in 10 parti uguali (il terzo decile D_3, ad es., è preceduto dai 3/10 dei dati ed è seguito dai 7/10);

— i *centili* (o percentili) sono definiti in modo analogo.

Per il calcolo dei quantili nel caso la distribuzione sia divisa in classi, si costruisce la distribuzione cumulativa di frequenze e si procede come per la mediana.

Ad esempio, se si indicano con x_{h-1} ed x_h i limiti della classe ove cade Q_1, per cui $N_{h-1} < \frac{N}{4} \leqslant N_h$, si ha

$$Q_1 = x_{h-1} + \frac{x_h - x_{h-1}}{N_h - N_{h-1}} \left(\frac{N}{4} - N_{h-1}\right).$$

come è evidente, la precedente è analoga alla formula che fornisce il valore della mediana.

2. - I rapporti statistici

Fra le tecniche di elaborazione dei dati statistici notevole interesse assumono i rapporti statistici.

I rapporti statistici sono rapporti che si istituiscono fra due grandezze A e B, di cui una almeno statistica, legate da una relazione logica. Ad es., se A rappresenta il numero di abitanti di un certo territorio e B i km² di quel territorio, non ha senso la differenza A - B, ma il rapporto $\frac{A}{B}$ fornisce la densità della popolazione di quel territorio. Anche nel caso che A e B siano due grandezze omogenee non sempre la differenza (B – A) può dare delle informazioni logiche. Infatti, se A e B rappresentano, ad es., il prezzo dell'oro in due epoche diverse, A_1 e B_1 il prezzo del pane nelle stesse due epoche, non è lecito confrontare B - A con $B_1 - A_1$ perchè il livello dei prezzi è diverso. In tal caso il confronto va fatto non fra le differenze, ma tra i rapporti.

I rapporti più usati sono:

$$\frac{B - A}{A}, \qquad \frac{B - A}{B}, \qquad \frac{B - A}{\frac{A + B}{2}}$$

Questi rapporti si chiamano anche differenze relative o variazioni relative.

Ovviamente, la scelta di una delle tre formule va fatta in relazione al problema da risolvere.

Ad es., se A = 7.685 è il numero di morti per incidenti stradali nel 1983 e B = 7.164 il numero di morti per incidenti stradali nel 1984 in Italia, poiché $\frac{B-A}{A} = \frac{7.164 - 7.685}{7.685} = -0{,}0678$,

si può dire che il numero dei morti per incidenti stradali ha subito un decremento, fra il 1983 e il 1984, del 6,78 %.

I rapporti si possono istituire fra grandezze omogenee o fra grandezze eterogenee e sono di vario tipo. Ne riportiamo i principali.

a) *Saggi d'incremento e di decremento.*

Se A e B rappresentano l'ammontare di un fenomeno rispettivamente ai tempi t_1 e t_2,

$$\frac{B - A}{A(t_2 - t_1)}, \qquad \frac{B - A}{B(t_2 - t_1)}, \qquad \frac{B - A}{\frac{(A+B)}{2}(t_2 - t_1)}$$

diconsi *saggi d'incremento* se B > A; invece, si chiamano *saggi di decremento* se B < A; se, poi, t_1 e t_2 sono espressi in anni, detti rapporti forniscono *gli incrementi o i decrementi medi annui* a seconda che sia B > A o B < A; analogamente se t_1 e t_2 sono espressi in mesi, i rapporti forniscono gli *incrementi* o i *decrementi medi mensili.*

Ad es., sia A = 3.582.787 l'ammontare della popolazione residente in Puglia al 24 - X - 1971, B = 3.871.617 l'ammontare della popolazione residente in Puglia al 25 - X - 1981, poiché $t_2 - t_1 = 1981 - 1971 = 10$ anni,

$$\frac{B - A}{A(t_2 - t_1)} = \frac{3.871.617 - 3.582.787}{3.582.787 \cdot 10} = 0{,}000874$$

è l'incremento medio annuo della popolazione residente in Puglia nel periodo 1971 - 1981.

b) *Rapporti di composizione.*

Questi sono rapporti di una parte del fenomeno al tutto: quindi esprimono la « composizione » percentuale di un fenomeno rispetto alle sue parti.

Ad es., nell'anno 1984 i suicidi in Italia, classificati secondo lo stato civile, sono stati: $n_1 = 888$ suicidi celibi o nubili, $n_2 = 1.552$ coniugati, $n_3 = 159$ separati o divorziati, $n_4 = 574$ vedovi, con un totale $n = 3.173$ suicidi. I suicidi celibi o nubili rispetto al totale sono $\frac{n_1}{n} = \frac{888}{3.173} = 0{,}28$ (sono cioé,

il 28% del totale), mentre i coniugati suicidi rispetto al totale sono $n_2/n=1552/3173=0,489$ (sono, cioé, il 48,9% del totale), e così via.

c) Rapporti di derivazione

Questi rapporti si istituiscono quando il fenomeno A deriva dal fenomeno B. Essi si ottengono ponendo A/B. Esempi sono dati dai quozienti demografici, di cui ne riportiamo alcuni.

Quoziente di natalità: se n_t=numero di nati vivi nell'anno t e P_t=popolazione media di quell'anno, il quoziente di natalità è dato da $(n_t/P_t)1.000$. Esso esprime il numero di nati vivi in quell'anno per 1.000 abitanti.

Quoziente di nuzialità: se M_t=numero di matrimoni celebrati nell'anno t e P_t=popolazione media di quell'anno, il quoziente di nuzialità è dato da $(M_t/P_t)1.000$. Esso esprime il numero di matrimoni celebrati in quell'anno per 1.000 abitanti.

Quoziente di fecondità: se n_t=numero di nati vivi nell'anno t e F_{15-50} è la popolazione femminile in età presunta feconda (in età, cioè, compresa fra i 15 ed i 50 anni) al 30 giugno dello stesso anno, il quoziente di fecondità è dato da $(n_t/F_{15-50})1.000$.

Quoziente di mortalità: esso è il rapporto, moltiplicato per 1.000, tra il numero dei morti D_t dell'anno t e la popolazione media P_t di quell'anno, cioè $(D_t/P_t)1.000$. Questo quoziente fornisce il numero di morti per 1.000 abitanti nell'anno t.

d) Rapporto di durata

Supponiamo di essere in presenza di una popolazione che si rinnova in un intervallo di tempo *t* a causa delle entrate E e delle uscite U di nuovi elementi, e di voler valutare il tempo medio di permanenza degli elementi nella popolazione. In tal caso occorrerebbe conoscere i tempi di permanenza di ciascun elemento per poi farne la media.

Poiché il più delle volte questi dati non sono disponibili, occorre stimare detto tempo medio in altra maniera, con i cosidetti *rapporti di durata.*

Per far ciò, indichiamo con C_0 l'ammontare della popolazione all'inizio dell'intervallo in esame (ad es., all'1 gennaio), con $C_1=C_0+E-U$ quello di fine intervallo (ad es., al 31 dicembre) e con $C=(C_0+C_1)/2=$ $=(2C_0+E-U)/2$ la consistenza media del periodo.

Il flusso medio di entrata è approssimativamente fornito dal numero di elementi che entrano nella popolazione nell'unità di

tempo, cioè $F_E=E/t$; analogamente il flusso medio in uscita è approssimativamente uguale a $F_U=U/t$.

Se supponiamo, ora, che in ogni istante il numero degli elementi della popolazione non si discosti in maniera significativa da C (e ciò implica che in ogni istante $E\cong U$), il flusso medio è approssimativamente $F=(F_E+F_U)/2=$ $=[(E+U)/2]/t$.

Il rapporto D=C/F fra la consistenza media del periodo ed il flusso medio del periodo (che ha la dimensione di un tempo) fornisce allora la *durata media* della permanenza di un elemento in quella popolazione.

Esempio

L'1 gennaio 1994 nel reparto di medicina interna di un ospedale erano ricoverati $C_0=20$ ammalati, il 31 dicembre dello stesso anno ve ne erano 18, mentre durante tutto il 1994 ci sono stati E=798 entrati ed U=800 usciti. Stimare il numero medio di giorni di ricovero degli ammalati di quel reparto.

Poiché il flusso annuale è $F=\frac{798+800}{2}\ \frac{\text{persone}}{\text{anno}}=799\ \frac{\text{persone}}{\text{anno}}$, mentre il flusso giornaliero è $F=\frac{799}{365}\ \frac{\text{persone}}{\text{giorni}}$, $C=\frac{20+18}{2}$ persone = $=19$ persone, si ha $D=C/F=(19\text{ persone})/\frac{799}{365}\ \frac{\text{persone}}{\text{giorni}}\cong 9$ giorni.

e) Rapporti indici

I rapporti indici (r.i.) sono particolari rapporti statistici (precisamente sono quei rapporti in cui le grandezze sono omogenee) che misurano l'intensità di un fenomeno in un dato periodo (indice temporale) o in un dato luogo (indice spaziale) rispetto all'intensità dello stesso fenomeno in un periodo diverso oppure rispetto ad un luogo diverso.

Il denominatore di questi rapporti, cioè, il valore che il fenomeno ha assunto nel tempo o nel luogo preso a riferimento, si chiama *base*.

Nella Tav. IV.3 sono riportate le ore di lavoro perdute per i conflitti di lavoro in Italia, e, nella terza colonna, i r.i., base 1980=100.

I rapporti indici sono stati ottenuti dividendo ciascun valore della seconda colonna per il valore iniziale (115.201) e moltiplicando per 100. In tal modo la *base* risulta essere il valore iniziale uguale a 100.

Tav. IV/3 - Ore di lavoro perdute per i conflitti di lavoro in Italia.

Anni	Ore di lavoro perdute (migliaia)	r.i. base 1980 = 100
1980	115.201	100
1981	73.691	64,0
1982	129.940	112,8
1983	98.021	85,1

Si deduce, ad es., che nel 1983 si è avuto un decremento, rispetto al 1980, delle ore di lavoro perdute in conflitti di lavoro del 14,9% (−14,9 = 85,1 − 100); mentre nel 1982 si è avuto un incremento del numero di ore di lavoro perdute in conflitti di lavoro, rispetto al 1980, del 12,8% (112,8 − 100 = 12,8).

f) *Numeri indici.*

La serie di rapporti indici permette di misurare le variazioni (nel tempo e nello spazio) di una sola grandezza: se vogliamo, invece, sintetizzare le variazioni (nel tempo e nello spazio) di più grandezze dovremo costruire altri tipi di indici, detti *numeri indici* (n. i.).

E' evidente che il r. i. è un particolare n. i., per cui, spesso nell'un caso si parla di *n. i. semplice* e nell'altro di *n. i. composto.*

Nel n. i. composto gli elementi vengono *ponderati*, per attribuire ad essi diversa importanza.

L'Istat, ad es., per costruire n. i. di prezzi, usa ponderare i prezzi con le quantità rilevate al tempo base (tempo che indicheremo con 0): usa, cioè, la formula di Laspeyres:

$$I = \frac{p_{1t} \cdot q_{10} + p_{2t} \cdot q_{20} + p_{3t} \cdot q_{30} + \ldots\ldots + p_{st} \cdot q_{s0}}{p_{10} \cdot q_{10} + p_{20} \cdot q_{20} + p_{30} \cdot q_{30} + \ldots\ldots + p_{s0} \cdot q_{s0}} = \frac{\sum_{i=1}^{s} p_{it} \cdot q_{i0}}{\sum_{i=1}^{s} p_{i0} \cdot q_{i0}}$$

ove, ad es., p_{1t} sta ad indicare il prezzo del bene 1 al tempo t, p_{10} sta ad indicare il prezzo del bene 1 al tempo 0, q_{10} sta ad indicare la quantità del bene 1 al tempo 0. Analogamente: p_{2t} sta ad indicare il prezzo del bene 2 al tempo t, e così via.

E' evidente che il denominatore della formula di Laspeyres indica il prezzo pagato nell'anno 0 per acquistare quelle quantità degli *s* beni, mentre il numeratore indica il prezzo pagato nell'anno t per acquistare le stesse quantità di beni presi in considerazione nell'anno 0. *I* fornisce, allora, la variazione, fra l'anno 0 e l'anno t, del prezzo necessario per acquistare quelle quantità di beni, rimaste immutate nel passare dall'anno 0 all'anno t.

Supponiamo, ad es., che l'Amministrazione del Policlinico di Bari, voglia calcolare come è variato il prezzo dell'olio di oliva nel 1974 rispetto al 1973.

Nella 5ª colonna della Tav. IV/4 sono riportati i prodotti $p_{i0} \cdot q_{i0}$ dei prezzi pagati nel 1973 per le quantità consumate nel 1973, mentre nella 6ª colonna compaiono i prodotti dei prezzi pagati nel 1974 moltiplicati per le quantità consumate nel 1973. Si ha, allora,

$$I = \frac{125.357 \cdot 500 + 136.835 \cdot 300 + 146.447 \cdot 150 + 169.209 \cdot 50}{72.948 \cdot 500 + 77.979 \cdot 300 + 87.616 \cdot 150 + 93.785 \cdot 50} = \frac{134.156.500}{77.699.350} = 1{,}727.$$

Il prezzo medio dell'olio d'oliva è variato, quindi, dal 1973 al 1974, del 72,7%.

Con questa tecnica l'Istat calcola vari indici come, ad es., quello dei prezzi al minuto, dei prezzi all'ingrosso, l'indice per il funzionamento della scala mobile, ecc..

Tav. IV/4 - Prezzi (in lire per quintale) e quantità (in quintali) dell'olio di oliva acquistato dal Policlinico di Bari. Calcoli per la costruzione del n. i.

Qualità dell'olio	Prezzi medi L. per quintale		Quantità di olio consumate nel '73	$p_{i0} \cdot q_{i0}$	$p_{it} \cdot q_{i0}$
	1973 (p_{i0})	1974 (p_{it})	(q_{i0})		
comune	72.948	125.357	500	36.474.000	62.678.500
fino	77.979	136.835	300	23.393.700	41.050.500
sopraffino	87.616	146.447	150	13.142.400	21.967.050
extra	93.785	169.209	50	4.689.250	8.460.450
Totale				77.699.350	134.156.500

3. - La variabilità e la sua misura

I valori medi non sono sufficienti a descrivere il fenomeno in quanto i dati si presentano molto diversi l'uno dall'altro sia perchè varia è la natura dei fenomeni e sia perchè si possono commettere degli errori nelle varie misurazioni di una stessa grandezza. Perciò è molto importante misurare la variabilità, misurare, cioè, di quanto le modalità siano diverse fra loro. Infatti se 20 individui percepiscono lo stesso salario, basta il salario di uno solo di essi per sintetizzare tutto il collettivo; ma se, invece, ci sono individui che percepiscono 100.000 lire al mese ed altri che ne percepiscono dieci milioni, allora qualsiasi valore medio non è adatto a descrivere il collettivo. Si aggiunga, inoltre, che distribuzioni diverse possono avere la stessa media: ad es., il salario di due individui che percepiscono entrambi 2 milioni di lire al mese e quello di altri due che percepiscono, rispettivamente, 1 milione e 3 milioni di lire al mese. E' evidente, allora, che più alta è la variabilità e meno rappresentativa è la media: per cui a fianco della media di una distribuzione è sempre opportuno scrivere un indice di variabilità.

Nel caso, poi, si tratti di misurazioni di una stessa grandezza gli errori che si possono commettere possono essere di varia natura: ad es., imperfezione del metodo usato, imprecisione da parte di chi effettua le misure, ed anche circostanze imprevedibili che denomineremo caso. Orbene, ogni qualvolta gli errori sono dovuti al *caso* (la differenza tra il valore osservato e quello reale si chiama *errore accidentale*), gli scarti positivi hanno la stessa probabilità di verificarsi di quelli negativi, per cui deve essere nulla la somma di questi errori: *la media aritmetica dei valori viene ad essere, dunque, assunta come valore reale del fenomeno.*

Per misurare la variabilità si costruiscono opportuni indici i quali assumono il valore 0 quando non esiste variabilità (cioè quando tutte le modalità sono uguali tra loro) e crescono al crescere della medesima.

Distingueremo due casi: il caso in cui si voglia misurare la *dispersione* dei valori osservati x_i intorno al valore reale (che abbiamo supposto essere uguale alla media $\bar{x}$), ed il caso in cui si voglia misurare la *disuguaglianza* dei vari termini fra loro.

a) *misura della dispersione*

Gli indici più usati si basano sugli scarti fra i valori osservati x_i e la media aritmetica $\bar{x}$. E' evidente che non possiamo assumere come indice di variabilità la media degli scarti perchè, per aver supposto che gli scarti siano di natura accidentale, risulta $\Sigma\,(x_i - \bar{x})\,n_i = 0$. Per cui gli indici di variabilità che si costruiscono considerano o la media dei *valori assoluti* degli scarti, oppure la media dei quadrati degli scarti.

L'indice

$$S = \frac{\sum_{i=1}^{s} | x_i - \mu | \cdot n_i}{n}, \qquad [13]$$

denominato *scostamento semplice medio*, è uguale, appunto, alla media dei valori assoluti degli scarti: esso fornisce una misura di quanto, in media, le varie determinazioni del carattere differiscono dalla media μ.

Tale indice è espresso nella stessa unità di misura dei dati originari.

Ad es., calcoliamo lo scostamento semplice medio dei voti assegnati (in sessantesimi) ad un candidato dai sei componenti una commissione esaminatrice agli esami di Maturità. I voti sono espressi nella seguente tabella

$$\begin{pmatrix} A & B & C & D & E & F \\ 48 & 46 & 47 & 48 & 49 & 50 \end{pmatrix}.$$

La precedente tabella può scriversi anche nel seguente modo

x_i = voto	46	47	48	49	50
n_i = frequenza	1	1	2	1	1

Poichè si può assumere la media $\mu = 48$ come effettivo valore della preparazione del candidato, calcoliamo la variabilità esistente fra i vari voti con l'indice

$$S = \frac{|46-48| \cdot 1 + |47-48| \cdot 1 + |48-48| \cdot 2 + |49-48| \cdot 1 + |50-48| \cdot 1}{6} = 1.$$

In media, quindi, i voti assegnati al candidato dai componenti la commissione differiscono di 1 voto dal voto medio $\mu = 48$.

L'indice di variabilità denominato *varianza* è ottenuto, invece, calcolando la media dei quadrati degli scarti. Esso è dato da

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 n_1 + (x_2 - \bar{x})^2 n_2 + \ldots + (x_s - \bar{x})^2 n_s}{n_1 + n_2 + \ldots + n_s} = \frac{\sum_{i=1}^{s} (x_i - \bar{x})^2 n_i}{n} =$$

$$= \frac{\sum (x_i^2 - 2\bar{x}x_i + \bar{x}^2) n_i}{n} = \frac{\sum x_i^2 n_i}{n} - 2\bar{x}\frac{\sum x_i n_i}{n} + \bar{x}^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2. \qquad [14]$$

ove $n = \sum n_i$ e $\frac{\sum x_i n_i}{n} = \bar{x}$.

I motivi che fanno preferire questo indice rispetto a quelli basati sui valori assoluti degli scarti sono due:

— negli sviluppi matematici della materia è più facile operare con l'indice [14] anzichè con quelli basati sui valori assoluti degli scarti;

— a causa della seconda proprietà della media, l'indice [14] è un minimo rispetto ad analogo indice ottenuto partendo da un valore $k \neq \bar{x}$.

Questo indice ha l'inconveniente, però, almeno dal punto di vista descrittivo, di essere espresso con una unità di misura uguale al quadrato dell'unità di misura con cui sono misurati i dati originari.

Per poter confrontare l'indice di variabilità con la media μ (in modo da poter valutare la precisione di tale media), nelle ricerche sociali si preferisce, quale misura tipica della variabilità, lo *scarto quadratico medio* σ, che si ottiene estraendo la radice quadrata della [14], ossia

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 n_i}{n}} = \sqrt{\frac{\sum x_i^2 n_i}{n} - \bar{x}^2}, \qquad [15]$$

in quanto il valore di σ è espresso nella stessa unità di misura dei dati originari.

Il numeratore della [14] si chiama *devianza* e si indica col simbolo Dev(X) quando il carattere è X (ovviamente con Dev(Y) quando il carattere è Y). Cioè,

$$\text{Dev}(X) = \sum_{i=1}^{s} (x_i - \bar{x})^2 n_i = \sum_{i=1}^{s} x_i^2 n_i - n\bar{x}^2.$$

Diciamo subito che nei capitoli sull'inferenza saranno usati, anche, σ^2 e Dev(X).

Esempio

Un analista effettua per 10 volte la conta dei globuli rossi (g.r.) esistenti in un campione di sangue ed ottiene i seguenti risultati (in migliaia di g.r. per mmc)

x_i = globuli rossi	4500	4600	4700	4800
n_i = osservazioni	1	4	4	1

Calcolare la media μ e lo scarto quadratico medio σ.

Si ha:

$$\mu = \frac{4500 \cdot 1 + 4600 \cdot 4 + 4700 \cdot 4 + 4800 \cdot 1}{1 + 4 + 4 + 1} \text{ g.r.} = 4650 \text{ g.r.},$$

e

$$\sigma = \sqrt{\frac{(4500 - 4650)^2 \cdot 1 + (4600 - 4650)^2 \cdot 4 + (4700 - 4650)^2 \cdot 4 + (4800 - 4650)^2 \cdot 1}{1 + 4 + 4 + 1}} =$$

$$= 80{,}6 \text{ g.r.}.$$

Supponiamo di considerare i voti conseguiti all'esame di maturità da quattro gruppi di giovani: il 1° gruppo formato da maschi figli di professionisti (X_1), il 2° gruppo da femmine figlie di professionisti (X_2), il 3° gruppo da maschi figli di operai (X_3) e il 4° gruppo da femmine figlie di operai (X_4).

$X_1 = 54, 48, 43, 60, 45, 53$

$X_2 = 52, 46, 44, 56, 40, 42, 42, 38$

$X_3 = 36, 36, 39, 44, 40, 42$

$X_4 = 38, 38, 40, 42, 42$

Se indichiamo con $\bar{x}_i$, n_i e Dev (X_i) la media, il numero dei termini e la devianza del gruppo i e con $\bar{x}$ e Dev (X) la media e la devianza di tutti i termini, controllare che

$\bar{x}_1 = 50{,}5$	$n_1 = 6$	$\text{Dev}(X_1) = 201{,}5$
$\bar{x}_2 = 45$	$n_2 = 8$	$\text{Dev}(X_2) = 264$

$\bar{x}_3 = 39{,}5$	$n_3 = 6$	$Dev(X_3) = 51{,}5$
$\bar{x}_4 = 40$	$n_4 = 5$	$Dev(X_4) = 16$
$\bar{x} = 44$	$n = 25$	$Dev(X) = 996$

Si può verificare, allora, la formula della scomposizione della devianza

$$Dev(X) = Dev(X_1) + Dev(X_2) + Dev(X_3) + Dev(X_4) + (\bar{x}_1-\bar{x})^2 n_1 + (\bar{x}_2-\bar{x})^2 n_2 + (\bar{x}_3-\bar{x})^2 n_3 + (\bar{x}_4-\bar{x})^2 n_4.$$

Infatti,

$$996 = 201{,}5 + 264 + 51{,}5 + 16 + (50{,}5 - 44)^2 \cdot 6 + (45 - 44)^2 \cdot 8 + (39{,}5 - 44)^2 \cdot 6 + (40 - 44)^2 \cdot 5.$$

La formula della scomposizione della devianza, se i gruppi sono s, può scriversi

$$Dev\ (X) = \sum_{i=1}^{s} Dev\ (X_i) + \sum_{i=1}^{s} (\bar{x}_i-\bar{x})^2 n_i\ ,$$

la quale pone in luce che la devianza totale è uguale alla somma delle devianze calcolate con i dati dei singoli gruppi aumentata della devianza cacolata con le medie dei singoli gruppi.

La formula è molto importante perché, come vedremo in seguito, consente di stabilire se i gruppi sono o non sono omogenei fra loro.

b) *misura della disuguaglianza*

Quando la media non è un valore reale, non sempre σ è un indice appropriato.

Ad es., se l'Amministrazione del Policlinico di Bari acquista diverse partite di olio di oliva da fornitori diversi a prezzi diversi, in genere non interessa sapere di quanto detti prezzi differiscano dalla media, ma di quanto i prezzi differiscano fra loro.

Un indice che si usa spesso nelle ricerche sociali è dato dalla media di tutte le possibili differenze d_i, prese in valore assoluto, fra ogni termine x_i e gli altri. Cioè

$$\Delta = \frac{\sum_{i=1}^{n(n-1)} |\ d_i\ |}{n\ (n\ -\ 1)}\ . \qquad [16]$$

Si noti che da ogni elemento vanno sottratti gli altri n-1 elementi: per cui, essendo n gli elementi, le differenze possibili sono n(n-1).

L'indice [16], che si chiama *differenza media del Gini*, è espresso nella stessa unità di misura dei dati e si usa, appunto, quando si vuol misurare la *disuguaglianza dei termini tra loro.*

Ad es., dati i salari giornalieri (in migliaia di lire, 1975) di tre operai 8, 10, 14,

tramite la [16] si ha

$$\Delta = \frac{|8 - 10|+|8 - 14|+|10 - 8|+|10 - 14|+|14 - 8|+|14 - 10|}{3 \cdot 2}\ 1000\ £ = £\ 4000.$$

Cioè, i salari giornalieri di quei tre operai, in media, differiscono fra di loro di £. 4.000.

Si consideri, ora, la distribuzione dei salari giornalieri, in migliaia di lire 1975, $\begin{pmatrix} 7 & 8 & 9 \\ 2 & 3 & 1 \end{pmatrix}$, **che può scriversi: 7,7,8,8,8,9.**

Nell'effettuare tutte le possibili differenze si può notare che la differenza $|7\text{-}8|$ compare $2 \cdot 3$ volte, mentre la differenza $|7\text{-}9|$ compare $2 \cdot 1$ volte: cioè, in generale, la differenza $|x_i\text{-}x_j|$ va ponderata col prodotto $n_i n_j$ delle frequenze con cui si presentano le due modalità.

Tutto ciò premesso, il numeratore di Δ vien dato dalla somma di tutte le possibili differenze fra le modalità, ponderate col prodotto delle frequenze di quelle modalità. Per cui

$$\Delta = \{[|7\text{-}8|\cdot 2\cdot 3+|7\text{-}9|\cdot 2\cdot 1+|8\text{-}7|\cdot 3\cdot 2+|8\cdot 9|\cdot 3\cdot 1+|9\text{-}7|\cdot 1\cdot 2+ +|9\text{-}8|\cdot 1\cdot 3]/(6\cdot 5)\}\cdot 1000\ £ = £\ 867,$$

cioè, i diversi salari differiscono, in media, di £ 867.

Osservando il numeratore della precedente frazione, appare evidente, però, che $|7\text{-}8| \cdot 2 \cdot 3 = |8\text{-}7| \cdot 3 \cdot 2$ (ossia, in generale, $|x_i\text{-}x_j|\ n_i n_j = |x_j\text{-}x_i|\ n_j n_i$) dal che si deduce che il numeratore di Δ può essere dato *dal doppio della somma delle differenze in valore assoluto distinte*, ovvero

$$\Delta = \{2\cdot\ [|7\text{-}8|\cdot 2\cdot 3+|7\text{-}9|\cdot 2\cdot 1+|8\cdot 9|\cdot 3\cdot 1]/(6\cdot 5)\}\cdot 1000\ £ = £\ 867.$$

Ovviamente, se la distribuzione è $\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ n_1 & n_2 & n_3 & n_4 \end{pmatrix}$, si ha

$$\Delta = 2 \cdot\ [|x_1 - x_2|n_1 n_2+|x_1 - x_3|n_1 n_3+|x_1 - x_4|n_1 n_4+|x_2 - x_3|n_2 n_3+ + |x_2 - x_4|n_2 n_4+|x_3 - x_4|n_3 n_4]\ /[n(n-1)] \qquad [17]$$

Si ricava, perciò, la seguente regola: *il valore di Δ si ottiene moltiplicando per due la somma delle differenze fra ogni modalità e le successive, ponderate col prodotto delle frequenze delle due modalità di cui si fa la differenza, e dividendo il risultato ottenuto per n(n-1).*

La [17] è facilmente generalizzabile nella seguente

$$\Delta = \frac{2}{n(n-1)} \cdot \sum_{i=1}^{s} \sum_{j=i+1}^{s} |x_i - x_j| n_i n_j \quad . \qquad [18]$$

Se la v.s. è divisa in classi, per il calcolo di Δ si considerano i valori centrali delle classi.

4. - La variabilità relativa

Gli indici di variabilità studiati nel paragrafo precedente sono espressi nella stessa unità di misura con cui si misurano le modalità del carattere. Così, ad esempio, se le misurazioni delle stature di un gruppo di individui sono espresse in metri, anche la differenza media Δ di tali misurazioni è espressa in metri. Da ciò appare evidente che tali indici non si prestano sempre bene per fare confronti fra la variabilità di due fenomeni di natura diversa: ad esempio, non possiamo confrontare il Δ calcolato sulle stature di un gruppo di 100 individui, col Δ calcolato sui pesi dello stesso gruppo di individui (per studiare se varia più la statura che il peso di quegli individui).

Per questo motivo si rapportano gli indici precedenti ad una grandezza espressa nella stessa unità di misura, costruendo, in tal modo, altri indici (chiamati indici di variabilità relativa) che sono dei numeri puri, cioè degli indici che non sono espressi con una unità di misura.

I più comuni di tali indici sono di due tipi a seconda che siano riferiti alla media μ oppure al massimo valore che può assumere l'indice.

a) *Indici di variabilità relativa riferiti alla media* μ.

Questi indici sono del tipo

$$V_r = \frac{V_a}{\mu} \qquad [19]$$

ove con V_a si è indicato uno qualunque degli indici di variabilità assoluta. Un indice di questo tipo, usato spesso nelle applicazioni socio-economiche, è il *coefficiente di variazione*

$$CV = \frac{\sigma}{\mu} 100.$$

L'indice [19] presenta, però, qualche inconveniente:

— poichè al diminuire di μ cresce il valore di V_r, tale indice non è limitato superiormente (nel senso che può assumere valori grandissimi), sicchè non risulta, in genere, chiaro quando la variabilità è elevata oppure è bassa;

— quando la v.s. prende anche valori negativi, come nel caso, ad es., dei profitti e delle perdite di un'impresa, allora μ può assumere il valore zero e, di conseguenza, V_r non può calcolarsi (in quanto non si può fare la divisione per zero).

b) *Indici di variabilità relativa riferiti al massimo.*

Questi indici sono del tipo

$$V_r' = \frac{V_a}{\text{Max } V_a} \quad . \qquad [20]$$

E' evidente che

$$0 \leqslant V_r' \leqslant 1.$$

Infatti, $V_r' = 0$ quando $V_a = 0$ (cioè quando le modalità x_i sono tutte eguali tra loro), mentre $V_r' = 1$ quando $V_a = \text{Max } V_a$ (cioè quando la variabilità della distribuzione è massima).

Vediamo con qualche esempio quando la variabilità della distribuzione è massima.

Se 10 individui percipiscono in complesso un reddito di £ 1.000.000 al giorno, è ovvio che la variabilità è massima quando 9 individui non hanno reddito ed 1 solo percepisce £ 1.000.000 giornaliere: si noti che in questo caso la media è sempre la stessa.

Se si vuol studiare, invece, la variabilità delle stature di un gruppo di 10 individui adulti, è evidente che è assurdo pensare che il massimo della variabilità si ha quando 9 individui hanno statura zero ed 1 solo la statura complessiva (così come si è fatto nell'esempio precedente): questa volta diremo che la variabilità è massima quando alcuni di quegli individui hanno una statura l pari alla più piccola statura riscontrabile in natura (ad es., la statura di un nano) e gli altri hanno una statura L pari alla più elevata statura riscontrabile in natura (ad es., la statura di un gigante), ferma restando la statura media.

Da quanto detto appare allora evidente che, in generale, la massima variabilità si avrà quando le frequenze sono tutte concentrate nei due valori estremi (l e L) che possono essere assunti dalle modalità della distribuzione: naturalmente l ed L non sempre coincidono con x_1 e x_s, sicchè, di volta in volta, a seconda del fenomeno, questi valori devono essere fissati dal ricercatore.

Una volta fissati l ed L, se con p indichiamo la frequenza che compete ad l e con q la frequenza che compete ad L, si tratta di stabilire i valori da dare a p e q perchè la variabilità sia massima.

Per far ciò, supporremo che la distribuzione che ha la massima variabilità (che chiameremo *distribuzione massimante*) abbia la stessa media μ della distribuzione empirica e lo stesso numero di elementi.

Si ha, allora, il sistema

$$\begin{cases} p + q = n \\ \dfrac{lp + Lq}{n} = \mu \end{cases},$$

da cui si ricavano

$$p = \frac{n(L-\mu)}{L-l}. \qquad q = \frac{n(\mu-l)}{L-l}.$$

Sostituendo questi valori di p e di q nelle formule che forniscono i valori di S, di σ e di Δ, si ottengono, con semplici passaggi, i massimi valori che possono assumere i tre indici studiati. Ossia:

$$\text{Max } S = \frac{|l-\mu|\,p + |L-\mu|\,q}{n} = 2\,\frac{(L-\mu)(\mu-l)}{L-l} \qquad [21]$$

$$\text{Max } \sigma = \sqrt{\frac{(l-\mu)^2 p + (L-\mu)^2 q}{n}} =$$

$$= \sqrt{\frac{(\mu-l)^2 n(L-\mu) + (L-\mu)^2 n(\mu-l)}{n(L-l)}} = \sqrt{(L-\mu)(\mu-l)} \qquad [22]$$

$$\text{Max } \Delta = \frac{2}{n(n-1)}(L-l)\,pq = \frac{n}{n-1}\cdot\frac{2(\mu-l)(L-\mu)}{L-l} \qquad [23]$$

5. - Uso degli indici di variabilità relativa

Abbiamo già detto degli scopi che si propone lo studio della variabilità relativa. Ora indichiamo, qui di seguito, i casi più frequenti in cui si ricorre agli indici di variabilità relativa:

a) le modalità delle due distribuzioni risultano espresse in differenti unità di misura (si vuol confrontare, ad es., la variabilità delle stature con quella dei pesi di un gruppo di individui);

b) le due v. s. di cui si vuol confrontare la variabilità sono formate da rapporti (es., quozienti demografici);

c) le modalità delle due distribuzioni a confronto, pur potendosi misurare con la stessa unità di misura, differiscono per motivi di carattere geografico, sociologico, ecc. (es., i consumi di due gruppi di individui appartenenti a due strati sociali diversi).

6. - La concentrazione

Diciamo che un fenomeno è molto concentrato se una frazione rilevante della sua intensità totale compete ad una piccola frazione di casi. Quindi la concentrazione è un caso particolare della variabilità e riguarda un qualunque fenomeno *trasferibile*.

Ad es., diremo che gli addetti alle industrie di un territorio sono molto concentrati quando la maggior parte di essi lavora in pochi centri industriali. Ed ancora: nel caso un tale lasci ai suoi figli la stessa quota parte della sua proprietà si ha *equidistribuzione* o *concentrazione nulla*, nel caso invece la proprietà la lasci ad un solo figlio si ha *concentrazione massima*.

Poichè è evidente che si tratta di *disuguaglianza* e non di *dispersione*, per misurare la concentrazione useremo l'indice $\Delta/\text{Max }\Delta$. Poichè la concentrazione massima si ha quando tutto il fenomeno è concentrato in un solo individuo, supposto

allora che si conosca la media μ del carattere, deve essere $l = 0$, $L = n\mu$ per cui, sostituendo nella [23], risulta $\text{Max}\,\Delta = 2\mu$. L'indice che misura la concentrazione diventa quindi:

$$R = \frac{\Delta}{2\mu}. \qquad [24]$$

Questo indice si chiama *rapporto di concentrazione*. Esso, ovviamente, è un numero puro variabile tra 0 e 1.

L'indice [24] presuppone che siano note le determinazioni della v. s., per poter calcolare Δ; non è, quindi, applicabile al caso di una v. s. divisa in intervalli. Vediamo, perciò, come si costruisce un indice che vada bene anche nel caso di v. s. divisa in classi. Supponiamo, ad es., di avere classi di reddito e il reddito complessivo di ogni classe (se non conoscessimo tale reddito, lo stimeremmo moltiplicando il valore centrale della classe per il numero di redditieri (frequenza) relativo a quella classe).

Riferiamoci all'esempio della Tav. IV/6. Si indichino con 100 il reddito medio (stimato) della 1ª classe e con 600 il reddito medio (stimato) dell'ultima classe (aperta a destra).

Nella terza colonna della tavola compaiono le frequenze relative dei redditieri che hanno un reddito minore del limite superiore di ciascuna classe: ad es., 0,75 sta ad indicare che il 75% dei redditieri ha un reddito meno di 250.000 lire al mese.

Tav. IV/6 - Calcoli per ottenere *R*.

Classi di reddito mensile (migliaia di L. 1971)	Redditieri n_i	$p_i = \frac{N_i}{N}$	Redditi stimati delle classi $x_i n_i$	A_i	$q_i = \frac{A_i}{A_n}$	$p_i - p_{i-1}$	$q_i + q_{i-1}$	$(p_i - p_{i-1}) \cdot (q_i + q_{i-1})$
meno di 150	100	0,50	10.000	10.000	0,263	0,50	0,263	0,132
150 - 250	50	0,75	10.000	20.000	0,526	0,25	0,789	0,197
250 - 350	30	0,90	9.000	29.000	0,763	0,15	1,289	0,193
350 - 450	15	0,975	6.000	35.000	0,921	0,075	1,684	0,126
450 e più	5	1	3.000	38.000	1	0,025	1,921	0,048
Σ	200		38.000					0,696

Nella quinta colonna compaiono i redditi posseduti da tutti gli individui con redditi minori del limite superiore della classe corrispondente: ad es., 29 milioni è il reddito posseduto (stima) dagli individui che hanno un reddito mensile meno di 350.000 lire; nella sesta colonna, invece, abbiamo riportato la frazione di reddito globale posseduta dagli individui che hanno reddito minore del limite superiore della classe corrispondente.

Si riportino in un sistema di assi cartesiani, i punti (p_i, q_i). Congiungendo detti punti si otterrà, in genere, una spezzata detta spezzata di concentrazione o di Lorenz (Fig. IV/1).

Nel caso di *equidistribuzione*, una frazione di redditieri deve possedere la stessa frazione di redditi, per cui $p_i = q_i$; i punti, allora, stanno su una retta detta *retta di equidistribuzione*.

Nel caso di massima concentrazione n - 1 punti hanno ordinata zero e uno solo (l'ultimo) ordinata 1: sicchè la spezzata coincide con i lati del quadrato.

Più piccola è l'ordinata dei punti, maggiore è la differenza $p_i - q_i$ (si rammenti, ad es., che, essendo il triangolo isoscele, $p_2 - q_2 = M_2N_2$) cioè, maggiore è la variabilità fra percentuale di redditieri più poveri e percentuale di reddito globale che ad essi compete: maggiore è, quindi, l'area (detta di concentrazione) racchiusa dalla retta di equidistribuzione e dalla curva di concentrazione. Da quanto detto balza evidente, allora, che quan-

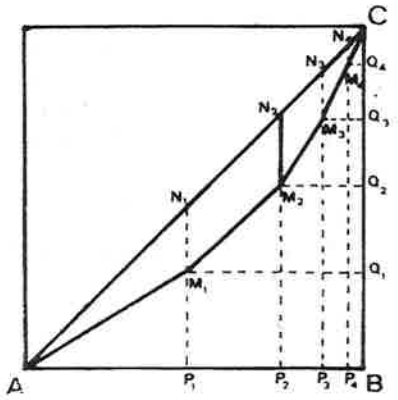


Fig. IV/1 - Spezzata di Lorenz

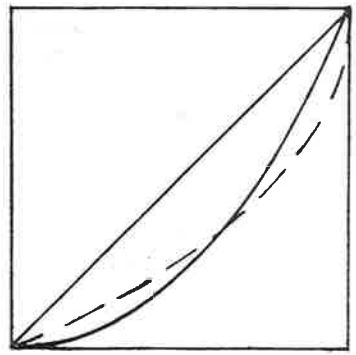


Fig. IV/2 - Curve con uguali aree di concentrazione

do la concentrazione è massima, l'area di concentrazione coincide con l'area del triangolo ABC. Il rapporto di concentrazione risulta, allora,

$$R^* = \frac{\text{Area di concentrazione}}{\text{Area del triangolo ABC}} .$$

Poiché l'area di concentrazione è data dall'area del triangolo ABC diminuita della somma delle aree

$$S(AP_1M_1) = p_1q_1/2 ,$$

$$S(P_1P_2M_2M_1) = (q_1 + q_2)\,(p_2 - p_1)/2, \text{ ecc.},$$

facilmente si ottiene

$$R^* = \frac{1/2 - \sum_{i=1}^{s} (p_i - p_{i-1})\,(q_i + q_{i-1})/2,}{1/2} =$$

$$= 1 - \sum_{i=1}^{s} (p_i - p_{i-1})\,(q_i + q_{i-1}) . \qquad [25]$$

ove $p_o = q_o = 0$.

Nell'esempio riportato è, dunque, $R^* = 1 - 0{,}696 = 0{,}304$. Il valore di R^* pone in luce una concentrazione dei redditi piuttosto tenue.

Le due formule [24] e [25], se applicate alla stessa v.s. discreta, forniscono risultati diversi. Si può facilmente verificare che fra i due indici sussiste la seguente relazione

$$R^* = [(s-1)/s]\, R .$$

I due indici [24] e [25] non tengono conto, però, della forma della curva di concentrazione, per cui, ad es., curve diverse possono avere lo stesso rapporto di concentrazione (vds., ad es., la Fig. IV/2): per i diversi tipi di curva di concentrazione sono stati proposti diversi indici il cui esame esula, però, dai limiti di questo corso.

7. - Relazione fra le medie e le varianze di variabili statistiche trasformate linearmente

Si consideri la variabile statistica X

$$X = \begin{pmatrix} 3 & 5 & 6 & 7 & 9 \\ 2 & 4 & 6 & 4 & 2 \end{pmatrix} ,$$

di media $\mu(X) = 6$ e varianza $\mathrm{Var}(X) = \frac{22}{9}$.

Si consideri, poi, la variabile statistica $Y_1 = X + 3$, che si ottiene aggiungendo 3 a tutte le modalità della variabile statistica X. Ossia

$$Y_1 = X + 3 = \begin{pmatrix} 6 & 8 & 9 & 10 & 12 \\ 2 & 4 & 6 & 4 & 2 \end{pmatrix} .$$

Si ricava facilmente che

$$\mu(X+3) = 6+3 = \mu(X)+3 \qquad \text{e} \qquad \mathrm{Var}(X+3) = \frac{22}{9} = \mathrm{Var}(X).$$

Generalizzando questo risultato si può dunque dire che la variabile statistica $Y_1 = X + a$, che si ottiene aggiungendo un numero a tutte le modalità della X, ha media e varianza fornite dalle relazioni

$$\mu(X + a) = \mu(X) + a \quad \text{e} \quad \mathrm{Var}(X + a) = \mathrm{Var}(X). \qquad [26]$$

Si consideri, ora, la variabile statistica $Y_2 = 2 \cdot X$, che si ottiene moltiplicando per 2 tutte le modalità della X, cioè

$$Y_2 = 2X = \begin{pmatrix} 6 & 10 & 12 & 14 & 18 \\ 2 & 4 & 6 & 4 & 2 \end{pmatrix} .$$

Facilmente si ricava

$$\mu(2X) = 2 \cdot 6 = 2 \cdot \mu(X) \quad \text{e} \quad \mathrm{Var}(2X) = 2^2 \cdot \frac{22}{9} = 2^2\,\mathrm{Var}(X).$$

Generalizzando questo risultato si può, allora, dire che la variabile statistica $Y_2 = b \cdot X$ (che si ottiene moltiplicando per b tutte le modalità della X) ha media e varianza fornite da

$$\mu(bX) = b \cdot \mu(X), \qquad \mathrm{Var}(bX) = b^2 \cdot \mathrm{Var}(X) . \qquad [27]$$

Si consideri, infine, la variabile statistica $Y = a + bX$ che si ottiene trasformando linearmente [9] la X (cioè ogni modalità della Y è legata alla corrispondente modalità della X dalla relazione $y_i = a + b\,x_i$).

Riunendo le [26] e le [27], si ottengono le relazioni

$$\mu(Y) = \mu(a+bX) = a+\mu(bX) = a+b\,\mu(X), \qquad [28]$$

$$\mathrm{Var}(Y) = \mathrm{Var}(a+bX) = \mathrm{Var}(bX) = b^2\,\mathrm{Var}(X) . \qquad [29]$$

[9] Una relazione del tipo $y = a + bx$ si dice lineare perchè in Matematica è l'equazione di una retta.

8. - Indici di mutabilità

Anche per i caratteri qualitativi si può misurare il grado di omogeneità degli elementi (*mutabilità*).

Detta omogeneità è massima se tutti gli elementi sono classificati con la stessa modalità (ad es., la distribuzione di un gruppo di diplomati quando i soggetti posseggano lo stesso diploma): ossia si ha massima omogeneità se la frequenza totale è concentrata in corrispondenza di una sola modalità; mentre l'omogeneità è minima (cioè l'eterogeneità è massima) quando le frequenze sono equidistribuite fra le modalità.

I fisici considerano la grandezza (*entropia*)

$$S(A) = k \cdot \log P(A)$$

(la quale mette in relazione lo stato di un sistema con la probabilità P (A) che il sistema sia in quello stato) come misura del grado di incertezza per un singolo elemento di appartenere ad uno stato del sistema (ove k è la costante di Boltzmann e la base dei log può essere anche 10).

Gli statistici, stimando P (A) — per ogni singola modalità — con la frequenza relativa della modalità e calcolando la media aritmetica ponderata delle stime dell'entropia relative a tutte le modalità (si è posto k = −1 per rendere l'entropia positiva), hanno introdotto l'*indice di entropia*

$$H = - \sum_{i=1}^{s} f_i \cdot \log f_i$$

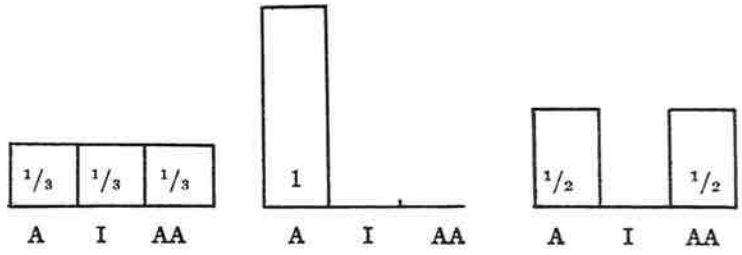
quale misura dell'eterogeneità del carattere qualitativo. Tale indice è 0 nel caso di massima omogeneità ed è massimo allorquando massima è l'eterogeneità.

Poiché si dimostra che

$$\text{Max } H = \log s,$$

si può costruire il corrispondente indice di mutabilità relativa, che varia, chiaramente, fra 0 e 1.

Si considerino, ad es., le distribuzioni degli addetti secondo i tre settori di attività economica Agricoltura (A), Industria (I) ed altre attività (AA), riportate in figura.



Nel caso della 1ª figura l'omogeneità è minima, per cui si ha massima incertezza per un individuo di appartenere ad uno dei tre stati (agricoltura, industria, altre attività): l'indice di entropia è

$$H = - (1/3) \cdot \log (1/3) - (1/3) \log (1/3) - (1/3) \log (1/3) = 0{,}477.$$

Nel caso della 2ª figura l'omogeneità è massima, perciò si ha la certezza per un individuo di appartenere all'agricoltura: l'indice di entropia risulta

$$H = - 1 \cdot \log 1 - 0 \cdot \log 0 - 0 \cdot \log 0 = 0$$

(ove si è posto $0 \cdot \log 0 = 0$, poiché in Matematica si dimostra che se x tende a 0 la funzione $x \cdot \log x$ tende anche a 0).

La 3ª figura rappresenta una situazione intermedia; in tal caso si ha

$$H = - (1/2) \cdot \log (1/2) - 0 \cdot \log 0 - (1/2) \log (1/2) = 0{,}301.$$

9. - Alcune notazioni fondamentali

Prima di procedere oltre è opportuno definire con chiarezza alcuni termini ed alcune notazioni che useremo in seguito.

Denomineremo *statistiche* (S) (in inglese *Statistics*) qualsiasi funzione dei dati campionari (come ad es., la media, lo

scarto quadratico medio o una frequenza); esse vengono indicate con le lettere dell'alfabeto latino. Invece i valori caratteristici della popolazione si indicano, generalmente, con le lettere dell'alfabeto greco. Ad es., l'età media al matrimonio di tutte le donne italiane coniugate è una caratteristica dello universo, invece l'età media al matrimonio di mille donne coniugate, estratte dalle varie regioni italiane, è una *statistica*.

I valori caratteristici dell'universo, generalmente sconosciuti, sono fissi, mentre le statistiche variano da campione a campione.

Le principali notazioni che useremo sono:

U = popolazione o universo;

N = numerosità dell'universo;

$\mu = \dfrac{\sum_{i=1}^{N} x_i}{N}$ = media del carattere X nell'universo;

$\hat{\mu}$ = stima di μ tramite il campione;

μ^* = valore di μ ipotizzato;

$\sigma^2 = \dfrac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$ = varianza del carattere X nell'universo;

$\hat{\sigma}^2$ = stima di σ^2 tramite il campione;

n = numerosità del campione;

$\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ = media del carattere X nel campione;

$s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$ = varianza del carattere X nel campione;

p = frequenza relativa del carattere X nell'universo;

$\hat{p}$ = stima di p tramite il campione;

p^* = valore ipotizzato di p;

f = frequenza relativa del carattere X nel campione.

10. - Media e varianza di caratteri qualitativi dicotomici

A questo punto è opportuno calcolare anche la media e la varianza, della popolazione e del campione, nel caso di caratteri qualitativi dicotomici (del tipo, ad es., maschi o femmine, occupati o non occupati, alfabeti o analfabeti, ecc.).

E' opportuno che lo studente impari subito i risultati che si otterranno, perché tali formule saranno utilizzate in seguito.

Cominciamo con l'assegnare agli elementi dell'universo:

— il valore 1 se l'elemento è portatore del carattere X,

— il valore 0 se l'elemento non è portatore del carattere X.

L'universo, in tal modo, è formato di 1 o di 0: per cui si possono trattare le scale nominali come scale ad intervalli.

Se indichiamo, allora, con p la frequenza dei portatori del carattere X nell'universo e, quindi, con 1–p la frequenza dei non portatori [di modo che Np è il numero degli elementi che portano il carattere X ed N(1–p) il numero degli elementi che non lo portano], la media e la varianza di tale popolazione sono fornite da

$$\mu = \frac{Np \cdot 1 + N(1-p) \cdot 0}{N} = p \qquad [30]$$

$$\sigma^2 = \frac{Np(1-p)^2 + N(1-p)\,(0-p)^2}{N} =$$

$$= \frac{Np(1-p)\,[(1-p) + p]}{N} = p(1-p)\,. \qquad [31]$$

Procedendo in modo analogo per il campione, si ha, poi,

$$\bar{x} = \frac{n\,f \cdot 1 + n\,(1-f) \cdot 0}{n} = f \qquad [32]$$

$$s^2 = \frac{n\,f\,(1-f)^2 + n\,(1-f)\;(0-f)^2}{n} = f\,(1-f)\,. \qquad [33]$$

ESERCIZI DA SVOLGERE

1) Calcolare la media aritmetica e la mediana dei dati riportati nell'esercizio 1) del Cap. II.

2) Calcolare la media aritmetica, la mediana e la classe modale delle risposte giuste date da 90 studenti e riportate nella Tavola ricavata coi dati dell'esercizio 1) del Cap. II.

3) Con riferimento ai dati dell'esercizio 8) del Cap. II, calcolare:

— l'età media dei 30 intervistati,

— l'età media degli occupati,

— l'età media degli operai,

— l'età media degli occupati nell'industria.

4) Calcolare la media aritmetica, la mediana e la classe modale dei guadagni di 80 individui riportati nella Tav. II/10.

5) Rilevando i dati dall'Annuario Statistico Italiano o dal Compendio Statistico Italiano, prevedere l'ammontare della popolazione pugliese al 25-X-1990.

6) Rilevando i dati dall'Annuario Statistico Italiano o dal Compendio Statistico Italiano, calcolare i quozienti di mortalità, per l'Italia, per classi d'età e sesso. Quale classe d'età è maggiormente colpita?

7) Calcolare l'ampiezza mediana delle famiglie italiane rilevate nell'esercizio 3) del Cap. II.

8) In Italia, nel 1975, negli istituti di cura si è registrato il seguente movimento dei ricoverati: 346.091 presenti al 1° gennaio, 9.702.083 entrati nell'anno e 9.714.207 dimessi. Calcolare la durata media di degenza.

9) Ricavando i dati dal Compendio Statistico Italiano, calcolare i rapporti indici (base 1973) del numero di posti letto e del numero di giornate di degenza negli istituti di cura italiani.

10) Ricavando i dati dal Compendio Statistico Italiano calcolare i rapporti indici (base 1973) del numero dei biglietti venduti in Italia per spettacoli, separatamente per teatro e cinematografo.

11) Ricavando dalle pubblicazioni ISTAT i dati relativi alla superficie e al numero di abitanti delle regioni italiane, dire qual è la regione con la più alta densità demografica.

12) Un commerciante all'ingrosso del settore alimentare desidera calcolare di quanto è variato il n.i. dei prezzi all'ingrosso dei prodotti acquistati nel 1977, rispetto all'anno precedente. A tal proposito considera le quantità acquistate di 8 prodotti nel 1976 e 1977 e i relativi prezzi medi di acquisto riportati nella tabella seguente

Prodotti	1976		1977	
	Prezzi medi per q.le	Quantità (in q.li)	Prezzi medi per q.le	Quantità (in q.li)
Salame	350.799	35	373.102	32
Mortadella	180.558	247	188.150	245
Prosciutto crudo	414.160	92	432.980	103
Formaggio grana vecchio	433.550	353	606.556	340
Pecorino romano tipo Italia	334.895	560	447.295	570
Provolone	205.019	96	234.751	100
Tonno all'olio	256.832	420	377.259	400
Conserva di pomodoro	57.291	575	74.492	580

13) Costruire l'indice più idoneo di variabilità assoluta e relativa dei dati riportati nell'esercizio 1) del Cap. II.

14) Con riferimento ai dati dell'esercizio 9) di questo capitolo calcolare il più idoneo indice di variabilità relativa.

15) Calcolare l'indice più idoneo di variabilità relativa dei quozienti di mortalità di cui all'esercizio 6) di questo capitolo.

16) Misurata per dieci volte la statura di una persona, si sono avuti i seguenti risultati in cm:
170,2; 169,8; 170; 170; 169,9; 169,9; 170,1; 170; 170,1; 170.
Calcolare il più idoneo indice di variabilità relativa.

17) Riferendosi ai dati dell'Es. 12) di questo capitolo, sono più variabili i prezzi del 1976 o del 1977?

18) Il numero di conflitti estranei al rapporto di lavoro in Italia nel 1977, per settore di attività economica, è fornito dalla seguente distribuzione

Agricoltura	Industrie	Altre attività
16	83	74

Calcolare un indice di variabilità assoluta.

19) Dividendo i dati dell'esercizio 1) del Cap. II nei tre gruppi: meno di 70, 70 - 90, 90 e più, verificare che la devianza totale è uguale alla somma delle devianze calcolate con i dati dei singoli gruppi aumentata della devianza calcolata con le medie dei singoli gruppi. Cioè:

$$\mathrm{Dev}(X) = \mathrm{Dev}(X_1) + \mathrm{Dev}(X_2) + \mathrm{Dev}(X_3) + (\bar{x}_1 - \bar{\bar{x}})^2 n_1 + (\bar{x}_2 - \bar{\bar{x}})^2 n_2 + (\bar{x}_3 - \bar{\bar{x}})^2 n_3 = \sum_{i=1}^{3} \mathrm{Dev}(X_i) + \sum_{i=1}^{3} (\bar{x}_i - \bar{\bar{x}})^2 n_i ,$$

ove con $\bar{x}_i$ e n_i si sono indicati, rispettivamente, la media e il numero dei termini del gruppo i e con $\bar{\bar{x}}$ si è indicata la media di tutti i termini.

20) Calcolare il rapporto di concentrazione della distribuzione dei comuni italiani per classi di ampiezza demografica (desumere i dati dal Compendio Statistico Italiano).

21) Calcolare il rapporto di concentrazione delle aziende agricole per classi di superficie (desumere i dati dal Compendio Statistico Italiano).

22) Calcolare la differenza media dell'indice generale dei prezzi all'ingrosso secondo i mesi (desumere i dati dal Compendio Statistico Italiano).

23) Ricavando i dati relativi alle provincie di Milano, Venezia, Bologna, Firenze, Roma e Cagliari dall'Annuario Statistico Italiano o dal Compendio Statistico Italiano, dire se sono più variabili le retribuzioni mensili lorde minime contrattuali degli impiegati di 1ª categoria delle costruzioni edilizie oppure del commercio.

24) Addizionando 4 ai dati dell'insieme numerico $X = \{2, 3, 6, 9, 10\}$, mostrare che si ottiene un nuovo insieme Y la cui media è data da quella del primo insieme più 4 e la cui varianza è uguale a quella del primo insieme. Ossia: $\mu(Y) = \mu(X) + 4$ e $\mathrm{Var}(Y) = \mathrm{Var}(X)$.

25) Moltiplicando per 3 i dati dell'insieme $X = \{2, 3, 6, 9, 10\}$, mostrare che si ottiene un nuovo insieme Y la cui media è data da quella del primo insieme moltiplicata per 3 e la cui varianza è data dalla varianza del primo insieme moltiplicata per 9: ossia, $\mu(Y) = 3\mu(X)$ e $\mathrm{Var}(Y) = 9 \cdot \mathrm{Var}(X)$.

26) Si consideri la variabile X di media $\mu = 6$ e $\sigma = 2$ e la variabile $Y = 3X + 4$. Qual è la media e lo scarto quadratico medio della variabile Y ?

27) L'età media dei 16 professori di lettere di una scuola media è di 40 anni, l'età media dei 4 professori di inglese è di 36 anni, l'età media degli 8 professori di matematica ed osservazioni scientifiche è di 38 anni, l'età media dei 4 professori di disegno è di 30 anni, l'età media dei 6 professori di educazione fisica è di 28 anni, l'età media degli 8 professori di applicazioni tecniche è di 30 anni e l'età media dei 2 professori di musica è di 26 anni. Calcolare l'età media di tutti i professori.

28) Un automobilista percorre con moto rettilineo uniforme quattro tratti di strada, rispettivamente di lunghezza s_1, s_2, s_3, s_4 e con velocità v_1, v_2, v_3, v_4. Sapendo che la legge del moto rettilineo uniforme è

$$s = vt,$$

trovare la velocità media con cui quell'automobilista ha effettuato lo intero percorso.

29) Un automobilista percorre con moto rettilineo uniforme (per cui $s = vt$) quattro tratti di strada, rispettivamente con velocità v_1, v_2, v_3, v_4 e con i tempi t_1, t_2, t_3, t_4. Trovare la velocità media con cui quell'automobilista ha effettuato l'intero percorso.