Appunti di STATISTICA corso di recupero

Docente Sciacchitano ANTONIA MARIA

Gli errori e le manchevolezze di questi appunti restano a mio carico.Sono grata a coloro che vorranno farmi pervenire,nella prospettiva di una sistemazione più adeguata,errori riscontrati,imprecisioni e giudizi di merito.

…Quelli che s'innamoran di pratica sanza scienza son come 'l nocchier ch'entra in naviglio sanza timone o bussola,che mai ha certezza dove si vada…

Leonardo Da Vinci

Aforismi sulla natura

Premessa

Pregiudizi sulla statistica

Ogni volta che si legge un giornale si incontra la Statistica!Non c'è scampo.Si cita la Statistica per sostenere qualsiasi tesi.Viviamo nell'era dei dati.I dati sono costantemente raccolti,talvolta ne siamo consapevoli altre volte no.

Lo scrittore inglese Herbert G.Wells sosteneva circa un secolo fa che il pensare in termini statistici sarà necessario un giorno per una vita civile ed efficiente quanto l'abilità del saper leggere e scrivere.Wells, in fondo,aveva visto giusto,ma non fu buon profeta.

Più di recente Ian Haching(epistemologo canadese)scrive che"la gente ha imparato ad usare i numeri…..e saper calcolare è considerato altrettanto importante che leggere e scrivere".Ma siamo ben lontani dall'intendere correttamente la Statistica e dall'acquisizione di uno stile di ragionamento statistico corretto.Sono ben note,d'altra parte,le canzonature che circolano da molto tempo sulla disciplina,che non servono a far chiarezza ma piuttosto ad alimentare confusione e opinioni distorte.

Citiamo tre esempi di pregiudizi che ricorrono spesso sulla statistica:

1-"Se mi rimanesse un'ora sola da vivere vorrei trascorrerla a una lezione di statistica perché sembrerebbe durare per sempre!"

Lamento di uno studente

2-"Ci sono tre tipi di menzogne-le bugie,le grandi bugie e la statistica".

Beniamin Disraeli

Tale aforisma è continuamente ripreso.E difatti è di pochi mesi fa un articolo dell'Espresso che riferiva come

il biologo Stephen Gould avesse riproposto la distinzione fra bugie,le dannate bugie e la Statistica.

3-“Se si muove è biologia,se cambia colore è chimica, se si rompe è fisica,se ti fa dormire è statistica”.

Bob Hogg,Università Dello Iowa

Anche voi potreste condividere una di queste idee sulla statistica.La statistica è noiosa e inutile!

“Nel breve periodo il pensiero statistico può migliorare la qualità delle decisioni prese;nel lungo periodo può trasformare le persone in leader”.

M.K.Pelosi-T.M.Sandifer

1-LA RILEVAZIONE DEI FENOMENI STATISTICI

1.1-INTRODUZIONE

La STATISTICA si configura come un momento importante della ricerca scientifica della pianificazione economica e dell’azione politica.La Statistica forma ed aiuta il momento conoscitivo dell’uomo fornendogli strumenti formali rigorosi e corroborati dalle osservazioni.

La parola STATISTICA ha perso il suo significato iniziale e viene impiegata per indicare la disciplina che analizza in termini quantitativi i fenomeni collettivi,ossia,i fenomeni il cui studio richiede l’osservazione di un insieme di manifestazioni individuali.

Intorno al 1600, nell'ambito di alcuni corsi universitari tedeschi venivano descritte le cose più rilevanti per la vita degli Stati.Questi argomenti venivano esaminati assieme a quelli di geografia,storia, di diritto pubblico ecc.. Si avvertiva l'esigenza di trattare autonomamente la descrizione dei fatti più rilevanti per uno Stato.Così nel 1660 il professore di Diritto pubblico Ermanno Coring tenne un corso di lezioni sulla descrizione sistematica dei fatti di uno Stato a cui diede il nome di

"Staats-Kunde".

Dalla descrizione qualitativa dei fenomeni dello Stato si passò un po' alla volta alla loro descrizione quantitativa che si manifestava in forma tabellare . Mentre in Germania si sviluppavano queste analisi, in Inghilterra si affermavano studi rivolti alla raccolta ed all'analisi di fenomeni sociali.Grande interesse suscitò nel 1662 la pubblicazione sugli atti della Società Reale di Londra, di una memoria del Capitano John Graunt.L'Autore utilizzava una massa di dati ricavati dalle registrazioni delle nascite e delle morti.Il Graunt, raggruppando i dati in classi omogenee, mise in evidenza regolarità e relazioni di notevole interesse demografico e sociale. Il Graunt ebbe seguaci sia in Inghilterra che in Germania.Ricordiamo fra questi il cappellano S$\ddot{u}$ssmilch(1707-1767) considerato il fondatore della demografia, che raccolse ed elaborò i dati demografici allora disponibili e diede un carattere di generalità ad alcune regolarità demografiche già riscontrate dal Graunt stesso.Lo sviluppo in Francia dell'indirizzo enciclopedico matematico ebbe il merito di introdurre nelle ricerche statistiche il calcolo delle probabilità.Ricordiamo Adolfo Quetelet(1796-1874) che riscontra nel campo sociale leggi che si descrivono con l'ausilio del calcolo delle probabilità.

K.Pearson(1854-1936),R.Fisher(1890-1962),J.Neyman(1894-1981)diedero un notevole impulso allo sviluppo dei metodi statistici nell'ambito della Statistica Inferenziale.Ricordiamo A.Wald(1902-1950) che ha introdotto i metodi sequenziali che permettono di ridurre i costi per il controllo della qualità dei prodotti industriali.

In Italia vanno ricordati Pareto,Benini,Boldrini.Notevole l'opera di Gini(1884-1965) che oltre ad occuparsi di problemi demografici,sociali ed economici, ha introdotto concetti originali nell'ambito della metodologia statistica.

La Statistica

non ha solo a che fare con la rilevazione e la rappresentazione dei dati, ma propone i metodi per analizzare gli stessi.(la Statistica viene spesso confusa con le Statistiche:dati, tabelle, grafici, indici , medie..).

L'arricchimento della conoscenza su un dato fenomeno del reale avviene mediante un processo di integrazione della conoscenza passata con le informazioni fornite da nuove osservazioni del fenomeno.Osservazioni che si concretizzano nella disponibilità di un numero più o meno elevato di dati statistici portatori di informazioni sul o sui caratteri in studio.In questo processo intervengono i metodi tipici di analisi e di sintesi di ciascuna scienza della natura,specializzata nello studio di particolari fenomeni del reale,quali la Fisica,la Chimica, la Medicina,e di due discipline che possono considerarsi ausiliarie delle precedenti che sono la Matematica e la Statistica.

Differenze

La Statistica a differenza della Medicina, dell'Ingegneria e della Giurisprudenza, non affronta, costruisce e risolve i casi singoli nel quadro di esperienze, normative o leggi generali.Al contrario, cerca di pervenire a leggi generali e alla discussione critica della loro validità a partire dalla conoscenza

aggregata risultante da tanti casi singoli accomunati da regole e definizioni.

La Matematica fornisce gli strumenti logico-linguistici per una corretta formulazione delle teorie consentendo di formalizzare in modelli e leggi le ipotesi suggerite dal mondo empirico.

La Statistica fornisce i metodi per una appropriata sintesi delle informazioni contenute nei dati osservati,nonché,insieme alla Matematica,il complesso dei procedimenti razionali per la verifica di conformità dei modelli alla realtà.

Le due discipline sono necessarie nel processo induttivo-deduttivo della conoscenza, ma i loro compiti sono ben distinti:

strettamente razionale quello della matematica perché elabora concetti, idee, ipotesi, cioè "enti astratti" per i quali è inconcepibile l'errore anche minimo pena la non verità delle proposizioni dimostrate,teoremi,lemmi, corollari,ecc.:

strettamente empirico quello della statistica chiamata ad elaborare dati osservati,fatalmente affetti da errori, dai quali si vuole estrarre la parte vera dell'*informazione* di cui sono portatori eliminando se è possibile, l'influenza degli errori.

La induzione da un collettivo limitato all'intera popolazione e la formalizzazione dei dati disponibili entro uno schema teorico di riferimento sono gli aspetti più affascinanti della ricerca statistica attuale.

Questi ultimi,infatti,sono direttamente rivolti alla elaborazione di teorie e leggi scientifiche sulla base di osservazioni sperimentali.

Nell' ambito della sua attività quotidiana, l'essere umano raccoglie informazioni sia per sete di conoscenza sia per raggiungere obiettivi concreti. In entrambi i casi l'obiettivo specifica la natura delle informazioni da raccogliere e gli strumenti con i quali esaminare i dati.

La Statistica prende le mosse dall'osservazione dei dati,sia osservazionali,sia sperimentali,osservazione che fornisce INFORMAZIONI sotto forma di DATI.

La STATISTICA:

1-Organizza tali dati;

2-li predispone per l'analisi;

3-li elabora per condensare le INFORMAZIONI in esse contenute in particolari VALORI TIPICI;

4-formula IPOTESI sui meccanismi che regolano i fenomeni;

5-verifica le teorie costruite sulla base delle IPOTESI formulate.

Stabilito quali siano i dati,disponibili od ottenibili,pertinenti all'obiettivo conoscitivo o decisionale,la Statistica interviene in modo astratto e generale,con principi e metodologie proprie.La natura dei dati (siano essi appunto demografici, economici, biologici, tecnologici, o di altro tipo)passa in secondo piano nel momento di astrazione. Non va trascurato,

tuttavia,che le conclusioni sostanziali vanno poi espresse nuovamente nell'ambito del contesto di partenza.

OSSERVAZIONE E SPERIMENTAZIONE

La distinzione fra dati sperimentali e dati di osservazione è cruciale per il loro conveniente trattamento statistico.La sperimentazione è contraddistinta dalla replicabilità a piacere,secondo una chiara definizione di operazioni e circostanze(protocollo dell'esperimento);l'osservazione è invece limitata a una specifica evenienza.Se si effettua una osservazione,il fenomeno oggetto di studio è precostituito,i dati esistono in natura,finiti,e sono solo rilevati o studiati,direttamente per come si presentano.Sono tipicamente di osservazione i dati relativi alle realtà antropometriche e demografiche esaminate asaustivamente tramite un censimento.

Nella sperimentazione,i dati sono per così dire creati,in circostanze controllate,in una successine di replicazioni dell'esperimento,potenzialmente infinita.Va precisato che *esperimento* in Statistica ha un'accezione specifica più ampia che nel linguaggio corrente,ove richiama l'idea di qualcosa che avviene in laboratorio.

Sono esempi di sperimentazioni:il lancio di un dado;l'estrazione di uno o più elementi da un'urna di elementi omogenei,e veri e propri esperimenti di laboratorio.

Occorre quindi una accurata definizione della terminologia statistica per l'apprendimento e la corretta applicazione dei metodi statistici.

Così quando parliamo di RILEVAZIONE STATISTICA intendiamo :

Il complesso di operazioni rivolte ad acquisire una o più Informazioni su un insieme di elementi oggetto di studio.

1.2-FASI DI UNA RILEVAZIONE

Come si intuisce si opera per stadi successivi che possiamo elencare:

1-Definizione degli Obiettivi
2-Rilevazione dei dati
3-Elaborazione metodologica
4-Presentazione ed Interpretazione dei dati
5-Utilizzazione dei risultati della ricerca

Esaminiamo i vari punti

1-Definizione degli obiettivi.

Come detto precedentemente,l'obiettivo specifica la natura delle informazioni da raccogliere e gli strumenti con i quali esaminare i dati.La selezione dell'informazione utile per gli obiettivi sostanziali dell'indagine è compito delle discipline specifiche(ad esempio Economia, Sociologia,Biologia ecc.) nel cui ambito interpretativo i dati si collocano.

ESEMPI

a-Ricerca sui consumi dei giovani

Occorre definire con esattezza:

- **fascia di età dei soggetti da intervistare;**
- **territorio di riferimento;**
- **tipologia dei consumi da richiedere;**
- **periodo temporale entro cui misurare tali consumi.**

b-Ricerca sui lavoratori pendolari

Definire:

- **mezzo di trasporto utilizzato;**
- **caratteristiche dell'utente(dipendente-imprenditore-rappresentante);**
- **la frequenza e la regolarità del suo recarsi al lavoro;**
- **altri aspetti(se trattasi di lavoratori saltuari,stagionali,fissi).**

La definizione degli obiettivi è il punto di partenza per ogni tipo di rilevazione statistica.Devono essere esplicitate le finalità conoscitive in maniera da rendere esattamente individuabile la popolazione alla quale si riferisce e le unità statistiche che la compongono.

2-RILEVAZIONE dei dati.

La rilevazione dei dati può essere totale(quando si esaminano tutti gli elementi oggetto di studio)o parziale(quando ci si limita a studiare un sottoinsieme,detto campione,dell'insieme di riferimento)diretta o indiretta.Considerazioni di natura pratica e teorica hanno spinto gli statistici a impiegare

sempre di più le indagini campionarie.I principali fattori che inducono a optare per l'indagine campionaria sono:a)la riduzione dei tempi e dei costi necessari all'effettuazione dell'indagine;b)l'attendibilità delle informazioni ottenibili da un campione può essere misurata e controllata;c)la qualità e la mole delle informazioni può essere maggiore di quella ottenibile da un'indagine totale.*(Alcune forme di campionamento saranno trattate più avanti)*

Infine è essenziale esplicitare il modo con cui si raccolgono le informazioni cioè tramite dichiarazioni (mediante questionario e intervista) ovvero misurazioni(con strumenti).Tra i metodi di acquisizione di dati da popolazioni particolare rilievo assume l'intervista, che può essere diretta,telefonica.Per fare un'intervista le domande vengono raccolte in un apposito modello detto questionario.

QUESTIONARIO-Un questionario è un insieme ragionato e ordinate di domande,che possono essere chiuse oppure aperte.Una domanda chiusa prevede una lista di risposte precodificate;l'intervistato sceglie quelle che meglio rispecchia il suo pensiero e alla quale è associato un codice numerico prestabilito.Una domanda aperta lascia invece all'intervistato la facoltà di rispondere come meglio crede,riservando alla fase di elaborazione dei dati il problema della codifica delle risposte stesse.Quando il questionario di intervista viene somministrato al soggetto non in un contatto diretto,ma per mezzo del telefono,si ricade nella tipologia delle interviste telefoniche,le quali hanno il vantaggio di

costare meno e di richiedere meno tempo rispetto alle interviste faccia a faccia.

Si sono realizzate alcune tecniche informatiche di supporto all'intervista,che permettono di ridurre i tempi e i costi dell'intervista,migliorando la qualità dei dati:parliamo del CATI.

CATI: COMPUTER AIDED TELEPHONIC INTERVIEWING:INTERVISTA Telefonica assistita dal computer.

CATI-Il metodo CATI funziona così:il questionario di intervista non é più cartaceo ma viene caricato su personal computer collegati con un elaboratore centrale;tutti gli intervistatori operano dalla sede centrale dell'istituto per mezzo di una postazione informatizzata composta da una cabina isolata acusticamente,un personal computer sul cui schermo compaiono le domande da porre e che può registrare le risposte mediante digitazione diretta sulla tastiera,e ovviamente il telefono;il computer seleziona casualmente,volta per volta,il numero telefonico dell'intervistato ed effettua la chiamata mettendolo in contatto con l'intervistatore;le domande compaiono sullo schermo del computer e l'intervistatore digita le risposte direttamente attraverso la tastiera;il computer avanza automaticamente alla domanda successiva a seconda del tipo di risposta ottenuta(minimizzando così gli errori di rilevazione);il computer centrale elabora le risposte in progress,immediatamente,tanto che alla fine dell'ultima intervista i risultati statistici sono già disponibili.I vantaggi temporali e di costo di una simile

tecnica sono evidenti.Con lo sviluppo del CATI,il mercato dei sondaggi di opinione(che non richiedono la visione di materiale da parte dei soggetti e possono benissimo articolarsi in semplici domande chiuse,come prevede appunto la metodologia CATI)ha ricevuto un notevole incremento.

Alcuni tipi di rilevazione non fanno uso del questionario,tra queste vi sono le rilevazioni che utilizzano strumenti automatici di registrazione.Per esempio,negli studi psicologici del comportamento a volte viene utilizzata la ripresa tramite telecamera,oppure nel campo meteorologico si utilizzano strumenti di rilevazione automatica di temperatura umidità ecc..

Oltre a poter acquisire i dati attraverso un esperimento o un'indagine,è possibile ricorrere a collezioni di dati predisposti da enti e società esterne e già pronti per essere analizzati.

FONTI Statistiche

Si chiamano "fonti statistiche" in senso lato gli enti che a vario titolo,forniscono dati statistici.La raccolta e la divulgazione dei dati statistici ufficiali sono coordinate nell'ambito del Sistema Statistico Nazionale(SISTAN).

Con il decreto legislativo del 6/9/89 è stato istituito il Sistema Statistico Nazionale-Sistan-,a cui è stato affidato il ruolo di coordinatore delle varie fonti di rilevazione.Include numerosi enti che rilevano e diffondono dati attraverso i loro Uffici statistici.La nuova legge ha creato un sistema flessibile e decentrato,costituito dagli uffici di statistica centrali e periferici delle Amministrazioni statali,delle Regioni e Province autonome,delle Prefetture,delle Province,dei Comuni,delle USL,delle aziende autonome e di altri enti pubblici.

ISTITUTO CENTRALE di STATISTICA

Nell'ambito del SISTAN,ha un ruolo fondamentale l'ISTAT,l'ente di produzione statistica

I dati diffusi dall'ISTITUTO Centrale di Statistica vengono considerati come la principale base di riferimento per ogni tipo di analisi statistica sul paese.In Italia,l'Istituto Nazionale di Statistica-ISTAT-attivo dal 1926,è l'organo della statistica ufficiale ed è il maggiore produttore di dati.

Ogni anno l'ISTAT svolge circa 200 indagini.Di queste il 3% riguarda l'ambiente il 5% è di tipo demografico,il 35% di tipo sociale,il 57% si occupa dell'area economica.

Le indagini censuarie svolte periodicamente sono:

-Censimento generale della popolazione e delle abitazioni,dal quale si traggono informazioni circa le caratteristiche della popolazione e delle abitazioni;

-Censimento generale dell’industria,del commercio,dei servizi e dell’artigianato,che rileva le caratteristiche strutturali delle imprese e delle unità locali;
-censimento generale dell’agricoltura,dal quale si ricavano le caratteristiche strutturali delle aziende agricole.

CARETTERISTICHE PRINCIPALI dei CENSIMENTI:

-UNIVERSALITA’,in quanto si estendono a tutte le Unità esistenti;

-SIMULTANIETA’,poiché l’enumerazione delle unità viene effettuata rispetto ad un dato istante di tempo;

-PERIODICITA’,in quanto la rilevazione viene ripetuta ad intervalli regolari e più precisamente decennali.

Tali rilevazioni censuarie presentano il pregio di essere ESAUSTIVE rispetto al fenomeno indagato,ma anche il LIMITE di non consentire sia analisi di breve periodo a causa della loro periodicità decennale,sia analisi sufficientemente dettagliate in quanto generalmente non rilevano specifiche sfaccettature della struttura produttiva.

Accanto alle indagini censuarie,si aggiungono alcune indagini campionarie a carattere periodico di rilevante importanza.Ricordiamo:

-indagine sulle forze di lavoro

-indagini sui consumi delle famiglie

-indagine multiscopo,effettuata in vari cicli,su qualche tematica di particolare interesse.Le indagini multiscopo mirano ad acquisire sull'unità campionaria(generalmente la famiglia)numerose informazioni(tempo libero,salute ecc.).

INDAGINI CONDOTTE DALL'ISTAT nel SETTORE TURISTICO.

Le indagini che l'Istat conduce direttamente o con la collaborazione dell'organizzazione turistica pubblica,consentono di quantificare la consistenza e la struttura dell'Offerta Ricettiva.Il termine offerta è molto lato ed è comprensivo delle strutture ricettive(Risorse riproducibili)e di tutto ciò che può stimolare le persone a frequentare determinati luoghi:comprende quindi tutte le preesistenze di carattere naturale,culturale e storico(Risorse non riproducibili).

Altre fonti di dati sono costituite dalle banche dati;queste raccolgono,su supporto magnetico,grosse moli di dati la cui organizzazione e gestione è controllata da società o enti tramite appositi software.

All'indirizzo: http://con.istat.it si può accedere alla banca dati dell'ISTAT solo dopo essersi registrati.Questa banca dati contiene più di 8000 serie

storiche d'indicatori congiunturali prodotti dall'ISTAT,articolati nei seguenti domini:

1-prezzi
2-attività delle imprese di servizi
3-occupazione
4-retribuzioni e altri indicatori del lavoro;
5-costruzioni;
6-attività delle imprese industriali;
7-commercio estero;
8-turismo.

In ciascun dominio è possibile rintracciare i dati relativi a una o più indagini.

Elementi di una rilevazione statistica.

Alcune definizioni si rendono opportune per una corretta applicazione dei metodi statistici,prima di passare ad esaminare il terzo punto, che attiene la elaborazione metodologica.

POPOLAZIONE-Tutte le indagini statistiche tendono ad acquisire conoscenze intorno alla distribuzione di uno o più caratteri in una popolazione.

Il termine Popolazione non è sinonimo di INSIEME; infatti con quest'ultimo termine ci si può riferire anche a Entità del tutto Eterogenee,mentre le UNITA' che costituiscono una popolazione devono possedere una o più caratteristiche comuni,che li facciano riconoscere come appartenenti alla data popolazione,e li facciano distinguere da unità che appartengono ad altre popolazioni.

Per POPOLAZIONE intendiamo quindi, l'insieme di Unità omogenee rispetto a una caratteristica comune a tutti i suoi componenti.

A tale riguardo, si distingue tra POPOLAZIONE Reale(effettivamente esistente e visibile) POPOLAZIONE Virtuale(definibile con accuratezza ma non è osservata né osservabile).Una popolazione non costituisce necessariamente un insieme biologico,essendo lecito pensare alla popolazione delle lampadine prodotte nell'ultimo mese da un'azienda X.E' una popolazione reale quella dei residenti maschi a Palermo in età compresa tra 16 e 65 anni.E' una popolazione virtuale quella delle possibili cinquine estraibili su una prefissata ruota del lotto.

UNITA'-E' l'elemento di base della popolazione sul quale viene effettuata la rilevazione o la misurazione di uno o più caratteri.L'unità statistica è definita in termini di occasione,tempo,durata,territorio.

Una volta individuata la Popolazione occorre osservare sulle Unità,che la compongono,alcuni Caratteri e specificare le Modalità del carattere.

CARATTERE-Classe di attributi,X,associabile alle unità statistiche.Se gli attributi sono misure numeriche si parlerà di caratteri quantitativi,VARIABILI QUANTITATIVE o più semplicemente,di VARIABILI.Se gli attributi sono qualificatori non metrici(aggettivi),si parlerà di caratteri qualitativi,Variabili qualitative o MUTABILI.

MODALITA'-Elementi x di una classe di attributi,X.In altre parole un carattere è definito dalla classe delle sue modalità.E' l'espressione concreta del carattere nelle unità statistiche,cioè il numero(per caratteri quantitativi) o l'attributo(per caratteri qualitativi) che l'unità statistica manifesta.L'elenco di tutte le possibili modalità di un carattere si dice ESAUSTIVO se è completo ossia deve essere capace di interpretare qualunque manifestazione del carattere,e le modalità si dicono Disgiunte se una unità statistica può manifestare il carattere in una ed una sola modalità tra quelle indicate.

LA POPOLAZIONE si specifica nella UNITA' STATISTICA mentre il CARATTERE(che varia nella Popolazione)si SPECIFICA nella MODALITA'(assunta nella Unità Statistica)

3-ELABORAZIONE METODOLOGICA-

In questa fase gioca un ruolo decisivo la distinzione tra caratteri qualitativi e quantitativi per la corretta applicazione del metodo statistico.

ESEMPI-

_ L'età,il peso, l'altezza, la temperatura,il tempo impiegato per raggiungere l'aeroporto,qualsiasi grandezza fisica misurabile,il numero di clienti in attesa in un sportello bancario,il numero dei passeggeri annui delle

linee Alitalia,etc.costituiscono esempi di caratteri quantitativi.

_ Il sesso,la religione,lo stato civile,la professione,la categoria sociale di appartenenza,il comune di nascita,lo stato d'animo dopo un esame,la reazione di fronte ad un'opera pittorica,il gradimento di una canzone etc.costituiscono esempi di caratteri qualitativi,perché per essi ogni associazione con numeri reali-se anche è possibile o talvolta conveniente-assume un connotato intrinseco di arbitrarietà.

All'interno del primo gruppo,i caratteri quantitativi, ossia le variabili si distinguono in DISCRETI e CONTINUI.

Le VARIABILI CONTINUE sono quelle capaci di assumere,in linea di principio,qualsiasi valore contenuto in un intervallo reale predefinito,le VARIABILI DISCRETE invece no:il che implica,che possono assumere al più un numero"discreto" di modalità,finito o infinito.In altri termini le variabili discrete assumono modalità che possono sempre essere poste in corrispondenza con l'insieme dei numeri naturali(1,2,..)o un suo sottoinsieme;invece,le variabili continue assumono modalità che possono essere poste in corrispondenza con l'insieme dei punti di un intervallo,finito o infinito della retta reale.

ESEMPI_

_L’età,il peso,l’altezza,la temperatura di una stanza,il tempo impiegato per raggiungere l’aeroporto,qualsiasi grandezza fisica misurabile,costituiscono variabili continue poiché possono assumere qualsiasi valore in un intervallo predefinito,né esistono motivi concettuali per cui,per esempio,fissati due pesi estremamente vicini,non sia possibile trovare una persona con peso intermedio a quei pesi.Agli effetti pratici,la nostra possibilità di valutare punti estremamente vicini dipende dagli strumenti di misura che consentono misure accurate solo fino ad un certo punto.

_Il numero di clienti in attesa ad uno sportello bancario,i passeggeri delle linee Alitalia,il numero dei componenti di una famiglia,etc.costituiscono esempi di variabili discrete perché concettualmente le possibili modalità di tali caratteri sono elencabili e quindi possono essere messe in corrispondenza con i numeri naturali.Il numero dei clienti in fila di attesa può assumere valori in (0,1,2,…..,M) essendo M un intero massimo predefinito in funzione dei potenziali utenti;il numero dei lanci necessari perché si verifichi(5) per la prima volta può assumere modalità contenute in (1,2,3,..)ed è quindi una variabile discreta con un numero infinito numerabile di modalità non essendo certi che esista un M finito tale che certamente la faccia con 5 punti si verifichi.

La distinzione tra variabili discrete e continue è anzitutto *concettuale* perché la limitatezza degli

strumenti di misura rende qualsiasi variabile,anche teoricamente continua,di fatto discreta.

E allora ,perché insistere a parlare di universo con infinite determinazioni?Perché accettare la finzione che qualunque valore reale può essere assunto dal fenomeno? Nessuno avrà mai la possibilità di eseguire misure esatte.Vi è un solo motivo:è più comodo lavorare con funzioni continue che con funzioni discrete,è più comodo assumere infinite possibili determinazioni invece che un numero finito.

La distinzione dei caratteri in qualitativi e quantitativi non è sufficiente ed è necessario specificare con maggiore dettaglio la natura del carattere.

Nell'ambito della teoria della misurazione,lo psicologo Stevens(1946) introdusse una classificazione dei caratteri che tuttora viene utilizzata nella letteratura internazionale.I caratteri si distinguono in funzione della scala di misurazione: nominale,ordinale,ad intervallo, di rapporto.

I caratteri nominali ed ordinali sono qualitativi(quindi sono delle mutabili)mentre i caratteri che ammettono una scala a intervallo o per rapporto sono quantitativi(quindi,sono delle variabili in senso stretto)

_I caratteri con scala nominale costituiscono mutabili le cui modalità(attributi)non assumono alcun ordine

precostituito.L’unico confronto ammissibile tra due unità statistiche rispetto a caratteri nominali consiste nello stabilire se possiedono o no lo stesso attributo,cioè se sono diversi oppure uguali rispetto a quella mutabile.

_I caratteri con scala ordinale costituiscono mutabili che assumono modalità logicamente sequenziali.Per tali caratteri,però,non è possibile attribuire un valore numerico alla distanza tra le modalità.

_I caratteri con scala ad intervallo sono variabili per le quali lo zero della scala delle misure non è assoluto ma convenzionale.

_I caratteri con scala di rapporto sono variabili per le quali è intrinseca ed univoca la definizione dello zero assoluto.

Tra le varie tipologie è implicita una gerarchia.Sia pure degradando l’informazione disponibile,le variabili quantitative continue possono essere discretizzate,le variabili quantitative discrete possono essere tradotte in variabili qualitative ordinali,le variabili qualitative ordinali possono essere considerate nominali.Non si può tuttavia ascendere nella gerarchia,ossia una scala nominale non può diventare ordinale e così via.Pertanto le analisi statistiche possibili rispecchiano la tipologia della variabile e sono per così dire più ricche di informazione ascendendo la gerarchia.

Relazioni tra le modalità delle scale di misurazione

Nella tabella si riporta uno schema riassuntivo delle principali scale di misurazione,in funzione delle relazioni che si possono instaurare tra le modalità.

	SCALE			
	QUALITATIVE		QUANTITATIVE	
RELAZIONI	**Sconnessa**	**Ordinale**	**Intervallare**	**Rapporto**
$x_i = x_j$	*	*	*	*
$x_i \neq x_j$	*	*	*	*
$x_i < x_j$		*	*	*
$x_i > x_j$		*	*	*
$x_i - x_j$			*	*
$x_i \diagup x_j$				*

In conclusione il tipo di scala di misura determina la quantità di informazione contenuta nei dati;altro fattore a determinare la quantità di informazione è l'errore presente nelle varie misure.L'errore può essere accidentale o sistematico.

L'errore accidentale è generato da un insieme di moltissimi fattori non controllabili e quindi non eliminabili.L'errore sistematico è dovuto a cause ben precise e quindi può essere corretto e rimosso una volta individuata la causa e l'entità.

4-Presentazione ed interpretazione dei risultati.

Lo statistico deve porre particolare cura nella presentazione dei risultati(sotto forma di tabelle,diagrammi,rapporti)e nella loro interpretazione.In assenza di una buona conoscenza dei

metodi utilizzati può recare danno,il ricorso ingenuo a software statistici. (Così la lettura di una ricerca sulla previsione della mortalità infantile in Italia suscita discussioni diverse da una ricerca sulla previsione delle presenze turistiche negli esercizi alberghieri,anche se entrambi gli studi potrebbero utilizzare metodi e software coincidenti).

5-Utilizzazione dei risultati della ricerca.

L'uso dei risultati va riferito al fenomeno per il quale la rilevazione è stata avviata.Così un'indagine telefonica condotta durante un giorno festivo tra le ore 10:00 e le ore 12:00 non è utile per conoscere le scelte politiche degli italiani,poiché una fetta consistente della popolazione non è presente in quelle ore.

In quale di queste fasi si gioca maggiormente la qualità scientifica di una ricerca?Qual è la fase più critica sul piano metodologico?

Secondo il Prof. Luigi Ferrari,presidente dell'ASSIRM,l'associazione che raggruppa gli istituti Italiani di Ricerche di Mercato,Sondaggi d'Opinione e Ricerche sociali,"la fase della raccolta dei dati è senz'altro quella più critica e determinante per la qualità della ricerca;se i dati raccolti sono scadenti,non c'è disegno campionario o tecnica di analisi che possa salvare il prodotto di ricerca.Al contrario,se i dati raccolti sono buoni,l'analisi potrà magari essere

mediocre,ma difficilmente potrà dar luogo a errori grossolani".

A conclusione di questa prima parte, segnaliamo due libri resi disponibili nella traduzione italiana,che trattano essenzialmente il problema dell'analfabetismo statistico e i pericoli che possono derivare dall'uso improprio di metodi e modelli statistici.Del primo è autore Gerd Gigerenzer(1),direttore del Center for Adaptive Behavior and Cognition del Max Plance Institut di Berlino.Punto di partenza del libro è la constatazione di come siano numerose le persone che non esitano quotidianamente ad utilizzare dati statistici di varia natura senza riuscire ad interpretarli criticamente.

Il secondo volume è di Nicholas Dunban(2),editor tecnico-scientifico della rivista "Risk".Si racconta come alcuni "visionari"(così li definisce l'autore)avevano accarezzato l'idea di rendere scientifico ciò che fino ad allora era stato pura speculazione e scommessa.

(1)G.Gigerenzer,CalculatedRisks,2002(nella trad.italiana:Quando i numeri ingannano.Imparare a vivere con l'incertezza,Milano,2003

(2)N.Dunbar,InventingMoney,Chichester,2000(traduz.italina:Anche I Nobel perdono.Idee,persone e fatti della finanza,Milano,2003.

1.3-Premettiamo alcuni richiami di calcolo combinatorio alle Indagini per campione

CALCOLO COMBINATORIO

Dato un insieme di oggetti,il calcolo combinatorio fornisce i criteri per configurare i raggruppamenti che si possono formare con tali oggetti.Siano ad esempio a,b,c e d quattro oggetti,si immagini di dover formare gruppi di due elementi.La costruzione dei gruppi presuppone la scelta di un criterio in base al quale si possa dire se due gruppi sono uguali oppure diversi.

Supponiamo,quindi, che lo spazio consista di N=4 elementi, che indichiamo con le prime quattro lettere dell'alfabeto {a,b,c,d}, e che da essi si intende sceglierne n=2.

Disposizioni con ripetizione

Si definiscono disposizioni con ripetizione i gruppi di N elementi scelti ad n ad n in modo che vengano ritenuti differenti *due gruppi se differiscono tra loro per un elemento oppure per l'ordine degli elementi.* Tutti i gruppi sono pari a

$$N^n$$

per i nostri dati $4^2 = 16$

Disposizioni senza ripetizione

Si definiscono disposizioni senza ripetizione di N elementi ad n ad n i *gruppi di n elementi che differiscono tra loro per almeno un elemento oppure per l'ordine degli elementi.*
Il numero delle disposizioni senza ripetizione di N elementi ad n ad n è dato dal prodotto dei primi n numeri interi decrescenti a partire da N:

$$D_{N,n} = N(N-1)(N-2).....(N-n+1) = \frac{N}{(N-n)}$$

per i nostri dati :

$D_{n,m} = 4(4-1) = 4$ **x3=12**

Permutazione

***Q*uando n=N si parla di permutazione di N elementi perché la *procedura equivale ad un rimescolamento dell'intero gruppo dove due elementi sono distinguibili solo per l'ordine;* in tal caso,ponendo n=N,si ottiene:**

$P_n = n(n-1)(n-2).....2$ **x1=n!**

cioè il prodotto dei primi n numeri interi.Tale prodotto è indicato con il simbolo di n!.

Per i nostri dati le permutazioni sono :

$P_4 =$ **4x3x2x1=4!=24**

Combinazioni

Si definiscono combinazioni di N elementi ad n ad n *i gruppi di n elementi differenti che si possono formare in modo che due gruppi differiscono tra loro per almeno un elemento.*

$$C_{n,m} = C_{N,n} = \left(\frac{N}{n}\right) = \frac{N}{n.(N-n)}$$

Ad esempio,date le quattro lettere a,b,c e d le combinazioni di classe due sono:

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

chiamiamo con ( ) le combinazioni di N oggetti di classe n,( ) prende il nome di coefficiente binomiale,

0!= =1

1.4-INDAGINI PER CAMPIONE

Le informazioni attorno alla popolazione,ossia attorno alle variabili che la caratterizzano,possono

essere il frutto di una rilevazione totale o censuaria,oppure di una rilevazione campionaria.Nel primo caso si procede alla disamina di tutte le unità della popolazione;la rilevazione totale presenta molte limitazioni dovute principalmente a tre fattori:costi,tempi di esecuzione e livello di precisione.Nel secondo caso si limita l'analisi ad una parte delle unità,ad un campione,avendo tuttavia come obiettivo lo studio dell'intera popolazione.

Nel linguaggio corrente la parola campione significa parte di un tutto,sottoinsieme di una totalità di elementi che viene assunto a"rappresentare"la totalità stessa.

Il concetto di campione non si discosta da questa nozione intuitiva:la totalità è qui costituita dalla popolazione oggetto dell'indagine;tuttavia,il campione statistico può anche non essere un sottoinsieme della popolazione in senso proprio.

Due problemi reali aiuteranno a capire il significato di "campione rappresentativo".

PRIMO ESEMPIO.

Agli albori delle trasmissioni televisive in Italia,un metodo seguito per qualche tempo per verificare l'audience e il gradimento di taluni programmi consisteva in una breve intervista a persone che si trovavano in talune strade o piazze di certe città fra le ore 9 e le ore 11 del giorno successivo alla messa in onda del programma in questione.Anche ammesso che la scelta del campione all'interno della popolazione osservata fosse casuale(ma non era certamente così),c'era una gran

maggioranza della popolazione che non aveva la possibilità di essere intervistata.

SECONDO ESEMPIO.

Per decidere sulla sede delle Nazioni Unite,una importante rete televisiva invitò il 20 settembre 1983, i suoi telespettatori a far conoscere per telefono la propria opinione in merito(Trattasi di un campionamento volontario).Le telefonate furono circa 180.000(un campione enorme);di queste il 33% erano a favore dell'idea che la sede dell'ONU restasse negli Stati Uniti,mentre l'ampia maggioranza del 67% sosteneva che l'ONU doveva andarsene.

Ma un campione veramente Casuale di 1200 interviste telefoniche eseguite nella stessa giornata ribaltò completamente la precedente conclusione:il 72% degli intervistati si dichiarò a favore della permanenza della sede ONU negli Stati Uniti e solo il 28% espresse l'opinione contraria.

Come per l'indagine totale,anche per l'indagine campionaria,è importante predisporre un piano di lavoro dove vengono definiti gli aspetti fondamentali dell'indagine.

Tali aspetti sono:

-formulazione degli obiettivi di lavoro;

-periodo di svolgimento e periodo di riferimento;

-determinazione della lista;

-scelta del piano di campionamento;

-metodo di raccolta dei dati;

-lavoro sul campo e addestramento dei rilevatori;

-elaborazione ed analisi dei dati;

-preparazione della relazione finale.

Il numero di unità che compongono il campione è detto *dimensione campionaria.*Il rapporto tra la dimensione campionaria *n* e quella della popolazione N viene chiamato *frazione di campinamento.*

L'elemento cruciale per la definizione del campione è dato dalla regola di selezione,ossia dalla procedura con la quale le unità campionarie sono estratte dalla popolazione.

Generalmente,la regola di selezione è di tipo probabilistico,cioè ogni unità della popolazione ha una probabilità prestabilita e non nulla di essere inserita nel campione.In tutti gli altri casi,si parla di campioni non probabilistici,in quanto prescindono dai criteri di casualità nella scelta delle unità campionarie.

CAMPIONI PROBABILISTICI.

Quando la procedura di scelta degli elementi dalla popolazione avviene mediante meccanismi di natura aleatoria si parla di campioni probabilistici.

Per definire una regola di selezione probabilistica si deve individuare:

.lo *spazio campionario*,formato da tutti i possibili campioni di numerosità n estraibili con una medesima tecnica da una popolazione(derivabili mediante un prefissato piano di campionamento definito su una popolazione finita);

.la probabilità di ogni campione di essere estratto.

La coppia ,(*spazio campionario,probabilità dei campioni)* è detta piano di campionamento; mentre lo

schema di campionamento è la procedura operativa per porre in essere un dato piano di campionamento cioè l'algoritmo mediante il quale si perviene alla effettiva selezione delle unità dalla popolazione(formalmente il piano di campionamento è l'associazione tra i campioni appartenenti allo spazio e le rispettive probabilità).

Un piano di campionamento è la definizione di una procedura di selezione di n unità statistiche,lo schema di campionamento è la procedura operativa.

Nel progettare un piano di campionamento sarà necessario tener conto del contesto in cui si opera e del tempo e delle risorse disponibili.
Illustriamo sinteticamente alcuni tra i principali piani campionari:

-campionamento casuale semplice con ripetizione;
-campionamento casuale semplice senza ripetizione;
-campionamento casuale stratificato;
-campionamento sistematico;
-campionamento casuale a grappoli;
-campionamento casuale a due stadi.

La realizzazione di un campionamento casuale può essere ricondotta agli schemi classici di estrazione da un'urna.

Operazione di estrazione delle unità.

Da un'urna contenente N palline di vario colore ne vengono estratte n.Le modalità di estrazione possono essere diverse:

1-Se dopo ogni estrazione la pallina viene reinserita nell'urna,i campioni distinti,ossia differenti tra loro per una o più unità,oppure per l'ordine di estrazione,sono tante quante le disposizioni con ripetizione di N elementi presi a n a n:sono cioè N^n.

2-Se dopo ogni estrazione la pallina non viene rimessa nell'urna,i campioni sono tanti quante le disposizioni di N elementi presi a n a n.Pertanto il numero dei punti campionari di S(SPAZIO CAMPIONARIO) è N(N-1)(N-2)…(N-n+1).

3-Se l'estrazione avviene in blocco,estraendo n palline dall'urna,per cui due gruppi differiscono per almeno un elemento,i campioni sono tanti quante le combinazioni di N elementi presi a n a n,cioè $\left(\frac{N}{n}\right)$.

La nostra attenzione sarà rivolta allo schema di estrazione con ripetizione e all'estrazione in blocco (cioè insieme,e non l'una dopo l'altra).

Nel primo caso i campioni sono tanti quante le disposizioni con ripetizione di N elementi presi a n a n,cioè N^n .

Nel secondo caso i campioni sono tanti quante le combinazioni di N elementi presi ad n a n,cioè $\left(\frac{N}{n}\right)$.

Nel primo caso una stessa unità può essere estratta e inserita nel campione più di una volta,talché il campione non è a stretto rigore un sottoinsieme della popolazione.

Con riferimento ai due schemi definiamo,per una popolazione composta di N=5 unità lo *spazio campionario.*

Esempio-Si consideri lo spazio campionario relativo ai campioni di 2 unità dalla popolazione composta da 5 persone aventi le seguenti stature:

166 169 172 173 178

I 25 campioni possibili campioni sono i seguenti:

166,166	169,166	172,166	173,166	178,166
166,169	169,169	172,169	173,179	178,169
166,172	169,172	172,172	173,172	178,172
166,173	169,173	172,173	173,173	178,173
166,178	169,178	172,178	173,178	178,178

Nel secondo caso lo spazio campionario sarà costituito da 10 campioni.Questo spazio differisce dal precedente per il fatto che in esso mancano i campioni i cui elementi sono uguali e i campioni i cui elementi sono permutati.osserviamo : se il campione che si vuole estrarre ha la stessa dimensione della popolazione cioè n=N lo spazio campionario sarà formato da $\binom{N}{N}$=1 campione. Schema di campionamento.

La procedura operativa per la selezione delle n unità statistiche dalla popolazione avviene tramite numeri casuali,in uno dei due modi seguenti:

a)in senso fisico,scegliendo n numeri da una tavola dei numeri casuali e individuando come unità statistiche da selezionare quelle aventi le etichette corrispondenti;

b)in senso algoritmo,utilizzando un software che genera n numeri pseudo-casuali tra 1 a N con probabilità costante $1/N$ e selezionando le unità statistiche corrispondenti agli n numeri pseudo-casuali.

Tavole dei numeri casuali

Le tavole dei numeri casuali forniscono serie di numeri disposti a caso tra lo 0 e il 9 aventi ciascuno la stessa frequenza.Esse possono essere utilizzate per estrarre a caso numeri di una o più cifre.Nello scegliere i numeri casuali il criterio da seguire può essere qualsivoglia,purché esso sia specificato prima di cominciare a impiegare la tavola.

Campionamento casuale stratificato

La popolazione iniziale costituita da N unità si suddivide in L sottopopolazioni o STRATI all'interno dei quali le unità siano omogenee secondo qualche criterio.

Da ciascun strato vengono poi estratte,tramite un campionamento casuale semplice,le unità da inserire nel campione.

Tale procedimento dà luogo al piano campionario noto come Campionamento Casuale STRATIFICATO.

Campionamento sistematico

Sia data una popolazione P le cui N unità sono numerate da 1 a N secondo un certo ordine.Si supponga che N sia multiplo di n, dove n è il campione che si vuole estrarre.Posto $N/n = k$ **,si estragga un numero casuale minore o uguale a k.Se r è il numero casuale estratto, si chiama Campione Sistematico l'insieme delle unità contraddistinte dalle etichette**

$$\{r, r+k, r+2k, ..., r+(n-1)k\}$$

Con questa tecnica,dunque,l'estrazione del numero casuale r identifica la prima unità che entra nel campione;le altre vengono individuate sequenzialmente:dopo l'unità r si contano k posizioni e si prende l'unità che occupa la posizione r+k,poi si prende l'unità r+2k,e così via,fino ad ottenere l'ampiezza voluta n.

Il numero casuale r verrà chiamato numero di partenza.Il rapporto $N/n = k$ **passo di campionamento(p.c.).Se N non è multiplo di n,come p.c. si assume l'intero k più vicino al rapporto** N/n **.**

Campionamento casuale a grappoli

Un GRAPPOLO(=CLUSTER) è un insieme di unità statistiche che sono "contigue" rispetto ad un criterio logico o naturale*.

Se una popolazione è suddivisa in N grappoli,il piano di campionamento a grappoli consiste nell'estrarre mediante un campionamento casuale semplice senza ripetizione n grappoli fra gli N possibili.Tutte le unità appartenenti ai grappoli prescelti fanno parte del campione.

***Il criterio che definisce un grappolo può essere di natura:**
a)fisica(edifici,comuni,regioni);b)amministrativa(residenti in un comune,appartenenti alle liste di collocamento);c)professionale(iscritti ad un ordine professionale);d)relazionale(familiari,amici..).

Campionamento a due stadi

Il piano consiste nell'estrarre,senza ripetizione,un campione casuale di grappoli,e nel selezionare,senza ripetizione,da ogni grappolo estratto un certo numero di unità elementari.

Come si vede,vi sono due livelli,due stadi,di campionamento:al primo vengono scelti i grappoli,al secondo le unità elementari.

I grappoli che costituiscono le unità di primo stadio,vengono anche chiamate unità primarie,mentre sono chiamate unità secondarie quelle estratte al secondo stadio e che costituiscono le unità elementari di interesse.

Osservazione.Il campionamento stratificato e quello a due stadi possono apparire in principio molto simili,perché ambedue creano

una partizione delle unità delle popolazione;in realtà sono diversi.Nel campionamento stratificato le unità all'interno dei strati sono il più possibile omogenee e si effettua l'estrazione da tutti gli strati.Nel campionamento a due stadi si formano dei raggruppamenti in modo che siano al loro interno il più possibile disomogenei e l'estrazione delle unità non viene effettuata da tutti i raggruppamenti bensì solo dai pochi raggruppamenti estratti.

Campioni non probabilistici

Vi sono altri metodi di formazione del campione,detti non probabilistici,poiché prescindono dai criteri di casualità nella scelta delle unità campionarie.Nel campionamento probabilistico ogni unità della popolazione ha una probabilità prestabilita e non nulla di essere inserita nel campione.Tra i metodi non probabilistici il più usato è il campionamento per quote.

Nei campioni per Quote le unità statistiche vengono scelte dal rilevatore in modo che il campione complessivo rispetti delle proporzioni predefinite da chi ha pianificato l'indagine ovvero secondo le proporzioni (quote)definite precedentemente per ogni variabile.

In sintesi,un campione per quote si costruisce come segue.La popolazione viene suddivisa in classi o sottogruppi omogenei,ad esempio secondo il sesso,secondo l'età,secondo la residenza,e così via.Dai dati censuari,o da altre fonti,si ricava il peso percentuale di ogni classe.Il totale delle unità nel viene poi suddiviso tra le classi in modo da rispecchiare le proporzioni esistenti nella popolazione.Si perviene dunque alla definizione delle quote,cioè del numero delle interviste

da effettuare in ciascuna classe.Ad esempio,quanti maschi e quante femmine da intervistare,quante persone per ciascuna delle classi di età definite,ecc.

L'elemento caratteristico del campionamento per quote è che la scelta delle unità da intervistare è demandata all'intervistatore stesso nell'ambito delle quote assegnate.

INDAGINI CAMPIONARIE NEL CAMPO TURISTICO

Le indagini campionarie nel campo turistico sono molto diffuse, e tale diffusione è ampiamente documentata.In campo turistico l'interesse prevalente che porta a realizzare un'indagine campionaria è quello di indagare sulle scelte,sui comportamenti,sulla soddisfazione di soggetti che si spostano in un'area per un certo periodo di tempo.Pertanto,la popolazione da indagare e formata prevalentemente da individui connotati dallo spazio e dal tempo.Dal punto di vista metodologico la popolazione di individui rientra nell'ambito delle popolazioni finite.

Spesso tuttavia non vengono descritti i piani di campionamento adottati,spesso si confonde l'unità di rilevazione,non sempre è chiaro il processo di selezione delle unità.Ciò è dovuto alle già esplicitate carenze conoscitive del metodo statistico,ma anche alle reali difficoltà che si devono affrontare per fare un piano di campionamento nel campo turistico.

L'individuazione delle unità statistiche che formano la popolazione finita costituisce un problema se si considera che nel campo turistico l'unità può essere

rappresentata da passeggeri, turisti, escursionisti, viaggiatori, ecc. che hanno una particolare collocazione nello spazio e nel tempo.Il carattere spazio-temporale della popolazione oggetto di studio costituisce un problema,soprattutto nell'interpretazione dei risultati,perché impone di analizzare flussi e dinamiche,spesso incognite,instabili e molto diverse rispetto alle aree di osservazione.In più,occorre riferirsi ad una precisa definizione della tipologia da osservare(turista,escursionista,ecc.).In generale esistono delle definizioni riguardanti il turista,l'escursionista o il passeggero,non sempre condivise.Ciò crea una confusione terminologica che non aiuta gli addetti ai lavori.Tale confusione cresce se si vuole analizzare un fenomeno turistico complesso,come ad esempio quello del sommerso.

L'identificazione delle unità che compongono il collettivo finito è un elemento necessario,ma non sufficiente per procedere con il campionamento.Per garantire la casualità dell'estrazione delle unità occorrerebbe disporre di una lista di tutte la unità che compongono la popolazione di riferimento:nessun ente preposto è in grado di fornire la lista completa di turisti,in una data area e/o in un dato periodo.

Vista l'impossibilità di ricorrere a campionamenti tradizionali(casuale,stratificato,a grappoli,ecc.),che necessitano della popolazione nota e fissa,in genere i campionamenti più opportuni,nel campo delle ricerche sul turismo,sono complessi,cioè derivati dalla combinazione di due o più tecniche di campionamento semplici o note.

2-Rappresentazioni delle Rilevazioni Statistiche

2.1-Introduzione

La semplice elencazione dei risultati di un’indagine definisce l’operazione di Rilevazione.La rilevazione associa a ciascuna Unità statistica una Modalità di X(la variabile che si intende studiare) ,indicata con il simbolo minuscolo

x_i , **i=1,2,…,N**

Rispetto alla semplice elencazione dei risultati di un’indagine,è preferibile presentare i dati statistici in una forma organizzata per semplificare confronti e analisi successive.

In generale,i dati statistici si possono presentare in forma enumerativa,tabellare o grafica.

Tra le rappresentazioni più comuni esamineremo le distribuzioni di frequenza,le matrici dei dati,le serie storiche e le serie territoriali.

Tra le rappresentazioni più comuni, la più importante è quella delle distribuzioni di frequenza le quali indicano come le unità della popolazione si distribuiscono rispetto alle modalità del carattere in esame.Se i caratteri rilevati sono più di uno,occorrerà esaminare la distribuzione congiunta delle coppie,terne,etc.di modalità all'interno dei soggetti che formano la popolazione.

Serie storiche.

L'insieme dei dati osservati in n tempi per p caratteri quantitativi si chiama serie storica (multipla se p>1).Nelle serie storiche il tempo rappresenta l'asse fondamentale lungo il quale può essere condotta l'analisi, cioè lo studio della dinamica del fenomeno nell'arco temporale considerato.

Serie territoriali

Le serie territoriali esprimono la distribuzione di un fenomeno in rapporto al territorio.

Matrice dei dati

E' una rappresentazione tabellare mediante la quale si schematizzano le informazioni(misure,risposte)raccolte su ciascuna unità in rapporto ad una molteplicità di fenomeni.Ogni colonna della matrice esprimerà una variabile o mutabile rilevata sulle diverse

unità;ciascuna riga esprimerà le misurazioni ottenute sulla singola unità.

2.2-Distribuzioni di Frequenza

La distribuzione di frequenza è una organizzazione dei dati in forma tabellare tale che ad ogni modalità di una certa variabile(qualitativa o quantitativa) si fa corrispondere la rispettiva frequenza(assoluta o relativa).Lo studio delle distribuzioni di frequenza ha senso compiuto sia per caratteri quantitativi (variabili)che qualitativi (mutabili) purché i metodi e le analisi si limitino,in questo secondo caso,ad operazioni aritmetiche sulle frequenze.(Se la variabile è qualitativa si parla di serie statistica,se la variabile è quantitativa si parla di seriazione).

Spoglio dei dati

Lo spoglio dei dati è l'operazione di riduzione statistica dell'informazione che consente di passare dalle osservazioni univariate o multivariate alla distribuzione di frequenza corrispondente.

Distribuzione di frequenza rispetto a un carattere quantitativo ovvero rispetto a una Variabile.

Una Variabile è una caratteristica delle unità statistiche che,al variare dell'Unità su cui è rilevata,può assumere una pluralità di valori.I valori assumibili da una Variabile sono detti Modalità della variabile.

Una ulteriore distinzione tra le distribuzioni di frequenza deriva dalla natura delle variabili.Distinguiamo:1)le distribuzioni di frequenza per variabili discrete;2)le distribuzioni di frequenza per variabili continue.

Distribuzioni di frequenza per variabili discrete

Se in una popolazione composta di N unità una Variabile X assume k modalità distinte $(x_1,..,x_i,..,x_k)$ **in modo tale che la modalità** x_1 **si presenta** n_1 **volte,….,la modalità** x_k **si presenta** n_k **volte,queste informazioni possono essere rappresentate in forma tabellare mediante la seguente distribuzione di frequenza della variabile X:**

Modalità della variabile X	Frequenze assolute	Frequenze relative
x_1	n_1	n_1/N
x_2	n_2	n_2/N
.		
.		
x_i	n_i	n_i/N
.		
.		
x_k	n_k	n_k/N

Totale N 1

Le quantità n_i **si definiscono frequenze assolute.Sono sempre numeri interi caratterizzati dalle proprietà:**

$$n_1 + n_2 + ... + n_k = \sum_{i=1}^{N} n_i = N; \qquad 0 \le n_i \le n \qquad \textbf{[2.1}$$

Le frequenze relative sono definite da

$$f_i = {}^{n_i}\!/_{N} \qquad \textbf{i=1,2,…,K}$$

$$f_1 + f_2 + ... + f_h = \sum_{i=1}^{K} f = 1; \qquad 0 \le f_i \le 1 \qquad \textbf{[2.2}$$

Distribuzioni di frequenza per Variabili Continue

Nel caso di una variabile continua non è possibile far corrispondere ai valori che essa assume le rispettive frequenze perché tra due modalità qualsiasi vene possono essere infinite altre;in questi casi occorre suddividere il campo di variazione di X in classi di modalità,cioè occorre esplicitare il criterio di chiusura delle classi. La generica classe viene definita come:

(x_{i-1}, x_i **], per** $i = 1,2,3,..,k$ **o notazioni simili** x_{i-1}**–|**x_i

ed in essa vanno incluse tutte le modalità di X comprese nell'intervallo reale $x_i < x \le x_i$

2.3-Funzione di ripartizione empirica

Il modo più semplice per presentare la distribuzione di una variabile statistica quantitativa X è attraverso la descrizione della frazione di unità statistiche per cui X non supera x_i**.**

Sia X una variabile statistica quantitativa univariata,sia $f(x_i)$ **la frequenza relativa della modalità** x_i**. Si dice funzione di ripartizione di X,e si indica con** $F_X(x)$**,la funzione definita da**

$$F_X(x) = (\,frequenzerelativedelleunitàcon\,\mathrm{mod}\,alità \le x) = \sum_{x_i \le x} f(x_i)$$

che associa ad ogni valore x_i **la frazione delle unità che sono inferiori o uguali(cioè, non superiori) a** x_i**.**

La funzione di ripartizione è solitamente espressa in termini di frequenze relative.Una funzione di ripartizione calcolata a partire da una variabile statistica osservata è detta empirica.

In pratica la funzione di ripartizione empirica è ottenuta cumulando progressivamente al crescere di X,le frequenze relative che per definizione,sono non negative. E' sempre non decrescente ed è compresa tra 0 e 1.

Pertanto la funzione di ripartizione empirica della variabile X soddisfa le seguenti proprietà:

- $0 \le F_x \le 1$
- F_x **è non decrescente** **[2.3**
- $F_{(-\infty)} = 0$
- $F_{(+\infty)} = 1$

La prima relazione afferma che F_x è compreso tra 0 e 1,limiti inclusi;la seconda precisa che la funzione non è decrescente.La non decrescenza della funzione deriva dal fatto che si cumulano quantità di valore positivo o al più nullo.Tale funzione è nulla per tutti i valori inferiori al più piccolo valore della variabile,di essere pari all'unità per tutti i valori superiori al massimo valore della variabile.

2.4-CONFRONTO TRA I DATI STATISTICI

Il confronto tra i dati statistici può essere realizzato per differenza oppure per rapporto;sono calcoli molto diffusi che si utilizzano per un esame preliminare dei dati statistici.Con il primo tipo di confronto emerge il divario assoluto tra i dati considerati,che si qualificherà in termini di eccedenza o di deficienza a seconda che la differenza risulti positiva oppure negativa;col secondo il divario relativo,cioè la proporzionalità tra numeratore e denominatore.

Il confronto per differenza implica,l'omogeneità delle grandezze comparate,condizione non necessaria per istituire un confronto per rapporto

DIFFERENZE

Tra le misure di due caratteri quantitativi le differenze possono essere assolute o relative.La differenza assoluta tra due modalità $x_1 e x_2$ di un carattere quantitativo X è definita mediante:

differenza= $x_2 - x_1$ [2.4

tale differenza è espressa nella stessa unità di misura del carattere.

La differenza assoluta,utile per il confronto tra fenomeni simili e stabili rispetto alle circostanze,non agevola il confronto tra fenomeni per i quali l'unità di misura oppure l'ordine di grandezza sono differenti.per questo si introduce la differenza relativa(o percentuale,se viene poi moltiplicata per 100)definita da:

$$\text{Differenza relativa}=\frac{x_2-x_1}{x_1} \qquad [2.5$$

Nella quale la modalità x_1 è antecedente sul piano logico(o temporale) rispetto a x_2 e,quindi,la variazione misurata dalla differenza relativa è espressa rispetto a x_1.*La differenza relativa è un numero puro,cioè non dipende dalla unità di misura delle modalità che si confrontano.*

Quando si osserva una variabile X rispetto al tempo,la differenza relativa tra la modalità assunta al tempo t_2 rispetto al tempo t_1, con $t_1<t_2$, viene definita tasso di variazione(percentuale,se il risultato è moltiplicato per 100) di X nel periodo $[t_1,t_2]$.

Quindi se a e b esprimono l'intensità o la frequenza di un fenomeno in due momenti diversi,avremo:

$$b-a= \text{ incremento assoluto} \qquad [2.6$$

$$\frac{b-a}{a}100 \text{= incremento relativo} \qquad [2.7$$

Esempio_ Il prezzo del biglietto per il trasporto urbano di una città è variato,da euro 1.4 a euro 1.5;la differenza assoluta di euro 0.100 corrisponde ad una differenza relativa di

$\left(\frac{(1.500-1.400)}{1.400}\right)=0.07143$ **cioè ad un incremento di circa il 7,1%.**

RAPPORTI STATISTICI

Una particolare categoria di misure elementari maggiormente consolidata è costituita dai rapporti statistici.

SI definiscono rapporti statistici quei rapporti istituiti tra quantità di cui almeno una possiede la caratteristica del dato statistico.

Essi devono riferirsi a fenomeni tra i quali sia ravvisabile un nesso logico:come di parte al tutto,di effetto a causa ecc..

I rapporti statistici non si possono "istituire indifferentemente fra termini presi ad arbitrio.Occorre,a fondamento del processo aritmetico,una RELAZIONE LOGICA"(Benini).

Elenchiamo i più importanti:

1 RAPPORTI DI COMPOSIZIONE;

2 RAPPORTI DI COESISTENZA;

3 RAPPORTI DI DERIVAZIONE;

4 RAPPORTI INDICI ovvero NUMERI INDICI.

1-RAPPORTI DI COMPOSIZIONE

Sono noti come rapporti di "parte al tutto" giacchè la grandezza posta al numeratore si quantifica come quota parte di quella che appare al denominatore.Il quoziente,quindi, è compreso tra 0 e 1 ed esprime pertanto la frazione relativa(o percentuale) posseduta o registrata dalla modalità rispetto al totale.

Esempio_Consideriamo 2.759 studentesse classificate secondo il diploma di provenienza

Diploma	Frequenza assoluta	Frequenza relativa
Classico	821	0,2976=$821/2.759$
Scientifico	637	0,2309=$637/2759$
Tecnico	1090	0,3951=$1.090/2759$
Altri	211	0,0765=$211/2.759$
TOTALE	2.759	1,0000

2_RAPPORTI DI COESISTENZA

Mettono a confronto le misure di un medesimo fenomeno in luoghi diversi o di fenomeni diversi nello stesso luogo.Riguardano ogni rapporto tra frequenza(o quantità) di una modalità rispetto alla frequenza(o quantità)corrispondente di un'altra modalità.

Un rapporto di coesistenza ragguaglia due parti di un complesso(che coesistono,appunto)illustrandone il reciproco ordine di grandezza.Tali indicatori sono molto utilizzati negli studi demografici ove assumono una terminologia specifica.

Tra i più comuni rapporti di coesistenza, ricordiamo i seguenti:

.rapporto di mascolinità definito

come il rapporto tra la frequenza della modalità "Maschi" e quella della modalità "Femmine" riferito ad un certo luogo,tempo o circostanza(per esempio alla nascita).

.indice di vecchiaia definito

come il rapporto tra i residenti con più di 65 anni e quelli con meno di di 15 anni;se moltiplicato per 100,esso misura quanti anziani vi sono per 100 giovani.

.indice di dipendenza degli anziani definito

"come il rapporto tra i residenti con età superiore a 65 anni e quelli di età compresa tra i 15 e 60 anni"; se moltiplicato per 100,esso misura quanti anziani sono economicamente sostenuti da 100 individui attivi.

3_RAPPORTI DI DERIVAZIONE

Con tali rapporti si confrontano dati statistici relativi a due fenomeni tra i quali intercorre una relazione,più o meno stretta,di casualità nel senso che uno di questi si qualifica come condizione generale di esistenza dell'altro.

Esempio:la popolazione è presupposto logico delle nascite delle morti,dei matrimoni,dei movimenti migratori.

Essi sono molto utilizzati in Demografia ove ricevono un nome specifico:*indice di natalità,indice di mortalità,etc.*Per tali indici,è importante stabilire correttamente il collettivo di riferimento più idoneo,ponendo al denominatore la effettiva popolazione che può aver generato il dato collocato al numeratore e non un collettivo generico di riferimento.

Ad esempio,osserva Piccolo,per il rapporto nati/popolazione bisognerebbe porre al denominatore le sole donne in età fertile stabilmente presenti nel territorio prefissato nei nove mesi precedenti il periodo cui fanno riferimento le nascite.

Poiché non sempre è possibile l'esatta individuazione del collettivo di riferimento,si distingue tra rapporti specifici e rapporti generici di derivazione.

4_RAPPORTI INDICI OVVERO NUMERI INDICI

Per una serie storica,*come sequenza di osservazioni di un fenomeno Y in T tempi*,é di grande interesse analizzare le variazioni tra due periodi di tempo contigui,attraverso rapporti, chiamati numeri indici semplici.

I numeri indici semplici si costruiscono,rapportando tra loro le intensità che uno stesso fenomeno presenta in tempi o in luoghi diversi.Il loro campo di applicazione è assai vasto potendo riguardare sia serie storiche che serie territoriali relative a fenomeni di varia natura.

Le serie dei numeri indici possono essere costruite in due modi diversi:*a base fissa o a base variabile*.

Una serie di numeri indici a base fissa si ottiene rapportando ciascun termine della successione ed un altro

della successione stessa che viene assunto come base.L'intensità del fenomeno posta a denominatore del rapporto,cioè la base,è abitualmente posta uguale a 100 onde interpretare il quoziente in termini percentuali.

Quindi un numero indice t con base s si ottiene dall'espressione $It/s = y_t / y_s .100$ **[2.8**

I numeri indici a base variabile sono ottenuti rapportando l'intensità o la frequenza di un certo periodo con l'intensità o la frequenza del periodo immediatamente precedente.

Quindi un numero indice a base variabile si ottiene dall'espressione $I_{t/\,t-1} = y_t / y_{t-1} .100.$ **[2.9**

L'andamento delle variazioni lungo il tempo di un certo fenomeno può essere descritta attraverso la serie percentuale dei numeri indici a base mobile.

Quindi,se la serie temporale degli indici a base fissa permette di stabilire di quanto è variato nel tempo il fenomeno considerato rispetto ad un unico tempo assunto come base,mediante la serie temporale degli indici a base variabile si determinano le variazioni successive che intervengono, di volta in volta,tra i tempi consecutivi che caratterizzano la serie.

Le relazioni che intercorrono tra una serie a base fissa e una serie a base mobile sono date dalle seguenti proprietà.

PROPRIETA' TRANSITIVA

La proprietà transitiva indica il procedimento mediante il quale si realizza la trasformazione di una

serie di numeri indici a base variabile in una serie di indici a base fissa.

In una serie storica di Numeri indici a Base Variabile,con base al tempo immediatamente precedente,il prodotto dell'indice relativo al tempo t(con base,al tempo t-1) con tutti gli indici precedenti determina l'indice relativo al tempo t con base al tempo iniziale:

$$I_{1/0}.I_{2/1}.I_{3/2}....I_{t/t-1} = I_{t/0}$$

Osserviamo che:
a-ciascun rapporto è una variazione relativa;
b-il prodotto delle variazioni relative riproduce la variazione relativa dell'intero periodo.

Prospetto 1-*Trasformazione della serie di indici semplici a base variabile in una serie di indici a base fissa*

Anni	Numeri indici a base variabile	Numeri indici a base fissa(base:2000=100)	
2000	-		=100
2001	121,2	121,2	=121,2
2002	122,9	$\frac{121,2\times122,9}{100}$	=149,0
2003	109,5	$\frac{121,2\times122,9\times109,5}{100^2}$	=163,2

PROPRIETA'CIRCOLARE-

La serie di indici a base fissa si trasforma in una serie di indici a base variabile dividendo ciascun indice della serie per quello che immediatamente lo precede. (Prospetto 2).

PROSPETTO 2-*Trasformazione della serie di indici a base fissa(base:2000=100) in una serie di indici a base variabile.*

ANNI	Numeri indici a base fissa(base:2000=100)	Numeri indici a base variabile
2000	100,0	-
2001	121,2	$\frac{121,2}{100}\times 100 = 121,2$
2002	149,0	$\frac{149,0}{121,2}\times 100 = 122,9$
2003	163,2	$\frac{163,2}{149,0}\times 100 = 109,5$

Dati tre tempi 0,s,t,vale la relazione

$$I_{s/0} \times I_{t/s} \times I_{0/t} = 1$$

Questa proprietà ci indica un principio di coerenza senza il quale non avrebbero senso operazioni,quali il cambio di base.

SLITTAMENTO DELLA BASE FISSA DI UNA SERIE DI INDICI SEMPLICI DAL TEMPO 0 AL TEMPO s

La serie di indici a base fissa,di cui al prospetto 3,per trasformarla in una nuova serie a base fissa,è sufficiente dividere ogni indice della serie iniziale per l'indice della nuova base

PROSPETTO 3-*Trasformazione della serie di indici,a base fissa (base:2000=100),in una nuova serie di indici a base fissa(base:2002=100).*

ANNI	Numeri indici a base fissa(base:2000=100)	Numeri indici a base fissa(base:2002=100)
2000	100,0	$\frac{100,0}{149,0}\times 100 = 67,1$

2001	**121,2**	$\frac{121,2}{149,0}\times 100 = 81,3$
2002	**149,0**	$\frac{149,0}{149,0}\times 100 = 100,0$
2003	**163,0**	$\frac{163,2}{149,0}\times 100 = 109,5$

3-Sintesi della distribuzione di un carattere-Le Medie

3.1- Medie analitiche

Uno dei compiti principali della Statistica consiste nel sintetizzare in alcune costanti, particolari aspetti del fenomeno in studio.Una prima categoria di tali costanti sintetiche è costituita dai valori medi,che se calcolati su dati quantitativi ,ne pongono in evidenza la dimensione,ossia il loro ordine di grandezza.

Una seconda categoria è costituita dagli indici di variabilità che misurano l'attitudine che hanno i caratteri ad assumere valori diversi;altri aspetti possono essere sintetizzati da altre costanti,quali asimmetria e kurtosi.

Di ciascun aspetto contenuto nei dati ed evidenziato dalle costanti sintetiche risulta importante il significato del tipo di informazione in essa contenuta perché aiuta a descrivere e a capire il fenomeno.

E’ importante quindi chiarire le idee di base cui si fondano i diversi processi razionali che guidano alla definizione di ciascun tipo di costante sintetica.

Il valor medio individua una delle possibili forme di sintesi operabili nei confronti delle distribuzioni statistiche:esso si qualifica come astrazione dal reale poiché,per il suo tramite,implicitamente si assume a modello distributivo una successione di intensità tutte uguali tra loro al posto della distribuzione variabile delle intensità rilevate.

Il termine *media* evoca il concetto che vuole il valore di sintesi di una pluralità di intensità intermedio tra l’intensità meno elevata e quella più elevata.Un tale concetto riconducibile a Cauchy non aiuta ad individuare la media,poiché tra due valori esistono un numero notevole di medie;non è idonea a fornire criteri operativi per la determinazione del valore medio da assumere a sintesi della distribuzione statistica.Movendo da queste critiche il Chisini propose una definizione che consentiva di pervenire ad una espressione analitica della media.

MEDIE DEL CHISINI

La definizione di *media* proposta dal Chisini trova la seguente formalizzazione matematica:
sia

1_ $f(x_{1,}x_2,..,x_k)$ **una prefissata funzione delle k modalità** x_i **si dice *media* rispetto alla funzione f,il valore costante** $\bar{x}$ **che,sostituito a ciascuna delle k modalità** x_i **che figurano nella 1 realizza la condizione:**

$$f(x_1,....,x_k) = f(\bar{x},.....,\bar{x}) \qquad [3.1$$

ESEMPIO-Ipotizziamo che la distribuzione del reddito annuo di 100 soggetti ha dato luogo alla seguente tabella

Tab. Distribuzione del reddito annuo (in migliaia di euro)percepito da 100 individui.

x_j	n_j	$x_j n_j$
7	40	280
10	25	250
15	20	300
25	8	200
30	4	120
50	3	150
	100	1.300

La definizione di media proposta dal Chisini,per i nostri dati comporta la seguente formalizzazione:

$$f(x_1.....x_k;n_1,...,n_k) = f(M,.....M;n_1,.....,n_k)$$

$$\sum_{j=1}^{6} x_j n_j = M\sum_{j=1}^{6} n_j$$

da cui: $$M=\frac{\sum_{j=1}^{6} x_j n_j}{\sum_{j=1}^{6} n_j}$$

In generale:

$M = \frac{\sum_{j=1}^{k} x_j}{N}$ **per una serie di valori;** **[3.2**

$M = \frac{\sum_{j=1}^{k} x_j n_j}{N}$ **per una distribuzione di frequenza** **[3.3**

LA MEDIA ARTMETICA trova il suo impiego in presenza Di un fenomeno per cui sia compatibile un modello di NATURA ADDITIVA,che prevede cioè che l'intensità Totale del fenomeno medesimo possa esprimersi come Somma di tutte le sue manifestazioni

Proprietà della Media Aritmetica

1- ***Proprietà di identità di somma***

Nel contesto della impostazione del Chisini,ritroviamo l'espressione della media aritmetica imponendo la condizione di invarianza della funzione"somma delle intensità"ossia

$$x_1 + x_2 + \cdots + x_k = M + M + \cdots + M$$
$$\sum_{i=1}^{k} x_i = kM \qquad \textbf{[3.4}$$

"Il valore M se viene sostituito alle singole intensità x_i***,in numero di N,lascia inalterata l'intensità globale che il carattere X presenta nel collettivo".***

2-Proprietà di nullità della somma algebrica degli scostamenti

La differenza $x_i - M$ **=scarto o scostamento individua uno degli N scostamenti che possono essere determinati nei confronti della media aritmetica.**

"La somma algebrica degli scostamenti dalla media aritmetica è nulla"

Poiché

$$kM = \sum_{i=1}^{k} x_i$$

$$kM - \sum_{i=1}^{k} x_i \qquad \textbf{[3.5}$$

0 anche

$\sum (x_i - M) = 0$ **[3.6 per una serie di valori**

$\sum (x_i - M) n_i = 0$ **[3.7 per una distribuzione di frequenza**

3-*Proprietà di minimo*

Consideriamo l'espressione

$\sum (x_i - A)^2$ **aggiungiamo e togliamo M**

$\sum (x_i - A)^2 = \sum (x_i - M + M - A)^2$ *sviluppiamo* **il quadrato a destra della relazione**

$$\sum (x_i - A)^2 = \sum [(x_i - M) + (M - A)]^2 =$$

$$= \sum (x_i - M)^2 + \sum (M - A)^2 + 2\sum (x_i - M)(M - A)$$

$$= \sum (x_i - M)^2 + k(M - A)^2 + 2(M - A)\sum (x_i - M)$$

per la prima proprietà della media aritmetica $\sum (x_i - M) = 0$

$$\sum (x_i - A)^2 = \sum (x_i - M)^2 + k(M - A)^2$$

poiché $k(M-A)^2 \geq 0$ **risulta**

$$\sum (x_i - A)^2 \geq \sum (x_i - M)^2 \qquad \textbf{[3.8}$$ **per una serie di valori**

$$\sum (x_i - A)^2 n_i \geq \sum (x_i - M)^2 n_i \qquad \textbf{[3.9}$$ **per una distribuzione di frequenza**

"la somma del quadrato degli scostamenti dalla media aritmetica è un minimo nel senso che è minore di tutte le altre somme di quadrati di scostamenti determinati nei confronti di qualsiasi altro valore diverso dalla media aritmetica".

4-*Proprietà di omogeneità*

Se X presenta modalità $\{x_1, x_2, \cdots, x_i, \cdots, x_k\}$ **con media aritmetica** $M(X) = \dfrac{(x_1 + \cdots + x_k)}{k}$ **se moltiplichiamo le modalità per b,ossia** bx_i**,si avrà**

$$M(bX) = bM(X) \qquad \textbf{[3.10}$$

"Se le intensità di una distribuzione vengono moltiplicate(o divise) tutte per una stessa quantità b si ottiene una nuova distribuzione la cui media aritmetica corrisponde alla media aritmetica della distribuzione moltiplicata(o divisa) per b"

5- ***Proprietà traslativa***

Se si aggiunge una stessa quantità a alle modalità x_i **,ossia** $x_i + a$ **si avrà che**

$$M(X + a) = M(X) + a \qquad \textbf{[3.11}$$

"Se alle intensità di una distribuzione si addiziona o si sottrae una stessa quantità a si ottiene una nuova

distribuzione la cui media aritmetica corrisponde alla media aritmetica della distribuzione originaria aumentata(o diminuita) della quantità a"

6-*Proprietà della linearità*

Se X presenta modalità x_i**,ed operiamo la trasformazione lineare** $y_i = a + bx_i$**,la media di**

$$M(Y) = a + bM(X) \qquad \textbf{[3.12}$$

dove a e b sono costanti.

Tale proprietà implica la proprietà di omogeneità(basta porre a=0) e la proprietà traslativa(basta porre b=1).

7-Proprietà associativa

Si abbiano k serie di dati ignoti,ognuna costituita da n_i **osservazioni ed avente per media** M_i

1^a *serie*	2^a *serie*	.....	k^a *serie*
x_{11}	x_{21}	.....	x_{k1}
x_{12}	x_{22}	.....	x_{k2}
. . .	. . .		. . .
x_{1i}	x_{2i}	.....	x_{ki}
. . .	. . .		. . .
x_{1n1}	x_{2N2}	.....	x_{knk}
M_1	M_2	.....	M_k
n_1	n_2	.....	n_k

Se si conoscessero tutti i dati,la media totale risulterebbe espressa da

$$M_{tot} = \frac{\sum x_{1i} + \sum x_{2i} + \cdots + \sum x_{ki}}{n_1 + n_2 + \cdots + n_k} \qquad \textbf{[3.13}$$

dove ciascuna sommatoria contiene tutti i dati relativi a ciascun gruppo parziale.

Note Le medie di questi gruppi parziali, espresse da

$$M_1 = \frac{\sum x_{1i}}{n_1} \ldots\ldots\ldots\ldots \quad M_k = \frac{\sum x_{ki}}{n_k}$$

La media totale è data da

$$M_{tot} = \frac{\sum M_i n_i}{N} \qquad \textbf{[3.14} \qquad \textbf{con} \qquad N = \sum n_i$$

"La media aritmetica complessiva di più gruppi parziali è uguale alla media aritmetica ponderata delle medie parziali,con pesi uguali al numero dei dati di ciascun gruppo"

Se tutti i gruppi hanno lo stesso numero di dati ($n_1 = n_2 = \cdots = n_k$ **),la media totale è uguale alla media semplice delle k medie parziali,ossia**

$$M_{tot} = \frac{\sum M_i n_i}{\sum n_i} = \frac{n \sum M_i}{kn} = \frac{\sum M_i}{k} \qquad \textbf{[3.15}$$

Media geometrica

Si supponga, che il capitale di 1 euro depositato presso una banca per un triennio frutti l'interesse i dell'8% il primo anno,del 10% il secondo anno e del 12%

il terzo anno.Si desidera valutare il tasso annuo che mediamente è stato corrisposto nei tre anni.

In tale situazione alla fine del primo anno la somma depositata C pari ad 1 euro ha maturato interessi pari 1×0,08,di modo che il montante risulta:

$$M_1 = 1 + (1 \times 0{,}08) = 1 + 0{,}08$$

Nel secondo anno questa somma ha maturato interessi pari a $M_1 \times 0{,}10$ **di modo che alla fine del secondo anno il montante risulta**

$$M_2 = M_1 + M_1 \times 0{,}10 = M_1(1 + 0{,}10) = (1 + 0{,}08)(1 + 0{,}10)$$

Analogamente,il montante alla fine del terzo anno risulta pari a:

$$M_3 = M_2 + M_2 \times 0{,}12 = M_2(1 + 012) = (1 + 0{,}08)(1 + 0{,}10)(1 + 0{,}12) = 1{,}3306$$

Desideriamo trovare il tasso di interesse i,costante,per tutto il triennio,capace di assicurare lo stesso montante M=1,3306.La soluzione è evidentemente la media geometrica dei montanti di fine anno e non dei tassi di interesse.

Deve essere allora:

(1,08)(1,10)(1,12)= $(1 + \bar{i})(1 + \bar{i})(1 + \bar{i}) = (1 + \bar{i})^3$

da cui

$$(1+\bar{i}) = \sqrt[3]{(1{,}08)(1{,}10)(1{,}12)} = 1{,}09988$$

ossia la media geometrica dei tre fattori di montanti annui:(1,08),(1,10),(1,12).

Il tasso di interesse medio composto da applicare al capitale unitario per ricevere dopo 3 anni ,lo stesso montante:

$\bar{i} = 1{,}09988 - 1 = 0{,}9988$ **ossia il 9,988% annuo.Tale tasso si chiama,appunto, tasso medio composto.**

Si noti che il tasso medio composto non coincide con la media geometrica dei tassi annui:

$$\sqrt[3]{0{,}08 \times 0{,}10 \times 0{,}12} = \frac{1(Log\,0{,}08 + Log\,0{,}10 + Log\,0{,}12)}{3} = -1{,}0059$$

$$M_g = 0{,}09865$$

ossia 9,865%

Un ragionamento analogo può essere fatto per le variazioni di popolazioni di cui si conoscono i tassi di accrescimento nei vari periodi e si vuole stimare un tasso medio dell'intero periodo.

"La media geometrica,pertanto,viene impiegata col proposito di sintetizzare l'ordine di grandezza di un fenomeno per cui sia compatibile un modello di NATURA MOLTIPLICATIVA,un modello cioè che prevede che l'intensità totale del fenomeno coincida con il prodotto delle sue singole manifestazioni"

In generale si chiama media geometrica il seguente valore:

$$M_g = \sqrt[k]{\prod_{i=1}^{k} x_i} \qquad \text{[3.15}$$

Una proprietà notevole della media geometrica è la seguente:

***"la media geometrica di più rapporti è uguale al rapporto tra la media geometrica delle grandezze poste al numeratore e la media geometrica di quelle che figurano al denominatore"*:**

$$\sqrt{\frac{x_1}{y_1} \cdots \frac{x_k}{y_k}} = \frac{\sqrt[k]{x_1 \cdots x_k}}{\sqrt[k]{y_1 \cdots y_k}} \qquad \text{[3.16}$$

In pratica,il calcolo della media geometrica viene realizzato,applicando i logaritmi ad entrambi i membri della relazione [3.15 e con l'impiego della seguente proprietà:

" il logaritmo di una radice è uguale al logaritmo del radicando diviso l'indice del radicale($\log_a \sqrt[n]{b} = \log_a (b)^{\frac{1}{n}} = \frac{1}{n} \log_a b$ ***)"***

Ossia

$$\log M_g = \frac{1}{k} \log \Pi x_i = \frac{1}{k} \sum \log x_i \qquad \text{[3.17}$$

da cui emerge che il logaritmo della media geometrica coincide con la media aritmetica delle k modalità x_i **(i**

logaritmi possono essere in base qualunque anche se la consuetudine preferisce quelli decimali o quelli neperiani).Tramite poi l'antilogaritmo si perviene alla media geometrica.

Si ricorda che il logaritmo di un numero b in base a è l'esponente x a cui occorre elevare a per avere b:

se $a^x = b \qquad \log_a b = x$

esempio: $\log_2 8 = 3$ **poiché** $2^3 = 8$

antilogaritmo:se $x = \log_a b \qquad b = a^{\log_a b}$

Media Armonica

Con riferimento alla seguente tabella

Distribuzione dei prezzi unitari p_j **(in euro) e delle spese** s_j **(in migliaia di euro)sostenute per l'acquisto delle quantità** q_j **di un prodotto di cinque marche diverse**

p_j	s_j	q_j
16	**240**	**15**
25	**450**	***18***
32	**640**	**20**
40	**400**	**10**
50	**400**	**8**
	2.130	**71**

nella quale sono riportati i prezzi unitari p_j **delle cinque marche e le corrispondenti spese** s_j **per l'acquisto delle quantità** q_j **si chiede di calcolare il prezzo medio.**

Il prezzo medio è pari alla media armonica dei prezzi ossia:

$$M_{-1} = \frac{2130}{71} = 30$$

Il motivo del ricorso alla media armonica dei prezzi è chiaro ricordando alcune relazioni.

La relazione fra spesa (s),prezzo(p) e quantità(q) è la seguente:

s=p.q da cui q=s/p

Il prezzo medio $\bar{p}$ **che lascia immutata la quantità q è la soluzione dell' equazione:**

$$\frac{s_1}{p_1}+\frac{s_2}{p_2}+....+\frac{s_5}{p_5}=\frac{s_1+s_2+....+s_5}{\bar{p}}$$

da cui

$$\bar{p}=\frac{\sum_{j=1}^{5} s_j}{\sum_{j=1}^{5}\frac{s_j}{p_j}}$$

che coincide con la media armonica dei prezzi p_j **avendo considerato quali frequenze le spese** $s_j\,(j=1,2,...,5)$**.**

In generale:

$$M_{-1}=\frac{N}{\sum_{j=1}^{k} 1/x_j}$$ **[3.18 per una serie di valori;**

$$M_{-1}=\frac{N}{\sum_{j=1}^{k} n_j/x_j}$$ **[3.19 per una distribuzione di frequenza.**

"La Media Armonica viene impiegata per sintetizzare l'ordine di grandezza di un fenomeno il quale si esprime come il reciproco di un altro di natura additiva"

Medie potenziate.

Tutte le medie sin qui esaminate sono casi particolari di una classe di medie,secondo il Chisini,definite come *medie potenziate.*

Si dice media potenziata di ordine r di X,per modalità tutte positive il valore

$$M_X = \left(\frac{\sum x_i^r n_i}{\sum n_i} \right)^{\frac{1}{r}} \qquad [3.20$$

cioè la radice r-esima della media aritmetica delle potenze r-esime delle osservazioni.

Si osservi che la media potenziata di ordine 1,per r=1,è la media aritmetica.

Per r=-1 si ottiene la media armonica.

Per r=2 si ottiene la media quadratica.

Per r che tende a zero si ottiene la media geometrica.

3.2-Medie di posizione:VALORE MEDIANO-MODA

QUANTILE: si definisce quantile ogni valore particolare di una successione ordinata di valori che suddivide la serie in q+1 parti di uguale numerosità.

Se q=1 avremo 1+1= parti uguali
Il Quantile corrispondente prende il nome di VALORE MEDIANO.

Se q=3 avremo 3+1= 4 parti uguali
I tre Quantili prendono il nome di Primo-Secondo-Terzo Quartile.

Il quantile-0,5,ossia la *MEDIANA* di X rappresenta quel valore che, rispetto all'ordinamento crescente delle osservazioni,risulta preceduto e seguito dalla stessa porzione di osservazioni(il 50%),a meno degli effetti della discretezza.Più precisamente,è *mediana* di X un valore $x_{0,5}$ che soddisfa la seguente doppia limitazione:

$$\sum_{m_j \le m_{0,5}} f_j \ge \frac{N}{2};$$

$$\sum_{m_j \ge m_{0,5}} f_j \ge \frac{N}{2}$$

dove con $\sum_{m_j \le m_{o,5}} f_j$ e $\sum_{m_j \ge m_{0,5}} f_j$ si intendono le somme delle frequenze associate a modalità m_j non maggiori e, rispettivamente,non minori di $m_{0,5}$

In pratica, tale definizione trova le specificazioni operative che seguono:

1-per il caso discreto,la posizione mediana è definita dal numero d'ordine $\frac{N+1}{2}$ nel caso in cui N è dispari,mentre è definita dalla posizione media tra $\frac{N}{2} e \frac{N}{2}+1$ nel caso pari.

In particolare,se n è un numero dispari,la *mediana* $m_{0,5}$ **è rappresentata dalla modalità associata all'unità che fra le N occupa la posizione centrale,ovvero la** $\left[\frac{(N+1)}{2}\right]$**-esima posizione.**
Detta unità,infatti,è preceduta da (N-1)/2unità caratterizzate da modalità $m_j \leq m_{0,5}$ **e seguita da altrettante con modalità** $m_j \geq m_{0,5}$**.**

Se N è un numero pari,invece,esistono due elementi separatori,ossia le due unità che occupano le posizioni(N/2)-esima e,rispettivamente (N/2)+1-esima.
Ora, se a dette unità corrisponde la medesima modalità,questa assume il significato di *mediana*;viceversa, se alle suddette due unità corrispondono modalità diverse è consuetudine identificarla con la semisomma delle predette modalità.

Quando i dati vengono presentati mediante una distribuzione di frequenze di un carattere quantitativo suddiviso in classi non e possibile individuare esattamente la mediana.Tuttavia,in questo caso,è possibile ottenere una sua approssimazione attraverso la seguente formula:

$$M_e \approx x_{i-1} + (x_i - x_{i-1})\frac{0,5 - F_{i-1}}{F_i - F_{i-1}} \qquad \textbf{[3.21}$$

dove

x_{i-1}: **estremo inferiore della classe mediana**

x_i: **estremo superiore della classe mediana**

$x_i - x_{i-1}$ **ampiezza della classe mediana**

F_{i-1} **è la frequenza relativa cumulata fino alla classe precedente a quella mediana**

F_i **è la frequenza relativa cumulata fino alla classe mediana.**

Moda

Nel linguaggio corrente si intende per moda un comportamento,un vestito,un oggetto e non la frequenza delle persone che adotta quel comportamento, indossa il vestito,compra quell'oggetto.Come tutti gli indici di posizione,la moda deve essere espressione del fenomeno oggetto di studio,quindi deve essere un valore scelto tra quelli manifestati dalla variabile nella popolazione:la moda è una modalità,non una frequenza ossia *è quella modalità cui corrisponde la massima frequenza.*

3.3-BOX e WHISKERS PLOT

Arrivati a questo punto,possiamo introdurre un metodo di rappresentazione grafica,detto box-plot,che si avvale di talune medie di posizione e che risulta utile nella comparazione di due o più collettivi.

I capisaldi per tale rappresentazione sono i seguenti valori:

$$Q_0 = \min(x_i);\ Q_1 = 1^0\, quartile;\ Q_2 = mediana \quad Q_3 = 3^0\, quartile \quad Q_4 = \max(x_i)$$

e la differenza interquatile

$$\textbf{IQR=}Q_3 - Q_1$$

Tramite tali quantità,si determina il box-plot

In un prefissato asse (orizzontale ma può anche essere verticale),la cui scala è quella della variabile in esame,si individuano barre in corrispondenza della mediana e del primo e terzo quartile,che poi si chiudono sino a formare una scatola.(La base della scatola rettangolare è posizionata sullo scarto interquatilico).

All'interno di questa scatola viene segnata con un segmento verticale,la *mediana*.Le tre linee verticali tracciate,una in corrispondenza del primo quartile,una in corrispondenza del terzo quartile e una in corrispondenza della mediana hanno la stessa altezza.

Tale diagramma può essere utilizzato per identificare eventuali osservazioni anomale dei dati.

In situazioni normali è lecito attendersi che la maggior parte delle osservazioni stiano all'interno di un intervallo.

Si sceglie l'intervallo che ha come estremi due barriere costruite a partire dallo scarto interquatilico(IQ) che indichiamo con B_{inf} e B_{sup} :

$$B_{inf} = Q_1 - 1,5IQ$$

$$B_{sup} = Q_3 + 1,5IQ$$

I valori esterni a queste barriere vengono considerate osservazioni anomale:

$$x > Q_3 + 1,5.IQ$$

$$x < Q_1 - 1,5Q$$

I baffi(=whiskers) sono due segmenti orizzontali,che si staccano dalla scatola(dal punto di mezzo dei lati del rettangolo)e raggiungono i valori minimo e massimo osservati oppure a distanza 1,5 IQR,rispettivamente,dal primo e terzo quartile.

4-Sintesi della distribuzione di un carattere-Indici di variabilità e di forma.

4.1-Variabilità-

Due distribuzioni di variabili statistiche quantitative univariate possono differire oltre che per la posizione, anche per la diversa variabilità;e a parità di valore medio di posizione, possono differire per la diversa variabilità.

Esempio-Ipotizziamo le due distribuzioni riportate in tabella

A **B**

x_i	n_i	x_i	n_i
6	**5**	**2**	**5**
8	**10**	**6**	**10**
10	**15**	**10**	**15**
12	**10**	**14**	**10**
14	**5**	**18**	**5**
Totale	**45**	**Totale**	**45**

Per le due distribuzioni media e mediana coincidono :

Distribuzione A:Media=10;Mediana=10;

Distribuzione B:media=10;Mediana=10.

Le due distribuzioni pur presentando valori medi uguali sono diverse.

Sul piano concettuale possiamo dire che la variabilità di un fenomeno è la sua attitudine ad

assumere differenti modalità.Operativamente,occorre pervenire ad una misura di tale attitudine.

E' importante distinguere due distinte famiglie di indici di variabilità:

-la prima famiglia attiene alla variabilità delle singole modalità x_i **rispetto ad un elemento della famiglia delle medie (ad esempio la media,la mediana etc.)mediante una sintesi degli scarti tra le singole modalità e il valore di riferimento,scarti chiamati scostamenti medi assoluti di ordine r, ossia** $|x_i - \alpha|^r$ **.**

-la seconda famiglia ricorre ad opportune medie potenziate costruite sulle differenze $|x_i - x_j|$ **intercorrenti tra le diverse modalità.**

Per la variabilità rispetto ad un elemento della famiglia delle medie,la più comune specificazione di un indice discende dalla media potenziata degli scarti assoluti di ordine r,ed assume la forma

$$\left\{\frac{1}{N}\sum|x_i - \alpha|^r\right\}^{1/r} \qquad \text{[4.1}$$

con r numero reale diverso da 0.Si controlla immediatamente che qualunque siano α e r la 1] assume valore 0 se, e solo se,il fenomeno non presenta variabilità;viceversa assume valori via via più alti quanto più gli scostamenti aumentano in valore assoluto.

La 1] assume forme diverse al variare di α e di r.In particolare:

se $\alpha = M_e$ **con r=1 si ha**

$\frac{\sum|x_i - M_e|}{N}$ **[4.2 lo scostamento semplice medio assoluto dalla mediana;**

se α=M con r=1 si ha

$\frac{\sum|x_i - M|}{N}$ **[4.3 lo scostamento semplice medio assoluto dalla media aritmetica;**

se α=M con r=2 si ha

$\left(\frac{\sum(x_i - M)^2}{N}\right)^{\frac{1}{2}}$ **[4.4 lo scarto quadratico medio,indicato convenzionalmente con σ.**

4.2-Varianza e scarto quadratico medio

L'indice più importante per misurare la variabilità di una distribuzione è espresso dalla media degli scarti al quadrato.Tale quantità si chiama varianza ed è indicata con il simbolo σ^2

$\sigma^2 = \frac{\sum(x_i - M)^2}{N}$ **[4.5 varianza**

Una difficoltà nella interpretazione della varianza deriva dal fatto che essa è espressa nella unità di misura del fenomeno al quadrato.Per questo si utilizza lo scarto quadratico medio poiché è espresso nella stessa unità di misura del carattere e misura di quanto in media quadrata i valori x_i **si discostano dalla loro media.**

Per calcolare la varianza σ^2 **mediante la formula che la definisce,(4.5)occorre calcolare la media M,poi tutti gli scarti** $(x_i - M)$**,quindi farne il quadrato e,infine,calcolare la media aritmetica di tali scarti.Si comprende che,anche se le modalità** x_i **sono numeri**

interi,la media M quasi sempre è un numero decimale per cui gli scarti al quadrato richiedono un numero doppio di decimali.

Per ridurre notevolmente i calcoli e le approssimazioni si può esprimere la varianza in modo diverso.Sviluppando il quadrato del numeratore della varianza,ossia la devianza, $\sum_{i=1}^{n}(x_i - M)^2 = \sum_{i=1}^{n} x_i^2 + nM^2 - 2M\sum_{i=1}^{n} x_i$ **e tenendo conto delle semplificazioni,si ha che la varianza:**

$$\sigma^2 = \frac{\sum x_i^2}{k} - \left(\frac{\sum x_i}{k}\right)^2 = M_q^2 - M^2$$ **[4.6 è pari alla media quadratica al quadrato meno il quadrato della media aritmetica.In pratica per ottenere** σ^2 **basta sommare i valori delle modalità ed i corrispondenti quadrati facendone poi le rispettive medie.**

Il ricorso al concetto di *momento* permette di esprimere la varianza diversamente.

Si chiama *momento di origine A e di grado r* la quantità:

$$\mu_{A,r} = \frac{\sum (x_i - A)^r n_i}{\sum n_i}$$ **[4.7**

Per ricavare alcune relazioni notevoli occorre fare alcune posizioni.

Per A pari zero otteniamo i *momenti di ordine r da zero*:

$$\mu_r = \frac{\sum x_i^r n_i}{\sum n_i}$$ **[4.8**

in particolare:

se r=0 otteniamo il momento di ordine 0 che vale 1 $\mu_{0,0} =$

$$\mu_0 = \frac{\sum_{i=1}^{n} x_i^0 n_i}{\sum_{i=1}^{n} n_i} \quad \text{[4.9}$$

se r=1 otteniamo il momento di ordine 1 che è la media aritmetica

$$\mu_{o,1} = \mu_1 = \frac{\sum_{i=1}^{n} x_i n_i}{\sum_{i=1}^{n} n_i} \quad \text{[4.10}$$

se r=2 otteniamo il momento di ordine 2 che è il quadrato della media quadratica

$$M_q^2 = \mu_{0,2} = \frac{\sum x_i^2 n_i}{\sum n_i} \quad \text{[4.11}$$

Se A coincide con la media aritmetica otteniamo l'r-esimo momento dalla media o momento centrale:

$$\mu_r = \frac{\sum (x_i - M)^r n_i}{\sum n_i} \quad \text{[4.12}$$

per r=2 otteniamo la varianza σ^2

$$\mu_2 = \sigma^2 = \frac{\sum (x_i - M)^2 n_i}{\sum n_i} \quad \text{[4.13}$$

Poniamo M-A=d,dove A è un valore diverso dalla media aritmetica,esplicitiamo rispetto ad M;dopo opportune semplificazioni troviamo che la varianza è pari:

$$\sigma^2 = \mu_2 = \frac{\sum (x_i - M)^2 n_i}{\sum n_i} = \frac{\sum (x_i - A)^2 n_i}{\sum n_i} - \left(\frac{\sum (x_i - A) n_i}{\sum n_i} \right)^2 \quad \text{[4.14}$$

$$\sigma_2 = \mu_2 = \mu_{A,2} - (\mu_{A,1})^2$$

se A=0 ritroviamo la relazione 6] per cui

$$\sigma^2 = \mu_2 = \mu_{0,2} - \left(\mu_{0,1}\right)^2 = M_q^2 - M^2 \qquad \textbf{[4.15}$$

Valutiamo lo scarto quadratico medio sulle due distribuzioni iniziali utilizzando la formula 14.

Tabella – Colonne di risultati parziali per il calcolo dello scarto quadratico medio

A **B**

x_i	n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$	x_i	$x_i n_i$	$x_i^2 n_i$
6	5	30	36	180	2	10	20
8	10	80	64	640	6	60	360
10	15	150	100	1500	10	150	1500
12	10	120	144	1440	14	140	1960
14	5	70	196	980	18	90	1620
Totale	45	450		4740		450	5460

Distribuzione A

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 n_i}{\sum_{i=1}^{n} n_i} - M^2} = \sqrt{\frac{4740}{45} - 10^2} = 2{,}3087$$

Distribuzione B

$$\sigma = \sqrt{\frac{5460}{45} - 10^2} = 4{,}6188$$

Confermiamo l'osservazione iniziale:a valori medi uguali corrispondono distribuzioni diverse.

4.3-Proprietà della varianza

Le principali proprietà della varianza sono:

$$\sigma^2(k + X) = \sigma^2(X) \qquad \textbf{(1)}$$

$$\sigma^2(kX) = k^2\sigma^2(X) \qquad \textbf{(2)}$$

$$\sigma^2(H + kX) = k^2\sigma^2(X) \qquad \textbf{(3)}$$

La prima afferma che:una costante k che si aggiunge a tutti i dati non si ripercuote sulla varianza*: la*

varianza è invariante rispetto ad un cambiamento di origine.

La seconda mette in evidenza che una costante che moltiplica tutti i dati si riproduce nella varianza con effetto quadratico:*effetto quadratico di una costante moltiplicativa.*

La terza,*varianza di una trasformazione lineare,* esplicita la varianza di un carattere Y,ottenuto attraverso la trasformazione Y=H+kX di un carattere X di media M e varianza $\sigma^2(X)$**.**

Varianza complessiva di k gruppi

Dati i 3 gruppi riportati in tabella valutare la varianza complessiva come somma della varianza entro i gruppi e della varianza tra i gruppi.

Primo gruppo	Secondo gruppo	Terzo gruppo
2	4	11
6	7	15
8	13	
14		
20		
Totali 50	24	26
Medie		
Parziali 10	8	13

Valutiamo la Media Generale.Per la proprietà associativa:

Media Generale= $\frac{10\times5+8\times3+13\times2}{5+3+2}=10$

La varianza Totale $\sigma^e=\frac{(2-10)^2+(6-10)^2+\cdots+(15-10)^2}{10}=28$

Valutiamo le varianze entro i gruppi : σ_i

$$\sigma_1^2=\frac{(2-10)^2+\cdots+(20-10)^2}{5}=\frac{200}{5}=40$$

$$\sigma_2^2=\frac{(4-8)^2+(7-8)^2+(13-8)^2}{3}=14$$

$$\sigma_3^2 = \frac{(11-13)^2 + (15-13)^2}{2} = 4$$

Valutiamo la varianza tra i gruppi:

$$\sigma_{TRA}^2 = \frac{\left[(10-10)^2 \times 5 + (8-10)^2 \times 3 + (13-10)^2 \times 2\right]}{10} = \frac{30}{10} = 3$$

La varianza complessiva dei 3 gruppi è la somma della varianza entro i gruppi e della varianza tra i gruppi ossia:

Varianza totale= $\frac{(40\times 5)+(14\times 3)+(4\times 2)}{10} + \frac{30}{10} = \frac{250}{10} + \frac{30}{10} = 28$

Var.totale=Var.entro i gruppi+var.tra i gruppi

In generale se abbiamo K gruppi con diversa numerosità,e sia x_{ij} **il valore j-esimo contenuto nel gruppo i-esimo per i=1,2,3,…,k e j=1,2,…,** n_i **,la varianza totale è uguale:**

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{k}(x_{ij}-M)^2}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-M_i)^2}{N} + \frac{\sum_{i=1}^{k}(M_i-M)^2 n_i}{N} \qquad \textbf{[4.16}$$

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-M)^2}{N} = \frac{\sum_{i=1}^{k}\sigma_i^2 n_i}{N} + \frac{\sum_{i=1}^{k}(M_i-M)^2 n_i}{N} \qquad \textbf{[4.17}$$

4.4-Differenze

Come si è già detto un differente modo per costruire indici di variabilità è quello di ricorrere ad opportune medie potenziate che si basano sulle differenze intercorrenti fra le diverse modalità con cui si è manifestato il fenomeno.

Con gli scostamenti da un valor medio si risponde al quesito"di quanto mediamente differiscono i singoli valori da un valor medio"

Con le differenze medie si risponde invece alla domanda"di quanto mediamente differiscono fra loro i singoli valori"

Nel primo caso una media degli scostamenti da detto valor medio può anche servire per giudicare se il valor medio sintetizza bene i dati.

Le differenze medie non sono altro che medie delle differenze in valore assoluto fra tutte le intensità.

Le più utilizzate sono le differenze medie,che sono definite come medie potenziate delle distanze $|x_i - x_j|$**.**

In particolare abbiamo:

1-*l'indice differenza media assoluta di ordine m senza ripetizione*,definito come media potenziata di tutte le differenze in valore assoluto tra le osservazioni

$$\Delta^m = \left[\frac{1}{N(N-1)} \sum_{i=1}^{k} \sum_{j \neq i}^{k} |x_i - x_j| n_i n_j \right]^{\frac{1}{m}} \qquad \textbf{[4.18}$$

2- *l'indice differenza media assoluta di ordine m con ripetizione*.Le differenze medie con ripetizione tengono conto delle differenze del tipo $|x_i - x_i|$**:**

$$\Delta^m = \left[\frac{1}{N^2} \sum_{i=1}^{k} \sum_{j=1}^{k} |x_i - x_j|^m n_i n_j \right]^{\frac{1}{m}} \qquad \textbf{[4.19}$$

Delle medie potenziate si usano soltanto quelle di ordine 1 e di ordine 2

Differenza semplice media di ordine 1 senza ripetizione

$$\Delta_1 = \frac{\sum_{i,j=1}^{k} |x_i - x_j|}{N(N-1)} \qquad \textbf{[4.20}$$

Differenza semplice media di ordine 1 con ripetizione

$$\Delta_1^R = \frac{\sum_{i,j=1}^{k} |x_i - x_j|}{N^2} \qquad \textbf{[4.21}$$

Dato che $|x_i - x_j| = |x_j - x_i|$

$\sum_{i,j=}^{k} |x_i - x_j| = 2\sum_{i<j} (x_j - x_i)$ **per una serie di valori**

$\sum_{i<j} |x_i - x_j| n_i n_j = 2\sum (x_j - x_i) n_i n_j$ **per una distribuzione di frequenza.**

La differenza semplice media senza ripetizione è pari:

$$\Delta_1 = \frac{2\sum_{i<j} (x_j - x_i) n_i n_j}{N(N-1)} \qquad \textbf{[4.22}$$

Fra la differenza semplice media senza ripetizione e la differenza semplice media con ripetizione sussiste la relazione:

$$\Delta_R = \frac{N-1}{N}\Delta_1 = 1 - \frac{1}{N}\Delta_1$$

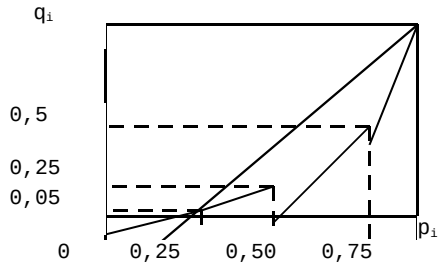
Inoltre:

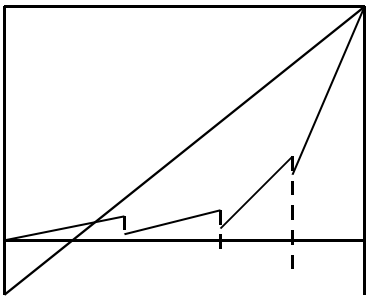
- $\Delta_R < \Delta_1$

-al tendere di $N \to \infty$ **si ha** $\lim_{N\to\infty} \Delta_R \to \Delta$**,ossia le due differenze medie assolute vengono a coincidere**

Esempio1-Voti conseguiti da uno studente in alcuni esami.Calcolare la differenza semplice media senza ripetizione Δ**.**

$x_i - x_j$	26	27	28	28	29	30
26	0					
27	1	0				
28	2	1	0			





28	2	1	0	0		
29	3	2	1	1	0	
30	4	3	2	2	1	0

La somma di tutte le differenze,in valore assoluto,è pari a 52

$$\Delta = \Delta_1 = \frac{\sum_{i,j} \left| x_i - x_j \right|}{N(N-1)} = \frac{52}{6 \times 5} = 1{,}73$$ **:**

le valutazioni differiscono in media di 1,73 trentesimi

Esempio 2-Sulla distribuzione di frequenza riportata in tabella,calcolare la differenza semplice media.

x_i	0	1	2	3	4	5	6	Totale
n_i	3	9	13	11	8	4	2	50

$$\Delta_1 = \frac{2\sum_{i>j}(x_i - x_j)n_i n_j}{N(N-1)}$$

predisponiamo la tabella per il calcolo delle differenze,riportando nella prima riga e nella prima colonna le modalità,e nella seconda riga e seconda colonna le frequenze.

$(x_i - x_j)n_i n_j$ per $i > j$		0	1	2	3	4	5	6
		3	9	13	11	8	4	2
0	3	0						
1	9	1x9x3	0					
2	13	2x13x3		0				
3	11				0			
4	8					0		
5	4						0	
6	2							0

Riportiamo in tabella il prodotto delle differenze con le frequenze

	27					
	78	117				
	99	198	143			
	96	216	208	88		
	60	144	156	88	32	
	36	90	104	66	32	8
$\sum_{i>j}(x_i - x_j)n_i n_j$	396	765	611	242	64	8

=2086

Applicando la 5)

$$\Delta_1 = \frac{2 \times 2.086}{50 \times 49} = 1{,}70286$$

4.5-Indici di asimmetria

Due distribuzioni possono differire per posizione,per variabilità.Tali differenze non esauriscono il complesso delle informazioni contenute nei dati.Ulteriori differenze nella distribuzione fanno riferimento al concetto di forma. Tra le diversità di forma sono importanti quelle riferibili a simmetria e asimmetria e alla curtosi(con riferimento ad una particolare distribuzione normale standardizzata).

In proposito una distribuzione si dice simmetrica rispetto ad un asse verticale di ascissa $x = x_0$ **quando,per tutti i valori** $\alpha > 0$ **si ha**

$$f(x_0 + \alpha) = f(x_0 - \alpha)$$

altrimenti si parla di distribuzione asimmetrica che può essere positiva o negativa.

Si può mostrare che in una distribuzione unimodale con asimmetria positiva,la distribuzione presenta più valori(si dice una coda)verso il semiasse positivo delle x e gli indici di posizione Media,Moda e Mediana soddisfano le diseguaglianze:

Moda≤Mediana≤Media

Viceversa in una distribuzione unimodale con asimmetria negativa,la distribuzione presenta una coda verso sinistra e gli indici Media,Moda,Mediana soddisfano le diseguaglianze:

Media≤Mediana≤Moda

Per avere informazioni sulle forme della distribuzione occorre ricorrere al calcolo di qualche indice di asimmetria.

Le misure dell'asimmetria sono equivoche.L'equivoco spesso nasce dalle stesse definizioni che vengono date di simmetria di una distribuzione,definizioni basate su particolari condizioni che si verificano per le distribuzioni simmetriche.

Poiché ,per esempio, per una distribuzione unimodale e simmetrica coincidono la media aritmetica,la mediana e la moda,alcuni considerando questa condizione sufficiente per la simmetria,mentre è solo necessaria,definiscono distribuzione simmetrica erroneamente una distribuzione in cui media aritmetica,mediana e moda coincidono;pertanto un indice adimensionale,in versione standardizzata,ottenuto considerando *Media-Mediana/sigma* si annulla se la distribuzione è simmetrica ma anche per distribuzioni che simmetriche non sono.

Altresì,una proprietà essenziale degli indici di asimmetria dovrebbe essere quella di annullarsi se e solo se la distribuzione è simmetrica;

La preferenza viene data a un indice che tiene conto della variabile standardizzata $Z = \frac{x - M}{\sigma}$ **ed è pari al rapporto del momento terzo sullo scarto quadratico medio**

al cubo,noto come indice di asimmetria di Fisher definito: $\gamma_1 = \frac{\mu_3}{\sigma^3}$

Esso è positivo,negativo o nullo per una distribuzione asimmetrica positiva,negativa o simmetrica,rispettivamente.

Valutiamo alcune misure di asimmetria sui dati riportati in tabella.

Tabella-Colonne di risultati parziali per il calcolo di alcuni indici di asimmetria

x_i	n_i	$x_i n_i$	f_i	F_i	$x_i - M$	$(x_i - M)^2 n_i$	$(x_i - M)^3 n_i$
2	6	12	0,12	0,12	-4	96	-384
4	11	44	0,22	0,34	-2	44	-88
6	14	84	0,28	0,62	0	0	0
8	15	120	0,30	0,92	2	60	120
10	4	40	0,08	1,00	4	64	256
Tot.	50	300	1,00			264	-96

Per i nostri dati:Media=6;Moda=8;Mediana=6

$$\frac{Media - Mediana}{\sigma} = \frac{6-6}{\sigma} = 0$$

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = -1{,}92$$

L’esempio mostra la necessità di non utilizzare acriticamente gli indici statistici di asimmetria e di completare l’informazione sulla forma facendo ricorso ad una rappresentazione grafica dei dati;dal confronto si 9evidenzierà la non corrispondenza tra il valore degli indici e la rappresentazione grafica dei dati.

4.6-Variabilità relativa

Si vuole conoscere quale fra i caratteri,riportati in tabella,presenta minore variabilità.
I caratteri sono: statura,diametro traverso del torace e peso.

Carattere	M	Scarto q.m.	c.v.=s.q.m./M
Statura	168,59cm	6,49cm	0,0385
Diametro…	28,58cm	2,35cm	0,0822
Peso	64,48k$_g$	7,27K$_g$	0,1127

Osservazioni

1-statura e peso non sono confrontabili sulla base dei rispettivi scarti quadratici medi,in quanto sono espressi in differenti unità di misura.
2-A ben riflettere, non è nemmeno opportuno confrontare il σ della statura con quello del diametro traverso,anche se questi sono espressi entrambi in cm.
Infatti,i due fenomeni,presentano valori medi molto diversi tra loro e una differenza di 1cm fra due altezze è meno rilevante della differenza di 1cm fra due diametri traversi.
Infatti,all'aumentare dell'ordine di grandezza di un fenomeno tende normalmente ad aumentare la sua variabilità

Il confronto di variabilità, sulla base degli indici di variabilità assoluta,non può avvenire:

1-Se le unità di misura di due variabili sono eterogenee;
2-se due variabili sono misurate sulla stessa scala,ma le loro medie differiscono sensibilmente,il confronto dei valori assunti da un indice assoluto di variabilità per l'una e l'altra variabile pur essendo lecito,non appare sempre utile.

Faremo riferimento a due categorie di indici relativi di variabilità:
1)indici di variabilità relativi ad un valore medio;
2)indici di variabilità relativi al massimo.

1-INDICI DI VARIABILITA' RELATIVI AD UN VALORE MEDIO

I più noti indici si basano sugli scostamenti in valore assoluto,relativi alla media aritmetica ossia

$$\frac{|x_i - M|}{M}$$

Si ottengono dei numeri puri in quanto l'unità di misura del numeratore è uguale a quella del denominatore.
Gli indici di variabilità relativi ad un valore medio si ottengono calcolando opportune medie degli scarti relativi.
Una espressione generale quale media potenziata di ordine m degli scarti relativi $\left|\frac{x_i - M}{M}\right|$ **è la seguente**

$$I_M = \left(\frac{\sum \left| \frac{x_i - M}{M} \right|^m n_i}{\sum n_i} \right)^{\frac{1}{m}} - \qquad \textbf{[4.24}$$

I due più usati indici percentuali di variabilità relativi ad un valore medio sono

$$100 \frac{\sum |x_i - M| n_i}{M \sum n_i} = 100 \frac{S_M}{M} \qquad \textbf{[4.25}$$

$$100 \sqrt{\frac{\sum \left(\frac{x_i - M}{M} \right)^2 n_i}{\sum n_i}} = \frac{100}{M} \sqrt{\frac{\sum (x_i - M)^2}{\sum n_i}} = 100 \frac{\sigma}{M} \qquad \textbf{[4.26}$$

Il più noto è $\frac{\sigma}{M}$ **,denominato *coefficiente di variazione.***

Il coefficiente di variazione è un numero puro che esprime σ **in termini di M**

Gli indici relativi ad un valore medio hanno il minimo uguale a zero e il massimo non definito.Possono assumere valore negativo se la media è minore di zero(profitti delle imprese,valori della temperatura)o addirittura infinito nel caso che sia la media uguale a zero.Quindi gli indici relativi ad un valore medio non vanno calcolati se $M \leq 0,$ **in quanto, se M<0 si otterrebbe un indice relativo negativo,mentre se M=0 il rapporto perde di significato.Consegue che è ragionevole impiegare gli**

indici di variabilità relativi ad M,solo per i caratteri che assumono valori positivi.Occorre quindi un indice relativo che assuma valore zero in assenza di variabilità e un valore pari ad un certo valore in presenza di massima variabilità; gli indici a cui faremo riferimento sono gli indici di variabilità relativi al massimo.

2-INDICI DI VARIABILITA'RELATIVI AL MASSIMO

Le due classi di indici danno informazioni diverse.Mentre le misure di variabilità relativi ad una media forniscono il valore dell'indice assoluto in termini del valore medio,le misure di variabilità relative al massimo di variabilità danno il valore dell'indice assoluto in termini del valore massimo che lo stesso può raggiungere.Anche qui si ottiene un numero puro che moltiplicato per 100 dà la percentuale del valore che l'indice assume nel caso reale rispetto al massimo teorico.

E' opportuno precisare che il concetto di massima variabilità non è legato ad uno schema.Per massima variabilità si intende quella configurazione di valori $(x_1, x_2, \ldots, x_n)$ **che,fra tutte le configurazioni possibili rende massimo il valore dell'indice considerato.**

La configurazione che andremo a presentare è la seguente:

Supponiamo di avere n dati disposti in ordine crescente

$$x_1 \leq x_2 \leq, \ldots, x_n$$

aventi come media M e il cui totale T sia

$$\mathbf{T=}\sum x_i$$

Se ammettiamo la trasferibilità del carattere-cioè la possibilità che un certo x_i perda una quantità h che viene acquisita dal termine successivo x_j-il massimo di variabilità(a parità di T) si ha quando i primi n-1 termini sono tutti nulli e l'ultimo è uguale a T.

x_i	n_i
0	**N-1**
$\sum x_i$	**1**
Totale	**N**

Rispetto a tale distribuzione valutiamo Max σ,MaxΔ

$$\text{Max}_{\sigma} = \sqrt{\frac{(0-M)^2(N-1)+(NM-M)^2}{N}} =$$

$$\sqrt{\frac{M^2(N-1)+N^2M^2+M^2-2NM^2}{N}} = \quad M\sqrt{N-1} \qquad [4.27$$

$$\text{Max}\,\Delta = \frac{2NM(N-1)}{N(N-1)} = 2M \qquad [4.28$$

$$\text{Max}\,\Delta_R = 2M\left(\frac{N-1}{N}\right) \qquad [4.29$$

Pertanto lo scostamento quadratico medio relativo è dato dal rapporto

$\frac{\sigma}{Max\sigma} = \frac{\sigma}{M\sqrt{N-1}}$ **ovvero è il coefficiente di variazione moltiplicato per** $1/\sqrt{N-1}$**.**

Mentre la differenza semplice relativa è pari al rapporto

$$\frac{\Delta}{2M} \qquad \textbf{[4.30}$$

Lo scostamento quadratico medio relativo $\frac{\sigma}{Max\sigma}$ **al crescere di N è decrescente ed è uguale al coefficiente di variazione per N=2**

Differenza semplice media relativa con ripetizione risulta

$$\frac{\Delta_R}{2M\frac{N-1}{N}} = \frac{\Delta_R}{2M}\frac{N}{N-1} \qquad \textbf{[4.31}$$

Poiché

$$\Delta_R = \frac{N-1}{N}\Delta$$

$$\frac{\Delta_R}{2M}\frac{N}{N-1} = \frac{\frac{N-1}{N}\Delta}{2M}\frac{N}{N-1} = \frac{\Delta}{2M} \qquad \textbf{[4.32}$$

ossia la differenza semplice media con ripetizione relativa è uguale alla differenza semplice media relativa.

4.7-Concentrazione

Consideriamo N unità,le quali posseggono rispettivamente le quantità(ordinate in valore crescente)

$$x_1 \le x_2 \le x_3 \le ..... \le x_n$$

di un certo carattere additivo trasferibile.
Sia T la quantità posseduta dalle N unità,ossia

$$\mathbf{T=} x_1 + x_2 + ... + x_n$$

Si chiama concentrazione di N quantità $x_1, x_2, ..., x_n$ **il modo con cui il totale** $T = \sum x_i$ **si distribuisce fra le unità stesse.**

Esistono due situazioni limite:

-CONCENTRAZIONE MINIMA: quando ogni unità avrà una quantità uguale alla media aritmetica

concentrazione nulla=equidistribuzione
concentrazione nulla=variabilità nulla

-CONCENTRAZIONE MASSIMA:

$$x_1 = x_2 = x_3 = ... = x_{n-1} = 0$$

quando:

$$x_n = T = \sum x_i$$

Sia

. $p_i = \frac{i}{N}$**,la frequenza relativa cumulata delle prime** i **unità**

. $q_i = \frac{\sum_{j=1}^{i} x_j}{T}$ **le quantità cumulate relative**

le q_i **rappresentano le quote relative all'intensità totale T detenute dalle prime i unità statistiche,ordinate in senso non decrescente,rapportate a T.**

E' sempre $p_i \geq q_i$ **[4.33**

Ogni coppia (p_i, q_i)**specifica la percentuale fornita dalle prime i unità ordinate della popolazione e la corrispondente percentuale di intensità totale T che tali unità posseggono.**

Curva di concentrazione:

-Portando in ascissa

$$p_i = \frac{i}{N}$$

-Portando in ordinata

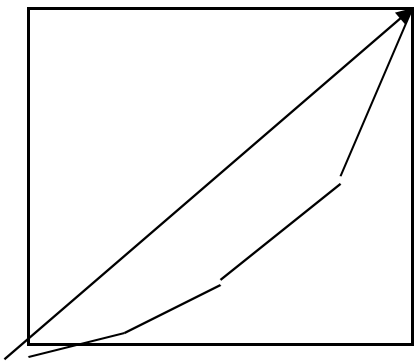
$$q_i = \frac{x_1 + x_2 + ... + x_i}{x_1 + x_2 + ... + x_n} = \frac{\sum_{j=1}^{i} x_j}{\sum_{i=1}^{n} x_i}$$

le n-1 coppie di frazioni (p_i, q_i) **per i=1,2,3,…,n-1 corrispondono in un piano cartesiano ad n-1 punti per i quali si può far passare una curva,detta di Lorenz,o di concentrazione,avente per estremi l'origine del sistema degli assi e il punto di coordinate** $p_n = q_n = 1$ **.**

Nelle situazioni concrete si considera un numero finito di unità statistiche,ad esempio i redditi sono riportati generalmente in una distribuzione per classi di intervallo.Conseguentemente,la curva teorica,detta curva di Lorenz,viene sostituita da una spezzata di concentrazione.

Esempio-Sui dati x_i**=1,4,5,10 trovare** p_i, q_i

Unità $\sum i$	x_i $\sum x_i$	$p_i = i/N$	$q_i = \frac{\sum_{j=1}^{i} x_j}{T}$
1 1	**1 1**	$1/4$ **= 0,25**	$1/20$ **= 0,05**
1 2	**4 5**	$2/4$ **= 0,5**	$5/20$ **= 0,25**
1 3	**5 10**	$3/4$ **= 0,75**	$10/20$ **=0,5**
1 4	**10 20**	$4/4$ **= 1**	$20/20$ **=1**



Tale curva è compresa in un triangolo rettangolo isoscele,avente base e altezza uguali a 1.

Una misura del grado di concentrazione è data dal rapporto di concentrazione R,introdotto dal Gini nel 1914 ed è pari a:

$$R = \frac{\sum_{i=1}^{n-1}(p_i - q_i)}{\sum_{i=1}^{n-1} p_i} \qquad \textbf{[4.34}$$

Sviluppando la sommatoria e ricordando che:

$$\sum_{i=1}^{n-1} p_i = \sum_{i=1}^{n-1} \frac{i}{N} = \frac{n-1}{2}$$

si ottengono le formule equivalenti:

$$R = 1 - \frac{\sum_{i=1}^{n-1} q_i}{\sum_{i=1}^{n-1} p_i} = 1 - \frac{2}{n-1}\sum_{i=1}^{n-1} q_i \qquad \textbf{[4.35}$$

E’ possibile dimostrare che il rapporto di concentrazione R **del Gini coincide con la differenza semplice media relativa** $R = \Delta / 2M$ **.**

Tale rapporto misura lo scostamento relativo della spezzata di concentrazione dalla retta di equidistribuzione:è il quoziente tra l’area di concentrazione e l’area di massima concentrazione.

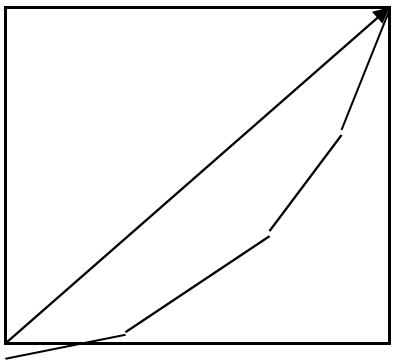
Il rapporto di concentrazione assume valori compresi nell’intervallo $[0,1]$ **e precisamente:**

R=0 se la concentrazione è nulla(equidistribuzione);

R=1 se la concentrazione è massima.

Nel caso di un carattere discreto si utilizza la formula precedente;nel caso in cui le modalità del carattere siano divise in classi,il calcolo del rapporto di concentrazione del Gini avviene mediante l'approssimazione della formula dei trapezi alla curva di Lorenz,e quindi con la seguente formula:

$$R = 1 - \sum_{i=1}^{n} (q_i + q_{i-1})(p_i - p_{i-1}) \qquad \textbf{[4.36}$$



Il poligono formato dalla retta di equidistribuzione e dalla spezzata di concentrazione presenta in generale una forma che non è riconducibile a dei poligoni notevoli della geometria elementare.Il problema si risolve calcolando l'area della parte del piano compresa tra la spezzata di concentrazione e l'asse orizzontale e sottraendo quest'ultima all'area del triangolo di coordinate $(0,0),(1,0),(1,1).$ **la cui area è uguale a** $1/2$**.Il poligono formato dalla spezzata e dall'asse orizzontale è**

scomponibile in una successione di trapezi rettangoli,il primo dei quali è degenere(Triangolo),le cui aree si calcolano tramite le note formule.

$$Areadiconcentrazione = \frac{1}{2} - \sum_{i=1}^{n} \frac{(q_i + q_{i-1})(p_i - p_{i-1})}{2} \qquad [4.37$$

$$R = \frac{Areadiconcentrazione}{1/2} \qquad [4.38$$

Esempio-Calcoli per determinare il rapporto di concentrazione della distribuzione dei redditi riportata in tabella.

Classe di reddito	Reddito complessivo della classe x_i	*Numero dei redditieri* n_i	Frequenze cumulate	Reddito cumulato
0 - 3	12.792	7.976		
3 - 6	40.650	8.763		
6 - 9	29.320	4.130		
9 - 15	12.932	1.176		
15 - 25	5.580	297		
25 - 50	3.405	105		
50 - 100	1.172	18		
>100	532	3		
	106.383	22.468		

R=1-0,6440452=0,354.8-

4.8-MUTABILITA’

Le varie definizioni di mutabilità(Zani,Piccolo..) sono riconducibili alla definizione data da Gini(1912).Gini ha definito la mutabilità come “l’attitudine di un carattere ad assumere differenti modalità”-Pertanto la mutabilità è la possibilità di variare per un carattere qualitativo tra una perfetta omogeneità(quando il carattere si manifesta mediante un

solo attributo) e una qualche eterogeneità(se nella popolazione vi sono almeno due attributi differenti).
Più specificatamente,una popolazione ripartita,rispetto ad un carattere,in K modalità,è omogenea, rispetto al carattere, od anche che le sue unità sono tutte omogenee fra loro, se tutte le unità presentano la stessa modalità del carattere:siamo in presenza della MASSIMA OMOGENEITA' (o MINIMA ETEROGENEITA'); se ciò non accade il collettivo è detto eterogeneo,e se tutte le unità del collettivo sono ripartite fra le K classi di modalità in parti uguali,parliamo di MASSIMA ETEROGENEITA'(o minima omogeneità).

Per fissare le idee si ammetta che un carattere X si manifesti con quattro modalità A,B,C,D,con le frequenze date nel seguente prospetto:

Modalità	A	B	C	D
frequenze	35	10	15	40

Vi sono numerosi modi di suddividere la frequenza totale di 100 fra le 4 modalità; due di essi configurano opposte situazioni limite di riferimento:

a) Massima omogeneità,tutti i 100 individui presentano la stessa modalità, ad esempio la modalità B;
b) Massima eterogeneità, quando il fenomeno non manifesta alcuna predilezione per l'una o l'altra modalità,che hanno quindi frequenze tutte uguali a 100/4=25.

Generalizzando a K modalità e utilizzando la proporzione della generica modalità sul totale, ovvero la frequenza relativa f_i, si ha che:

- **massima omogeneità(o minima eterogeneità) quando f_i=1 per qualche i, e f_j=0 per i≠j;**
- **massima eterogeneità(o minima omogeneità) quando f_i= $1/k$ per i=1,2,…,k.**

La <u>eterogeneità</u> ,come concetto opposto alla <u>omogeneità</u>, misura la variabilità delle frequenze relative f_i,i=1,2, …,k, senza coinvolgere le modalità del carattere;i concetti di eterogeneità o omogeneità non richiedono alcun ordinamento delle modalità,e quindi sono applicabili a qualunque tipo di fenomeno.

La problematica della <u>eterogeneità</u> coinvolge differenti settori disciplinari, dall'Ecologia alla Sociologia,dalla Linguistica alla Demografia, dalla Biologia alla Statistica economica.Tra le problematiche più rilevanti ne segnaliamo alcune.

. Sul fronte degli studi ecologici, la problematica della eterogeneità è connessa alla diversità delle specie animali e vegetali presenti in un determinato territorio:gli squilibri dell'ecosistema rendono urgente la determinazione della differenziazione genetica tra le specie e dei rischi di una loro riduzione progressiva:occorre misurare la diversità ecologica.

. Nelle analisi delle preferenze elettorali tra k partiti,i risultati possono oscillare tra la massima eterogeneità(tutti i k partiti ricevono le stesse preferenze) ed una totale polarizzazione(pochissimi partiti ricevono la totalità delle preferenze).I problemi della polarizzazione del consenso elettorale nel tempo e sul territorio,e dei sistemi elettorali che lo accentuano, possono essere misurati tramite indici di polarizzazione.

. La diversità razziale presente in un continente, la mutevolezza delle specie vegetali, la ricchezza del vocabolario linguistico di un Autore sembrano problematiche differenti,ma la cui misura presenta aspetti tutti assimilabili agli indicatori preposti nell'ambito della misura della eterogeneità.

Un indice generico I,quindi deve soddisfare alcune proprietà:

1) **$I(f) \geq 0$**
2) **I(f) è minimo quando (k-1)/f_i sono uguali a zero e una sola è uguale a 1;**
3) **I(f) è massimo se $f_i = 1/k$ per ogni i quando il fenomeno non manifesta alcuna predilezione per l'una o l'altra modalità e quindi le unità si ripartiscono in parti uguali fra le modalità.**
4) **L'indice deve essere sensibile ai trasferimenti di frequenza.**

INDICI PER SERIE SCONNESSE

Tra gli indici di mutabilità più diffusi,consideriamo:

a)l'indice di Gini;

b)la cosiddetta Entropia;

c)l'indice di eterogeneità di Frosini.

a)INDICE DI GINI

L'indice di mutabilità di Gini(o di eterogeneità) può farsi derivare dalla seguente argomentazione: nella seriazione(a_i,n_i), i=1,2,..,k della mutabile A che assume gli attributi a_i con frequenze assolute n_i e frequenze relative $f_i = n_i / n$ per i=1, 2,..,k costruiamo un indice di variabilità mediante la differenza semplice media con ripetizione tra tutte le n^2 coppie di modalità a_i ed a_j

,assumendo come misura della diversità d(a_i,a_j) la più elementare funzione possibile cioè

$$d(a_i, a_j) = \begin{pmatrix} 0, sei = j \\ 1, sei \neq j \end{pmatrix}$$

$$\mathbf{I}_1 = \sum_{i=1}^{k} \sum_{i=1}^{k} d(a_i, a_j)/n^2 =$$

$$\mathbf{I}_1 = 1 - \sum_{i=1}^{k} (f_i)^2 \qquad \textbf{[4.39}$$

L'indice di Gini vale 0 nel caso di minima eterogeneità(l'intera distribuzione si accentra in un solo attributo) mentre il suo massimo si verifica quando ciascuno dei k possibili attributi raccoglie esattamente n/k **delle frequenze totali delle popolazioni per cui** $f_i = 1/k, \forall i = 1,2,...,k.$ **Pertanto:**

$$G_{\max} = 1 - \sum_{i=1}^{k} (\frac{1}{k})^2 = 1 - (\frac{1}{k})^2 = 1 - \frac{1}{k} \qquad \textbf{[4.40}$$

Quindi,l'indice di mutabilità normalizzato di Gini,compreso in $[0,1]$**,è:**

$$\mathbf{I}_1{}^* = \left(1 - \sum_{i=1}^{k} (f_i)^2\right) \frac{k}{k-1} \qquad \textbf{[4.41}$$

b)<u>ENTROPIA</u>

La cosiddetta Entropia è uguale:

$$\mathbf{I}_2 = -\sum_{i=1}^{k} f_i \log f_i \qquad \textbf{[4.42}$$

Anche per $\mathbf{I}_2$ **il minimo è zero** $(= -\log 1)$, **mentre**

$$I_{2\max} = \log k$$

La base dei logaritmi può essere qualsiasi; comunque, le basi più usate sono 2 ed e(base dei logaritmi naturali). Anche $\mathbf{I}_2$ **può essere normalizzato semplicemente dividendolo per il suo massimo.**

L'indice normalizzato $\mathbf{I}_2^*$, **compreso in** $[0,1]$, **è definito da:**

$$\mathbf{I}_2^* = \frac{-\sum_{i=1}^{k} f_i \log(f_i)}{\log(k)} . \qquad \textbf{[4.43}$$

(L'entropia è un concetto della meccanica statica e della termodinamica perché si collega al disordine di un sistema e al calore dissipato in ogni trasformazione che avviene nell'Universo.Tale indice è noto come indice di diversità di Shannon,connesso alla teoria dell'informazione)

INDICE DI FROSINI

Un approccio più generale è stato proposto dal Frosini che introduce indici di omogeneità(ed i loro complementari indici di eterogeneità) basati sul concetto di distanza(1)(***La nozione di distanza fu introdotta nella matematica nel 1905-1906 dal grande matematico francese M.Frèchet,che fu anche studioso insigne di statistica matematica.A lui si deve il concetto di spazio metrico ossia di spazio basato sul concetto di distanza.Con essa si può intendere una misura della diversità fra unità statistiche riguardo a caratteri di***

natura sociale, demografica, economica,…Per i particolari si rimanda a G. Leti:Distanze e indici statistici,La Goliardica,Roma 1979) **tra il vettore delle frequenze relative osservate** $(f_1, f_2,....., f_k)'$ **ed il corrispondente vettore delle frequenze relative nell'ipotesi di massima eterogeneità** $(\frac{1}{k},...,\frac{1}{k})'$**.**

Tra le misure tipiche della distanza tra entità si utilizza la distanza Euclidea,cioè la radice quadrata della somma delle differenze al quadrato tra le coordinate corrispondenti. L'indice di omogeneità deve variare fra il minimo e il massimo proporzionalmente alla distanza euclidea fra la generica distribuzione f= $(f_1, f_2,....., f_k)$ **e la situazione di massima eterogeneità** $\left(\frac{1}{k},...,\frac{1}{k}\right)$**; tale distanza euclidea è**

$$d=\sqrt{\sum_{i=1}^{k}\left(f_i-\frac{1}{k}\right)^2}=\sqrt{\sum_{i=1}^{k}f_i^2-\frac{1}{k}} \qquad [4.44$$

e varia tra 0(quando l'omogeneità è minima) e $\left(1-\frac{1}{k}\right)^{1/2}=$ $\left(\frac{k-1}{k}\right)^{1/2}$ **(quando l'omogeneità è massima ossia quando un solo** f_i**=1,e gli altri sono =0).**

Un indice di omogeneità normalizzato fra 0 e 1 è allora $\frac{d}{\max d}$ **e prendendo il suo complementare, si perviene all'indice normalizzato di eterogeneità di Frosini**

$$I_3^* = 1 - \sqrt{\frac{k}{k-1}\sum_{i=1}^{k}\left(f_i^2 - \frac{i}{k}\right)} \qquad [4.45$$

La classe di indici proposti da Frosini è molto ampia e modificando la metrica si possono dedurre numerose altre misure(Per i particolari si rimanda a Frosini B.V. (1981),*Heterogeneity indices and distances between distributions,*in Metron,39,n.3-4

Applichiamo i tre indici alla distribuzione che ci è servita per introdurre l'argomento,e cioè f=$(0.35;0,10;0,15;0,40)$. **Si ha rispettivamente**

$$I_1 = 1 - \left(0,35^2 + 0,10^2 + 0,15^2 + 0,40^2\right) = 0,685$$

da cui l'indice normalizzato

$$I_1^* = \frac{4}{3} I_1 = 0,913$$

$$I_2 = -\left(0,35\log 0,35 + 0,1\log 0,1 + 0,15\log o,15 + 0,4\log 0,4\right)$$

Prendendo i logaritmi naturali si ottiene

$$I_2 = 1,2488$$

e l'indice normalizzato

$$I_2^* = \frac{I_2}{\log 4} = 0,9008$$

$$I^*_3 = 1 - \left\{\frac{4}{3}\left(0,35^2 + 0,1^2 + 0,15^2 + 0,4^2 - \frac{1}{4}\right)\right\}^{1/2} = 0,7056$$

Da un esame qualitativo la distribuzione f non appare tanto prossima alla situazione di massima eterogeneità $(0,25,0,25;0,25;0,25)$ **come sembrerebbe dai valori dei primi due**

indici.Si ritrova in I_3^* una maggiore validità, l'unico ad essere definito in termini di distanza.

---

1)*Si dice distanza, o metrica, una misura tra entità caratterizzata dalle seguenti proprietà:*

- la distanza d_{ii} tra un'entità e se stessa è nulla:

$$d_{ii} = 0 \qquad (i = 1,...,n);$$

- ***la distanza tra due entità qualsiasi i e j è non negativa:***

$$d \geq 0 \qquad (i, j = 1,.....,n);$$

- ***la distanza tra due entità i e j è la stessa se si misura da i a j oppure da j a i (simmetria)***

$$d_{ij} = d_{ji} \qquad (i \neq j = 1,....,n);$$

- ***la distanza tra due entità è non superiore alla somma delle distanze tra queste entità e una terza entità (disuguaglianza triangolare)***

$$d_{ij} \leq d_{ik} + d_{jk} \qquad (i \neq j \neq k = 1,....,n);$$

dove d_{ik} e d_{jk} sono le distanze tra la terza entità k e le entità i e j

INDICE DI DIVERSITA' PER DATI ORDINALI

Siano $a_1 < a_2 < \ldots < a_k$ le k modalità di un carattere A con scala ordinale.Per i caratteri ordinali occorre introdurre misure che siano funzione dei gradi di diversità esistenti fra le modalità,oltre che del modo di distribuirsi delle frequenze fra le modalità.

Una misura di diversità è quella proposta dal Gini sin dal 1912 che è la seguente:

$$D = 2\sum_{j=1}^{k-1} F_j(1 - F_j)$$ **dove** $F_j = \frac{N_j}{N}$ **è la frequenza relativa cumulata**

A tale espressione è pervenuto il Leti facendo osservazioni sul comportamento delle frequenze cumulate e retrocumulate in presenza di massima omogeneità e minima omogeneità.

Si considerino le 5 distribuzioni che seguono, relative alla popolazione in età da 6 anni in poi classificate secondo il grado di istruzione(esempio tratto da Zenga"Introduzione alla statistica descrittiva" pag.200 Vita e pensiero,1990).

Grado istruzione	*Distribuzione reale*	*Distribuzione massima omogeneità*	*Distribuzione minima omogeneità*	*Distribuzione massima diversità*	*Distribuzione prossima alla massima omogeneità*
Analfabeti	1.144	0	68.942	206.826	1
Alfabeti	34.987	0	68.942	0	0
Lic.elem.	200.058	413.651	68.942	0	0
Lic.media	125.434	0	68 942	0	0
Diploma	42.078	0	68.942	0	0
Laurea	9.950	0	68.942	205.825	413.650
TOTALI	413.651	413.651	413.651	413.651	413.651

La situazione di massima omogeneità può manifestarsi in 6 modi diversi a seconda della modalità su cui si concentra la totalità delle frequenze.La situazione di massima omogeneità coincide con quella di minima eterogeneità e si può anche ritenere di minima diversità. La situazione di massima eterogeneità delle frequenze si raggiunge invece quando le frequenze si trasferiscono ai gradi di istruzione estremi e ,in particolare, quando si equiripartiscono fra le modalità estreme.Infatti, se le frequenze,pur concentrandosi solo sulle modalità estreme, non vi si ripartissero in parti uguali, si potrebbero avere situazioni assai prossime alla massima omogeneità,

come si evince dall'ultima colonna del prospetto sopra riportato.

Il valore minimo di D(uguale a 0) si realizza se $F_1 = F_2 = ...F_{k-1} = 0$ **cioè se** $n_1 = n_2 = ...n_{k-1} = 0$ **e** $n_k = N$**. In altre parole D=0 se si è in presenza della massima omogeneità, ovvero della minima diversità., mentre nel caso di massima diversità**

D=$\left\{\frac{K-1}{2},\right\}$**per N pari,** **D=**$\left\{\frac{(K-1)}{2}(1-1/N^2)\right\}$**per N dispari**

Per N sufficientemente grande il valore massimo di D può assumersi pari a $\frac{K-1}{2}$ **qualunque sia N.**

Si definisce il seguente indice di dispersione relativo

$$d=\frac{D}{\frac{K-1}{2}} \qquad \text{con } d \in (0,1) \qquad [4.47$$

Recentemente (Statistica 2002 n.1)se ne dimostra la scomposizione ,in analogia con la scomposizione della varianza.