conoscenza è la vera protagonista nell'economia contemporanea e
me il ruolo di fattore primario nei più svariati contesti. Sono gli in-
dui a detenere il potere della conoscenza, che è necessaria per ge-
e anticipare i continui cambiamenti e le condizioni di incertezza
caratterizzano la nostra epoca. Questo volume rappresenta una gui-
uno strumento concreto per approcciarsi alla conoscenza della realtà
rigorosità scientifica.
approccio statistico alla conoscenza dei fenomeni reali consente di
ogliere informazioni, elaborarle e analizzarle, guidando appunto l'in-
duo alla presa di decisioni. Un sapere puramente teorico può facil-
te divenire obsoleto, ma l'applicabilità promossa dal testo consente
cquisire strumenti atti a gestire il cambiamento e a prevedere l'im-
edibile, vivendo così un processo di crescita continua. Il percorso
nosso da questo libro non sarà, quindi, quello tradizionale di un
siasi manuale di statistica, ma guiderà in modo graduale all'utilizzo
i strumenti specifici di questa disciplina, attraverso una presenta-
e e articolazione dei metodi che ne sono alla base, prediligendo
giormente un approccio di tipo deduttivo ai problemi reali.
scopo del volume è quello di mettere il lettore in condizione di po-
acilmente trasferire le conoscenze statistiche acquisite in campo so-
economico, nel settore di istruzione e formazione e più in generale
diversi indirizzi della ricerca scientifica.

nio Di Battista è professore ordinario di statistica. Dal 2004 è Presi-
te dei Corsi di Laurea di Scienze dell'Educazione e della Formazione
Scienze Pedagogiche della Facoltà di Scienze della Formazione nel-
iversità degli Studi "G. d'Annunzio" di Chieti-Pescara, dove detiene i
i di metodi e tecniche della valutazione e analisi e valutazione dei
essi formativi. È Direttore del Centro di Ricerca Universitario per la
tazione e lo Sviluppo (CERVAS). È membro del Consiglio Direttivo
a Società Italiana di Statistica dal 2010. I principali temi di ricerca
ardano lo studio della biodiversità, del campionamento statistico e
metodi e delle tecniche di valutazione dei servizi pubblici.

FrancoAngeli
La passione per le conoscenze

ISBN 978-88-204-0397-3

9 788820 403973

€ 39,00 (U)

367.71 T. DI BATTISTA METODI E TECNICHE PER LA VALUTAZIONE

Economia

Tonio Di Battista

Metodi e tecniche per la valutazione

Un approccio statistico

FrancoAngeli

Tonio Di Battista

Metodi e tecniche per la valutazione

Un approccio statistico

Coordinamento editoriale di Stefano Oronzo

FrancoAngeli

Stefano Oronzo, autore di diverse pubblicazioni, è direttore della rivista on line *Sportlex* (www.sportlex.net), dedicata alla ricerca scientifica nello sport.

Ristampa	Anno
0 1 2 3 4 5 6 7 8 9	2012 2013 2014 2015 2016 2017 2018 2019 2020 2021

Stampa: Global Print s.r.l., Via degli Abeti n. 17/1, 20064 Gorgonzola (MI).

A chi ha sempre creduto in me...
a Mia Madre.

"I often say that when you can measure what you are speaking about,
and express it in numbers, you know something about it;
but when you cannot measure it, when you cannot express it in
numbers, your knowledge is of a meager and unsatisfactory kind:
it may be the beginning of knowledge, but you have scarcely,
in your thoughts, advanced to the stage of science,
whatever the matter may be."
(Sir William Thomson Lord Kelvin, 1824-1907)

Indice

Prefazione

"Scopo basilare della valutazione è stimolare la crescita e il miglioramento.
Tutte le altre finalità, pur rispettabili, sono solo sfaccettature
dello sforzo generale che consiste nel valutare le condizioni presenti
come base per migliorare.
Una valutazione che non porti a un perfezionamento delle pratiche è sterile."
(Kempfer H. H., 1955)

Gli ultimi decenni sono stati caratterizzati da profonde, significative e repentine trasformazioni in campo tecnologico e scientifico, che hanno positivamente influenzato la crescita economica e sociale a livello mondiale. Viviamo oggi il periodo storico dell'economia fondata sulla conoscenza, siamo cittadini di quella che E. Morin definisce "*era planetaria*", caratterizzata da continui sviluppi scientifici, tecnologici ed economici che producono un *divenire planetario* comune per tutti gli esseri umani. Il business, come afferma Daft (2004), sta diventando un'unica arena globale tanto che nel ventunesimo secolo le organizzazioni dovranno "*sentirsi a casa*" in ogni parte del mondo. Con lo sviluppo delle tecnologie della comunicazione e dell'informazione è stata superata, infatti, la barriera del *qui ed ora* e siamo entrati nell'era del *sempre e ovunque*. Il quadro ambientale, dunque, sta diventando complesso e competitivo, di conseguenza le organizzazioni devono mettersi in condizione di poter attraversare confini culturali e geografici, cogliendo i vantaggi dell'interdipendenza globale e minimizzandone gli svantaggi. Nell'ambiente attuale, "*iperturbolento ed incerto*" (Emery e Trist, 1960), nel quale i processi di cambiamento risultano rapidi e dinamici, le organizzazioni devono maturare conoscenze che consentano di anticipare il cambiamento stesso e di guidare tali cambiamenti nella direzione più vantaggiosa per l'organizzazione;

devono, quindi, maturare le conoscenze necessarie a saper prendere decisioni in condizione di incertezza.

La *conoscenza* diventa la vera protagonista nell'economia contemporanea e, di conseguenza, assume il ruolo di fattore primario di produzione; i dirigenti, in questo nuovo panorama, sono chiamati ad incrementare la conoscenza stessa all'interno delle organizzazioni, favorendo in tal modo un processo di valorizzazione di tutti i loro collaboratori. Sono, infatti, gli individui e non i macchinari ad avere il potere della conoscenza necessaria a mantenere le organizzazioni competitive in quanto, se le macchine possono facilmente divenire obsolete, la conoscenza di contro, quando è in grado di gestire il cambiamento, vive un processo di crescita continua. Deve maturare negli individui una forma mentis che consenta di prevedere l'imprevedibile e di essere preparati al cambiamento.

In questo contesto gioca un ruolo fondamentale la ***Statistica*** in quanto scienza che, basandosi su un metodo rigoroso, raccoglie informazioni, le elabora, le analizza e guida alla presa di decisioni. In tutti i contesti organizzativi, in campo sociale, economico, politico, educativo, sanitario, ecc. ci si trova giornalmente a dover gestire l'incertezza e a dover prendere decisioni che devono essere fondate su previsioni che siano garanti di un metodo rigoroso ed attendibile. La statistica può essere quindi intesa come la scienza necessaria per guidare l'assunzione di decisioni. La stessa, quindi, consente di effettuare le scelte che si presentano come più vantaggiose. La statistica non è una scienza esatta, se con questo termine intendiamo una scienza in grado di dare risposte universali ed assolute e, consentiteci di aggiungere "fortunatamente", poiché se così fosse non sarebbe perfettibile e non avrebbe alcun senso continuare a far ricerca in questo campo. La statistica consente, infatti, di fare previsioni sull'evolversi di un fenomeno osservato e fonda le sue previsioni su indagini basate su un metodo rigoroso scientifico, valutando il livello di confidenza e di affidabilità in relazione all'incidenza sulla previsione stessa di un certo grado di errore. La stima dei livelli di confidenza e il calcolo dell'incertezza di misura sono gli elementi essenziali che fanno della statistica una scienza. La statistica, infatti, consente di fare previsioni ed è sempre in grado di associare ad esse una stima di errore; se così non fosse la previsione diventerebbe mera predizione! È proprio questo il senso

della statistica: prevedere, stimando un certo grado di errore, e non predire (... lavoro da presunti maghi e fattucchiere!)

Le facoltà che afferiscono al settore delle ***Scienze Umanistiche*** hanno un ruolo essenziale nell'attuale assetto dell'economia della conoscenza poiché si pongono l'obiettivo di formare i propri studenti e al tempo stesso di dare loro gli strumenti necessari per promuovere la conoscenza ad altri individui. La caratteristica saliente delle facoltà umanistiche innovative è senza dubbio quella di aver superato la tradizionalmente affermata dicotomia tra conoscenze umanistiche e conoscenze scientifiche, promuovendo la conoscenza in tutta la sua complessità e valorizzandone tutte le sfaccettature.

Lo scopo di questo manuale è appunto quello di contribuire al superamento della scissione umanistico-scientifica, dando agli studenti le conoscenze necessarie per osservare ed analizzare fenomeni reali in campo educativo, formativo e sociale, attraverso il metodo statistico e, di conseguenza, di guidare alla previsione di scelta in condizioni di incertezza.

Il percorso che ci accingiamo a compiere non sarà quello tradizionale di un vero e proprio corso di statistica, ma impareremo ad utilizzare gli strumenti specifici della statistica al fine di trasferire le conoscenze acquisite in campo umanistico.

Quanti di noi usano la macchina? Moltissimi, ma non per questo siamo ingegneri meccanici... Quanti di noi usano il telefono? Ancora di più, ma non per questo siamo esperti in telecomunicazioni... Quanti tra voi lettori, alla fine di questo percorso, diventeranno esperti statistici? Sicuramente nessuno, ma sarete però in grado di utilizzare alcuni strumenti statistici per soddisfare necessità di ordine pratico nell'esperienza di vita quotidiana.

Questo manuale vuole, quindi, rappresentare uno strumento pratico per coloro che abbiano la necessità di applicare, in campo umanistico appunto, un approccio scientifico che consenta di osservare e valutare fenomeni legati al mondo dell'istruzione, della formazione e in tutti i campi della ricerca sociale.

Capitolo 1

La valutazione

Prima di aprire un discorso sulla valutazione occorre riflettere su quale sia il significato profondo del "valutare". Quali sono i fini della valutazione oggi? Quali i mezzi più idonei e le strategie più sofisticate per valutare in un'epoca così cómplessa? Chi deve valutare? Chi o che cosa deve essere valutato? Dove? Quando? Come? Perché?

Il termine *valutare* deriva dal latino *vàlitus*, che significa valere, avere prezzo, stimare. La valutazione è uno dei compiti fondamentali dai quali nessun insegnante, manager, sociologo, psicologo e, più in generale, qualsiasi professionista può esimersi, pertanto è necessario che, chi è chiamato a valutare, abbia ben chiaro il fine che intende perseguire, l'oggetto al quale deve essere dato un valore e quale valore sia opportuno attribuirgli. Questo stesso principio vale per la valutazione di una qualsiasi organizzazione che, per crescere e mantenersi competitiva sul mercato, deve saper condurre analisi mirate alla rilevazione continua delle risorse, sia in termini economici e finanziari, che in riferimento alle risorse strutturali e soprattutto del capitale umano che la costituisce. In campo sociologico risulta essenziale comprendere le interconnessioni e le relazioni causa-effetto di fenomeni storico-culturali che caratterizzano le società complesse. E, ancora, in campo psicologico il settore Ricerca & Sviluppo è alla continua ricerca di metodi psicometrici sempre più precisi e mirati alla conoscenza della psiche umana. Insomma, ogni settore, in campo scientifico ed umanistico, non può esimersi da compiere continue valutazioni finalizzate alla comprensione di specifici fenomeni e alla presa di decisione

in condizione di incertezza.

La statistica gioca un ruolo fondamentale nell'analisi dei più variegati fenomeni poiché, attraverso approcci metodologici scientifici, rigorosi e matematicamente formalizzati, ne indaga le caratteristiche proprie ed indotte e, al tempo stesso, consente di fare previsioni sull'evolversi degli stessi, considerando, come vedremo nel dettaglio, anche il margine di errore nelle misurazioni e nelle previsioni. La statistica è, quindi, caratterizzata da una fortissima componente trasversale che consente il suo impiego praticamente in tutti i settori dello scibile umano: medicina, istruzione, economia, evoluzione storia e sociale. In seguito useremo, pertanto, più in generale l'espressione *"fenomeno reale"* per riferirci ad un qualunque fenomeno che possa essere indagato e conosciuto.

Il Consiglio Europeo straordinario di Lisbona, tenutosi nei giorni 23 e 24 marzo 2000, è nato dalla volontà di imprimere un nuovo slancio alle politiche comunitarie attraverso la progettazione di un obiettivo strategico per l'Unione finalizzato alla risoluzione del problema occupazionale, alla realizzazione di riforme economiche e alla coesione sociale nel contesto di un'economia basata sulla conoscenza. Il Consiglio Europeo ha avvertito, quindi, l'esigenza di divenire una potenza competitiva nello scenario dell'economia della conoscenza.

Le economie e le società contemporanee, quindi, stanno vivendo continui e profondi cambiamenti che possono essere riassunti come segue:

- affermarsi della globalizzazione in tutti i settori economici. A seguito della quale l'Europa è chiamata ad essere all'avanguardia in tutti i settori nei quali è forte l'intensificarsi della concorrenza;
- arrivo repentino ed affermarsi massivo delle tecnologie dell'informazione e delle comunicazioni sia nella sfera professionale che in quella privata. A seguito della quale è nata l'esigenza, in Europa come nel resto del mondo, di promuovere forme di alfabetizzazione informatica.

I sistemi di istruzione dell'Unione Europea si sono, di conseguenza, dovuti impegnare ad una revisione completa dei sistemi stessi per garantire l'accesso, per tutti i cittadini della comunità, alla formazione

lungo tutto l'arco della vita. Il compito che si è dato l'Unione Europea è quello di modellare questi cambiamenti in modo coerente con i propri valori. Il Consiglio Europeo di Lisbona, dunque, ha cercato di formulare orientamenti in grado di cogliere le opportunità offerte dalla nuova economia, prefissandosi, quindi, l'obiettivo strategico di "diventare l'economia basata sulla conoscenza più competitiva e dinamica del mondo, in grado di realizzare una crescita economica sostenibile con nuovi e migliori posti di lavoro e una maggiore coesione sociale". Quelli di Lisbona erano obiettivi di medio e lungo termine che richiedevano un eccezionale impegno politico ed organizzativo. Lo spazio europeo dell'istruzione e della formazione è oggi il punto ideale di raccordo dei progetti culturali, tecnologici e scientifici e punto di partenza di ogni progetto sociale ed economico. L'Italia in questa prospettiva, negli ultimi anni, si è impegnata a rinnovare profondamente scuola, università e strutture di formazione. Qualcosa si è fatto ma molto è ancora da fare.

È tempo di guardare lontano e di trovare i modi e i mezzi più idonei a garantire nelle nuove generazioni lo sviluppo di un pensiero razionale, critico e, soprattutto, creativo che consenta a tutti e a ciascuno di vivere positivamente e serenamente la complessità della nostra era. I docenti di ogni ordine e grado hanno la responsabilità di essere protagonisti attivi nella costruzione di un sistema di istruzione innovativo e competitivo. Per essere in grado di affrontare è gestire il cambiamento, è necessario saper valutare il cambiamento stesso perchè, come vedremo, solo chi è in grado di operare valutazioni attendibili ed oggettive è in grado di compiere scelte ottimali.

Il termine valutazione è oggi sempre più utilizzato, sia nel linguaggio comune che in quello specifico dei diversi settori disciplinari. Si parla di valutazione, infatti, in tutti i settori economici, sociali, politici e più in generale in ogni contesto organizzativo di qualunque natura e genere.

Ma cos'è la valutazione? E soprattutto, quando è opportuno ricorrere a strumenti specifici di valutazione? Se riflettiamo un attimo su quanto ci circonda, ci accorgiamo che è difficile trovare un fenomeno reale o una circostanza che non sia sottoposta ad un criterio di valutazione, fondato sulla nostra esperienza o basato su strumenti specifici.

Quando ci svegliamo la mattina la prima cosa che facciamo è decidere cosa indossare e la nostra decisione è subordinata ad una valutazione, per esempio in relazione alle condizioni atmosferiche o al luogo nel quale ci dobbiamo recare. Così, decidiamo di indossare un abito elegante per andare ad una cerimonia, comodo e pratico per affrontare un lungo viaggio, di lana in caso di freddo, di lino se fa particolarmente caldo, ecc. Stessa cosa vale per la scelta della colazione, per il pranzo, per il mezzo di trasporto, insomma ognuno di noi durante la giornata ricorre sovente a valutazioni che condizionano le scelte della vita quotidiana. Si utilizzano ancora criteri di valutazione anche quando si deve fare un acquisto importante, per esempio la casa o l'auto, quando si valuta una proposta di lavoro, quando si progetta il proprio futuro. E, ancora, chi di noi non è stato sottoposto qualche volta nella sua vita ad una valutazione? Persino le nostre mamme, sin dai primissimi giorni di vita, ci hanno valutato: "Oh! Come è buono!"; "Assomiglia al padre."; "È così capriccioso, non dorme mai!", ecc. Per non parlare poi di quando si va a scuola: chi di noi non è mai stato sottoposto ad una prova di verifica, ad un'interrogazione, ad un esame? Chi non ha fatto mai il confronto con i voti ottenuti da un compagno?

Oggi ogni organizzazione, pubblica o privata che sia, ha un comitato di valutazione preposto a stabilire l'efficienza e l'efficacia del proprio operato, sia esso rappresentato da servizi o da prodotti. Si parla di valutazione dei dirigenti, del personale, di ascolto del cliente per prendere decisioni (*customer satisfaction*). Insomma sembrerebbe che in ogni ambito reale non si possa fare a meno di ricorre alla valutazione.

L'atto del valutare è necessario, quindi, ogni qualvolta ci troviamo in una situazione nella quale dobbiamo prendere una decisione. In tutti gli esempi riportati abbiamo parlato di valutazione, ma allora è possibile affrontare il tema della valutazione da un punto di vista scientifico? Esiste un metodo trasversale a diversi ambiti di applicazione, che consenta di osservare in modo corretto un fenomeno reale ed aiuti a prendere la decisione più vantaggiosa? Quando osserviamo un fenomeno reale è necessario avere strumenti che ci consentano un'astrazione formale del fenomeno stesso, mirata alla comprensione profonda di esso. Questa fase della nostra indagine conoscitiva di un fenomeno reale possiamo chiamarla con il termine di *formalizzazione*.

Solo dopo aver formalizzato un fenomeno reale, possiamo *individuare il modello* matematico o statistico che ci consente di misurare il fenomeno stesso e di procedere, quindi, ad una sua *valutazione*.

Dopo una disamina di alcuni dei principali problemi connessi alla valutazione, affronteremo l'argomento fornendo un metodo generale che sia valido per una pluralità di casi, riservandoci di trattare metodi specifici solo a casi di studio concreti.

Capitolo 2

La rilevazione dei fenomeni reali

Sai ched'è la statistica? È na' cosa
che serve pe fà un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.
Ma pè me la statistica curiosa
è dove c'entra la percentuale,
pè via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perch'è c'è un antro che ne magna due.
("La Statistica" - Trilussa)

In questo capitolo cominceremo a muovere i primi passi nel mondo della statistica avvicinandoci prima di tutto al linguaggio specifico di questa scienza per inoltrarci a piccoli passi all'interno del suo metodo.

L'esperienza di questi anni ci ha guidato alla consapevolezza di un approccio diffidente degli studenti, che frequentano facoltà ad indirizzo umanistico, alle discipline scientifiche. Abbiamo spesso

constatato, infatti, che l'approccio alla statistica viene spesso vissuto con ansia e timore, a causa di una radicata convinzione che perpetua in molti, circa l'incapacità di comprendere ed apprendere concetti legati al *mondo dei numeri*.

La comprensione è l'elemento chiave dell'apprendimento; se lo studente pensa di poter apprendere la statistica cercando di memorizzare formule e procedure si avvia ad un lavoro tanto arduo, quanto assolutamente inutile. Lo scopo di questo manuale è, appunto, quello di guidare lo studente nel vivo della statistica a brevi passi, sostenendo nella comprensione di ogni elemento che incontra nel suo percorso formativo e consentendogli di costruire il proprio apprendimento in modo graduale e consapevole.

Gli studenti, soprattutto di ambito umanistico, considerano in generale la matematica, e tutte le discipline ad essa connessa, come una delle materie tradizionalmente più ostiche; una materia considerata accessibile solo a persone particolarmente dotate di capacità logiche, che per i più resta una disciplina appresa con meccanica ripetizione di procedure delle quali non si comprendono le logiche e le interconnessioni.

Numerosi studi, condotti nel campo della Ricerca & Sviluppo su soggetti in età neo-natale, mettono in evidenza, invece, una naturale predisposizione del cervello umano alle leggi della matematica. Brian Butterworth, professore di Neuropsicologia Cognitiva presso l'Istituto di Neuroscienze Cognitive dell'Università di Londra, ha condotto indagini su neonati che hanno dimostrato un'innata attitudine del cervello umano di conoscenza della realtà che li circonda attraverso un approccio matematico.

Ciò che deve essere messo in discussione, partendo da queste premesse, è allora l'impianto metodologico e didattico che sottende ai processi di insegnamento-apprendimento delle discipline scientifiche. È di estrema importanza, infatti, che il docente, di qualunque ordine e grado di scuola, utilizzi strategie mirate ad un approccio motivante alla propria disciplina al fine di consentire la maturazione del desiderio di scoperta e di conoscenza. Proprio questo ci proponiamo di fare noi nei confronti del metodo statistico.

2.1 Considerazioni generali sul metodo statistico

La ricerca scientifica applicata in campo educativo, come peraltro nei più svariati ambiti, si fonda principalmente su indagini strutturate secondo rigorosi metodi statistici. Per fare previsioni di intervento che risultino mirate, adeguate ed efficienti è necessario prima di tutto conoscere in modo dettagliato il fenomeno reale sul quale si intende intervenire. Quest'ultima affermazione porta ad una riflessione profonda sulla professionalità di chi è coinvolto nella conduzione di una ricerca scientifica in campo umanistico. Riteniamo, infatti, che un'indagine corretta, e dal punto di vista epistemologico e dal punto di vista scientifico, necessita di un pluralità di competenze globali e specifiche che difficilmente possono concentrarsi in una sola persona. Risulta pertanto essenziale, nel settore di Ricerca & Sviluppo in campo umanistico, un lavoro coordinato di esperti in pedagogia, psicologia, sociologia, didattica e statistica, che riesca ad analizzare tutte le molteplici sfaccettature del fenomeno reale osservato al fine di trarne le informazioni e le conoscenze necessarie alla presa di decisione corretta sugli interventi attuativi.

La statistica è un metodo scientifico che risiede nella logica dell'Illuminismo. Noi riteniamo, infatti, che il primo statistico sia stato Galileo Galilei, che per primo ha introdotto il metodo sperimentale. Il metodo statistico consente, attraverso un metodo di studio scientifico rigoroso, di sintetizzare le informazioni che si rilevano tramite l'osservazione di un fenomeno reale e di estendere induttivamente i risultati a casi più generali. Il metodo scientifico, infatti, procede dall'osservazione di alcune caratteristiche di un fenomeno, registrando in modo dettagliato tutte le repliche che si manifestano di ogni specifica caratteristica. La sintesi consente di semplificare e rendere di più immediata comprensione le informazioni che, in caso contrario, risulterebbero di per sé troppo articolate e complesse. La generalizzazione consente poi di estendere il risultato dell'analisi effettuata su un gruppo limitato di unità statistiche (campione) ad un intera collettività di appartenenza (universo, popolazione).

La statistica, quindi, tratta *caratteri*, cioè aspetti della realtà osservabili (lo stato di una spiaggia, la professione di una persona, la valutazione degli apprendimenti e di fenomeni sociali, ecc.) e *variabili*,

nel senso che questi caratteri possono assumere espressioni differenti (balneabile, inquinata e altro ancora; calzolaio, scrittore, deputato, regista, ecc.). I caratteri devono poter essere rilevati sui soggetti che li esprimono, che identifichiamo come *unità statistiche*, le quali devono appartenere ad una *collettività* (un unico dato rilevato su un singolo individuo è privo di interesse per la statistica). Le scienze umanistiche sono classificabili ed individuabili come espressione di *fenomeni reali*, che investono l'individuo in quanto essere sociale in continua evoluzione e trasformazione. Il processo di insegnamento-apprendimento è, ad esempio, un fenomeno reale socialmente e storicamente contestualizzato, pertanto l'obiettivo più generale è quello di conoscere il comportamento di questo fenomeno nelle sue diverse manifestazioni. Ciò che caratterizza un fenomeno reale, come vedremo meglio nel dettaglio, è la sua caratteristica di *replicabilità* ed ogni volta che si replica individueremo una "*unità di replica*", ossia un nuovo individuo sul quale andiamo ad osservare le caratteristiche di uno specifico fenomeno reale. L'insieme di tutte le repliche andrà a costituire il *collettivo*, o *popolazione*, su cui il fenomeno reale si manifesta.

Gli scopi della statistica, come vedremo meglio nel dettaglio, sono quindi di duplice natura: *sintetizzare* e *generalizzare*. Sintetizzare significa predisporre i dati raccolti in una forma (tabelle, grafici, sintesi numeriche) che consenta di comprendere meglio i fenomeni rispetto ai quali è stata eseguita la rilevazione. La sintesi viene incontro all'esigenza di semplificare, che a sua volta deriva dalla limitata capacità della mente umana di gestire informazioni articolate, complesse o multidimensionali. I metodi orientati a soddisfare questa finalità appartengono alla statistica descrittiva. Il secondo scopo della statistica è quello di estendere il risultato dell'analisi effettuata sui dati di un gruppo limitato di unità statistiche (campione) all'intera collettività di appartenenza (universo, popolazione). L'estensione avviene secondo metodi di induzione che rappresentano il contenuto della statistica inferenziale.

Nella realtà esistono molteplici fenomeni per i quali è necessario trovare la misura di una o più caratteristiche. Tante più caratteristiche riusciamo a misurare tanto più diamo concretezza al fenomeno reale oggetto di studio. La misurazione di un fenomeno reale deve essere oggettiva, ripetibile e convertibile. Il metodo statistico, appunto,

consente di estrapolare un'astrazione formale da un fenomeno reale. Un processo decisionale non può essere attendibile se non si sottopongono i fenomeni reali ad una *formalizzazione*, alla *scelta di un modello* e ad un'attenta *valutazione* degli stessi fenomeni osservati, al fine di conoscerli e comprenderli.

Il metodo statistico ha, quindi, come obiettivo lo studio della distribuzione di un fenomeno reale attraverso una *sintesi*, una *variabilità* e una *forma*.

Nei prossimi capitoli saranno analizzate nel dettaglio, per ogni tipologia di carattere statistico, la *sintesi*, la *variabilità* e, dove possibile, la *forma*; tuttavia, diamo qui un breve accenno introduttivo a questi tre concetti per fare il quadro della situazione.

Per *sintesi* intendiamo il valore più rappresentativo di una distribuzione, ossia fra tutte le repliche quella che ti aspetti sia la più frequente, la più probabile, la più rappresentativa.

La *variabilità* indica la differenza di atteggiamenti rispetto al valore di sintesi, ossia il modo in cui si comportano le nostre unità statistiche in riferimento al valore espresso dalla sintesi. La sintesi è robusta se e solo se è caratterizzata dall'avere una bassa variabilità.

La *forma* indica il modo in cui sono distribuiti i dati e ci dice se la variabilità è rivolta più a valori alti o bassi rispetto alla sintesi. La forma ci consente, quindi, di rappresentare graficamente, di disegnare il nostro collettivo e di verificare se possa essere modellato attraverso un modello matematico, economico, statistico, ecc.

2.2 I concetti che stanno alla base del linguaggio statistico: fenomeno reale, collettivo, unità statistica, carattere e modalità

Introducendo il discorso sulla valutazione abbiamo fatto riferimento all'etimologia di questo termine che deriva dal latino *vàlitus* "dare il prezzo, stimare". Alla base della valutazione dobbiamo, quindi, porre il concetto di *misura*, ossia un criterio convenzionale che stabilisca un valore quantitativo o qualitativo da assegnare ad un'unità d'osservazione che sia invariante nel tempo e nello spazio. La realtà che ci circonda è ricca dei più svariati fenomeni che si possono osservare e

valutare ed ognuno di essi è, a sua volta, definito da tutta una serie di caratteristiche specifiche. È necessario trovare per ognuna di queste caratteristiche la misura più idonea, poiché tante più caratteriste del fenomeno stesso riusciamo a misurare, tanto più diamo ad esso concretezza. La misurazione di un fenomeno deve essere, come vedremo meglio di seguito, oggettiva, ripetibile e convertibile. La statistica consente, attraverso un metodo di studio rigoroso, di sintetizzare le informazioni che si rilevano dell'osservazione di un fenomeno reale e di estendere induttivamente i risultati a casi più generali.

Abbiamo ribadito più volte che la statistica indaga *fenomeni reali*, risulta pertanto opportuno precisare a cosa ci si riferisca con questa espressione. Tutto ciò che ci circonda può essere catalogato come un fenomeno reale: la produzione metalmeccanica; lo stato di salute di una popolazione; il livello di apprendimento di un gruppo di studenti; la soddisfazione dei clienti rispetto ad un servizio erogato; e si potrebbe continuare con esempi infiniti. Ogni fenomeno è, infatti, un evento che si manifesta attraverso una serie di caratteristiche specifiche che possono essere indagate e misurate.

Prima di parlare di valutazione di un fenomeno è necessario, quindi, definire dettagliatamente il fenomeno reale sul quale effettuare la valutazione. In genere il compito di questa fase della ricerca è assegnata ad un esperto, ossia ad uno studioso della materia oggetto di studio, che in base alla sua esperienza e competenza definisce l'impianto del protocollo del disegno della ricerca.

Dovendo, tuttavia, valutare un fenomeno reale, si può pensare che il fenomeno sia replicabile sotto determinate condizioni ad un insieme di *unità*, ossia un insieme di oggetti o persone da sottoporre all'analisi di valutazione. In linea del tutto esemplificativa, possiamo immaginare di osservare il fenomeno reale riferito ad un gruppo di unità accomunate dalla stessa caratteristica e su cui convergono gli interessi della ricerca. Chiameremo detto insieme di unità *collettivo statistico* o *popolazione* o ancora *universo*. L'insieme di un certo numero di unità statistiche che siano classificabili in relazione ad uno o più caratteri costituisce, quindi, il collettivo statistico o la popolazione. Per fare un esempio possiamo dire che un alunno rappresenta un'unità statistica del collettivo, o della popolazione, classe I A, ma è anche un'unità statistica della popolazione maschile degli alunni del "Liceo

G. Galilei" e, ancora, è un'unità statistica dei residenti di 15 anni di età del comune di "Chieti".

Un'*unità statistica* di osservazione è, quindi, dal punto di vista statistico, un oggetto (fisico) o soggetto (persona) su cui si vuole effettuare una valutazione in relazione ad uno specifico fenomeno reale. L'unità statistica rappresenta, dunque, l'unità elementare su cui vengono osservati i caratteri oggetto di studio, intendendo per caratteri, come vedremo meglio più avanti, gli aspetti della realtà osservabili. A titolo di esempio rappresentano unità statistiche: gli alunni di una scolaresca; i pazienti di un ospedale; gli studenti di una facoltà; le merci di un supermercato; i banchi di una scuola, ecc. È bene tuttavia precisare che il concetto di unità è strettamente legato al fenomeno che si vuole valutare, nel senso che "l'unità banco" e l'unità "alunno", pur essendo riferiti alla stessa classe, non possono appartenere allo stesso fenomeno che si vuole studiare.

Ciascuna unità del collettivo è identificabile da una serie di caratteristiche necessarie ai fini della valutazione del fenomeno reale che si vuole studiare. In statistica le diverse caratteristiche, che sono osservabili su ciascuna unità del collettivo preso in esame, vengono identificate con il termine *carattere*. Così, per esempio, a ciascuna unità di una scolaresca possiamo rilevare il carattere "sesso", il carattere "età", il carattere "voto in condotta", il carattere "religione professata", il carattere "giudizio espresso nei confronti di un cartone", ecc.

A ciascun carattere viene assegnato un criterio di misura che sarà associato ad ogni unità del collettivo. Il modo con cui tale misura si manifesta sull'unità si chiama *modalità*. Pertanto se il carattere preso in considerazione è, ad esempio, il *sesso* allora le possibili modalità saranno *uomo* o *donna*. Se il carattere, invece, è l'*età* allora le modalità saranno il *numero* di anni associato all'individuo osservato; mentre, se il carattere è il *voto* in condotta allora le possibili modalità saranno espresse da un *numero* della scala dei numeri naturali compreso tra zero e dieci. E, ancora, se il carattere è il *giudizio* dato ad un film allora la modalità è l'*aggettivo* esprimente il parere espresso dagli spettatori che lo hanno visto.

Come possiamo notare dagli esempi riportati, il concetto di misura adottato è leggermente diverso da quello che comunemente utilizziamo; infatti, nell'immaginario collettivo, la misura è espressa sempre

da una quantità numerica. Nel nostro contesto la misura è vista in una forma più generale, cioè come la valutazione fatta su un'unità. In questo senso la misura può essere sia *quantitativa* che *qualitativa*. Quindi, essere uomo per il carattere *sesso* rappresenta il valore che assumerà l'unità osservata rispetto alla caratteristica oggetto di studio, cioè il sesso.

Sostanzialmente dire "Marco è un maschio e pesa 84 kg" significa aver fatto una rilevazione sull'*unità statistica* Marco per la quale il *carattere qualitativo* sesso si manifesta nella *modalità* maschio (attributo, aggettivo) e il *carattere quantitativo* peso in kg si manifesta nella *modalità* 84 (quantità numerica).

Da una più attenta riflessione sul comportamento delle misure dei caratteri, ci accorgiamo che in molte circostanze la modalità del carattere dipende dall'istante in cui viene osservato. In altri termini, se è intuibile il fatto che il sesso è invariante nel tempo, altrettanto non si può dire per il peso di un individuo, per il voto preso da uno studente in una materia, ecc. C'è comunque da precisare che vi sono caratteri la cui invarianza è temporanea come lo stato civile, l'attività professionale, ecc. Così come quelli dipendenti dal tempo possono essere di *stato*, come il peso o la statura, in cui la modalità del carattere deve essere individuata nelle unità facendo riferimento ad un istante di tempo, altri caratteri sono *dinamici*, come il consumo di energia elettrica, in cui la modalità del carattere va osservata nell'unità facendo riferimento ad un intervallo di tempo.

Volendo riassumere attraverso un esempio concreto:

- *Fenomeno reale*: andamento corso di studi degli studenti della Facoltà di Scienze della Formazione;
- *Collettivo o Popolazione*: tutto l'insieme degli studenti iscritti alla Facoltà di Scienze della Formazione;
- *Unità statistica*: ognuno degli studenti della Facoltà di Scienze della Formazione;
- *Carattere*: tutte le caratteristiche che vogliamo osservare di ogni unità statistica (ad es. sesso, età, scuola secondaria di provenienza, anno di corso, voti riportati agli esami, numero di crediti maturato, ecc.);

- *Modalità* per i caratteri osservati: il sesso si può esprimere con le modalità "maschio" o "femmina"; l'età con modalità diverse rappresentate dai numeri da 19 in avanti; la scuola di provenienza con le modalità "liceo scientifico", "liceo pedagogico", "istituto tecnico industriale"; l'anno di corso con le modalità "primo", "secondo", "terzo"; i voti con le modalità espresse dai numeri da 18 a 30; e così via.

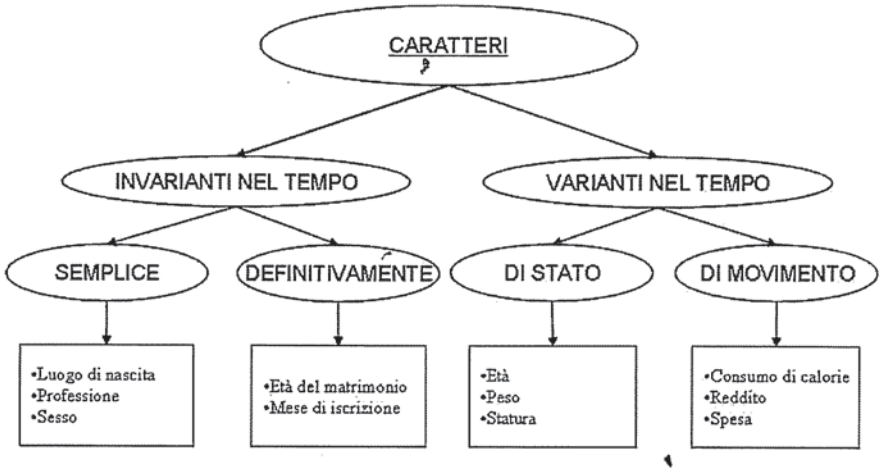


Figura 2.1: I caratteri statistici.

2.3 Classificazione dei caratteri statistici

Un fenomeno reale oggetto di osservazione ed indagine presenta tutta una serie di *caratteristiche* che devono essere misurate in maniera corretta. Se, quindi, il peso di un individuo è esprimibile in *kg* (chilogrammi), la professione e il sesso sono misurabili attraverso un attributo, ossia un aggettivo o un nome che specifica la modalità del carattere osservato sull'unità. In questo senso diciamo che i caratteri si dividono essenzialmente in: *caratteri quantitativi*, quando le *modalità* sono esprimibili con una misura numerica, ossia una quantità associata su una scala di misura predefinita; *caratteri qualitativi*, quando le

modalità sono attributi, ossia espressioni verbali, in genere aggettivi, che hanno la funzione di esprimere un giudizio sul carattere che si vuole misurare.

Volendo utilizzare un linguaggio più specifico dobbiamo dire che un carattere si definisce *variabile* se le sue modalità si esprimono con caratteri *quantitativi* ossia tramite numeri reali, mentre si definisce *mutabile*, oppure *variabile qualitativa*, se le sue modalità si esprimono con attributi, aggettivi.

Il numero di figli, l'età, l'altezza, il peso, il voto preso ad un esame, il numero di vani che compone un appartamento, i mq di una stanza, la quantità di macchine prodotte da un'azienda, e molti altri ancora, sono tutti esempi di *variabili*, ossia di caratteri di natura quantitativa. Il sesso, il colore degli occhi, il titolo di studio, l'attività lavorativa, la tipologia di manufatti prodotti da un'azienda, e molti altri ancora, sono esempi di *mutabili*, ossia di caratteri di natura qualitativa.

La classificazione di un carattere in variabile o mutabile ci dà l'immediata percezione che i caratteri non possono essere trattati allo stesso modo quando si affronta il problema della loro misurazione. Vedremo di seguito nel dettaglio le caratteristiche proprie di ognuno di essi. Per il momento ci limitiamo ad anticipare che i *caratteri qualitativi* possono essere distinti, in relazione alle loro specifiche caratteristiche, in *sconnessi* quando le modalità dalle quali possono essere misurati non sono ordinabili (sesso, religione, cittadinanza,...) o in *ordinabili* quando sono misurati attraverso un attributo che esprime la qualità, la performance, il gradimento, per il quale si può stabilire un ordine (voto: non sufficiente, sufficiente, buono, distinto, ottimo; gradimento: mi piace moltissimo, mi piace poco, non mi piace affatto, ...). I caratteri *quantitativi*, invece, sono sempre ordinabili poiché espressi tramite numeri. Tuttavia, possono essere ulteriormente distinti in: *discreti*, quando sono misurabili con numeri finiti entro un intervallo di variazione, ossia numeri naturali (numero dei figli: 1, 2, ..., 10, ...); o *continui*, se comunque si fissino due valori, entro l'intervallo in cui il carattere è osservabile, tutti i valori intermedi possono essere assunti come modalità del carattere (altezza: 1,80 m - 1,815 m - ... - 1,88 m - 1,894 m - ... - 1,90 m).

Una volta stabilito il criterio di misurazione resta da stabilire la potenzialità operativa per ogni tipo di carattere. In particolare potrebbe essere utile stabilire fra le modalità le relazioni esistenti o effettuare operazioni che consentano di interpretare i dati rilevati. In questo contesto facciamo notare che non tutti i caratteri hanno la stessa potenzialità operativa, nel senso che se si fa riferimento a modalità quantitative è possibile operare facendo ricorso agli strumenti aritmetici, altrettanto non è possibile tra le modalità di un carattere qualitativo. Anche in questo caso resta utile fare una classificazione dei caratteri a seconda del tipo e della potenzialità operativa (vd. Tabella "Potenzialità operativa dei caratteri statistici"). Dallo schema si mette in luce

Potenzialità operativa per tipo di carattere			
	Caratteri Qualitativi		Caratteri Quantitativi
Operatività	*Sconnessi*	*Ordinabili*	*Discreti* e *Continui*
Diversità	SI	SI	SI
Ordinamento	NO	SI	SI
Operazioni aritmetiche	NO	NO	SI

Tabella 2.1: Potenzialità operativa dei caratteri statistici.

una sorta di gerarchia tra i caratteri. Il più "basso" in graduatoria, il *meno potente* dal punto di vista operativo, è il carattere qualitativo sconnesso, sul quale è possibile effettuare solo confronti di uguaglianza e diseguaglianza tra le modalità. Poi si passa ai caratteri qualitativi ordinabili, sui quali, oltre alle diversità tra le modalità, è anche possibile operare appunto un loro ordinamento; infine, c'è il carattere quantitativo, possiamo dire il *più potente* dal punto di vista operativo, che si presta appunto ad ogni tipo di operazione. Infatti, essendo la modalità un numero allora è possibile eseguire sia il confronto sia l'ordinamento sia le operazioni aritmetiche. Concludendo, proprio in virtù della potenzialità operativa che abbiamo sopra indicato, diciamo che tutte le metodologie applicabili ai caratteri qualitativi sconnessi basati sui confronti tra le modalità, sono applicabili sia ai caratteri qualitativi ordinabili sia ai caratteri quantitativi; mentre, i metodi di analisi per i caratteri qualitativi ordinabili basati sull'ordinamento delle modalità, non sono applicabili ai caratteri qualitativi sconnessi ma sono applicabili ai caratteri quantitativi. Infine, i metodi per i caratteri

quantitativi basati su operazioni algebriche, non sono replicabili per nessun tipo di carattere qualitativo.

2.3.1 La misura di un carattere qualitativo

Un carattere qualitativo abbiamo detto che si definisce *mutabile*, oppure *variabile qualitativa*, e che le sue modalità si esprimono con attributi, aggettivi.

I caratteri qualitativi a loro volta sono classificabili in *sconnessi*, quando le modalità qualitative del carattere non seguono un ordinamento logico, nel senso non rispettano un criterio di ordinamento oggettivamente riconosciuto. I caratteri qualitativi sconnessi sono definiti anche *variabili nominali*. Esempi di questi caratteri sono: il sesso, con le modalità "maschio" o "femmina"; il tipo di diploma conseguito alla maturità, esprimibile con le modalità "diploma liceale", "perito tecnico", "ragioniere"; l'attività lavorativa dei genitori di una scolaresca, con le modalità "disoccupato", "operaio", "impiegato". Posso determinare se un soggetto è maschio o femmina, se due soggetti presentano o meno lo stesso attributo, ma non posso stabilire un ordine logico ponendo su una scala gli attributi maschio - femmina; non possiedo, infatti, nessun elemento che mi consenta di stabilire se su una scala ordinata viene prima l'attributo maschio o l'attributo femmina. Un carattere qualitativo sconnesso si misura su *scala nominale*, detta anche *categoriale non ordinata*.

Un carattere qualitativo è invece *ordinabile* quando le modalità possono essere logicamente poste in un ordine universalmente riconosciuto, sia esso crescente o decrescente; ossia quando tra le diverse modalità, con cui si può esprimere un carattere, esiste una relazione d'ordine. Un carattere qualitativo ordinabile si misura su *scala ordinale*, detta anche *categoriale ordinata*. Esempi di questo genere di caratteri sono: il giudizio espresso ad una verifica di esame, con le modalità "non sufficiente", "sufficiente", "buono", "distinto" e "ottimo"; il grado di istruzione, con le modalità "analfabeta", "licenza elementare", "licenza media", "diploma", "laurea di I livello", "laurea magistrale", "dottorato"; il giudizio espresso dall'utenza in relazione ad un servizio erogato, con le modalità "non soddisfatto", "poco soddisfatto", "soddisfatto", "molto soddisfatto". Alcuni autori definiscono questi caratteri

"semi-quantitativi" in relazione all'utilizzo degli stessi nel settore di Ricerca & Sviluppo.

Circa i caratteri qualitativi ordinabili possiamo fare una ulteriore classificazione distinguendoli in *caratteri qualitativi ordinabili rettilinei* e *caratteri qualitativi ordinabili ciclici*. Esempi di caratteri ordinabili rettilinei sono oltre a quelli sopra elencati, ossia il giudizio attribuito ad una verifica di esame, il grado di istruzione, la classe frequentata, espressa con le modalità "prima", "seconda", la posizione occupazionale con le modalità "operaio generico", "operaio specializzato", "capo squadra", "impiegato generico", "impiegato di concetto", "funzionario", "dirigente", "capo area". Invece, come esempi di caratteri qualitativi ordinabili ciclici troviamo: il giorno della settima stabilito per una attività, il mese di nascita.

2.3.2 La misura di un carattere quantitativo

Un carattere quantitativo è esprimibile, quindi, attraverso modalità numeriche, ma occorre fare qualche ulteriore precisazione poiché non tutti i caratteri quantitativi presentano le medesime caratteristiche. Per far comprendere al meglio questi concetti, procederemo facendo alcuni esempi.

Il numero dei figli, dei clienti, degli studenti, il voto conseguito ad un esame universitario, sono senza alcun dubbio caratteri quantitativi elencabili ed ordinabili su una scala numerica; tuttavia, essi sono rappresentati dall'insieme dei numeri naturali e di conseguenza non ammettono intervalli di misura tra di loro. Posso dire di aver 2 figli, ma non 2 figli e mezzo, posso prendere 25 all'esame di statistica, ma non 25 e qualcosa, e così via. Un carattere quantitativo esprimibile con i numeri naturali (0, 1, 2, ...) si definisce in statistica *variabile discreta*.

Il peso, l'altezza, lo stipendio, il tempo, sono anch'essi caratteri quantitativi ordinabili, ma la scala numerica sulla quale si possono rappresentare può assumere un qualsiasi valore contenuto in un intervallo reale. Posso dire di essere alto 1,78 m, di pesare 78,5 kg, ecc. Se, comunque si fissino due valori, entro l'intervallo in cui il carattere è osservabile, tutti i valori intermedi possono essere assunti come modalità di carattere (peso, altezza), l'insieme delle modalità assumibili può essere messo in corrispondenza biunivoca con l'insieme dei numeri

reali. Un carattere quantitativo esprimibile con l'insieme dei numeri reali, quindi, si definisce in statistica *variabile continua*.

I caratteri quantitativi, inoltre, possono essere classificati in: *rettilinei*, quando le modalità sono logicamente disponibili su un asse orientato, ossia con modalità numeriche che hanno un origine di riferimento; o *ciclici*, quando, invece, l'origine delle modalità coincide con l'ultima. A titolo di esempio i caratteri quantitativi rettilinei sono: età, peso, altezza, reddito, voto in trentesimi ad un esame, ecc. Sono, invece caratteri quantitativi ciclici: latitudine del luogo di nascita, ora della giornata, ecc.

Abbiamo detto che le relazioni di disuguaglianza, di ordinamento e le operazioni aritmetiche consentono di esprimere la valutazione di un carattere rispetto ad una unità di misura prestabilita.

Ma come si stabilisce una unità di misura? La vita quotidiana ci ha abituato a concetti di misura che hanno sostituito in tutto e per tutto il fenomeno che si vuole misurare e spesso si confonde il fenomeno che si vuole misurare con la misura stessa; in altri termini è come se la misura rappresentasse il fenomeno stesso. Ad esempio, il fenomeno reale statura di un individuo è, nell'immaginario collettivo, sostituito dalla misura in centimetri corrispondente. In pochi sanno qual è l'unità di riferimento della statura ma tutti, senza esitazione, sanno che *180 cm* corrisponde ad una statura alta e che *150 cm* ad una bassa. Se qualcuno ci interroga sul qual è l'altezza di un individuo non gli è sufficiente sapere che è alto o basso ma vuole saper quanto è alto o basso. Insomma, come se fosse nata prima la misura e poi la statura. Com'è facile intuire la realtà non è questa. La misura è una *convenzione* e la *misurazione* è un confronto rispetto alla convenzione. Per fare un esempio relativo ad una unità di misura convenzionale, prendiamo in considerazione il *metro*. Il termine *metro* è stato coniato nel 1675 da Tito Livio Burattini (al quale si deve un primo tentativo di definizione basato sulla lunghezza di un pendolo che batte il secondo). La prima definizione originale del metro, basata sulle dimensioni della Terra, viene fatta risalire al 1791, stabilita dall'Accademia francese delle scienze come 1/10000000 della distanza tra Polo Nord ed Equatore, lungo la superficie terrestre, calcolata sul meridiano di Parigi, ma solo il 7 aprile 1795 la Francia adottò il metro come unità di misura ufficiale. L'incertezza nella misurazione della distanza portò l'Ufficio

internazionale dei pesi e delle misure (BIPM) a ridefinire nel 1889 il metro come la distanza tra due linee incise su una barra campione di platino-iridio conservata a Sèvres presso Parigi. In Italia il metro è attuato mediante il campione dell'Istituto Nazionale di Ricerca Metrologica di Torino, nato dall'unione dell'ex Istituto Metrologico "Gustavo Colonnetti" (IMGC-CNR) e dell'ex Istituto Elettrotecnico Nazionale "Galileo Ferraris" (IEN). Nel 1960, con la disponibilità dei laser, l'undicesima "Conferenza generale di pesi e misure" cambiò la definizione del metro in *"la lunghezza, pari a* 1650763,73 *lunghezze d'onda nel vuoto, della radiazione corrispondente alla transizione fra i livelli* $2p^{10}$ *e* $5d^5$ *dell'atomo di kripton-86"*. Ma la storia non finisce qui. Nel 1983, infatti, la XVII "Conferenza generale di pesi e misure" definì il metro come la distanza percorsa dalla luce nel vuoto in 1/299792458 di secondo (ovvero, la velocità della luce nel vuoto venne definita essere 299792458 metri al secondo). Poiché si ritiene che la velocità della luce nel vuoto sia la stessa ovunque, questa definizione è più universale della definizione basata sulla misurazione della circonferenza della Terra o della lunghezza di una specifica barra di metallo e il metro campione può essere riprodotto fedelmente in ogni laboratorio appositamente attrezzato. Sempre grazie agli esperimenti in laboratorio, dalla fine del 1997 è possibile raggiungere un ordine di accuratezza dell'ordine di 10^{-10}m. Questo risultato è ottenibile sfruttando la relazione $\lambda = c/\nu$ (dove, λ=lunghezza d'onda, c=velocità della luce, ν=frequenza della radiazione), utilizzando oscillatori laser stabilizzati a frequenza conosciuta la cui radiazione viene utilizzata in sistemi di misura interferometrici. Il sistema numerico decimale ha ovviamente completato l'opera, definendo i multipli e i sottomultipli a cui, sempre per convenzione, è stato dato il nome: *centimetro, decimetro, chilometro*, ecc.. Stesso discorso vale per il *Chilogrammo* come unità di misura del peso; il *litro* come unità di misura della capacità; *gradi Celsius* per la temperatura ecc. Le unità di misura del Sistema Internazionale sono riportate nella tabella 2.3.2. L'effettiva misurazione viene fatta attraverso uno strumento, ossia un apparecchio tarato sul campione di riferimento. Compiere una misura con un apparecchio tarato significa leggere la posizione che assume un indice su una scala, che è stata già graduata mediante l'unità di misura scelta per il carattere da misurare. Ad esempio, nel caso di una bilancia si può stabilire

Grandezza fisica	Simbolo della grandezza	Nome dell'unità	Simbolo dell'unità
lunghezza	l	metro	m
massa	m	chilogrammo	kg
intervallo di tempo	t	secondo	s
Intensità di corrente	i	ampère	A
temperatura assoluta	T	kelvin	K
quantità di sostanza	n	mole	mol
intensità luminosa	Iv	candela	cd

Tabella 2.2: Unità di misura del Sistema Internazionale.

una scala ad **intervalli** di un grammo, nel caso del metro una scala ad **intervalli** di un millimetro. Questo implica che lo strumento nel rilevare la misura produce un'approssimazione; infatti la bilancia che, ad esempio, indica il peso di 854g ci dice che il peso dell'oggetto è più vicino a 854g piuttosto che 853g e a 855g. Insomma lo strumento indica che il peso x dell'oggetto è:

$$853,5\text{g} < x < 854,5\text{g}$$

Questa relazione sta a significare che la reale misura dell'oggetto è un qualsiasi valore compreso tra 853,5g e 854,5g, e che se si introducesse uno strumento con una taratura più precisa si avrebbe un numero comunque interno a quell'intervallo ma diverso, anche se molto vicino a 854g. Quindi, fissato lo strumento, quello che possiamo dire è che il rilevatore, affermando che il peso è di 854g, sostiene che il vero valore è più vicino a 854g piuttosto che 853g o a 855g. In altri termini l'errore massimo compiuto è di $\pm$ 0,5g e tutto questo implica che le misurazioni non sono mai esatte, ma comunque affette da un errore dovuto allo strumento di misura.

Alla naturale imprecisione dello strumento c'è tuttavia da aggiungere due possibili errori: l'*errore accidentale* e l'*errore sistematico*. L'errore accidentale è, per definizione, incontrollabile ed è ritenuto essere di media zero, nel senso che in prove ripetute esso si annulla presupponendo che lo stesso sia una volta in eccesso e un'altra in difetto. L'errore sistematico è sicuramente più grave e sta ad intendere che

per cause da accertare la misura tende ad essere prevalentemente in eccesso o in difetto conducendo il rilevatore ad un errore certo. Come avremo modo di precisare, le misure dei caratteri quantitativi, dette anche **a scala ad intervalli** sono, ai fini dell'analisi di una valutazione, da preferirsi.

Facciamo tuttavia notare che non sempre è possibile definire un'unità di misura che sia univocamente riconosciuta e standardizzata come quelle appena riportate.

Ad esempio il voto ad un esame, pur avendo una precisa scala di riferimento, può essere influenzato da tutta una serie di circostanze che possono andare dallo stato d'animo del docente, all'ordine di chiamata, alla presenza di pregiudizi, ecc.

Da quanto abbiamo visto emerge che la misura di un fenomeno reale è una fase delicata e complessa che non va assolutamente trascurata e che merita approfondimenti metodologici che dovranno essere meglio specificati in seguito.

2.4 La misura di una variabile latente

Il concetto di *variabile latente* è forse uno dei più affascinanti e dibattuti degli ultimi cinquant'anni. Le variabili latenti sono variabili non direttamente osservabili in quanto rappresentano concetti molto generali o complessi. Una delle maggiori difficoltà per un ricercatore di scienze dell'educazione e della formazione nell'esplicitare un modello statistico che descriva i nessi causali tra variabili, deriva dal fatto che le variabili oggetto dell'analisi non sono sempre direttamente osservabili; si pensi ad esempio alla valutazione della competenza che è fatta da una pluralità di aspetti: conoscenza, attitudine, abilità, relazionalità e altro ancora; o all'intelligenza che risulta ancora più complessa da descrivere e misurare direttamente.

Quando vogliamo definire una variabile latente diamo, quindi, alla nostra definizione un'accezione negativa poiché diciamo che una variabile latente è una variabile non direttamente osservabile e misurabile in quanto manca di una origine e di un intervallo di misurazione.

Molto intuitivamente possiamo dire che le variabili latenti possono essere osservate e misurate tramite una serie di variabili osservabili

direttamente, le quali possono essere aggregate in un modello per rappresentare in modo esplicito una data teoria. In questo senso le variabili latenti possono essere pensate come la rappresentazione di osservazioni su un fenomeno reale e sulle correlazioni di dati osservabili nell'ambiente.

L'uso di opportuni indicatori, quindi, può aiutare la misurazione delle variabili latenti. Per avere una descrizione più accurata di un concetto non misurabile direttamente facciamo un esempio concreto: il fenomeno del "bullismo" non può essere osservato direttamente poiché è estremamente complesso, esso tuttavia è determinato da tutta una serie di variabili, come ad esempio l'autostima, la capacità relazionale, la propensione ad assumere un ruolo da gregario o da leader in un gruppo, il rendimento scolastico, l'aspetto fisico, ... Tutte queste variabili sono direttamente osservabili e misurabili attraverso l'utilizzo di test standardizzati; l'applicazione di uno specifico modello statistico può, inoltre, consentire di valutare i nessi causali tra le diverse variabili osservate consentendoci così di effettuare una valutazione di fenomeno, quale quello del bullismo, di per sé non direttamente misurabile.

Il concetto di variabile latente ha riscosso un enorme successo nelle discipline statistiche, dando luogo ad una vasta letteratura sia di indirizzo teorico, sia in campo applicativo. In particolare, nelle scienze sociali e in psicometria, l'uso del concetto di variabile latente è stato largamente adottato per far fronte al problema di misurare quantità che, in natura, non possono essere direttamente osservate.

Il primo autore ad introdurre il concetto di variabile latente è stato Charles Spearman nel suo articolo del 1904 sul American Journal of Psychology per definire il concetto di intelligenza generale. Tuttavia, fu durante la seconda guerra mondiale che la metodologia statistica per lo studio delle variabili latenti venne formalizzata teoricamente. Il contributo di Paul F. Lazarsfeld al team multidisciplinare impiegato dal Dipartimento della Guerra del governo americano per gli studi sociali e psicologici del personale militare consistette nella formulazione della teoria e la dimostrazione dell'uso dei modelli a struttura latente nel quarto volume del The American Soldier: Studies in Social Psychology in WW II (Stouffer, 1949-50), intitolato "Measurement and Prediction". Più tardi, Lazarsfeld contribuì ad un capitolo sull'analisi a struttura

latente nel monumentale lavoro "Psychology: A Study of A Science" (1959) fino alla stesura del libro con Neil W. Henry, "Latent Structure Analysis" (1968), che colleziona e raffina i progressi fatti in questa metodologia statistica in venticinque anni. Negli ultimi quaranta anni, numerosi ricercatori di statistica, psicologia e sociologia hanno contribuito allo studio dei modelli riconducibili all'analisi a struttura latente.

Il passaggio dalle variabili osservate a quelle latenti non è un processo banale e richiede una particolare attenzione, considerando il fatto che gli indicatori osservabili sono solo approssimazioni dei costrutti latenti, pertanto in questo manuale ci limiteremo solo a dare questo accenno sul concetto di variabile latente e non ci addentreremo nella metodologia di misurazione delle stesse.

2.5 La misura dell'incertezza: la probabilità

Tra i fenomeni reali che richiedono di dover prendere una decisione ce ne sono alcuni che possono presentare diversi scenari ed alternative. Questi tipi di scenari prevedono che il verificarsi della loro evoluzione sia incerto, pertanto necessitano di una misura dell'incertezza del verificarsi dell'evento. Lo strumento di misura che ci consente di valutare l'incidenza con la quale si può verificare un evento è il *calcolo della probabilità*.

L'approccio corretto è, quindi, quello di descrivere un fenomeno reale, i cui risultati siano incerti, tramite una misura quantitativa chiamata *variabile casuale* con una specifica *distribuzione di probabilità*. Tale distribuzione sarà chiamata *modello decisionale*; nel senso che può essere utilizzata per assumere decisioni sulla base di una valutazione quantitativa. In questo contesto, si intuisce che la conoscenza dei vari modelli decisionali viene subordinata alla molteplicità di fattori che chi opera nel campo dell'educazione e della formazione è chiamato a gestire.

In questo paragrafo verranno proposti solo alcuni modelli di carattere generale che possono risolvere un'ampia gamma di problemi. In sostanza si propone essenzialmente un approccio metodologico più che un'elencazione di strumenti. Ampio spazio sarà dato, invece, al

metodo statistico ritenuto necessario per una corretta valutazione del modello decisionale.

Molte attività in campo umanistico, e non solo, si basano spesso sulla necessità di dover prendere decisioni in tempi rapidi per decidere tra quali alternative scegliere affinché il risultato di un processo educativo, formativo e sociale sia più efficace. In questo contesto, si possono incontrare due tipi di formatori: quello più sprovveduto, che basa la sua scelta sull'intuito o se vogliamo sull'esperienza; quello più competente che, invece, farà la sua scelta con l'ausilio di strumenti decisionali basati su valutazioni quantitative. Naturalmente ci piace pensare che le azioni importanti, in ogni settore che si preoccupa del welfare, vengano prese da questo secondo tipo di persona. Immaginiamo per un attimo che il manager in questione sia un medico che deve diagnosticare una nostra malattia sulla base di un sintomo. Preferiremmo il medico che diagnostica la malattia solo sull'indicazione del sintomo, casomai guardandoci solo in faccia o quello che dopo un'attenta visita ci prescrive una serie di analisi, dai cui risultati quantitativi ricava gli elementi necessari per stabilire una diagnosi? Anche in campo formativo ci auguriamo che le scelte dei formatori siano basate su una dettagliata osservazione ed analisi di fenomeni reali, al fine di essere in grado di prendere la decisione ottimale per colui che deve essere formato.

Il formatore deve essere in grado di riscrivere il problema reale attraverso una o più misure espresse su scala quantitativa, o anche su scala ordinale o nominale, purché ad ogni scelta sia associabile un valore numerico corrispondente alle modalità della misura proposta. Nella quasi totalità dei casi, le decisioni vengono prese in condizioni di incertezza. Ciò implica che ad ogni modalità della variabile scelta, può essere associata una probabilità. In questo paragrafo, sulla base di questi presupposti, svilupperemo i concetti essenziali del *calcolo delle probabilità*, ossia della disciplina che fornisce gli elementi formali per trattare in modo corretto il processo logico dei modelli di decisione in condizioni di incertezza.

Essenzialmente il calcolo delle probabilità può essere riassunto in tre elementi.

1. L'*esperimento casuale*, definito anche *prova* o *gioco*: ogni fenomeno

reale su cui prendere una decisione può essere visto come una partita (the game), ossia un gioco con precise regole da rispettare. Per meglio comprendere quanto vogliamo dire, pensiamo ad una azienda che vuole fare una politica di investimento finanziario acquistando i titoli in borsa. L'azienda, attraverso il suo manager, dovrà eseguire un'attività di acquisto dei titoli, ma acquistare i titoli significa stabilire un contratto (il prospetto informativo) che stabilisce le regole del gioco tra l'acquirente e l'istituto di credito delegato all'acquisto. In altri termini la compra-vendita dei titoli in borsa (naturalmente questo è uno degli innumerevoli esempi che in tal senso possono essere fatti) può essere visto come un esperimento casuale le cui prove definiscono dei risultati o, come vedremo avanti, degli *eventi aleatori*.

2. Gli *eventi*: il manager dovrà prendere una decisione in condizioni di incertezza, cioè ad ogni azione (scelta) corrisponde un possibile risultato. Diremo che ogni gioco prevede una serie di possibili risultati che chiameremo eventi. Quindi gli eventi possono essere visti come i possibili risultati del gioco (outcomes): volendo fare un altro esempio più semplice ma sicuramente più diretto del precedente, pensiamo ad una partita di calcio: essa convenzionalmente ha delle regole che stabiliscono i vincitori sulla base dei risultati della partita, ossia ad ogni risultato è associato un evento (vince la squadra che gioca in casa, vince la squadra che gioca fuori casa, pareggio). L'insieme degli eventi di un gioco viene chiamato spazio degli eventi, spazio campionario, spazio fondamentale o con terminologia anglosassone *outcomes space*.

3. La *misura degli eventi*: se l'evento è aleatorio, ossia il suo risultato è incerto, allora abbiamo bisogno di una misura della sua incertezza, che dia un peso diverso a tutte le possibili scelte e che colleghi le sottili sfumature tra gli eventi appartenenti allo stesso spazio. Tuttavia, cosìccome accade per tutti i concetti di misura (il metro, il chilogrammo, il litro, ecc.), si rende necessario stabilire una convenzione che renda la misura proposta riconosciuta e universale.

In questo paragrafo parleremo di gioco nell'accezione più generale del termine, intendendo con ciò un qualsiasi esperimento casuale. Ci riferiamo, quindi, a tutti quei problemi reali che quotidianamente si possono presentare e che, in qualche modo, definiscono uno spazio di eventi aleatori.

Per quanto concerne il concetto di eventi, dovendo operare all'interno di uno spazio, riteniamo necessario definire alcuni principi della sua algebra, in modo tale da avere gli strumenti per eseguire le principali operazioni. La parte più ampia della trattazione, infine, sarà quella riferita alla misura dell'incertezza, ossia quello che nel linguaggio diffuso, viene chiamato calcolo delle probabilità.

2.5.1 Lo spazio degli eventi

Da quanto appena detto, ad ogni esperimento casuale è associata una pluralità di possibili risultati che chiameremo eventi. L'insieme di tutti gli eventi è detto *spazio degli eventi* o *spazio fondamentale* o, per usare un linguaggio più statistico, *spazio campionario*.

Lo spazio degli eventi può essere costituito da un numero finito di eventi, in tal caso diremo che lo spazio è *discreto*, o da un infinità di eventi, in tal caso diremo che lo spazio è *infinito*. In questo lavoro detto spazio sarà indicato con il simbolo Ω. Ogni risultato possibile, indicato con le lettere dell'alfabeto scritte in maiuscolo, è detto *evento elementare* cosicchè l'insieme di tutti gli eventi elementari definisce lo spazio degli eventi.

Un sottoinsieme di Ω può essere ricavato da uno o più eventi elementari che, per ragioni di semplicità, sarà rappresentato attraverso un diagramma di Venn così come riportato in figura 2.2. Lo spazio degli eventi può essere visto, quindi, come un contenitore all'interno del quale ciascun evento è identificato con un insieme. Le operazioni tra gli eventi necessitano di un'algebra le cui operazioni elementari sono le seguenti.

1. Definiamo *evento complementare* dell'evento A, che indichiamo con $\bar{A}$, l'insieme dei punti di Ω che non appartiene all'evento A (2.3).

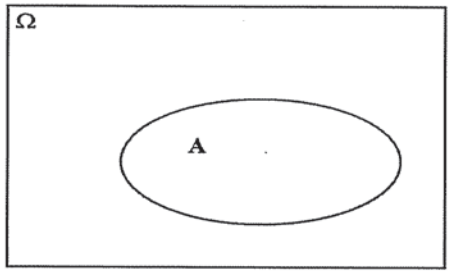


Figura 2.2: Lo spazio degli eventi(Ω).

2. Siano A e B due eventi di Ω, nel senso che sono due eventi dell'esperimento casuale che definisce lo spazio Ω, allora definiamo *evento unione* e lo indichiamo con $C = A \cup B$, l'insieme dei punti di Ω che appartengono o all'evento A, oppure all'evento B, ossia $C = \{x : x \in A \quad o \quad x \in B\}$, dove con x abbiamo indicato un punto qualsiasi dello spazio degli eventi Ω (2.4).

3. Siano A e B due eventi di Ω allora definiamo *evento intersezione* e lo indichiamo con $C = A \cap B$, l'insieme dei punti di Ω che appartengono sia all'evento A, sia all'evento B ossia $C = \{x : x \in A \quad e \quad x \in B\}$ (2.5).

4. Due eventi di Ω, A e B, sono *incompatibili* o *disgiunti* se non hanno elementi in comune, ovvero quando la loro intersezione è l'*insieme vuoto*. Formalmente l'*insieme vuoto* viene scritto come $\emptyset$, quindi l'intersezione tra due eventi incompatibili sarà indicata come $A \cap B = \emptyset$. Di seguito è riportato un esempio grafico dell'intersezione vuota tra due insiemi (2.6).

5. Siano A e B due eventi di Ω, definiamo l'*evento differenza* e lo indichiamo con $C = B - A$, l'insieme costituito dai punti di B che non appartengono ad A, ossia $C = \{x : x \in B \quad e \quad x \notin A\}$ (2.7).

Alle definizioni appena date sono associate alcune proprietà che possono risultare utili nel prosieguo della trattazione.

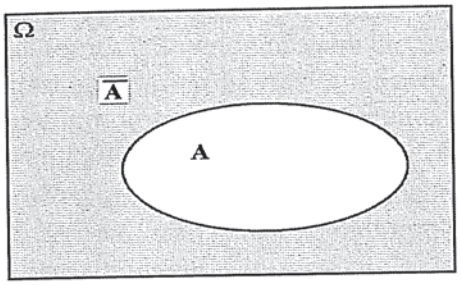


Figura 2.3: Rappresentazione grafica dell'evento complementare.

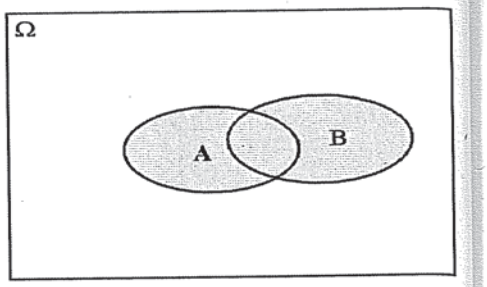


Figura 2.4: Rappresentazione grafica dell'evento unione.

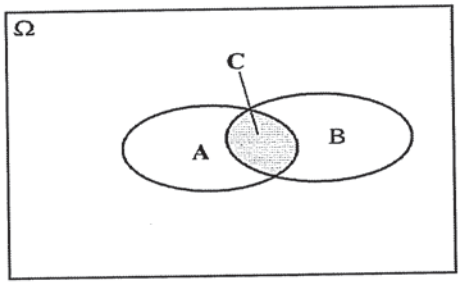


Figura 2.5: Rappresentazione grafica dell'evento intersezione.

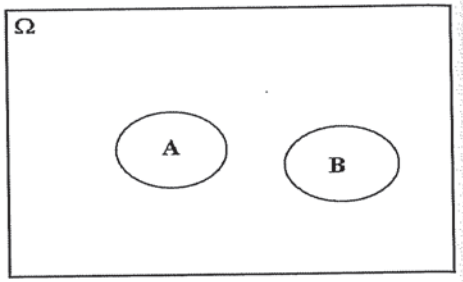


Figura 2.6: Rappresentazione grafica di due eventi incompatibili.

1. Idempotenza

$$A \cup A = A$$
$$A \cap A = A$$

2. Commutativa

$$A \cup B = B \cup A$$
$$A \cap B = B \cap A$$

3. Associativa

$$A \cup (B \cup C) = (A \cup B) \cup C$$
$$A \cap (B \cap C) = (A \cap B) \cap C$$

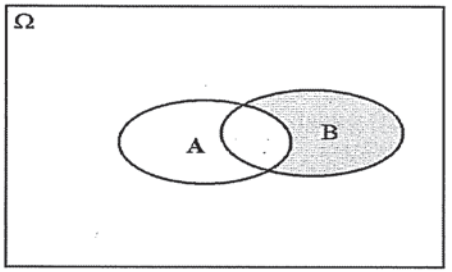


Figura 2.7: Rappresentazione grafica dell'evento differenza di due eventi.

4. Distributiva

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

5. Elemento neutro

$$A \cup \emptyset = A$$
$$A \cap \emptyset = A$$
$$A \cap \Omega = A$$

6. Unione e intersezione di un insieme con il suo complementare

$$A \cup \bar{A} = \Omega$$
$$A \cap \bar{A} = \emptyset$$

2.5.2 La probabilità

Una volta definito lo spazio degli eventi aleatori, si rende necessario introdurre una misura dell'incertezza. Come abbiamo più volte rimarcato, essendo la probabilità una misura al pari di altre misure come il chilogrammo, il metro, ecc. si dovranno stabilire delle regole convenzionali che siano universalmente riconosciute. In tal senso, le

regole per il calcolo della probabilità vengono ricavate attraverso tre diverse impostazioni che, anche se in alcuni casi mostrano delle lacune, sono ormai riconosciute universalmente. Per rendere più intuitivo il concetto di probabilità, possiamo immaginare di poter rappresentare graficamente la probabilità come un linea direzionalmente orientata che parte da un *evento impossibile*, al quale associamo una probabilità nulla di verificarsi $P(x) = 0$, e avanza verso un *evento sicuro*, certo, al quale associamo una probabilità di verificarsi $P(x) = 1$. Associamo il valore 0 all'evento "impossibile" proprio perché è un evento che non potrà mai accadere e usiamo il valore 1 per l'evento "certo" perché, se si verifica, accade uno e un solo evento. Tutti i valori compresi tra 0 e 1 saranno i valori che associamo, di volta in volta, agli eventi che possono verificarsi con una certa probabilità.

Indicheremo, quindi, la probabilità con la lettera P e indicheremo tra parentesi l'evento a cui la probabilità si riferisce. In tal modo, $P(x) = 1$ indicherà che l'evento x è sicuro (certo) e $P(x) = 0$ che l'evento x è impossibile.

Di seguito verrà svolto un breve excursus sulle principali concezioni della probabilità, ossia la *concezione classica*, quella *frequentista* e quella *soggettivista*.

La *concezione classica* interpreta la probabilità come un semplice rapporto. Così, ad esempio, indicando con n il numero dei possibili risultati a priori, con uno dei quali un gioco di sorte deve necessariamente terminare, e con m il numero di quelli favorevoli all'evento prefissato ($m \leqslant n$), la probabilità che questo si verifichi trova il suo significato nel rapporto m/n nell'ipotesi che gli n risultati siano fra loro equipossibili. Così, ad esempio, nel lancio di un dado da gioco, le cui facce sono numerate da uno a sei, la probabilità del verificarsi in un lancio dell'evento E: *"faccia con un numero di punti non inferiore a* 5" è il valore 2/6, pari al rapporto fra il numero dei casi favorevoli (facce con un numero di punti non inferiore a 5, cioè le facce 5 e 6) e il numero dei casi possibili (le sei facce del dado).

Avremo, quindi: dato l'evento E: *"faccia con un numero di punti non inferiore a* 5", la probabilità del verificarsi dell'evento stesso sarà

$$p(E) = \frac{\text{numero dei casi favorevoli}}{\text{numero dei casi possibili}} = \frac{m}{n} = \frac{2}{6} = \frac{1}{3}$$

Occorre precisare che la definizione classica richiede che l'insieme dei casi possibili sia finito in modo da consentirne la loro enumerazione. Inoltre, tutti i casi si assumono equipossibili.

La *concezione frequentista*, la più ampiamente discussa e utilizzata, interpreta la probabilità sulla base dell'esperienza che si acquisisce con l'osservazione della frequenza relativa osservata. Essa è un concetto applicabile ad esperimenti e osservazioni che possono essere replicati un gran numero di volte in condizioni uniformi. Allora, dopo una serie di replicazioni, i possibili risultati sono caratterizzati da frequenze relative che, dopo oscillazioni più o meno marcate, tendono a stabilizzarsi ad una costante p che, evidentemente, non potrebbe che essere un valore dell'intervallo $[0, 1]$. Un tale comportamento della frequenza relativa è il motivo che ha indotto a interpretare la suddetta costante p come la probabilità con cui, da una prova, può scaturire un determinato evento. Procedendo con un esempio, consideriamo il caso in cui un dado sia stato lanciato un elevato numero di volte e che la frequenza con cui si è presentato il valore 3 sia stata di $\frac{1}{6}$. Dirò, allora, che la probabilità dell'evento E: *"esce il 3"* è $p = \frac{1}{6}$. In generale, quindi, nella concezione frequentista si definisce la probabilità come la frequenza con cui un evento si presenta quando l'esperimento viene ripetuto un elevato numero di volte, sempre nelle stesse condizioni.

La *concezione soggettivista*, infine, identifica la probabilità nel grado di fiducia che un individuo ha nel verificarsi di un evento. La pietra angolare di tale nuovo modo di pensare non s'identifica più in una semplice regola matematica, come nel caso dei classici, non è più legata alla possibilità di replicare le esperienze come accade per i frequentisti, ma è data dall'individuo. Poniamoci, ad esempio, la domanda "Qual è la probabilità dell'evento E: *"il Napoli quest'anno vince lo scudetto"*. Per dare una risposta non posso usare né la concezione classica, né quella frequentista, ma posso prendere in considerazione solo quella soggettivista. La concezione soggettivista, quindi, ha nel soggetto il punto di riferimento. Naturalmente il soggettivismo non ammette che le valutazioni siano arbitrarie o irragionevoli: esse, anzi, devono tenere conto di tutte le circostanze note. Per i soggettivisti la locuzione "la probabilità dell'evento E è p" non ha alcun senso avendo senso, viceversa, quest'altra: "è p_A la probabilità che il soggetto A - stante le informazioni in suo possesso, le sue esperienze ed il suo modo di

vedere le cose - assegna all'evento in questione". Cosicché il soggetto B potrà anche attribuire ad E probabilità p_B diversa da p_A.

Tutte le possibili impostazioni della probabilità possono essere riassunte in un'unica teoria della probabilità proposta da Kolmogorov, detta *impostazione assiomatica*, dove la teoria delle probabilità viene formulata sulla base di alcuni assiomi. La formalizzazione matematica consente di ottenere i medesimi risultati indipendentemente dal tipo di approccio probabilistico che il ricercatore predilige. Gli assiomi sui quali si fonda il modello matematico per la probabilità sono di seguito elencati.

Assioma 1 - *Gli eventi dello spazio formano un'algebra di Boole completa.*

Assioma 2 - *La misura di probabilità di un evento è unica.*

Assioma 3 - *Principio della misura: La misura della probabilità di un evento è sempre non negativa.*

$$P(A) \geqslant 0$$

Assioma 4 - *La probabilità dell'evento certo è uguale a 1*

$$P(\Omega) = 1$$

dalla fusione dell'assioma 3 e dell'assioma 4 si desume che la probabilità è una misura compresa tra 0 (evento impossibile) e 1 (evento certo).

$$0 \leqslant P(A) \leqslant 1$$

Assioma 5 - *Principio delle probabilità totali per eventi incompatibili*: siano A e B due eventi incompatibili dello spazio Ω, nel senso che $A \cap B = \emptyset$, allora

$$P(A \cup B) = P(A) + P(B)$$

Il principio delle probabilità totali per eventi incompatibili può essere esteso a più di due eventi. In tal caso, si avrà che la probabilità dell'unione di più eventi incompatibili è uguale alla somma delle

probabilità di ciascun evento. Dagli assiomi ricaviamo alcuni risultati immediati.

- Ad ogni evento A di Ω è associato l'evento negato la cui probabilità è data da:
$$P(\bar{A}) = 1 - P(A)$$
- Sia $\emptyset$ l'evento impossibile allora
$$P(\emptyset) = 0$$
- Principio delle probabilità totali per eventi qualsiasi. Siano A e B due eventi qualsiasi dello spazio Ω, nel senso che $A \cap B \neq \emptyset$, allora
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

2.5.3 La probabilità condizionata

Uno dei concetti più importanti del calcolo delle probabilità è la probabilità condizionata. Parliamo di probabilità condizionata quando siamo interessati alla probabilità di un evento il cui risultato è influenzato dal verificarsi di un altro evento, detto condizionante. Per meglio chiarire questo concetto facciamo un semplice esempio. Supponiamo che i due eventi siano A (Paolo esce di casa); B (giornata piovosa). L'evento $A \mid B$ indica l'evento: "esce Paolo dato che piove". È naturale pensare che una giornata piovosa possa influire sulla voglia di uscire. Quello che bisogna stabilire è quanto una giornata piovosa scoraggia l'uscita di casa di Paolo. È facile intuire che, se fuori casa ad aspettare Paolo c'è la fidanzata, allora l'effetto della pioggia è poco influente sulla sua decisione. Mentre se Paolo vuole uscire per fare una passeggiata, allora la pioggia condiziona fortemente la sua azione. Si capisce che, anche in questo caso, possiamo trovare una misura dell'incertezza che chiameremo appunto probabilità condizionata. Siano A e B due eventi di Ω e $A \mid B$ l'evento condizionato, definiamo probabilità condizionata la quantità

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

sotto la condizione che $P(B) > 0$.

Questa definizione di probabilità può intuitivamente essere letta come rapporto tra eventi favorevoli su eventi possibili. Infatti riferendoci all'esempio di Paolo, il numeratore esprime la probabilità dell'evento favorevole alle volte che Paolo esce nelle giornate piovose, mentre il denominatore esprime la probabilità degli eventi possibili (la probabilità che piova), ossia la probabilità dell'evento che circoscrive la volontà di uscita di Paolo ai giorni di pioggia. Un'immediata conseguenza della probabilità condizionata è il principio delle probabilità composte per eventi dipendenti, la cui espressione formale è immediatamente ricavabile dalla precedente dopo semplici passaggi algebrici

$$P(A \cap B) = P(A \mid B)P(B).$$

Può accadere che Paolo non sia affatto influenzato dall'evento meteorologico in quanto ci sono altri fattori diversi dalla pioggia che possono condizionare la sua uscita di casa. In questo caso, si dice che l'evento A (Paolo esce di casa) è indipendente dall'evento B (giornata piovosa). Da un punto di vista formale, si ha evidentemente che la probabilità condizionata:

$$P(A \mid B) = P(A)$$

ossia la probabilità che Paolo esca dato che piove è circoscritta alla sola probabilità che Paolo esca. Sostituendo opportunamente quest'ultima relazione nel principio delle probabilità composte per eventi dipendenti [che ricordiamo esser $P(A \cap B) = P(A \mid B)P(B)$], si ricava immediatamente l'espressione

$$P(A \cap B) = P(A)P(B)$$

Questa uguaglianza è anche detta *principio delle probabilità composte per eventi indipendenti*. Quindi diremo che due eventi A e B, entrambi appartenenti allo spazio Ω sono *stocasticamente indipendenti* se per essi vale il principio delle probabilità composte per eventi indipendenti.

In generale il concetto di indipendenza a più di due eventi è formalizzabile nel modo che segue: siano $A_1, A_2, \ldots, A_n$ n eventi appartenenti all'insieme Ω, allora detti eventi saranno indipendenti tra loro se è rispettata la seguente condizione:

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = \Pi_{i=1}^{n} P(A_i)$$

dove il simbolo Π indica la produttoria dei fattori $P(A_i)$; essa è equivalente, quindi, a $P(A_1) \cdot P(A_2) \cdot, \ldots, \cdot P(A_n)$.

È opportuno far notare che il concetto di indipendenza non è in nessun modo collegato al concetto di incompatibilità, nel senso che l'incompatibilità non implica l'indipendenza e viceversa. Infatti, c'è incompatibilità se il verificarsi di un evento esclude il verificarsi dell'altro, il che fa intravedere un legame di dipendenza, essendo uno degli eventi fortemente condizionato dal verificarsi dell'altro.

2.5.4 Il Teorema di Bayes

Un'immediata applicazione del concetto della probabilità condizionata risiede nel teorema di Bayes. Per meglio chiarire quanto vogliamo ricavare in questo paragrafo, immaginiamo di avere uno spazio Ω costituito da un numero di n urne indicate con $U_1, U_2, \ldots, U_n$, che definiscono una partizione completa di eventi incompatibili dello spazio Ω, tale che $\Omega = \cup_{i=1}^{n} U_i$; quanto detto implica che l'intersezione di eventi incompatibili sia l'insieme vuoto, ossia $U_i \cap U_j = \emptyset \quad \forall i, j$.

Un evento A di Ω si può esprimere, quindi, come:

$$A = A \cap \Omega = A \cap (U_1 \cup U_2 \cup, \ldots, \cup U_n)$$

da cui, per la proprietà distributiva, sarà

$$(A \cap U_1) \cup (A \cap U_2) \cup \cdots \cup (A \cap U_n)$$

La domanda a cui risponde il teorema di Bayes è la seguente: supponiamo di avere estratto un elemento di A; qual è la probabilità che esso appartenga ad un urna generica U_i?

Procedendo con un esempio, supponiamo di avere tre scatole di cioccolatini tali che: la prima contiene 10 cioccolatini al latte e 10 fondenti; la seconda 5 al latte e 15 fondenti; la terza 20 cioccolatini al latte. Il teorema di Bayes risponde alla seguente domanda: "Sapendo che mi è uscito un cioccolatino fondente, qual è la probabilità che esso provenga dalla prima scatola". Formalmente, U_1, U_2 e U_3 sono le tre scatole di cioccolatini, A= "è uscito un cioccolatino fondente". Da un punto di vista formale, il problema può essere posto nel seguente modo: $P(U_i \mid A)$.

Richiamando il concetto di probabilità condizionata, si ha che:

$$P(U_i \mid A) = \frac{P(U_i \cap A)}{P(A)}$$

Ricordando che $P(U_i \cap A) = P(U_i)P(A \mid U_i)$, si otterrà

$$P(U_i \mid A) = \frac{P(U_i)P(A \mid U_i)}{P(A)}$$

il denominatore può, invece, essere espresso nel seguente modo:

$$P(A) = P(A \cap U_1) + P(A \cap U_2) + \cdots + P(A \cap U_n)$$

in quanto gli eventi U_i sono a due a due incompatibili e di conseguenza lo sono anche gli eventi $(A \cap U_i)$.

In conclusione si ha che:

$$P(A) = \sum_{i=1}^{n} P(U_i)P(A \mid U_i)$$

dove il simbolo $\sum$ indica la somma dei termini $P(U_i)P(A \mid U_i)$, cioè $P(U_1)P(A \mid U_1) + \cdots + P(U_n)P(A \mid U_n)$. Da quest'ultima equazione, con le opportune sostituzioni, si ottiene la nota formula di Bayes:

$$P(U_i \mid A) = \frac{P(U_i)P(A \mid U_i)}{\sum_{i=1}^{n} P(U_i)P(A \mid U_i)}$$

Nel linguaggio ormai acquisito, la $P(U_i \mid A)$ è chiamata *probabilità a posteriori*, mentre $P(U_i)$ *probabilità a priori* e, infine, $P(A \mid U_i)$ è chiamata *probabilità probativa* o anche *verosimiglianza*.

Per esplicare meglio quanto detto facciamo un esempio:

Supponiamo di avere quattro urne la cui relativa probabilità di scelta è: $P(U_1) = 0.2$, $P(U_2) = 0.4$, $P(U_3) = 0.25$ e $P(U_4) = 0.15$.

Sia A l'evento pallina arancione contenuta nelle urne, la probabilità di estrarre una pallina arancione è rispettivamente data dalle seguenti probabilità: $P(A \mid U_1) = 0.2$, $P(A \mid U_2) = 0.6$, $P(A \mid U_3) = 0.3$ e $P(A \mid U_4) = 0.9$.

Se il nostro esperimento consiste nell'estrazione di una pallina da una delle quattro urne, avremo che la probabilità di estrarre una pallina arancione è pari a:

$$P(A) = \sum_{i=1}^{4} P(U_i)P(A \mid U_i) =$$

$$= (0.2 \cdot 0.2) + (0.4 \cdot 0.6) + (0.25 \cdot 0.3) + (0.15 \cdot 0.9) = 0.49$$

Qual è la probabilità che, essendo stata estratta una pallina arancione, la stessa provenga dall'urna U_4, ovvero $P(U_4 \mid A)$?

$$P(U_4 \mid A) = \frac{0.15 \cdot 0.9}{0.49} = 0.27$$

Quindi, sapere che è stata estratta una pallina arancione modifica la probabilità a priori di scelta dell'urna U_4, $P(U_4) = 0.15$, in quella a posteriori $P(U_4 \mid A) = 0.27$. Prima di effettuare l'estrazione, la probabilità di scegliere l'urna era pari alla probabilità a priori. Grazie al teorema di Bayes è possibile vedere come questa probabilità si modifichi quando si ha l'informazione addizionale che la pallina estratta è arancione.

2.5.5 Le variabili casuali

Da quanto detto nei paragrafi precedenti, il concetto di probabilità può essere schematizzato nella terna $(\Omega, B, P(A))$, dove con $\Omega = (\omega_1, \omega_2, ...)$ abbiamo indicato lo spazio degli eventi, con B, l'*algebra di Boole*, che contiene il complesso delle operazioni eseguibili sugli eventi appartenenti a Ω, mentre con $P(A)$, la misura dell'evento aleatorio esprimibile con una delle impostazioni della probabilità precedentemente definite. Nelle applicazioni su casi reali, può risultare complesso definire gli elementi appartenenti alla terna $(\Omega, B, P(A))$. Inoltre, molti fenomeni sono esprimibili con modalità quantitative o sono riconducibili a modalità quantitative e, in questo contesto, è utile introdurre il concetto di variabile casuale. La variabile casuale, v.c., è una variabile quantitativa che associa ad ogni elemento ω_i dello spazio Ω un numero reale. Il nuovo spazio degli eventi, così definito,

lo indichiamo con Ω_R. Più formalmente si dirà che una v.c. $X(\omega)$ è un'applicazione di Ω in Ω_R, cioè

$$X(\omega) : \Omega \to \Omega_R.$$

Per meglio capire quanto appena detto, supponiamo che il nostro esperimento casuale sia il lancio di un dado. Allora l'insieme degli eventi nello spazio Ω è dato da:
ω_1=faccia del dato con un puntino
ω_2=faccia del dato con due puntini
ω_3=faccia del dato con tre puntini
ω_4=faccia del dato con quattro puntini
ω_5=faccia del dato con cinque puntini
ω_6=faccia del dato con sei puntini.

Una variabile casuale associata allo spazio appena definito può essere:
$X(\omega)$=*numero reale associato a ciascuna faccia del dado*, ovvero:
$X(\omega_1) = 1$
$X(\omega_2) = 2$
$X(\omega_3) = 3$
$X(\omega_4) = 4$
$X(\omega_5) = 5$
$X(\omega_6) = 6$
Di conseguenza il nuovo spazio $\Omega_R = 1, 2, 3, 4, 5, 6$

In generale, da un punto di vista grafico, una v.c. può essere schematizzata come segue: si possono costruire diversi tipi di variabile casuale anche riferite allo stesso esperimento. Per esempio, nel lancio del dado di cui sopra un'altra v.c. potrebbe essere $X(\omega) = 1$ se $\omega_1 \geqslant 3$, $X(\omega) = 0$ altrimenti. In altri termini, una v.c. è una congettura formale che è in grado di descrivere lo spazio degli eventi di un esperimento casuale in un nuovo spazio di numeri reali. I vantaggi che scaturiscono dall'utilizzo delle variabili casuali sono molteplici, non ultima la possibilità di introdurre un'algebra applicata ai numeri reali che indichiamo con B_R,la quale risulta essere più completa di quella associata all'algebra B potendo far ricorso agli strumenti di analisi matematica.

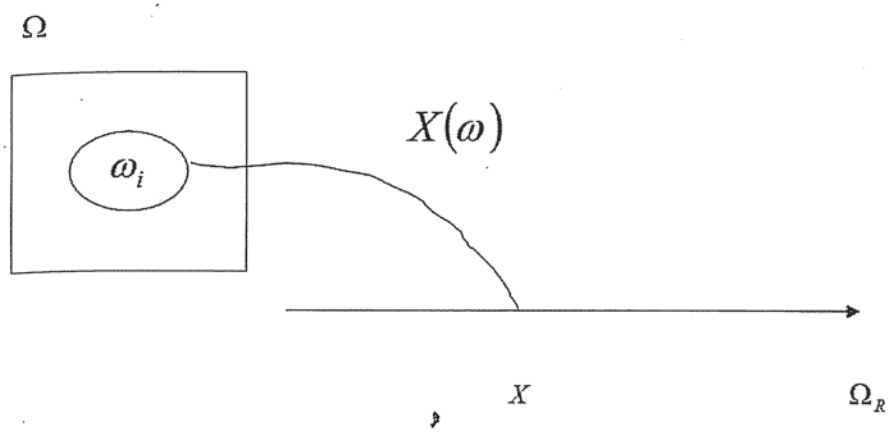


Figura 2.8: Schematizzazione grafica di una variabile casuale.

Le variabili casuali sono classificabili in due grosse categorie:

- *v.c. discrete*, cioè costituite da un numero finito o infinità numerabile di valori;
- *v.c. continue*, ovvero costituite da un numero infinito di valori compresi in un intervallo di ampiezza finita o infinita.

Le variabili casuali discrete

Da quanto detto sopra, i valori di una v.c. sono numeri reali legati al risultato di un esperimento aleatorio; di conseguenza, una variabile casuale X è sempre accompagnata dalla sua *funzione di probabilità* $P(x)$. Questa corrispondenza tra valori di X e probabilità definisce la *distribuzione di probabilità di* X.

Nel caso di una v.c. discreta, la corrispondente funzione di probabilità è ancora discreta ed è data da:

$$P(x) = P(X = x_i) = p_i \quad \forall i$$

Graficamente, una v.c. discreta può essere rappresentata come in figura 2.9:

La funzione di probabilità $P(x)$ appena definita soddisfa le seguenti condizioni:

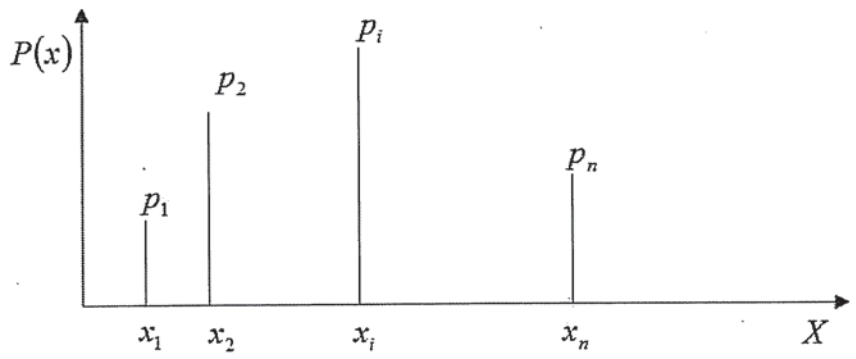


Figura 2.9: Rappresentazione grafica di una variabile casuale discreta.

- $P(x) \geqslant 0 \ \forall x$;
- $\sum_{\forall x} P(x) = 1$.

In taluni casi può essere utile trovare la probabilità che la v.c. X assuma un valore inferiore o uguale ad un valore dato; tale probabilità viene descritta dalla *funzione di ripartizione* indicata con:

$$F(x_i) = P(X \leqslant x_i) = \sum_{x \leqslant x_i} P(x_i).$$

La funzione di ripartizione è caratterizzata dalle seguenti proprietà:

- $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$ e $F(+\infty) = \lim_{x \to +\infty} F(x) = 1$
- $F(x)$ è non decrescente (è una probabilità cumulata), ossia per ogni coppia di numeri a e b tali che $a < b$ si ha $F(a) < F(b)$;
- $F(x)$ è continua a destra ossia $\lim_{h \to 0^+} F(x+h) = F(x)$;
- $P(a < X \leqslant b) F(b) - F(a)$.

Esempio: pensiamo all'esperimento di due lanci di una moneta regolare, cioè $P(T) = P(C) = \frac{1}{2}$. L'insieme dei risultati possibili, cioè lo spazio campionario dell'esperimento è $\Omega = TT, TC, CT, CC$.

$\omega \in \Omega$	TT	TC	CT	CC
$X(\omega) = x$	2	1	1	0
$P(\omega)$	1/4	1/4	1/4	1/4

Tabella 2.3: Spazio campionario dell'esperimento "Due lanci di una moneta".

Se introduciamo la v.c. $X(\omega)$= *numero di teste ottenute ad ogni lancio*, possiamo costruire la tabella 2.3.

Come si vede, X rappresenta il risultato numerico di un esperimento. Possiamo sintetizzare la tabella sopra raccogliendo i valori distinti assunti da X (0,1,2) con le rispettive probabilità. X assume valore 0 solo in corrispondenza dell'evento CC, il quale ha probabilità $\frac{1}{4}$; quindi $P(X = 0) = P(CC) = \frac{1}{4}$. Lo stesso avviene per il valore 2, infatti, $X = 2$ solo quando si verifica l'evento TT, quindi $P(X = 2) = P(TT) = \frac{1}{4}$. Il valore $X = 1$ si realizza, invece, in due casi: quando si realizza TC o quando si realizza CT, dunque $P(X = 1) = P(CT \cup TC) = P(CT) + P(TC) = 1/4 + 1/4 = 1/2$.

Riassumendo, abbiamo la seguente distribuzione di probabilità per la v.c. discreta X (vd. Tab. 2.4):

$X = x$	0	1	2
$P(X = x)$	1/4	1/2	1/4
$F(x)$	1/4	3/4	1

Tabella 2.4: Distribuzione di probabilità.

Le variabili casuali continue

In molti problemi empirici, conviene pensare ad una variabile casuale come la realizzazione di un processo continuo. La lunghezza ed il trascorrere del tempo sono esempi tipici di grandezze che possono, teoricamente, assumere ogni possibile valore in un intervallo finito o infinito. Nella pratica poi, si opera necessariamente una discretizzazio-

ne della variabile che, invece, per sua natura è continua. Ad esempio, il tempo si approssima ai minuti secondi o ai decimi di secondo, la lunghezza ai millimetri o ai decimillimetri e così di seguito. Esistono quindi, nella realtà, molti fenomeni che vengono descritti da un numero finito di possibilità, in quanto i sistemi di misura che si utilizzano, per quanto precisi, richiedono un'approssimazione all'unità di misura utilizzata. Per tale motivo, considerando le v.c. continue, si passerà al concetto di area ovvero ad assegnare le probabilità a singoli intervalli piuttosto che a singoli punti (in quanto i singoli punti che compongono un intervallo sono infiniti).

Una definizione di v.c. continua è la seguente:
Una variabile casuale X *è continua se esiste una funzione* $f(x)$ *tale che*:

$$f(x) = P(x \leqslant X \leqslant x + dx) = \int_x^{x+dx} f(t)dt \quad \forall x$$

La funzione f viene chiamata *funzione di densità di probabilità* (f.d.p.) di X. In questo caso, tuttavia, la funzione $f(x)$ non può essere interpretata come la $P(X = x)$, in quanto tale probabilità sarà sempre nulla, per le v.c. di tipo continuo, infatti l'integrale calcolato nel punto è per definizione sempre uguale a zero.

Si può determinare, invece, la probabilità di osservare un valore compreso nell'intervallo (a, b), cioè $P(a \leqslant X \leqslant b)$ attraverso l'integrale

$$P(a \leqslant X \leqslant b) = \int_a^b f(x)dx.$$

Possiamo, infine, definire la *funzione di ripartizione* di una v.c. continua come:

$$F(x) = P(X \leqslant x) = \int_{-\infty}^{x} f(x)dx \quad \text{per } -\infty < x < +\infty.$$

La funzione di densità $f(x)$ gode delle seguenti proprietà:

- $f(x) \geqslant 0; \ -\infty < x < +\infty$
- $\int_{-\infty}^{+\infty} f(x)dx = 1$

Rappresentando graficamente una v.c. continua, si vede immediatamente che la probabilità associata all'intervallo $[a, b]$, con $a < b$, è data dall'area compresa tra l'asse orizzontale e la funzione $f(x)$.

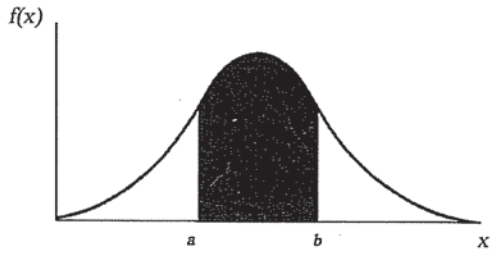


Figura 2.10: Rappresentazione di una v.c. continua.

Variabili casuali bi-dimensionali

Ogni qualvolta andiamo a considerare congiuntamente due o più fenomeni quantitativi, la rappresentazione degli eventi avviene in termini di due o più variabili casuali. Consideriamo dapprima il caso più semplice di due soli fenomeni. In tal caso, la coppia di fenomeni trova la sua interpretazione nella funzione reale misurabile (X, Y), definita a partire dallo spazio probabilistico $(\Omega, B, P(A))$ Tale funzione, per ogni ω in Ω, identifica nel piano reale $\mathbb{R}^2$ un punto di coordinate (x, y).

Una coppia di variabili casuali semplici definisce una variabile casuale doppia. Tale v.c. è discreta se X e Y sono discrete, continua se X e Y sono continue, mista se una v.c. è di tipo discreto e l'altra è di tipo continuo. La distribuzione di probabilità che ne consegue sarà una distribuzione di probabilità bivariata.

Si chiama funzione di probabilità congiunta delle due variabili discrete X e Y la funzione

$$p(x, y) = P(X = x, Y = y).$$

Le condizioni che essa soddisfa sono:

- $p(x, y) \geqslant 0$
- $\sum_x \sum_y p(x, y) = 1 \ \forall x, y.$

Conseguentemente, possiamo definire la funzione di ripartizione congiunta delle due v.c. discrete X e Y come:

$$F(x_i, y_j) = P(X \leqslant x_i, Y \leqslant y_j) = \sum_{x \leqslant x_i} \sum_{y \leqslant y_j} p(x, y)$$

Distribuzioni marginali e distribuzioni condizionali

Dalla distribuzione di probabilità della v.c. doppia (X,Y) è possibile ottenere le distribuzioni di probabilità di ogni v.c. semplice che compone la coppia (X,Y), ovvero le distribuzioni marginali di X rispetto a Y e di Y rispetto a X. La *funzione di probabilità marginale* di X è data da:

$$p(x) = \sum_{\forall y} p(x, y),$$

ossia la probabilità che $X = x$, quando Y assume un valore qualsiasi. In modo analogo si definisce la funzione di probabilità marginale della variabile Y:

$$p(y) = \sum_{\forall x} p(x, y) = P(Y = y).$$

Le funzioni di probabilità marginali servono, inoltre, per definire le distribuzioni di *probabilità condizionali*. Nel caso di v.c. discrete, la distribuzione di probabilità della v.c. X, condizionata dalla realizzazione $Y = y$, è definita come:

$$P(x|y) = P(x|Y = y) = \frac{P(x, y)}{P(x)}$$

che esprime la probabilità di x subordinata all'evento $Y = y$.

Facciamo il seguente esempio.

Supponiamo che il nostro esperimento consista nel lancio di una moneta e di un dato. Il nostro spazio degli eventi Ω sarà definito dai singoli eventi:

ω_1 =Testa, numero 1	ω_2 =Testa, numero 2
ω_3 =Testa, numero 3	ω_4 =Testa, numero 4
ω_5 =Testa, numero 5	ω_6 =Testa, numero 6

ω_7 =Croce, numero 1	ω_8 =Croce, numero 2
ω_9 =Croce, numero 3	ω_{10} =Croce, numero 4
ω_{11} =Croce,numero 5	ω_{12} =Croce, numero 6

La distribuzione di probabilità congiunta sarà la seguente 2.5.

X, Y	P(x, y)
T,1	1/12
T,2	1/12
T,3	1/12
T,4	1/12
T,5	1/12
T,6	1/12
C,1	1/12
C,2	1/12
C,3	1/12
C,4	1/12
C,5	1/12
C,6	1/12

Tabella 2.5: Distribuzione di probabilità congiunta.

Se volessimo determinare la funzione di probabilità condizionata di X, associata a $Y = 5$, avremo che:

$$p(T|y = 5) = \frac{p(T, 5)}{p(5)} = \frac{1/12}{1/6} = 0.5$$

$$p(C|y = 5) = \frac{p(C, 5)}{p(5)} = \frac{1/12}{1/6} = 0.5$$

2.5.6 Momenti delle variabili casuali

Abbiamo visto che a ciascuna v.c. è associata una distribuzione di probabilità. Si rende, quindi, necessario disporre di uno strumento operativo per lo studio di dette distribuzioni di probabilità.

A tal proposito introduciamo il calcolo dei *momenti*. Questi ultimi hanno la capacità di caratterizzare ogni singola distribuzione di

probabilità negli aspetti più essenziali. In particolare, i momenti ci permettono di definire dei parametri o delle grandezze caratteristiche di una distribuzione di probabilità, i quali hanno la capacità di riassumere in modo immediato e sintetico l'informazione relativa alla distribuzione.

In generale, si distinguono i momenti dall'origine, dai momenti centrati rispetto ad un'origine arbitraria.

Per una v.c. discreta, si definisce *momento di ordine r* dall'origine la seguente espressione:

$$E(X^r) = \sum_{\forall x} x^r p(x) = \mu_r.$$

In particolare, per $r = 1$, si ha la speranza matematica o valore atteso della v.c. X dove, in pratica, $E(X)$ è la media delle diverse realizzazioni di X ponderate con le probabilità corrispondenti e si indica con μ.
Per $r = 2$ avremo $\mu_2 = \sum_{\forall x} x^2 p(x)$ che è la media quadratica al quadrato.

Mentre, per un v.c. discreta, si definisce *momento centrato di ordine r un'origine arbitraria a* l'espressione:

$$E[(X-a)^r] = \sum_{\forall x} (x-a)^r p(x) = \bar{\mu}_r(a)$$

Quale origine arbitraria, in genere, viene assunto il momento primo dall'origine, ossia la media μ o anche la mediana o la moda. A seconda del valore che daremo all'origine arbitraria a, avremo diverse misure caratteristiche della distribuzione:

- per $r = 1, a = \mu$, si ha $\bar{\mu} = \sum_{\forall x}(x-\mu)p(x) = 0$, ovvero la somma degli scarti dalla media è pari a zero, che, come vedremo in seguito, rappresenta una nota proprietà della media aritmetica;
- per $r = 2, a = \mu$, si ha $\bar{\mu}_2 = \sum_{\forall x}(x-\mu)^2 p(x) = \sigma^2$, che, come approfondiremo, rappresenta un indice di variabilità nota come varianza;
- per $r = 3, a = \mu$, si ha $\bar{\mu}_3 = \sum_{\forall x}(x-\mu)^3 p(x)$;

- per $r = 4, a = \mu$, si ha $\bar{\mu}_4 = \sum_{\forall x}(x-\mu)^4 p(x)$.

In particolare, standardizzando la v.c. X attraverso la trasformazione lineare $Z = \frac{X-\mu}{\sigma}$ e calcolando il momento terzo e il momento quarto della nuova v.c. Z, ossia

$$\delta_1 = E(Z^3) = \sum_{\forall x} \left(\frac{x-\mu}{\sigma}\right)^3 p(x) = \frac{\bar{\mu}_3}{\sigma^3},$$

$$\delta_2 = E(Z^4) = \sum_{\forall x} \left(\frac{x-\mu}{\sigma}\right)^4 p(x) = \frac{\bar{\mu}_4}{\sigma^4},$$

si hanno rispettivamente un indice di asimmetria e di curtosi della distribuzione di probabilità della v.c. discreta X.

In particolare, l'*asimmetria* è la valutazione del comportamento che ha una distribuzione di probabilità rispetto ad un punto di simmetria che spesso è identificato con μ.

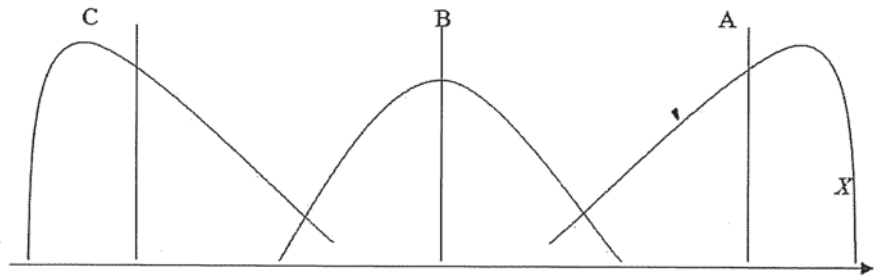


Figura 2.11: Asimmetria.

- se $\delta_1 > 0$ la curva ha *asimmetria positiva* (la curva C è spostata verso destra);
- se $\delta_1 = 0$ la curva è *simmetrica* (curva B);
- se $\delta_1 < 0$ la curva ha *asimmetria negativa* (curva A spostata verso sinistra), cioè ha la coda di sinistra allungata.

La *curtosi*, invece, è la valutazione del comportamento della distribuzione di probabilità alle code. Infatti, essendo δ_2 funzione del momento quarto, si avrà un alto valore della curtosi quanto più i valori nelle code, cioè distanti dalla media, hanno una probabilità di verificarsi elevata. Nella figura abbiamo tre diverse distribuzioni con probabilità che si verifichino valori estremi alti (curva C), intermedi (curva B) e bassi (curva A).

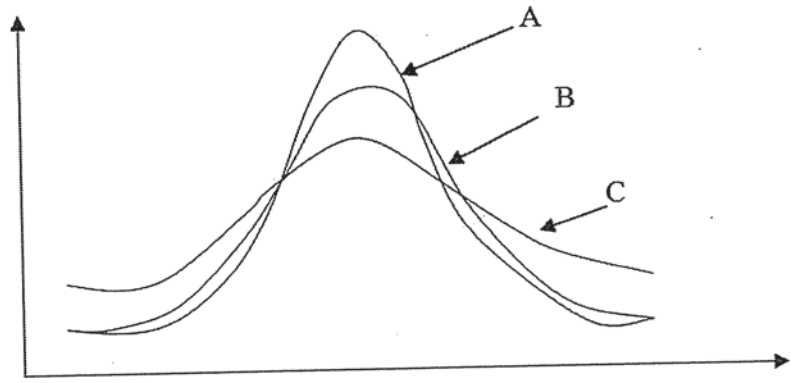


Figura 2.12: Curtosi.

Se la v.c. è continua, allora i momenti di ordine r dall'origine sono espressi dalla relazione:

$$\mu_r = \int_{\forall x} x f(x) dx;$$

mentre quelli da una origine arbitraria a da:

$$\bar{\mu}_r = \int_{\forall x} (x-a)^r f(x) dx;$$

Da quanto detto sinora, si intuisce l'importanza del calcolo dei momenti di una distribuzione di probabilità. É opportuno, quindi, acquisire un'adeguata abilità nell'operare con i momenti. In particolare, è utile ricavare una relazione che lega i momenti da un'origine arbitraria dai momenti dall'origine. Infatti, ricordando che $\bar{\mu}_r = E(x-\mu)^r$

altro non è che il valore atteso di un binomio di ordine r, ricorrendo allo sviluppo in serie di detto binomio, si ottiene immediatamente la seguente relazione:

$$\bar{\mu}_r = \sum_{b=0}^{r} (-1) \binom{r}{b} \mu^b \mu_{r-b},$$

dalla quale è immediato ricavare la relazione: $\sigma^2 = \mu_2 - \mu^2$.

2.5.7 Funzione generatrice dei momenti

Dai paragrafi precedenti, appare evidente l'importanza del calcolo dei momenti di una generica distribuzione di probabilità. In molti casi, il calcolo dei momenti risulta essere immediato, in special modo quando si dispone di una distribuzione empirica di probabilità. In altri casi, come vedremo più avanti, quando si ha una distribuzione di probabilità teorica, il calcolo dei momenti potrebbe essere difficoltoso dovendo ricorrere a strumenti sofisticati di analisi matematica. Per dare risoluzione a questo tipo di problemi, si introduce la *funzione generatrice dei momenti (f.g.m.)*.

Essa, appunto, ha il compito di ricavare i momenti di una distribuzione di probabilità, sia nel caso in cui la variabile casuale è discreta, sia quando la v.c. è continua. In particolare, definiamo funzione generatrice dei momenti di una v.c. X la funzione della variabile reale t, data da $M_x(t) = E(e^{tx})$.

Per come è stata definita, la funzione generatrice dei momenti esiste per ogni valore reale di t. Detta funzione prende il nome di funzione generatrice dei momenti in quanto permette di calcolare tutti i momenti dall'origine. In particolare, la funzione generatrice dei momenti per una v.c. discreta è definita come:

$$M_x(t) = E(e^{tx}) = \sum_{\forall t} e^{tx} p(x),$$

mentre per una v.c. continua, la funzione generatrice dei momenti è definita come:

$$M_x(t) = E(e^{tx}) = \int_{\forall x} e^{tx} f(x) dx.$$

Approssimando la funzione $Y = e^{tx}$ attraverso lo sviluppo in serie di Mc Laurin, cioè:

$$Y = e^{tx} = 1 + tX + \frac{t^2}{2!}X^2 + \frac{t^3}{3!}X^3 + \ldots + \frac{t^r}{r!}X^r + \ldots$$

è possibile ricavare i momenti di ogni ordine della v.c. X attraverso il calcolo del suo valore atteso ossia:

$$E(e^{tx}) = E\left(1 + tX + \frac{t^2}{2!}X^2 + \frac{t^3}{3!}X^3 + \ldots + \frac{t^r}{r!}X^r + \ldots\right),$$

che risulta essere uguale a:

$$E(e^{tx}) = 1 + tE(X) + \frac{t^2}{2!}E(X^2) + \frac{t^3}{3!}E(X^3) + \ldots + \frac{t^r}{r!}E(X^r) + \ldots$$

Infine si ha che:

$$E(e^{tx}) = 1 + t\mu + \frac{t^2}{2!}\mu_2 + \frac{t^3}{3!}\mu_3 + \ldots + \frac{t^r}{r!}\mu_r + \ldots$$

Per isolare ogni momento, si ricorre alla derivata di ordine r della funzione generatrice dei momenti, calcolata nel punto $t = 0$. Di conseguenza, il momento di ordine r di una variabile casuale sarà calcolabile attraverso la seguente espressione:

$$\mu_r = \frac{d^r}{dt^r}M_x(t)|_{t=0} = M_x^r(0) \qquad (2.5.7.1)$$

In particolare,

$$\frac{d}{dt}M_x(t) = \mu + \frac{2t}{2!}\mu_2 + \frac{3t^2}{3!}\mu_3 + \cdots + \frac{rt^{r-1}}{r!}\mu_r \cdots =$$

$$= \mu + t\mu_2 + \frac{1}{2}t^2\mu_3 + \ldots$$

Ponendo $t = 0$ si ottiene il momento primo.
Allo stesso modo, procedendo alla derivata seconda, si ha

$$\frac{d^2}{dt^2}M_x(t) = \mu_2 + \frac{2}{2}t\mu_3 + \cdots + \frac{(r-1)t^{r-2}}{(r-1)!}\mu_r \cdots =$$

$$= \mu_2 + t\mu_3 \ldots$$

ponendo $t = 0$ si ottiene il momento secondo.

In seguito riprenderemo l'espressione (2.5.7.1) appena data per ricavare i momenti delle distribuzioni di probabilità che, di volta in volta, introdurremo per descrivere fenomeni reali i cui eventi sono aleatori.

Anche per variabili casuali bi-dimensionali è possibile calcolare i momenti. In particolare, definiamo il momento centrato di ordine $r+s$ della distribuzione di probabilità congiunta X e Y, il valore atteso di X^rY^s:

$$\mu_{rs} = E(X^rY^s).$$

Si definirà, invece, momento centrato di ordine $r+s$ rispetto alle medie della X e di Y, il valore atteso di $(X-\mu_x)^r(Y-\mu_y)^s$, ossia:

$$\bar{\mu}_{rs} = E\{(X-\mu_x)^r(Y-\mu_y)^s\}.$$

Si fa notare che, per diversi valori di r e di s, si hanno differenti valori dei momenti della variabile casuale doppia. In particolare, per $r = 1$ e $s = 0$, ricaviamo il momento primo della variabile casuale X; mentre, per $r = 0$ e $s = 1$, ricaviamo il momento primo della variabile casuale Y. Nel caso dei momenti centrati, invece, ricaviamo le seguenti importanti relazioni:

- per $r = 2$ e $s = 0$: si ricava $Var(X)$;
- per $r = 0$ e $s = 2$: si ricava $Var(Y)$;
- per $r = 1$ e $s = 1$: si ricava $Cov(X, Y)$.

Estremamente utile risulterà, in seguito, il calcolo dei momenti di particolari combinazioni lineari di variabili casuali. A questo fine, riportiamo i seguenti principali risultati.

- **Valore atteso della combinazione lineare di variabili casuali:** siano $X_1, X_2, \ldots, X_n$, n variabili casuali e $c_1, c_2, \ldots, c_n$, n costanti reali, si definisce $Y = \sum_{i=1}^{n} c_iX_i$ combinazione lineare delle variabili casuali $X_1, X_2, \ldots, X_n$. Si ha che:

$$E(Y) = E\left(\sum_{i=1}^{n} c_iX_i\right) = \sum_{i=1}^{n} c_iE(X_i).$$

Un caso particolare, conseguente a quest'ultimo risultato, è la media aritmetica di n variabili casuali, ottenuta dalla precedente ponendo i pesi $c_i = \frac{1}{n}$ $\forall i$.

- **Varianza di una combinazione lineare di variabili casuali:** Sotto le stesse condizioni del precedente punto si ha che, la varianza della variabile casuale Y, combinazione lineare di n variabili casuali $X_1, X_2, \ldots, X_n$, è uguale a:

$$Var(Y) = Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 Var(X_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j Cov(X_i Y_j).$$

Un particolare risultato utile in seguito, si ha quando le n variabili casuali sono tutte indipendenti tra di loro. In tal caso, si ha che: $Cov(X_i Y_j) = 0 \quad \forall i, j$ e, di conseguenza, $Var(Y) = \sum_{i=1}^{n} c_i^2 Var(X_i)$.

2.6 Le scale di misura

Una misura è tale se è il risultato del confronto di un dato osservato con una posizione identificabile su una scala. L'operazione preliminare alla misurazione vera e propria, pertanto, è l'esplicitazione della scala di cui si serve chi valuta.

La teoria della misurazione consente di distinguere tra quattro tipi di scale diverse, che sono classificabili, dalla più semplice alla più complessa:

- la *scala nominale*;
- la *scala ordinale*;
- la *scala ad intervalli*;
- la *scala di rapporti*.

Quest'ultima scala che, partendo da un punto di valore zero, ci consente di determinare se una quantità posta su essa è multipla di un'altra, tuttavia non è utilizzabile, come vedremo meglio di seguito, in campo educativo per misurare il raggiungimento di obiettivi cognitivi,

sia perchè risulta molto complesso determinare la totale assenza di conoscenza, sia perchè al comportamento umano, caratterizzato da una varietà di differenze individuali, è pressochè impossibile attribuire una misura assoluta alla quale la letteratura in materia riconosca concordemente un significato quantitativamente univoco.

La *scala nominale* non è una vera e propria scala ma rappresenta semplicemente un modo di *categorizzare*; è chiamata così proprio perchè si costruisce attribuendo nomi a determinate qualità che vengono assunte come criterio di riferimento. Essa consente di operare classificazioni sulla base della presenza o dell'assenza della qualità considerata, senza dar luogo a graduazioni che stabiliscano la misura maggiore o minore della qualità stessa. Questo tipo di scala è utile per individuare chi possieda una certa abilità e chi no. Un esempio può essere la classificazione in alfabeti/analfabeti, possesso/non possesso di un diploma, presenza/assenza di un elenco di caratteristiche in relazione ad un tema specifico (es. malattie avute nell'infanzia: morbillo, rosolia, parotite, scarlattina, ecc.). Questo tipo di scala, quindi, non consente la registrazione di possibili gradazioni di una qualità ma stabilisce solo la presenza o l'assenza della qualità stessa. Una particolare scala nominale è quella che si definisce *dicotomica* poichè ad essa si possono attribuire due soli valori: "1" per indicare che il soggetto ha quella modalità di un carattere; "0" per indicare che il soggetto non ha quella modalità del carattere. Quando, invece, la scala nominale si esprime per mezzo di più modalità, possiamo definirla *politomica* o *multinomiale*; in questo caso alle diverse modalità si possono attribuire più valori al fine di categorizzare i soggetti in relazione ad una caratteristica (es: 0, 1, 2, 3, ecc.). Tuttavia, l'attribuzione di un numero, per rappresentare le modalità con cui si esprime un carattere su scala nominale, non ha alcun significato quantitativo in quanto ha il solo scopo di identificare la categoria alla quale appartiene un individuo. La scala nominale permette, quindi, di registrare i *caratteri qualitativi sconnessi*, cioè non ordinabili, quali ad esempio il sesso, il credo religioso e così via. In valutazioni di questa natura, quindi, manca l'indicazione della gradualità, cioè delle differenze nel possesso di una caratteristica tra soggetti diversi o rispetto ad un livello considerato standard di riferimento.

In una scala nominale, come vedremo meglio in seguito quando

analizzeremo nel dettaglio i diversi tipi di carattere, è possibile solamente registrare le *frequenze assolute*, ossia la numerosità dei casi che rientrano in ciascuna modalità con cui è stato classificato il carattere relativo al fenomeno reale che si vuole studiare. Quando questo tipo di scale viene utilizzato per una valutazione può, inoltre, essere utile calcolare anche in che proporzione, *frequenza relativa*, o in che *percentuale* gli individui valutati rientrano in una categoria piuttosto che in un'altra. Le sole proprietà formali di cui gode la scala nominale sono quelle dell'*equivalenza* e della *non equivalenza* tra gli individui della popolazione oggetto di indagine, in relazione alle categorie di una variabile presa in considerazione.

La *scala ordinale consente*, rispetto a quella nominale, di stabilire delle graduatorie d'ordine, cioè delle relazioni di maggioranza e di minoranza che si riferiscono ad una determinata qualità osservata. Le diverse modalità di un carattere qualitativo prese in esame da un'indagine statistica, se distribuite su una scala ordinale, non staranno tutte su uno stesso piano, ma saranno ordinate gerarchicamente in relazione ad un valore che viene arbitrariamente attribuito rispetto alla proprietà considerata. Su una scala ordinale possono essere rappresentati caratteri di tipo qualitativo e di tipo quantitativo. Nel primo caso le modalità di un carattere saranno espresse per mezzo di attributi, aggettivi che descrivono una qualità logicamente ordinabile; nel secondo caso le modalità con cui misuriamo un fenomeno saranno espresse con un numero reale. Un esempio in campo valutativo può essere quello di ordinare gli studenti di una classe in relazione al voto ottenuto in una disciplina, sia esso espresso con aggettivi (es: insufficiente, sufficiente, buono, ecc.) o con quantità numeriche (0, 1, 2, 3, 4, 5, 6, ecc.).

Su una scala di tipo ordinale, il fatto che le categorie siano ordinate non ci assicura, però, che le distanze tra esse siano uguali. Questo è un dato di fatto di non poca rilevanza in campo educativo poichè i sistemi di valutazione utilizzati, indipendentemente se siano voti numerici o giudizi sintetici (non sufficiente, sufficiente, buono, distinto, ottimo), ci danno indicazioni su un ordinamento del livello di apprendimento ma non garantiscono scientificamente quale sia la distanza tra una modalità e l'altra o, ancora, quale valore attribuisca ad ogni modalità il singolo docente. In parole povere non è possibile determinare nè

se tutti i docenti attribuiscano ad un voto lo stesso valore, nè se la distanza tra i singoli voti sia costante.

A livello di scala ordinale, infatti, possiamo applicare la proprietà dell'ordinamento tra le categorie, ovvero possiamo dire che, rispetto alla caratteristica misurata, una persona che in graduatoria ha una posizione x, ha un valore più elevato rispetto ad una persona in posizione $x-1$, e che quest'ultima ha un valore più elevato rispetto ad una persona in posizione $x-2$. Inoltre se $x > x-1$ e $x-1 > x-2$, se ne deduce che $x > x-2$, nota come proprietà transitiva.

In questa scala, però, non siamo in grado di quantificare la distanza tra il valore x e il valore $x-1$, e non siamo in grado di dire se tra x e $x-1$ da un lato, ed $x-1$ e $x-2$ dall'altro vi sia la stessa distanza.

La scala rappresentata dai voti scolastici si può definire molto particolare poichè l'ordine che gli studenti assumono nella valutazione degli apprendimenti non è costante ma varia in relazione a diversi fattori: al peso che chi valuta dà al voto; al tipo di prova; a fenomeni che possono influenzare il giudizio; al contesto socio-culturale; ecc. Usualmente l'insegnante attribuisce un voto più alto alla prestazione migliore e un voto basso ad una prestazione scarsa, ma non tutti gli insegnanti utilizzano lo stesso range di voti, che solo teoricamente va da 0 a 10; inoltre occorre chiedersi se tra 8 e 9, ad esempio, esiste lo stesso scarto che c'è tra il voto 5 e il 6 o fra un 1 e un 2. Quando siamo in grado di dare questa informazione, infatti, siamo in presenza di un livello di misurazione che può essere garantito dalla scala ad intervalli.

È proprio l'impossibilità della scala ordinale di definire intervalli, che rende assolutamente erroneo il suo utilizzo nella valutazione scolastica per effettuare la media dei voti conseguiti negli apprendimenti. Quest'uso distorto dell'applicazione di una misura statistica alla valutazione è invece largamente diffuso in ambito scolastico, e ancor più grave è il fatto che venga utilizzato per avere una misurazione di caratteri eterogenei, come i risultati conseguiti in una prova scritta e in una prova orale, o ancora tra discipline epistemologicamente assolutamente diverse.

Sostanzialmente fare una media aritmetica tra voti espressi con numeri non è concettualmente diverso dal fare la stessa cosa con aggettivi qualificativi: che media c'è tra sufficiente e buono? Per contrastare

questa pericolosa equivocità la valutazione scolastica dovrebbe, come anticipato poc'anzi, utilizzare prove di verifica oggettive basate esclusivamente su scale ad intervalli.

La *scala ad intervalli* possiede le stesse caratteristiche delle scale ordinali, con il vantaggio che l'intervallo fra i valori distribuiti su essa rimane costante per tutta la sua estensione; un esempio può essere rappresentato dalla scala utilizzata nei termometri. Un'altra caratteristica delle scale ad intervalli è che lo zero non rappresenta la mancanza della qualità che si va ad osservare, ma è una delle modalità convenzionalmente definite. Per avere un'idea più immediata, pensiamo ad una gara che si svolge su una percorso fissato e che per la determinazione della classifica di arrivo si basa sui tempi di percorrenza del percorso stesso. Quando si rileva il tempo preciso nel quale un atleta, un motociclista, un automobilista taglia il traguardo si utilizzano cronometri ad altissima precisione che sono in grado di misurare il millisecondo, che per avere un'idea corrisponde all'incirca al tempo impiegato per sbattere le palpebre, e si potrebbero usare strumenti ben più sofisticati, precisi e dettagliati, fino ad arrivare a misurare lo yoctosecondo! L'unità di misura standard del tempo, utilizzata dal Sistema Internazionale, è quindi il secondo. In base ad esso sono definite misure sottomultiple (millisecondo, microsecondo, nanosecondo ... zeptosecondo e yoctosecondo) e multiple (minuto, giorno, settimana, mese, anno, secolo, millennio ... yottasecondo, che corrisponde a 32 biliardi di anni). Il secondo è definito come la durata di 9.192.631.770 periodi della radiazione corrispondente alla transizione tra due livelli iper-fini dello stato fondamentale dell'atomo di cesio-133. Sostanzialmente, quindi, l'intervallo rappresenta il grado di precisione della misura; maggiore è l'intervallo minore è la precisione e viceversa.

La scala ad intervalli risulta essere più idonea della scala ordinale per attribuire punteggi ad una prova di verifica. In questo caso l'intervallo della scala andrà da 0, nel caso non si sia risposto ad alcuna domanda, al punteggio massimo teorico, quello che coincide con l'aver dato tutte le risposte esatte. Se, ad esempio, si considera una prova a scelta multipla costituita da 30 items con una scelta alternativa di 4 risposte e si attribuisce il punteggio 3 alla risposta esatta e 0 a quella sbagliata, o omessa, la scala di valori andrà da 0 a 90.

Questa tipologia di scala consente di determinare con discreta precisione in che misura uno studente possieda una specifica abilità o conoscenza. Il vantaggio principale, comunque, sta nel fatto che, utilizzando una scala ad intervalli, si possono effettuare confronti dettagliati fra i risultati conseguiti da alunni diversi. Alle misurazioni consentite per la scala ordinale, si possono aggiungere, infatti, indagini relative alla *media* dei risultati ottenuti dal singolo alunno o dall'intera classe, *indici di variabilità* e di *dispersione*, e ancora si possono valutare le differenze del livello di apprendimento conseguito dai diversi alunni valutati, attraverso il *calcolo della deviazione standard*.

La *scala di rapporti*, ancora, si può definire come un ulteriore sviluppo della scala precedente poichè, pur presentando tutte le qualità di essa, consente anche di stabilire un rapporto quantitativo preciso tra le diverse posizioni. Nella scala di rapporti, contrariamente a quella di intervalli nella quale il punto zero non coincide con la mancanza di una qualità ma con un valore generalmente scelto in modo convenzionale (ad esempio lo zero del termometro corrisponde convenzionalmente alla temperatura in cui l'acqua solidifica), il punto zero sta ad indicare l'assenza di una qualità. Un esempio di scala di rapporti è il sistema metrico decimale che si presta alla misurazione puramente quantitativa.

In campo educativo questo tipo di scala non trova applicazione poichè, come abbiamo anticipato, non si può stabilire l'assoluta mancanza di abilità o conoscenza, che coinciderebbe con uno zero assoluto, nè tantomeno, in relazione ai voti conseguiti in una prova di verifica, si può affermare che chi ottiene la valutazione 10 possiede il doppio delle abilità o conoscenze di chi consegue 5.

Possiamo dunque affermare che le scale di valutazione utilizzate in campo scolastico sono una sorta di *scala ibrida* che presenta alcune delle caratteristiche delle scale studiate, ma non è paragonabile in senso stretto a nessuna di esse. Non è una scala di rapporti, poichè non si può definire lo zero assoluto ne confrontare quantitativamente il possesso di abilità o conoscenze. Non è una scala ad intervalli poichè è impossibile stabilire se la distanza tra un voto e il suo precedente sia identica a quella tra lo stesso voto e il successivo, motivo per cui spesso gli insegnanti utilizzano una serie di mezzi punti o simboli per indicare le variazioni di intervallo tra un voto e l'altro. Si potrebbe, quindi, dire

che si tratta di una scala ordinale poichè le valutazioni si presentano in un ordine progressivo crescente, ma occorre precisare che la scala di valutazione utilizzata in ambito scolastico presenta particolarità ben precise che, statisticamente parlando, non le consentono di rientrare a pieno titolo neanche nella scala ordinale. L'intervento di fattori soggettivi che influenzano la valutazione scolastica, infatti, determina una non unicità del valore attribuito ai voti da insegnanti diversi, e ancora bisogna tener presente che ogni atto valutativo è fortemente influenzato dal contesto nel quale si esplica. La scala dei voti può quindi essere definita come una *scala ordinale speciale*.

Volendo uscire dalla specificità della valutazione scolastica, possiamo dire che le scale di misura ad intervalli e/o a rapporti esprimono, in genere, la misura di caratteri quantitativi continui come, ad esempio, le misure antropometriche (peso, altezza, ecc.) e le misure delle leggi fisiche (temperatura, pressione, potenza, ecc.). È intuitivo comprendere che un fenomeno reale, misurato su scala ad intervalli o a rapporti, presenta grandi vantaggi dal punto di vista valutativo in ragione del fatto che garantiscono oggettiva comparabilità nel tempo e nello spazio.

Capitolo 3

L'indagine statistica

L'*indagine statistica* è uno strumento che un numero crescente di settori e di discipline utilizzano in modo sempre più diffuso per soddisfare un fabbisogno informativo relativo ad un fenomeno reale. Il fabbisogno riguarda, più nello specifico, la stima di una o più caratteristiche di una popolazione oggetto di indagine.

Le informazioni relative ad uno specifico fenomeno reale possono essere acquisite osservando tutte le unità componenti la popolazione o soltanto alcune di esse. Nel primo caso l'indagine è detta *completa* o *censuaria*, nel secondo, *parziale* o *campionaria*. Un lettore che abbia poca dimestichezza con lo strumento dell'indagine sarà facilmente portato a pensare che se si osserva l'intera popolazione sicuramente si otterranno informazioni più complete del fenomeno oggetto di studio. Nel corso di questa trattazione, che illustrerà concettualmente i principi fondamentale dell'indagine statistica e non ha pretesa alcuna di essere esaustiva di tutta la "Teoria dei campioni", vedremo che quanto creduto dai più in relazione alle pratiche di indagine non è esattamente sempre così vero.

Da un punto di vista teorico un'indagine statistica completa può apparire apparentemente semplice ma all'atto pratico scopriremo che presenta molti lati negativi. La numerosità elevata della popolazione, ad esempio, può comportare l'impiego di notevoli risorse economiche, di un elevato numero di collaboratori e di tempi lunghi. Esistono poi alcune indagini che per le caratteristiche specifiche della popolazione

potrebbero indurre a pensare che non possano essere svolte; si tratta di quelle indagini rivolte a popolazioni non finite.

Cerchiamo di approfondire meglio il concetto di popolazione in relazione al requisito di finitezza. In generale, una popolazione statistica è costituita da un collettivo di unità statistiche aggregabili per una o più caratteristiche. Se, ad esempio, volessimo indagare il livello di presenze/assenze mensile di un gruppo di impiegati di un'azienda, la nostra popolazione è rappresentata da tutti gli impiegati dell'azienda; ogni impiegato rappresenta una unità statistica della nostra popolazione. La caratteristica oggetto di indagine è, in questo esempio, rappresentata da "presenze/assenze mensile" del dipendente.

Una popolazione può essere definita finita se costituita da un numero finito di unità statistiche, che utilizzando una simbologia statistica si esprime con N (nel nostro esempio N indica il numero complessivo degli impiegati). Indicheremo, invece, con x un qualsiasi carattere osservato sulle N unità statistiche che costituiscono la popolazione finita, ad esempio la presenza/assenza mensile di ogni dipendente dell'azienda. È possibile, su ogni unità statistica registrare la caratteristica "presenza/assenza" e indicheremo con x_1 la registrazione corrispondente all'individuo 1, con x_2 la presenza/assenza dell'individuo numero 2 e così via, fino ad arrivare all'ultimo dipendente dell'azienda. Sulle N osservazioni, che rappresenteranno i singoli dipendenti dell'azienda, è possibile calcolare alcuni indici che descrivono gli aspetti salienti del carattere x preso in esame sulle N unità statistiche della popolazione finita. Tali indici, che studieremo nel dettaglio in seguito, sono delle costanti che in statistica si definiscono *parametri*.

Sono tipici esempi di parametri la media, la varianza, i quartili, lo scarto quadratico medio ecc.

Il concetto di popolazione infinita è meno immediato di quello di popolazione finita. Una popolazione infinita è formata da un numero potenzialmente infinito di unità statistiche che in un certo istante possono anche non esistere fisicamente. La popolazione infinita, quindi, è una popolazione ipotetica. Ad esempio, se si vuole misurare il tempo medio di percorrenza di 100 metri in una gara di velocità, le unità statistiche di riferimento sono rappresentate dalle singole prove di velocità nelle quali misuro il tempo di percorrenza; la popolazione,

invece, è costituita da tutte le prove che sono potenzialmente infinite.

Nelle popolazioni infinite avremo modo di vedere che il carattere d'interesse può essere rappresentato da una variabile casuale alla quale è associata una funzione di probabilità o una funzione di densità, a seconda che tale variabile casuale sia discreta o continua. Generalmente una funzione di probabilità o di densità è caratterizzata da uno o più parametri che la specificano completamente. In questo contesto la variabile casuale x rappresenta il modello statistico che interpreta la caratteristica di interesse sugli elementi della popolazione infinita. Quindi, la distribuzione di x rappresenta la distribuzione teorica del carattere, non quella reale. Nell'ambito delle popolazioni infinite i parametri che specificano completamente la distribuzione della variabile casuale x sono delle costanti che sintetizzano alcuni aspetti caratteristici della x stessa, quali ad esempio la media e la varianza. Un esempio di popolazione infinita è rappresentato dalla cosiddetta *popolazione normale* che è appunto una popolazione teorica di unità statistiche. Un'indagine statistica campionaria offre, soprattutto nei casi di popolazioni non finite, ma non solo, una serie di vantaggi che vedremo più approfonditamente nel dettaglio.

Definiamo *campione* un qualunque sottoinsieme della popolazione contenente un certo numero di unità statistiche. Affinché il campione sia utile all'analisi statistica deve essere rappresentativo della popolazione, cioè deve "riprodurre" tutte le caratteristiche della popolazione. In altri termini deve assomigliare il più possibile alla popolazione nel suo collettivo. La possibilità di limitare la rilevazione ad un insieme di unità di dimensione ben inferiore a quella della popolazione consente sicuramente di contenere i costi dell'indagine entro limiti accettabili, di svolgere l'indagine in tempi relativamente brevi, di raccogliere per ogni unità inclusa nell'indagine un maggior numero di informazioni, di raccogliere le informazioni con maggior accuratezza grazie all'utilizzazione di personale qualificato e/o di tecniche specialistiche. Sul piano teorico occorre, tuttavia, tener presente che l'indagine campionaria presenta due notevoli problemi: il primo, legato al modo in cui deve essere scelto il campione; il secondo, relativo ai procedimenti da adottare per estendere l'evidenza campionaria alla popolazione. Lo studio di questi problemi, che come si vedrà sono strettamente collegati, costituisce l'oggetto della teoria del campionamento statistico.

3.1 Le fasi di una indagine statistica

La statistica per acquisire informazioni su uno o più fenomeni reali utilizza lo strumento dell'*indagine statistica*. Occorre ricordare che i fenomeni reali sono caratterizzati dall'essere fenomeni complessi e, di conseguenza, per essere indagati nella loro complessità è necessario che siano analizzati da un'équipe di persone, che possieda specifiche competenze relative ai diversi aspetti che si vogliono conoscere. Così se si vuole indagare un fenomeno relativo al settore della formazione, in relazione alla specificità del fenomeno osservato, potrebbe essere vantaggioso che cooperino nell'indagine non solo statistici ma anche docenti, psicologi, sociologi, pedagogisti e altri specialisti.

Per condurre una corretta analisi di valutazione basata sul metodo statistico si devono percorrere sei fasi essenziali.

1. *Progettazione*, ossia la fase in cui si procede all'elaborazione del disegno di indagine, si stabiliscono le metodologie e le strategie da utilizzare per condurre l'indagine, si fissano i tempi, si suddividono i compiti tra i ricercatori, si strutturano gli strumenti di rilevazione e si testano per verificarne la validità.

2. *Rilevazione*, ossia la fase in cui si utilizzano gli strumenti strutturati nella progettazione della ricerca al fine di reperire i dati.

3. *Controllo e correzione* dei dati, ossia la fase in cui si individuano gli errori e si procede all'eliminazione degli errori non campionari. Gli errori non campionari sono quelli che si riferiscono direttamente ai dati elementari e che si manifestano come differenze tra valori "veri" e valori "osservati" di una variabile di interesse. Gli errori non campionari possono essere *sistematici* e *casuali*. Si dicono sistematici quegli errori dovuti a difetti strutturali o organizzativi del processo di produzione dell'informazione statistica. Si dicono casuali o stocastici quegli errori la cui origine è da attribuirsi a fattori non direttamente individuabili. In realtà, date le diverse tipologie di errore non campionario che possono simultaneamente contaminare un insieme di dati, diverse sono le metodologie e le tecniche che possono essere utilizzate in modo integrato all'interno della procedura complessiva di controllo e correzione.

4. *Elaborazione*, ossia la fase in cui i dati raccolti vengono elaborati con metodi statistici per ricavare da essi le informazioni sul fenomeno che si vuole studiare.

5. *Interpretazione*, ossia la fase in cui, dai risultati ottenuti e sulla base di conoscenze teoriche, viene data risposta e/o giustificazione alle assunzioni fatte sul problema che si sta esaminando.

6. *Presentazione*, ossia la fase in cui si provvede alla spiegazione ed illustrazione dei risultati ottenuti dall'utilizzo del metodo statistico che non sempre appare chiaro ed immediato ai non addetti ai lavori.

Nella fase di progettazione è fondamentale attuare un processo di astrazione del fenomeno reale oggetto di analisi, individuando gli aspetti più salienti del fenomeno oggetto di studio. È nella progettazione che si costruisce un modello di riferimento che si ritiene idoneo a rappresentare, in una versione semplificata, il fenomeno su cui si deve effettuare l'indagine. Il documento di progettazione deve indicare in modo dettagliato la popolazione di riferimento, ossia l'insieme di unità statistiche individuate quali soggetti dell'indagine. Si deve, inoltre, identificare quali siano le variabili rispetto alle quali si effettuerà la rilevazione delle informazioni. Fissati questi aspetti si procede al disegno dell'indagine che consiste nell'effettuare una scelta della metodologia, degli strumenti di indagine e delle strategie di somministrazione, prestando particolare attenzione alla stima dei costi e dei tempi. Il disegno può prevedere tipologie di indagini differenti.

- *Occasionali*: si tratta di indagini che hanno lo scopo di ottenere stime riferite a caratteristiche possedute da una specifica popolazione in un preciso istante di tempo (es.: distribuzione per età degli studenti di un ateneo che hanno partecipato ad un evento specifico, che può essere ad esempio un seminario) o riferite a un periodo (es.: distribuzione per consumo di bevande energizzanti degli iscritti ad un ateneo in un semestre accademico).

- *Ripetute*: chiamate anche indagini periodiche o ricorrenti poiché vengono effettuate con intervalli di tempo predefiniti. Un esempio può essere un'indagine effettuata trimestralmente sul fattu-

rato o sulla forza lavoro di uno specifico settore di produzione industriale.

- *Longitudinali senza rotazione*: sono indagini predisposte con lo scopo di seguire un particolare gruppo di unità nel tempo, in modo da creare un record longitudinale per ogni unità osservata. L'obiettivo è quello di studiare le modificazioni intervenute nel collettivo durante il tempo, utilizzando i cambiamenti avvenuti sui record individuali. La popolazione che partecipa all'indagine deve rimanere stabile nel tempo, quindi non è previsto inserimento di nuove unità in essa.

- *Longitudinali con rotazione*: indagini disegnate per seguire un particolare gruppo di unità per un periodo di tempo, introducendo nuove unità nel campione in occasioni specificate. Mediante l'ingresso periodico di nuove unità nel campione è possibile mantenere il campione stesso rappresentativo della popolazione, in quanto in questo modo si tiene conto che nel tempo il collettivo di interesse si modifica con l'ingresso di nuove unità (es.: nascite o immigrazioni).

- *Indagine totale*: rilevazione in cui tutte le unità, delle quali si possiede un indirizzo nei propri archivi di base, sono interessate dalla rilevazione. La più importante fra le rilevazioni totali è senz'altro il censimento. Anche se dal punto di vista teorico con un'indagine totale si riescono ad ottenere misure precise dei paramenti di interesse, tuttavia queste indagini presentano un enorme costo di rilevazione e trattamento dei dati nonché problemi connessi alla qualità dei dati, primo fra tutti l'incompletezza della rilevazione dovuta all'incapacità di raggiungere tutte le unità statistiche.

- *Indagini campionarie*: sono caratterizzate dal fatto che solo una parte delle unità statistiche componenti la popolazione viene selezionata ed indagata (campione). Questo espediente, diminuendo l'onere della rilevazione, consente di destinare maggiore attenzione a tutte le attività connesse al miglioramento e al controllo della qualità dei dati raccolti. Se la selezione del campione

viene effettuata con scelta rigorosamente casuale, è possibile misurare il livello di precisione delle stime ottenute rispetto al vero valore del parametro di interesse nella popolazione.

La rilevazione delle informazioni avviene sulla base delle specifiche contenute nel documento di progettazione. Nella fase di rilevazione, le unità selezionate per l'indagine vengono contattate allo scopo di raccogliere l'informazione rilevante ai fini dello studio. Le modalità di contatto e raccolta dati presso le unità di rilevazione dipendono dalla tecnica di indagine adottata. Indipendentemente dalla tecnica adottata, la rilevazione ha come obiettivi l'individuazione delle unità di rilevazione e la raccolta delle informazioni in modo neutrale, senza cioè distorsioni dovute all'influenza dello strumento utilizzato o dell'intervistatore sul rispondente. Affinché tali obiettivi siano raggiunti occorre che l'attività di rilevazione sia preparata con cura, attraverso la formazione dei rilevatori nelle indagini dirette o telefoniche e la predisposizione di strumenti di rilevazioni di facile comprensione, siano essi somministrati direttamente, recapitate a domicilio o tramite i diversi canali delle nuove tecnologie dell'informazione e della comunicazione. Inoltre, si devono predisporre meccanismi di controllo durante la rilevazione stessa per correggere eventuali distorsioni che possono verificarsi in itinere. Quando la registrazione dei dati avviene per mezzo di un intervistatore non è trascurabile considerare il margine di errore che è rappresentato da un'interpretazione sbagliata di quanto riferisce l'intervistato o da errori nell'immissione dei dati nel dataset che si costruisce per la rielaborazione degli stessi. D'altra parte, un margine d'errore si riscontra anche quando la registrazione dei dati avviene direttamente su supporto informatico (es. CATI o CAPI) in quanto potrebbero essere presenti interferenze dovute ad un lavoro poco accurato di predisposizione del tracciato record, che deve riportare le etichette delle variabili, la posizione e la larghezza dei campi, il tipo di campo (numerico, alfanumerico, data, ecc.), i codici da utilizzare nell'immissione delle risposte. Occorre, inoltre, prevedere un codice non ambiguo per indicare le mancate risposte (es. 999) dal momento che altrimenti potrebbero sorgere ambiguità qualora non si faccia esplicitamente la distinzione fra tali codici, gli zeri e i blank. Ai fini della registrazione delle risposte alle domande

aperte si deve, ancora, predisporre un'opportuna codifica.

Cominceremo qui con il trattare la fase della *rilevazione*, tralasciando quella di *progettazione* che sarà ripresa ed ampliata quando prenderemo in considerazione gli strumenti di valutazione. Diciamo innanzitutto che una ricerca può essere eseguita sia a partire dai dati ricavati dalle fonti ufficiali (Istat, Unioncamere, Banche dati ufficiali, ecc.) sia a partire da indagini progettate per il problema specifico.

In genere una rilevazione dei dati viene sviluppata in due fasi:

- determinazione del piano di rilevazione;
- raccolta dei dati.

Nel piano di rilevazione si deve anzitutto conoscere lo scopo a cui mira la rilevazione, i mezzi finanziari ed il capitale umano disponibile. Si deve porre particolare attenzione ai seguenti punti:

1. definire con precisione l'unità statistica o unità di rilevazione;
2. stabilire i caratteri quantitativi e qualitativi che interessano ai fini della ricerca;
3. indicare i mezzi tecnici per raccogliere le informazioni sui dati;
4. fissare l'estensione della rilevazione sulla base della disponibilità finanziaria.

In merito a quest'ultimo punto distinguiamo due tipi di rilevazioni:

- *censuarie*, dette anche totali o universali: in questo genere di rilevazioni l'indagine viene estesa a tutte le unità appartenenti al collettivo o popolazione;
- *campionarie* o parziali: in questo caso viene scelto un numero ridotto di unità del collettivo sulla base di un criterio di scelta che può essere, come vedremo meglio in seguito, casuale o ragionato.

Circa il concetto di popolazione o collettivo, già delineato nel precedente capitolo, aggiungiamo che le popolazioni si possono distinguere in:

- *popolazioni reali;*

- *popolazioni derivanti da sperimentazioni;*
- *popolazioni teoriche.*

Le popolazioni reali sono quelle effettivamente esistenti; si dicono anche popolazioni finite perché sono composte da un numero finito anche se molto elevato di unità. Esse sono accompagnante da una lista che etichetta ciascuna unità della popolazione.

Sono popolazioni reali gli alunni di una scuola, gli iscritti ad una facoltà, le auto prodotte da un stabilimento nel corso di un periodo, le aziende manifatturiere di una regione, i comuni di una provincia ecc.

Le popolazioni derivate da un esperimento sono invece virtuali e si dicono anche infinite ed in genere non hanno una *lista* di riferimento.

Ad esempio, i malati di AIDS costituiscono una popolazione sconosciuta non definita nel numero e nelle unità; non è infatti noto il numero dei malati perché non tutti i malati di AIDS sanno di esserlo, molti sani di oggi potrebbero essere malati in futuro. In questo caso un insieme di unità "malati di AIDS" si deve considerare un campione di una popolazione infinita detta anche *super popolazione.*

Le popolazioni teoriche sono quelle descritte da modelli matematici. Come meglio vedremo più avanti, si suppone che il carattere osservato sulla popolazione abbia una ben specifica distribuzione descritta da un modello matematico. Questo genere di popolazioni è utilizzato per descrivere fenomeni reali complessi, come ad esempio il movimento di una perturbazione atmosferica, ma anche fenomeni più semplici, come il modello di sviluppo di crescita dei bambini o la distribuzione del reddito di una nazione ecc.

Da quanto detto è ovvio desumere che se la popolazione è reale, allora è sempre possibile scegliere se condurre un'indagine censuaria o parziale, mentre negli altri due casi si deve sempre ricorrere ad un'indagine campionaria.

3.2 Gli strumenti di rilevazione

Con il termine *tecnica di indagine* si intende l'insieme delle modalità di contatto delle unità statistiche interessate dalla rilevazione e di reperimento delle informazioni oggetto di interesse. La scelta della

tecnica di indagine più idonea a raccogliere le informazioni oggetto della ricerca è uno degli aspetti di maggiore importanza nella pianificazione e nell'esecuzione di una indagine ed è strettamente connessa ad altre caratteristiche quali il fenomeno indagato, gli archivi di base, la strategia di campionamento, l'organizzazione del personale sul campo, i costi e i tempi attesi. Le diverse tecniche di indagine sono di seguito elencate.

- L'*intervista diretta* (o faccia a faccia) viene condotta da un rilevatore che legge le domande e le opzioni di risposta nell'esatto ordine e con lo stesso linguaggio adottati nel questionario, riportandovi quindi le risposte così come sono fornite dal rispondente. Se tale indagine è svolta tramite l'ausilio del computer si parlerà di CAPI (Computer Assisted Personal Interviewing). Questo tipo di rilevazione si presta meglio ad alcuni disegni di indagine (es.: censimenti) e offre una maggiore possibilità di contattare e convincere il rispondente a collaborare in quanto il rispondente è identificato esattamente. L'intervistatore deve chiaramente esplicitare all'intervistato la motivazione e gli obiettivi dell'indagine e deve dare istruzioni dettagliate e precise sul significato delle domande e sul modo corretto di fornire le risposte. Tuttavia, questa modalità di raccolta delle informazioni presenta alcuni svantaggi che sono da ricercare soprattuto nei costi e nella difficoltà di una organizzazione capillare sul territorio. Inoltre, richiede tempi più lunghi di altri metodi per la raccolta dei dati e comporta maggiori rischi di condizionamento delle risposte.

- L'*intervista telefonica* viene condotta appunto al telefono da un intervistatore che legge le domande e le opzioni di risposta nell'esatto ordine e con lo stesso linguaggio adottati nel questionario, riportandovi quindi le risposte così come sono fornite dal rispondente. Se tale indagine è svolta tramite l'ausilio del computer si parlerà di CATI (Computer Assisted Telephone Interviewing). Sicuramente, rispetto all'intervista faccia a faccia, si abbattono i costi e la raccolta dei dati è più tempestiva in quanto non è richiesta un'organizzazione sul territorio. Anche i rischi di condizionamento sono ridotti. Tuttavia, occorre tener presente

che alcuni individui non sono raggiungibili al telefono e non si ha la certezza dell'identità dell'intervistato.

- Il *questionario postale autocompilato* si invia a mezzo posta o corriere e deve essere compilato e rispedito indietro o eventualmente a riconsegnarlo ad un addetto che lo ritira a domicilio. Questo strumento sicuramente abbassa i costi di realizzazione e richiede un'organizzazione minore. I rischi di condizionamento sono molto bassi e consente anche di porre quesiti delicati poiché l'intervistato può rispondere senza sentirsi imbarazzato dalla presenza dell'intervistatore. Tuttavia, gli svantaggi sono molti poiché i tempi di raccolta possono essere molto lunghi e si incorre nel rischio di una scarsa partecipazione all'indagine.

- Il *diario* è un particolare tipo di questionario strutturato appositamente per registrare eventi frequenti e di scarsa importanza quali spese di bassa entità o attività quotidiane. L'organizzazione di tale strumento è tale da permettere la registrazione degli eventi nel momento della giornata in cui essi avvengono in modo tale da non dover ricorrere ad uno sforzo di memoria, con una conseguente sottonotifica degli eventi, nello svolgimento di una intervista di tipo classico. Questo tipo di strumento non è affetto da problemi di memoria per la rilevazione di eventi poco rilevanti e ad elevata frequenza (ad esempio: spese giornaliere, uso del tempo, visione di programmi TV), tuttavia, la sua struttura è molto complessa e facilmente condizionata dei comportamenti da registrare.

3.3 Il questionario

Il questionario, in tutte le tipologie di indagine, rappresenta lo strumento di misura designato a raccogliere le informazioni sulle variabili qualitative e quantitative oggetto di indagine.

Il questionario deve essere visto come uno strumento di comunicazione finalizzato a facilitare l'interazione fra il ricercatore, il rilevatore e il rispondente. Affinché possa svolgere il suo ruolo occorre che il questionario sia uno strumento standardizzato; ovvero domande e

comunicazione devono essere identiche per tutti i rispondenti in modo tale da garantire la confrontabilità delle informazioni raccolte.

La realizzazione del questionario si articola in diverse fasi:

1. Progettazione:
   - progettazione del questionario in relazione al modello di ricerca di riferimento.
2. Strutturazione del questionario:
   - scelta del tipo di domande e formulazione dei quesiti;
   - redazione del questionario.
3. Verifica del questionario, strutturata in tre fasi:
   - prima fase: la bozza del questionario viene provata su un campione ragionato di unità;
   - seconda fase: confronto sperimentale di due o più versioni del questionario, o di diverse tecniche di indagine, diverse sequenze di domande;
   - terza fase: prova generale dell'indagine, rivolta a valutare la bontà del questionario ma anche di tutti gli altri aspetti della ricerca.

Per la stesura di un questionario è opportuno far riferimento ad alcuni principi essenziali mirati a garantire una facile fruizione da parte dei soggetti intervistati. Prima di tutto è consigliabile stabilire una sorta di gerarchia nella strutturazione delle domande, in relazione alla quale raggruppare le domande in sotto-aree omogenee per tematica ed individuare la collocazione ottimale delle stesse, adottando una successione logica dei temi.

Ogni domanda deve essere fine a se stessa e non deve in alcun modo condizionare la risposta alle successive, salvo il caso in cui si tratti di domande filtro costruite per evitare che gli intervistati siano costretti a rispondere a domande per le quali non possiedono i requisiti e per consentire di indagare alcuni aspetti più specifici in maniera più dettagliata. I diversi item, quindi, non devono concatenarsi o essere

l'uno conseguente dell'altro, pertanto si deve fare in modo che i quesiti siano tra loro indipendenti.

Particolare attenzione deve essere prestata al linguaggio utilizzato che deve tener conto del substrato socio-culturale della popolazione di indagine e deve utilizzare costrutti sintattici comprensibili e non ambigui. Le domande devono essere costruite con una forma semplice ed esplicita, e formulate per mezzo di una frase in forma interrogativa o affermativa. È necessario, inoltre, evitare di inserire negazioni semplici o doppie nel corpo domanda; se fosse proprio indispensabile, bisognerebbe avere l'accortezza di evidenziare la negazione in neretto, con sottolineatura o usando la lettera maiuscola. Quando si propongono risposte che indicano cifre o quantità è opportuno disporle in ordine crescente o decrescente.

Il questionario, infine, deve essere ben calibrato anche nella dimensione, evitando domande ridondanti che possano impegnare per tempi troppo lunghi l'intervistato.

Le domande contenute nel questionario possono essere classificate in relazione all'apertura/chiusura delle domande e delle risposte. I questionari non strutturati sono caratterizzati dall'avere uno stimolo aperto che consente una risposta aperta dell'intervistato. Questo tipo di questionario, utilizzato soprattutto quando il fenomeno reale oggetto di studio è parzialmente o completamente sconosciuto, presenta notevoli difficoltà nella trattazione attendibile e oggettiva dei dati poiché è caratterizzato, appunto, dalla presenza di una forte componente soggettiva sia nella compilazione sia nell'interpretazione, che comporta una delicata fase di codifica. I questionari semi-strutturati si possono differenziare in due tipologie: stimolo chiuso - risposta aperta o stimolo aperto - risposta chiusa. La presenza di vincoli nella domanda o nella risposta garantisce una maggior oggettività sia nella compilazione sia nella valutazione. I questionari strutturati, infine, caratterizzati dalla presenza di stimolo chiuso - risposta chiusa, rappresentano sicuramente lo strumento più utilizzato poiché garantisce oggettività e affidabilità. Generalmente sono usate, infatti, per accertare la presenza-assenza di un carattere, e spesso rappresentano domande filtro, utili per individuare eventuali sottogruppi ai quali saranno sottoposte ulteriori domande più specifiche. Sono classificabili in questa categoria i quesiti vero/falso, a scelta multipla, a

corrispondenze, a completamenti.

Nella figura sottostante si può cogliere il livello di strutturazione di un questionario.

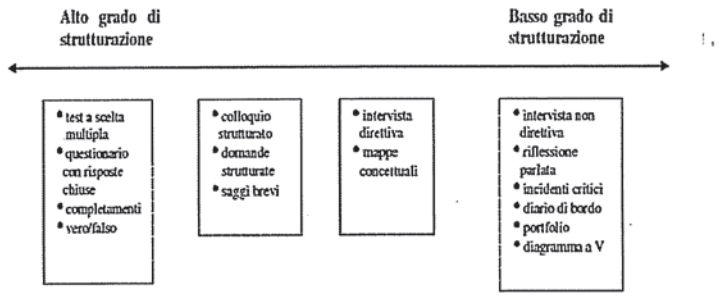


Figura 3.1: Livello di strutturazione del questionario.

3.4 Le rilevazioni campionarie o parziali

Da quanto detto appena sopra, si intuisce che molto spesso il ricercatore si trova nelle condizioni di lavorare su un sottoinsieme di unità, chiamato campione, appartenente al collettivo di riferimento. In genere se la popolazione è finita si indica con N la numerosità del collettivo e con n quella del campione. Definita la *popolazione obiettivo* di un'indagine statistica, è necessario verificare la disponibilità di una base di campionamento (*frame*) che le corrisponda perfettamente. In altri termini, occorre disporre di una lista completa delle sue unità. Per *lista* si intende un insieme ordinato di contrassegni (*label*) delle unità della popolazione, registrati su un supporto informatizzato che ne consenta la trattazione dei dati. La lista identifica, quindi, in modo dettagliato la *popolazione di selezione*. Purtroppo, talvolta, si possono verificare casi in cui non esiste perfetta coincidenza tra popolazione di selezione e popolazione obiettivo in quanto, soprattutto in campo sociale ed economico, gran parte delle liste oggi disponibili presentano difetti tra i quali il più grave è quello della incompletezza, ossia della presenza di dati mancanti. Una volta che si è selezionato il campione,

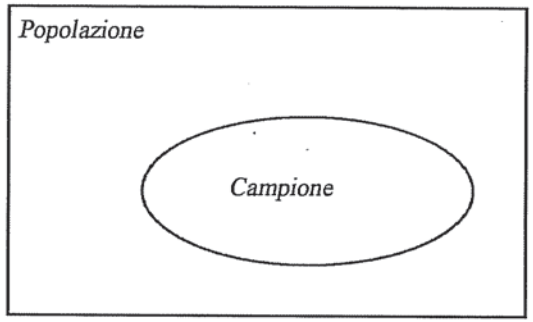


Figura 3.2: Rappresentazione grafica del sottoinsieme "campione".

potrà ancora accadere di non poterne osservare tutte le unità, in quanto si possono verificare casi in cui si sia impossibilitati a contattare tutte le unità o in cui siano le unità contattate a rifiutarsi di partecipare all'indagine. Il fenomeno della mancata osservazione di un'unità che fa parte della popolazione di selezione prende il nome di *non risposta* o *mancata risposta* e rappresenta la *popolazione di indagine*. In presenza di questo fenomeno il campione fornisce evidenze soltanto sull'insieme di coloro che sarebbe stato possibile osservare se l'indagine fosse stata completa. L'insieme delle unità che effettivamente sono state sottoposte ad indagine costituisce, appunto, la popolazione di indagine (*survey population*). Di fatto, quindi, accade frequentemente che la popolazione obiettivo, cui è direttamente interessato chi svolge l'indagine, differisca da quella di selezione a causa dell'incompletezza della lista. La popolazione di selezione differisce a sua volta da quella di indagine a causa della non risposta. Nella pratica spesso non si tiene adeguatamente conto di tali differenze o si ipotizza implicitamente che siano trascurabili.

Se fosse sempre possibile effettuare rilevazioni sul totale della popolazione si avrebbe sicuramente una conoscenza esaustiva e completa del fenomeno reale oggetto di studio; tuttavia, molto spesso vincoli di tempo e di costi, nonché difficoltà concrete di progettazione e strutturazione di indagini censuarie, portano ad effettuare scelte che risultino più efficaci nella conoscenza del fenomeno stesso in relazione

al problema costo-beneficio. Occorre, inoltre, considerare che esistono alcune popolazioni per le quali non è possibile definire una lista completa delle unità e che per esse sarebbe quindi impossibile procedere ad una rilevazione totale. La costruzione e lo studio delle modalità di campionamento è noto come *teoria dei campioni*. In particolar modo ci interesseremo di disegni molto semplici che si basano sulla scelta casuale degli elementi dalla popolazione.

L'indagine campionaria, o parziale, è prevalentemente usata, quindi, per ridurre il costo di un'indagine e/o per ridurre i tempi di elaborazione della ricerca; altre volte è l'unica forma di indagine in virtù del tipo di popolazione virtuale o teorica a cui si è fatto riferimento sopra.

In ogni caso il campione deve essere rappresentativo della popolazione, ossia deve essere una immagine ridotta in termini di unità della popolazione, ma deve rappresentare con uguale proporzione tutte le modalità del carattere che si vuole valutare. La rappresentatività del campione, tuttavia, vedremo che non è da sola sufficiente a descrivere la popolazione intera nel suo complesso se non è seguita da una stima accurata degli indici di sintesi, i *parametri*, e se non si procede alla *verifica delle ipotesi* per valutare la congruità tra ciò che si è teoricamente ipotizzato e ciò che si è concretamente rilevato. In modo più specifico diremo che gli approcci ad un'indagine campionaria si suddividono in due grandi filoni: un primo approccio si definisce *basato su modello* ed è applicabile allorquando si hanno a disposizione solide conoscenze sulla struttura dell'universo, in quanto sono proprio le caratteristiche della variabile universo che danno le informazioni necessarie alla costruzione di un *modello statistico*; un secondo approccio si definisce *basato su disegno* e si fonda sulle caratteristiche delle variabili casuali campionarie che sono determinate appunto dal disegno di campionamento.

Allo scopo di rendere il campione rappresentativo, le unità appartenenti ad un campione possono essere scelte attraverso un *piano di campionamento* che può essere basato su criterio di scelta casuale o su criterio di scelta ragionata. Un campione è casuale se a ciascuna unità della popolazione è associata una probabilità di estrazione diversa da zero definita dal disegno campionario scelto. In altri termini diciamo che un campione è casuale se tutte le unità della popolazione

hanno probabilità non nulla di essere incluse nel campione; questa probabilità è chiamata appunto *probabilità di inclusione*. Le modalità di estrazione del campione da una popolazione, come vedremo nel dettaglio, possono seguire uno schema probabilistico, quando ogni elemento della popolazione ha una probabilità nota di essere estratto, o diversamente uno schema non probabilistico ma opportunamente disegnato.

Più nel dettaglio si parla di campione o di *campionamento probabilistico* quando è possibile definire l'insieme C (spazio campionario) di tutti i campioni distinti estraibili dalla popolazione ed a ciascun membro, c (campione), di tale insieme è assegnabile a priori una probabilità di selezione, indicata con $p(c)$. Il campione viene selezionato mediante un meccanismo casuale che consenta di associare ad ogni campione c una probabilità di estrazione esattamente pari a $p(c)$.

Naturalmente il piano di campionamento ha per definizione le seguenti proprietà:

$$p(c) \geqslant 0, \forall c \in C \qquad \text{e} \qquad \sum_{c \in C} p(c) = 1$$

Tutti i campioni probabilistici vengono formati, quindi, ricorrendo ad un meccanismo di selezione casualizzata o casuale. Tale meccanismo, che è sintetizzabile in un insieme di regole e/o algoritmi, viene denominato *schema di campionamento*. Sono *non probabilistici* i campioni che non hanno i requisiti suddetti. Le forme più comuni di campione non probabilistico che possono essere raggruppate in due categorie: campioni ragionati e campioni fortuiti. I campioni a scelta ragionata sono formati senza alcun ricorso a meccanismi di casualizzazione, in quanto la scelta delle unità da includere nel campione è affidata al ricercatore ed è operata il più delle volte con obiettivi di rappresentatività di certi aspetti strutturali della popolazione. Un altro tipo diffuso di campione ragionato è quello formato da unità tipo, unità cioè che a giudizio di un esperto, cui è demandata la loro selezione, possiedono caratteristiche ritenute più frequenti nella popolazione. I campioni selezionati fortuitamente non devono essere confusi con i campionamenti casuali poiché non sono formati né con l'ausilio di tecniche di casualizzazione, né seguendo procedimenti che implicano la preferenza da parte del rilevatore per certe unità anziché per altre.

Appartengono, infatti, a questa categoria i campioni formati da volontari, da unità che transitano da passaggi obbligati come frontiere, ingressi di edifici, le casse di un supermercato, ecc. e, ancora, le parti più accessibili di popolazioni generalmente formate da oggetti (si pensi all'estrazione manuale di un campione di riso da un sacco), molti tipi di campioni di animali, ad esempio: pesci catturati con reti, cavie estratte manualmente da una gabbia, ecc.

Si ribadisce che il ricorso al campionamento è necessario sia per ridurre i costi di un'indagine, sia quando le unità che costituiscono la nostra popolazione sono illimitate e, quindi, non è possibile osservarle direttamente tutte. A titolo esplicativo, una popolazione illimitata può essere rappresentata dagli individui affetti da AIDS sulla popolazione mondiale; tale popolazione è illimitata, non finita, poiché non è possibile effettuare il conteggio di tutte le persone malate in quanto, ad esempio, molte di esse potrebbero al momento dell'indagine non essere consapevoli della propria condizione di malattia. Si dirà che gli indici, o parametri, che osserveremo sul nostro campione rappresentano *stime* dei corrispondenti indici reali che avremmo ottenuto osservando l'intera popolazione; tuttavia, tali stime possono variare in relazione alle specifiche registrazioni effettuate sul campione preso in esame. Vedremo, in seguito, che esiste una specifica branca della statistica, denominata con il termine di *Inferenza* che ha, appunto, l'obiettivo di valutare la variabilità della stima di un campione e il suo livello di bontà.

La probabilità di inclusione

Per procedere ad un campionamento casuale non si può prescindere dal rispetto di alcune regole attraverso le quali le singole unità di selezione possono entrare a far parte del campione. Lo schema di campionamento e il piano ad esso associato, quindi, consentono anche di associare ad ogni unità della popolazione una prestabilita probabilità di selezione. Dati, quindi, una popolazione di N unità, che enumeriamo con l'indice i ($i = 1, 2, \ldots, N$), uno schema di campionamento ed un piano di campionamento $p(c)$, la probabilità che una generica unità della popolazione entri a far parte del campione è detta

probabilità di inclusione ed è definita come segue:

$$\pi_i = \sum_{i \in c} p(c)$$

in cui con $i \in c$ si indicano i campioni che includono l'$i-esima$ unità, ai quali si estende la sommatoria. Tale probabilità di inclusione è detta *semplice* o del *primo ordine* in quanto riferita a una singola unità della popolazione.

La probabilità di inclusione è definibile anche per coppie di unità ed è detta, in questo caso, *probabilità di inclusione congiunta* o del *secondo ordine* ed è valida, più in generale, per $n-ple$ (con $n \geqslant 2$). La probabilità di inclusione congiunta, cioè la probabilità che le unità i e j siano ambedue incluse nel campione è definita come segue:

$$\pi_{i,j} = \sum_{iej \in c} p(c)$$

Capitolo 4

La distribuzione di un carattere statistico

Da quanto detto nei capitoli precedenti si comprende che la valutazione di un collettivo può passare attraverso diversi tipi di misura, che possono essere quantitativi o qualitativi, ma anche sconnessi o ordinabili. I dati di un'indagine statistica possono essere rappresentati in una tabella di raccolta dati e analizzati in modo da ottenere una opportuna *sintesi*, *variabilità* e *forma*.

Nell'immaginario comune la statistica è semplicemente un insieme di tabelle, qualche percentuale e un grafico rappresentativo della distribuzione dei dati rilevati. Il tentativo di questo libro è, invece, quello di dimostrare che la statistica è un approccio metodologico all'analisi di un fenomeno reale, in quanto si fonda su una filosofia di ragionamento consolidato da evidenza scientifica. L'obiettivo primario della statistica non è quello di comunicare dati sterili, ma di dare informazioni utili sul collettivo che si sta valutando in modo da aiutare il decisore a fare scelte il più possibile ponderate e corrette. La Statistica Descrittiva nasce, infatti, dalla necessità di estrarre e riassumere le informazioni rilevanti contenute in un grande volume di dati. Questa necessità è motivata dalla incapacità della mente umana di comprendere le informazioni contenute in un insieme molto grande di dati se questi sono organizzati semplicemente come un'elencazione.

Un *dato* è la formalizzazione di un'osservazione realizzata, in re-

lazione ad una variabile che è interesse di indagine, su un individuo che appartiene alla popolazione oggetto di studio. Un qualunque insieme congiunto di dati ottenuto da una rilevazione si chiama *distribuzione*. In questo capitolo studieremo nello specifico i metodi per descrivere le variabili in una distribuzione. In particolare, useremo tabelle e grafici per descrivere la *distribuzione di frequenza* di variabili. Puntualizziamo che con il termine *variabile* si indica una caratteristica osservata di un fenomeno reale che può assumere, in un intervallo fissato, differenti valori. A titolo di esempio possiamo dire che il fenomeno reale "Condizione lavorativa delle donne in Italia" può essere osservato attraverso una serie di variabili, ad esempio "età", "voto di laurea", ecc., ognuna delle quali presenta diverse modalità: "30", "40", ecc., "80", "95", ecc. In seguito vedremo che le modalità possono anche essere raggruppate per classi (età: $20-30$, $30-55$, ecc.), garantendo una maggiore sintesi a discapito però della perdita di alcune informazioni. Con il termine *mutabile*, invece, si indica una caratteristica osservata di un fenomeno reale che può assumere, in un intervallo fissato, differenti aspetti qualitativi. A titolo di esempio possiamo dire che il fenomeno reale "Condizione lavorativa delle donne in Italia" può essere osservato attraverso una serie di mutabili, ad esempio "orientamento religioso", "stato civile", ecc., ognuna delle quali presenta diverse modalità: "cristiano", "induista", ecc., "nubile", "coniugata", ecc. Una variabile e una mutabile rappresentano, quindi, le caratteristiche che osserviamo e differiscono tra loro in relazione al livello di misura che utilizziamo. Le variabili, infatti, si misurano a livello di scala ordinale, ad intervalli e a rapporti, mentre le mutabili esclusivamente su scala nominale.

Affermare che oggi si vive nell'era dell'informazione significa sostenere che conoscere equivale ad avere potere. Le organizzazioni che riscuotono successo sono, in generale, quelle capaci di raccogliere, analizzare e utilizzare in maniera efficiente ed efficace le informazioni. Tuttavia, una recente ricerca condotta da "*Businessweek*" rivolta ai dirigenti di azienda, ha messo in evidenza che almeno il 50% delle decisioni prese dai due terzi degli intervistati si basa su sensazioni piuttosto che su fatti concreti. Dall'indagine emerge, inoltre, che il 77% delle decisioni errate proviene da una carenza di informazioni.

Una simile realtà induce sempre più l'attività scientifica a svilup-

pare metodi complessi di analisi dei dati a supporto delle decisioni, traendo suggerimenti dalla metodologia statistica. Le valutazioni relative all'apertura di un nuovo impianto produttivo, all'entrata sul mercato con un nuovo prodotto o alla valutazione di un piano di offerta formativa rimandano a problemi che richiedono analisi preliminari e specifiche metodologie.

Lo sviluppo della *information technology* ha prodotto un incremento notevole delle attività di ricerca in questo settore, anche con il sostegno delle aziende, delle amministrazioni pubbliche e dei centri di ricerca che, con regolarità, raccolgono ed elaborano ingenti masse di dati. I supermercati, ad esempio, emettono quotidianamente migliaia di scontrini. Il contenuto di ciascuno di essi riflette le esigenze di spesa, le propensioni ed il comportamento economico del consumatore. I dati raccolti rappresentano un utile strumento per orientare le politiche di vendita e di approvvigionamento delle merci. In microbiologia l'analisi delle sequenze di porzioni di *DNA* porta alla costruzione di gigantesche tabelle dette *DNA microarray*, in cui ciascuna colonna costituisce una sequenza di alcune migliaia di valori numerici. Tali dati possono essere analizzati per evidenziare le relazioni tra il *DNA* e la presenza di malattie. In ambito ambientale i moderni strumenti di monitoraggio permettono la raccolta continuativa di grandi quantità di dati di natura fisica, chimica e biologica. L'obiettivo di un loro studio potrebbe essere quello di valutare la qualità della vita nei centri urbani, l'evoluzione della bio-diversità attraverso le abbondanze delle specie, le variazioni climatiche, ecc.

L'organizzazione strutturata di tali dati e la loro analisi sistematica sono strumenti essenziali per il manager d'azienda, il ricercatore scientifico, l'amministratore pubblico, i quali possono disporre, in tal modo, di un quadro completo ed obiettivo del fenomeno oggetto d'attenzione e prendere adeguate decisioni.

Dovendo ricavare valutazioni di sintesi dai dati, si intuisce come la statistica viene configurandosi strumento strategico che mette a disposizione validi e potenti metodologie. Spesso si ritiene che la statistica possa descrivere solo situazioni semplici; al contrario, si tratta di una scienza che aiuta ad esprimere giudizi essenziali che facilitano il *decision maker* nelle scelte tra scenari alternativi e complessi. È necessario, quindi, ricordare che il metodo statistico è un costrutto logico

induttivo ottenuto su base sperimentale; è un metodo di investigazione scientifica che, a partire da una misura associata ad ogni fenomeno reale, giunge, attraverso le osservazioni empiriche, ad argomentazioni logiche ed a principi di validità generale. Ad esempio: la statura degli individui è misurabile in centimetri; la compravendita di un titolo in borsa è misurabile con il prezzo in euro fissato per l'acquisto, ecc. Si intuisce che un fenomeno reale è costituito da un numero di repliche finite o infinite sulle quali si manifesta una modalità della misura ad esso associata. La domanda che ci si pone è: Il fenomeno reale che si vuole studiare si manifesta in modo identico in tutte le sue repliche? Qualora ciò non accada, si impone la necessità di studiare il comportamento del fenomeno e ricavare, se possibile, una regola o un'interpretazione della sua variabilità.

L'approccio è, quindi, quello del metodo sperimentale che fonda le sue radici nella filosofia dell'umanesimo e del rinascimento, basato sull'osservazione e sulla replica dell'esperimento per ricavarne le regolarità.

Nel senso più ampio si ritengono repliche le diverse misure associate alle caratteristiche di una scolaresca, dei prodotti industriali, le rilevazioni demografiche su aree geografiche, le spese per consumi in mesi o anni diversi ecc. In ogni caso il focus è sempre lo stesso: ricavare regole generali sulla base dei dati osservati. Generalizzando, si intuisce come sia difficile trovare un esempio che non possa essere studiato con metodo statistico.

È, tuttavia, necessario precisare che il metodo statistico a supporto delle decisioni sfocia in due grossi ambiti applicativi: quello basato sui dati e quello basato sul modello. Il primo si limita all'analisi descrittiva del fenomeno reale che si vuole studiare in termini di indicatori quali, ad esempio, indici, tassi, ecc. Un esempio di applicazioni basato sui dati è il tasso d'inflazione ottenuto sulla base di un paniere di beni monitorato in tempi e/o luoghi diversi.

Un approccio sistematico a sostegno del processo decisionale, sostanzialmente diverso da quello basato sui dati, è quello di costruire un modello capace di descrivere, comprendere, prevedere, simulare e controllare un fenomeno reale.

Il metodo statistico basato sul modello tenta, a partire dai dati osservati, di definire un modello generale ossia un regola che sia in

grado di sostituire in tutto e per tutto i dati ottenuti dalle repliche sperimentali. Ad esempio, in campo dei processi formativi si ipotizzi che si voglia sperimentare un proprio *progetto formativo*. Allo scopo si supponga di applicare il metodo formativo a un collettivo di studenti. Da quanto detto nei precedenti capitoli, lo sperimentatore dovrà stabilire una misura delle risposte agli stimoli formativi previsti dal processo. I risultati della sperimentazione descriveranno una distribuzione dei risultati al variare degli stimoli. Nel metodo statistico basato sul modello, quindi, la distribuzione dell'esperimento rappresenta la base dei dati necessaria per definire il modello o la regola stessa. In questo ambito il modello ricavato descriverà il fenomeno reale "processo formativo progettato" e sostituirà lo stesso sino a quando nuove fasi sperimentali forniranno dati tali da sostituire o modificare il modello precedentemente accettato.

L'uso di tali modelli si è sostanzialmente diffuso in diversi contesti applicativi, tra cui quello scientifico-tecnologico e quello economico-aziendale. Esso è una riproduzione del fenomeno reale e ne emula gli aspetti essenziali al fine di fornire utili indicazioni per chi deve assumere una decisione. Il modello è una struttura fondamentale di ragionamento della ricerca scientifica e della tecnologia perché consente di esaminare la complessità mediante l'analisi di relazioni semplici. Pertanto un modello potrebbe essere accolto, anche se palesemente falso, solo perché comodo ed universalmente accettato.

In economia, ad esempio, è nota la relazione che lega i livelli di consumo ed i corrispondenti redditi delle famiglie (le famiglie ricche spendono più delle famiglie povere). Tuttavia, supporre che il consumo dipenda in forma lineare dal reddito, cioè che ad aumenti di reddito corrispondano aumenti proporzionali del consumo, è una semplificazione del fenomeno "comportamento del consumo delle famiglie" che potrebbe sicuramente essere meglio rappresentato da un modello più complesso. Il ricorso alla esemplificazione è una necessità metodologica; infatti, adottare un modello scarsamente leggibile complicherebbe la sua interpretazione, facendo perdere di vista il concetto di sintesi di cui ha bisogno il decisore. La costruzione di un sistema di supporto alle decisioni basato sul modello deve essere sostenuta da una rigorosa metodologia statistica, fondata su un protocollo sperimentale che rispetti le ipotesi di lavoro in tutte le sue fasi.

4.1 I dati per un'indagine statistica

Abbiamo detto che la valutazione di un fenomeno passa attraverso lo studio di un collettivo di unità sul quale vengono individuate alcune caratteristiche, *caratteri*, osservate tramite opportune scale di misura. Si intuisce, quindi, che per effettuare una corretta valutazione è necessario eseguire un'accurata e attenta rilevazione dei dati. Facciamo, tuttavia, osservare che questa fase della ricerca viene troppe volte sottovalutata e lasciata all'attenzione degli informatici a cui è dato il compito di organizzare i *data base*, ossia tabelle di dati multidimensionali raccolti.

Nella maggior parte dei casi essi non si prestano ad una immediata elaborazione, in quanto raccolgono informazioni diverse e tra loro eterogenee. L'assunzione "dati uguale informazione" non è quasi mai vera. Si rende necessaria una rielaborazione dei primi in data-base operativi opportunamente sviluppati e specificatamente orientati allo scopo. In altri termini, è fondamentale costruire un *data-base strategico*, detto *data warehouse*, creato con lo specifico scopo di supportare le attività decisionali e contenere tutti i dati di interesse "ripuliti e resi omogenei". Questa fase operativa è chiamata in letteratura *Extract, Transformation and Load (ETL)*. Rientrano in questa tipologia i Sistemi Informativi Esecutivi (*Executive Information Systems-EIS*), i Sistemi Informativi Geografici (*Geographic Information Systems-GIS*) e i Sistemi di Intelligenza Aziendale (*Business Intelligence Systems-BIS*). Un primo

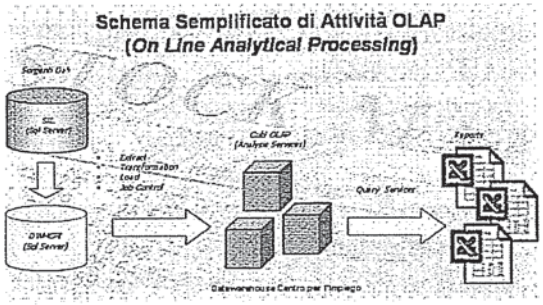


Figura 4.1: Olap (On Line Analytical Processing).

utilizzo del data warehouse conduce alla stesura di reporting di natura semplificata. L'attività di questo tipo viene chiamata *OLAP* (*On Line Analytical Processing*) ed è finalizzata ad estrarre informazioni di sintesi dai dati contenuti nel warehouse. In genere l'attività di *OLAP* consiste nel ricavare opportune tabelle ad entrata multipla dette iper-cubi in cui ogni spigolo rappresenta una variabile e ad ogni cella interna è assegnata una frequenza o un indicatore sintetico corrispondente alle modalità delle variabili considerate.

A titolo di esempio, si supponga che una tabella multipla rappresenti gli studenti universitari: uno spigolo dell'iper-cubo potrebbe essere la variabile "Comune di residenza", un altro spigolo la variabile "Facoltà", un altro "il tipo di diploma di scuola secondaria", ecc. Questa tabella si presta a diverse elaborazioni, alcune delle quali sono di immediata lettura. Si pensi alla ripartizione degli studenti per luogo di residenza, alla distribuzione del tipo di diploma di scuola secondaria nelle diverse Facoltà, e così via. Ma sono possibili anche analisi più complesse che, come meglio vedremo inseguito, mettono in evidenza utili e simultanee informazioni su tutte le variabili in esame.

La costruzione di un *data-base OLAP*, quindi, consente di effettuare una sorta di fotografia dei dati e trasformarli in informazioni multidimensionali. L'applicazione delle tecniche *OLAP* può rappresentare una risposta in diversi ambiti della ricerca e della gestione di una organizzazione; si pensi, ad esempio, al crescente bisogno di informazioni sull'occupazione e all'importanza dell'analisi del mercato del lavoro in un determinato ambito territoriale, nonché alla necessità di disporre di informazioni sui disoccupati, sulle assunzioni, sui licenziamenti, sulle nuove opportunità di lavoro, e così via.

In questo capitolo ci proponiamo di affrontare la fase della raccolta dei dati cercando di fornire una metodologia semplice ed operativa. Per ragioni di opportunità il nostro punto di partenza è una sola dimensione del *data warehouse* che chiameremo *matrice dei dati*.

La matrice dei dati è una tabella a doppia entrata composta da tante righe quante sono le unità osservate e tante colonne quanti sono le caratteristiche degli individui ritenute utili per la valutazione del collettivo. Ad esempio, se l'oggetto della nostra valutazione è un insieme di individui iscritti all'università, allora le righe saranno rappresentate dagli studenti mentre le colonne individueranno le

caratteristiche osservate, come ad esempio gli esami sostenuti, la residenza, il sesso, la professione del padre o della madre, il credo religioso, l'orientamento politico e così via in relazione all'obiettivo della nostra indagine (Tabella 4.1).

Chiamiamo la matrice dei dati *"matrice unitaria"*, poiché descrive i dati raccolti unità per unità ma, come vedremo tra breve, la sua funzione è esclusivamente quella di organizzare i dati affinché questi possano essere analizzati e studiati con metodo statistico, in quanto la sua lettura non ci consente di ottenere in modo immediato informazioni circa la popolazione di riferimento.

Unità	Caratteri					
		Ex.	Prof.			Sc.
Stud.	Sex	Stat.	Padre	Religione	...	Superiore
E.R.	*M*	*26*	*Impiegato*	*Cristiana*	...	*L. Scientifico*
F.M.	*F*	*30*	*Operaio*	*Ebraica*	...	*L. Pedagogico*
G.V.	*M*	*27*	*Medico*	*Islamica*	...	*Ist. Tecnico*
M.R.	*M*	*18*	*Pensionato*	*Induista*	...	*L. Classico*
⋮	⋮	⋮	⋮	⋮	⋮	⋮
S.E.	*F*	*18*	*Notaio*	*Ebraica*	...	*I. Tecnico*
T.I.	*F*	*30*	*Militare*	*Buddista*	...	*L. Scientifico*
Z.A.	*M*	*24*	*Disoccupato*	*Cristiana*	...	*L. Classico*

Tabella 4.1: Esempio di distribuzione unitaria.

Ricordiamo, quindi, che si tratta di una tabulazione dei dati raccolti all'interno di una tabella costituita da tante righe quante sono le unità statistiche e da dante colonne quante sono le caratteristiche osservate di un fenomeno reale, ossia le variabili prese in esame. Naturalmente la matrice dei dati può contenere uno o più tipi di caratteri e può essere studiata considerando un carattere per volta o due o più caratteri insieme.

Nel primo caso diremo che si condurrà un'analisi *univariata*. In molti studi, però, è necessario valutare la relazione esistente tra i caratteri, che potrebbe essere fatta considerando due caratteri per volta o più caratteri per volta. Se prendiamo in considerazione due

caratteri per volta, diremo che si condurrà un'analisi *bivariata*, mentre più in generale, se si coinvolgeranno più di due caratteri, diremo che si condurrà un'analisi *multivariata*.

La tabella 4.1 descrive nel dettaglio un esempio specifico, tuttavia è necessario conoscere la forma generalizzata di questa tabella, ossia la struttura che formalizza le variabili e le caratteristiche osservate in modo da utilizzare una simbologia convenzionale.

La matrice dei dati può essere, infatti, convenzionalmente rappresentata come indicato nella tabella seguente: dove $X_1, X_2, \ldots, X_j, \ldots, X_k$

	Caratteri					
Unità	X_1	X_2	...	X_j	...	X_k
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1k}
2	x_{21}	x_{12}	...	x_{2j}	...	x_{2k}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ik}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	x_{N1}	x_{N2}	...	x_{Nj}	...	x_{Nk}

Tabella 4.2: La matrice dei dati.

rappresentano i caratteri che vogliamo andare ad analizzare; x_{ij}, $\forall i, j$, sono le diverse *modalità* con cui si esprime il carattere X. [1]

L'approccio statistico sarà nel testo affrontato in modo graduale, garantendo un'ampia e dettagliata trattazione con rigore metodologico, partendo dall'analisi unitaria per arrivare ad analisi più complesse. Saranno trattati separatamente i metodi statistici idonei all'analisi di

[1] Da osservare le coppie di numeri e le lettere poste a pedice: il primo numero fa riferimento all'unità statistica che va da 1 ad N, cioè al numero totale delle unità presenti nella nostra matrice; il secondo numero indica la posizione del carattere preso in considerazione, quindi, nella casella in cui troviamo la modalità x_{11} facciamo riferimento alla prima unità statistica e alla misurazione del primo carattere; la lettera j indica, invece, il generico carattere, j-esimo, che prendiamo in considerazione, mentre la lettera k indica che abbiamo osservato un numero k di caratteri dove il valore di k è il numero di caratteri presi in esame.

ciascun tipo di carattere, mettendo in evidenza per ognuno di essi le informazioni ricavabili in funzione della propria potenzialità operativa, così com'è stato specificato nei paragrafi precedenti.

4.2 Distribuzioni unitarie e distribuzioni di frequenza

La matrice che abbiamo analizzato nel paragrafo precedente può presentarsi come una lunga stringa di dati, dalla quale è difficoltoso trarre un'idea riassuntiva ed immediata del fenomeno che si sta indagando. Occorre, quindi, predisporre opportune matrici di dati, o tabelle, che consentano di avere una lettura più immediata.
La prima operazione che si conduce è quella di costruire la *distribuzione di frequenza*. Dalla matrice dei dati, quindi, si estrae una colonna corrispondente al carattere che si vuole esaminare, scrivendo su una prima colonna tutte le modalità diverse con cui si è manifestato il carattere nel collettivo. Nella colonna a fianco si riportano, per ciascuna modalità, il numero di volte con cui la modalità si è ripetuta nel collettivo. Tale numero prende il nome di *frequenza assoluta*.
La distribuzione di frequenza può essere genericamente riassunta nella tabella seguente, dove, come abbiamo detto, nella prima colonna sono riportate le modalità del carattere, mentre nelle colonne successive le *frequenze assolute*, le *frequenze relative* e le *frequenze percentuali*.

Descrizione della tabella:

- $x_1, x_2, \ldots, x_i, \ldots, x_k$ rappresentano le *modalità* con cui si misura il carattereX;
- $n_1, n_2, \ldots, n_i, \ldots, n_k$ sono le *frequenze assolute*;
- $N = \sum_{i=1}^{k} n_i$;
- $f_1, f_2, \ldots, f_i, \ldots, f_k$ sono le *frequenze relative*;
- $\sum_{i=1}^{k} f_i = 1$;
- $p_1, p_2, \ldots, p_i, \ldots, p_k$ sono le *frequenze percentuali*;
- $\sum_{i=1}^{k} p_i = 100$.

	Distribuzione di frequenza		
Carattere	Freq. assoluta	Freq. relativa	Freq. percentuale
X	$\mathbf{n_i}$	$\mathbf{f_i}$	$\mathbf{p_i}$
x_1	n_1	f_1	p_1
x_2	n_2	f_2	p_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_i	f_i	p_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	p_k
	N	1	100

Tabella 4.3: La distribuzione di frequenza.

La *frequenza assoluta* indica il numero di osservazioni rilevate per ogni modalità della variabile presa in esame. È opportuno, comunque, far notare che nella stragrande maggioranza di casi, in particolare quando il collettivo di riferimento è molto grande, la lettura delle frequenze assolute è poco informativa per il ricercatore. In questi casi si ricorre, quindi, al calcolo delle frequenze relative indicate con la lettera f_i. Per calcolare la frequenza relativa per ogni modalità è necessario fare un rapporto tra la frequenza assoluta, della modalità presa in esame, ed il totale del collettivo. Sarà, quindi:

$$f_i = \frac{n_i}{N} \forall i$$

La frequenza relativa f_i è sempre un numero compreso tra 0 e 1, poiché rappresenta la frazione di un intero (il collettivo, cioè N), e può essere letta come il peso che ha la modalità *i-esima* nella distribuzione.
È facile verificare che

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

sarà, infatti:

$$\sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{N} = \frac{1}{N} \sum_{i=1}^{k} n_i = \frac{1}{N} N = 1$$

A fini di una maggiore efficacia comunicativa si usa calcolare la percentuale che indichiamo con p_i. Essa viene calcolata moltiplicando per 100 la frequenza relativa, ossia:

$$p_i = fi \cdot 100$$

Facciamo osservare, tuttavia, che il calcolo della frequenza relativa e conseguentemente della percentuale, potrebbe dare per risultato un numero con diversi decimali. La cosa, naturalmente compromette la chiarezza interpretativa dei risultati. In questo caso si ricorre al calcolo per approssimazione. Ossia al troncamento ad un numero ridotto di decimali attraverso il criterio sopra a 5 o sotto 5. In pratica, supponendo che si voglia troncare al secondo decimale, ossia al centesimo, allora si controlla il decimale successivo, cioè il millesimo; se quest'ultimo risulta maggiore o uguale a 5 allora il numero si approssima al centesimo superiore (approssimazione per eccesso), in caso contrario si approssima per difetto. Facendo un esempio concreto possiamo dire che il valore 0,386, essendo $6 > 5$, si approssima a 0,39, mentre il valore 0,382, essendo $2 < 5$, si approssima a 0,38. Occorre prestare particolare attenzione a quei numeri decimali che necessitano di un "doppio arrotondamento", come nell'esempio seguente: 0,396 si arrotonda a 0,40, in quanto la cifra dei millesimi 6 si arrotonda per eccesso, essendo $6 > 5$, e di conseguenza la cifra dei centesimi, 9, diventa 0 e incrementa di 1 la cifra dei decimi che diventa 4; in questi casi lo zero, pur essendo posto dopo la virgola, non si omette mai per rendere più comprensibile la procedura di approssimazione.

Da quanto detto sopra, la fase preliminare di uno studio di valutazione è l'analisi *univariata* dei dati. In sintesi, quindi, si tratta di estrarre dalla matrice dei dati una sola colonna per volta. In sintesi, quindi, è opportuno riscrivere i dati in una *tabella di frequenza*, ossia una tabella composta da due colonne: nella prima vengono riportate diverse modalità di un carattere, osservate nel collettivo; nella seconda il numero di unità statistiche (*frequenza assoluta*) che hanno la stessa

modalità. Facciamo un esempio concreto del passaggio dalla matrice dei dati alla distribuzione di frequenza, supponiamo di isolare dalla tabella 4.1 la colonna del carattere "Sesso". Una tabella di questo tipo

Unità	Sesso
1	*M*
2	*F*
3	*M*
4	*M*
5	*F*
6	*F*
7	*M*

Tabella 4.4: Distribuzione unitaria di un carattere.

non consente, però, una comprensione immediata dei dati, allora costruiamo la distribuzione di frequenza (Tab. 4.5) che, invece, permette di riassumere i dati e di interpretarli più facilmente.

	Distribuzione di frequenza		
Sesso	Freq. assoluta	Freq. relativa	Freq. percentuale
X	$\mathbf{n_i}$	$\mathbf{f_i}$	$\mathbf{p_i}$
Maschio	4	0,57	57
Femmina	3	0,43	43
Totale (N)	7	1	100

Tabella 4.5: Esempio di distribuzione di frequenza di un carattere.

È immediato verificare, infatti, che gli individui, che rappresentano le unità di indagine, sono messi in relazione alle sole modalità con le quali si esprime il carattere "sesso". Sulla base di questo possiamo dire che i $\frac{57}{100}$ (cinquantasette centesimi) del totale sono maschi e i $\frac{43}{100}$ sono femmine o ancora, con maggiore facilità di comprensione, che il 57% del nostro collettivo è composto da maschi e che il restante 43% è costituito da femmine.

Quando il carattere sottoposto all'indagine è continuo, come ad esempio l'età, il peso, l'altezza, ecc., allora può essere conveniente

raggruppare le modalità in classi, che vengono propriamente dette *classi di modalità*. In questo caso la distribuzione viene chiamata *distribuzione di frequenza divisa in classi*. In altre parole il carattere viene suddiviso in opportuni intervalli di modalità la cui ampiezza viene stabilita a seconda delle esigenze dell'indagine che si sta effettuando. Schematicamente se il collettivo viene misurato nell'intervallo chiuso $[x_{min}; x_{max}]$ allora esso viene suddiviso opportunamente in k classi di ampiezza, non necessariamente uguali, nel modo seguente:

$$[(x_{min} \mapsto x_2), (x_2 \mapsto x_3), \ldots, (x_i \mapsto x_{i+1}), \ldots, (x_{k-1} \mapsto x_{max})]$$

Di conseguenza la distribuzione di frequenza assume la forma rappresentata nella tabella seguente:

Carattere X	Freq. Assoluta n_i
$x_{min} \mapsto x_2$	n_1
$x_2 \mapsto x_3$	n_2
$\vdots$	$\vdots$
$x_i \mapsto x_{i+1}$	n_i
$\vdots$	$\vdots$
$x_{k-1} \mapsto x_{max}$	n_k
Pop. totale	N

Tabella 4.6: Distribuzione di frequenza di un carattere quantitativo suddiviso in classi.

È necessario far notare che il simbolo $\mapsto$ indica che la classe è chiusa all'estremo inferiore, ad esempio se il carattere è l'età e la classe è $5 \mapsto 10$ anni allora apparterranno a questa classe tutti i bambini che hanno compiuto, alla data dell'indagine, 5 anni fino a quelli che non hanno compiuto 10 anni. Come esplicitato nella tabella che riassume il numero di studenti iscritti presso un Istituto Comprensivo raggruppati per classe di età (vd. Tabella 4.7).

In molti casi il carattere continuo di rilevazione non è limitato in un intervallo chiuso $[x_{min}; x_{max}]$ bensì l'intervallo di misurazione del carattere è aperto da $[-\infty; \infty]$. In questi casi è necessario fissare

Età alunni X	Numero alunni n_i
$3 \mapsto 5$	126
$5 \mapsto 10$	213
$10 \mapsto 15$	176
Tot. alunni	515

Tabella 4.7: Esempio di distribuzione di frequenza di un carattere quantitativo suddiviso in classi.

l'ampiezza della prima classe e quella dell'ultima. A questo proposito va notato che non esiste una regola standard per operare, ma la stessa viene lasciata al rilevatore che di volta in volta stabilirà, sulla base delle circostanze il valor di x_{min} e quello di x_{max}. Ulteriori ampliamenti in merito a questo argomento saranno trattati più nello specifico quando si tratteranno i *caratteri quantitativi*.

4.3 Metodi empirici per la suddivisione in classi di un carattere discreto o continuo

Il problema della determinazione del numero opportuno di classi in cui suddividere i dati, discreti o continui, di un determinato collettivo non è banale e non può avvenire in modo casuale. Occorre, infatti, considerare che un numero troppo ridotto di classi può determinare un eccessivo raggruppamento di dati, che determina consequenzialmente un'importante perdita di informazioni. Allo stesso modo, un numero eccessivo di classi può determinare l'inutilità di aver effettuato un raggruppamento.

Diversi sono stati negli anni i metodi empirici proposti per garantire un criterio di oggettività nella suddivisione in classi di un insieme di dati misurati su scala discreta o continua. Qui ricordiamo una metodologia che riteniamo particolarmente semplice ed intuitiva nell'utilizzo: il *metodo di Sturges*.

Il metodo proposto da H. Sturges, nel 1926, determina il numero ottimale di classi k in cui suddividere un carattere misurato su scala

discreta in relazione al numero di osservazioni effettuate, secondo la seguente formula:

$$k = 1 + \frac{10}{3} \cdot \log_{10}(N) \qquad (4.3.0.1)$$

dove k è il numero delle classi in cui suddividere la scala usata ed N è il numero complessivo delle osservazioni effettuate. Se si vuole utilizzare questo metodo si procederà come segue:

- dato un insieme di osservazioni $X = (x_1, x_2, \ldots, x_n)$, occorre distribuire i valori registrati su ogni unità statistica in ordine crescente (o decrescente);
- costruire la tabella della distribuzione di frequenza assoluta dei dati registrati;
- calcolare il *Range* (ossia il campo di variazione dei nostri valori che indichiamo con R) dei valori della distribuzione come differenza tra il valore massimo e il valore minimo registrato:

$$R = (x_{max} - x_{min}) = (x_n - x_1);$$

- calcolare, con l'applicazione della formula di Sturges, il numero ottimale di classi in cui suddividere l'intervallo, applicando le regole di arrotondamento nel caso in cui il risultato presenti dei decimali:

$$k = 1 + \frac{10}{3} \cdot \log_{10}(N);$$

- stabilire l'ampiezza della classe, W, secondo il rapporto tra il campo di variazione e il numero delle classi:

$$W = \frac{R}{K};$$

- occorre notare che, nel caso in cui il valore di W presenti dei numeri decimali, indipendentemente dalle regole di arrotondamento, si approssima al numero intero seguente (ad esempio se $W = 3,2$, allora si considera come ampiezza della classe il valore 4; se $W = 3,8$, allora si considera come ampiezza della classe sempre il valore 4);

- costruire le classi nel modo seguente: la prima classe avrà come estremo inferiore il valore minimo della distribuzione, $x_{min} = x_1$, e come estremo superiore il valore ottenuto sommando al valore minimo della distribuzione l'ampiezza della classe. Quest'ultimo valore sarà a sua volta l'estremo inferiore della seconda classe, al quale si aggiungerà il valore dell'ampiezza per ottenere il valore dell'estremo superiore della classe, e così via fino ad arrivare al valore massimo della distribuzione, $x_{max} = x_n$, o a superarlo leggermente.

  Per esempio, se $x_{min} = 18$, $x_{max} = 30$ e $N = 25$ avremo:

$$R = (x_{max} - x_{min}) = 12$$

$$k = 1 + \frac{10}{3} \cdot \log_{10}(25) = 5.66 \cong 6$$

$$W = \frac{R}{K} = \frac{12}{6} = 2$$

  allora dovremo costruire 6 classi di ampiezza 2 a partire dal valore 18.

  Sarà, quindi:

$$x_1 \mapsto x_2 = 18 \mapsto 20$$

$$x_2 \mapsto x_3 = 20 \mapsto 22$$

$$x_3 \mapsto x_4 = 22 \mapsto 24$$

$$x_4 \mapsto x_5 = 24 \mapsto 26$$

$$x_5 \mapsto x_6 = 26 \mapsto 28$$

$$x_6 \mapsto x_7 = 28 \mapsto 30$$

- ovviamente il numero di classi ottenuto deve corrispondere al numero di classi calcolato con la formula di Sturges.

Unità	Sex	Soddisfazione Abitazione	Età	Numero Figli	Reddito in euro
1	*M*	*Per niente soddisfatto*	*25*	*0*	12.000
2	*F*	*Molto soddisfatto*	*26*	*1*	21.000
3	*M*	*Abbastanza soddisfatto*	*26*	*2*	*32.000*
4	*M*	*Molto soddisfatto*	33	*2*	*15.000*
5	*M*	*Molto soddisfatto*	*35*	*3*	*20.000*
6	*F*	*Poco soddisfatto*	*36*	*2*	22.000
7	*F*	*Molto soddisfatto*	*26*	*2*	*22.000*
8	*M*	*Molto soddisfatto*	*31*	*1*	*26.000*

Tabella 4.8: Esempio di distribuzione di frequenza dei residenti in una piccola palazzina.

4.4 Riepilogando

Per riepilogare quanto è stato detto in questo capitolo facciamo qualche semplice esempio. Per i diversi esempi che proponiamo faremo riferimento alla tabella 4.8 di distribuzione unitaria multivariata.

Questa tabella sarà utilizzata anche nei capitoli seguenti ogni qualvolta si ravvederà la necessità di fare qualche esempio concreto. Procediamo, tuttavia, con ordine partendo da una breve analisi della tabella.

Nella prima colonna troviamo le *unità statistiche* che compongono il nostro *collettivo*, ossia la nostra *popolazione* oggetto di indagine; qui le unità statistiche sono state indicate con numeri in ordine crescente, ma la scelta di come identificare le unità è lasciata al ricercatore, che può liberamente decidere se utilizzare un sistema di codifica, un codice univoco che identifica ogni individuo, le iniziali del nome o ancora il nome per esteso.

Nella seconda colonna troviamo il *carattere qualitativo sconnesso* "Sesso", misurato su *scala nominale*, che si esprime con le *modalità* "M: maschio" e "F: femmina".

Nella terza colonna è inserito il *carattere qualitativo ordinabile* "Soddisfazione per l'abitazione", misurato su *scala ordinale*, che si esprime con

le *modalità* "Per niente soddisfatto", "Poco soddisfatto", "Abbastanza soddisfatto", "Molto soddisfatto" .

Nella quarta colonna è posto il *carattere quantitativo continuo* "Età", che si esprime con diverse *modalità* le quali indicano l'età cronologica espressa in anni.

Nella penultima colonna troviamo il *carattere quantitativo discreto* "Numero di figli", che si esprime con le *modalità* "0, 1, 2, . . . ".

Nell'ultima colonna, infine, troviamo il *carattere quantitativo continuo* "Reddito espresso in euro", che si può esprimere con le tutte le possibili *modalità* che indicano il valore del reddito.

Per ognuno dei caratteri presi in considerazione possiamo costruire la relativa *tabella di frequenza* in modo da avere informazioni più immediate sul collettivo preso in esame. Indipendentemente dal tipo di carattere preso in considerazione il procedimento che riguarda la strutturazione della tabella di frequenza è pressoché identico; per questo motivo illustreremo solo alcuni esempi, lasciando poi al lettore la libertà di concludere il lavoro per ogni carattere.

1. Carattere qualitativo sconnesso: "Sesso"

   - Prima di tutto si procede alla compilazione della colonna della *Frequenza assoluta*, conteggiando quanti "Maschi" e quante "Femmine" costituiscono il collettivo ed effettuando la somma totale delle due categorie prese in esame, per verificare di aver preso in considerazione tutti gli individui. Nel nostro esempio gli individui di sesso maschile risultano essere 5 e quelli di sesso femminile sono 3, quindi, riprendendo un po' di simbologia convenzionale sarà: $n_M = 5$ e $n_F = 3$, da cui $N = \sum_{i=1}^{k} n_i = 5 + 3 = 8$. Con un carattere qualitativo sconnesso, come quello preso in esame, possiamo effettuare una categorizzazione che ci consente di stabilire, appunto, quanti individui appartengono all'una o all'altra categoria. L'operazione di categorizzazione, essendo valida, per questo tipo di carattere che presenta una bassissima potenzialità operativa, sarà valida per tutti gli altri caratteri statistici poiché, come abbiamo specificato nel paragrafo 2.3, esiste tra i caratteri statistici una sorta di

gerarchia operativa. La frequenza assoluta evidenzia, appunto, il dato assoluto delle volte che una specifica modalità si ripete nel collettivo, ma non ci dà nessuna indicazione di questo dato in riferimento al collettivo stesso.

- La frequenza relativa consente, invece, di fare un rapporto tra il numero di volte che è ripetuta una modalità rispetto al collettivo oggetto di indagine. Per ogni modalità si procederà, quindi, a dividere la rispettiva frequenza assoluta per il totale in quanto $f_i = \frac{n_i}{N}$. In questo modo il totale del collettivo rappresenterà un intero frazionato in tante parti quante sono le modalità; questo è il motivo per cui la somma delle frequenze relative è sempre uguale a 1.
- La frequenza percentuale, infine, si ottiene moltiplicando per 100 il valore della frequenza relativa. La frequenza percentuale è sicuramente quella che risulta più facilmente fruibile anche da utenti meno esperti poiché largamente usata nel linguaggio comune.

	Distribuzione di frequenza		
Sesso	Freq. assoluta	Freq. relativa	Freq. percentuale
X	n_i	f_i	p_i
Maschio	5	0,625	62,5
Femmina	3	0,375	37,5
Totale(N)	8	1	100

Tabella 4.9: Esempio distribuzione di frequenza: carattere qualitativo sconnesso.

2. Carattere qualitativo ordinabile: "Soddisfazione per l'abitazione"

3. Carattere quantitativo discreto: "Numero di figli"

4. Carattere quantitativo continuo: "Reddito in euro"

   Utilizzando il metodo di Sturges, procediamo alla suddivisione in classi. Sarà

	Distribuzione di frequenza		
Soddisf.	Freq. assoluta	Freq. relativa	Freq. percentuale
X	n_i	f_i	p_i
Per niente	1	0,125	12,5
Poco	1	0,125	12,5
Abbast.	1	0,125	12,5
Molto	5	0,625	62,5
Totale(N)	8	1	100

Tabella 4.10: Esempio distribuzione di frequenza: carattere qualitativo ordinabile.

	Distribuzione di frequenza		
Num. di figli	Freq. assoluta	Freq. relativa	Freq. percentuale
X	n_i	f_i	p_i
0	1	0,125	12,5
1	2	0,25	25
2	4	0,5	50
3	1	0,125	12,5
Totale(N)	8	1	100

Tabella 4.11: Esempio distribuzione di frequenza: carattere quantitativo discreto.

$x_{min} = 12.000$, $x_{max} = 32.000$ e $N = 8$, quindi avremo:

$$R = (x_{max} - x_{min}) = 20.000$$

$$k = 1 + \frac{10}{3} \cdot \log_{10}(8) = 4,01 \cong 4$$

$$W = \frac{R}{K} = \frac{20.000}{4} = 5.000$$

allora dovremo costruire 4 classi di ampiezza 5.000 a partire dal valore 12.000.

Sarà, quindi:

$$x_1 \mapsto x_2 = 12.000 \mapsto 17.000$$

$$x_2 \mapsto x_3 = 17.000 \mapsto 22.000$$

$$x_3 \mapsto x_4 = 22.000 \mapsto 27.000$$

$$x_4 \mapsto x_5 = 27.000 \mapsto 32.000$$

	Distribuzione di frequenza		
Reddito in euro	Freq. assoluta	Freq. relativa	Freq. percentuale
X	n_i	f_i	p_i
$12.000 \mapsto 17.000$	2	0,25	25
$17.000 \mapsto 22.000$	2	0,25	25
$22.000 \mapsto 27.000$	3	0,375	37,5
$27.000 \mapsto 32.000$	1	0,125	12,5
Totale(N)	8	1	100

Tabella 4.12: Esempio distribuzione di frequenza: carattere quantitativo continuo in classi.

Capitolo 5

Analisi univariata di caratteri statistici misurati su scala di diversa natura

In questo paragrafo ed in quelli che seguiranno svilupperemo, dal punto di vista applicativo, le procedure metodologiche necessarie per raggiungere gli obiettivi che ci siamo prefissati. Tuttavia prima di procedere dal punto di vista formale, ci sembra necessario precisare cosa prevede il protocollo del metodo statistico. Dato un fenomeno reale composto da un collettivo di N unità e supposto che lo stesso collettivo sia stato misurato attraverso una delle misure introdotte, come più volte detto, è immediato ricavare la distribuzione di frequenza, che si assume essere la base dei dati per avviare il procedimento del metodo statistico. In linea di principio generale il metodo statistico consiste nella ricerca delle seguenti tre elementi:

- *sintesi*;
- *variabilità*;
- *forma*.

Oltre a questi elementi, come vedremo, è molto utile associare uno studio basato sul *confronto* tra due o più distribuzioni relative ad uno stesso fenomeno reale o alla stessa distribuzione nel tempo. D'altra

parte, quante volte nella vita ci è capitato di confrontarci con noi stessi o con altre persone? Pensiamo ad esempio a tutte le volte che la nostra mamma ha misurato la nostra altezza per vedere se eravamo cresciuti; oppure basti pensare a tutte le volte che, per sfida tra coetanei, si confrontava l'altezza propria con quella di un altro bambino.

Nella vita reale i confronti servono a cogliere le differenze tra individui, società, nazioni e così via. Un modo per poter cogliere tali differenze deriva dalle necessità dello studio che si vuole fare. Se ad esempio fossimo interessati a verificare l'apprendimento di una classe, allora potremmo agire secondo due modalità: sulla stessa classe possiamo somministrare due questionari di valutazione, ad esempio uno prima e l'altro al termine di uno specifico percorso formativo, per verificare in che modo siano modificate conoscenze, abilità e comportamenti; oppure si può effettuare uno stesso test di apprendimento su due classi diverse e paragonare i risultati ottenuti. A questo punto si può notare una somiglianza di procedura tra l'esempio della misurazione dell'altezza e quello della valutazione di apprendimento di una classe: i confronti si effettuano o a distanza di tempo sullo stesso individuo/classe, oppure in uno stesso istante temporale tra due individui/classi diverse.

Non è banale sottolineare che le differenze tra le due procedure di confronto sono molto diverse poiché nel primo esempio si valuta il confronto tra due bambini, nel secondo il confronto viene effettuato tra due gruppi di bambini. Quest'ultimo aspetto è molto importante poichè, se si lavora tra collettivi e si vuole eseguire un'analisi per confrontare le caratteristiche tra due gruppi di unità statistiche, è necessario scegliere un criterio per poter rendere confrontabili i due collettivi oggetto di studio. Il primo esempio è il caso più semplice in quanto, una volta rilevate le due altezze allora si può affermare quale dei due bambini è più alto; il secondo esempio, invece, fornisce due distribuzioni dei punteggi dei test di apprendimento. Sulla base di cosa si può affermare se le due classi hanno un diverso livello di apprendimento? Quanto sono diversi i due livelli di apprendimento? Occorre scegliere un criterio che ci permetta di fare tali affermazioni. In statistica non vi è mai una scelta giusta o sbagliata degli strumenti statistici da utilizzare per prendere decisioni, ma esiste una scelta migliore, ottimale, da adoperare per raggiungere determinati obiettivi.

5.1 Sintesi e variabilità

In generale per ricerca della *sintesi*, come anticipato, si intende la ricerca della modalità più rappresentativa della distribuzione, ossia quel valore che, più di altri, è nelle condizioni di rappresentare il collettivo di riferimento. Il concetto di sintesi è di fatto innato nell'essere umano. In diverse circostanze siamo abituati a esprimerci con sentenze categoriche su fenomeni che invece hanno un pluralità di eventi. Ad esempio l'insegnante spesso si pronuncia nei confronti di una sua classe: *"La 3ª A è una classe eccellente"* oppure *"In 1ª B non hanno le basi per seguire la mia lezione"*. Nel primo caso sembrerebbe che tutti gli alunni della 3ª A siano eccellenti e che nessuno della 1ª B abbia le basi per seguire la lezione del docente. In realtà non è così, in quanto sia nell'una quanto nell'altra classe si possono incontrare situazioni diverse da quelle sentenziate. Ad esempio in 3ª A, pur essendoci diversi casi di eccellenza, si potrebbero incontrare anche alunni di un più basso profilo che comunque non fanno cambiare l'opinione generale dell'insegnante.

Facciamo un esempio in campo economico. In economia, a proposito dell'inflazione, l'Istituto di Statistica periodicamente comunica allarmato che il tasso di inflazione riferito ad un determinato istante temporale è per esempio del 3,1% . Anche in questo caso sembrerebbe che tutti i beni siano aumentati della stessa percentuale. Ma il danno economico dell'inflazione è tanto maggiore quanto più la variazione dei prezzi è differente. Il dato pubblicizzato rappresenta il valore di sintesi, ossia il più indicativo di un andamento generale dei prezzi.

Naturalmente, per quanto appena detto, la sintesi di una distribuzione non è in grado di dare tutte le informazione necessarie sulla distribuzione. Occorre, quindi, valutare quanto la sintesi di una distribuzione sia rappresentativa della distribuzione stessa, introducendo un criterio di valutazione che sia in grado di cogliere le diversità manifestate dalle singole unità statistiche nel collettivo di riferimento. In altri termini, è necessario introdurre una misura del grado di *variabilità*.

La variabilità esprime l'attitudine di una misura ad assumere le proprie modalità. Gli indici di variabilità assumeranno valori prossimi allo zero quando nella distribuzione tutte le unità assumeranno valori

molto simili e crescerà man mano che la diversità tra le modalità aumenta. Si intuisce come la misura della variabilità sia altamente informativa al fine di stabilire la significatività del valore di sintesi. Una distribuzione con un valore di variabilità modesta rafforza il significato di rappresentatività della sintesi, nel senso che in una tale circostanza si assisterà ad una distribuzione con valori molto simili e prossimi a quelli di sintesi. Al contrario una distribuzione con alto valore di variabilità indicherà che vi sono unità che assumono valori molto diversi tra loro, ciò implica che il valore di sintesi non è in grado pienamente di rappresentare l'intero collettivo.

In questo senso, rispettando gli esempi fatti in precedenza, si intuirà che l'affermazione di eccellenza della classe è tanto più vera quanto più in essa saranno prevalenti i casi di eccellenza; l'inflazione sarà tanto più dannosa, da un punto di vista economico, quanto più la variazione dei prezzi è diversa tra bene e bene. Quindi la considerazione congiunta del valore di sintesi con quello di variabilità, permettono al ricercatore di possedere elementi di valutazione più completi sul fenomeno reale che si sta studiando.

5.2 Forma

Oltre alla sintesi e alla variabilità, è interessante conoscere la *forma* della distribuzione dei dati. Ossia il modo in cui le unità del collettivo si distribuiscono tra le diverse modalità del carattere misurato. In un collettivo è interessante conoscere come si comportano le unità rispetto al valore di sintesi e di variabilità.

Due sono gli aspetti fondamentali nello studio della forma di una distribuzione: la *simmetria* e il comportamento della distribuzione alle code (la *curtosi*).

Una prima valutazione della forma è data, quindi, dalla *Simmetria*, ossia dal comportamento della distribuzione rispetto alla modalità di sintesi. Cosa diversa è sapere che in una classe, giudicata eccellente, vi sono pochi casi di eccellenza e molti casi di mediocrità, piuttosto che molti casi di eccellenza e pochi elementi di mediocrità.

Una distribuzione, come approfondiremo meglio in seguito trattando la forma più nello specifico, potrà dimostrare asimmetria negativa o

positiva, a seconda che siano più presenti nella distribuzione modalità più grandi o più piccole.

I grafici A, B, C, riportati di seguito, aiutano a comprendere in modo intuitivo i concetti di simmetria/asimmetria. Nel caso C la

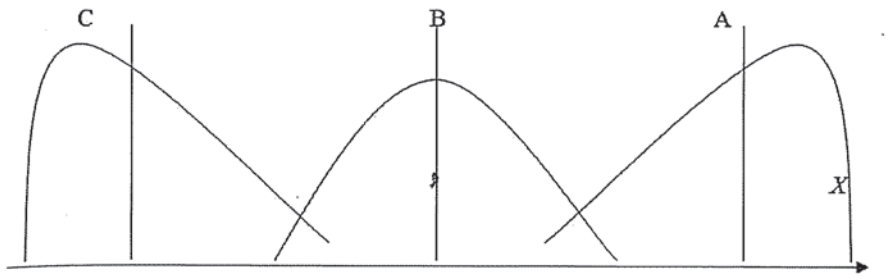


Figura 5.1: Confronto simmetria/asimmetria.

distribuzione mette in evidenza unità con modalità prevalentemente basse, in questo caso si dirà che c'è un *asimmetria positiva*.

Contrariamente, nel caso A la distribuzione manifesta una prevalenza di modalità con valori alti, si dirà, allora, che c'è *asimmetria negativa*.

La distribuzione B, infine, si equidistribuisce tra i valori minori e maggiori della sintesi e in questo caso si dirà che la distribuzione è *simmetrica*.

Anticipiamo, ancora, che un'altra valutazione della forma, che tratteremo nel dettaglio nei prossimi capitoli, è lo studio della *Curtosi*, ossia la valutazione del comportamento della distribuzione alle code e nella zona centrale. La curtosi, infatti, fornisce una valutazione sulle frequenze corrispondenti alle modalità estreme e centrali rispetto a quelle mediane di una distribuzione unimodale e simmetrica.

L'obiettivo è quello di valutare appunto il comportamento della distribuzione nei valori estremi. In altri termini si vuole stabilire il peso che hanno le modalità molto basse e quelle molto alte rispetto a quelle centrali. Nell'esempio della classe riportata sopra si vuole,

quindi, sapere se il numero degli eccellenti e di coloro con scarsa abilità scolastica rappresentano o meno un'elevata quota di alunni.

Di seguito il concetto di curtosi è stato rappresentato graficamente confrontando tre tipi di curve. In particolare una curtosi alta è rappresentata da valori estremi alti (curva C), mentre una curtosi bassa da valori estremi bassi (curva A). Lo studio nella forma di una distribu-

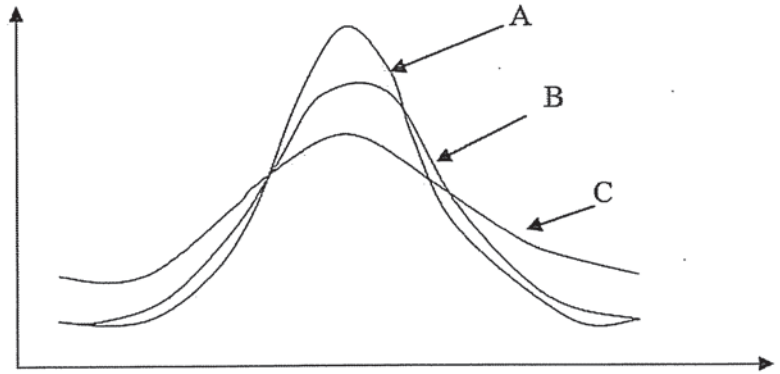


Figura 5.2: Confronto tra livelli diversi di Curtosi.

zione si conclude con la ricerca di un modello, in genere matematico, che ha l'obiettivo di descrivere il fenomeno reale attraverso una regola ovvero una norma che sia in grado di sostituire lo studio del fenomeno attraverso le misure della caratteristica del collettivo.

Capitolo 6

La valutazione di un carattere misurato su scala nominale

6.1 Sintesi

Da quanto appreso nei precedenti capitoli, se un carattere è di tipo qualitativo sconnesso viene misurato attraverso una scala nominale. L'unica operazione che possiamo effettuare con questa tipologia di caratteri è quella di stabilire se tra le unita del collettivo ci sia diversità o uguaglianza, ossia si possono compiere solo dei confronti. In queste circostanze ci chiediamo: Come il metodo statistico procede per la ricerca della sintesi della variabilità e della forma?. In realtà, nella trattazione di questo tipo di carattere scopriremo che, proprio per la sua specifica caratteristiche di non ordinabilità, non sarà possibile studiare la forma della distribuzione di un carattere qualitativo sconnesso.

6.1.1 La moda

Da un punto di vista metodologico la *moda* è un indice di posizione che può essere individuato per qualsiasi tipo di carattere ed è il solo che può essere utilizzato per i caratteri qualitativi sconnessi.

La *moda* è, tra tutte le modalità osservate, *quella modalità che presenta la frequenza più alta nel collettivo osservato.*

Da un punto di vista interpretativo, il concetto di moda è molto intuitivo poiché si può associare all'uso che si fa del termine nel linguaggio comune: ad esempio, dire che quest'anno vanno di moda i jeans a vita bassa, sta a significare che la maggioranza dei giovani porta jeans che presentano questa caratteristica. Formalmente si avrà che nella distribuzione di frequenze di un collettivo di giovani in relazione al tipo di pantaloni indossati, si verificherà che la modalità "jeans" avrà la massima frequenza e quindi rappresenterà la *moda* della distribuzione.

Diremo, quindi, che la sintesi di un carattere qualitativo sconnesso misurato su scala nominale è *la moda, ossia la modalità prevalente del carattere, ovvero quella a cui è associata la massima frequenza.*

Per comprendere meglio quanto detto procediamo ad un semplice esempio.

	Frequenze		
Risultato scolastico	*Assolute*	*Relative*	*Percentuali*
Promossi	**137**	**0,61**	**60,9**
Rimandati	65	0,29	28,9
Bocciati	23	0,10	10,2
Totale	225	1	100

Tabella 6.1: Esempio di distribuzione di frequenza.

La moda della nostra distribuzione è la modalità "Promossi" che è presente in 137 unità ossia al 60.9% degli studenti della scuola.

Si possono presentare, tuttavia, casi di ambiguità nello stabilire un valore di sintesi univoco di una distribuzione in quanto si possono verificare distribuzioni che non presentano una sola modalità prevalente, ma due o più di esse. È immediato intuire come, in questi casi, il valore della sintesi perda il suo significato più puro del termine. Definiamo la distribuzione che presenta due modalità con lo stesso maggior numero di frequenze, quindi, come *distribuzione bimodale;*

quella che presenta più modalità con lo stesso maggiore numero di frequenze come *plurimodale.*

Tuttavia, come vedremo più avanti, queste situazioni di ambiguità saranno messe in evidenza dalla ricerca della variabilità.

6.2 Variabilità

In una distribuzione statistica possiamo trovare casi in cui alcune modalità presentano elevate frequenze e le altre basse, in altri casi, invece, tutte le modalità possono presentare la stessa frequenza. Diciamo che l'attitudine di un carattere ad assumere diverse modalità si dice generalmente *variabilità*, anche se con i caratteri qualitativi sarebbe più corretto parlare di *mutabilità.*

Il concetto di variabilità di un carattere qualitativo sconnesso misurato su scala nominale passa attraverso due concetti: quello dell'*omogeneità* e quello dell'*eterogeneità*. Non potendo, infatti operare algebricamente tra le modalità ci si potrebbe chiedere se nel collettivo esaminato c'è una convergenza di comportamento delle unità su una o poche modalità, in tal caso diremo che il collettivo è omogeneo (*omogeneità*), oppure se le unità si equidistribuiscono tra la pluralità delle modalità con cui il carattere è stato misurato e in tal caso diremo che il collettivo è eterogeneo (*eterogeneità*). La variabilità è tanto maggiore quanto più sono grandi le differenze che i singoli casi individuali presentano tra loro oppure rispetto ad un valore caratteristico del fenomeno considerato.

Per chiarirci meglio le idee facciamo un semplice esempio molto intuitivo; consideriamo tre differenti distribuzioni univariate relative al colore dei capelli:

d_1 *(rilevazione del colore dei capelli nel gruppo A):*

biondi – cenere – bianchi – castani – neri – rossi

d_2 *(rilevazione del colore dei capelli nel gruppo B):*

neri – neri – neri – neri – neri – neri

d_3 *(rilevazione del colore dei capelli nel gruppo C):*

biondi – biondi – biondi – neri – neri – neri

Nel caso del gruppo con distribuzione d_2 si comprende subito che la variabilità è nulla poiché tutte le modalità sono uguali tra loro e quindi la distribuzione presenta massima omogeneità; nella distribuzione d_1, al contrario, c'è un'altissima variabilità poiché tutte le modalità sono diverse tra loro, quindi possiamo dire che c'è massima eterogeneità; nel caso della distribuzione d_3 è opportuno individuare quale sia il livello di omogeneità e/o eterogeneità.

In linea di principio diremo che *un collettivo, misurato su una scala nominale, è omogeneo rispetto ad un carattere se tutte le unità presentano la stessa modalità.*

6.2.1 Omogeneità ed eterogeneità assoluta e relativa

È immediato intuire che molto difficilmente un collettivo può essere considerato omogeneo nel senso assoluto così come appena definito. Solitamente si presentano casi in cui si rende necessario misurare il grado di omogeneità della distribuzione o in caso contrario di eterogeneità. In altri termini, si rende necessario trovare una misura di omogeneità o di eterogeneità. L'omogeneità e l'eterogeneità sono, quindi, due concetti complementari poiché quando ci sarà alta omogeneità, di conseguenza, registreremo una bassa eterogeneità e viceversa.

In termini generali diremo che un indice di omogeneità assumerà il suo valore massimo in caso di perfetta omogeneità, ossia quando valgono le condizioni sopra indicate, ossia che tutte le unità assumono la stessa modalità (v. Tab. 6.2); assumerà viceversa il suo minimo quando il collettivo è perfettamente eterogeneo, ossia quando tutte le unità si *equidistribuiscono* tra tutte le possibili modalità con cui il carattere può essere misurato. Da un punto di vista operativo, data una distribuzione si avrà massima omogeneità (minima eterogeneità) se una sola frequenza relativa è diversa da zero o, se preferiamo, solo una frequenza relativa vale 1 e tutte le altre valgono 0.

Al contrario, in una distribuzione si avrà minima omogeneità (massima eterogeneità) se tutte le modalità hanno un valore pari alla frequenza media $n = \frac{N}{k}$, dove N rappresenta il totale del collettivo e k il numero di modalità. Conseguentemente la frequenza relativa

Carattere X	Numero n_i	Frequenza relativa f_i
x_1	0	0
x_2	0	0
$\vdots$	$\vdots$	$\vdots$
x_i	n_i	1
$\vdots$	$\vdots$	$\vdots$
x_k	0	0
	$N = \sum_{i=1}^{k} n_i$	1

Tabella 6.2: Caso di massima omogeneità (minima eterogeneità).

media è $f = \frac{1}{k}$ (v. Tab. 6.3). Si intuisce che un indice di omogeneità

Carattere X	Numero n_i	Frequenza relativa f_i
x_1	n	$\frac{1}{k}$
x_2	n	$\frac{1}{k}$
$\vdots$	$\vdots$	$\vdots$
x_i	n	$\frac{1}{k}$
$\vdots$	$\vdots$	$\vdots$
x_k	n	$\frac{1}{k}$
	N	1

Tabella 6.3: Caso di minima omogeneità (massima eterogeneità).

ed il complementare indice di eterogeneità si basano su opportune operazioni sulle frequenze relative. In particolare, un primo indice assoluto di omogeneità è dato dalla somma delle frequenze relative al quadrato.

$$OM_1 = f_1^2 + f_2^2 + \ldots + f_k^2 = \sum_{i=1}^{k} f_i^2 \qquad (6.2.1.1)$$

L'opportunità di porre le frequenze al quadrato ha una duplice ragione. La prima, di carattere operativo, risiede nel fatto che la semplice somma delle frequenze relative risulterebbe sempre e comunque

uguale ad 1, così come evidenziato nei capitoli precedenti, indipendentemente dal valore delle frequenze stesse; la seconda ragione, di carattere più metodologico, risiede nel fatto che l'elevazione a potenza di una frazione tende a distinguere i valori prossimi all'unità (caso di massima omogeneità) rispetto a quelli prossimi allo zero, in altri termini l'elevazione a potenza ha la stessa funzione di una lente di ingrandimento per l'effetto dell'omogeneità.

È opportuno far notare che in caso di massima omogeneità l'indice OM_1 vale sempre 1 infatti

$$max(OM_1) = 0^2 + 0^2 + \ldots + 1 + \ldots + 0^2 = 1$$

Mentre in caso di minima omogeneità esso vale $\frac{1}{k}$

$$min(OM_1) = \left(\frac{1}{k}\right)^2 + \left(\frac{1}{k}\right)^2 + \ldots + \left(\frac{1}{k}\right)^2 = k\left(\frac{1}{k}\right)^2 = k\frac{1}{k^2} = \frac{1}{k}$$

quindi il campo di variazione dell'indice assoluto di omogeneità OM_1 sarà: $\frac{1}{k} \leqslant OM_1 \leqslant 1$.

Un secondo indice di omogeneità si ottiene come prodotto delle frequenze relative elevate per se stesse.

$$OM_2 = f_1^{f_1} \cdot f_2^{f_2} \cdot \ldots \cdot f_k^{f_k} = \prod_{i=1}^{k} f_i^{f_i} \qquad (6.2.1.2)$$

Come meglio potremo vedere più avanti quest'ultimo indice si ottiene come media geometrica delle frequenze relative associate a ciascuna modalità ponderata con le corrispondenti frequenze assolute. Questa operazione, allo steso modo di prima, ha come obiettivo quello di amplificare l'effetto della concentrazione delle unità su poche modalità.

In caso di massima omogeneità l'indice vale 1. Infatti [1]

$$max(OM_2) = 0^0 \cdot 0^0 \cdot \ldots \cdot 1^1 + \ldots + 0^0 = 1$$

Mentre il minimo di omogeneità vale ancora una volta $\frac{1}{k}$. Infatti

$$min(OM_1) = \left(\frac{1}{k}\right)^{\frac{1}{k}} \cdot \left(\frac{1}{k}\right)^{\frac{1}{k}} \cdot \ldots \cdot \left(\frac{1}{k}\right)^{\frac{1}{k}} =$$

[1] Si deve ricordare che $0^0 = 1$ perchè $\lim_{x \to 0} x^x = 1$

$$= \left(\frac{1}{k}\right)^{\frac{1}{k}+\frac{1}{k}+\ldots+\frac{1}{k}} = \left(\frac{1}{k}\right)^1 = \frac{1}{k}$$

La difficoltà di operare con prodotti e potenze di frazioni, suggerisce una semplificazione operativa. Con questo fine si introduce il terzo indice di omogeneità ottenuto come logaritmo in base a di OM_2. Ricordiamo che la funzione logaritmo in base a è la funzione inversa rispetto alla funzione esponenziale in base a. Si dice, quindi, logaritmo in base a di un numero x l'esponente da dare ad a per ottenere x (x viene chiamato *argomento del logaritmo*). In altre parole, se

$$x = a^y$$

segue che:

$$y = \log_a x$$

(si legge: y è il logaritmo in base a di x).

Per esempio, $\log_3 81 = 4$ perché $3^4 = 81$.

Il logaritmo è utile soprattutto perché trasforma prodotti in somme, i rapporti in differenze, elevamenti a potenza in moltiplicazioni e radicali in divisioni.

Un terzo indice di omogeneità (OM_3) si otterrà, quindi, come logaritmo in base a, con $a > 1$, dell'indice di omogeneità OM_2:

$$\log_a OM_2 = OM_3 = f_1 \log_a f_1 + f_2 \log_a f_2 + \ldots + f_k \log_a f_k \qquad (6.2.1.3)$$

da cui si ricava che

$$OM_3 = \sum_{i=1}^{k} f_i \log_a f_i \qquad (6.2.1.4)$$

Il massimo valore dell'indice OM_3 è 0, infatti

$$max(OM_3) = \log_a 1 = 0$$

mentre il minimo di OM_3 è $-\log_a k$, infatti

$$min(OM_3) = \frac{1}{k}\log_a\left(\frac{1}{k}\right) + \frac{1}{k}\log_a\left(\frac{1}{k}\right) + \ldots + \frac{1}{k}\log_a\left(\frac{1}{k}\right)$$

Ricordando che $\log_a \frac{1}{k} = -\log_a k$ si ha

$$min(OM_3) = \frac{-\log_a k}{k} + \frac{-\log_a k}{k} + \ldots + \frac{-\log_a k}{k} = -\log_a k$$

C'è da notare che gli indici di omogeneità che abbiamo introdotto sono espressi in valore assoluto, nel senso che hanno un proprio campo di variazione e dipendono dal numero di modalità del carattere sotto osservazione. Per avere una maggiore capacità informativa ed una immediatezza interpretativa, si introdurranno tra breve i corrispondenti indici relativi.

6.3 Gli indici relativi

Prima di procedere nello studio degli indici relativi di omogeneità, riteniamo utile specificare meglio il significato di indice relativo, concetto che come vedremo sarà ripreso più volte in avanti.

In genere un indice usato per interpretare un aspetto della distribuzione viene calcolato sulle modalità del carattere oggetto di studio. Questo implica che l'indice sarà espresso con la misura del carattere considerato. Ad esempio, se facciamo la media delle stature espresse in centimetri di tre bambini, la statura media risultante è una statura ancora espressa in centimetri. Ciò implica che l'indice ottenuto non può essere confrontato con un'altra misura per esempio il peso dei tre bambini, così come non può essere paragonato con altri collettivi per esempio le stature dei genitori.

La soluzione a questo problema è data dalla introduzione degli indici relativi. Diremo che *un indice è relativo se il suo campo di variazione è compreso tra* 0 *e* 1 *e se è un numero puro ossia privo di dimensione o, se vogliamo, espresso in nessuna delle scale di misura introdotte precedentemente.*

Per ottenere un indice con queste caratteristiche, si procede introducendo il seguente principio generale:

$$I_{rel} = \frac{I_{ass} - min(I_{ass})}{max(I_{ass}) - min(I_{ass})} \qquad (6.3.0.5)$$

Nel caso specifico dell'indice di omogeneità OM_1, per esempio, il valore minimo dell'indice assoluto sarà $min(I_{ass}) = \frac{1}{k}$, mentre il massimo sarà $max(I) = 14$, quindi applicando il citato principio generale:

$$I_{min(rel)} = \frac{\frac{1}{k} - \frac{1}{k}}{1 - \frac{1}{k}} = \frac{0}{1 - \frac{1}{k}} = 0$$

il valore massimo che può assumere l'indice di omogeneità relativo sarà:

$$I_{max(rel)} = \frac{1 - \frac{1}{k}}{1 - \frac{1}{k}} = 1$$

L'indice di omogeneità relativo, quindi, sarà sempre un numero puro, non affetto dall'unità di misura presa in considerazione, compreso tra i valori 0 e 1.

Allora sarà:

$$0 \leqslant I_{rel} \leqslant 1$$

Per comprendere meglio i passaggi appena visti, precisiamo che I indica l'indice, il pedice rel indica relativo, il pedice ass sta per assoluto, min e max indicano rispettivamente il valore minimo e il valore massimo che può assumere l'indice assoluto.

È piuttosto facile verificare che il principio appena introdotto rispetta le caratteristiche richieste all'indice relativo. Infatti nel caso che l'indice assoluto calcolato sulla distribuzione assuma valori molto vicini al minimo, allora il numeratore diventerà sempre più piccolo fino ad assumere il valore 0 quando l'indice sarà talmente piccolo da coincidere con il suo valore minimo. Al contrario se l'indice è molto grande al più può raggiungere il massimo valore possibile e in tal caso il numeratore coinciderà con il denominatore, quindi l'indice relativo assumerà valore pari a 1.

L'indice relativo è un numero puro in quanto rappresenta un rapporto tra due quantità espresse nella stessa unità di misura, ciò implica che l'effetto della misura si semplifica, rendendo il numero privo di dimensione così come desiderato.

È necessario puntualizzare che l'indice relativo permette di dare un valore interpretativo all'indice proposto, in quanto indica nella scala dei valori tra 0 e 1 quanta parte di esso è stato raggiunto dal collettivo studiato.

6.4 Gli indici relativi di omogeneità e di eterogeneità

Tornando alla misura di omogeneità, una volta noti i valori massimi e minimi che può assumere una distribuzione, è immediato

trovare le corrispondenti misure relative, infatti, indicando con le lettere minuscole dell'alfabeto gli indici relativi, si ha:

$$om_1 = \frac{\sum_{i=1}^{k} f_i^2 - \frac{1}{k}}{1-\frac{1}{k}} = \frac{k\sum_{i=1}^{k} f_i^2 - 1}{k-1} \tag{6.4.0.6}$$

$$om_2 = \frac{\prod_{i=1}^{k} f_i^{f_i} - \frac{1}{k}}{1-\frac{1}{k}} = \frac{k\prod_{i=1}^{k} f_i^{f_i} - 1}{k-1} \tag{6.4.0.7}$$

$$om_3 = \frac{\sum_{i=1}^{k} f_i \log_a f_i + \log_a k}{\log_a k} = 1 + \sum_{i=1}^{k} f_i \frac{\log_a f_i}{\log_a k} \tag{6.4.0.8}$$

Per ragioni computazionali abbiamo per prima introdotto le misure di omogeneità anche se, ai fini della ricerca della variabilità il concetto di eterogeneità sembra essere più idoneo a catturare la diversità esistente tra le modalità.

Tuttavia, dalla definizione data sopra è immediato ricavare i corrispondenti indici assoluti e relativi di eterogeneità in quanto l'eterogeneità è il complementare dell'omogeneità.

Il primo indice di eterogeneità assoluto si ricava, appunto, come differenza da 1 di OM_1.

$$ET_1 = 1 - OM_1 = 1 - \sum_{i=1}^{k} f_i^2 \tag{6.4.0.9}$$

Questo indice è conosciuto anche come *indice di eterogeneità di Gini.*

Il secondo come rapporto a 1 di OM_2

$$ET_2 = \frac{1}{OM_2} = \frac{1}{\prod_{i=1}^{k} f_i^{f_i}} \tag{6.4.0.10}$$

Il terzo come logaritmo di ET_2

$$ET_3 = \log_a ET_2 = -\sum_{i=1}^{k} f_i \log_a f_i \tag{6.4.0.11}$$

Eterogeneità assoluta	Minima eterogeneità	Massima eterogeneità	Indice relativo di eterogeneità
$ET_1 = 1 - \sum_{i=1}^{k} f_i^2$	$\min(ET_1) = 0$	$\max(ET_1) = \frac{k-1}{k}$	$et_1 = \frac{k}{k-1}\left[1 - \sum_{i=1}^{k} f_i^2\right]$
$ET_2 = \frac{1}{\prod_{i=1}^{k} f_i^{f_i}}$	$\min(ET_2) = 1$	$\max(ET_2) = k$	$et_2 = \frac{1-\prod_{i=1}^{k} f_i^{f_i}}{(k-1)\prod_{i=1}^{k} f_i^{f_i}}$
$ET_3 = -\sum_{i=1}^{k} f_i \log_a f_i$	$\min(ET_3) = 0$	$\max(ET_3) = \log_a k$	$et_3 = -\sum_{i=1}^{k} f_i \frac{\log_a f_i}{\log_a k}$

Tabella 6.4: Indici di eterogeneità ed intervalli di variazione.

Dopo opportune operazioni algebriche, si possono ricavare i corrispondenti intervalli di variazione degli indici di eterogeneità che, per sinteticità, sono riportati in tabella 6.4. È importante sapere che l'indice ET_3 ed et_3 sono anche noti, in statistica, come indice di *entropia assoluta* e indice di *entropia relativa di Shannon*.

6.5 L'indice di dissomiglianza

In particolare, date due distribuzioni distinte di uno stesso carattere nominale possiamo effettuare un confronto attraverso un indice che ci dica come differiscono tra loro, ossia quanto sono dissomiglianti. Date due distribuzioni secondo un medesimo carattere, che indicheremo con A e B, con f_{iA} e f_{iB}, rispettivamente, le frequenze relative dell'i-esima modalità della prima e della seconda distribuzione, l'indice di dissomiglianza è rappresentato da una funzione dei valori assoluti degli scarti tra f_{iA} e f_{iB}.

Vediamo meglio come si deve procedere nel confronto proponendo alcuni esempi. Prendiamo, quindi le nostre due distribuzioni di frequenza relativa A e B degli iscritti alle diverse facoltà, che indicheremo con f_{iA} e con f_{iB}, dell'Università agli Studi "G. D'Annunzio" negli anni accademici 2009/2010 e 2010/2011.

A: Anno Accademico 2009/2010	**Numero iscritti (f_i)**
Medicina$_1$	f_{1A}
Psicologia$_2$	f_{2A}
⋮	⋮
Scienze della Formazione$_i$	f_{iA}
⋮	⋮
Lettere$_k$	f_{kA}
	1

Tabella 6.5: Tabella di frequenza relativa degli scritti dell'Università agli Studi "G. D'Annunzio" a.a. 2009/2010.

Le due distribuzioni sono simili se:

B: Anno Accademico 2010/2011	**Numero iscritti (f_i)**
Medicina$_1$	f_{1B}
Psicologia$_2$	f_{2B}
⋮	⋮
Scienze della Formazione$_i$	f_{iB}
⋮	⋮
Lettere$_k$	f_{kB}
	1

Tabella 6.6: Tabella di frequenza relativa degli scritti dell'Università agli Studi "G. D'Annunzio" a.a. 2010/2011.

$\forall i$, $f_{iA} = f_{iB}$, cioè $f_{iA} - f_{iB} = 0$ oppure $| f_{iA} - f_{iB} | = 0$ [2]

Se, al contrario, esiste almeno una differenza $f_{iA} - f_{iB} \neq 0$, allora le due distribuzioni sono dissimili.

L'indice di dissomiglianza sarà dato, quindi, dalla somma degli scarti tra le frequenze delle due distribuzioni, prese in valore assoluto. Formalmente, sarà:

$$Z = \sum_{i=1}^{k} | f_{iA} - f_{iB} | \qquad (6.5.0.12)$$

L'indice Z, ovviamente, rappresenta un indice di dissomiglianza assoluto, ma abbiamo più volte avuto modo di constatare che per effettuare confronti, che non siano influenzati dalle caratteristiche specifiche della distribuzione (numerosità, unità di misura, ...), occorre sempre far riferimento ad un indice relativo.

Procediamo, quindi, con un esempio numerico per introdurre l'indice di dissomiglianza relativo. Riprendendo l'esempio delle facoltà dell'Università "D'Annunzio", consideriamo, al fine di rendere maggiormente semplice la spiegazione, il caso estremo in cui nei due diversi anni accademici presi in considerazione gli studenti siano tutti iscritti ad un unico corso di laurea.

[2] L'operazione modulo $|x|$ è quella operazione che assegna valore positivo all'argomento x. Di conseguenza per $x > 0$ sarà $|x| > 0$, per $x < 0$ sarà $|x| > 0$.

È abbastanza semplice intuire che, se raffrontiamo i due anni accademici, non si registrano variazioni nel numero degli iscritti alle diverse facoltà, l'indice di dissomiglianza risulterà nullo poiché si verificherà la condizione per cui $f_{iA} = f_{iB}$.

Al contrario scopriremo che nel caso di massima dissomiglianza, in cui per ognuna delle due distribuzioni tutte le unità presentano una stessa modalità, diversa nei due casi, gli indici assumono valore massimo pari a 2. In questo caso tutti gli studenti sono iscritti alla Facoltà di Scienze della Formazione e non vi sono iscritti alle altre facoltà.

A: Anno Accademico 2009/2010	**Numero iscritti (f_i)**
Medicina$_1$	$f_{1A} = 0$
Psicologia$_2$	$f_{2A} = 0$
⋮	⋮
Scienze della Formazione$_i$	$f_{iA} = 1$
⋮	⋮
Lettere$_k$	$f_{kA} = 0$
	1

Tabella 6.7: Esempio 1: distribuzione teorica di massima dissomoglianza.

Nel secondo caso, relativo all'anno accademico 2010/2011, al contrario tutti gli studenti dell'Università "G.D'Annunzio" risultano essere iscritti alla facoltà di Medicina.

Se procediamo ad effettuare la somma degli scarti, presi in valore assoluto, delle coppie di frequenze relative di ogni corso di laurea verifichiamo quanto espresso nella tabella 6.9: Da cui si deduce, quindi, che il valore massimo che può assumere l'indice di dissomiglianza assoluto è 2.

Abbiamo ora tutti gli strumenti necessari per costruire l'*indice di dissomiglianza relativo*. Sappiamo, infatti, che l'indice di dissomiglianza assoluto varia in un intervallo compreso tra i valori $Z_{min} = 0$ e

B: Anno Accademico 2010/2011	**Numero iscritti (f_i)**
Medicina$_1$	$f_{1B} = 1$
Psicologia$_2$	$f_{2B} = 0$
⋮	⋮
Scienze della Formazione$_i$	$f_{iB} = 0$
⋮	⋮
Lettere$_k$	$f_{kB} = 0$
	1

Tabella 6.8: Esempio 2: distribuzione teorica di massima dissomoglianza.

Facoltà	$\lvert f_{iA} - f_{iB} \rvert$
Medicina$_1$	$\lvert f_{iA} - f_{iB} \rvert = 1$
Psicologia$_2$	$\lvert f_{iA} - f_{iB} \rvert = 0$
⋮	⋮
Scienze della Formazione$_i$	$\lvert f_{iA} - f_{iB} \rvert = 1$
⋮	⋮
Lettere$_k$	$\lvert f_{iA} - f_{iB} \rvert = 0$
$Z = \sum_{i=1}^{k} \lvert f_{iA} - f_{iB} \rvert$	2

Tabella 6.9: Indice di dissomiglianza.

$Z_{max} = 2$; allora sarà:

$$z = \frac{Z - \min(Z_{ass})}{\max(Z_{ass}) - \min(Z_{ass})} \tag{6.5.0.13}$$

da cui

$$z = \frac{Z - 0}{2 - 0} = \frac{1}{2} \cdot Z$$

e quindi sostituendo Z, ne deriva che l'*indice relativo di dissomiglianza* sarà:

$$z = \frac{1}{2} \sum_{i=1}^{k} \lvert f_{iA} - f_{iB} \rvert \tag{6.5.0.14}$$

L'indice di dissomiglianza ci consente di fare anche confronti due a due (ad esempio facoltà a facoltà) in una distribuzione consentendoci di individuare quali siano più simili o dissimili tra loro rispetto ad un carattere osservato: dove, ad esempio, $z_{BF} = z_{FB}$ indicano l'indice

Facoltà	A	B	...	F	P
A: Medicina	0	z_{AB}	$z_{A...}$	z_{AF}	z_{AP}
B: Psicologia	z_{BA}	0	$z_{B...}$	z_{BF}	z_{BP}
⋮	⋮	⋮	⋮	⋮	⋮
F: Scienze della Formazione	z_{FA}	z_{FB}	$z_{F...}$	0	z_{FP}
⋮	⋮	⋮	⋮	⋮	⋮
P: Lettere	z_{PA}	z_{PB}	$z_{P...}$	z_{PF}	0

Tabella 6.10: Confronto tra "Indici di dissomiglianza.

relativo di dissomiglianza tra la Facoltà di Scienze della Formazione e quella di Psicologia per quanto riguarda il numero degli iscritti.

6.6 Rappresentazioni grafiche di caratteri misurati su scala nominale

La rappresentazione sintetica della distribuzione di un carattere preso nell'indagine di un fenomeno reale può avvenire oltre che attraverso le distribuzioni di frequenza, che abbiamo visto, anche per mezzo dell'uso di grafici. Queste due tipologie di strumenti sono concettualmente equivalenti poichè trasmettono le stesse informazioni, ma spesso i grafici sono di più immediata e semplice lettura rispetto ai numeri.

La rappresentazione grafica, infatti, ha l'obiettivo di porre in luce alcune caratteristiche salienti del fenomeno reale oggetto di studio, in una forma immediatamente percepibile e facilmente fruibile dai più, anche in assenza di specifiche competenze matematiche.

I grafici trovano maggiore applicazione soprattutto nella fase dell'analisi preliminare dei dati rilevati, ossia quando cerchiamo di capire cosa ci raccontano i dati del fenomeno reale indagato. Risultano, inoltre, molto utili nella presentazione dei risultati, ossia nella fase in cui si abbia la necessità di comunicare le informazioni ottenute dall'analisi dei dati in riferimento al fenomeno reale oggetto di studio.

Un grafico deve presentare alcune caratteristiche ottimali che sono da ricercare nell'accuratezza, nella semplicità e nella chiarezza con le quali è in grado di trasmettere le informazioni contenute. Il titolo del grafico deve descrive a quale carattere si riferisce specifico si riferisce la distribuzione, su quale collettivo di unità è stato misurato e, possibilmente, quando ed in quale contesto è stata effettuata la rilevazione (distribuzione del livello di gradimento del Servizio Sanitario della regione Abruzzo, Anno 2012).

Fondamentale è l'utilizzo di "Etichette opportune che diano indicazioni chiare delle modalità attraverso le quali si esprime il carattere, ossia che indichino la tipologia di frequenze alle quali ci si riferisce (assolute, relative, percentuali).

Un elemento fondamentale di un grafico è la legenda che ha il compito di illustrare tutti gli elementi presenti nel grafico stesso.

6.6.1 Diagramma a settori circolari

Nel *Diagramma a settori circolari*, più comunemente conosciuto come *torta*, a ciascuna modalità i x_i viene associato un settore circolare avente area proporzionale alla frequenza f_i. Questo tipo di grafico è particolarmente utile per i caratteri qualitativi sconnessi. Nella rappresentazione circolare le modalità, infatti, non sono ordinate ma ognuna rappresenta semplicemente la parte di un tutto. Supponendo di dover rappresentare la seguente distribuzione, utilizzeremo il grafico che segue in figura:

Professione	2010
Commessi	51.890
Contabili	29.840
Muratori	26.870
Camerieri	21.380
Atri	99.440

Le "fette" di una torta circolare rappresentano le modalità del carattere e sono costruite in modo che la loro dimensione sia pro-

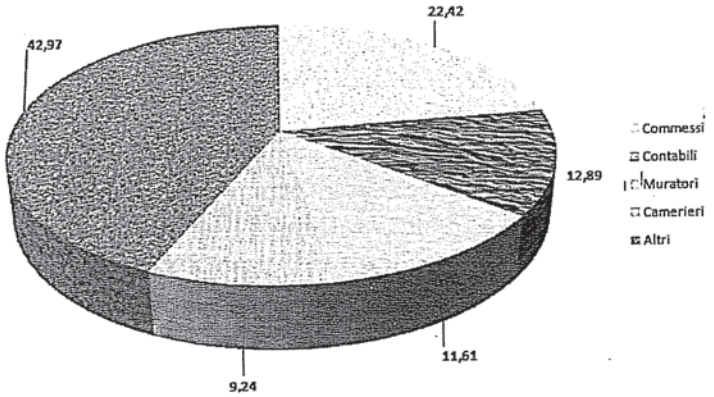


Figura 6.1: Esempio di distribuzione di frequenza.

porzionale alla corrispondente frequenza. Sostanzialmente ogni fetta rappresenta una parte del collettivo e, più nello specifico, quella parte che si esprime con una determinata modalità. L'ampiezza della "fetta" sarà detrminata, quindi, in relazione alla frequenza che la modalità da essa descritta ha rispetto a tutto il collettivo preso in esame. Vediamo, allora, in che modo si costruisce l'ampiezza dell'angolo di una "fetta" in relazione alla frequenza con cui si esprime la modalità.

Se lavoriamo sulle frequenze assolute di una distribuzione, l'ampiezza dell'angolo α_i della fetta di torta che rappresenta la modalità k_i, si definisce per mezzo della proporzione:

$$n_i : N = \alpha_i : 360°$$

da cui

$$\alpha_i = \frac{n_i \cdot 360°}{N}$$

Lo stesso vale se ci esprimiamo in termini di frequenza relativa, infatti possiamo scrivere:

$$f_i : 1 = \alpha_i : 360°$$

6.6.2 Diagrammi a barre e a nastro

Nel *grafico a barre* o *a nastro* a ciascuna modalità i x_i si associa un rettangolo avente base costante ed un'altezza proporzionale alla

frequenza f_i registrata per ogni modalità. Questo tipo di grafico, oltre che per i caratteri qualitativi sconnessi, può essere come vedremo in seguito anche per caratteri qualitativi ordinabili. L'ordine delle barre, con questi caratteri infatti, può essere corrispondente all'ordinamento delle modalità del carattere. Con questo tipo di grafico possiamo porre sull'asse delle ascisse le modalità del carattere, su quello delle ordinate le frequenze (assolute, relative o percentuali). I *Grafici a nastro* si costruiscono con le stesse modalità dei grafici a barre ma in questo caso le barre che rappresentano le frequenze sono poste in modo orizzontale.

Sia i grafici a barre che quelli a nastro possono essere suddivisi in modo che ogni barra rappresenti il modo in cui una specifica modalità osservata su un caratteresi distribuisce in relazione a due o più caratteristiche del collettivo preso in esame, ad esempio se si osserva il carattere gradimento della mensa scolastica espresso con la modalità buono possiamo rappresentare graficamente il numero, o la percentuale di maschi e femmine che hanno espresso tale opinione sul totale del collettivo (vedi esempio 6.2).

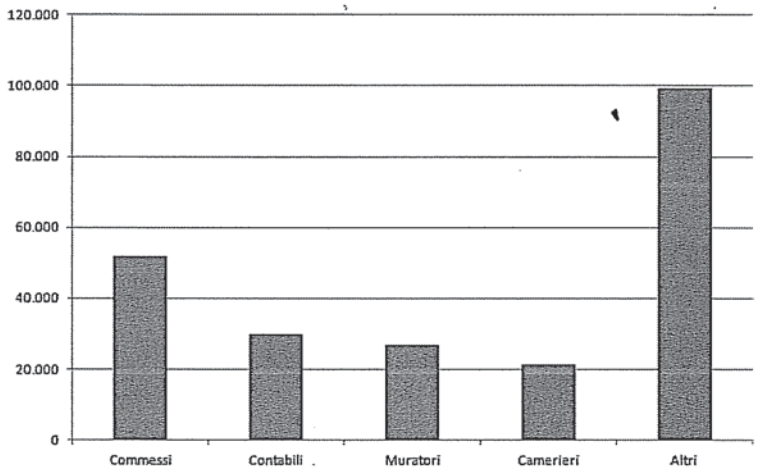


Figura 6.2: Esempio di Grafico a barre.

Supponendo di dover rappresentare la seguente distribuzione delle professioni in due anni successivi, possiamo utilizzare un *grafico*

a barre in pila che raffigura la variazione nella composizione delle diverse professioni rispetto al totale, come illustrato in figura 6.3. Come possiamo vedere il grafico viene costruito mettendo appunto *in pila* le barre corrispondenti alle frequenze delle singole modalità della distribuzione relative alle sotto-popolazioni di rilievo, che nel nostro caso sono rappresentate dalle diverse professioni.

Professione	2010	2009
Commessi	51.890	55.980
Contabili	29.840	24.220
Muratori	26.870	22.180
Camerieri	21.380	21.920
Atri	99.440	108.400

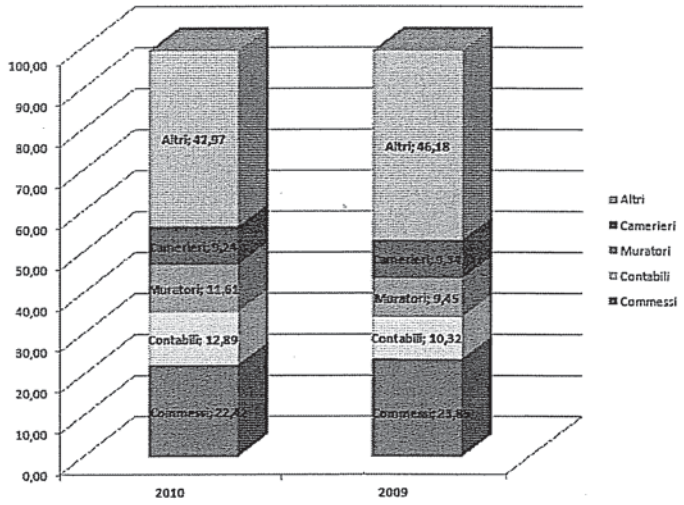


Figura 6.3: Esempio di Grafico a barre in pila.

6.6.3 Grafici figurativi o pittogrammi

Possiamo utilizzare questo tipo di grafico per rappresentare, attraverso l'uso di disegni, la distribuzione presa in esame. Ciascun

disegno rappresenta una modalità e la sua dimensione è proporzionale alla sua frequenza. Sono spesso grafici di tipo divulgativo e non molto rigorosi dal punto di vista scientifico nella rappresentazione, ma di semplice ed immediata lettura. Supponendo di dover rappresentare la seguente distribuzione del numero di bottiglie di vino vendute in diversi paesi, potremmo utilizzare un pittogramma in cui ogni bottiglia di vino rappresenta 100.000 bottiglie vendute (6.4):

Stato	Numero bottiglie vendute
Italia	700.000
Grecia	550.000
Portogallo	100.000
Spagna	1.450.000

Figura 6.4: Esempio di Pittogramma.

6.7 Riepilogando

Per riassumere quanto è stato detto del carattere qualitativo sconnesso in relazione alle sua peculiarità e alla sua operatività, procediamo con un esempio concreto di analisi di dati.

Si osservi la tabella 6.11, relativa ad un'indagine condotta su base nazionale dal Sistema Informativo Excelsior di Unioncamere per valutare le professioni più richieste nel 2010 dalle aziende in Italia. La

ricerca, condotta all'inizio del 2010, mira a conoscere le professioni più richieste nel 2010, in relazione ai dati del 2009. I dati riferiti al 2010, quindi, sono delle previsioni: si tratta di assunzioni programmate, in altre parole di richieste di personale previste dalle aziende per il 2010.

Professioni	2010	2009	Trend
Commessi	51.890	55.980	−
Contabili	29.840	24.220	+
Muratori	26.870	22.180	+
Camerieri	21.380	21.920	−
Conduttori mezzi	14.400	18.080	−
Tecnici vendita	11.970	10.710	+
Gestione magazzini	11.860	16.590	−
Professioni sanitarie	11.140	10.770	+
Elettricisti	10.280	9.840	+
Cuochi	10.160	9.340	+
Imp. segreteria	9.640	14.020	−
Idraulici	6.660	6.910	−
Informatici e telematici	5.820	5.610	+
Tecnici informatici	5.760	4.300	−
Ingegneri meccanici	1.750	2.230	−

Tabella 6.11: Esempio di un carattere qualitativo sconnesso. Fonte: Sistema Informativo Excelsior di Unioncamere.

Il carattere Professioni è un carattere di tipo qualitativo sconnesso misurabile su scala nominale. Possiamo, quindi, procedere con la ricerca della sintesi che, per questa tipologia di carattere, abbiamo visto essere la *moda*.

Osservando le frequenze assolute registrate in tabella, individuiamo che la moda del carattere Professioni, sia per l'anno 2009 che per l'anno 2010, è rappresentata dalla modalità "Commessi" poiché questa modalità presenta la frequenza assoluta più elevata. Ci esprimeremo, quindi, dicendo che *la sintesi della distribuzione relativa al carattere "Professioni" è "Commessi" in entrambe gli anni presi in considerazione.*

Costruiamo adesso la tabella di frequenza della nostra distribuzione prendendo in considerazione l'anno 2010.

Professioni	2010	Freq. Relativa (f_i)	Freq. Percentuale (p_i)
Commessi	51.890	0,23	23
Contabili	29.840	0,13	13
Muratori	26.870	0,12	12
Camerieri	21.380	0,09	9
Conduttori mezzi	14.400	0,06	6
Tecnici vendita	11.970	0,05	5
Gestione magazzini	11.860	0,05	5
Professioni sanitarie	11.140	0,05	5
Elettricisti	10.280	0,04	4
Cuochi	10.160	0,04	4
Imp. segreteria	9.640	0,04	4
Idraulici	6.660	0,03	3
Informatici e telematici	5.820	0,03	3
Tecnici informatici	5.760	0,03	3
Ingegneri meccanici	1.750	0,01	1
Totale	229.420	1	100

Tabella 6.12: Distribuzione di frequenza delle professioni più richieste nell'anno 2010. Fonte. Sistema Informativo Excelsior di Unioncamere.

Consideriamo adesso la variabilità della nostra distribuzione. Ai fini del nostro esempio, tra tutti i differenti metodi da noi utilizzati, considereremo gli indici OM_1, ET_1, om_1 ed et_1.

Nel caso dell'indice di omogeneità assoluta, sarà:

$$OM_1 = f_1^2 + f_2^2 + \ldots + f_k^2 = \sum_{i=1}^{k} f_i^2$$

Costruiamo la tabella per calcolare l'indice di omogeneità assoluta:

Professioni	2010	Freq. Relativa f_i	f_i^2
Commessi	51.890	0,23	0,051
Contabili	29.840	0,13	0,017
Muratori	26.870	0,12	0,014
Camerieri	21.380	0,09	0,009
Conduttori mezzi	14.400	0,06	0,004
Tecnici vendita	11.970	0,05	0,003
Gestione magazzini	11.860	0,05	0,003
Professioni sanitarie	11.140	0,05	0,002
Elettricisti	10.280	0,04	0,002
Cuochi	10.160	0,04	0,002
Imp. segreteria	9.640	0,04	0,002
Idraulici	6.660	0,03	0,001
Informatici e telematici	5.820	0,03	0,001
Tecnici informatici	5.760	0,03	0,001
Ingegneri meccanici	1.750	0,01	0,000
Totale	229.420	1	0,110

Tabella 6.13: Calcolo OM_1 della distribuzione di frequenza delle professioni più richieste nell'anno 2010.

L'indice di omogeneità assoluta della nostra distribuzione, sarà:

$$OM_1 = f_1^2 + f_2^2 + \ldots + f_k^2 = \sum_{i=1}^{k} f_i^2 = 0,110$$

Allora, ricordando che il valore massimo assunto da OM_1 è 1 e il valore minimo è pari ad $\frac{1}{k}$, che nel nostro caso sarà $\frac{1}{15} = 0,067$, diremo che l'omogeneità assoluta della nostra distribuzione registra un valore basso, per cui la nostra distribuzione è caratterizzata da una bassa omogeneità (di conseguenza, da un'alta eterogeneità).

A conferma di quanto detto, calcoliamo l'indice di eterogeneità

assoluta ET_1 nel modo seguente:

$$ET_1 = 1 - OM_1 = 1 - \sum_{i=1}^{k} f_i^2 = 1 - 0,110 = 0,890$$

da cui si evidenzia un'alta eterogeneità assoluta nella distribuzione, per cui un'elevata variabilità tra le diverse modalità. Infatti l'indice di eterogeneità di Gini, ET_1, varia tra $min(ET_1) = 0$ e $max(ET_1) = \frac{k-1}{k}$, ossia tra $min(ET_1) = 0$ e $max(ET_1) = 0,933$.

C'è da ricordare che gli indici calcolati sono espressi in valore assoluto, nel senso che hanno un proprio campo di variazione e dipendono dal numero di modalità del carattere sotto osservazione. Per avere una maggiore capacità informativa ed una immediatezza interpretativa, si introducono i corrispondenti indici relativi.

$$om_1 = \frac{\sum_{i=1}^{k} f_i^2 - \frac{1}{k}}{1 - \frac{1}{k}} = \frac{k\sum_{i=1}^{k} f_i^2 - 1}{k-1}$$

da cui

$$\frac{15 * 0,110 - 1}{15 - 1} = 0,047$$

e poiché sappiamo che *un indice è relativo ha un campo di variazione compreso tra 0 e 1 ed è un numero puro, ossia privo di dimensione*, allora possiamo confermare la bassa omogeneità della nostra distribuzione.

Procediamo, in fine, alla verifica del livello di eterogeneità relativa

$$et_1 = \frac{k}{k-1}\left[1 - \sum_{i=1}^{k} f_i^2\right]$$

da cui

$$et_1 = \frac{15}{14}[1 - 0,110] = 0,831$$

che conferma un'alta eterogeneità della distribuzione.

Si ricorda che per caratteri qualitativi sconnessi non è possibile studiare la forma della distribuzione.

Passiamo adesso ad effettuare un confronto tra le distribuzioni delle professioni in relazione all'anno di rilevamento dati, utilizzando

l'*indice di dissomiglianza*. A tale scopo utilizzeremo la tabella dell'indice di dissomiglianza tra due distribuzioni (cfr tab.6.14).

Professioni	2010	2009	Z
	f_A	f_B	$\mid f_{iA} - f_{iB} \mid$
Commessi	0,23	0,24	0,01
Contabili	0,13	0,10	0,03
Muratori	0,12	0,10	0,02
Camerieri	0,09	0,09	0,00
Conduttori mezzi	0,06	0,08	0,02
Tecnici vendita	0,05	0,05	0,00
Gestione magazzini	0,05	0,07	0,02
Professioni sanitarie	0,05	0,05	0,00
Elettricisti	0,04	0,04	0,00
Cuochi	0,04	0,04	0,00
Imp. segreteria	0,04	0,06	0,02
Idraulici	0,03	0,03	0,00
Informatici e telematici	0,03	0,02	0,01
Tecnici informatici	0,03	0,02	0,01
Ingegneri meccanici	0,01	0,01	0,00
$Z = \sum_{i=1}^{k} \mid f_{iA} - f_{iB} \mid$			0,14

Tabella 6.14: Indice di dissomiglianza tra due distribuzioni.

Ricordiamo che l'indice di dissomiglianza assoluto è dato dalla somma degli scarti tra le frequenze delle due distribuzioni, prese in valore assoluto. L'indice di dissomiglianza assoluto, quindi, è dato da:

$$Z = \sum_{i=1}^{k} \mid f_{iA} - f_{iB} \mid$$

Adesso abbiamo tutti gli strumenti necessari per individuare l'indice di dissomiglianza relativo poiché sappiamo che l'indice di dissomiglianza assoluto varia in un intervallo compreso tra i valori $Zmin = 0$ e $Zmax = 2$; allora sarà:

$$z = \frac{Z - min(Z_{ass})}{max(Z_{ass}) - min(Z_{ass})}$$

da cui

$$z = \frac{Z-0}{2-0} = \frac{1}{2} \cdot Z$$

e quindi sostituendo Z, ne deriva che

$$z = \frac{1}{2} \cdot 0,14 = 0,07$$

Ricordando che

$$0 \leqslant z \leqslant 1$$

possiamo concludere che tra le due distribuzioni la dissomiglianza è molto bassa, ossia le due distribuzioni sono molto simili tra loro.

Capitolo 7

Caratteri misurati su scala qualitativa ordinabile

Un secondo tipo di carattere, che si potrebbe incontrare durante una ricerca della valutazione di un fenomeno reale replicabile su un collettivo di unità, è quello misurato su scala qualitativa ordinabile. Alcuni esempi di questo tipo di carattere sono: la valutazione scolastica espressa nelle modalità *non sufficiente, sufficiente, buono, distinto e ottimo*; il livello di gradimento di un servizio erogato espresso nelle modalità *per niente soddisfatto, poco soddisfatto, soddisfatto, molto soddisfatto*; l'espressione del proprio accordo rispetto ad un'opinione espresso nelle modalità *per niente d'accordo, poco d'accordo, d'accordo, pienamente d'accordo*; e così via.

In ognuno di questi casi tra le modalità è possibile definire un ordine gerarchico di valori. In linea generale indichiamo, come abbiamo fatto in precedenza:

- il carattere con la lettera dell'alfabeto romano in maiuscolo;
- la misura del carattere con la corrispondente lettera in minuscolo, ponendo al suo pedice l'indice tra parentesi. Formalmente $x_{(i)}$ indica, quindi, la modalità che occupa la posizione *i*-esima.

7.1 La frequenza cumulata

Oltre alle distribuzioni di frequenze viste in precedenza per i caratteri qualitativi sconnessi, ossia le frequenze assolute (n_i con $i = 1, 2, \ldots, k$), le frequenze relative (f_i con $i = 1, 2, \ldots, k$) e le frequenze percentuali (p_i per $i = 1, 2, \ldots, k$), risulta molto utile ricavare la distribuzione delle *frequenze cumulate*.

Da un punto di vista formale indichiamo la frequenza assoluta cumulata con N_i ($i = 1, 2, \ldots, k$); dal punto di vista operativo, invece, le *frequenze assolute cumulate* si ottengono attraverso la somma delle frequenze assolute minori o uguali a quella associata alla modalità *i*-esima (v. Tab. 7.1). Possiamo, quindi, definire la frequenza assoluta cumulata come la somma delle frequenze assolute di tutti i valori inferiori o uguali al valore della posizione considerata. Sarà quindi:

$$N_1 = n_1$$

$$N_2 = n_1 + n_2$$

$$\ldots$$

$$N_i = n_1 + n_2 \ldots + n_i$$

$$\ldots$$

$$N_k = N$$

In generale, quanto abbiamo appena esplicitato, si formalizza come segue:

$$N_i = \sum_{k=1}^{i} n_k \qquad (7.1.0.1)$$

dove N_i indica il valore della frequenza assoluta cumulata della posizione *i*-esima.

Quanto appena detto e sostanzialmente replicabile per quanto riguarda le *frequenze relative cumulate*, che si ottengono attraverso la somma delle frequenze relative minori o uguali a quella associata alla modalità *i*-esima (v. Tab. 7.1). Possiamo, quindi, definire la frequenza relativa cumulata come la somma delle frequenze relative di tutti i

valori inferiori o uguali al valore della posizione considerata. Sarà quindi:

$$F_1 = f_1$$

$$F_2 = f_1 + f_2$$

$$\ldots$$

$$F_i = f_1 + f_2 \ldots + f_i$$

$$\ldots$$

$$F_k = 1$$

In generale, quanto abbiamo appena esplicitato, si formalizza come segue:

$$F_i = \sum_{k=1}^{i} f_k \qquad (7.1.0.2)$$

dove F_i indica il valore della frequenza assoluta cumulata della posizione *i*-esima ed ha il significato di: aliquota sul totale delle unità del collettivo che presentano modalità minore o uguale a...

È opportuno notare che in corrispondenza della *k*-esima modalità la frequenza relativa cumulata assume sempre vare 1, poiché in questa ultima posizione sarà vero che tutto il collettivo preso in esame, nelle posizioni precedenti, presenta modalità minori rispetto all'unità, che nel caso delle frequenze relative rappresenta l'intero di riferimento.

La figura 7.1, invece, riporta la rappresentazione delle frequenze cumulate nel caso di una generica variabile che può assumere k modalità: dove con N_i , F_i e P_i sono indicate, rispettivamente, le frequenze cumulate assolute, relative e percentuali in corrispondenza della generica i-esima modalità.

La frequenza cumulata in corrispondenza di una data modalità del carattere, indica il numero (in caso di frequenze assolute) o la frazione (in caso di frequenze relative o percentuali) delle unità della popolazione considerata che presentano un valore della variabile minore o uguale (ovvero non superiore) alla modalità in questione.

Dal punto di vista interpretativo, quindi, la frequenza assoluta cumulata risponde alla domanda: In quanti hanno avuto un risultato minore di ...?

X	n	f	N	F
$x_{(1)}$	n_1	f_1	$N_1 = n_1$	$F_1 = f_1$
$x_{(2)}$	n_2	f_2	$N_2 = n_1 + n_2$	$F_2 = f_1 + f_2$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$x_{(i-1)}$	n_{i-1}	f_{i-1}	$N_{i-1} = n_1 + \ldots + n_{i-1}$	$F_{i-1} = f_1 + \ldots + f_{i-1}$
$x_{(i)}$	n_i	f_i	$N_i = n_1 + \ldots + n_i$	$F_{i-1} = f_1 + \ldots + f_i$
$x_{(i+1)}$	n_{i+1}	f_{i+1}	$N_{i-1} = n_1 + \ldots + n_{i+1}$	$F_{i-1} = f_1 + \ldots + f_{i+1}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$x_{(k-1)}$	n_{k-1}	f_{k-1}	$N_{k-1} = n_1 + \ldots + n_{k-1}$	$F_{k-1} = f_1 + \ldots + f_{k-1}$
$x_{(k)}$	n_k	f_k	$N_k = n_1 + \ldots + n_k = N$	$F_k = f_1 + \ldots + f_k = 1$
	N	1		

Figura 7.1: Frequenze cumulate.

Carattere	Freq. assolute	Freq. relative	Freq. assoluta cumulata	Freq. relative cumulata
X	(n_i)	(f_i)	N_i	F_i
$x_{(1)}$	n_1	f_1	N_1	F_1
$x_{(2)}$	n_2	f_2	N_2	F_2
⋮	⋮	⋮	⋮	⋮
$x_{(i)}$	n_i	f_i	N_i	F_i
⋮	⋮	⋮	⋮	⋮
$x_{(k)}$	n_k	f_k	$N_k = N$	$F_k = 1$
Totale	N	1	-	-

Tabella 7.1: Distribuzione di frequenza di un carattere qualitativo ordinabile.

Proseguendo con lo studio della sintesi della distribuzione di un carattere qualitativo ordinabile, vedremo come possa tornare utile l'individuazione delle frequenze relative cumulate.

Per esplicare quanto detto, facciamo un esempio concreto riguardo alla frequenza cumulata. Supponiamo che si sia misurato il carattere "Risultati del test di ingresso al I liceo Luigi Buffon di Cepagatti", con la scala ordinale riportata in tabella seguente, e che si sia ottenuta la seguente distribuzione di frequenze assolute e frequenze assolute cumulate (v. Tab.15.3).

7.2 La frequenza retrocumulata

Oltre alla frequenza cumulata, è possibile definire la *frequenza retrocumulata* in corrispondenza di una data modalità come il numero (nel caso di frequenze assolute) o la frazione (nel caso di frequenze relative o percentuali) di unità che hanno un valore del carattere maggiore o uguale (ovvero non minore) della modalità in questione.

La seguente tabella riporta la definizione delle frequenze retrocumulate nel caso di una generica variabile che può assumere k modalità: dove con RN_i , RF_i e RP_i sono indicate, rispettivamente, le frequen-

X	n	f	RN	RF
$x_{(1)}$	n_1	f_1	$RN_1 = n_1 + ... + n_k = N$	$RF_1 = f_1 + ... + f_k = 1$
$x_{(2)}$	n_2	f_2	$RN_2 = n_2 + ... + n_k$	$RF_2 = f_2 + ... + f_k$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$x_{(i-1)}$	n_{i-1}	f_{i-1}	$RN_{i-1} = n_{i-1} + ... + n_k$	$RF_{i-1} = f_{i-1} + ... + f_k$
$x_{(i)}$	n_i	f_i	$RN_i = n_i + ... + n_k$	$RF_i = f_i + ... + f_k$
$x_{(i+1)}$	n_{i+1}	f_{i+1}	$RN_{i+1} = n_{i+1} + ... + n_k$	$RF_{i+1} = f_{i+1} + ... + f_k$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$x_{(k-1)}$	n_k	f_k	$RN_k = n_k$	$RF_k = f_{k-1} + f_k$
$x_{(k)}$	n_{k-1}	f_{k-1}	$RN_{k-1} = n_{k-1} + n_k$	$RF_k = f_k$
	N	1		

Figura 7.2: Frequenze retrocumulate.

Risultati test di Ingresso X	Freq. assolute (n_i)	Freq. relative (f_i)	Freq. assoluta cumulata N_i	Freq. relative cumulata F_i
Insufficiente	9	0,10	9	0,09
Sufficiente	21	0,24	30	0,34
Buono	40	0,45	70	0,80
Distinto	10	0,11	80	0,91
Ottimo	8	0,09	88	1
Totale	88	1	-	-

Tabella 7.2: Frequenze cumulate: esempio n. 1

ze retrocumulate assolute, relative e percentuali in corrispondenza dell'i-esima modalità.

Dalle definizioni di frequenze retrocumulate e cumulate è possibile ricavare la seguente relazione che permette di calcolare le prime una volta note le seconde:

$$RN_i = n_i + ... + n_k = N - (n_1 + ... + n_{i-1}) = N - N_{i-1} \qquad (7.2.0.3)$$

ovvero di calcolare la frequenza retrocumulata in corrispondenza dell'i-esima modalità sottraendo da N il valore della frequenza cumulata in corrispondenza della modalità precedente. Valgono ovviamente anche le seguenti due relazioni per le frequenze retrocumulate relative e percentuali:

$$RF_i = 1 - F_{i-1} \qquad (7.2.0.4)$$

$$RP_i = 100 - P_i \qquad (7.2.0.5)$$

È chiaro che in corrispondenza della prima modalità si ha sempre: $RN_1 = N$, $RF_1 = 1$ e $RP_1 = 100$.

Le frequenze cumulate e retrocumulate possono essere calcolate, chiaramente, anche per una variabile quantitativa. Non risultano invece applicabili nel caso di una variabile qualitativa nominale o sconnessa, anche se dal punto di vista meramente algebrico risulta possibile calcolarle.

7.3 Sintesi

Dopo una preliminare presentazione tabellare del carattere si rende necessario passare all'analisi del fenomeno reale, come precedentemente detto, attraverso il metodo statistico che in prima istanza si concretizza nella ricerca della sintesi. Il fattore aggiuntivo rispetto al carattere sconnesso risiede nella possibilità computazionale di sfruttare l'ordinamento della modalità. Ciò ovviamente non esclude la possibilità di considerare come valore di sintesi la *Moda*, ossia la frequenza prevalente, alla stessa stregua del carattere misurato su scala nominale. Ma, come abbiamo avuto modo di sottolineare in precedenza, questo valore di sintesi è poco indicativo della distribuzione.

Un valore di sintesi sicuramente più adatto per questa specie di caratteri, che tiene anche conto dell'ordinamento delle modalità, è la *Mediana*.

La mediana è, in una distribuzione di un carattere le cui modalità sono ordinabili, la modalità che divide la distribuzione esattamente in due parti uguali.

Per parti uguali si intende che le frequenze cumulate dalla modalità che chiameremo mediana sono esattamente il 50% del totale.

Il grafico qui riportato esplicita il significato delle affermazioni fatte.

In altri termini la mediana è la modalità che lascia alla sua sinistra esattamente il 50% delle frequenze.

Per individuare la mediana è necessario, prima di tutto, disporre le modalità in maniera ordinata. Dal punto di vista computazionale, poi, il calcolo della mediana va distinto a seconda se la numerosità del collettivo sia pari o dispari.

Nel caso di un collettivo composto da un numero dispari di unità, la mediana sarà la modalità che occupa la posizione $\frac{N+1}{2}$-esima. Quindi la mediana, sarà:

$$Me = x_{(\frac{N+1}{2})} \tag{7.3.0.6}$$

Mentre se il collettivo è composto da un numero pari di termini allora si intuisce facilmente che non esiste un sola modalità per la quale sia possibile cumulare esattamente il 50% delle unità. La banale

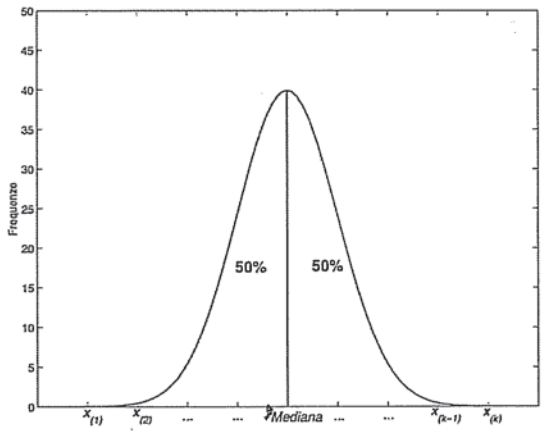


Figura 7.3: Rappresentazione grafica della mediana.

soluzione è quella di individuare due mediane: la prima che occupa la posizione $\frac{N}{2}$ e l'altra che occupa la posizione $\frac{N}{2}+1$; ossia:

$$Me_1 = x_{(\frac{N}{2})} \text{ e } Me_2 = x_{(\frac{N}{2}+1)} \tag{7.3.0.7}$$

Quando si dispone di una distribuzione di frequenze, così come riportato nel secondo esempio del Liceo "Luigi Buffon" di Cepagatti (Tab. 7.3), per individuare la modalità mediana si procede attraverso l'analisi delle frequenze cumulate.

Vediamo come operativamente si individua la mediana di una distribuzione. Le modalità vengono poste in tabella in modo ordinato. Il valore della frequenza cumulata, ossia N_1, indica che la prima modalità $x_{(1)}$ è occupata da N_1 posizioni e che la modalità $x_{(2)}$ occupa la posizione da N_1+1 a N_2, e così via per la generica frequenza cumulata N_i per $i = 1, 2, \ldots, k$. In questo modo per sapere la modalità mediana, dapprima si verifica se N è pari o dispari. Se è dispari si calcola il valore $\frac{N+1}{2}$ oppure se pari i valori $\frac{N}{2}$ e $\frac{N}{2}+1$, poi si va sulla colonna delle frequenze cumulate e si trova la frequenza cumulata appena superiore al valore o ai valori trovati. La modalità nel caso dispari o le modalità nel caso pari che corrispondono a dette frequenze rappresentano la mediana desiderata. Nell'ipotesi di un

collettivo di un numero pari di unità, capita spesso che le due modalità $Me_1 = x_{(\frac{N}{2})}$ e $Me_2 = x_{(\frac{N}{2}+1)}$ coincidano con la stessa modalità e in tal caso si concluderà che $Me_1 = x_{(\frac{N}{2})} = Me_2 = x_{(\frac{N}{2}+1)} = Me$ quindi ritorniamo ad un solo valore di mediana.

Facciamo un esempio:

Risult. test di ingresso	Freq. assolute	Freq. relative	Freq. assolute cumulate	Freq. relative cumulata
	n_i	f_i	N_i	F_i
Insufficiente	9	0,10	9	0,09
Sufficiente	21	0,24	30	0,34
Buono	40	0,45	**70**	0,80
Distinto	10	0,11	80	0,91
Ottimo	8	0,09	88	1
Totale	88	1		

Tabella 7.3: Frequenze cumulate: esempio n. 2

In questo caso, siccome $N = 88$ è pari, si avrà $\frac{N}{2} = \frac{88}{2} = 44$ mentre $\frac{N}{2} + 1 = 44 + 1 = 45$, dove i due valori di mediana sono le modalità che occupano rispettivamente la 44-esima e la 45-esima posizione. Se si controlla sulla colonna delle frequenze cumulate quella appena superiore è 70 ciò significa che la modalità "buono" occupa le posizioni da 31-esima fino alla 70-esima; quindi anche la 44-esima e la 45-esima. In conclusione la mediana è la modalità *Buono*.

In una distribuzione di frequenze di un carattere qualitativo misurato su scala ordinale, può essere interessante conoscere la modalità che occupa un predefinito posto in graduatoria. Per esempio la modalità che occupa il primo quarto di posto o il primo quinto ecc. In altri termini la modalità cha lascia alla sua sinistra o che cumula il 25% delle unità o il 20% delle unità. Nel metodo statistico queste modalità sono chiamate *quantili*.

In particolare i quantili si distinguono in *centili* (o *percentili*), *decili* o *quartili*. Data una distribuzione naturalmente avremo 100 centili nel senso che la distribuzione può essere divisa in cento parti dove il

primo percentile è la modalità che cumula 1% delle unità, il secondo centile è la modalità che cumula il 2% delle unità, il cinquantesimo centile il 50% e corrisponde alla mediana, ecc.

I decili sono 10 posto $N = 100$ nel senso che la distribuzione può essere divisa in dieci parti, dove il primo decile è la modalità che cumula il 10% delle unità, il secondo decile è la modalità che cumula il 20% delle unità, il quinto il 50% e corrisponde alla mediana, ecc.

I quartili dividono la distribuzione in 4, nel senso che la distribuzione può essere divisa in quattro parti dove il primo quartile è la modalità che cumula 25% delle unità, il secondo quartile è la modalità che cumula il 50% delle unità e corrisponde alla mediana; il terzo quartile è la modalità che cumula il 75% delle unità e infine il quarto quartile è la modalità che cumula il 100% delle unità. Da un punto di vista formale indicheremo con C_i per $i = 1, 2, \ldots, 100$ l'*i*-esimo percentile; con D_i con $i = 1, 2, \ldots, 10$ l'*i*-esimo decile; con Q_i, per $i = 1, 2, 3, 4$, l'*i*-esimo quartile. La procedura metodologica per calcolare i quantili è la stessa adottata per la mediana. Per esempio se si volesse calcolare il primo quartile Q_1, bisognerebbe dividere per 4 il collettivo, ossia $\frac{N}{4}$ e controllare sulla colonna delle frequenze cumulate la modalità corrispondente, mentre per il terzo quartile Q_3 bisogna dividere per $\frac{3}{4}$ il collettivo, ossia $\frac{3}{4} \cdot N$ e controllare sulla colonna delle frequenze cumulate la modalità corrispondente. I quartili, quindi, possiamo definirli come quantili di ordini 1/4, 2/4 e 3/4. La mediana corrisponde al quantile di ordine $\frac{1}{2}$.

Nell'esempio precedente, $Q_1 =$ *Sufficiente*, mentre $Q_2 =$ *Buono* che rappresenta appunto la mediana, $Q_3 =$ *Buono* e $Q_4 =$ *Ottimo*.

7.4 Variabilità

Dovrebbe ormai risultare chiaro che, dopo la valutazione della sintesi di un distribuzione, la fase successiva è quella della misura della variabilità, aspetto che, come abbiamo più volte discusso, è essenziale per la valutazione del collettivo. Naturalmente anche in questo caso, così come si è detto per la sintesi, potendo disporre dell'ordinamento delle modalità, oltre agli indici di eterogeneità e di omogeneità già in-

trodotti per i caratteri misurati su scala nominale possiamo introdurre indici che tengano conto dell'aspetto dell'ordinamento delle modalità.

A questo proposito introduciamo la misura di *dispersione* che ci consente di valutare appunto in che modo si "disperdono", si distribuiscono, le modalità del collettivo.

Per rendere più semplice l'approccio a quanto stiamo per spiegare, riprendiamo per un attimo il concetto già visto della omogeneità. *Diremo che in una distribuzione c'è minima dispersione quando il collettivo è perfettamente omogeneo, ossia tutte le unità sono concentrate su una sola modalità.* Al contrario diremo che *in una distribuzione di un carattere ordinato c'è massima dispersione se tutte le unità sono ripartite equamente nelle due modalità estreme.*

Per ottenere un indice di dispersione di un carattere qualitativo misurato su scala ordinale, possiamo ricorrere alle frequenze cumulate; infatti, tenendo conto della definizione di dispersione data sopra, ci si aspetterà un'elevata dispersione se le frequenze più alte saranno concentrate verso le modalità estreme.

Per evidenziare questo effetto si propone il seguente indice:

$$DIS = 2\sum_{i=1}^{k} F_i(1-F_i) \qquad (7.4.0.8)$$

Nel caso di minima dispersione l'indice *DIS* vale zero. In particolare in caso di massimo l'indice vale:

$$max(DIS) = \frac{k-1}{2} \text{ se N è pari}$$

mentre

$$max(DIS) = \frac{k-1}{2}\left(1-\frac{1}{N^2}\right) \text{ se Nè dispari.}$$

In definitiva, conoscendo il valore di minima dispersione e quello di massima, possiamo facilmente ricavare i corrispondenti indici relativi nei casi di numerosità del collettivo pari o dispari.

Nel caso di *indice di dispersione relativo con un collettivo di* N *pari* avremo quindi:

$$dis = \frac{DIS - min(DIS)}{max(DIS) - min(DIS)} =$$

$$= \frac{2\sum_{i=1}^{k} F_i(1-F_i) - 0}{\frac{k-1}{2} - 0} =$$

$$= \frac{4\sum_{i=1}^{k} F_i(1-F_i)}{k-1}$$

Nel caso di *indice di dispersione relativo con un collettivo di* N *dispari*, sarà invece:

$$dis = \frac{DIS - min(DIS)}{max(DIS) - min(DIS)} =$$

$$= \frac{2\sum_{i=1}^{k} F_i(1-F_i) - 0}{\frac{k-1}{2}\left(1 \pm \frac{1}{N^2}\right) - 0} =$$

$$= \frac{4\sum_{i=1}^{k} F_i(1-F_i)}{k-1\left(1-\frac{1}{N^2}\right)}$$

È necessario tuttavia far notare che quando il collettivo è molto numeroso, allora la quantità $\frac{1}{N^2}$ diventa molto piccola, tale da poter essere trascurata. La conseguenza di questo fatto è che per grandi collettivi l'unico indice di dispersione relativo da prendere in considerazione sarà:

$$dis = \frac{4\sum_{i=1}^{k} F_i(1-F_i)}{k-1} \qquad (7.4.0.9)$$

Per concludere questa sezione facciamo un piccolo esempio concreto che riassumiamo nella tabella(7.4).

Indice gradimento servizi erogati	Freq. assolute (n_i)	Freq. relative (f_i)	Freq. relative cumulate (F_i)	$(1-F_i)$	$F_i(1-F_i)$
Non soddisf.	12	0,40	0,40	0,60	0,2400
Poco soddisf.	7	0,23	0,63	0,37	0,2331
Abbastanza soddisf.	4	0,13	0,76	0,24	0,1824
Molto soddisf.	5	0,17	0,93	0,07	0,0651
Pienamente soddisf.	2	0,07	1,00	0,00	0,0000
Totale	30	1			0,7206

Tabella 7.4: Calcolo indici di dissomiglianza: esempio n. 1

Da cui si ricava che la dispersione assoluta è data da:

$$DIS = 2\sum_{i=1}^{k} F_i(1-F_i) = 2 \cdot 0,7206 = 1,4412$$

Passando alla dispersione relativa si ha:

$$dis = \frac{4\sum_{i=1}^{k} F_i(1-F_i)}{k-1} = \frac{4 \cdot 0,7206}{5-1} = 0,7206$$

dal valore dell'indice di dissomiglianza si evince che nella distribuzione considerata c'è una variabilità piuttosto elevata, in quanto ricordiamo che un indice relativo ha sempre un campo di variazione $0-1$.

7.5 Forma

Contrariamente a quanto abbiamo visto per i caratteri qualitativi misurati su scala nominale, per i caratteri qualitativi ordinabili, che possiedono la prerogativa dell'ordinamento tra le modalità, è possibile considerare una misura della forma della distribuzione. Ossia possiamo studiare come si distribuiscono le frequenze tra le diverse modalità, allo scopo di valutare la possibilità di presenza di regolarità o di particolari comportamenti su modalità diverse rispetto alla sintesi individuata.

7.5.1 Indici di asimmetria

In questo contesto una prima valutazione è quella della *simmetria* della distribuzione. In particolare diremo che, in una distribuzione di un carattere qualitativo misurato su scala ordinale, c'è simmetria se, data la distribuzione, si verifica la seguente situazione:

$$\begin{aligned} n_1 &= n_k \\ n_2 &= n_{k-1} \\ &\cdots \\ n_i &= n_{k-i+1} \end{aligned}$$

Tra le forme di asimmetria più interessanti, dal punto di vista interpretativo di una distribuzione, ci sono l'*asimmetria positiva* e l'*asimmetria negativa*. Diremo che in un distribuzione c'è *asimmetria positiva* se la maggior parte delle unità è concentrata nelle modalità più piccole e di conseguenza c'è una maggiore variabilità tra quelle più grandi; in caso contrario, cioè se la maggior parte delle unità è concentrata tra le modalità più grandi, diremo che c'è *asimmetria negativa*.

Il modo più semplice per trovare una misura dell'asimmetria di una distribuzione è quello di considerare una particolare versione dell'indice di dispersione.

In particolare, supponiamo di dividere la distribuzione dei dati in due parti uguali: la prima costituita dalle più piccole modalità che cumulano $\frac{N}{2}$ unità se N è pari o $\frac{N+1}{2}$ se N è dispari. La seconda distribuzione costituita dalle modalità più grandi per le restanti unità. Su ciascuna di esse è possibile calcolare la misura della dispersione, che chiameremo dispersione destra della distribuzione (DIS_d) e dispersione sinistra della distribuzione (DIS_s). Naturalmente se nella distribuzione c'è simmetria allora $DIS_d = DIS_s$; mentre se c'è asimmetria positiva $DIS_d > DIS_s$, in quanto, essendoci più unità concentrate tra le modalità più piccole, si registrerà anche una maggiore variabilità tra quelle grandi. Viceversa nel caso di asimmetria negativa le unità saranno concentrate nelle modalità più grandi, quindi risulterà $DIS_d < DIS_s$.

Introduciamo, quindi, il seguente indice (ASI), quale *indice assoluto di asimmetria* di un distribuzione di un carattere qualitativo misurato su scala ordinale:

$$ASI = DIS_d - DIS_s \qquad (7.5.1.1)$$

Diremo che se $ASI = 0$ allora la distribuzione è simmetrica, mentre se $ASI > 0$ allora $DIS_d > DIS_s$ e si avrà asimmetria positiva; se, infine, $ASI < 0$ allora $DIS_d < DIS_s$ e si avrà asimmetria negativa.

Anche in questo caso, per avere una misura dell'asimmetria che sia confrontabile e che esprima il grado della misura di asimmetria, è necessario introdurre il corrispondente *indice relativo di asimmetria*. Da quanto visto sopra, per fare ciò abbiamo necessità di conoscere il $\min(ASI)$ ed il $\max(ASI)$. Circa il valore minimo dell'indice di asimmetria relativa, come abbiamo più volte rimarcato in caso di

assenza di asimmetria, ossia in presenza di simmetria, si osserverà $DIS_d = DIS_s$ quindi $ASI = 0$. Mentre il valore massimo sarà dato da $DIS_d + DIS_s$. In questo modo l'indice relativo di asimmetria per un carattere qualitativo misurato su scala ordinale sarà:

$$asi = \frac{DIS_d - DIS_s}{DIS_d + DIS_s} \qquad (7.5.1.2)$$

L'indice asi assumerà valore 1 quando $DIS_s = 0$, cioè in caso di massima asimmetria positiva, mentre assumerà valore -1 quando $DIS_d = 0$, cioè in caso di massima asimmetria negativa.

7.6 Rappresentazioni grafiche

7.6.1 Il Box-plot

Per rappresntare graficamente i caratteri qualitativi ordinabili, oltre a tutti i grafici che possiamo utilizzare con i caratteri qualitativi sconnessi, possiamo utilizzare il *box-plot*. In statistica il *box-plot*, detto anche *box and whiskers plot, diagramma a scatola e baffi*, è una rappresentazione grafica utilizzata per descrivere la distribuzione di un collettivo, mettendone in evidenza la dispersione rispetto ad un di posizione.

Il box-plot può essere rappresentato orientato orizzontalmente o verticalmente, per mezzo di un rettangolo diviso in due parti, da cui fuoriescono due segmenti. Più nel dettaglio diremo che il rettangolo rappresenta la "scatola" ed è delimitato dal primo e dal terzo quartile della distribuzione. Al suo interno esso è diviso dalla mediana, ossia dal secondo quartile della distribuuzione. I segmenti che fuoriescono, detti anche comunemente "baffi", sono delimitati dal minimo e dal massimo dei valori della distribuzione. Con questa rappresentazione grafica vengono, quindi, rappresentati i quattro intervalli della distribuzione delimitati dai quartili.

Osservano la figura 7.4 osserviamo, quindi, che la linea interna alla scatola rappresenta la Mediana della distribuzione. Le linee estreme della scatola rappresentano il primo ed il terzo quartile. La distanza tra il terzo ed il primo quartile, detta appunto *Distanza Interquartilica*, è una misura della dispersione della distribuzione. Il

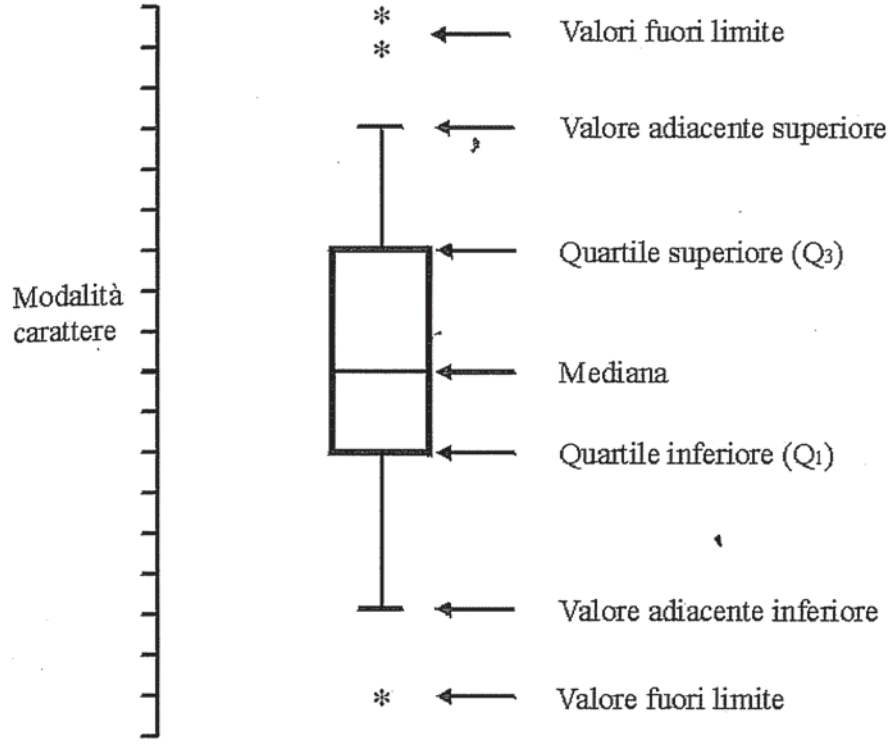


Figura 7.4: Box-plot.

50% delle osservazioni si trovano comprese tra questi due valori. Se l'intervallo interquartilico è piccolo, tale metà delle osservazioni si trova fortemente concentrata intorno alla mediana; all'aumentare della distanza interquartilica aumenta la dispersione del 50% delle osservazioni centrali intorno alla mediana. Le distanze tra ciascun quartile e la mediana forniscono informazioni relativamente alla forma della distribuzione. Se una distanza è diversa dall'altra allora la distribuzione è asimmetrica. Le linee che si allungano dai bordi della scatola, i baffi, individuano gli intervalli in cui sono posizionati i valori rispettivamente minori del primo quartile, Q1, e maggiori del terzo, Q3; i punti estremi dei baffi evidenziano i valori adiacenti. Se si indica con $r = (Q3 - Q1)$ la differenza interquartilica, il valore adiacente inferiore è il valore più piccolo tra le osservazioni che risulta maggiore o uguale a $Q1 - 1,5r$. Il valore adiacente superiore, invece, è il valore più grande tra le osservazioni che risulta minore o uguale a $Q3 + 1,5r$. Pertanto se gli estremi della distribuzione sono contenuti tra $Q1 - 1,5r$ e $Q3 + 1,5r$ essi coincideranno con gli estremi dei baffi, altrimenti come estremi verranno usati i valori $Q1 - 1,5r$ e $Q3 + 1,5r$. I valori esterni a questi limiti (esterni rispetto ai valori adiacenti), chiamati in genere valori anomali, vengono segnalati individualmente nel boxplot per meglio evidenziarne la presenza e la posizione. Questi valori infatti costituiscono una "anomalia" rispetto alla maggior parte dei valori osservati e pertanto è necessario identificarli per poterne analizzare le caratteristiche e le eventuali cause che li hanno determinati. Essi forniscono informazioni ulteriori sulla dispersione e sulla forma della distribuzione. Quando i valori adiacenti, superiore e inferiore, coincidono con gli estremi della distribuzione non comparirà alcun valore fuori limite. I valori adiacenti inferiore e superiore forniscono informazioni sulla dispersione e sulla forma della distribuzione ed anche sulle code della distribuzione.

7.7 Riepilogando

Per riassumere quanto è stato detto del carattere qualitativo ordinabile, in relazione alle sua peculiarità e alla sua operatività, procediamo con un esempio concreto di analisi di dati.

Si osservi la tabella 7.5, relativa ad un'indagine condotta dall'A.Di.S.U. Ateneo "Federico II" di Napoli da gennaio del 2003 a dicembre del 2010 sul gradimento del servizio ristorazione di un gruppo di studenti universitari.

Giudizio	(n_i)	(f_i)	(p_i)	(N_i)	(P_i)
Insufficiente	236	0,010	1%	236	1%
Sufficiente	1.835	0,085	8,5%	2.071	9,5%
Buono	12.767	0,588	58,8%	14.838	68,3%
Ottimo	6.871	0,317	31,7%	21.709	100%
Totale	21.709	1	100%	—	—

Tabella 7.5: Esempio di un carattere qualitativo ordinabile. Fonte: A.Di.S.U. Ateneo "Federico II" - Napoli

Il carattere "Giudizio" è un carattere di tipo qualitativo ordinabile, poiché è misurato attraverso le modalità Insufficiente, Sufficiente, Buono, Ottimo, che sono aggettivi che esprimono un grado di percezione della qualità di un servizio. Possiamo procedere con la ricerca del più idoneo indice di sintesi che, per questa tipologia di carattere, abbiamo visto essere la *mediana*. Ricordando, tuttavia, che anche per questo tipo di carattere, essendo operativamente superiore rispetto al carattere qualitativo, possiamo individuare la moda della distribuzione.

Osservando le frequenze assolute registrate in tabella, scopriamo che la moda del carattere "Giudizio" è rappresentata dalla modalità 'Buono" poiché questa modalità presenta la frequenza assoluta più elevata. Ci esprimeremo, quindi, dicendo che *la moda della distribuzione relativa al carattere "Giudizio" è "Buono"* .

Andiamo, quindi, ad individuare la mediana della distribuzione. Il totale del nostro collettivo è dispari, di conseguenza per individuare la modalità che divide esattamente a metà le due distribuzioni individueremo la modalità posta alla $\frac{N+1}{2}$-esima posizione. Quindi la mediana sarà la modalità che ci trova alla posizione:

$$Me = x_{(\frac{N+1}{2})} = \frac{21.709 + 1}{2} = 10.855$$

ossia, osservando la colonna delle frequenze cumulate in tab.??, la mediana della distribuzione corrisponde alla modalità "Buono", in quanto questa modalità raggruppa tutti i giudizi espressi dalla posizione 2.072 alla posizione 14.838, di conseguenza anche la posizione individuata 10.855. Si osservi che avremmo ottenuto lo stesso risultato anche se avessimo lavorato con le frequenze percentuali cumulate, infatti in questo caso avremmo dovuto individuare la modalità che divideva la distribuzione al 50%. La modalità "Buono" è stata espressa da tutti i soggetti che stanno nella fascia dal 9,6% fino al 68,3%, quindi anche il 50% della distribuzione ricade in questa modalità.

Consideriamo adesso la variabilità della nostra distribuzione. A questo proposito introduciamo la misura di *dispersione* che ci consente di valutare appunto in che modo si "disperdono", si distribuiscono, le modalità del collettivo.

Per ottenere un indice di dispersione di un carattere qualitativo misurato su scala ordinale, possiamo ricorrere alle frequenze cumulate; infatti, tenendo conto della definizione di dispersione data sopra, ci si aspetterà un'elevata dispersione se le frequenze più alte saranno concentrate verso le modalità estreme.

Applichiamo al nostro esempio il calcolo dell'indice di dispersione utilizzato nella trattazione teorica:

$$\text{DIS} = 2\sum_{i=1}^{k} F_i(1-F_i)$$

ricordando che nel caso di minima dispersione l'indice *DIS* vale zero.

In caso di massimo l'indice vale, invece:

$$\max(\text{DIS}) = \frac{k-1}{2} \text{ se N è pari}$$

mentre

$$\max(\text{DIS}) = \frac{k-1}{2}\left(1-\frac{1}{N^2}\right) \text{ se Nè dispari.}$$

Nel nostro caso specifico costruiamo una tabella che ci possa aiutare nel procedere con i calcoli (cfr. 7.6):

Giudizio	(f_i)	(F_i)	$(1-F_i)$	$F_i \cdot (1-F_i)$
Insufficiente	0,010	0,010	0,990	0,010
Sufficiente	0,085	0,095	0,905	0,086
Buono	0,588	0,683	0,317	0,217
Ottimo	0,317	1,000	0,000	0,000
Totale	1	—	—	0,312

Tabella 7.6: Esempio di calcolo della dispersione assoluta (DIS) di un carattere qualitativo ordinabile.

La formula $\text{DIS} = 2\sum_{i=1}^{k} F_i(1-F_i)$ indica che è necessario costruire la colonna delle frequenze relative cumulate, eseguire la differenza $1-F_i$ per ciascuna modalità, eseguire la moltiplicazione $F_i(1-F_i)$ per ciascuna modalità, sommare questi risultati ottenuti ed infine moltiplicare il risultato per 2. Sarà, quindi:

$$\text{DIS} = 2\sum_{i=1}^{k} F_i(1-F_i) = 2 \cdot 0,312 = 0,625$$

Cerchiamo adesso di comprendere cosa significa un valore DIS = 0,625. Sappiamo che il valore minimo della DIS è sempre zero, individuiamo quindi il valore massimo che può ottenere DIS nella nostra distribuzione. Ricordando che la nostra distribuzione è dispari poiché N = 21.709 allora dobbiamo considerare la formula

$$\max(\text{DIS}) = \frac{k-1}{2}\left(1-\frac{1}{N^2}\right)$$

da cui, sostituendo, otteniamo

$$\max(\text{DIS}) = \frac{4-1}{2}\left(1-\frac{1}{21.709^2}\right) = \frac{3}{2}\left(1-\frac{1}{471.280.681}\right) = \frac{3}{2}\cdot 1 = 1,5$$

Il valore DIS = 0,625 indica una dispersione modesta, infatti $0 < 0,625 < 1,5$. Ricordiamo però che DIS è un valore assoluto, quindi riferito alla nostra distribuzione specifica e che di conseguenza non ci consente di paragonare la nostra distribuzione ad altre distribuzioni.

A questo scopo occorre individuare l'indice di dispersione relativo (dis).

Passando alla dispersione relativa si ha:

$$dis = \frac{4\sum_{i=1}^{k} F_i(1-F_i)}{k-1} = \frac{4 \cdot 0,312}{4-1} = 0,417$$

dal valore dell'indice di dissomiglianza si evince che nella distribuzione considerata c'è una variabilità discreta, in quanto ricordiamo che un indice relativo ha sempre un campo di variazione $0-1$.

Contrariamente a quanto abbiamo visto per i caratteri qualitativi misurati su scala nominale, per i caratteri qualitativi ordinabili, che possiedono la prerogativa dell'ordinamento tra le modalità, è possibile considerare una misura della forma della distribuzione. Ossia possiamo studiare come si distribuiscono le frequenze tra le diverse modalità, allo scopo di valutare la possibilità di presenza di regolarità o di particolari comportamenti su modalità diverse rispetto alla sintesi.

Per fare un esempio concreto dello studio della forma di una distribuzione procediamo a piccoli passi riprendendo l'aspetto teorico e applicandolo al nostro caso pratico.

Abbiamo detto che, per i caratteri qualitativi ordinabili, il modo più semplice per trovare una misura dell'asimmetria di una distribuzione è quello di considerare una particolare versione dell'indice di dispersione.

Dopo avere costruito la tabella delle frequenze ponendo le modalità in modo ordinato, dividiamo la distribuzione dei dati in due parti uguali: la prima costituita dalle più piccole modalità che cumulano $\frac{N}{2}$ unità se N è pari o $\frac{N+1}{2}$ se N è dispari. La seconda distribuzione costituita dalle modalità più grandi per le restanti unità. Nel nostro esempio, poiché $N = 21.709$, divideremo la distribuzione dei dati in due parti considerando $\frac{N+1}{2}$. Per facilitarci nella comprensione costruiamo due tabelle distinte, una per ogni metà della nostra distribuzione.

Giudizio	(n_i)	(f_i)	(F_i)	$(1-F_i)$	$F_i \cdot (1-F_i)$
Insufficiente	236	0,022	0,022	0,978	0,021
Sufficiente	1.835	0,169	0,191	0,809	0,154
Buono	8.784	0,809	1	0,000	0,000
Totale	10.855	1	–	–	0,175

Tabella 7.7: Esempio di studio della forma di un carattere qualitativo ordinabile: dispersione sinistra.

Giudizio	(n_i)	(f_i)	(F_i)	$(1-F_i)$	$F_i \cdot (1-F_i)$
Buono	3.984	0,367	0,367	0,633	0,232
Ottimo	6.871	0,633	1,000	0,000	0,000
Totale	10.855	1	–	–	0,232

Tabella 7.8: Esempio di studio della forma di un carattere qualitativo ordinabile: dispersione destra.

Su ciascuna di esse è possibile calcolare la misura della dispersione, che chiameremo dispersione destra per la distribuzione (DIS_d) e dispersione sinistra per la distribuzione (DIS_s). Ricordiamo che se nella distribuzione c'è simmetria allora $DIS_d = DIS_s$; mentre se c'è asimmetria positiva $DIS_d > DIS_s$, in quanto, essendoci più unità concentrate tra le modalità più piccole, si registrerà anche una maggiore variabilità tra quelle grandi. Viceversa nel caso di asimmetria negativa le unità saranno concentrate nelle modalità più grandi, quindi risulterà $DIS_d < DIS_s$.

Introduciamo, quindi, il seguente indice (ASI), quale *indice assoluto di asimmetria* di un distribuzione di un carattere qualitativo misurato su scala ordinale:

$$ASI = DIS_d - DIS_s = 0,464 - 0,350 = 0,114$$

Ricordiamo che se $ASI = 0$ allora la distribuzione è simmetrica, mentre se $ASI > 0$ allora $DIS_d > DIS_s$ e si avrà asimmetria positiva; se, infine, $ASI < 0$ allora $DIS_d < DIS_s$ e si avrà asimmetria negativa. Osserviamo allora che in questo caso abbiamo un asimmetria

leggermente positiva positiva poiché il valore della dispersione destra è superiore al valore della dispersione sinistra.

Sappiamo, però, che per avere una misura dell'asimmetria che sia confrontabile con altre distribuzioni, è necessario introdurre il corrispondente *indice relativo di asimmetria*. Da quanto visto sopra, per fare ciò abbiamo necessità di conoscere il $\min(ASI)$ ed il $\max(ASI)$. Circa il valore minimo dell'indice di asimmetria relativa, come abbiamo più volte rimarcato in caso di assenza di asimmetria, ossia in presenza di simmetria, si osserverà $DIS_d = DIS_s$ quindi $ASI = 0$. Mentre il valore massimo sarà dato da $DIS_d + DIS_s$, che nel nostro esempio assumerà il valore 0,814. Calcoliamo, quindi, l'indice relativo di asimmetria:

$$asi = \frac{DIS_d - DIS_s}{DIS_d + DIS_s} = \frac{0,464 - 0,350}{0,464 + 0,350} = \frac{0,114}{0,814} = 0,140$$

Poiché abbiamo visto che l'indice asi assumerà valore 1 in caso di massima asimmetria positiva, mentre assumerà valore -1 in caso di massima asimmetria negativa, allora diciamo che nel nostro esempio siamo di fronte ad un caso di una leggera asimmetria positiva.

Capitolo 8

La valutazione di un carattere quantitativo

Come già più volte detto, l'analisi di un fenomeno reale passa attraverso la replica di misure osservate su un collettivo di unità. Tuttavia, quando si parla di misure, l'abitudine ci induce a pensare che esse siano riferite alle sole scale quantitative ad intervalli e di rapporti. Sebbene non sia sempre così, come abbiamo visto nei capitoli precedenti, occorre sottolineare che, quando ciò è possibile, è sicuramente auspicabile e preferibile riferirsi alle scale quantitative piuttosto che alle scale nominali e ordinali. Il motivo della preferenza risiede essenzialmente nel fatto che, in una distribuzione di un carattere quantitativo, le modalità sono numeri, ciò implica che gli indici di sintesi, di variabilità e di forma possono essere calcolati tenendo conto dei valori quantitativi di tutte le modalità. Ciò premesso va altresì precisato che in virtù di quanto detto nei capitoli precedenti, tutte le misure di sintesi (moda e mediana), di variabilità (omogeneità, eterogeneità e dispersione) e di forma (asimmetria), introdotte per le misure su scala nominale ed ordinale, possono essere ripetute per i caratteri quantitativi. Tuttavia, prima di procedere dal punto di vista metodologico alla ricerca delle misure di sintesi, di variabilità e di forma di caratteri quantitativi, è necessario precisare che, per ragioni di semplicità, la trattazione dei metodi e degli indici, relativi a questi caratteri, verranno introdotti preventivamente sulla distribuzione unitaria semplice dei dati. Prima

di procedere facciamo una dovuta distinzione tra caratteri quantitativi discreti e caratteri quantitativi continui. Nella pratica comune molti fenomeni reali vengono misurati su scala quantitativa discreta. Rientrano in queste categorie le cosiddette variabili conteggio: ad esempio la misura della grandezza di un appartamento in base al numero dei vani o di un'azienda in riferimento al numero dei dipendenti, una scuola per numero di classi, ecc. Talvolta è pratica comune associare numeri a grandezze qualitative ordinali per facilitare l'individuazione degli indici di sintesi, variabilità e forma; un esempio classico è rappresentato dall'uso dei voti nella valutazione scolastica che pur essendo espressi per mezzo di numeri di fatto non descrivono una specifica quantità di apprendimento; lo stesso discorso vale per i giudizi di qualità di un bene o di un servizio. A nostro giudizio la distribuzione di frequenza del carattere misurato su scala quantitativa discreta dovrebbe essere considerato alla stessa stregua dei caratteri misurati su scala qualitativa ordinale. Ci esprimiamo, tuttavia, in forma ipotetica in quanto è prassi comune trattare la valutazione di tali caratteri come caratteri quantitativi. Ad ognuno è capitato nella sua carriera scolastica di effettuare la media aritmetica dei propri voti e di aver ottenuto nella maggior parte dei casi un numero razionale al quale non corrispondeva di fatto nessuno dei voti assegnati dal professore. Per esplicitare meglio il nostro pensiero vi invitiamo ad immaginare di fare una media del numero di vani per appartamento del vostro palazzo; immaginate di ottenere ad esempio una media di $4,29$ vani, che significato ha il vostro risultato? Abbiamo ben chiaro quanti sono 4 vani ma $0,29$ vani a quanto corrispondono? Tornando all'esempio dei voti, quanto vale un $6,49$? È più tendente al 6 o al 7? E se tuttavia riuscissimo ad avere un'idea chiara del suo valore perchè siamo condizionati dal consolidato uso delle medie in campo scolastico, possiamo trovare un'idea oggettiva che ci dica univocamente quanto vale un $6--$ oppure un $7++$. Dire $6--$ equivale a dire $5\frac{1}{2}$? E se volessimo fare una media dei voti $6--$, $7++$ e $5\frac{1}{2}$, che voto daremmo allo studente? La media aritmetica deve essere universale e trasferibile, quindi l'uso nella valutazione scolastica risulta altamente impreciso, mentre sarebbe più opportuno individuare la mediana come indice di sintesi. Nell'analisi degli indici di sintesi, variabilità e forma dei caratteri quantitativi, tuttavia, pur essendo consapevoli delle differen-

ze concettuali che stanno alla base dei caratteri quantitativi discreti e continui, tratteremo entrambe le tipologie di caratteri quantitativi al medesimo modo. Riprendendo il discorso sulla distribuzione unitaria semplice osserviamo come si rappresenta schematicamente:

$$X\,[x_1, x_2, \ldots, x_n]$$

dove con X abbiamo indicato il nome del carattere, con x_i la misura espressa su scala quantitativa osservata sulla i-esima unità ed infine con il pedice $i = 1, 2, \ldots, N$ l'unità del collettivo osservato.

Dopo una trattazione completa degli indici, ricavati sulla distribuzione semplice, si passerà all'estensione del calcolo degli stessi su distribuzioni di frequenza.

Va, a questo proposito, ricordato che le distribuzioni di frequenza vengono ricavate da quelle semplici raggruppando le modalità uguali; quindi, da un punto di vista interpretativo, tra esse non si ravvisa nessuna differenza concettuale. Da quanto detto in precedenza, dobbiamo ricordare che un carattere quantitativo può essere misurato su una scala discreta (composta da un numero finito di modalità) o su scala continua; in ognuno dei due casi le distribuzioni schematicamente si presenteranno come illustrato nella Tabelle 8.1.

Carattere X	**Numero** n_i
x_1	n_1
x_2	n_2
$\vdots$	$\vdots$
x_i	n_i
$\vdots$	$\vdots$
x_k	n_k
	N

Tabella 8.1: Distribuzione di frequenza.

Talvolta può risultare più opportuno evitare una distribuzione di frequenza come quella rappresentata nella tabella 8.1 al fine di ottenere un maggior raggruppamento delle modalità. Per far comprendere meglio quanto detto si procede con un semplice esempio.

Se si vuole osservare su un collettivo il fenomeno altezza, possiamo procedere in due modi diversi:

1. possiamo registrare in una tabella di frequenza tutte le altezze rilevate in ordine crescente e associare ad ognuna di esse le relative frequenze assolute, ma in questo caso otterremmo molte modalità diverse con basse frequenze;

2. possiamo raggruppare le altezze registrate in classi (ad esempio 130-140, 140-150, ecc.) e registrare le relative frequenze assolute.

Quando si opera sulla distribuzione di frequenza divisa in classi nasce il problema di quale modalità della classe deve essere presa in considerazione nei calcoli degli indici. In questi casi si assume l'ipotesi che il sottoinsieme del collettivo (ossia la frequenza assoluta) si equidistribuisca all'interno della classe. Ciò implica che la modalità più rappresentativa sia data dal suo valore centrale $c_i = \frac{x_{i+1}-x_i}{2}$ per $i = 1, 2, \ldots, k$.

Operativamente si procede, quindi, per ciascuna classe come segue:

Carattere X	**Numero n_i**
$x_{min} \mapsto x_2$	n_1
$x_2 \mapsto x_3$	n_2
$\vdots$	$\vdots$
$x_i \mapsto x_{i+1}$	n_i
$\vdots$	$\vdots$
$x_{k-1} \dashv x_{max}$	n_k
	N

Tabella 8.2: Distribuzione di frequenza divisa in classi.

Quando ci riferiamo a caratteri quantitativi raggruppati in classi occorre tener ben presente qual è il limite inferiore o superiore della classe per evitare situazioni in cui non sappiamo con precisione dove collocare ogni unità oservata. Se per esempio, infatti, definiamo le classi della altezze $140-150$, $150-160$, ... abbiamo indecisione in

Carattere X	**Valore centrale c_i**	**Numero n_i**
$x_{min} \mapsto x_2$	$c_1 = \frac{x_{min}+x_2}{2}$	n_1
$x_2 \mapsto x_3$	$c_2 = \frac{x_2+x_3}{2}$	n_2
$\vdots$	$\vdots$	$\vdots$
$x_i \mapsto x_{i+1}$	$c_i = \frac{x_{i+1}+x_i}{2}$	n_i
$\vdots$	$\vdots$	$\vdots$
$x_{k-1} \dashv x_{max}$	$c_k = \frac{x_{k-1}+x_k}{2}$	n_k
		N

Tabella 8.3: Distribuzione di frequenza divisa in classi con valore centrale.

quale classe collocare i soggetti alti 150 cm. Per non avere ambiguità di questo tipo possiamo "chiudere" la classe ad un estremo o all'altro come segue.

1. Se la classe è *chiusa a sinistra* tra i suoi due estremi si pone il simbolo $\vdash$, in questo caso significa che la classe contiene tutti i valori di x tali che $a \leqslant x < b$, quindi b non appartiene alla classe. Considerando il nostro esempio, quindi, si scriverà $140 \vdash 150$, $150 \vdash 160$, ... ed apparterranno alla prima classe i soggetti alti da 140 a 149, nella seconda quelli alti da 150 a 159, e così via.

2. Se la classe, invece, è *chiusa a destra* tra i due estremi si pone il simbolo $\dashv$, in questo caso significa che la classe contiene tutti i valori di x tali che $a < x \leqslant b$, quindi a non appartiene alla classe. Considerando il nostro esempio, quindi, si scriverà $140 \dashv 150$, $150 \dashv 160$, ... ed apparterranno alla prima classe i soggetti alti da 141 a 150, nella seconda quelli alti da 151 a 160, e così via.

3. Se la classe, invece, è *chiusa* tra i due estremi si pone il simbolo $\vdash\dashv$, in questo caso significa che la classe contiene tutti i valori di x tali che $a \leqslant x \leqslant b$, quindi a e b appartengono entrambe alla classe.

8.1 Rappresentazioni grafiche dei caratteri quantitativi

Non sempre quando leggiamo i dati presenti in una tabella riusciamo ad avere informazioni immediate della distribuzione che stiamo analizzando. Abbiamo visto, infatti, che l'uso di rappresentazioni grafiche consente una lettura funzionale più immediata per l'impatto visivo che hanno.

L'uso di grafici consente di ottenere una sintesi delle informazioni contenute in una tabella di frequenza, quindi siamo portati a credere che sintetizzando si possa incorrere nella perdita di informazioni. Per questo è opportuno che l'uso di rappresentazioni grafiche sia sempre accompagnato dalla tabella che contiene i dati che hanno consentito di generare il grafico stesso. L'uso associato di tabelle e grafici consente di avere una visione chiara e globale del fenomeno che si sta osservando.

Abbiamo già avuto modo di vedere alcune rappresentazioni grafiche studiando i caratteri qualitativi; procediamo, quindi a vedere quali tipologie di grafico sono più opportune da utilizzare per caratteri quantitativi.La scelta del grafico più idoneo a rappresentare una distribuzione non è casuale, ma è dettata dalla specificità della distribuzione stessa, in particolar modo dal tipo di carattere che stiamo analizzando.

Alcune tipologie di rappresentazioni grafiche possono essere utilizzate sempre con ogni tipo di carattere. Fanno parte di questa categoria i grafici a barre, a nastri e le torte.

Quando operiamo su caratteri quantitativi discreti possiamo utilizzare un *diagramma a pettine* (Fig. 8.1). Per costruire questo diagramma operiamo su un piano cartesiano. Nella rappresentazione poniamo sull'asse delle ascisse le modalità per mezzo delle quali si esprime il carattere, mentre sull'asse delle ordinate poniamo le frequenze assolute, relative o percentuali che abbiamo registrato in tebella per igni modalità.

Il grafico viene costruito disegnando, quindi, in corrispondenza di ogni modalità, un segmento la cui altezza sia pari alla frequenza assoluta, relativa o percentuale di ogni modalità per mezzo della quale si esprime il carattere.

Un grafico molto utilizzato per caratteri quantitativi è sicuramente l'*istogramma* che possiamo pensare come un diagramma a barre nel

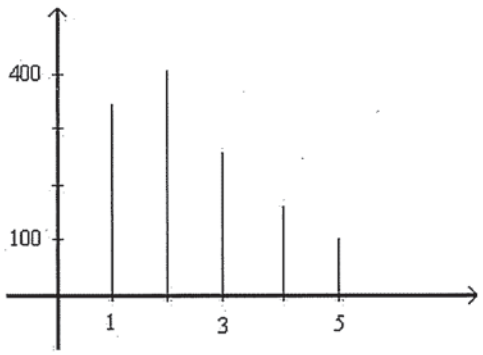


Figura 8.1: Diagramma a pettine.

quale le barre non siano distanziate. Nella sua costruzione, tuttavia, occorre fare un'importante distinzione tra carattere qualitativo ordinabile, quantitativo discreto e quantitativo continuo. Per il primo sappiamo che si può operare solo un ordinamento, pertanto per convenzione possiamo rappresentare le modalità per mezzo di rettangoli che abbiano una base di dimensione fissata e altezza corrispondente alla frequenza assoluta, relativa o percentuale per mezzo della quale decidiamo di descrivere il carattere preso in esame. Per i caratteri quantitativi discreti e per quelli continui che abbiano un ampiezza di intervallo fissa, possiamo dire che sostanzialmente operiamo con le stesse modalità appena descritte per quelli qualitativi ordinabili.

La situazione cambia quando operiamo con caratteri quantitativi continui per i quali le modalità siano raggruppate in classi di ampiezza diversa. In questo caso le basi dei rettangolo saranno diverse in relazione all'ampiezza specifica di ogni classe. L'istogramma si presenta, quindi, come una serie di rettangoli uniti affiancati tra loro per i quali l'area di ognuno è proporzionale alla frequenza assoluta, relativa o percentuale che decidiamo di utilizzare. Per comprendere in che modo si costruiscono le aree dei rettangoli, occorre introdurre il concetto di *densità* di una classe. L'altezza di ogni rettangolo sarà, infatti, rappresentata proprio dalla densità assoluta, relativa o percentuale della classe. Per individuare la misura della densità assoluta della classe, che indicheremo con h_i, occorre fare un rapporto tra tra

la frequenza assoluta e l'ampiezza della classe, ossia $h_i = n_i/a_i$, dove a_i sta ad indicare appunto l'ampiezza della classe. L'area di ogni rettangolo che esprime ognuna delle modalità del carattere, sarà data da $A_i = a_i \times h_i$.

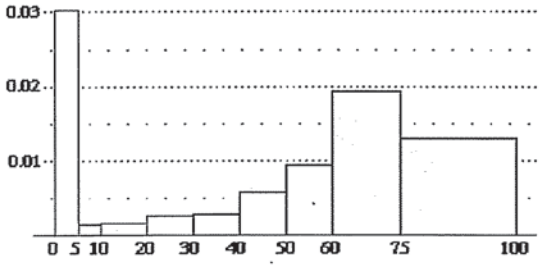


Figura 8.2: Istogramma per un carattere quantitativo continuo.

Nella figura 8.2 è illustrata la distribuzione delle età dei morti in Italia nel 1951 per classi di età con diverse ampiezze delle classi (inferiore a 5 anni, tra 5 anni e 10 anni, da 10 a 20 anni e così via). Le aree dei rettangoli rappresentano la frequenza di morti rispetto alle diverse classi di età. Sull'asse verticale sono riportate, come detto sopra, le densità di frequenza per ogni classi. Leggendo il grafico diremo, ad esempio, che i morti tra gli 0 e i 5 anni (intervallo di 5 anni) costituiscono circa il 15% dei casi in quanto $5 \cdot 0.03 = 0.15 = 15\%$.

Un'altra tipologia di grafico che può essere utilizzata per i caratteri continui è il *diagramma ramo-foglia*. Questo diagramma consente di vedere con immediatezza la distribuzione di frequenza di un carattere. Per costruire questo diagramma è necessario ordinare in modo crescente i dati oggetto di studio. Il diagramma si costruisce dividendo ciascuna osservazione nella sua parte principale, ramo, e in quella secondaria, foglia. Riportiamo, per fare un esempio concreto, il diagramma ramo-foglia dei risultati standardizzati conseguiti alla prova di matematica dell'esame di stato del primo ciclo di istruzione dell'anno scolastico 2009-2010 (dati del Rapporto Nazionale dell'INVALSI).

Nella parte sinistra del grafico si legge la distribuzione dei punteggi standardizzati ottenuti dagli studenti che hanno sostenuto l'esame di stato, mentre nella parte destra la collocazione sulla stessa scala

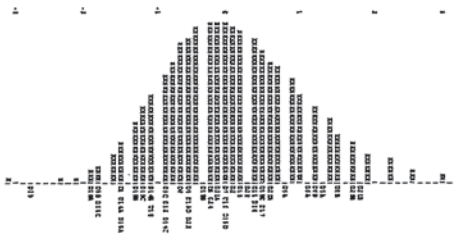


Figura 8.3: Diagramma ramo-foglia.

(asse verticale tratteggiato) della difficoltà delle domande. Il grafico mostra la probabilità di rispondere correttamente ad una domanda in relazione al suo livello di difficoltà. Sostanzialmente tutti gli studenti che si trovano ad un livello inferiore rispetto a quello occupato da una specifica domanda avranno una scarsa probabilità di rispondere in modo corretto alla stessa (probabilità di risposta corretta inferiore al 50%), mentre gli studenti che si trovano in una posizione superiore rispetto alla domanda presa in esame avranno la probabilità di rispondere in modo corretto (probabilità di risposta corretta superiore al 50%); questa probabilità aumenta all'aumentare della distanza studente-domanda.

Quando lavoriamo con caratteri quantitativi possiamo ancora rappresentare le *serie storiche*. Una serie storica rappresenta la descrizione dell'andamento di un fenomeno osservato in un determinato intervallo temporale. Le serie storiche vengono studiate sia per interpretare un fenomeno, sia per fare previsioni sul suo andamento futuro. Una serie consente di classificare diverse osservazioni di un fenomeno rispetto ad una variabile. Nel caso delle serie storiche la variabile presa in considerazione è il tempo, la serie viene quindi definita storica o temporale. Un fenomeno può essere osservato in determinati istanti di tempo o alla fine di specifici periodi. Non entreremo qui nel dettaglio dell'analisi delle serie storiche, ma ci limiteremo a dare un esempio di come poter rappresentare graficamente una serie storica. Osserviamo, quindi, il grafico 8.4 relativo alla registrazione degli negli anni 2001 e 2002 (fonte: www.numedionline.it). Il grafico ci consente, con immediata percezioine, di vedere come gli infortuni denunciati

all'INAIL siano passati dai 109 dei primi nove mesi del 2001 ai 52 dello stesso periodo del 2002, con una riduzione superiore al 50%.

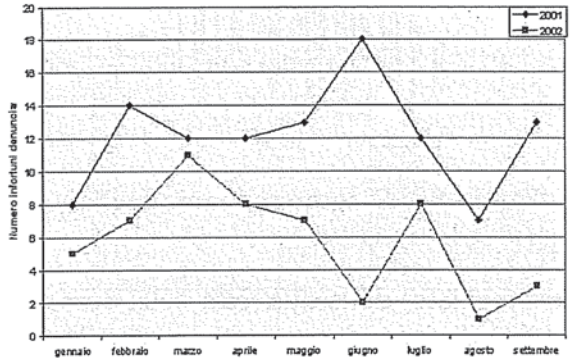


Figura 8.4: Diagramma serie storica.

8.2 Sintesi

8.2.1 La classe mediana

Ricordiamo che la mediana è la modalità che ripartisce equamente le unità del collettivo, si intuisce di conseguenza che, nel caso di una distribuzione divisa in classi, difficilmente si verifica il caso in cui la modalità mediana cada in corrispondenza di un valore centrale o estremo di una classe. Allo scopo di ottenere una procedura semplice costruiamo un grafico delle frequenze cumulate associato a ciascuna classe.

Nel grafico abbiamo riportato la spezzata delle frequenze cumulate. Tracciando una retta parallela all'asse delle ascisse in corrispondenza della frequenza cumulata $N/2$ si intersecherà la spezzata nel punto D, la cui proiezione sull'asse della ascisse rappresenta la modalità mediana *Me*. Dallo stesso grafico si capisce, allora, che la mediana è data dalla modalità x_i più il segmento che parte da x_i fino a *Me*. Dato che x_i è noto, essendo il valore estremo inferiore della classe che contiene la mediana, quello che resta da calcolare è il segmento

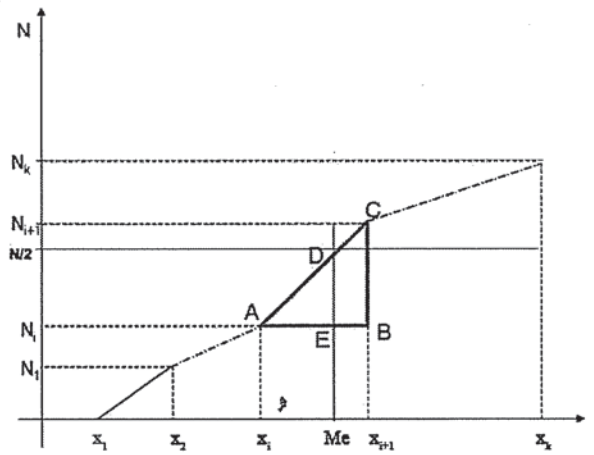


Figura 8.5: Individuazione della posizione mediana all'interno di una classe mediana.

che nel grafico abbiamo indicato con $AE = x$. Da una attenta lettura del grafico vediamo che, nella classe mediana di estremi x_i e $x_i + 1$, si formano due triangoli rettangoli simili ABC e ADE, di cui un lato AE rappresenta proprio la nostra incognita. Grazie alla condizione di similitudine è possibile quindi impostare la seguente proporzione

$$AE : AB = DE : BC$$

dove AE, come si è detto, è il segmento che ci interessa calcolare, mentre AB, DE e BC sono segmenti la cui lunghezza è nota. Infatti è immediato verificare che

$$AB = x_{i+1} - x_i, DE = N/2 - N_i, BC = N_{i+1} - N_i$$

Calcolando la proporzione e risolvendo rispetto al segmento incognito AE si ha

$$AE \cdot BC = AB \cdot DE \Rightarrow AE = \frac{AB \cdot DE}{BC}$$

sostituendo a ciascun segmento la rispettiva lunghezza, si ottiene quella del segmento desiderato

$$AE = \frac{AB \cdot DE}{BC} = \frac{(x_{i+1} - x_i)(N/2 - N_i)}{N_{i+1} - N_i}$$

In definitiva la mediana sarà data da

$$Me = x_i + \frac{(x_{i+1} - x_i)(N/2 - N_i)}{N_{i+1} - N_i} \qquad (8.2.1.1)$$

Facciamo un piccolo esempio su una distribuzione per età. Il

Carattere X	n_i	N_i
$10 \mapsto 20$	34	34
$20 \mapsto 60$	23	57
$60 \mapsto 80$	15	72
$80 \mapsto 100$	10	82
Totale	82	

Tabella 8.4: Esempio di distribuzione di frequenza.

primo calcolo da fare è $N/2 = 82/2 = 41$, il che individua la classe mediana di estremi $20 \mapsto 60$ il che implica che $x_i = 20, x_{i+1} = 60, N_i = 34, N_{i+1} = 57$. La mediana risulta essere

$$Me = 20 + \frac{(60 - 20)(41 - 34)}{57 - 34} = 32,17$$

Puntualizziamo che qualora la distribuzione del carattere fosse stata misurata su scala ad intervalli con modalità discrete, per calcolare la mediana avremmo impiegato la stessa procedura utilizzata per i caratteri misurati su scala qualitativa ordinale, vista nel capitolo precedente.

8.2.2 Il concetto di media

Le misure specifiche di sintesi di un carattere misurato su scala quantitativa sono le *medie*. La media, infatti, è quel valore che, tra

tutti gli altri, meglio rappresenta una distribuzione misurata su scala quantitativa ad intervalli. Il concetto di media è largamente diffuso nel linguaggio e nell'uso comune anche se spesso questo strumento è utilizzato in modo errato. Chi di noi, ad esempio, non ha mai fatto la media dei voti a scuola? Partiamo proprio da un esempio scolastico che identifica uno di quei casi più diffusi di uso improprio della media poiché, come abbiamo più volte ribadito, i voti scolastici pur essendo espressi con numeri non sono misurabili su una scala ad intervalli pertanto, pur se il largo uso che si fa delle medie in campo scolastico ha portato nel tempo ad una sorta di legittimazione di questa pratica, sappiamo che concettualmente sarebbe più corretto considerare i voti come caratteri qualitativi ordinabili e di conseguenza individuare la mediana come valore di sintesi. Supponiamo, tuttavia, che uno studente in tre compiti in classe abbia preso rispettivamente questi voti: 6, 4, e 5. Tutti noi concludiamo che sulla pagella lo studente avrà come voto 5. Qualora il professore decida di mettergli 6, ci sarà sicuramente uno studente che si lamenterà della decisione; probabilmente lo stesso farà illazioni sull'avvenuto. Al contrario, se il professore deciderà di mettergli 4, lo studente si lamenterà del fatto che il professore lo abbia preso di mira. Eccetto nel caso della decisione del 5, negli altri due gli studenti hanno ragione di lamentarsi ed in entrambi i casi essi troveranno molte persone che gli daranno ragione. Ma allora perché solo la decisione del 5 è quella giusta? Molti di voi risponderanno perché il voto 5 rappresenta la media. Ma allora che cosa è la media? Facciamo questo semplice ragionamento. Rileggiamo i voti presi come punti acquisiti, diremo allora che al primo compito lo studente ha acquisito 6 punti, al secondo 4 punti ed infine al terzo 5 punti. Lo studente nelle tre prove ha acquisito complessivamente 15 punti e, nel calcolo della media, per nessuna ragione è disposto a perderne qualcuno. Se lo studente avesse preso a ciascun compito 5 punti comunque, sommandoli, cumulerebbe 15 punti. Mentre con 4 cumulerebbe 12 punti e con 6 cumulerebbe 18 punti (meno di quanti realmente accumulati nel primo caso e troppi nel secondo). In questo senso diremo che la media è il valore che, se sostituito a ciascuna modalità, lascia invariata la somma.

Come meglio preciseremo più avanti, potremmo trovare casi in cui il valore della media, che meglio rappresenta la distribuzione, potrebbe

essere una modalità che, se sostituita a tutte le altre della distribuzione stessa, lascia invariato il prodotto o la somma dei reciproci ecc. In sintesi, contrariamente a quanto in genere si crede, il concetto di media è piuttosto complesso e variegato e non si riduce al solo caso, forse più comune, della somma dei valori diviso il numero degli stessi. Facciamo un piccolo esempio in cui la media, per l'appunto, non corrisponde alla somma dei dati diviso il loro numero.

Supponiamo che una coppia di fidanzati decida di sposarsi tra tre anni e che vogliano avviare un piano di accumulo di capitale da utilizzare il giorno del loro matrimonio. Supponiamo inoltre che i due riescano a racimolare tra i parenti complessivamente 10.000 euro. I due allora decidono di trovare un istituto di credito dove depositare la cifra raccolta che gli faccia un piano di accumulo soddisfacente. Per prima si rivolgono ad una banca (Banca A), parlano con il direttore il quale, non potendogli garantire un tasso di interesse unico per i tre anni, gli propone il seguente piano a capitalizzazione composta con tassi d'interesse: 4% il primo anno; 6% il secondo anno e 5% il terzo anno. I ragazzi, al fine di valutare la convenienza della proposta, si rivolgono ad un altro istituto di credito del centro città (Banca B), nel quale gli offrono un piano di capitalizzazione semplice a tasso fisso per tre anni del 5%.

Infine si rivolgono ad una Banca (Banca C) dove lavora un amico del fidanzato il quale propone loro un piano di accumulo del 5% fisso a capitalizzazione composta.

I due ragazzi tornando a casa riflettono sulle tre proposte e ad un'analisi frettolosa, la ragazza, d'istinto, dice che esse sono perfettamente identiche essendo il 5% giusto la media di 4%, 6% e 5%. Dovendo preferire una delle tre Banche sceglie quella del centro, essendo situata vicino ai negozi le permetterebbe una bella passeggiata dopo il ritiro del denaro. Il ragazzo preoccupato della passeggiata del "dopo ritiro del denaro" ed avendo fatto studi di statistica, riflette sulle tre proposte e prova una comparazione, producendo i seguenti conti:

Capitalizzazione Banca A

Alla fine del primo anno il capitale maturato sarà il capitale versato più il 4% di interessi cioè: $10.000 + 0,04 \times 10.000 = 10.400$. Essendo un piano di capitalizzazione composto, si assume che il capitale maturato non sia ritirato e che sia il nuovo capitale per l'anno successivo. Così

alla fine del secondo anno si avrà: $10.400 + 0,06 \times 10.400 = 11.024$. il quale per lo stesso principio descritto sopra sarà il capitale di partenza per l'ultimo anno che alla fine maturerà: $11.024 + 0,05 \times 11.024 =$ **11.575,2**.

Capitalizzazione Banca B

La Banca B invece gli ha proposto un interesse fisso del 5% sul capitale di 10.000 euro con capitalizzazione semplice ; ciò implica che la quota maturata a fine anno non è reimpiegata per il calcolo degli interessi, cioè gli interessi maturati ogni anno restano fissi e sempre pari a 500 euro, ne consegue: $10.000 + 0,05 \times 10.000 = 10.500$. In conclusione il capitale finale da utilizzare per il matrimonio è:$10.000 + 1.500 =$ **11.500**.

Capitalizzazione Banca C

La terza Banca infine propone il seguente piano di accumulo: alla fine del primo anno $10.000 + 0,05 \times 10.000 = 10.500$ alla fine del secondo anno $10.500 + 0.05 \times 10.500 = 11.025$ alla fine del periodo $11025 + 0.05 \times 11.025 =$ **11.576,25**.

Il fidanzato, orgoglioso dei suoi studi, richiama l'attenzione della fidanzata facendo notare che non è affatto vero che i tre piani sono identici e che il più conveniente è quello della Banca C del suo amico. Ma l'idea che ci lavora un amico e, ancora di più, il fatto di non poter fare la passeggiata dopo il ritiro del denaro, fa avviare un discussione tra i due fidanzati che finisce come spesso accade recentemente nel non si fa più niente.

Ora vediamo cosa è successo da un punto di vista formale e perché si raggiungono risultati diversi I tre tassi di interesse del 4%, del 5% e del 6% proposti dalla banca A non corrispondono al tasso di interesse medio del 5% anche se il piano di capitalizzazione è il medesimo come quello della Banca C. Infatti, applicando in generale il procedimento adottato nel caso Banca A e semplificando, senza perdere in generalità, nel caso in cui il capitale sia di 1 euro si ha:

- primo anno: $(1 + 0,04) = M_1$
- secondo anno: $M_1 + 0,06 \times M_1 = M_1(1 + 0,06) = (1 + 0,04)(1 + 0,06) = M_2$

- terzo anno: $M_2 + 0,05 \times M_2 = M_2(1+0,05) = (1+0,04)(1+0,06)(1+0.05) = M_3$

Volendo trovare un tasso di interesse medio bisognerebbe sostituire per ciascun anno un tasso diciamo x tale che costituirebbe lo stesso capitale. Cioè:

$$(1+0,04)(1+0,06)(1+0.05) = (1+x)^3$$

ossia

$$x = \sqrt[3]{(1+0,04)(1+0,06)(1+0.05)} - 1 = 0.049948 = 4,995\%$$

che, sebbene non molto dissimile dal 5%, è diverso. La diversità risiede nel fatto che la media non è stata ottenuta come somma dei termini $(4+6+5)/3$ ma attraverso la radice cubica del prodotti dei termini diminuita di 1.

Allo stesso modo si potrebbero fare esempi in cui il calcolo della media si potrebbe ottenere sulla base di calcoli diversi da quelli assunti dall'opinione pubblica.

Tenuto conto di quanto appena detto, è nostro obiettivo individuare un procedimento generale che conglobi l'insieme di tutte le medie come sintesi di un carattere quantitativo misurato su scala ad intervalli.

In generale la media è la modalità del carattere che lascia invariata una definita operazione logica sui dati, (la somma dei punteggi nel primo caso, il prodotto dei tassi di interesse nel secondo ecc.). Adottando per "operazione sui dati" il termine più corretto di "funzione dei dati", si ricava la base essenziale del calcolo delle medie attraverso quella che viene chiamata *funzione invariante della media*. Ossia:

$$f(x_1, x_2, \ldots, x_N) = f(\bar{x}, \bar{x}, \ldots, \bar{x})$$

L'espressione per l'appunto richiama tutti i principi generali espressi sopra. Infatti, se la funzione dei dati è la *somma* allora si otterrà la più nota delle medie cioè la *media aritmetica*.

La *media aritmetica*, indicata con la lettera greca "μ", può essere espressa come:

$$x_1 + x_2 + \ldots + x_N = \bar{x} + \bar{x} + \ldots + \bar{x}$$

in modo più sintetico

$$\bar{x} = \mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (8.2.2.1)$$

Se la funzione dei dati è il *prodotto* allora si otterrà la *media geometrica*

$$x_1 \cdot x_2 \cdot \ldots \cdot x_N = \bar{x} \cdot \bar{x} \cdot \ldots \cdot \bar{x} \quad \Rightarrow \quad \prod_{i=1}^{N} x_i = \bar{x}^N$$

da cui

$$M_g = \sqrt[N]{\prod_{i=1}^{N} x_i} \qquad (8.2.2.2)$$

Se la funzione dei dati è la *somma dei reciproci* allora si otterrà la *media armonica*

$$\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_N} = \frac{1}{\bar{x}} + \frac{1}{\bar{x}} + \ldots + \frac{1}{\bar{x}} \quad \Rightarrow \quad \sum_{i=1}^{N} \frac{1}{x_i} = \frac{N}{\bar{x}}$$

da cui

$$M_{ar} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \qquad (8.2.2.3)$$

Se la funzione è la *somma dei quadrati* allora si otterrà la *media quadratica*

$$x_1^2 + x_2^2 + \ldots + x_N^2 = \bar{x}^2 + \bar{x}^2 + \ldots + \bar{x}^2 \quad \Rightarrow \quad \sum_{i=1}^{N} x_i^2 = N\bar{x}^2$$

da cui

$$M_q = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \qquad (8.2.2.4)$$

Se la funzione è la *somma di potenza di ordine r* allora si otterrà la *media di potenza r-esima*

$$x_1^r + x_2^r + \ldots + x_N^r = \bar{x}^r + \bar{x}^r + \ldots + \bar{x}^r \quad \Rightarrow \quad \sum_{i=1}^{N} x_i^r = N\bar{x}^r$$

da cui

$$M_r = \sqrt[r]{\frac{\sum_{i=1}^{N} x_i^r}{N}} \qquad (8.2.2.5)$$

La media di potenza assume il ruolo di generatrice delle medie, infatti da essa è possibile ricavare tutte le medie appena enunciate; in particolare per $r = -1$ si ottiene la media armonica, per $r \to 0$ si ottiene la media geometrica, per $r = 1$ la media aritmetica e infine per $r = 2$ la media quadratica. La grandezza di r stabilisce anche un ordine delle medie; infatti è possibile dimostrare che dato un insieme di dati $x_1, x_2, \ldots, x_N$, con x_{min}, x_{max} rispettivamente il valore più piccolo e quello più grande, allora si ha la seguente relazione:

$$x_{min} \leqslant M_{ar} \leqslant M_g \leqslant \mu \leqslant M_q \leqslant x_{max}$$

Quest'ultima relazione è anche nota come proprietà di *internalità* delle medie, nel senso che qualsiasi media è sicuramente un valore maggiore o al minimo uguale della modalità più piccola e minore o al massimo uguale alla modalità più grande.

La media aritmetica

Da quanto appena detto, il valore di sintesi più comune calcolato su un carattere misurato su scala ad intervalli è la media aritmetica. Essa rappresenta la modalità del carattere che lascia invariata la somma dei dati; i motivi per cui questo valore di sintesi è maggiormente usato risiedono nelle proprietà di cui esso gode.

Proprietà 1

La somma degli scarti dalla media aritmetica è uguale a zero.

Per dimostrare questa proprietà introduciamo la quantità *scarto dalla media* $s_i = x_i - \mu$ per $i = 1, 2, \ldots, N$, allora la somma di tutti gli

scarti è espressa da

$$\sum_{i=1}^{N} (x_i - \mu)$$

che può essere scritta anche come

$$\sum_{i=1}^{N} (x_i - \mu) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \mu = \sum_{i=1}^{N} x_i - N\mu$$

Ora ricordando che

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$

e che

$$N\mu = \sum_{i=1}^{N} x_i$$

sostituendo a μ si avrà

$$\sum_{i=1}^{N} x_i - \sum_{i=1}^{N} x_i = 0$$

Dal punto di vista interpretativo, la proprietà appena enunciata significa che la media aritmetica è quel valore numerico che bilancia in eccesso ed in difetto tutte le modalità. Il senso del bilanciamento è da intendersi come la compensazione dei valori più grandi della media rispetto a quelli più piccoli. Nell'esempio dei tre compiti in classe dello studente il voto 6 eccede di 1 voto la media che è 5. Il punto in eccesso serve per compensare il difetto di un voto nella prova in cui lo studente ha preso 4. Come meglio preciseremo più avanti, questa caratteristica è di esclusiva prerogativa della media aritmetica.

Proprietà 2

La somma dei quadrati degli scarti dalla media aritmetica è un minimo. Questo significa che se, negli scarti, sostituiamo il valore della media aritmetica con qualsiasi altro numero, anche minore della media stessa, otterremo sempre come risultato un numero maggiore di quello che otteniamo sommando gli scarti dalla media elevati al quadrato.

Per fare un esempio pratico torniamo a prendere in considerazione i tre voti dello studente (4, 5, 6), sommando gli scarti dalla media elevati al quadrato otteniamo $[(4-5)^2+(5-5)^2+(6-5)^2]=2$. Ora se al valore della media, che nel nostro caso vale 5, sostituiamo un qualsiasi numero maggiore o minore di 5, avremo sempre un risultato maggiore di 2. Infatti se prendiamo, ad esempio, $3<5$ avremo $[(4-3)^2+(5-3)^2+(6-3)^2]=14$, dove $14>2$; se prendiamo $7>5$ avremo $[(4-7)^2+(5-7)^2+(6-7)^2]=14$, dove $14>2$.

Di seguito dimostreremo, appunto, che la media aritmetica è quel valore numerico che rende minima la somma degli scarti al quadrato.

Consideriamo sempre l'insieme dei dati $x_1, x_2, \ldots, x_n$ la cui media aritmetica semplice è μ; i quadrati degli scarti lineari sono: $(x_1-\mu)^2, (x_2-\mu)^2, \ldots, (x_n-\mu)^2$. Indicando con a un qualunque numero diverso da μ, definiamo gli scarti da tale numero: scarto $= x_i - a$ con $i=1,2,\ldots,n$.

La somma dei quadrati di tali scarti sarà:

$$s=(x_1-a)^2+(x_2-a)^2+\cdots+(x_n-a)^2$$

Essendo a diverso dalla media aritmetica μ, differirà da essa di una certa quantità d, in altre parole: $a-\mu=\pm d$ e anche: $a=\mu\pm d$.

Sostituendo tale espressione nella somma dei quadrati degli scarti dal numero a si avrà:

$$s=\sum_{i=1}^{n}(x_i-a)^2=$$

$$=\sum_{i=1}^{n}[(x_i-\mu)\pm d]^2=$$

$$=\sum_{i=1}^{n}(x_i-\mu)^2\pm 2d\sum_{i=1}^{n}(x_i-\mu)+nd^2=$$

$$=\sum_{i=1}^{n}(x_i-\mu)^2+nd^2$$

Resta così provato che:

$$s>\sum_{i=1}^{n}(x_i-\mu)^2$$

e possiamo quindi affermare che: la somma dei quadrati degli scarti dalla media aritmetica è un valore minimo rispetto alla somma dei quadrati degli scarti da un qualsiasi altro numero.

Dal punto di vista interpretativo questa proprietà indica che la media aritmetica è il valore numerico che più di altri è vicino a tutte le modalità, o, se vogliamo, è il valore meno distante; quindi quello da preferirsi per rappresentare la distribuzione del carattere quantitativo osservato.

Proprietà 3

La media aritmetica è invariante per la trasformazione lineare $y_i=a+bx_i$ per $i=1,2,\ldots,N$.

Questa proprietà, come vedremo con un esempio concreto, risulta essere molto utile da un punto di vista operativo, o se vogliamo, computazionale. Essa sostanzialmente afferma che se ai dati della distribuzione viene operata una trasformazione lineare arbitraria, cioè se i dati vengono moltiplicati per un numero b, scelto a piacere, e poi a ciascuno viene aggiunto un altro numero, anch'esso scelto arbitrariamente, allora la media dei dati trasformati sarà moltiplicata per b e vi sarà aggiunto a.

Formalmente diremo: sia $y_i=a+bx_i$ per $i=1,2,\ldots,N$, allora $\mu_y=a+b\mu_x$.

La dimostrazione di questa proprietà è molto semplice; infatti calcoliamo il valore della media dei dati trasformati

$$\mu_y=\frac{1}{N}\sum_{i=1}^{N}y_i=$$

$$=\frac{1}{N}\sum_{i=1}^{N}(a+bx_i)\Rightarrow\frac{1}{N}\left(Na+b\sum_{i=1}^{N}x_i\right)=$$

$$=a+b\frac{1}{N}\sum_{i=1}^{N}x_i=$$

$$=a+b\mu_x$$

Per capire il senso di questa proprietà facciamo un esempio: supponiamo che uno studente voglia calcolare la media dei suoi primi quattro esami all'università. Supponiamo che essi siano 26; 28; 30; 26. Lo studente che non conosce la proprietà che abbiamo appena

enunciato procede alla somma dei dati per poi dividerli per 4 cioè $(26+28+30+26)/4$. Si capisce che, sebbene abbiamo solo pochissimi dati, senza l'aiuto di una calcolatrice il calcolo della media risulta difficoltoso. Procediamo, in alternativa applicando la proprietà 3 attraverso la seguente trasformazione di comodo: $y_i = 30 - x_i (a = 30, b = -1)$.

Logicamente la trasformazione indica quanto è mancato a ciascun voto per essere il massimo; se vogliamo quanto si è perso in ciascun esame se l'obiettivo fosse stato quello di ottenere a ciascun esame il massimo. In questo modo possiamo dire che al primo esame abbiamo perso voti 4=30-26; al secondo 2=30-28; al terzo 0=30-30 al quarto 4=30-26. Calcolando la media dei voti persi si ha $(4+2+0+4)/4 = 10/4 = 2,5$ il che significa che nei quattro esami abbiamo perso in media 2,5 punti cioè $\mu_y = 2,5$. Ricorrendo alla formula inversa si ha immediatamente che

$$\mu_x = \frac{\mu_y - a}{b} = \frac{2,5-30}{-1} = 27,5$$

Il ricorso alla proprietà ha semplificato notevolmente il calcolo della media aritmetica dei voti d'esame dello studente.

In letteratura si trovano altre proprietà per la media aritmetica. In questa trattazione esse sono, per ragioni di sinteticità, rinviate a testi il cui argomento viene trattato con maggiore rigore metodologico. I valori medi che abbiamo introdotto sono stati ottenuti per una serie di dati $x_1, x_2, \ldots, x_N$, ma cosa succede se disponiamo di una distribuzione di frequenza come quella riportata nelle tabelle 8.1 e 8.2?

Ricordiamo che la frequenza assoluta indica il numero delle volte che nel collettivo si ripete la stessa modalità. Per come è stata introdotta la media aritmetica, quindi, il suo calcolo prevede che ogni modalità prima di essere sommata debba essere moltiplicata per il numero di volte che si ripete nel collettivo, ossia per la corrispondente frequenza. Vale a dire:

$$\mu = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{N} = \frac{1}{N} \sum_{i=1}^{k} x_i n_i \qquad (8.2.2.6)$$

Ricordando che $f_i = \frac{n_i}{N}$, il calcolo della media si può semplificare

come segue

$$\mu = \frac{1}{N} \sum_{i=1}^{k} x_i n_i = \sum_{i=1}^{k} x_i \frac{n_i}{N} = \sum_{i=1}^{k} x_i f_i \qquad (8.2.2.7)$$

Nel caso di una distribuzione di un carattere quantitativo diviso in classi, il calcolo della media passa prima per il calcolo del valore centrale di ciascuna classe e ricordando lo schema fatto in precedenza per questo genere di distribuzioni si ha:

$$\mu = \sum_{i=1}^{k} c_i f_i \qquad (8.2.2.8)$$

dove $c_i = \frac{x_{i-1}+x_i}{2}$ per $i = 1, 2, \ldots, k$ indica appunto il valore centrale di ciascuna classe.

Per fare un esempio concreto consideriamo la tabella 8.5. Aiutandosi con la costruzione della tabella e svolgendo tutti i calcoli si otterrà, per la formula 8.2.2.8, che la media aritmetica dei dati in tabella sarà $\mu = 18$ anni.

Classe di età	Valore centrale della classe $c_i f_i$	Frequenza assoluta n_i	Frequenza relativa f_i	$c_i f_i$
$5 \mapsto 10$	7,5	35	0,075	0,566
$10 \mapsto 13$	11,5	84	0,181	2,082
$13 \mapsto 18$	15,5	136	0,293	4,543
$18 \mapsto 25$	21,5	107	0,231	4,958
$25 \mapsto 28$	26,5	96	0,207	5,483
$28 \mapsto 30$	29	6	0,013	0,375
Totale		464	1,000	18

Tabella 8.5: Esempio: Distribuzione del carattere "età".

La media geometrica

La media geometrica è data dal valore che lascia invariato il prodotto delle modalità. Ricordiamo che formalmente essa si presenta

come

$$M_g = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

Dall'espressione appena scritta si intuisce che questa media può essere calcolata se e solo se i valori della modalità sono tutti positivi. In quanto un solo valore negativo renderebbe il radicando negativo e quindi la radice immaginaria. Dal punto di vista operativo, il calcolo della media geometrica potrebbe riservare alcune difficoltà. Infatti dovendo estrarre la radice N-esima del prodotto di N numeri, si capisce come questa operazione diventa improbabile già quando essi sono nell'ordine di una decina. Provate per esempio a calcolare la media geometrica delle età di 25 alunni di una classe di liceo; vi accorgerete che si lavorerà con numeri troppo grandi e forse incalcolabili. Ma allora se vi sono tutte queste controindicazioni come può essere applicata?

La risposta a questa ovvia e condivisa domanda risiede nella seguente proprietà: *il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi dei valori di cui si vuole calcolare la media geometrica.*

Formalmente si ha

$$\log M_g = \frac{1}{N} \sum_{i=1}^{N} \log x_i$$

Quest'ultima espressione che, a prima vista sembrerebbe complicare la vita, di fatto è una notevole semplificazione se si tiene conto che il logaritmo è semplicemente un operazione algebrica che va fatta eseguire ad una calcolatrice.

Facciamo un esempio elementare. Sia data una distribuzione semplice di 5 punteggi di cui si vuole calcolare la media geometrica.

Nella prima colonna sono riportati i dati osservati dei punteggi, il cui prodotto (ultimo elemento della colonna) è pari a $2,36 \cdot 10^{12}$, come vedete, un numero grandissimo, difficile da gestire sebbene si stia trattando di pochissimi numeri. Nella seconda colonna abbiamo calcolato i corrispondenti logaritmi di base 10 (in genere su una calcolatrice questo tipo di calcolo è indicato con il simbolo log). Il

Carattere X	$\log x_i$
165	2,22
186	2,27
143	2,16
654	2,82
823	2,92
2,36217E+12	12,37

Tabella 8.6: Esempio: calcolo della media geometrica.

calcolo del logaritmo della media geometrica consiste nel dividere la somma della seconda colonna per il numero delle modalità, cioè $\log M_G = \frac{12,37}{5} = 2,47$ che, come si può vedere, semplifica non di poco i conti.

Occorre precisare, tuttavia, che il valore 2,47 non è la media geometrica ma il logaritmo della media geometrica, quindi, per ottenere il valore della media geometrica, è necessario fare l'operazione inversa del logaritmo che, in algebra, è chiamata *esponenziale* (per noi un altro tasto della calcolatrice in genere indicato con il simbolo 10^x). Quindi la media geometrica dei dati riportati in tabella è:

$$M_g = 10^{2,47} = 298,31$$

Una seconda proprietà della media geometrica è che la media geometrica dei rapporti è uguale al rapporto delle medie geometriche dei termini al numeratore ed al denominatore.

Formalmente dati i rapporti $\frac{x_1}{y_1}, \frac{x_2}{y_2}, \ldots, \frac{x_N}{y_N}$ la loro media geometrica è data da

$$M_g = \sqrt[N]{\prod_{i=1}^{N} \frac{x_i}{y_i}} = \frac{\sqrt[N]{\prod_{i=1}^{N} x_i}}{\sqrt[N]{\prod_{i=1}^{N} y_i}}$$

Infine, analogamente a come abbiamo fatto per la media aritmetica, estendiamo il calcolo della media geometrica alle distribuzioni di frequenze di un carattere misurato su scala quantitativa con modalità sia discrete che per classi. Ricordandoci che in tali casi la modalità

deve essere ponderata per le rispettive frequenze

$$M_g = \sqrt[k]{\prod_{i=1}^{k} x_i^{n_i}} \qquad \text{per distribuzioni discrete}$$

$$M_g = \sqrt[k]{\prod_{i=1}^{k} c_i^{n_i}} \qquad \text{per distribuzioni divise in classi}$$

e, per quanto detto sopra, il calcolo passa attraverso la trasformata logaritmica, ossia

$$\log M_g = \frac{1}{k}\sum_{i=1}^{k} n_i \log x_i \qquad \text{per distribuzioni discrete}$$

$$\log M_g = \frac{1}{k}\sum_{i=1}^{k} n_i \log c_i \qquad \text{per distribuzioni divise in classi}$$

La media armonica

Come ultimo tipo di media, introduciamo la media armonica. Essa, come abbiamo visto, rappresenta la modalità che lascia invariata la somma dei reciproci delle misure quantitative di un carattere.

Formalmente ricordiamo che

$$M_{ar} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

Per capire l'utilità e i campi di applicazione di questo tipo di media facciamo un esempio. Supponiamo che si voglia conoscere il consumo medio annuo di toner per stampante di un gruppo di dipendenti. La domanda diretta "Quanti toner consumi in media all'anno?" implica una risposta stimata e quindi imprecisa. Contrariamente sapere quanto dura un toner risulta essere sicuramente più semplice da rilevare, potendo per esempio calcolare i giorni trascorsi dall'intervento del tecnico preposto alla sostituzione del toner. Quindi, dati N dipendenti, è facile rilevare per ciascuno di essi la seguente tabella dei dati

Dipendente	1	2	...i...	N	Totale
Durata toner	x_1	x_2	$...x_i...$	x_N	$\sum_{i=1}^{N} x_i$

Tabella 8.7: Esempio: Calcolo della media armonica.

dove la modalità x_i indica la durata media di toner calcolata da ciascun dipendente.

La media aritmetica dei dati, in questo caso, è inadeguata. Dai dati, infatti, è facilmente ottenibile il consumo totale annuo di toner attraverso il seguente calcolo ripetuto per ciascun dipendente

Dipendente	1	2	i	N	Totale
Durata toner	$360/x_1$	$360/x_2$	$360/x_i$	$360/x_N$	$360/\sum_{i=1}^{N} \frac{1}{x_i}$

Tabella 8.8: Applicazione della media armonica

dove $360/x_i$indica il consumo annuo di toner dell'i-esimo dipendente mentre $360\sum_{i=1}^{N} \frac{1}{x_i}$ indica il numero totale di toner consumati dalla ditta. Se si volesse saper il consumo medio pro-capite basterebbe dividere $360\sum_{i=1}^{N} \frac{1}{x_i}$ per il totale dei dipendenti N, ossia $\frac{360\sum_{i=1}^{N} \frac{1}{x_i}}{N}$ mentre la durata media di ogni toner si ottiene dividendo 360 per il consumo pro-capite; ossia:

$$\frac{360}{\frac{360\sum_{i=1}^{N} \frac{1}{x_i}}{N}} = 360 \cdot \frac{N}{360\sum_{i=1}^{N} \frac{1}{x_i}} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} = M_{ar}$$

che risulta essere proprio la media armonica dei dati.

Facciamo un semplice esempio:

Dipendente	1	2	3	4	*Totale*
Durata toner	12	35	43	71	
Consumo medio giornaliero	0,08	0,03	0,02	0,01	0,15

Tabella 8.9: Esempio: la media armonica.

da cui si evince che la durata media di ogni toner è di 26,80 giorni, in quanto $M_{ar} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} = \frac{4}{0,15} = 26,80$. Mentre il consumo totale di toner sarà dato da: $360 \times 0,15 \cdot 4 = 53,73$, da cui si può dedurre che il consumo medio pro-capite è di 13,43 toner, poiché $53,73/4 = 13,43$. Se si volesse calcolare la media armonica nel caso di una distribuzione di frequenze allora la formula della media armonica diventa

$$M_{ar} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}} \text{ per distribuzioni discrete}$$

$$M_{ar} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{c_i}} \text{ per distribuzioni divise in classi}$$

8.3 Variabilità

In un'accezione generale si definisce variabilità di un carattere misurato su scala quantitativa, l'attitudine del carattere ad assumere le diverse modalità. Il termine attitudine vuole indicare proprio la capacità di un collettivo di assumere diverse misure quantitative di un carattere, quindi, in linea di principio, diremo che in una distribuzione c'è molta variabilità se le unità assumono modalità tra loro diverse; al contrario diremo che in una distribuzione non c'è, o c'è scarsa, variabilità se le unità assumono modalità con misura simile.

Come più volte rimarcato, è importante, in questo contesto, sottolineare il fatto che la misura di variabilità è indicativa della dispersione rispetto al valore di sintesi, nel senso che una maggiore variabilità riduce il valore rappresentativo della sintesi mentre una bassa variabilità la rafforza. Ad esempio due alunni in tre prove d'esame prendono rispettivamente i voti 6, 6, 6 e 5, 6, 7, per entrambi il valore di sintesi è $\mu = 6$. Ma è palese che il valore della media rappresenta meglio il primo alunno rispetto al secondo.

In virtù di quest'ultimo principio, prima di introdurre una misura di variabilità, distinguiamo due situazioni distinte.

- Le modalità del carattere assumono valori diversi per effetto di misure affette da errori accidentali. Per esempio la misura di una montagna eseguita con strumenti che ne stimano l'altezza. In

questo caso infatti l'altezza della montagna è un dato incognito ma reale e le N misure ripetute sono affette da errori accidentali in alcuni casi con valori in eccesso ed in altri con valori in difetto. La media in questo caso assume il significato di valore N volte più preciso del valore reale dell'altezza della montagna.

- Le modalità del carattere assumono valori distinti e ciascuno di essi assume un valore proprio. Il valore di sintesi non è attribuibile a nessuna delle unità del collettivo ed il suo significato interpretativo è semplicemente indicativo del comportamento in media della distribuzione. Nell'esempio dei voti di una scolaresca, la media esprime il comportamento generale degli alunni della classe, ma con molta probabilità non rappresenta nessuno di essi.

Nel primo caso il problema della variabilità è di valutare di quanto in media le quantità rilevate differiscono dalla grandezza effettiva del carattere. Nel secondo caso la variabilità ha lo scopo di misurare quanto mediamente le varie grandezze osservate differiscono tra loro.

Per risolvere il primo dei due casi introduciamo la misura dello *scostamento*

$$s_i = x_i - m \text{ per } i = 1, 2, \ldots, N$$

dove con m indichiamo genericamente il valore di sintesi della distribuzione.

Come misura di variabilità allora si potrebbe adottare una media della suddetta misura.

In forma molto generale, introduciamo le seguenti misure di variabilità: *Lo scostamento medio di ordine r dalla sintesi* m espresso da

$$^{r}S_m = \sqrt[r]{\frac{1}{N} \sum_{i=1}^{N} |x_i - m|^r} \qquad (8.3.0.9)$$

che esprime appunto la media di ordine r degli scostamenti.

In termini particolari se m rappresenta la media aritmetica μ e $r = 1$ otteniamo l'*indice scostamento semplice medio dalla media aritmetica*

espresso da

$$S_\mu = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu| \qquad (8.3.0.10)$$

Mentre se m rappresenta la mediana Me e $r = 1$ otteniamo l'*indice scostamento semplice medio dalla mediana* espresso da

$$S_{Me} = \frac{1}{N}\sum_{i=1}^{N}|x_i - Me| \qquad (8.3.0.11)$$

Se infine m rappresenta la media aritmetica μ e $r = 2$ otteniamo l'*indice scarto quadratico medio dalla media aritmetica* espresso da

$${}^2S_\mu = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} = \sigma \qquad (8.3.0.12)$$

Tra tutti, quest'ultimo, è sicuramente l'indice di variabilità più usato. La ragione sta nel fatto che la quantità $(x_i - \mu)^2$ può essere vista come distanza euclidea della modalità x_i dalla media aritmetica μ. Ricordando che la distanza euclidea è il percorso minimo da fare per raggiungere un punto nel piano; allora possiamo attribuire all'indice σ il significato di media delle distanze minime intercorrenti tra tutte le modalità rispetto alla media aritmetica. Lo scarto quadratico medio è la misura di variabilità ritenuta essere più sensibile rispetto alle altre sopraindicate quando le misure sono affette da errori accidentali. Inoltre si può dimostrare che, in generale, ${}^2S_\mu < S_\mu$.

Lo scostamento medio di ordine r *dalla sintesi* m ${}^rS_m$ gode delle seguenti proprietà.

Proprietà 1

Se la sintesi m è invariante per cambiamenti di unità di misura, cioè se per i dati, a cui viene applicata la trasformazione $y_i = ax_i$ per $i = 1,2,\ldots,N$ vale la relazione $\mu_y = a\mu_x$ per $i = 1,2,\ldots,N$ allora anche la trasformazione per ${}^rS_m$ è invariante per la stessa traslazione, cioè vale ${}^rS_{\mu_y} = a \cdot {}^rS_{\mu_x}$.

Proprietà 2

Se la sintesi m è invariante per traslazioni, cioè se per i dati, dopo la trasformazione $y_i = b + x_i$ per $i = 1,2,\ldots,N$ vale la relazione $\mu_y = b + \mu_x$ per $i = 1,2,\ldots,N$, allora ${}^rS_{\mu_y} = {}^rS_{\mu_x}$.

Proprietà 3

Data la proprietà della mediana cioè $\sum_{i=1}^{N}|x_i - Me| = min$, ne consegue che s_{me} è minimo, cioè che tra gli scostamenti semplici medi il minore è quello dalla mediana.

Proprietà 4

Data la proprietà della media aritmetica $\sum_{i=1}^{N}(x_i - \mu)^2 = min$ ne consegue che ${}^2S_\mu = \sigma$ è minimo, cioè che tra gli scostamenti quadratici medi il minore è quello dalla media aritmetica.

Nelle applicazioni statistiche frequentemente si calcola la variabilità attraverso il quadrato dello scarto quadratico medio ossia

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad (8.3.0.13)$$

Questo indice è noto con il nome di *varianza*. Dal punto di vista interpretativo è identico allo scarto quadratico medio con la variante che l'ordine di misura è espressa al quadrato.

Nel secondo caso, cioè quando le modalità del carattere assumono valori distinti e ciascuno di essi assume un valore proprio, introduciamo la misura della *differenza tra l'unità* i*-esima e l'unità* j*-esima*

$$d_{ij} = x_i - x_j \text{ per } i,j = 1,2,\ldots,N$$

Come misura di variabilità allora si potrebbe adottare una media delle N^2 differenze che si possono stabilire fra tutte le modalità di un carattere, che abbiamo indicato genericamente con $d_{ij} = x_i - x_j$. In forma molto generale introduciamo le seguenti misure di variabilità chiamata *Differenza media di ordine r con ripetizione*, la cui espressione è:

$${}^r\Delta_R = \sqrt[r]{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|^r}{N^2}} \qquad (8.3.0.14)$$

Il termine con ripetizione sta ad indicare che vengono considerate tutte le coppie di differenze possibili; ossia anche quelle composte dalla differenza di ogni modalità con se stessa.

A titolo esemplificativo supponiamo di aver osservato su tre unità A, B e C i seguenti valori: A=3; B=5; C=7. Allora tutte le possibili differenze possono essere disposte in una tabella del tipo: dove in ogni

	A	B	C
A	3-3	3-5	3-7
B	5-3	5-5	5-7
C	7-3	7-5	7-7

Tabella 8.10: Esempio: differenza media di ordine r con ripetizione.

casella è ripetuta la differenza tra la modalità osservata e le due unità riferite alla riga e alla colonna di incrocio. » immediato constatare che sulla diagonale sono riportate le differenze di modalità associate alla stessa unità. È altrettanto immediato constatare che il numero totale di possibili differenze è $3 \cdot 3 = 3^2 = 9$.

Nell'indice di variabilità, basato sulla *differenza media*, bisogna innanzitutto calcolare il valore assoluto di ciascuna differenza (cioè cambiare il segno negativo in positivo), poi elevarle alla potenza di ordine r e sommarle; infine il risultato viene diviso per quante sono le differenze. In questa sequenza di operazione ci si rende conto che le differenze con se stesse, essendo comunque uguali a zero, non inficiano l'operazione complessiva. Quindi si potrebbe optare per l'ipotesi di non considerarle affatto nella operazione. In questo caso le differenze si riducono diventando $N(N-1)$ nell'esempio $3 \cdot 2 = 6$. In questo caso si introduce un altro indice di variabilità chiamato *differenza media di ordine r senza ripetizione* la cui espressione è:

$$ {}^{r}\Delta = \sqrt[r]{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|^r}{N(N-1)}} \qquad (8.3.0.15) $$

È immediato verificare che dall'una si passa alla seconda attraverso la seguente semplice operazione:

$$ {}^{r}\Delta = \sqrt[r]{\frac{N}{N-1}}\,{}^{r}\Delta_R $$

Per valori di $r = 1, r = 2$ si ottengono rispettivamente:

- *Differenza semplice con ripetizione*

$$ \Delta_R = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|}{N^2} \qquad (8.3.0.16) $$

- *Differenza semplice senza ripetizione*

$$ \Delta = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|}{N(N-1)} \qquad (8.3.0.17) $$

- *Differenza quadratica con ripetizione*

$$ {}^{2}\Delta_R = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2}{N^2}} \qquad (8.3.0.18) $$

- *Differenza quadratica senza ripetizione*

$$ {}^{2}\Delta = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2}{N(N-1)}} \qquad (8.3.0.19) $$

In generale, per gli indici ${}^{r}\Delta$ e ${}^{r}\Delta_R$ e valgono le stesse proprietà degli indici basati sugli scostamenti medi di ordine r; ossia sono invarianti per cambiamenti di scala di misura e assumono lo stesso valore per distribuzioni ottenute dopo una traslazione delle modalità. Sulla base dell'esempio esposto sopra calcoliamo i differenti indici di variabilità basati sulle differenze. Innanzitutto calcoliamo le differenze in valore assoluto i cui valori servono sia per calcolare Δ e Δ_R

	A	B	C
A	0	2	4
B	2	0	2
C	4	2	0

Tabella 8.11: Esempio: differenza semplice.

La somma dei dati in tabella è $\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j| = 16$ da cui $\Delta_R = \frac{16}{9} = 1,78$ mentre $\Delta = \frac{16}{6} = 2,67$

Allo stesso modo possiamo calcolare ${}^{2}\Delta$ e ${}^{2}\Delta_R$, in questo caso bisogna calcolare le differenze al quadrato ossia:

	A	B	C
A	0	4	16
B	4	0	4
C	16	4	0

Tabella 8.12: Esempio: differenza quadratica.

La somma dei dati in tabella è $\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2 = 48$ da cui ${}^2\Delta_R = \sqrt{\frac{48}{9}} = 2,31$ mentre ${}^2\Delta = \sqrt{\frac{48}{6}} = 2,83$.

Nel corso degli anni si è acceso un grosso dibattito intorno all'opportunità di usare l'uno o l'altro degli indici di variabilità fin qui considerati. Sebbene sarebbe troppo lungo raccogliere tutti i risultati, si può concludere che per quanto gli indici basati sulle differenze sono concettualmente diversi da quelli basati sugli scostamenti, essi, almeno in certe particolari condizioni, conducono alle stesse informazioni. Si può dimostrare infatti che ${}^2\Delta_R = \sqrt{2}\sigma$, il che significa che a meno di una costante di proporzionalità, fissati i dati, i due indici possono essere ottenuti l'uno dall'altro moltiplicando o dividendo per $\sqrt{2}$.

Dimostrazione

Partiamo dall'indice

$${}^2\Delta_R = \sqrt{\frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2}{N^2}}$$

elevandolo al quadrato si ha

$${}^2\Delta_R^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2}{N^2} \Rightarrow N^2\,{}^2\Delta_R^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - x_j)^2$$

aggiungendo e sottraendo la media aritmetica μ si ha

$$N^2\,{}^2\Delta_R^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i - \mu + \mu - x_j)^2$$

da cui, raccogliendo i termini $(x_i - \mu)(\mu - x_j)$ e sviluppando il quadrato del binomio ottenuto si ottiene

$$\sum_{i=1}^{N}\sum_{j=1}^{N}[(x_i - \mu) - (\mu - x_j)]^2 =$$

$$= \sum_{i=1}^{N}\sum_{j=1}^{N}[(x_i - \mu)^2 + (\mu - x_j)^2 - 2(x_i - \mu)(\mu - x_j)]$$

ricordando la proprietà della media aritmetica si ha

$$2\sum_{i=1}^{N}(x_i - \mu)\sum_{j=1}^{N}(\mu - x_j) = 0$$

quindi si può riscrivere sinteticamente

$$N^2\,{}^2\Delta_R^2 = N\sum_{i=1}^{N}(x_i - \mu)^2 + N\sum_{j=1}^{N}(\mu - x_j)^2$$

dividendo tutto per N^2 si ha

$${}^2\Delta_R^2 = \frac{N\sum_{i=1}^{N}(x_i - \mu)^2 + N\sum_{j=1}^{N}(\mu - x_j)^2}{N^2} =$$

$$= \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N} + \frac{\sum_{j=1}^{N}(\mu - x_j)^2}{N}$$

in conclusione si ottiene ${}^2\Delta_R^2 = 2\sigma^2$ da cui è immediato ricavare la dimostrazione.

La dimostrazione appena formulata mette in chiara luce che la differenza concettuale tra le misure di variabilità basate sulle differenze e quelle basate sugli scostamenti è meno profonda di quanto appaia a prima vista. Nel campo della metodologia empirica, quindi, si possono immaginare più indici di variabilità, purché aderiscano alla particolare situazione concreta. Possiamo considerare, dunque, non solo la differenza media con o senza ripetizione, ma anche lo scarto quadratico medio, lo scarto semplice medio e tutte le espressioni sintetiche di variabilità.

Anche per la misura di variabilità gli indici appena proposti sono stati ricavati considerando una serie di dati $x_1, x_2, \ldots, x_N$. Nel caso si disponga di una distribuzione di frequenza come quella riportata nelle tabelle 8.1 e 8.2, allora ogni scarto $x_i = x_i - m$ per $i = 1, 2, \ldots, k$ deve essere ponderato con la frequenza n_i, quindi lo scostamento di ordine r diventa

$$ {}^{r}S_m = \sqrt[r]{\frac{1}{N}\sum_{i=1}^{k} |x_i - m|^r n_i} \qquad (8.3.0.20) $$

da cui possiamo facilmente ricavare i casi particolari per i diversi valori che assumono r e m. Per esempio, per $r = 2$ ed $m = \mu$ lo scarto quadratico medio

$$ \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k} (x_i - \mu)^2 n_i} \qquad (8.3.0.21) $$

Allo stesso modo per le differenze medie di ordine r con ripetizione, ciascuna differenza $d_{ij} = |x_i - x_j|$ va ponderata con n_i e n_j

$$ {}^{r}\Delta_R = \sqrt[r]{\frac{\sum_{i=1}^{k}\sum_{j=1}^{k} |x_i - x_j|^r n_i n_j}{N^2}} \qquad (8.3.0.22) $$

8.3.1 Il Range e il campo di variazione interquartile

Un'altra impostazione per lo studio della variabilità è basata sugli *intervalli di variazione*. Si fa ricorso, in questo caso, alla differenza tra la modalità più grande e quella più piccola

$$ R = x_{(N)} - x_{(1)} $$

Questa misura è chiamata *campo di variazione* o *Range (R)* ed è prevalentemente impiegata nel campo industriale ed in particolare nel controllo statistico della qualità. È usata anche nella teoria dei campioni, in particolare nel calcolo della numerosità campionaria. Va tuttavia precisato che esso presenta molte difficoltà applicative, specie quando in una distribuzione di dati si è in presenza di dati anomali, in questo caso infatti si amplificherebbe notevolmente la misura di variabilità.

Per alleviare il problema molti autori suggeriscono l'uso della differenza interquartilica

$$ DQ = Q_3 - Q_1 $$

Sebbene semplici e facilmente calcolabili sia *R* che *DQ* sono difficilmente utilizzati nella pratica comune. La causa risiede nel fatto che essi non tengono conto di tutte le modalità, quindi fortemente condizionati dal valore delle code della distribuzione. In conclusione possiamo dire che la misura della variabilità, per mezzo dell'intervallo di variazione e della differenza interquartilica, altro non sono che misure grezze di variabilità e cattive sostitute degli indici basati sugli scostamenti medi e sulle differenze medie di cui se ne preferisce l'uso.

8.3.2 La variabilità relativa

Gli indici di variabilità sopra esposti sono tutti espressi nella stessa unità di misura del carattere; quindi essi sono detti assoluti. Per quanto già noto dai precedenti capitoli, tale caratteristica impedisce il confronto tra distribuzioni di caratteri diversi o di stesso carattere in circostanze diverse. Per chiarire il problema facciamo un esempio. Supponiamo di voler calcolare lo scarto quadratico medio dei pesi di un gruppo di puerpere ricoverate al reparto di ginecologia dell'ospedale di Chieti. È immediato verificare che esso è espresso in chilogrammi; infatti la media μ necessaria per il calcolo dell'indice, essendo la somma dei pesi diviso il numero di puerpere, sarà ancora un peso espresso in kg. Per calcolare lo scarto quadratico medio bisogna eseguire la differenza di ciascun peso dalla media, ma differenze tra pesi sono ancora una volta pesi espressi in chilogrammi, infine lo scarto quadratico medio è la media quadratica delle differenze, che è ancora un peso espresso in chilogrammi, quindi lo scarto quadratico medio è una misura espressa in chilogrammi. Se si volesse paragonare il dato ottenuto con quello delle altezze, ci si accorgerebbe immediatamente che il confronto non avrebbe senso, in quanto il kg è una unità di misura diversa da quella del centimetro (cm). Ma anche volendo confrontare le variabilità dei pesi delle mamme con quelle dei neonati, essendo il collettivo diverso, il confronto risulterebbe errato nonostante si stiano utilizzando le stesse unità di misura.

In questi casi si ricorre agli indici di variabilità relativi che, come abbiamo visto, sono misure pure dalla dimensione e comprese tra zero (assenza di variabilità) e uno (massima variabilità).

Ne consideriamo due tipi: uno basato sul confronto tra il minimo e il massimo; l'altro basato sul rapporto dello scarto quadratico medio dalla media aritmetica

$$CV = \frac{\sigma}{\mu} \cdot 100 \qquad (8.3.2.1)$$

Quest'ultimo chiamato *coefficiente di variazione* (*CV*), sebbene è più semplice da calcolarsi, è molto meno indicativo della variabilità relativa; in quanto non possiede un limite inferiore ed un limite superiore. È altrettanto facile verificare che esso perde di significato se la media μ è uguale a zero e comunque tende ad esplodere per valori della media prossima allo zero.

Per queste ragioni è preferibile far riferimento agli indici di variabilità relativi basati sul confronto tra il minimo ed il massimo

$$Vr = \frac{Va - min(Va)}{max(Va) - min(Va)} \qquad (8.3.2.2)$$

dove con Vr indichiamo la variabilità relativa, Va quella assoluta e con $min(Va)$ e $max(Va)$ il minimo ed il massimo della variabilità assoluta, calcolabile sulla distribuzione data. In particolare, e da quanto già detto sopra, il minimo della variabilità si ha quando tutte le modalità sono uguali e quindi, in tal caso, ogni indice sopra definito varrà sempre zero $min(Va) = 0$. In maniera più semplice diciamo che l'indice di variabilità relativa si ottiene dividendo la variabilità assoluta per il suo massimo.

$$\frac{Va}{max(Va)} \qquad (8.3.2.3)$$

A questo punto si tratta di calcolare il massimo dell'indice di variabilità assoluta che può essere ottenuto in due modi: il primo consiste nel fissare a priori, sotto ipotesi particolari, quella che chiameremo la *distribuzione massimante*, successivamente calcolare su di essa il valore dell'indice di variabilità desiderato che risulterà essere massimo. Il secondo consiste nel determinare analiticamente il massimo sotto

particolari vincoli, come ad esempio lo stesso numero di termini e la stessa media della distribuzione data.

Il primo procedimento può portare all'inconveniente di non poter ricavare, per casi concreti, il massimo, cosicché il corrispondente indice relativo non varrà mai uno. Critiche simili si possono fare anche per il secondo procedimento che risulta condizionato dai vincoli posti in partenza.

Da un punto di vista operativo, la cosa più semplice è quella di definire a priori una distribuzione massimante, cioè una distribuzione teorica in cui, dato un carattere misurato su scala quantitativa, risulta esserci massima variabilità. Una tale distribuzione ha la caratteristica di concentrare tutte le unità tra la minore delle modalità $x_{(1)}$ e la maggiore $x_{(k)}$, possiamo esplicare più chiaramente quanto detto nel grafico 8.6.

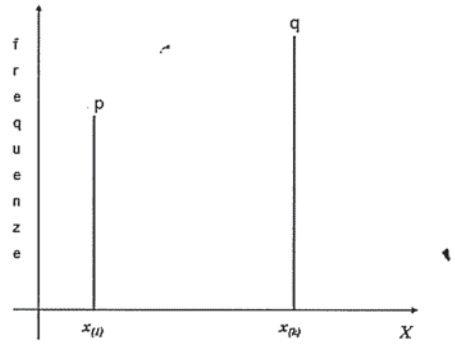


Figura 8.6: Esempio: concentrazione di tutte le unità tra la minore delle modalità $x_{(1)}$ e la maggiore $x_{(k)}$.

Si assume che questa distribuzione sia somigliante a quella osservata, nel senso che si suppone abbia la stessa media e la stessa numerosità. Formalmente le due condizioni sono:

$$\frac{x_{(1)}p + x_{(k)}q}{N} = \mu$$

dove μ è la media calcolata sulla distribuzione dei dati osservati e di

cui si vuole ottenere la variabilità relativa; la seconda invece

$$p + q = N$$

dove N è la numerosità del collettivo ricavata ovviamente dalla stessa distribuzione. Nella distribuzione massimante le quantità p e q sono incognite quindi devono essere calcolate.

Partendo dalle due condizioni appena esplicitate si ricava un sistema di equazioni nelle incognite p e q infatti

$$\begin{cases} \frac{x_{(1)}p + x_{(k)}q}{N} = \mu \\ p + q = N \end{cases}$$

ricavando dalla seconda equazione p e sostituendola alla prima equazione si ha

$$\begin{cases} \frac{x_{(1)}(N-q) + x_{(k)}q}{N} = \mu \\ p = N - q \end{cases}$$

da cui risolvendo la prima equazione nell'incognita p si ha

$$Nx_{(1)} - qx_{(1)} + x_{(k)}q = N\mu$$

segue

$$-qx_{(1)} + x_{(k)}q = N\mu - Nx_{(1)}$$

raccogliendo il fattore comune q e N si ha

$$q(x_{(k)} - x_{(1)}) = N(\mu - x_{(1)})$$

da cui

$$q = \frac{N(\mu - x_{(1)})}{x_{(k)} - x_{(1)}}$$

sostituendo infine q nella seconda equazione del sistema si ricava immediatamente il valore di p

$$p = \frac{N(x_{(k)} - \mu)}{x_{(k)} - x_{(1)}}$$

Nell'ipotesi che si volesse calcolare il massimo dello scarto quadratico medio, si deve calcolare il valore di σ sulla distribuzione massimante che da quanto detto diventa

$$max(\sigma) = \sqrt{(x_{(1)} - \mu)^2 \frac{N(x_{(k)} - \mu)}{x_{(k)} - x_{(1)}} (x_{(k)} - \mu)^2 \frac{(\mu - x_{(1)})^2}{x_{(k)} - x_{(1)}}}$$

infine lo scarto quadratico medio relativo sarà

$$\sigma_{rel} = \frac{\sigma}{\sqrt{(x_{(1)} - \mu)^2 \frac{N(x_{(k)} - \mu)}{x_{(k)} - x_{(1)}} (x_{(k)} - \mu)^2 \frac{(\mu - x_{(1)})^2}{x_{(k)} - x_{(1)}}}} \qquad (8.3.2.4)$$

Da una prima lettura, l'espressione appena scritta sembra essere molto complessa ed articolata. Di seguito riportiamo alcune esemplificazioni che rendono il calcolo della misura di variabilità relativa più agevole. È tuttavia necessario sottolineare che le semplificazioni sono ottenute dopo una serie di elaborazioni algebriche che in questa sede sono state totalmente omesse.

In sintesi diciamo che dopo lunghi passaggi si può ricavare che il massimo dello scarto quadratico medio si semplifica nell'espressione:

$$max(\sigma) = \sqrt{(\mu - x_{(1)})(x_{(k)} - \mu)}$$

da cui il corrispondente indice relativo risulta essere

$$\sigma_{rel} = \frac{\sigma}{\sqrt{(\mu - x_{(1)})(x_{(k)} - \mu)}} \qquad (8.3.2.5)$$

Analogamente si può ricavare il massimo dello scarto semplice medio che ricordiamo essere $S_\mu = \frac{1}{N}\sum_{i=1}^{N} |x_i - \mu|$ la cui espressione semplificata è

$$max(S_\mu) = \frac{2(\mu - x_{(1)})(x_{(k)} - \mu)}{x_{(k)} - x_{(1)}}$$

da cui il corrispondente indice relativo risulta essere

$$S_{\mu_{rel}} = \frac{S_\mu}{\frac{2(\mu - x_{(1)})(x_{(k)} - \mu)}{x_{(k)} - x_{(1)}}} \qquad (8.3.2.6)$$

Passando alle misure di variabilità basate sulle differenze ed in particolare la differenza semplice media senza ripetizione il cui indice assoluto è

$$\Delta = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|}{N(N-1)}$$

L'espressione del massimo dopo una lunga semplificazione si riduce a

$$max(S_\Delta) = 2\mu$$

da cui il corrispondente indice relativo diventa

$$\Delta_{rel} = \frac{\Delta}{2\mu} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}|x_i - x_j|}{2\mu N(N-1)} \qquad (8.3.2.7)$$

8.4 Indici di forma: asimmetria e curtosi

Oltre agli indici di tendenza centrale e di dispersione che abbiamo descritto in precedenza, ci sono altri due valori caratteristici che permettono di farci un'idea di altri aspetti della distribuzione dei dati di una variabile quantitativa su scala ad intervalli. Uno è la *simmetria/asimmetria* e l'altro è la *curtosi*.

Come sottolinea Leti, "*una distribuzione è simmetrica (rispetto alla mediana) se le modalità che sono equidistanti dalla mediana hanno la stessa frequenza*" (da Marrani, 1993, p. 114).

Nel caso delle variabili quantitative su scala ad intervalli, anziché far riferimento alla mediana, si può far riferimento alla media in quanto in una distribuzione perfettamente simmetrica la media e la mediana coincidono. Sulla base di quanto detto è necessario disporre di un indice capace di rilevare la presenza o l'assenza di simmetria in una distribuzione e, in caso negativo, occorre inoltre che lo stesso indice sia in grado di descrivere quanta asimmetria c'è rispetto alla media. Ovvero l'indice di asimmetria dovrà rispondere a due domande:

1. La distribuzione è simmetrica?

2. Nel caso in cui ci fosse asimmetria, quale coda è più lunga dell'altra?

Quest'ultima domanda si pone come obiettivo quello di classificare l'asimmetria in due modi: asimmetria negativa e asimmetria positiva.

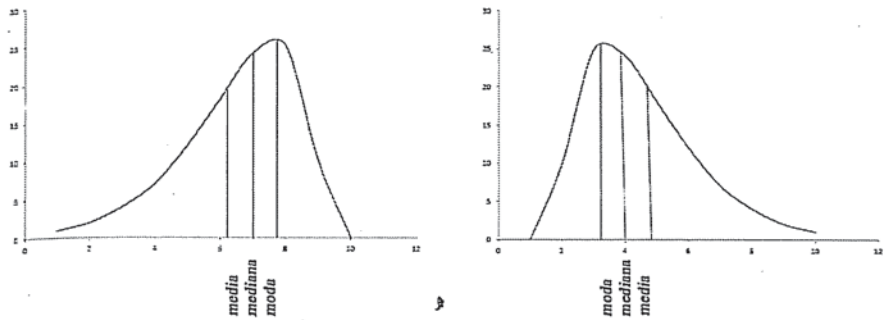


Figura 8.7: Asimmetria negativa e asimmetria positiva.

La prima si ha nel caso in cui risulta essere più lunga la coda che si trova a sinistra della media; mentre il secondo tipo di asimmetria si avrà quando la coda che si trova a destra della media è più lunga.

L'*indice di asimmetria* viene calcolato attraverso la formula di Pearson:

$$ASI = \frac{\sum\left(\frac{x-\mu}{\sigma}\right)^3}{N} \qquad (8.4.0.8)$$

dove μ è la media e σ è lo scarto quadratico medio (o deviazione standard) della distribuzione. Nel caso di una distribuzione di frequenza diventa

$$ASI = \frac{\sum\left(\frac{x-\mu}{\sigma}\right)^3 n_i}{N} \qquad (8.4.0.9)$$

Ragionando su questo indice si può affermare che il suo segno è determinato dal numeratore poiché sia σ che N sono positivi. Quindi, se si è in presenza di asimmetria negativa prevarranno gli scarti di segno negativo, mentre se si avranno più scarti positivi allora si avrà asimmetria positiva. Detto ciò si può concludere che l'indice di asimmetria sarà positivo se la distribuzione è asimmetrica positiva, mentre sarà negativo se la distribuzione è asimmetrica negativa.

Per quanto riguarda la prima domanda, invece, l'indice di asimmetria sarà pari a zero se c'è simmetria. Infatti si è in presenza di simmetria se le due code hanno uguale lunghezza rispetto alla media,

ovvero quando il numero degli scarti negativi è uguale a quello degli scarti positivi.

L'*indice di curtosi* mira, invece, a rilevare quanto una distribuzione è piatta oppure appuntita. Le distribuzioni piatte con code piccole sono chiamate platicurtiche, quelle appuntite con code ampie sono chiamate leptocurtiche. Una distribuzione con la stessa curtosi della distribuzione normale [1] è chiamata mesocurtica.

L'indice di curtosi viene calcolato attraverso la formula di Fisher:

$$CUR = \frac{\sum \left(\frac{x-\mu}{\sigma}\right)^4}{N} - 3 \qquad (8.4.0.10)$$

Come si può notare, la prima parte della formula (la frazione) sarà sempre positiva poiché sia lo scarto dalla media che il σ sono elevati alla quarta, quindi il valore di riferimento di tale frazione sarà il -3. In caso di distribuzione di frequenza, sarà

$$CUR = \frac{\sum \left(\frac{x-\mu}{\sigma}\right)^4 n_i}{N} - 3 \qquad (8.4.0.11)$$

Osservando il disegno 8.9, notiamo che il caso della distribuzione normale, con curva mesocurtica, si avrà solo quando il valore della frazione sarà pari a 3 e l'indice di curtosi assumerà, quindi, valore zero; il caso della distribuzione con curva platicurtica, si avrà quando il valore della frazione sarà minore di 3 e quindi il valore dell'indice di curtosi sarà negativo; mentre il caso della distribuzione con curva leptocurtica, si verificherà quando il valore della frazione sarà superiore a 3, il che implica che l'indice sarà positivo.

8.5 Riepilogando

Facciamo un esempio riepilogativo di alcune delle principali analisi proposte in questo capitolo. Supponiamo, dunque, di aver rilevato dalla segreteria studenti della Facoltà di Scienze della Formazione la distribuzione dei voti conseguiti all'esame di statistica dagli studenti per l'anno accademico 2010/11, i cui dati sono riportati nella tabella e nel grafico che seguono. Ci poniamo l'obiettivo di analizzare i

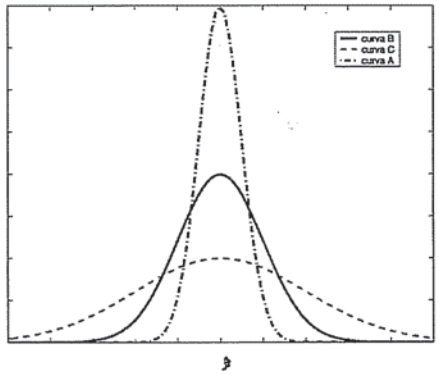


Figura 8.8: Confronto tra curve con un diverso grado di curtosi.

Voti X	Num. studenti n_i
18	17
19	23
20	19
21	35
22	20
23	45
24	56
25	62
26	73
27	40
28	53
29	12
30	20
TOTALE	475

Tabella 8.13: Esempio di distribuzione di frequenza dei voti.

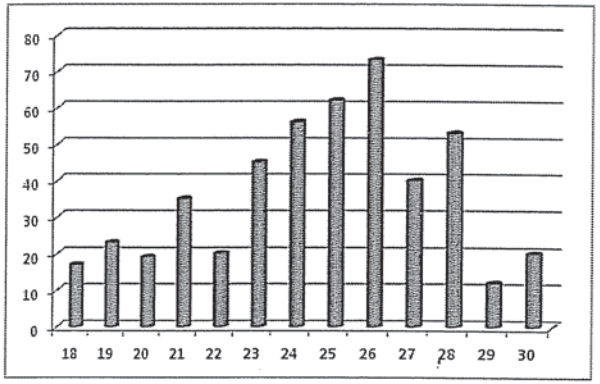


Figura 8.9: Esempio di grafico della distribuzione di frequenza dei voti.

dati con metodo statistico determinando, come al solito, la sintesi, la variabilità e la forma della distribuzione. Procediamo, prima di tutto, alla ricerca dell'indice di sintesi. In particolare, essendo una variabile quantitativa, possiamo calcolare la media aritmetica che riteniamo opportuno essere l'indice di sintesi più valido. Per il calcolo della media utilizziamo i procedimenti indicati nella tabella 8.14 dove, oltre alla colonna delle frequenze relative, $f_i = \frac{n_i}{N}$, nell'ultima colonna abbiamo riportato il prodotto $x_i f_i$ necessario per il calcolo della media aritmetica la cui espressione è

$$\mu = \sum_{i=1}^{k} x_i f_i = 24,50$$

Per valutare la variabilità calcoliamo, invece, lo scarto quadratico medio e la varianza le cui espressioni sono rispettivamente:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 n_i}$$

[1]La curva normale è una distribuzione teorica molto nota in satistica. Questo argomento sarà approfondito nel capitolo "Distribuzioni teoriche".

Voti X	Num. studenti n_i	f_i	$x_i f_i$
18	17	0,04	0,64
19	23	0,05	0,92
20	19	0,04	0,80
21	35	0,07	1,55
22	20	0,04	0,93
23	45	0,09	2,18
24	56	0,12	2,83
25	62	0,13	3,26
26	73	0,15	4,00
27	40	0,08	2,27
28	53	0,11	3,12
29	12	0,03	0,73
30	20	0,04	1,26
TOTALE	475	1	24,5

Tabella 8.14: Procedimento di calcolo della media aritmetica.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \mu)^2 n_i$$

I procedimenti di calcolo sono, invece, riportati nella tabella 8.15

Da quanto esplicitato in tabella, ricaviamo facilmente che $\sigma^2 = \frac{4394,75}{475} = 9,25$ mentre lo scarto quadratico medio $\sigma = \sqrt{9,25} = 3,04$. Volendo trovare una misura relativa per la variabilità si deve innanzitutto procedere al calcolo del massimo valore teorico dello scarto quadratico medio, in particolare ricordiamo che

$$max(\sigma) = \sqrt{(\mu - x_1)(x_k - \mu)}$$

dove con x_1 abbiamo indicato la modalità minore della distribuzione, quindi $x_1 = 18$ e con x_k la modalità maggiore, quindi $x_k = 30$.

Per cui sulla base dei dati disponibili otteniamo,

$$max(\sigma) = \sqrt{(24,5 - 18)(30 - 24,5)} = 5,98.$$

Voti X	Num. studenti n_i	f_i	$(x_i - \mu)^2 n_i$
18	17	0,04	718,02
19	23	0,05	695,48
20	19	0,04	384,57
21	35	0,07	428,49
22	20	0,04	124,89
23	45	0,09	101,11
24	56	0,12	13,94
25	62	0,13	15,57
26	73	0,15	164,48
27	40	0,08	250,21
28	53	0,11	649,64
29	12	0,03	243,11
30	20	0,04	605,23
TOTALE	475	1	4394,75

Tabella 8.15: Procedimento di calcolo della varianza.

Infine, l'indice di variabilità relativa è dato dall'espressione

$$\sigma_{rel} = \frac{\sigma}{max(\sigma)} = \frac{3,04}{5,98} = 0,51$$

che evidenzia una discreta variabilità tra i voti riportati dagli studenti di questo corso di laurea.

Procediamo, infine, alla ricerca degli indici di forma, ossia l'asimmetria e la curtosi. Da quanto detto nella parte teorica, ponderando gli indici con le frequenze, avremo le seguenti due espressioni

$$ASI = \frac{\sum \left(\frac{x-\mu}{\sigma}\right)^3 n_i}{N}$$

$$CUR = \frac{\sum \left(\frac{x-\mu}{\sigma}\right)^4 n_i}{N} - 3$$

I procedimenti di calcolo sono illustrati nella tabella 8.16 dove

Voti X	Num. studenti n_i	$\left(\frac{x-\mu}{\sigma}\right)^3 n_i$	$\left(\frac{x-\mu}{\sigma}\right)^4 n_i$
18	17	-165,81	354,27
19	23	-135,90	245,68
20	19	-61,48	90,93
21	35	-53,27	61,28
22	20	-11,09	9,11
23	45	-5,39	2,65
24	56	-0,25	0,04
25	62	0,28	0,05
26	73	8,77	4,33
27	40	22,24	18,28
28	53	80,82	93,02
29	12	38,88	57,54
30	20	118,31	213,96
TOTALE	475	-163,89	1151,15

Tabella 8.16: Procedimento di calcolo di asimmetria e curtosi.

i valori di μ e di σ, ovviamente, sono quelli già calcolati nelle fasi precedenti.

In conclusione gli indici di asimmetria e di curtosi risultano essere $ASI = -0,35$ e $CUR = -0,5$. Dagli indici calcolati rileviamo una modesta asimmetria negativa ed una certa importanza dei valori della distribuzione alle code, ossia una buona presenza di voti bassi (intorno al 18) e voti alti (intorno al 30).

Capitolo 9

Distribuzioni teoriche

Nei capitoli precedenti abbiamo trattato la valutazione di un fenomeno reale proponendo il metodo statistico. In particolare per ciascuna scala di misura abbiamo introdotto gli indici di sintesi di variabilità e di forma. In questo capitolo estendiamo il metodo statistico introducendo un approccio che possiamo definire globale in quanto ci permetterà di trattare le distribuzioni empiriche, cioè osservate, non più attraverso i soli indici precedentemente introdotti, come le medie, la variabilità, la simmetria ecc., ma sfruttando l'intera distribuzione dei dati.

Possiamo dire in maniera semplice che quanto stiamo per sviluppare rientra nella valutazione della forma della distribuzione intesa, questa volta, come lo studio del comportamento globale di tutte le osservazioni effettuate sul fenomeno reale posto a valutazione. Pertanto, anche in questo caso, il punto di partenza dell'analisi statistica sarà la distribuzione di frequenze assolute o relative introdotte nel capitolo 4. Il metodo che stiamo per approfondire ci consente di valutare l'esistenza di regolarità o meno del fenomeno osservato grazie all'uso di modelli matematici. Lo studio e l'uso di questo metodo nasce dalla necessità di verificare se esiste o meno un'associazione tra una distribuzione osservata, quindi empirica, e la corrispondente distribuzione attesa o teorica. Nell'analisi di questo metodo prenderemo in esame due diverse tipologie di approccio:

- approccio parametrico;

- approccio non-parametrico.

Vedremo in particolare come nell'approccio parametrico la regolarità del fenomeno reale viene approssimata da una funzione matematica, che dipende da uno o più parametri. In questo contesto il modello utilizzato risponde ad un'ipotesi predefinita del comportamento del fenomeno, pertanto la diversità e l'interpretazione dello stesso sarà data dalla valutazione da una prefissata famiglia di funzioni che chiameremo *distribuzioni teoriche* e che si distingueranno, caso per caso, dal valore stimato dei suoi parametri sulla base dei dati osservati. Per chiarire meglio il concetto facciamo un semplice esempio: si supponga di voler valutare il comportamento dei consumi delle famiglie di una specifica regione e che si voglia studiare il comportamento di questo fenomeno reale per le diverse fasce di reddito. È abbastanza intuitivo pensare che il consumo sia misurabile con la spesa sostenuta, in moneta corrente, da ciascuna famiglia per il consumo di beni e sevizi. Analogamente, le fasce di reddito possono essere misurate con la stessa scala di misura, ossia con la quantità di moneta corrente disponibile proveniente dal lavoro e degli investimenti di capitali di ciascuna famiglia. Seguendo la più semplice teoria economica in materia, si può assumere che il comportamento del consumo è una funzione lineare del reddito. Ossia si assumerà che il modello teorico, che sottace il fenomeno reale in oggetto, è esplicabile per mezzo un retta di equazione $c_s = \alpha + \beta y + e$, dove con c_s indichiamo la spesa per consumo delle famiglie, con y il reddito delle famiglie e con e una componente accidentale, non specificata dal modello, che esprime le diversità di spesa per consumo di ciascuna famiglia appartenente alla stessa fascia di reddito e che il modello non è in grado di valutare. Come si può osservare il modello teorico rappresenta una famiglia di funzioni che generano un fascio di rette al variare dei due suoi parametri α e β. Nell'approccio parametrico ci proponiamo, quindi, di stimare i parametri sulla base dei dati empirici osservati. Naturalmente la valutazione del fenomeno reale sarà data dalla valutazione dei parametri stimati del modello.

Nell'approccio non parametrico, invece, la regolarità del fenomeno reale viene stabilita da una approssimazione della distribuzione empirica, utilizzando funzioni generiche o più in generale curve che

non rispondono necessariamente ad una ipotesi teorica del comportamento del fenomeno. Si sceglieranno quelle che sono più opportune all'approssimazione dei dati. In questo caso la valutazione sarà data dalla funzione stimata nella sua interezza e dal suo comportamento grafico; quindi l'approccio non parametrico, diversamente da quello parametrico, non prevederà la stima dei parametri della funzione scelta, ma stabilirà metodi, che vedremo meglio in seguito, che stimeranno complessivamente la curva interpretativa del fenomeno reale sottoposto all'analisi. In altri termini l'obiettivo è quello di cercare tra tutte le funzioni matematiche quella che meglio approssima i dati. Riprendendo l'esempio precedente una volta osservato il fenomeno reale nelle due misure del consumo e del reddito, si cercherà di individuare una funzione che non necessariamente risponde ad una teoria economica ma che meglio di tutte interpreta il comportamento empirico dei dati.

9.1 L'approccio parametrico

Per affrontare l'approccio parametrico, sulla base di quanto abbiamo appena detto, chiariamo meglio il concetto di distribuzione teorica e, in questo contesto, ci riferiamo a quelle utilizzate con maggior frequenza, allo scopo di avere gli strumenti utili per trattare un'ampia classe di fenomeni reali che si possono incontrare nei vari campi della ricerca.

In generale e per quanto premesso, il nostro scopo è quello di esplicitare meglio la forma di una distribuzione. È, infatti, necessario puntualizzare che il concetto di forma è molto più ampio e complesso di quanto abbiamo trattato in precedenza. La forma di una distribuzione, infatti, coinvolge diversi aspetti che cercheremo di sviluppare e chiarire nel dettaglio in questo capitolo.

Abbiamo detto che lo spirito che muove il ricercatore in questo contesto è quello di trovare, qualora ce ne fossero, delle regolarità che sulla base della osservazione empirica tenta di prendere una forma a volte descrivibile con una funzione matematica del tipo $y = f(x)$, ossia una funzione reale di variabili reali che permette di racchiudere tutte le informazioni del fenomeno reale oggetto di studio in una sola fun-

zione matematica. A prima vista questo sembra complicare lo studio, ma di fatto ne agevola notevolmente la comprensione e l'interpretazione potendo ricorrere agli strumenti di analisi matematica. È necessario puntualizzare che i fenomeni reali sono osservati sulle unità statistiche che, per quanto si possono immaginare numerose, sono sempre un numero finito. Mentre le distribuzioni teoriche, in quanto tali, non corrispondono ad un collettivo di unità statistiche ma rappresentano il fenomeno reale nel suo complesso e nella sua essenza intrinseca rappresentabile con un modello matematico. Tuttavia il ricercatore che usa il metodo statistico e che vuole valutare un fenomeno reale in questo contesto, ossia attraverso una modello/distribuzione teorico, non può prescindere dalla fase della osservazione empirica del fenomeno attraverso un collettivo di unità. In quest'ambito il collettivo viene inteso come l'insieme di repliche del fenomeno reale. Quindi posto vero il modello teorico, le repliche vengono assunte avere le stesse regolarità descrivibili dal modello ipotizzato. Infatti, sebbene le osservazioni sono affette da errori di diversa natura, si ritiene che il modello teorico sia la causa generatrice delle osservazioni empiriche. Sulla base di questo principio molto verosimilmente la distribuzione empirica assumerà una forma simile a quella teorica. Come vedremo più avanti, l'obiettivo è quello di ricavare la distribuzione teorica stabilendo le intrinseche caratteristiche del fenomeno osservato ed usare la distribuzione empirica solo per la stima dei suoi parametri. In altri termini nell'approccio parametrico il modello teorico viene stabilito a priori seguendo una predefinita ipotesi del comportamento del fenomeno reale. Una volta validato il modello teorico e stimati i suoi parametri, esso sostituirà in tutto e per tutto lo studio di valutazione del fenomeno reale: ciò significa che le consuete tre fasi del metodo statistico ossia sintesi, variabilità e forma, che in questo caso è implicita nel modello, vengono ricavate direttamente dalla distribuzione teorico stimata.

Da un punto di vista più formale, le distribuzioni teoriche possono essere classificate in diversi tipi. Di seguito sono illustrate le caratteristiche essenziali delle distribuzioni teoriche.

1. La distribuzione teorica è esplicitata da una funzione matematica del tipo $y = f(x)$, in cui il dominio della funzione è espresso

dalle modalità in cui è stato misurato il fenomeno reale tale che sia possibile esprimere una legge che associa ad ogni valore del dominio, ossia ogni modalità di misura, un valore reale in genere espresso da una frequenza assoluta o relativa. Schematicamente:

$$\left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_k \\ f_1 & f_2 & \dots & f_k \end{array} \right\} \qquad (9.1.0.1)$$

dove k è un numero finito di modalità diverse di misurazione del fenomeno reale; schematicamente la possiamo indicare $\{x_i, f_i\}$ con $i = 1, 2, \dots, k$.

2. La distribuzione teorica è esplicitata da una funzione matematica del tipo $y = f(x)$, in cui il dominio della funzione è espresso da una successione infinita numerabile di modalità di misura a cui fa corrispondere un'altra successione di valori reali ossia

$$\left\{ \begin{array}{cccc} x_1 & x_2 & \dots & x_i \dots \\ f_1 & f_2 & \dots & f_i \dots \end{array} \right\} \qquad (9.1.0.2)$$

schematicamente possiamo indicare

$$\{x_i, f_i\} \text{ con } i = 1, 2, \dots, i, \dots$$

3. La distribuzione teorica è esplicitata da una funzione matematica del tipo $y = f(x)$, in cui il dominio della funzione è espresso da un intervallo limitato o illimitato dell'asse reale a cui fa corrispondere una curva che nell'accezione del linguaggio matematico può essere considerata continua, ma anche con punti di discontinuità, schematicamente la possiamo indicare $\{x, f(x)\}$.

Nel primo e secondo caso c'è poco da aggiungere dal punto di vista del calcolo della sintesi e della variabilità e della forma, in quanto non si riscontra niente di diverso rispetto alle distribuzioni empiriche viste nei capitoli precedenti. Mentre nell'ultimo caso le valutazioni delle distribuzioni teoriche vengono fatte attraverso lo studio dei momenti.

9.1.1 Momenti di una distribuzione teorica

Abbiamo detto che la distribuzione teorica sostituisce in tutto e per tutto la distribuzione empirica, in altri termini possiamo dire che essa è un surrogato della distribuzione dei dati osservati. Quindi, similmente a quanto fatto in precedenza, si rende necessario acquisire strumenti per lo studio delle distribuzioni ricavate. L'approccio seguito è, al solito, quello del metodo statistico cioè del calcolo della sintesi della variabilità e della forma.

Lo strumento metodologico è il calcolo dei *momenti* che, come vedremo più avanti, rappresentano univocamente la distribuzione teorica. In particolare, i momenti ci permettono di definire delle grandezze caratteristiche della distribuzione teorica ed hanno per questo la capacità di riassumere in modo immediato e sintetico l'informazione relativa alla distribuzione oggetto di studio. In termini generali, si distinguono i *momenti dall'origine*, dai *momenti centrati* rispetto ad un'*origine arbitraria*.

Se la distribuzione è discreta, ossia è del primo tipo e per estensione al secondo tipo, si definisce momento di ordine r dall'origine la seguente espressione:

$$\mu_r = \sum_{i=1}^{k} x_i^r f_i \qquad (9.1.1.1)$$

In particolare, per $r = 1$, ricaviamo la ormai nota media aritmetica indicata semplicemente con la lettera greca μ. Per $r = 2$, avremo

$$\mu_2 = \sum_{i=1}^{k} x_i^2 f_i \qquad (9.1.1.2)$$

ossia la media quadratica al quadrato.

Si definisce, invece, momento centrato di ordine r da un'origine arbitraria a l'espressione:

$$\bar{\mu}_r(a) = \sum_{i=1}^{k} (x_i - a)^r f_i \qquad (9.1.1.3)$$

In genere come origine arbitraria viene assunto il momento primo dall'origine, ossia la media μ o anche la mediana o la moda. Quindi, a

seconda del valore che daremo all'origine arbitraria a, avremo diverse misure caratteristiche della distribuzione:

- per $r = 1$, $a = \mu$, si ha $\bar{\mu} = \sum_{i=1}^{k}(x_i - \mu)f_i$, ovvero la somma degli scarti dalla media che, come noto, è pari a zero (cfr. prima proprietà della media aritmetica);
- per $r = 2$, $a = \mu$, si ha $\bar{\mu}_2 = \sum_{i=1}^{k}(x_i - \mu)^2 f_i = \sigma^2$, che come noto rappresenta la misura della *varianza*;
- per r=3, $a = \mu$, si ha $\bar{\mu}_3 = \sum_{i=1}^{k}(x_i - \mu)^3 f_i$;
- per r=4, $a = \mu$, si ha $\bar{\mu}_4 = \sum_{i=1}^{k}(x_i - \mu)^4 f_i$.

Da quanto abbiamo appreso in precedenza, standardizzando la misura quantitativa x attraverso la trasformazione lineare $z = \frac{x-\mu}{\sigma}$ e calcolando il momento terzo e il momento quarto della nuova variabile standardizzata, si ricavano rispettivamente gli indici di *asimmetria*

$$\delta_1 = \sum_{i=1}^{k} \left(\frac{x - \mu}{\sigma}\right) f_i = \frac{\mu^3}{\sigma^3} \qquad (9.1.1.4)$$

e di *curtosi*

$$\delta_1 = \sum_{i=1}^{k} \left(\frac{x - \mu}{\sigma}\right) f_i = \frac{\mu^4}{\sigma^4} \qquad (9.1.1.5)$$

le cui interpretazioni sono le stesse già esplicitate nei capitoli precedenti.

Nel caso di distribuzioni teoriche continue, cioè nel terzo tipo, il *momento di ordine* r *dall'origine* è definito come:

$$\mu_r = \int_{\forall x} x^r f(x)\,dx \qquad (9.1.1.6)$$

il *momento centrato di ordine* r *dall'origine* a, invece, è definito come:

$$\mu_r = \int_{\forall x} (x - a)^r f(x)\,dx \qquad (9.1.1.7)$$

Da quanto sinora detto, si intuisce l'importanza del calcolo dei momenti di una distribuzione teorica, quali parametri rappresentativi

ed interpretativi di una distribuzione. Per questo si rende necessario acquisire un'adeguata abilità nell'operare con i momenti. In particolare, è facile ricavare una relazione che lega i momenti da un'origine arbitraria dai momenti dall'origine ricordando l'espansione in serie del binomio di potenza r da cui si ottiene la seguente relazione:

$$\bar{\mu}_r = \sum^{r} (-1) \binom{r}{b} \mu^b \mu_{r-b} \qquad (9.1.1.8)$$

dalla quale è immediato ricavare l'importante relazione: $\sigma^2 = \mu_2 - \mu^2$. Che definisce la varianza come differenza del momento secondo dall'origine dalla media aritmetica al quadrato.

Seguendo lo stesso principio si possono facilmente estendere le altre misure di sintesi, variabilità e forma già introdotte per le distribuzioni empiriche che per semplicità le riassumiamo nella tabella 9.1.

Indici	*Distribuzioni teoriche del primo e secondo tipo*	*Distribuzioni teoriche del terzo tipo*
Media armonica	$M_{ar} = \frac{1}{\sum_{i=1}^{k} \frac{1}{x_i} f_i}$	$M_{ar} = \frac{1}{\int_{\forall x} \frac{f(x)dx}{x}}$
Media quadratica	$M_q = \sqrt{\sum_{i=1}^{k} x_i^2 f_i}$	$M_q = \sqrt{\int_{\forall x} x^2 f(x)dx}$
Media di potenze di ordine r	$M_r = \sqrt[r]{\sum_{i=1}^{k} x_i^r f_i}$	$M_r = \sqrt[r]{\int_{\forall x} x^r f(x)dx}$
Scarto semplice medio dalla media aritmetica	$s_\mu = \sum_{i=1}^{k} \lvert x_i - \mu \rvert f_i$	$s_\mu = \int_{\forall x} \lvert x_i - \mu \rvert f(x)dx$
Scostamento semplice medio dalla mediana	$S_{Me} = \sum_{i=1}^{k} \lvert x_i - Me \rvert f_i$	$S_{Me} = \int_{\forall x} \lvert x_i - Me \rvert f(x)dx$
Scostamento quadratico medio	$\sigma = \sqrt{\sum_{i=1}^{k} (x_i - \mu)^2 f_i}$	$\sigma = \sqrt{\int_{\forall x} (x_i - \mu)^2 f(x)dx}$
Differenza semplice media con ripetizione	$\Delta_R = \sum_{i=1}^{k} \sum_{j=1}^{k} \lvert x_i - x_j \rvert f_i f_j$	$\Delta_R = \int_{\forall x} \int_{\forall y} \lvert x - y \rvert f(x) f(y) dx dy$

Tabella 9.1: Misure di sintesi, variabilità e forma

9.1.2 Distribuzioni teoriche di uso più frequente

Modello di Bernoulli

Per ragioni di opportunità, cominciamo con il trattare le distribuzioni teoriche che appartengono al primo gruppo. Tra i più elementari modelli appartenenti a questa classe c'è il modello di Bernoulli. Immaginiamo di disporre di un caso sperimentale i cui possibili risultati sono dati da una modalità di risposta che chiamiamo successo ed un'altra che chiameremo insuccesso. In altri termini, diciamo che lo spazio delle possibili misure del carattere è banalmente ripartibile nelle due sole modalità A =successo e B =insuccesso. Assumendo che θ sia la frequenza relativa del successo, ovvero del verificarsi della modalità A, è immediato ricavare che $1-\theta$ è la frequenza relativa dell'insuccesso, ossia del verificarsi della modalità B. Una variabile quantitativa che rappresenti questo schema può essere banalmente indicata:

$$x_i = \begin{cases} x_1 = 1 & \textit{se si verifica l'evento A} \\ x_2 = 0 & \textit{se si verifica l'evento B} \end{cases} \qquad (9.1.2.1)$$

La distribuzione teorica associata a questa variabile è:

$$f_i = \theta^{x_i}(1-\theta)^{1-x_i} \qquad (9.1.2.2)$$

il cui unico parametro d'interesse è θ. In forma compatta, il modello bernoulliano lo indicheremo $x \approx B(\theta)$. È immediato ricavare il momento di ordine r, che risulta essere pari a:

$$\mu_r = 1^r\theta + 0^r(1-\theta) = \theta \qquad (9.1.2.3)$$

da cui si ha immediatamente che la media risulta essere pari a

$$\mu = \theta$$

mentre la varianza sarà

$$\sigma^2 = \mu_2 - \mu^2 = \theta - \theta^2 = \theta(1-\theta)$$

è facile intuire che, per trattare il modello di Bernoulli, è necessario trovare una stima adeguata del parametro θ, il cui valore dovrà essere opportunamente stimato sulla base dei dati ricavati dal sondaggio.

9.1.3 Distribuzione binomiale

Tra le distribuzioni teoriche che appartengono al primo gruppo della nostra classificazione e che maggiormente si incontrano nella valutazione di fenomeni reali c'è sicuramente la distribuzione binomiale o detta delle prove ripetute. Essa fa riferimento a uno schema teorico dove si deve immaginare il fenomeno reale diviso in due gruppi, il primo misurabile con la modalità A, il cui numero di unità è pari a n_A, mentre l'altro misurabile con la modalità B, il cui numero di unità è pari a n_B. Naturalmente la somma $n_A + n_B = N$ rappresenta il complesso del fenomeno reale che, come abbiamo detto, è composto da un numero finito di misure. In questo schema teorico è quindi possibile calcolare le frequenze relative $\theta = \frac{n_A}{N}$ e $(1-\theta) = \frac{n_B}{N}$. è immediato verificare che $\theta + (1-\theta) = \frac{n_A}{N} + \frac{n_B}{N} = 1$.

Supponiamo che il collettivo sia osservato n volte e ogni volta l'osservazione sia indipendente dalle altre. Inoltre, supponiamo che la composizione del collettivo non cambi da un'osservazione all'altra, ossia, la numerosità del collettivo osservato rimane N e le frequenze θ e $(1-\theta)$. Il disegno sperimentale può essere semplificato supponendo di estrarre n volte un'unità a caso dal collettivo di riferimento ed avere l'accortezza di reinserire l'unità estratta nella lista originale.

A titolo di esempio supponiamo che il collettivo N di riferimento siano gli studenti di una scuola e supponiamo che $\theta = \frac{n_A}{N}$ di essi siano stranieri mentre $(1-\theta) = \frac{n_B}{N}$ hanno la nazionalità italiana. Supponiamo di essere interessati alla misura x del numero di studenti di nazionalità straniera che si possono ottenere in un gruppo composto da n studenti di quella scuola. È immediato intuire che detta misura può assumere le modalità $x_i = 0, 1, 2, \ldots, n$; 0 nel caso non ci siano studenti di nazionalità straniera; 1 se c'è un solo studente straniero; 2 se ce ne sono due, ecc. Nella tabella 9.2 siamo interessati a conoscere le relative frequenze associate a ciascuna modalità $x_i = 0, 1, 2, \ldots, n$.

La risposta a questo disegno sperimentale è data dalla distribuzione teorica chiamata *binomiale* o anche delle prove ripetute ed indipendenti, la cui espressione formale è:

$$f_i = \binom{n}{x_i}\theta^{x_i}(1-\theta)^{n-x_i} \qquad (9.1.3.1)$$

Modalità della misura numero di stranieri in n estrazioni ripetute ed indipendenti x	Frequenze relative f_i
0	f_0
1	f_1
2	f_2
⋮	
i	f_i
⋮	
n	f_n

Tabella 9.2: Es. stima frequenze relative attraverso la v.c. Binomiale

Fissati, infatti, n e θ, facendo variare x_i da 0 a n si ottengono tutte le frequenze teoriche schematicamente rappresentate della tabella 9.2. È evidente che l'espressione 9.1.3.1 può essere univocamente determinata solo se si conoscono il valore delle estrazioni n e della frequenza relativa degli studenti stranieri θ.

Questo problema al momento esula dalla nostra trattazione in quanto appartiene ad un altro ramo del metodo statistico noto come *inferenza statistica* che, sulla base dell'estrazione di un campionamento casuale ricavato dal collettivo sottoposto allo studio, permette di ottenere una stima nelle quantità richieste.

Nel nostro caso diciamo che n e θ sono i parametri della distribuzione teorica ed i loro valori determinano la forma della stessa. Per questa ragione è utile indicare sinteticamente la distribuzione teorica binomiale con l'espressione $x = B(n, \theta)$ dove x, come abbiamo detto, indica la misura del fenomeno reale "numero di eventi favorevoli in n estrazioni casuali ed indipendenti" (nell'esempio numero di stranieri estratti).

Con n e θ sono indicati rispettivamente il numero delle estrazioni e la frequenza degli eventi favorevoli nel collettivo (nell'esempio quanti studenti sul totale sono stranieri nel collettivo di riferimento).

È facile dimostrare che al variare di x_i la distribuzione teorica ricavata è una distribuzione di frequenza così come è stata intesa nel

capitolo in cui è stata introdotta. Infatti ciascuna f_i è un numero sicuramente maggiore di 0 in quanto si ottiene dal prodotto di numeri positivi: $\binom{n}{x_i} > 0$ (esso è, infatti, il coefficiente binomiale ottenuto dalle combinazione di unità prese x_i a x_i[1] che è sicuramente maggiore di 0). Allo stesso modo si può dimostrare che la somma delle frequenze f_i è uguale a 1; infatti, calcolando la somma $\sum_{i=1}^{k} \binom{n}{x_i}\theta^{x_i}(1-\theta)^{n-x_i}$ essa corrisponde allo sviluppo in serie del binomio $(\theta + (1-\theta))^n = 1$. Di conseguenza si ha che $0 \leqslant f_i \leqslant 1$. Ciò dimostra che l'espressione 9.1.3.1 genera una distribuzione di frequenze.

Come di consueto, a questo punto della trattazione, devono essere introdotti gli indici di sintesi, di variabilità e di forma allo scopo di studiare il fenomeno reale rimpiazzato dal modello teorico appena introdotto. Pertanto applicando le espressioni introdotte nella Tab. 9.1 si dimostra che la media aritmetica è:

$$\mu = \sum_{i=1}^{k} x_i \binom{n}{x_i} \theta^{x_i}(1-\theta)^{n-x_i} = n\theta$$

mentre la varianza è:

$$\sigma^2 = \sum_{i=1}^{k} (x_i - n\theta)^2 \binom{n}{x_i} \theta^{x_i}(1-\theta)^{n-x_i} = n\theta(1-\theta).$$

Di conseguenza lo scarto quadratico medio è $\sigma = \sqrt{n\theta(1-\theta)}$.
Infine l'asimmetria e la curtosi possono essere calcolate con indici:

$$ASI(X) = \frac{1-2\theta}{\sqrt{n\theta(1-\theta)}}$$

$$CUR(X) = 3 + \frac{1-6\theta+6\theta^2}{n\theta(1-\theta)}$$

Il grafico della distribuzione teorica varia al variare di θ: è simmetrico per $\theta = 0.5$, asimmetrico negativo per $\theta < 0.5$, asimmetrico positivo per $\theta > 0.5$. Di seguito riportiamo una rappresentazione grafica del comportamento del modello teorico per tre diversi valori di θ e n (Fig. 9.1).

[1] Per maggiore chiarezza si consulti un testo di calcolo combinatorio.

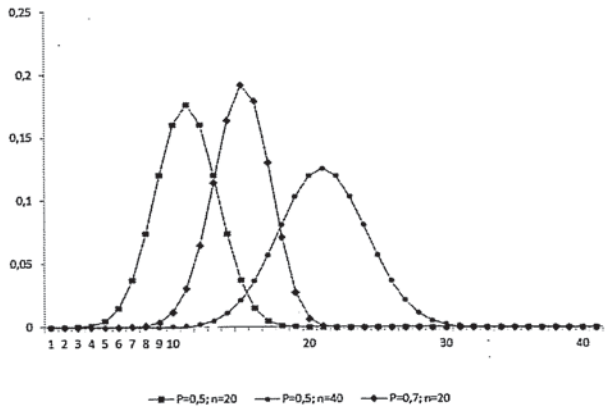


Figura 9.1: Rappresentazione modello binomiale per tre diverse coppie di valori di θ e n

Quale esempio di applicazione del modello binomiale consideriamo che il fenomeno reale sul quale si vuole prendere una decisione sia l'elezione a presidente di una regione, supponiamo che ci siano solo due candidati: il candidato A, la cui probabilità incognita di successo sia θ e il candidato B, la cui probabilità di successo sia $1-\theta$. Supponiamo, inoltre, che venga commissionato un sondaggio di opinione a n elettori scelti con criterio di estrazione bernoulliana (ossia con ripetizione) dalla popolazione.

Il disegno sperimentale è assimilabile al modello teorico binomiale appena introdotto, dove $x_i = (0,1,2\ldots,n)$ esprime il numero dei possibili voti riportati dal candidato A nel sondaggio. Naturalmente, è immediato intuire che l'analisi del fenomeno è strettamente legata alla conoscenza dei parametri θ e $1-\theta$. Una valutazione completa dell'atteggiamento degli elettori verso i due candidati si può ottenere solo attraverso l'analisi del modello proposto, ossia attraverso la ricerca di un valore che misuri la sintesi, la variabilità, nonché la forma. La soluzione a tale problema è data dal calcolo della media $n\theta$ e della varianza $n\theta(1-\theta)$. In questo contesto il problema si riduce alla conoscenza dei parametri n e θ che dovranno essere opportunamente stimati sulla base dei dati ricavati dal sondaggio.

9.1.4 Distribuzione geometrica

Partendo dal modello bernoulliano introdotto precedentemente, ci si può chiedere quante prove bisogna ripetere per avere il primo successo. Per esempio, nel caso dell'elezione a presidente della regione di cui al paragrafo precedente, si vuole sapere con quale frequenza si avrà la prima scheda votata con il nominativo del candidato A nella procedura di spoglio. Indicando, come al solito, con θ la probabilità del successo e con $1-\theta$ la probabilità dell'insuccesso, allora il modello teorico geometrico assumerà la seguente forma:

$$f_i = (1-\theta)^k\theta \quad \text{per} \quad k = 0,1,2,\ldots,$$

dove k è il numero delle prove.

Il modello geometrico dipende da θ e lo indicheremo in sintesi con

$$x \approx G(\theta).$$

9.1.5 Distribuzione Binomiale Negativa

Una generalizzazione della distribuzione geometrica è data dalla distribuzione Binomiale negativa che ci fornisce la frequenza relativa di ottenere un successo dopo che in $x+r-1$ prove il numero di successi è pari a $r-1$. In tal senso, la distribuzione Binomiale negativa coincide con quella geometrica quando $r = 1$. Il modello della Binomiale negativa assume la seguente forma:

$$f_i = \binom{x_i+r-1}{x_i}\theta^r(1-\theta)^{x_i} \quad \text{per} \quad x_i = 0,1,2,\ldots \quad \text{e} \quad r = 0,1,2,\ldots$$

che, come detto, ci fornisce la frequenza relativa di ottenere x insuccessi prima di ottenere l'r-esimo successo. In questo senso, la distribuzione Binomiale Negativa viene talvolta chiamata *Tempo di Attesa* poichè può intendersi l'interprete di quanto è necessario attendere in termini di insuccessi, prima che si verifichino esattamente r successi.

Senza entrare troppo nei dettagli diciamo che la media è:

$$\mu = \frac{r(1-\theta)}{\theta}$$

mentre la varianza è :

$$\sigma^2 = \frac{r(1-\theta)}{\theta^2}.$$

Continuando con la notazione solita, indicheremo il modello binomiale negativo come:

$$x \approx Bn(r, \theta).$$

9.1.6 Distribuzione ipergeometrica

Una distribuzione teorica che rappresenta uno schema simile a quello binomiale è la distribuzione ipergeometrica. Essa si distingue dal caso precedente in quanto si assume che il collettivo sia finito e composto da N unità, di cui r di un tipo e $N-r$ di un altro tipo e che l'estrazione delle unità dal collettivo viene fatta n volte senza ripetizione (reintroduzione della unità nel collettivo). In questo schema la frequenza relativa di trovare per estrazione senza ripetizione un elemento del primo tipo è

$$\theta = \frac{r}{N}$$

e la frequenza di trovare per estrazione senza ripetizione un elemento del secondo tipo è $1-\theta$ ottenuta da

$$1-\theta = \frac{N-r}{N}$$

L'espressione formale della distribuzione ipergeometrica può essere sinteticamente riassunta nell' espressione $x_i \sim H(N, n, \theta)$
In dettaglio la frequenza relativa di ottenere unità del primo tipo da un campione di numerosita n estratto dalla popolazione di ampiezza N è:

$$f_i = \frac{\binom{r}{x_i}\binom{N-r}{n-x_i}}{\binom{N}{n}} \quad \text{per} \quad max(0, n-N+r) \leqslant x_i \leqslant min(n, r).$$

Valore medio e varianza valgono rispettivamente:

$$\mu = \frac{nr}{N} = n\theta$$

$$\sigma^2 = \frac{nr}{n}\frac{N-r}{N}\frac{N-n}{N-1} = n\theta(1-\theta)\frac{N-n}{N-1}$$

Dato che l'asimmetria e la curtosi hanno espressioni complesse vengono tralasciate in questa trattazione.

Si dimostra facilmente che per N molto grande ed n piccolo, questa distribuzione si approssima alla binomiale. Infatti, quando la numerosità del collettivo è molto grande, lo schema di estrazione con ripetizione e lo schema di estrazione senza ripetizione tendono a coincidere, in quanto la probabilità di ripescare la stessa unità ad ogni estrazione tende a zero. Si dice, in questo contesto, che i due schemi sono somiglianti. Da un punto di vista formale, si ha che la distribuzione ipergeometrica tende, al crescere di N, alla distribuzione binomiale ossia:

$$\lim_{N\to+\infty} \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \binom{N}{x}\theta^x(1-\theta)^{n-x}$$

9.1.7 Distribuzione di Poisson

Come abbiamo visto il fenomeno reale schematizzato dalla distribuzione binomiale dipende dalla numerosità delle ripetizioni indipendenti indicato con n e dalla frequenza relativa θ dei successi della misura di cui si vuole conoscere il numero degli esiti favorevoli. In questo contesto si può immaginare che il numero delle ripetizioni n sia elevato, ossia un numero molto grande di ripetizioni e che la frequenza relativa sia piccola nel senso che la misura di cui si vuole conoscere il numero degli esiti sia rara nel collettivo posto a valutazione.

In questo caso è utile introdurre una nuova distribuzione teorica nota in letteratura con il nome di distribuzione di *Poisson* detta anche *legge degli eventi rari.* Essa schematizza una variabile di fondamentale importanza ed è utile per determinare il numero di volte in cui una misura qualitativa d'interesse poco frequente si verifica in un dato intervallo di tempo (o spazio). Per meglio chiarire lo schema richiamato si supponga di essere interessati a conoscere il numero di incidenti stradali che possono avvenire in un arco di tempo su un tratto di autostrada. È facilmente intuibile che sul totale di transito di autovetture, nel tratto di indagine, la frequenza di incidenti è piccola e prossima

allo zero (una frequenza moderatamente elevata ad esempio $\theta = 0.3$ significherebbe che su cento macchine transitate 30 sarebbero vittima di un inciderete, situazione ovviamente da scongiurare e comunque fortemente inverosimile). Visto che tale frequenza è sensibilmente bassa, per poter assistere ad un caso di incidente si rende necessario considerare un numero elevato di transiti di auto. È, infatti, inutile soffermarsi a riflettere che i casi di incidente stradale sono maggiori all'aumentare del traffico stradale. Il modello si dice degli eventi rari appunto perchè è adatto a descrivere i fenomeni in cui si ha un grande numero di prove e la frequenza del successo è piccola.

Formalmente indicando la media del modello binomiale con $\lambda = n\theta$, con n indipendente da θ si ha che:

$$\theta = \frac{\lambda}{n} \to 0 \quad \text{per} \quad n \to \infty$$

quindi, svolgendo il limite

$$\lim_{n\to\infty} \binom{n}{x_i} \theta_i^x (1-\theta)^{n-x_i}$$

e ponendo al posto di $\theta = \frac{\lambda}{n}$ si può dimostrare che:

$$\lim_{\theta\to 0, n\to\infty} \binom{n}{x_i} \frac{\lambda^{x_i}}{n} \left(1-\frac{\lambda}{n}\right)^{n-x_i} = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \qquad (9.1.7.1)$$

Dimostrazione: come abbiamo detto la distribuzione binomiale, dopo le sostituzioni $\theta = \frac{\lambda}{n}$, può essere riscritta come segue:

$$f_i = \binom{n}{x_i} \binom{\lambda}{n}^{x_i} \left(1-\frac{\lambda}{n}\right)^{n-x_i} = \frac{n!}{x_i!(n-x_i)!} \binom{\lambda}{n}^{x_i} \left(1-\frac{\lambda}{n}\right)^{n-x_i} =$$

$$= \frac{(n)(n-1)(n-2)...(n-x_i+1)(n-x_i)!}{x_i!(n-x)!} \binom{\lambda}{n}^{x_i} \left(1-\frac{\lambda}{n}\right)^{n-x_i} =$$

$$= \frac{(n)(n-1)(n-2)...(n-x_i+1)}{n^{x_i}} \frac{\lambda^{x_i}}{x_i!} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-x_i}$$

Calcolando il limite di ogni membro della funzione si ha:

$$\lim_{n\to\infty} \frac{n(n-1)(n-2)...(n-x_i+1)}{n^{x_i}} = 1$$

in quanto limite per $n \to \infty$ di un rapporto di polinomi di grado x è:

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n} = e^{-\lambda}$$

e trattandosi di limite notevole si ha:

$$\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{x_i} = 1.$$

in quanto il limite per $n \to \infty$ di $\frac{\lambda}{n} \to 0$. Di conseguenza 1 elevato alla $-x_i$ tende a 1.

In conclusione il limite del modello binomiale, per n che tende ad infinito, si riduce a:

$$\lim_{n\to\infty} \frac{(n)(n-1)(n-2)...(n-x_i+1)}{n^{x_i}} \frac{\lambda^{x_i}}{x!} \left(1-\frac{\lambda}{n}\right)^{n} \left(1-\frac{\lambda}{n}\right)^{-x_i} = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Quindi il modello di Poisson ha la seguente espressione finale:

$$f_i = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} \quad \text{per} \quad x_i = 0, 1, 2, ..., .$$

Facciamo notare che il modello decisionale di Poisson è caratterizzato dal solo parametro λ. Da cui si deduce che le frequenze relative teoriche della distribuzione ricavata (nell'esempio il numero di incidenti nel tratto autostradale nel periodo di tempo esaminato riportato in tab. 9.3) sono calcolabili dall'espressione seguente:

$$f_i = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

Questa distribuzione si indica come $x \sim P(\lambda)$ e, per quanto appena detto, il limite della distribuzione binomiale per $n\theta = \lambda$ e $n \to \infty$, il che significa che la *Poisson* è un caso limite della binomiale e, quindi, associati allo stesso schema sperimentale.

In effetti se θ è molto piccolo, il numero medio di eventi sarà molto più piccolo di n, quindi il numero di successi x sarà estremamente più piccolo di n. È altrettanto dimostrabile (la dimostrazione viene omessa ma rinviata a testi specializzati) che la media e la varianza coincidono

numero di incidenti x	Frequenze relative f_i
0	f_0
1	f_1
2	f_2
⋮	
i	f_i
⋮	
n	f_n

Tabella 9.3: Esempio di distribuzione di Poisson

e risultano essere uguali all' unico parametro della distribuzione ossia a λ.

$$\mu = \lambda$$
$$\sigma^2 = \lambda$$

Mentre gli indici di forma, simmetria e curtosi risultano essere:

$$ASI(x) = \frac{1}{\sqrt{\lambda}}$$

$$CUR(x) = 3 + \frac{1}{\sqrt{\lambda}}$$

Va sottolineato che il grafico illustrato nella figura 9.2 non è simmetrico e che l'asimmetria è positiva. Si nota, inoltre, che al crescere di λ la distribuzione tende alla simmetria. Similmente alla distribuzione binomiale, il ricercatore che vuole utilizzare la distribuzione teorica di Poisson ha bisogno di conoscere il suo parametro λ. Anche in questo caso quest'aspetto viene risolto attraverso la stima campionaria la cui trattazione esula questa parte di programma.

9.1.8 Applicazione di alcune distribuzioni teoriche discrete

Nell'analisi statistica dei fenomeni territoriali esiste una gran varietà di configurazioni relative alla distribuzione degli oggetti sul

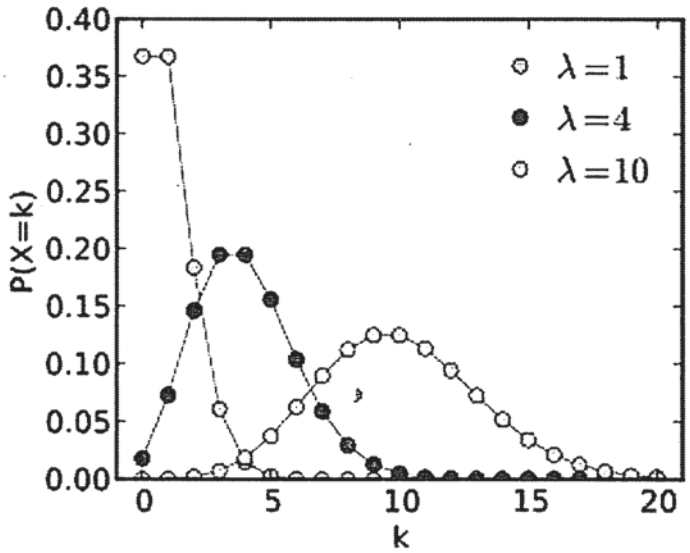


Figura 9.2: Funzione di distribuzione per tre diversi λ

territorio. Nella maggior parte dei casi, dette configurazioni sono formate dalla disposizione di un insieme di punti (pattern puntuali). I pattern sono, quindi, delle mappe che forniscono una rappresentazione spaziale di informazioni. Nel caso in cui le informazioni sono le coordinate geografiche di determinati oggetti, la mappa è detta pattern di punti. In realtà, si considerano degli oggetti come i plessi scolastici, le residenze di extra comunitari, i supermercati e altro, che non sono adimensionali come i punti, ma possono essere considerati come oggetti infinitesimali. La rappresentazione in un pattern di tali oggetti è comunque possibile dato che la loro dimensione fisica è trascurabile sia rispetto alla distanza che li separa, sia rispetto all'area geografica che occupano. Nelle configurazioni puntuali si riscontra che, così come esistono particolari oggetti che tendono ad agglomerarsi o a raggrupparsi, esistono altri che tendono a diffondersi in modo molteplice nello spazio.

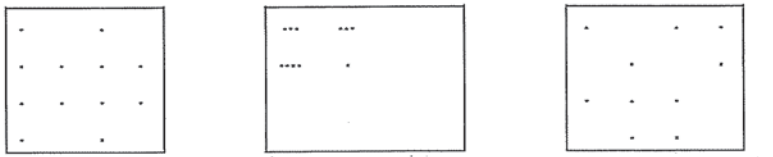
È possibile considerare la superficie come divisa in celle al fine di riuscire a discretizzare il territorio, individuando aree omogenee di pari dimensioni e pari superfici. Diciamo che la ripartizione territoriale

divide lo spazio continuo in N repliche del fenomeno reale.

Così facendo si ottiene una griglia, composta da celle omogenee, ciascuna comprendente, ad esempio, un km^2 di superficie. Si ipotizza che la disposizione degli oggetti nelle celle, che d'ora in poi si considerano associati ai punti in cui sono localizzati, sia il risultato di un processo stocastico spaziale, nel nostro caso a distribuzioni teoriche la cui variabile d'interesse è il numero di punti osservati nello spazio.

Quanto detto significa che i punti-oggetto si disporranno non secondo una regola deterministica, ma tenendo conto delle forze di natura fisica, economica e sociale presente sul territorio. Ad esempio, se si considerano le residenze degli extra-comunitari in una determinata area geografica, si osserva facilmente che esse sono disposte sul territorio secondo scelte di convenienza. La loro disposizione non è certamente stabilita deterministicamente, ad esempio, da un piano predefinito da un organo di governo. Quindi, se è in atto una forza esogena di tipo competitiva, questa darà luogo ad una insieme di punti disposti lontano l'uno dagli altri producendo una sorta di equidistribuzione. Se, invece, in alcune zone è in atto una forza *attrattiva*, si osserverà una tendenza alla concentrazione di punti in tali zone. è facile immaginare, nel caso degli extra comunitari, di osservare una agglomerazione della stessa etnia nella stessa zone. Infine, se non vi è alcuna tendenza alla localizzazione, i punti tenderanno a collocarsi casualmente.

In accordo con quanto detto, si possono rappresentare graficamente degli esempi di collocazione di punti (Fig. 9.3).



a)localizzazione a seguito della presenza di una forza competitiva..

b) localizzazione addensata

c)localizzazione casuale

Figura 9.3: Rappresentazione grafica dei diversi tipi di collocazione dei punti

L'obiettivo del ricercatore è quello di analizzare, con procedure statistiche il criterio di allocazione dei punti sul territorio allo scopo di valutarne la dinamica e, semmai, controllare e condizionare il fenomeno in oggetto. L'approccio che proponiamo è quello di formalizzare i fenomeni territoriali attraverso un modello teorico, che spieghi verosimilmente la disposizione dei punti sul territorio. Da quanto detto è evidente che, soprattutto nel caso in cui la conoscenza del fenomeno è limitata, il point-pattern assume un ruolo decisivo in quanto costituisce la fase preliminare di un processo di descrizione e formalizzazione del fenomeno stesso.

A seconda dell'area territoriale, è possibile associare modelli di tipo regolare, concentrato o accidentale (casuale). Se, nella loro collocazione, i punti manifestano tra loro indifferenza, il processo si definisce *accidentale*; se i punti tendono a respingersi, si avrà un processo definito *regolare* (*ogni punto è equidistribuito sul territorio*); se, infine, i punti manifestano reciproca attrazione, il processo è detto *concentrato*.

La domanda da porsi a questo punto è la seguente: Qual è il modello teorico che genera una situazione di tipo regolare, oppure concentrata, o ancora una situazione spontanea?

Per quanto riguarda il processo accidentale è possibile dimostrare che il relativo modello sarà quello di Poisson, in quanto possiamo considerare il piano composto da infiniti punti. La frequenza relativa che a caso ci si collochi in un dato punto del territorio è molto piccola, prossima allo 0. Il modello che genera gli eventi rari è quello di Poisson, mentre il modello teorico ad un processo regolare sarà quello Binomiale. Infine, per un processo concentrato si dimostra che il modello è quello della Binomiale negativa.

In sintesi, possiamo riassumere quanto detto con la tabella 9.4.

Processo	Modello	Modello teorico	Parametri	Media	Varianza
Casuale	Poisson	$f_i = \frac{\lambda_i^x}{x_i!}e^{-\lambda}$	λ	$\mu = \lambda$	$\sigma^2 = \lambda$
Regolare	Binomiale	$f_i = \binom{n}{x_i}\theta^{x_i}(1-\theta)^{n-x_i}$	n, θ	$\mu = n\theta$	$sigma^2 = n\theta(1-\theta)$
Concentrata	Binomiale Negativa	$f_i = \binom{x_i+r-1}{x_i}\theta^r(1-\theta)^{x_i}$	r, θ	$\mu = \frac{r(1-\theta)}{\theta}$	$\sigma^2 = \frac{r(1-\theta)}{\theta^2}$

Tabella 9.4: Modelli da applicare nei diversi tipi di collocazione dei punti

L'obiettivo sarà, quindi, quello di stimare adeguatamente i parametri del modello sulla base dei dati osservati e calcolare opportuni indicatori che ci aiutano a stabilire quale processo determina la disposizione dei punti sul territorio. Un metodo è quello di conteggiare il numero di punti osservati nei quadrati posizionati nell'area di studio per poi ricavarne la distribuzione di frequenza empirica ottenuta dai conteggi.

A partire dalla distribuzione empirica si può calcolare un indice per stabilire il tipo di processo presente in una determinata area, ad esempio l'indice di Fisher

$$\varphi = \frac{\sigma^2}{\mu_1}.$$

Infatti, bisogna ricordare che, se il processo sottostante è accidentale, ovvero generato da un modello di Poisson, si avrà che:

$$\mu = \sigma^2.$$

Se il processo, invece, è concentrato, la variabilità del numero dei punti è più elevata di quell'attesa nel caso di accidentalità completa e di conseguenza:

$$\mu < \sigma^2.$$

Infine, nel caso di regolarità, la variabilità del numero di punti sarà inferiore a quella attesa nel caso di accidentalità completa e di conseguenza:

$$\mu > \sigma^2.$$

Sulla base di queste considerazioni, è naturale che l'indice di Fisher assumerà valore 1, $\varphi = 1$, nel caso di processo accidentale (Modello di Poisson), valore maggiore di 1, $\varphi > 1$, nel caso di processi concentrati (modello Binomiale Negativa), e valori minori di 1, $\varphi < 1$, nel caso di processi regolari (Modello Binomiale).

Una volta stabilito il tipo di modello teorico seguono le analisi interpretative del fenomeno, attraverso lo studio dei momenti.

9.2 Modelli teorici per variabili continue

Nel corso di questo capitolo abbiamo introdotto le distribuzioni teoriche esplicitate da una funzione matematica del tipo: $y = f(x)$ in

cui il dominio della funzione è espresso da un intervallo limitato o illimitato dell'asse reale a cui fa corrispondere una curva che nell'accezione del linguaggio matematico può essere considerata continua, ma anche con punti di discontinuità, schematicamente la possiamo indicare $\{x, f(x)\}$. Di seguito portiamo alcune tra i più comuni modelli teorici che rientrano in questa categoria.

9.2.1 Distribuzione uniforme

Il più banale dei modelli per le variabili continue è quello uniforme. Questo tipo di modello ha scarsa applicazione in ambito reale, ma viene più volte utilizzato nelle simulazioni quando si vogliono generare numeri casuali in un definito intervallo di numeri reali.

Da un punto di vista formale, la variabile uniforme la pensiamo definita nell'intervallo chiuso $[a, b]$ e diremo che, in detto intervallo, la variabile assumerà per ogni punto la stessa frequenza relativa. Mentre, all'esterno di questo intervallo, suddetta variabile sarà nulla.

Formalmente scriveremo:

$$f_i = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \textit{altrimenti} \end{cases} \qquad (9.2.1.1)$$

Da un punto di vista grafico il modello decisionale uniforme è presentato in fig. 9.4.

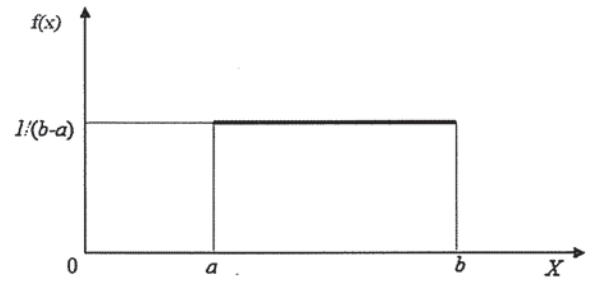


Figura 9.4: Distribuzione uniforme

Così come abbiamo fatto in precedenza il modello teorico uniforme lo indicheremo $x \approx U(a, b)$. Per calcolare i valori di sintesi e di

variabilità di questo modello, vista l'estrema banalità del modello stesso possiamo calcolare i momenti direttamente sul modello.

In particolare, si ha:

$$\mu = \int_a^b x \frac{1}{b-a} dx =$$

$$= \frac{1}{2(b-a)} x^2 \Big]_a^b =$$

$$= \frac{1}{2(b-a)} (b^2 - a^2) =$$

$$= \frac{1}{2(b-a)} (b-a)(b+a),$$

pertanto la media del modello decisionale uniforme è: $\mu = \frac{b+a}{2}$. Proseguendo allo stesso modo possiamo calcolare il momento secondo che, dopo il calcolo dell'integrale, risulta essere:

$$\mu = \int_a^b x \frac{1}{b-a} dx = \frac{1}{2(b-a)} x^2 \Big]_a^b$$

$$= \frac{1}{2(b-a)} (b^2 - a^2) = \frac{1}{2(b-a)} (b-a)(b+a)$$

Infatti:

$$\mu_2 = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x^2 dx = \frac{1}{b-a} \left[\frac{1}{3} x^3 \right]_a^b$$

$$= \frac{1}{b-a} \left(\frac{1}{3} b^3 - \frac{1}{3} a^3 \right) = \frac{1}{3(b-a)} (b^3 - a^3)$$

$$= \frac{1}{3(b-a)} (b-a)(b^2 + ab + a^2)$$

Ricordando infine che la varianza è uguale al momento secondo meno il momento primo al quadrato, si ricava che la varianza del modello uniforme è:

$$\sigma^2 = \frac{b^3 - a^3}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}$$

Infatti:

$$\sigma^2 = \frac{b^2 + ab + a^2}{3} - \frac{(a+b)^2}{4}$$

$$= \frac{b^2 + ab + a^2}{3} - \frac{a^2 + 2ab + b^2}{4} = \frac{(b-a)^2}{12}$$

Allo stesso modo si possono calcolare gli altri momenti necessari per ottenere i corrispondenti indici forma.

Il modello decisionale uniforme è un modello di scarsa utilità pratica, ma presenta una forte valenza teorica.

9.2.2 Distribuzione normale o di Gauss

Il modello teorico sicuramente più famoso per le molteplici applicazioni sia a fenomeni reali, sia per le notevoli proprietà asintotiche è il *modello normale* (o *curva di Gauss*). Questo particolare modello assume un duplice aspetto. Da un lato può essere utilizzato, così come abbiamo visto finora, come un modello che approssima molto verosimilmente una moltitudine di casi reali. In particolare, la maggior parte delle misure antropometriche si suppongono distribuirsi secondo questo modello. Dall'altro lato, invece, suddetto modello assume un'importanza fondamentale nell'ambito della teoria dell'inferenza statistica, argomento che si vedrà in seguito. Limitandoci, dunque, al primo aspetto, diremo che questo tipo di modello bene interpreta le misure quantitative di fenomeni reali che si distribuiscono simmetricamente intorno ad un valore rappresentativo, per esempio la media aritmetica.

Da un punto di vista formale, il modello decisionale si presenta come segue:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad (9.2.2.1)$$

dove μ e σ rappresentano, rispettivamente, i parametri del modello e, a loro volta, coincidono con la media e la varianza. Da un punto di vista compatto, indicheremo detto modello come $x \approx N(\mu, \sigma^2)$. In particolare, la media è anche il punto di ascissa del massimo della funzione, in altri termini, il valore a cui corrisponde la massima frequenza relativa; mentre σ, ossia lo scarto quadratico medio, è la distanza che

intercorre tra la media aritmetica e i flessi della distribuzione. Da un punto di vista interpretativo, lo scarto quadratico medio è una misura della dispersione della misura della variabile osservata rispetto al valore di sintesi, ovvero la media.

Il modello di Gauss può essere visto anche come distribuzione di una variabile casuale che chiameremo variabile casuale normale. Un risultato molto utile in particolare quando tratteremo dell'inferenza statistica è la combinazione lineare di variabili casuali normali, dove diciamo che una combinazione di variabili casuali indipendenti è ancora una variabile casuale la cui espressione formale è:

$$Y = \sum_{i=}^{n} a_i X_i \qquad (9.2.2.2)$$

dove a_i, per $i = 1, 2, \ldots, n$, sono costanti chiamate anche pesi.

In particolare si dimostra il seguente risultato generale: siano $X_1, X_2, \ldots, X_n$ le n variabili casuali normali indipendenti distribuite con media $\mu_1, \mu_2, \ldots, \mu_n$ e varianza $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, allora la variabile casuale

$$Y = \sum_{i=1}^{n} a_i X_i$$

combinazione lineare di dette variabili casuali, è ancora una variabile casuale normale con media

$$\mu_Y = \sum_{i=1}^{n} a_i \mu_i$$

e varianza

$$\sigma^2 = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

In sintesi, sarà quindi:

$$Y = N\left(\mu_y \sigma_y^2\right) \qquad (9.2.2.3)$$

Nell'ipotesi particolare che i pesi siano $a_1 = \frac{1}{n}, \forall i$ e che la media e la varianza di ciascuna variabile casuale siano costanti e uguali a μ e a σ^2, allora la variabile casuale combinazione lineare di variabili casuali

indipendenti ed identicamente distribuite è la media delle n variabili casuali $X_1, X_2, \ldots, X_n$ che indichiamo, per uniformità a quanto seguirà, $\hat{\mu} = \frac{1}{n} \sum x_i$ (i.i.d.).

Sostituendo ai pesi a_i del risultato generale $\frac{1}{n}$ si ricava il risultato utile ai fini dell'inferenza statistica, che vedremo in seguito:

$$\hat{\mu} = N\left(\mu, \frac{\sigma_2}{n}\right) \qquad (9.2.2.4)$$

Variabile normale standardizzata

In molti casi applicativi, dove il modello più appropriato è quello binomiale, ci si scontra con il problema di effettuare un numero di repliche indipendenti molto elevato (si pensi ad esempio ad un'elezione elettorale dove il numero delle repliche, ossia degli elettori, è dell'ordine dei milioni). Quando si opera con il modello binomiale, ci si accorge che il calcolo del coefficiente binomiale è intrattabile da un punto di vista computazionale. Per risolvere questo problema, si ricorre ad un risultato asintotico dovuto a De Moivre, che afferma: quando $p \approx 0,5$, la distribuzione della binomiale è simmetrica rispetto alla media, cioè np. Pertanto, effettuando un'opportuna trasformazione lineare, detta anche standardizzazione, della variabile x nella variabile casuale $z = \frac{x-\mu}{\sigma}$ si ha che, per il modello binomiale $z = \frac{x-np}{\sqrt{np(1-p)}}$, detta variabile tende, per $n \to \infty$, ad una distribuzione normale di media zero e varianza 1 ossia $z \approx N(0,1)$. Questo risultato, come si intuisce, è molto importante dal momento che il modello binomiale è, sotto queste condizioni, approssimabile al modello normale.

9.2.3 Distribuzione Gamma

Il modello gamma assume un'importanza fondamentale nella trattazione di molte variabili continue. La sua importanza risiede nel fatto che da esso si generano diversi modelli teorici, così come vedremo in seguito. Da un punto di vista applicativo, il modello gamma bene si adatta all'interpretazione di fenomeni riguardanti il campo della ricerca operativa e del controllo statistico della qualità. Da un punto

di vista formale, si ha la seguente distribuzione:

$$f(x) = \frac{\theta}{\Gamma(\alpha)} x^{\alpha-1} e^{-\theta x} \quad 0 \leqslant x \leqslant \infty \tag{9.2.3.1}$$

Il nome gamma deriva dall'integrale omonimo che, come noto, ha la seguente espressione:

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx \tag{9.2.3.2}$$

È facile verificare che, dopo opportune integrazioni per parti, risulta essere uguale a $\Gamma(\alpha) = (\alpha - 1)$. In sintesi, il modello gamma lo scriveremo:

$$x \approx Ga(\alpha, \theta) \tag{9.2.3.3}$$

dove α e θ sono i parametri del modello da stimare.

Di notevole interesse metodologico è, come abbiamo più volte visto, il calcolo dei momenti da cui è possibile ricavare indicazioni sulla sintesi, la variabilità e la forma.

A tal fine senza entrare nei dettagli che esulano la trattazione di questo testo diamo i principali risultati

$$\mu = \frac{\alpha}{\theta}; \quad \sigma^2 = \frac{\alpha}{\theta^2} \tag{9.2.3.4}$$

Da un punto di vista applicativo, il modello gamma trova ampia applicazione come modello di "sopravvivenza" nello studio della durata di apparecchi soggetti a logorio, oppure nella teoria delle "file di attesa".

9.2.4 Distribuzione chi-quadrato

La prima diretta derivazione del modello gamma è il modello *chi-quadrato*. Esso si ricava ponendo $\theta = \frac{1}{2}$ ed $\alpha = \frac{r}{2}$, con r numero intero positivo chiamato "gradi di libertà". In particolare, sostituendo quanto appena detto al modello gamma, si ha la nota espressione formale del modello chi-quadrato che risulta essere la seguente:

$$f(x) = \frac{1}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})} x^{\frac{r}{2}-1} e^{-\frac{x}{2}} \tag{9.2.4.1}$$

In sintesi, il modello chi-quadrato lo scriviamo $x \approx \chi^2(r)$, da cui si evince che l'unico parametro da stimare, per definire la funzione della distribuzione, è il numero dei gradi di libertà.

Analogamente a quanto detto sopra, è possibile ricavare la media e la varianza che risaltano essere:

$$\mu = r$$

e

$$\sigma^2 = 2r$$

Come meglio vedremo più in avanti, il modello chi-quadrato assume una notevole importanza nell'ambito dell'inferenza statistica.

A questo proposito, risulta essenziale derivare un suo risultato asintotico. Infatti, si può dimostrare che la variabile casuale $\sqrt{2\chi^2}$ si distribuisce asintoticamente al erescere dei gradi di libertà come una curva normale i cui parametri sono $\mu = \sqrt{2r-1}$ e $\sigma^2 = 1$. Cioè, in sintesi:

$$\sqrt{2\chi^2} \approx N(\sqrt{2r-1}, 1) \quad 0 \leqslant x \leqslant \infty \tag{9.2.4.2}$$

Questo risultato, in altri termini, lega la forma distribuzionale del chi-quadrato a quella della normale il cui significato sarà meglio specificato nel capitolo relativo all'inferenza statistica.

Un altro risultato estremamente importante ai fini dell'inferenza statistica è la variabile somma di normali standardizzate al quadrato. A questo proposito, si può dimostrare il seguente teorema.

Siano $X_1, X_2, \ldots, X_n$, n *variabili indipendenti ed identicamente distribuite (i.i.d.) secondo una normale* $N(\mu, \sigma^2)$, *allora la variabile così ottenuta*

$$Y = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma} = \frac{nS^2}{\sigma^2} \tag{9.2.4.3}$$

si distribuisce secondo una chi-quadrato, con n *gradi di libertà.*

Il significato intrinseco di questo risultato sarà chiarito in seguito nell'ambito della teoria della stima.

9.2.5 Distribuzione di Snedecor-Fisher

Siano $\chi^2_{r_1}$ e $\chi^2_{r_2}$ due chi-quadrato fra loro indipendenti aventi funzione di distribuzione:

$$f(x) = \frac{1}{2^{r_i/2}\Gamma(r_i/2)} x_i^{r_i/2-1} e^{-x_i/2} \qquad x_i > 0,$$

allora la distribuzione

$$F_{r1,r2} = \frac{\chi^2_{r_1}/r_1}{\chi^2_{r_2}/r_2} = \frac{\chi^2_{r_1}}{\chi^2_{r_2}} \frac{r_2}{r_1}$$

è chiamata di Snedecor con r_1 e r_2 gradi di libertà.

In sintesi, possiamo scrivere il suddetto modello nella forma

$$x \approx F(r_1, r_2)$$

da cui si evince come gli unici parametri da dover stimare sono, per l'appunto, i gradi di libertà r_1 e r_2.

9.2.6 Distribuzione esponenziale

Un'altra distribuzione che può essere derivata dalla Gamma e che riveste un interesse particolare è la distribuzione Esponenziale, la cui funzione di distribuzione si ricava ponendo $\alpha = 1$ e $\theta = \lambda$, ottenendo così:

$$f(x) = \lambda e^{-\lambda x} \qquad 0 \leqslant x \leqslant \infty$$

Questa distribuzione è anche nota come modello di Laplace o Esponenziale negativa, essendo negativo l'esponente del numero "e" base dei logaritmi neperiani che in essa figura. Il modello esponenziale gode della proprietà di assenza di memoria, nel senso che i valori della variabile assunti nel futuro non sono condizionati dai valori assunti nel passato (questo concetto sarà meglio chiarito nell'esempio che segue).

In sintesi, il modello decisionale esponenziale possiamo scriverlo come $x \approx Exp(\lambda)$, da cui si evince che l'unico parametro da dover stimare è appunto λ . Anche in questo caso la media e la varianza sono:

$$\mu = \frac{1}{\lambda}; \qquad \sigma^2 = \frac{1}{\lambda^2}.$$

La distribuzione esponenziale può considerarsi come un'estensione del modello di Poisson, dal momento che ha come obiettivo quello di definire la distribuzione del tempo di attesa della successiva realizzazione.

A titolo di esempio, supponiamo di studiare la durata in vita, espressa in ore, di una lampadina. Questo tipo di variabile è approssimabile attraverso la funzione esponenziale. Supponiamo, per semplicità, che il parametro del modello sia stato adeguatamente stimato e che sia pari a $\lambda = 0.002$, ossia che in media la lampadina duri 50 ore. Posto che la lampadina sia accesa da oltre 45 ore, si vuole calcolare la probabilità che la stessa resti in vita almeno altre 30 ore, vale a dire:

$$P[X > (45 + 30) | X > 45]$$

Dato che il modello esponenziale gode, come si è detto, della proprietà di assenza di memoria, calcolare la probabilità di cui sopra equivale a calcolare la sola probabilità

$$P(X > 30).$$

In conclusione, applicando la funzione esponenziale si ha che:

$$P(X > 30) = 0.02e^{-(0.002)(30)} = 0.02e^{-0.6} = 0.011$$

In altre parole, la probabilità che la lampadina non si fulmini, rimanendo accesa per 75 ore (45+30) consecutive, condizionata dal fatto che la medesima è già rimasta in vita oltre 45 ore, risulta uguale alla probabilità non condizionata che essa possa rimanere accesa senza fulminarsi per almeno 30 ore. Così, date due lampadine di questo tipo, l'una nuova e l'altra già accesa da un certo numero di ore, entrambe hanno la medesima probabilità di non fulminarsi entro un determinato numero di ore di funzionamento. È infatti noto che il fulminarsi di una lampadina non è determinato dall'usura dei materiali che la compongono, che risulta trascurabile, ma dai fattori che intervengono dall'esterno come, ad esempio, le attivazioni e disattivazioni o, ovviamente, il frantumarsi del vetro, gli sbalzi di temperatura e corrente. In questo senso va interpretata la locuzione "assenza di memoria".

9.2.7 Distribuzione di Weibull

Una diretta derivazione del modello esponenziale è quello di Weibull, introdotto dallo svedese omonimo per interpretare il fenomeno della resistenza alla rottura dei materiali.

Esso si ottiene dalla esponenziale ponendo $\lambda = \theta_2$ e trasformando la variabile $y = x^{\frac{1}{\theta_1}}$, da cui si ricava che $x = y^{\theta_1}$. A seguito di questa trasformazione, il differenziale è $dx = \theta_1 y^{\theta_1 - 1} dy$. Sostituendo opportunamente al modello $f(x) = \lambda e^{-\lambda x}$, si ottiene immediatamente il modello di Weibull che risulta essere il seguente:

$$f(y) = \theta_1 \theta_2 y^{\theta_1 - 1} e^{\theta_2 y^{\theta_1}} \quad \text{per} \quad y > 0$$

In sintesi, il modello di Weibull lo indicheremo $y \approx W(\theta_1, \theta_2)$, dove θ_1 e θ_2 sono rispettivamente i parametri da stimare. Anche per questo modello è possibile ricavare i momenti. Per semplicità riportiamo soltanto i risultati della media e della varianza:

$$\mu = \theta_2^{-\frac{1}{\theta_1}} \Gamma(\theta^{-1} + 1), \qquad \sigma^2 = \theta_2^{-\frac{2}{\theta_1}} \{\Gamma(2\theta_1{}^{-1} + 1) - [\Gamma(\theta_1^{-1} + 1)]^2\}$$

dove, al solito, $\Gamma(.)$ rappresenta l'integrale gamma introdotto nell'ambito del modello gamma.

9.2.8 Distribuzione Beta

Il modello, la cui funzione di densità è

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} \quad 0 < x < 1,$$

$$\text{con} \quad (\alpha > 0, \beta > 0) \quad \text{e con} \quad B(\alpha, \beta) = \int_0^1 z^{\alpha - 1} (1 - z)^{\beta - 1} dz$$

integrale euleriano di prima specie, prende il nome modello Beta.

Esso può ricavarsi dalla Gamma e, precisamente, se x_1 e x_2 hanno distribuzione Gamma indipendenti, con funzione di densità rispettivamente

$$f(x_1) = \frac{1}{\Gamma(\alpha)} x_1^{\alpha - 1} e^{-x_1} \quad \text{e} \quad f(x_2) = \frac{1}{\Gamma(\beta)} x_2^{\alpha - 1} e^{-x_2},$$

allora la variabile

$$x = \frac{x_1}{x_1 + x_2}$$

è di tipo Beta con appunto i parametri α e β.

I momenti si ricavano agevolmente in via diretta considerando:

$$B(\alpha, \beta) = \int_0^1 z^{\alpha - 1} (1 - z)^{\beta - 1} dz = \frac{B(\alpha + r, \beta)}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + r)}{\Gamma(\alpha)\Gamma(\alpha + \beta + r)},$$

da cui per $r = 1$ e $r = 2$, si ottiene:

$$\mu = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + r)}{\Gamma(\alpha)\Gamma(\alpha + \beta + r)} = \frac{\alpha}{\alpha + \beta}$$

$$\mu^2 = \frac{\Gamma(\alpha + \beta)\Gamma(\alpha + r)}{\Gamma(\alpha)\Gamma(\alpha + \beta + r)} = \frac{\alpha(\alpha + 1)}{(\alpha + \beta + 1)(\alpha + \beta)}$$

$$\sigma^2 = \mu_2 - \mu^2 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}.$$

9.2.9 Distribuzione di Pareto

L'economista Pareto ha proposto una distribuzione teorica che ha largo impiego negli studi sulla ricchezza e sulla distribuzione dei redditi. Sotto il profilo analitico, anche la distribuzione di Pareto è un caso particolare della Beta.

Infatti, ponendo $\alpha_1 = 1$, $\alpha_2 = \theta$ e $\alpha_3 = \frac{\alpha_4(1 - \theta_2)}{\theta_2}$, con la trasformazione:

$$y = \frac{\alpha_4}{\alpha_4 - x},$$

$$\text{da cui} \quad x = \frac{\alpha_4(y - 1)}{y} \quad \text{e} \quad dx = \frac{\alpha_4}{y^2} dy,$$

si ottiene il modello di Pareto:

$$f(x) = \theta_1 \theta_2{}^{\theta_1} y^{-(\theta_1 + 1)} \quad \text{con} \quad y > \theta_2 \quad \text{e} \quad (\theta_1, \theta_2 > 0)$$

Va ricordato che numerose applicazioni del modello in questione su dati dei redditi hanno messo in luce la tendenza della medesima ad interpretare con buona approssimazione i redditi più bassi, comunque

non inferiori a θ_2 , e quelli più alti, mentre presenta scarso adattamento per i redditi intermedi.

In sintesi, il modello di Pareto lo indicheremo $y \approx P(\theta_1, \theta_2)$, dove θ_1 e θ_2 sono, rispettivamente, i parametri da stimare.

Anche per questo modello è possibile ricavare i momenti. Tuttavia, per semplicità, riportiamo soltanto i risultati della media e della varianza:

$$\mu = \frac{\theta_1 \theta_2}{\theta_1 - 1},$$

$$\sigma^2 = \frac{\theta_1 {\theta_2}^2}{(\theta_1 - 1)^2(\theta_1 - 2)}.$$

9.2.10 Distribuzione Lognormale

Sia y una variabile con distribuzione normale $N(\mu, \sigma^2)$ e x una variabile che può assumere soltanto valori positivi e tali che:

$$y = \ln x,$$

allora, la variabile casuale x prende il nome di Lognormale e la sua funzione si ottiene dalla normale mediante la seguente trasformazione:

$$x = e^y,$$

$$\text{da cui} \quad y = \ln x \quad \text{e} \quad dy = \frac{dx}{x},$$

risultando:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2} \qquad (-\infty < \mu < \infty, \quad \sigma > 0).$$

In sintesi, il modello Lognormale lo indicheremo $y \approx Lg(\theta_1, \theta_2)$, dove θ_1 e θ_2 sono, rispettivamente, i parametri da stimare. Anche per questo modello è possibile ricavare i momenti ottenedendo:

$$\mu = e^{\mu + \frac{\sigma^2}{2}}, \mu_2 = e^{2(\mu+\sigma^2)} \quad \text{e} \quad \sigma^2 = e^{2(\mu+\sigma^2)} - \left(1 - e^{-\sigma^2}\right).$$

9.3 Interpolazione analitica

Da quanto è emerso nei paragrafi precedenti si intuisce che l'analisi di un fenomeno reale tramite una distribuzione teorica passa attraverso la stima dei parametri dei modelli sopra indicati. Questa fase può essere risolta usando un approccio inferenziale, cioè estraendo un campione casuale dalla popolazione di riferimento (approccio che vedremo meglio in futuro) o attraverso un approccio di interpolazione analitica. Se ci poniamo in quest'ottica allora l'obiettivo è quello di determinare una funzione $y = f(x)$, dove con x indichiamo la variabile del fenomeno reale intesa come variabile indipendente e y la variabile risposta, ad esempio la frequenza associata a ciascuna modalità della variabile x, ritenuta essere dipendente attraverso la funzione $f(x)$ dalla variabile x. Sulla base dell'esperimento empirico il ricercatore disporrà solo delle k coppie di osservazioni $(x_1; y_1), (x_2; y_2), \ldots, (x_k; y_k)$. Ossia le coppie dei dati composte da ciascuna modalità e la corrispondente frequenza assoluta/relativa.

Il primo passo della interpolazione analitica è quello di rappresentare graficamente le coppie dei dati osservati. Questa fase è estremamente utile per stabilire il modello teorico più opportuno o, più in generale, la funzione matematica che potrebbe essere più adatta ad interpolare i dati. Una volta stabilita la funzione si noterà che esistono diverse funzioni dello stesso tipo che potrebbero interpolare i dati. Ad esempio se la funzione scelta è la parabola di equazione $y = ax^2 + bx + c$ allora è immediato capire che ci saranno infinite interpolazioni associati agli infiniti valori dei parametri a b e c. Diremo che a ciascuna funzione scelta come interpolante corrisponde una famiglia parametrica che indicheremo sinteticamente con $y = f(x; \theta_1, \theta_2, \ldots, \theta_s)$ dove $\theta_1, \theta_2, \ldots, \theta_s$ sono appunto i parametri della funzione che dovranno essere stimati per stabilire la specifica funzione interpolante dei dati osservati.

È facilmente intuibile che il passaggio essenziale dell'interpolazione analitica è la stima dei parametri. Questa fase può essere condotta attraverso due procedimenti essenziali.

1. Imponendo che la funzione interpolante passi per i punti definiti dalle coppie di osservazioni $(x_1; y_1), (x_2; y_2), \ldots, (x_k; y_k)$. Questo tipo di interpolazione è scarsamente utilizzato per diversi motivi,

tra i quali il più importante è che in questo contesto si ha bisogno di un modello matematico con un numero di parametri pari alle k coppie di coordinate individuate dai dati osservati. Questo implica che il modello interpolante non è facilmente interpretabile per valori diversi da quelli osservati e non è parsimonioso nel senso che dipende da un numero di parametri elevato.

2. Imponendo che la funzione interpolante si approssima il meglio possibile alle coppie di coordinate osservate si rende necessario definire un criterio di ottimizzazione che restituisca la migliore funzione di interpolazione.

Questo secondo procedimento è sicuramente il più utilizzato e per questo gli daremo maggior risalto, in quanto, sebbene i dati interpolati siano affetti da errori, ci permette di scegliere la più opportuna funzione che interpreta il fenomeno oggetto di studio. Inoltre il metodo è parsimonioso nel senso che non è necessario elevare il numero di parametri da stimare, il che rende più facile l'interpretazione del fenomeno reale che si vuole studiare. Un esempio aiuterà a capire quanto appena detto.

Supponiamo che una azienda di abbigliamento voglia conoscere la distribuzione teorica delle taglie di uno specifico capo. Dai dati aziendali e di produzione/vendita gli sarà facile ricavare per ciascuna taglia, a partire dalle taglie più piccole 38/40 alle più grandi 60/62, il numero (frequenza) di capi venduti. È altrettanto intuibile che le taglie molto piccole così come quelle molto grandi saranno vendute di meno di quelle intermedie. È facile immaginare che la rappresentazione grafica della distribuzione assumerà una forma campanulare che ci farà sospettare una distribuzione interpolante appartenete alla famiglia delle curve normali che possiamo genericamente scrivere come:

$$y = \frac{1}{\sqrt{2\pi c}} e^{-b(x-a)^2}.$$

Per poter stabilire la specifica distribuzione teorica relativa ai dati osservati, si tratterrà di stimare solo i tre parametri a b e c.

Al fine di ottenere una funzione teorica ottimale, nel senso che passi il più vicino possibile ai punti osservati, chiamati anche empirici, bisogna scegliere una misura dello scostamento tra il valore

teorico, che indichiamo con $y_i^* = f(x_i; \theta_1; \theta_2; \ldots; \theta_s)$, ossia il valore della frequenza relativa calcolata sulla funzione teorica scelta, e il valore empirico, che indichiamo con y_i , ossia la frequenza relativa osservata per ciascuna modalità x_i. Indicando con $e_i = y_i^* - y_i$ per $i = 1, 2, \ldots, k$ ciascuna misura dello scostamento, si ha che la funzione teorica ottimale sarà ottenuta minimizzando una opportuna funzione degli scostamenti.

9.3.1 Il metodo dei minimi quadrati

Nel paragrafo precedente abbiamo indicato con $e_i = y_i^* - y_i$ lo scostamento tra il valore teorico e quello empirico e abbiamo detto che la funzione teorica che passa il più vicino possibile ai punti empirici y_i si ottiene stabilendo un criterio che minimizzi un'opportuna funzione degli scostamenti. Il metodo più noto, che dà soluzione a questo problema, è il *metodo dei minimi quadrati*. Esso impone come funzione degli scostamenti la somma degli scarti al quadrato; formalmente si ha:

$$g(\theta_1, \theta_2, \ldots, \theta_s) = \sum_{i=1}^{k} [f(x_i; \theta_1, \theta_2, \ldots, \theta_s) - y_i]^2 \qquad (9.3.1.1)$$

Essendo x_i e y_i, con $i = 1, 2, \ldots, k$, valori osservati empiricamente (quindi dati), la funzione $g(\theta_1, \theta_2, \ldots, \theta_s)$ dipende solo dai parametri della funzione teorica interpolante $y_i^* = f(x_i; \theta_1, \theta_2, \ldots, \theta_s)$.

Ricordando che l'obiettivo è quello di trovare la funzione che passa il più vicino possibile ai punti osservati, il problema si riduce nel trovare per quali valori dei parametri $\theta_1, \theta_2, \ldots, \theta_s$ la 9.3.1.1 è minima.

In definitiva si tratta di risolvere un problema di ricerca del minimo di una funzione s-variata (a s variabili) $\theta_1, \theta_2, \ldots, \theta_s$ le cui soluzioni le indichiamo con $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_s$.

Senza entrare troppo nel merito del metodo analitico della ricerca del minimo di una funzione, ricordiamo che un tale problema viene risolto ricorrendo a strumenti di analisi matematica ed in particolare al ricorso delle derivate parziali[2].

[2]Per maggiori chiarezza si consiglia di consultare un testo specifici di analisi matematica.

In particolare la condizione necessaria ma non sufficiente è che siano nulle simultaneamente le derivate prime parziali $\frac{\partial g(\theta_1,\theta_2,\ldots,\theta_s)}{\partial \theta_1} = 0$, $\frac{\partial g(\theta_1,\theta_2,\ldots,\theta_s)}{\partial \theta_2} = 0, \ldots, \frac{\partial g(\theta_1,\theta_2,\ldots,\theta_s)}{\partial \theta_s} = 0$. In definitiva una soluzione di minimo è data dalla soluzione del sistema

$$\begin{cases} \frac{\partial \sum_{i=1}^{k} [f(x_i;\theta_1,\theta_2,\ldots,\theta_s)-y_i]^2}{\partial \theta_1} = 0 \\ \frac{\partial \sum_{i=1}^{k} [f(x_i;\theta_1,\theta_2,\ldots,\theta_s)-y_i]^2}{\partial \theta_2} = 0 \\ \vdots \\ \frac{\partial \sum_{i=1}^{k} [f(x_i;\theta_1,\theta_2,\ldots,\theta_s)-y_i]^2}{\partial \theta_s} = 0 \end{cases} \qquad (9.3.1.2)$$

Tuttavia, è anche noto che il sistema appena impostato non garantisce che la soluzione sia un minimo, in quanto si potrebbero individuare anche punti di massimo della funzione $g(\theta_1, \theta_2, \ldots, \theta_s)$. Per l'individuazione del minimo, quindi, è necessaria una valutazione delle derivate seconde parziali. Tale analisi però complicherebbe alquanto i calcoli e la procedura. È, però, immediato notare che nel caso specifico la funzione $g(\theta_1, \theta_2, \ldots, \theta_s)$ non può avere un massimo, infatti, dato il problema, è sempre possibile assumere una funzione $f(x_i; \theta_1, \theta_2, \ldots, \theta_s)$ che sia lontana quanto si vuole dai punti le cui coordinate sono $(x_1; y_1), (x_2; y_2), \ldots, (x_k; y_k)$. Ciò ci porta a concludere che non potendo il sistema 9.3.1.2 ammettere soluzioni per l'individuazione di un massimo allora, se il sistema ammette soluzioni finite, esse contraddistinguono necessariamente un punto di minimo della funzione $g(\theta_1, \theta_2, \ldots, \theta_s)$.

Il metodo dei minimi quadrati appena descritto si basa quindi sulla soluzione del sistema di equazioni 9.3.1.1. Va sottolineato che esso non sempre è di facile soluzione. In particolare se la funzione interpolante è lineare o linearizzabile nei parametri, cioè è del tipo:

$$f(x_i; \theta_1, \theta_2, \ldots, \theta_s) = \theta_1 h_1(x) + \theta_2 h_2(x) + \cdots + \theta_s h_s(x)$$

allora il sistema si riduce ad un sistema di s equazioni di primo grado con s incognite e, salvo casi eccezionali, il sistema ammette una ed una sola soluzione.

Nei casi in cui la funzione interpolante non è lineare nei parametri, allora il sistema non è di immediata soluzione. Si deve ricorrere a metodi intensivi che esulano dal contenuto di questo testo. In questi casi il metodo dei minimi quadrati è talmente complesso che se ne sconsiglia l'uso.

A titolo esemplificativo supponiamo di dover interpolare una distribuzione empirica attraverso una parabola che passi il più vicino possibile ai punti $(x_1; y_1), (x_2; y_2), \ldots, (x_k; y_k)$. Da quanto sopra detto la famiglia delle parabole è data da

$$y = ax^2 + bx + c$$

dove a, b e c sono rispettivamente i parametri da stimare. L'obiettivo è quello di scrivere una funzione teorica che a ciascuna x_i, per $i = 1, 2, \ldots, k$, associ un valore teorico $y_i^* = ax_i^2 + bx_i + c$. In sostanza il problema si riduce nello stimare per mezzo del metodo dei minimi quadrati i parametri che indicheremo $\hat{a}$, $\hat{b}$ e $\hat{c}$.

Ripercorrendo le fasi sopra descritte si ha che la funzione da minimizzare è

$$g(a, b, c) = \sum_{i=1}^{k} (ax_i^2 + bx_i + c - y_i)^2 = \min$$

In particolare le derivate parziali rispetto ai tre parametri a, b e c sono

$$\frac{\partial g}{\partial a} = \sum_{i=1}^{k} (ax_i^2 + bx_i + c - y_i)x_i^2 =$$

$$= a\sum_{i=1}^{k} x_i^4 + b\sum_{i=1}^{k} x_i^3 + c\sum_{i=1}^{k} x_i^2 - \sum_{i=1}^{k} x_i^2 y_i$$

$$\frac{\partial g}{\partial b} = \sum_{i=1}^{k} (ax_i^2 + bx_i + c - y_i)x_i =$$

$$= a\sum_{i=1}^{k} x_i^3 + b\sum_{i=1}^{k} x_i^2 + c\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} x_i y_i$$

$$\frac{\partial g}{\partial c} = \sum_{i=1}^{k} (ax_i^2 + bx_i + c - y_i) =$$

$$= a\sum_{i=1}^{k} x_i^2 + b\sum_{i=1}^{k} x_i + kc - \sum_{i=1}^{k} y_i$$

Da cui il sistema delle derivate parziali da risolvere è:

$$\begin{cases} a\sum_{i=1}^{k} x_i^4 + b\sum_{i=1}^{k} x_i^3 + c\sum_{i=1}^{k} x_i^2 - \sum_{i=1}^{k} x_i^2 y_i \\ a\sum_{i=1}^{k} x_i^3 + b\sum_{i=1}^{k} x_i^2 + c\sum_{i=1}^{k} x_i - \sum_{i=1}^{k} x_i y_i \\ a\sum_{i=1}^{k} x_i^2 + b\sum_{i=1}^{k} x_i + kc - \sum_{i=1}^{k} y_i \end{cases}$$

Ricordando che le x_i e le y_i sono dati osservati si vede subito che il sistema appena scritto è un sistema di equazioni di primo grado di tre equazioni in tre incognite, la cui soluzione può essere ottenuta usando uno dei metodi a scelta del lettore, come per esempio il metodo di Kramer.

Facciamo un piccolo esempio numerico. Supponiamo di aver rilevato le temperature medie notturne di una località montana nei tre mesi di ottobre, novembre e dicembre i cui risultati empirici sono contenuti nelle prime due colonne della distribuzione riportata in Tab. 9.5.

	x Temperature medie	Giorni Frequenza assoluta	y Frequenza relativa	x_i^2	x_i^3	x_i^4	$x_i y_i$	$x_i^2 y_i$	Frequenza teorica
	-5	3	0,02	25	-125	625	-0,10	0,50	0,00
	-4	6	0,04	16	-64	256	-0,16	0,64	0,07
	-3	15	0,10	9	-27	81	-0,30	0,90	0,12
	-2	26	0,17	4	-8	16	-0,35	0,69	0,16
	-1	30	0,20	1	-1	1	-0,20	0,20	0,17
	0	25	0,17	0	0	0	0,00	0,00	0,17
	1	20	0,13	1	1	1	0,13	0,13	0,15
	2	18	0,12	4	8	16	0,24	0,48	0,11
	3	7	0,05	9	27	81	0,14	0,42	0,05
Somma	-9	150	1,00	69	-189	1077	-0,59	3,97	1

Tabella 9.5: Esempio per la stima dei minimi quadrati

Sapendo che le coppie di punti sono $k = 9$ si ha il sistema

$$\begin{cases} 1077a - 189b + 69c = 3,97 \\ -189a + 69b - 9c = -0,59 \\ 69a - 9b + 9 = 1 \end{cases}$$

La cui soluzione richiede sicuramente uno strumento di calcolo adeguato.

Applicando il metodo di Kramer si hanno le stime dei parametri

$$\hat{a} = -0,009 \quad \hat{b} = -0,012 \quad \hat{c} = 0,17$$

che definiscono la seguente distribuzione teorica $y_i^* = -0,009x_i^2 - 0,012x_i + 0,171$.

La rappresentazione grafica in fig. 9.5 riporta simultaneamente la distribuzione empirica rappresentata dai rombi e quella teorica osservata linea continua.

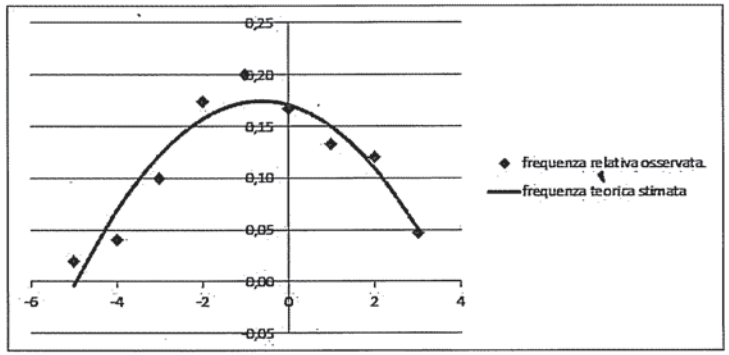


Figura 9.5: Rappresentazione grafica interpolazione

9.4 L'approccio non-parametrico

9.4.1 Stima kernel univariata

Nella *stima non parametrica* non si fa alcuna assunzione a priori sulla distribuzione da cui sono estratti i dati; questi vengono utilizzati,

attraverso tecniche inferenziali, per stimare tutta la funzione di densità $f(x)$.

Uno dei metodi non parametrici è la *stima kernel*.

Supponiamo di avere un campione o un insieme di n osservazioni reali $X_1, ..., X_n$, di cui si voglia stimare la funzione di densità. L'idea è quella di considerare una funzione K, detta appunto *kernel*, simmetrica intorno a 0 e tale che $\int K(x)dx = 1$. Spesso si utilizza come kernel una funzione di densità. La stima della densità $f(x)$ sarà una media degli n kernel, centrati in X_i $(i = 1, ..., n)$:

$$\hat{f}(x; h) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i).$$

La quantità $\frac{1}{n}K_h(x - X_i)$ è detta *bump*, perchè ha la forma di una "gobba"; in tal modo, la stima $\hat{f}(x; h)$ è data dalla somma dei bump. K_h rappresenta il kernel riscalato, la sua relazione con K è

$$K_h(x) = \frac{1}{h} K\left(\frac{x}{h}\right),$$

h è un parametro che determina la larghezza del bump, è detto *ampiezza di banda*, o *parametro di smoothing*. Così, lo stimatore $\hat{f}$, in termini di K, può essere riscritto come

$$\hat{f}(x; h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

in tal modo si evidenzia il ruolo fondamentale del parametro di smoothing.

Se, ad esempio, si utilizza come kernel la normale standard, avremo che il kernel riscalato sarà una normale a media 0 e varianza h^2. Quindi h riveste lo stesso ruolo dello scarto quadratico medio, ovvero, all'aumentare di h, aumenterà la larghezza della campana, mentre un valore piccolo di h indica una funzione kernel molto concentrata intorno al suo valore centrale e quindi una campana stretta e alta.

Per avere un'idea di come il parametro h influenzi la stima della densità $\hat{f}(\cdot; h)$, consideriamo l'esempio di $n = 9$ dati estratti da una mistura di normali, ovvero supponiamo che la funzione da stimare

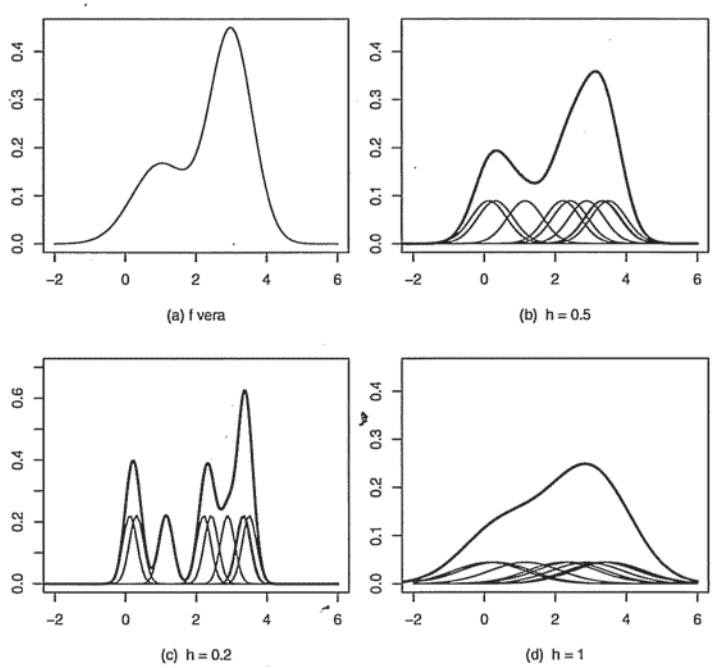


Figura 9.6: Stima di densità kernel basata su 9 osservazioni per diversi valori di h.

sia $f(x) = \frac{1}{3}\phi(1, 1) + \frac{2}{3}\phi\left(3, \frac{1}{4}\right)$, dove $\phi(\mu, \sigma^2)$ indica la densità normale con media μ e varianza σ^2. Scegliamo come kernel la normale $N(0, 1)$, e parametri di smoothing rispettivamente: (b) $h = 0.5$, (c) $h = 0.2$, (d) $h = 1$. Graficamente le stime $\hat{f}(x; h)$ che si ottengono nei tre casi si possono vedere in Figura 9.6.

Il grafico (a) rappresenta la densità vera f da stimare; le figure (b), (c), (d) rappresentano i 9 bump, rispettivamente centrati intorno a ciascun dato X_i, con $h = 0.5$ e, con una linea più spessa, la stima kernel di f.

Si nota visivamente come il valore più adeguato del parametro sia quello del caso (b), $h = 0.5$, in quanto la stima risultante meglio rappresenta la f vera.

Nel caso (c) si è utilizzato $h = 0.2$ e si nota, come conseguenza di un h troppo piccolo, la presenza di più mode, in numero eccessivo rispetto

alla bimodalità della f vera. Si parla in questo caso di *undersmoothing*.

Nel caso (d) si è utilizzato un valore di h più grande, pari a 1, e ciò porta ad un eccessivo smussamento della stima $\hat{f}(\cdot; h)$, venendo a mancare la bimodalità che invece caratterizza la funzione da stimare. Si parla in questo caso di *oversmoothing*.

Il fenomeno appare più evidente quando si utilizza un maggior numero di dati, ad esempio $n = 200$, ottenendo graficamente le $\hat{f}(\cdot; h)$ rappresentate in Figura 9.7.

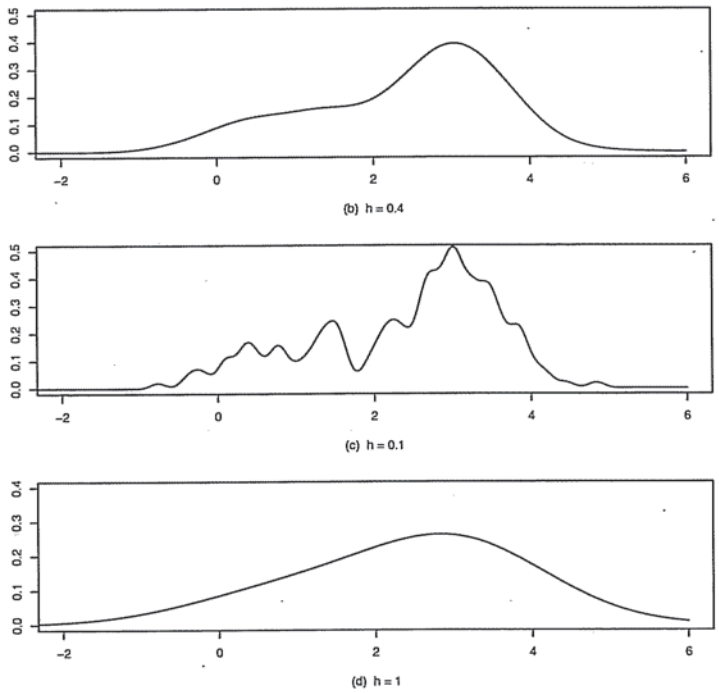


Figura 9.7: Stima di densità kernel basata per diversi valori di h.

Si capisce intuitivamente come un h troppo piccolo generi un problema di undersmoothing, caratterizzato dalla presenza di molti picchi non strutturali e che possono cambiare posizione al variare del campione di dati. Tali picchi, quindi, non caratterizzano il fenomeno reale, ma sono legati al particolare campione estratto e non sono rappresentativi del fenomeno oggetto di studio.

Al contrario, un h troppo grande provoca un problema di oversmoothing, deprivando la stima di plurimodalità eventualmente presenti nella f.

Di qui la necessità di individuare un parametro di smoothing ottimale. Abbiamo bisogno allora di un criterio per scegliere adeguatamente il parametro.

Kernel Uniforme

Prima di esaminare un criterio per la scelta di h, consideriamo un semplice esempio numerico, in cui cerchiamo di stimare una densità partendo da $n = 6$ dati e considerando come kernel la distribuzione uniforme:

$$K(x) = \begin{cases} \frac{1}{2} & -1 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti} \end{cases}$$

Si tratta di una densità simmetrica rispetto a 0 e tale che

$$\int_{-1}^{1} \frac{1}{2} dx = \frac{1}{2} x \Big|_{-1}^{1} = \frac{1}{2}(1+1) = 1.$$

La K_h riscalata sarà:

$$K_h(x) = \begin{cases} \frac{1}{2h} & X_i - h \leqslant x \leqslant X_i + h \\ 0 & \text{altrimenti} \end{cases}$$

Si veda la Figura 9.8.

Supponiamo che i valori di X siano $X = (1, 1.5, 1.6, 2, 4, 4.8)$ e scegliamo $h = 0.5$. Rappresentiamo dapprima i bump centrati in X_i, larghi 0.5 a sinistra di X_i e 0.5 a destra; essi avranno altezza pari a $\frac{1}{n} \cdot \frac{1}{2h} = \frac{1}{6 \cdot 2 \cdot 0.5} = \frac{1}{6}$. Si veda la Figura 9.9.

Il primo rettangolo è centrato in $X_1 = 1$ e vale come ordinata $\frac{1}{6}$, dal punto $x = 0.5$ al punto $x = 1.5$. Il secondo rettangolo è centrato in $X_2 = 1.5$ e vale $\frac{1}{6}$ da $x = 1$ a $x = 2$ e 0 altrove. Il terzo, centrato in $X_3 = 1.6$ vale $\frac{1}{6}$ da $x = 1.1$ a $x = 2.1$. Analogamente si costruiscono gli altri tre bump.

Operando la somma dei rettangoli, equivalente alla media dei kernel, si ottiene la stima finale, rappresentata in Figura 9.10. Da $x = 0.5$ a $x = 1$ si incontra un solo rettangolo, quindi la somma vale $\frac{1}{6}$;

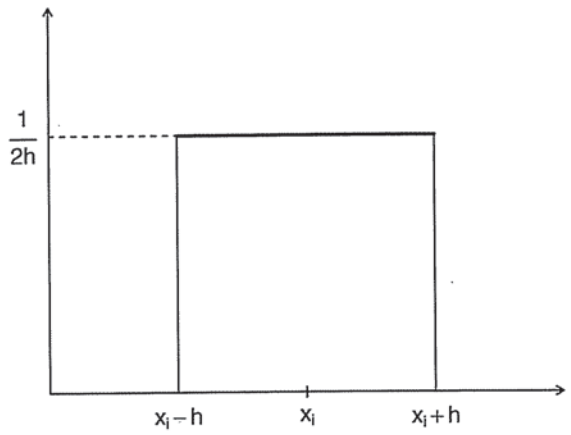


Figura 9.8: Kernel uniforme nell'intervallo $X_i - h, X_i + h$.

tra $x = 1$ e $x = 1.1$ si incontrano due rettangoli, quindi la somma vale $\frac{2}{6}$; tra $x = 1.1$ e $x = 2$ si incontrano tre rettangoli, quindi la somma vale $\frac{3}{6}$ e così via.

Il risultato è una funzione a gradini, derivante dal tipo di kernel "squadrato" che abbiamo utilizzato.

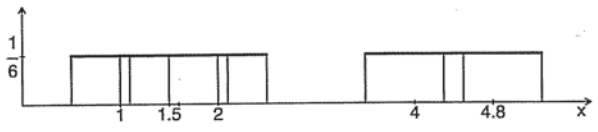


Figura 9.9: Bump uniformi centrati in X_i.

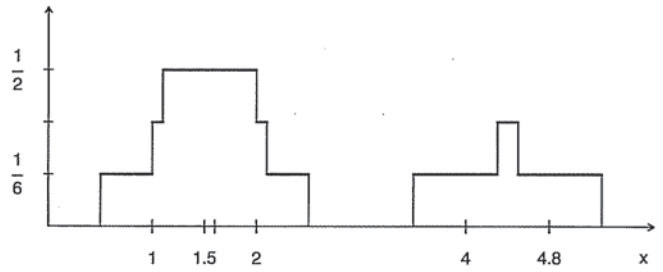


Figura 9.10: Stima $\hat{f}(x; h)$, media dei 6 kernel uniformi.

Kernel triangolare

Proviamo a ripetere l'esercizio utilizzando un altro tipo di kernel, quello triangolare:

$$K(x) = \begin{cases} 1 - |x| & -1 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti} \end{cases}$$

e quindi il bump che ne deriva sarà

$$(nh)^{-1} K\left(\frac{x - X_i}{h}\right) = \begin{cases} \frac{1}{nh}\left(1 - \left|\frac{x - X_i}{h}\right|\right) & X_i - h \leqslant x \leqslant X_i + h \\ 0 & \text{altrimenti} \end{cases}$$

Il grafico è riportato in Figura 9.11.

Scegliamo gli stessi valori $X = (1, 1.5, 1.6, 2, 4, 4.8)$ e lo stesso valore $h = 0.5$.

Nella Figura 9.12 sono rappresentati i bump; si tratta di triangoli isosceli, centrati nei punti X_i, base pari a 1 e altezza pari a $\frac{1}{nh} = \frac{1}{6 \cdot 0.5} = \frac{1}{3}$.

Calcolare la somma in questo caso è più complicato in quanto, per ogni $x \in \mathbb{R}$, dovremmo vedere quali bump sono diversi da zero, calcolare il loro valore e sommarli. Ad esempio, se guardiamo gli ultimi due triangoli sulla destra, riportati anche in Figura 9.13, da $x = 3.5$ a $x = 4.3$ e poi da $x = 4.5$ a $x = 5.3$ si incontra un solo triangolo, quindi la somma coincide col bump stesso; mentre tra $x = 4.3$ e

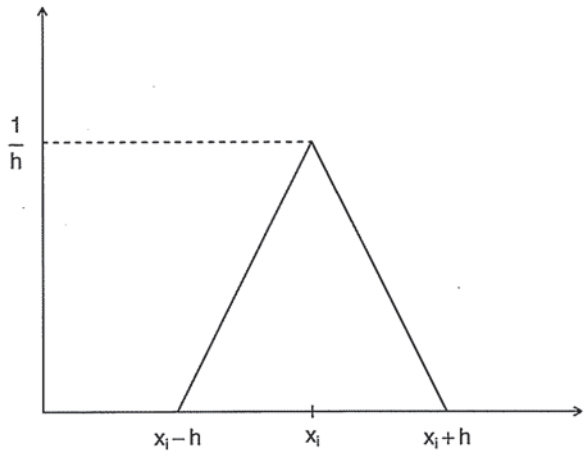


Figura 9.11: Kernel triangolare centrato in X_i.

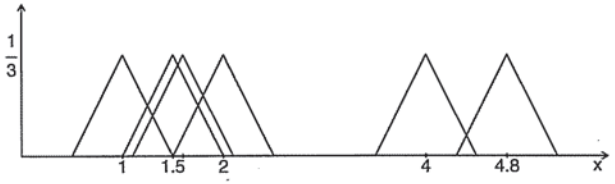


Figura 9.12: Bump triangolari centrati in X_i.

$x = 4.5$ si incontrano due bump, quindi la stima $\hat{f}(x; h)$ in tale intervallo sarà data dalla somma dei due bump; notiamo che, mentre un lato scende, l'altro sale con la stessa inclinazione, quindi in tale intervallo la somma rimane costante, pari a

$$\hat{f}(4.3; h = 0.5) = \frac{1}{6 \cdot 0.5}\left(1 - \left|\frac{4.3 - 4}{0.5}\right|\right) +$$

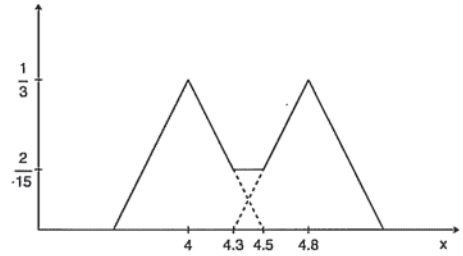


Figura 9.13: Somma dei due triangoli più a destra.

$$+\frac{1}{6 \cdot 0.5}\left(1 - \left|\frac{4.3 - 4.8}{0.5}\right|\right) = 0.1\overline{3}$$

(Verificare che anche in $x = 4.4$ si ottiene lo stesso valore della stima).

Analogamente si procede per l'intervallo tra $x = 0.5$ e $x = 2.5$, sommando i bump che si incontrano per ciascuna x. Ad esempio, per $x = 1.6$, la stima risulta:

$$\hat{f}(1.6; h = 0.5) = \frac{1}{6 \cdot 0.5}\left(1 - \left|\frac{1.6 - 1.5}{0.5}\right|\right) + \frac{1}{6 \cdot 0.5}\left(1 - \left|\frac{1.6 - 1.6}{0.5}\right|\right) +$$

$$+\frac{1}{6 \cdot 0.5}\left(1 - \left|\frac{1.6 - 2}{0.5}\right|\right) = 0.2\overline{3} + 0 + 0.0\overline{6} = 0.\overline{3}$$

La stima risultante è rappresentata in Figura 9.14. Il risultato ottenuto in tale figura è alquanto diverso dal risultato di Figura 9.10, capiamo così come entrambi i kernel utilizzati siano insoddisfacenti.

Si comprende da questi due esempi come le cose si complichino sia all'aumentare del numero dei dati e sia scegliendo kernel più smussati dei due esaminati negli esempi. Si capisce anche come sia indispensabile l'uso di uno strumento informatico per il calcolo delle densità e delle loro medie.

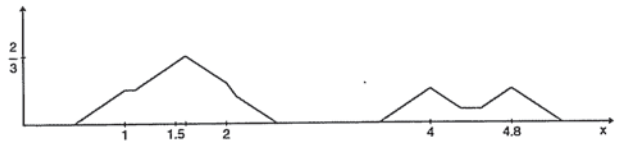


Figura 9.14: Stima con kernel triangolare.

Kernel uniforme con h maggiore

Riprendiamo il primo esempio del kernel uniforme e ripetiamo l'esercizio usando un parametro di smoothing maggiore, diciamo $h = 1$. L'altezza del rettangolo sarà pari a $\frac{1}{2h} = \frac{1}{2}$.

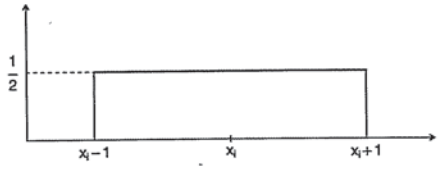


Figura 9.15: Kernel uniforme con $h = 1$.

Il kernel è rappresentato in Figura 9.15. Ripetendo tutti i calcoli precedenti aggirnati col nuovo h, si ottiene la $\hat{f}(x; h)$ rappresentata in Figura 9.16.

Confrontando questo risultato con la Figura 9.10, si nota un maggiore smussamento della stima $\hat{f}(x; h)$; in questo senso la stima kernel appare troppo appiattita e ci rendiamo conto che non si può stabilire un valore di h a priori. Bisogna, anzi, avere un criterio per individuare la h più adeguata.

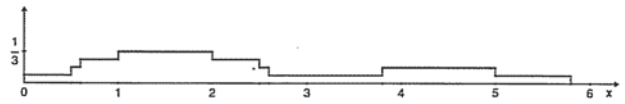


Figura 9.16: Media dei 6 kernel uniformi con $h = 1$.

Kernel di Epanechnikov

Come ultimo esempio, consideriamo il kernel di Epanechnikov, il quale risulta uno dei kernel più efficienti:

$$K(x) = \begin{cases} \frac{3}{4}(1-x^2) & -1 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti} \end{cases}$$

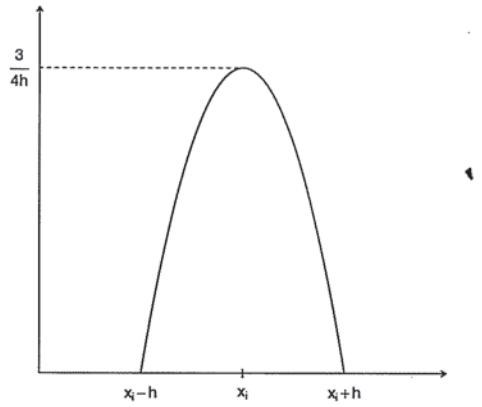


Figura 9.17: Kernel di Epanechnikov.

Si tratta di una parabola con la concavità verso il basso, vertice in $\left(0, \frac{3}{4}\right)$, ma diversa da 0 solo tra -1 e 1. Il grafico è riportato in Figura 9.17.

Scegliamo gli stessi valori $X = (1, 1.5, 1.6, 2, 4, 4.8)$ degli esempi precedenti. Scegliamo $h = 0.5$ e calcoliamo il valore della stima $\hat{f}(x; h = 0.5)$ in alcuni punti, ad esempio in $x = 0.8$, $x = 1.2$, $x = 2.5$, $x = 4.4$.

Riscaliamo il kernel e calcoliamo l'espressione dei bump per x e X_i generici:

$$n^{-1}K_h(x - X_i) = \frac{1}{nh}K\left(\frac{x - X_i}{h}\right) = \begin{cases} \frac{1}{6 \cdot 0.5}\frac{3}{4}\left[1 - \left(\frac{x - X_i}{0.5}\right)^2\right] & X_i - 0.5 \leqslant x \leqslant X_i + 0.5 \\ 0 & \text{altrimenti} \end{cases}$$

Per $x = 0.8$:
dobbiamo calcolare i 6 bump, uno per ciascun X_i, e poi sommarli. Notiamo che $x + h = 0.8 + 0.5 = 1.3$, quindi solo il primo kernel è diverso da 0 in $x = 0.8$, il kernel successivo è centrato in $X_2 = 1.5$ ed è diverso da 0 tra 1 e 2, quindi in 0.8 è nullo, a maggior ragione saranno nulli i successivi kernel.

$$X_1 = 1 \quad : \quad n^{-1}K_h(0.8 - 1) = \frac{1}{4}\left[1 - \left(\frac{0.8 - 1}{0.5}\right)^2\right] = 0.21,$$
$$X_i, i = 2, \ldots, 6 \quad : \quad n^{-1}K_h(0.8 - X_i) = 0,$$
$$\Rightarrow \quad \hat{f}(0.8; h = 0.5) = 0.21.$$

Calcoliamo ora la stima per $x = 1.2$:
$x - h = 1.2 - 0.5 = 0.7$, $x + h = 1.7$, rientrano in questo intervallo i primi tre kernel, gli altri saranno nulli (per intenderci, quello centrato in $X_4 = 2$ sarà diverso da 0 solo in $[2 - 0.5, 2 + 0.5] = [1.5, 2.5]$, quindi in $x = 1.2$ esso è nullo, a maggior ragione lo saranno i kernel successivi).

$$X_1 = 1 \quad : \quad n^{-1}K_h(1.2 - 1) = \frac{1}{4}\left[1 - \left(\frac{1.2 - 1}{0.5}\right)^2\right] = 0.21,$$
$$X_2 = 1.5 \quad : \quad n^{-1}K_h(1.2 - 1.5) = \frac{1}{4}\left[1 - \left(\frac{1.2 - 1.5}{0.5}\right)^2\right] = 0.16,$$
$$X_3 = 1.6 \quad : \quad n^{-1}K_h(1.2 - 1.6) = \frac{1}{4}\left[1 - \left(\frac{1.2 - 1.6}{0.5}\right)^2\right] = 0.1875,$$
$$X_i, i = 4, \ldots, 6 \quad : \quad n^{-1}K_h(1.2 - X_i) = 0,$$
$$\Rightarrow \quad \hat{f}(1.2; h = 0.5) \triangleq 0.21 + 0.16 + 0.1875 = 0.5575.$$

Per $x = 2.5$:
$x - h = 2.5 - 0.5 = 2$, $x + h = 3$, in tale intervallo è non nullo solo il quarto kernel:

$$X_4 = 2 \quad : \quad n^{-1}K_h(2.5 - 2) = \frac{1}{4}\left[1 - \left(\frac{2.5 - 2}{0.5}\right)^2\right] = 0,$$
$$X_i, i = 1, \ldots, 6, i \neq 4 \quad : \quad n^{-1}K_h(2.5 - X_i) = 0,$$
$$\Rightarrow \quad \hat{f}(2.5; h = 0.5) = 0.$$

Infine, per $x = 4.4$:
$x - h = 4.4 - 0.5 = 3.9$, $x + h = 4.9$, rientrano nell'intervallo gli ultimi due kernel:

$$X_5 = 4 \quad : \quad n^{-1}K_h(4.4 - 4) = \frac{1}{4}\left[1 - \left(\frac{4.4 - 4}{0.5}\right)^2\right] = 0.09,$$
$$X_6 = 4.8 \quad : \quad n^{-1}K_h(4.4 - 4.8) = \frac{1}{4}\left[1 - \left(\frac{4.4 - 4.8}{0.5}\right)^2\right] = 0.09,$$
$$X_i, i = 1, \ldots, 4 \quad : \quad n^{-1}K_h(4.4 - X_i) = 0,$$
$$\Rightarrow \quad \hat{f}(4.4; h = 0.5) = 0.18.$$

Abbiamo calcolato il valore della stima $\hat{f}(x : h)$ solo in cinque punti dell'asse reale. Se volessimo rappresentare graficamente la stima di

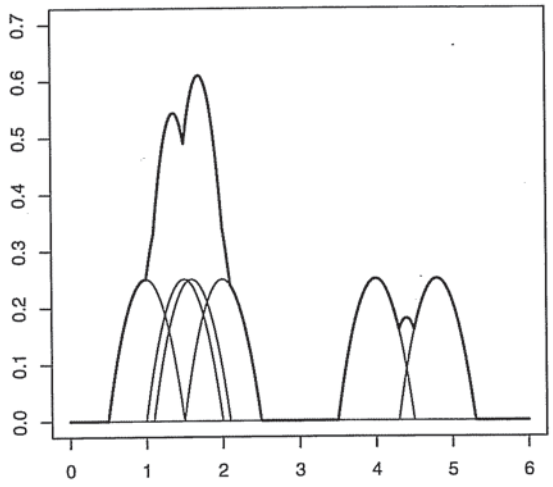


Figura 9.18: Stima con kernel di Epanechnikov.

f dovremmo calcolarci il suo valore in un numero ben maggiore di punti. Capiamo così come nel problema della stima kernel si renda necessario l'utilizzo di un pacchetto software che calcoli tale stima per noi. Nel caso del kernel di Epanechnikov, il grafico risultante per la stima è rappresentato in Figura 9.18.

Per esercizio, il lettore può rifare i calcoli negli stessi valori di x provando a modificare il valore di h, oppure può calcolare la stima in diversi valori di x.

9.4.2 Stima kernel multivariata

Nel caso di stima kernel di una densità multidimensionale, le cose si complicano in quanto non bisogna specificare un solo parametro di smoothing h, ma tutta una matrice di ampiezza di banda **H**. Inoltre, la scarsità dei dati in uno spazio multidimensionale rende difficoltosa la stima.

Supponiamo di avere un campione o un insieme di osservazioni d-variato $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n$ con funzione di densità f. Gli $\mathbf{X}_i$ scritti in grassetto sono vettori di dimensione d, $\mathbf{X}_i = (X_{i1}, X_{i2}, \cdots, X_{id})$.

In generale, **H** è una matrice simmetrica completa, con $\frac{d(d+1)}{2}$ elementi indipendenti, il che comporta, anche per piccole dimensioni d, un numero elevato di parametri da stimare; ma tale **H** è quella che ci dà maggiore flessibilità e quindi maggiore probabilità di avvicinarci alla funzione vera f da stimare.

Tuttavia è possibile avere due semplificazioni per **H**, entrambe hanno gli elementi fuori della diagonale uguali a zero, nella prima semplificazione nella diagonale principale abbiamo d valori di h_i, $i = 1, ..., d$, nella seconda semplificazione gli elementi della diagonale sono tutti uguali, abbiamo così un unico parametro h da stimare. È chiaro che l'ultima semplificazione, anche se rende più agevoli le stime, è un po' troppo restrittiva per poter ottenere una stima adeguata di f.

Nel caso più generale, diremo che $\mathbf{H} \in \mathcal{F}$:

$$\mathbf{H} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1d} \\ h_{12} & h_{22} & \cdots & h_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ h_{1d} & h_{2d} & \cdots & h_{dd} \end{pmatrix}$$

nel secondo caso diremo che $\mathbf{H} \in \mathcal{D}$:

$$\mathbf{H} = \begin{pmatrix} h_1^2 & 0 & \cdots & 0 \\ 0 & h_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_d^2 \end{pmatrix}$$

nel caso più semplice diremo che $\mathbf{H} \in \mathcal{S}$:

$$\mathbf{H} = \begin{pmatrix} h^2 & 0 & \cdots & 0 \\ 0 & h^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h^2 \end{pmatrix}$$

Nel caso più generale, lo stimatore kernel di densità è dato da

$$\hat{f}(\mathbf{x}; \mathbf{H}) = n^{-1} \sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{x} - \mathbf{X}_i),$$

dove

$$K_H(\mathbf{x}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{x})$$

e K è una funzione kernel d-variata che soddisfa la proprietà $\int K(\mathbf{x})d\mathbf{x} = 1$;
nel caso $\mathbf{H} \in \mathcal{D}$, avremo

$$\hat{f}(\mathbf{x};\mathbf{h}) = n^{-1}\left(\prod_{l=1}^{d} h_l\right)^{-1} \sum_{i=1}^{n} K\left(\frac{x_1 - X_{i1}}{h_1}, \cdots, \frac{x_d - X_{id}}{h_d}\right),$$

infine, nel caso più semplice in cui $\mathbf{H} \in \mathcal{S}$, si ha

$$\hat{f}(\mathbf{x};h) = n^{-1}h^{-d} \sum_{i=1}^{n} K\left(\frac{\mathbf{x} - \mathbf{X}_i}{h}\right).$$

Limitandoci al caso più semplice, con un unico parametro di smoothing, mostriamo una rappresentazione grafica di una stima di densità per $n = 200$ osservazioni estratte da una mistura di normali bivariate e vediamo come, al variare di h, si modifichi il risultato per la $\hat{f}(\mathbf{x};h)$. Si vedano le Figure 9.19-9.22.

Osservando la f vera e le tre stime successive appare evidente come il valore di $h = 0.4$ sia troppo piccolo e generi delle "gobbe" indesiderate, in numero maggiore rispetto alla bimodalità della funzione vera; il valore $h = 0.6$ sembra il più adeguato, mentre $h = 1$ sembra un valore che smussa troppo la stima, rendendola troppo liscia e con una sola moda apparente.

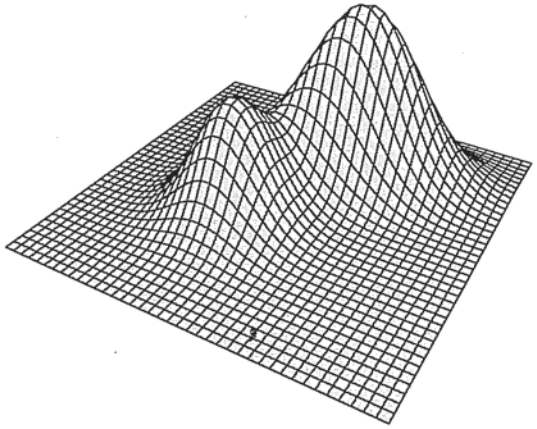


Figura 9.19: Mistura di normali bivariate, funzione f vera da stimare.

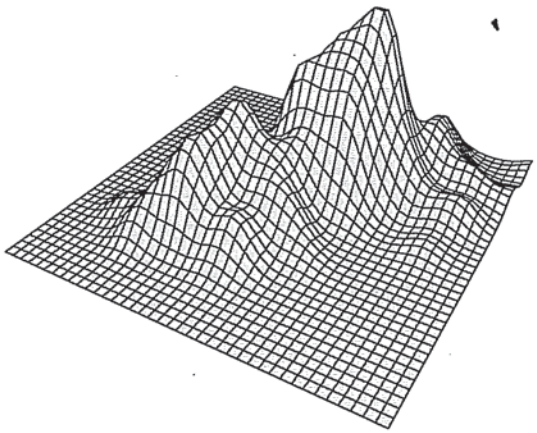


Figura 9.20: Stima di densità, con h=0.4

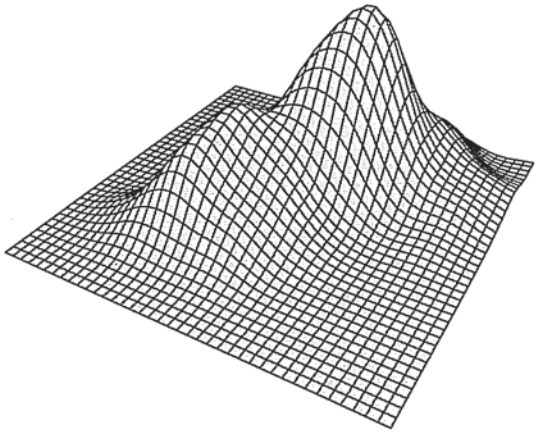


Figura 9.21: Stima di densità, con h=0.6

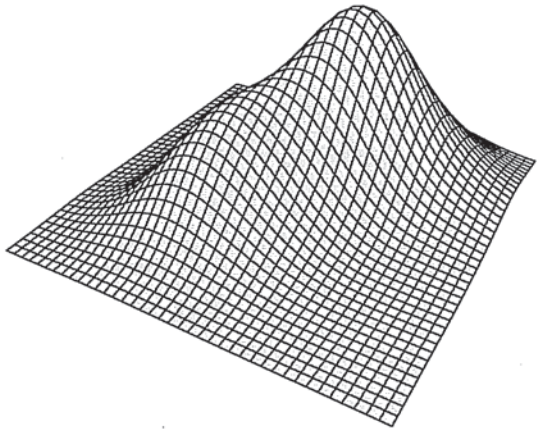


Figura 9.22: Stima di densità, con h=1

Ovviamente questi erano dati simulati, in cui conoscevamo la funzione f da stimare, nella realtà questo è proprio l'obiettivo dell'analisi non parametrica, ovvero stimare una f non nota. Anche nel caso multivariato si ha, quindi, la necessità di stabilire un criterio per la scelta ottimale di h, o in generale, di **H**.

Capitolo 10

Analisi delle relazioni tra due variabili

Quando ci esprimiamo in termini di *statistica multivariata* ci riferiamo allo studio delle relazioni esistenti tra più caratteri osservati rispetto ad uno stesso fenomeno reale. L'obiettivo, nello specifico, è quello di studiare l'esistenza di relazioni di *dipendenza* o di *associazione* tra diversi caratteri presi in esame su uno stesoo fenomeno reale, nonché di descrivere ed interpretare tali relazioni e di effettuare previsioni. Per analizzare le relazioni causa-effetto non è possibile limitarsi allo studio delle singole distribuzioni di frequenza ma sarà necessario analizzare contemporaneamente due o più caratteristiche osservate di un fenomeno reale. In generale ci si esprime sempre in termini *distribuzione multivariata* quando si prendono in esame contemporaneamente più variabili, anche se possiamo parlare anche *distribuzione bivariata* nel caso di due caratteri, *distribuzione tripla* nel caso di tre caratteri e così via. In questo capitolo approfondiremo l'analisi della statistica bivariata ed affronteremo più nel dettaglio l'aspetto multivariato nel capitolo seguente. Nel corso della trattazione prenderemo in considerazione, quindi, due caratteri che indicheremo rispettivamente con X e con Y.

Come accadeva in ambito univariato, anche nel caso di una distribuzione bivariata le metodologie di analisi si differenzieranno in relazione alla specifica natura dei caratteri che prenderemo in esame; si può

presentare, infatti, il caso in cui entrambe i caratteri siano di tipo qualitativo o di tipo quantitativo, o ancora siano uno qualitativo e uno quantitativo.

Più nello specifico possiamo dire che abbiamo una *distribuzione bivariata quantitativa* se i due caratteri presi in considerazione sono tutti e due misurati su una scala ad intervalli, sia nel caso che siano entrambi quantitativi discreti (numero di stanze di una abitazione e numero di figlio), quantitativi continui (reddito e spese mensili di un collettivo di famiglie) o l'uno discreto e l'altro continuo (numero di stanze di un'abitazione e reddito). Abbiamo una *distribuzione bivariata qualitativa* quando entrambi i caratteri sono di tipo qualitativo; anche in questo caso possiamo trovarci di fronte a diverse combinazioni: tutti e due i caratteri sono qualitativi sconnessi, quindi misurati su scala nominale (ad esempio, colore degli occhi e colore dei capelli), sono tutti e due misurati su scala ordinabile (titolo di studio e la professione) e, infine, sono uno sconnesso ed uno ordinabile (religione e titolo di studio). Si può verificare ancora che i due caratteri presi in considerazione siano uno di tipo qualitativo (sconnesso o ordinabile) e uno di tipo quantitativo(discreto o continuo), in questo caso ci riferiamo ad una *distribuzione bivariata mista* (ad esempio, il titolo di studio e il reddito).

10.1 Distribuzione doppia di frequenza

Prendiamo in considerazione la tabella 10.1, relativa ad una distribuzione unitaria di un collettivo di N individui.

Se volessimo studiare la relazione tra due caratteri, ad esempio il titolo di studio e la professione, avremmo a che fare con una *variabile doppia* (X, Y) sulle N unità statistiche, ossia per ogni unità statistica avremo una coppia ordinata di dati che andranno ad identificare la successione seguente:

$$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots, (x_N, y_N)$$

Nel nostro esempio specifico, quindi, la coppia (x_1, y_1) individuerà l'unità statistica 1 che sarà "Laureato ed Impiegato", la coppia (x_2, y_2) individuerà l'unità statistica 2 che sarà "Laureato ed Operaio", e così di seguito sino ad arrivare all'unità statistica N.

UNITÀ	CARATTERI					
Indiv.	**Sex**	**Età**	**Professione**	**...**	**Titolo di Studio**	**Reddito Famiglia (€)**
1	*M*	*26*	*Impiegato*	...	*Laurea*	*25.000*
2	*F*	*30*	*Operaio*	...	*Laurea*	*55.000*
3	*M*	*27*	*Studente*	...	*Diploma*	*85.000*
4	*M*	*18*	*Studente*	...	*Licenza Media*	*40.000*
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n_i	*F*	*18*	*Operaio*	...	*Licenza Media*	*18.000*
⋮	⋮	⋮	⋮	⋮	⋮	⋮
n_N	*M*	*24*	*Studente*	...	*Diploma*	*45.000*

Tabella 10.1: Distribuzione unitaria.

La matrice dei dati può essere convenzionalmente rappresentata come indicato nella tabella 10.2: In questo caso su ogni unità statistica

UNITÀ	(X, Y)
1	(x_1, y_1)
2	(x_2, y_2)
⋮	⋮
i	(x_i, y_i)
⋮	⋮
N	(x_N, y_N)

Tabella 10.2: Distribuzione doppia unitaria.

rileviamo, quindi, una coppia di modalità, quella di X e quella di Y. Abbiamo, quindi, un campione o un insieme di N coppie ordinate di valori $\{(x_i; y_i); i = 1, \ldots, N\}$ dove nell'$i - esima$ coppia le due coordinate rappresentano rispettivamente le modalità di X e Y assunte dall'unita statistica i.

Solitamente quando si conduce lo studio di un fenomeno reale abbiamo a che fare con un elevato numero di unità statistiche, pertanto risulta opportuno organizzare i nostri dati in una *tabella a doppia entrata*, che rappresenta appunto la nostra matrice dei dati (cfr. 10.3).

X, Y	y_1	y_2	$\dots$	y_j	$\dots$	y_h	**TOTALI riga**
x_1	n_{11}	n_{12}	$\dots$	n_{1j}	$\dots$	n_{1h}	$n_{1.}$
x_2	x_{21}	x_{22}	$\dots$	x_{2j}	$\dots$	x_{2h}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	x_{i1}	x_{i2}	$\dots$	x_{ij}	$\dots$	x_{ih}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	x_{k1}	x_{k2}	$\dots$	x_{kj}	$\dots$	x_{kh}	$n_{k.}$
TOTALI colonna	$x_{.1}$	$x_{.2}$	$\dots$	$x_{.j}$	$\dots$	$x_{.h}$	N

Tabella 10.3: La matrice dei dati di una variabile doppia.

Le due variabili considerate vengono indicate con X e Y; le modalità con cui si esprimono vengono indicate rispettivamente con x_i, dove $(i = 1, 2, \dots, k)$ e y_j, dove $(j = 1, 2, \dots, k)$; ad ogni coppia (x_i, y_j) si fa corrispondere nella tabella la sua frequenza assoluta associata n_{ij}, cioè il numero di elementi, tra gli n della popolazione, che possiedono contemporaneamente la modalità x_i di X e y_j di Y.

$$n_{i.} = \sum_{j=1}^{k} n_{i,j}$$

dove $(i = 1, 2, \dots, h)$ rappresenta le frequenze marginali assolute di X

$$n_{.j} = \sum_{i=1}^{h} n_{i,j}$$

dove $(j = 1, 2, \dots, k)$ rappresenta le frequenze marginali assolute di Y.

Ovviamente, sommando tutte le frequenze assolute presenti nella tabella, troveremo la numerosità N della popolazione.

L'approccio della statistica multivariata, quindi, utilizza un insieme di metodologie utili per analizzare congiuntamente due o più fenomeni di interesse, osservati su di un collettivo di unità statistiche.

Gli obiettivi fondanti possono essere riassunti nei seguenti due punti:

- studiare eventuali relazioni di *dipendenza* o *associazione* tra diversi fenomeni osservati;
- descrivere ed interpretare le relazioni tra fenomeni ed effettuare previsioni.

Come abbiamo fatto nell'approccio di tipo univariato, differenzieremo le metodologie da utilizzare in base alla natura dei caratteri che prenderemo in analisi. In particolar modo analizzeremo i seguenti casi:

- dipendenza o associazione tra due caratteri qualitativi;
- dipendenza o associazione tra due caratteri quantitativi;
- dipendenza o associazione tra due caratteri di cui uno qualitativo e uno quantitativo.

10.2 Tabella di contingenza: distribuzioni congiunte, marginali e condizionate

Per aiutarci nella comprensione dell'argomento partiamo dall'analisi di alcuni dati.

Consideriamo un'azienda che abbia 1256 dipendenti, dei quali possediamo diverse informazioni riguardanti, ad esempio, lo stato civile (coniugato, non coniugato), il mezzo di trasporto utilizzato per raggiungere la sede di lavoro (auto, autobus, treno), il sesso, l'età, ecc. Se prendiamo in considerazione, ad esempio, i caratteri "stato civile" e "mezzo di trasporto", otteniamo una *distribuzione bivariata* (10.4).

Dipendente	Stato Civile	Mezzo di trasporto
1	non coniugato	auto
2	coniugato	auto
3	non coniugato	treno
⋮	⋮	⋮
1256	coniugato	autobus

Tabella 10.4: Esempio distribuzione bivariata

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	108	168	469
non coniugato	112	157	518	787
Totale	305	265	686	1256

Tabella 10.5: Esempio Tabella di contingenza: frequenze assolute

Al fine di leggere meglio i dati possiamo sintetizzarli costruendo la *tabella a doppia entrata*, chiamata anche *tabella di contingenza* (tab. 10.5).

Da questa tabella, di più immediata ed intuitiva lettura, possiamo desumere, ad esempio, che 193 dipendenti coniugati viaggiano abitualmente con la propria auto; 518 non sono coniugati e solitamente viaggiano in autobus; in totale 469 dipendenti dell'azienda sono coniugati, 265 dipendenti viaggiano in treno e così via. Una tabella di contingenza, quindi, contiene numerose informazioni:

- la parte centrale della tabella contiene la *distribuzione di frequenza assoluta congiunta delle variabili*, come evidenziato nella tabella 10.6;

- l'ultima riga e l'ultima colonna della tabella contengono le *distribuzioni di frequenze assolute marginali* delle due variabili, dove ciascun totale rappresenta la marginale di una variabile.

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	**193**	**108**	**168**	469
non coniugato	**112**	**157**	**518**	787
Totale	305	265	686	1256

Tabella 10.6: Esempio distribuzione di frequenza assoluta congiunta

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	108	168	**469**
non coniugato	112	157	518	**787**
Totale	305	265	686	**1256**

Tabella 10.7: Esempio distribuzione margimale di riga

Dobbiamo precisare che la distribuzione marginale della variabile "Stato civile" non fornisce alcuna informazione circa la variabile "Mezzo di trasporto", ossia tratta la condizione di stato civile indipendentemente dal mezzo che utilizzano gli impiegati per spostarsi (Tab. 10.7). Nella tabella 10.8, invece, è evidenziata la distribuzione margi-

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	108	168	469
non coniugato	112	157	518	787
Totale	**305**	**265**	**686**	**1256**

Tabella 10.8: Esempio distribuzione margimale di colonna

nale della variabile "Mezzo di trasporto", che non ci dà informazioni della variabile "Stato civile".

- Se, invece, prendiamo in considerazione una sola riga o una sola colonna ci riferiamo alla *distribuzione di frequenza assoluta di una*

variabile condizionata ad una modalità specifica dell'altra variabile, come illustrato ad esempio nelle tabelle 10.9 e 10.10.

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	**193**	**108**	**168**	469
non coniugato	112	157	518	787
Totale	305	265	686	1256

Tabella 10.9: Esempio 1 di distribuzione di una variabile condizionata ad un'altra

In particolare, nella distribuzione della variabile "Mezzo di trasporto" condizionata alla modalità a *Coniugato* della variabile "Stato civile", osserviamo la distribuzione di frequenze assolute del "Mezzo di trasporto" limitando l'attenzione ai soli individui coniugati 10.9. Attraverso il condizionamento, quindi, non consideriamo più l'intero collettivo (1256), ma restringiamo l'analisi al sottogruppo di unità che assumono una determinata modalità a della variabile per cui condizioniamo (469).

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	**108**	168	469
non coniugato	112	**157**	518	787
Totale	305	265	686	1256

Tabella 10.10: Esempio 2 di distribuzione di una variabile condizionata ad un'altra

Nella tabella 10.10, invece, è rappresentata la distribuzione della variabile "Stato civile" condizionata alla modalità a *Treno* della variabile "Mezzo di trasporto". Consideriamo, quindi, lo stato civile solo tra coloro che viaggiavano in treno, restringendo pertanto il campo di analisi a 265 dipendenti.

La *distribuzione congiunta è bivariata*poichè prende in considerazione due variabili contemporaneamente. Le *distribuzioni marginali* e le

condizionate sono invece *distribuzioni univariate*, poichè prendono in considerazione una sola variabile.

Distribuzioni di frequenza relative

Le tabelle di contingenza viste fino ad ora, contengono le frequenze assolute delle distribuzioni prese in esame. Possiamo ancora costruire per una tabella di contingenza la distribuzione delle frequenze relative. Continuando ad utilizzare il nostro esempio avremo una rappresentazione come illustrato in tabella 10.11.

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	0,15	0,09	0,13	0,37
celibe/nubile	0,09	0,13	0,41	0,63
Totale	0,24	0,21	0,55	1

Tabella 10.11: Esempio Tabella di contingenza: frequenze relative

Le *frequenze relative marginali di* X si indicano genericamente come segue: $f_{1\cdot}, \ldots, f_{i\cdot}, \ldots, f_{r\cdot}$. Mentre le *frequenze relative marginali di* Y con $f_{\cdot 1}, f_{\cdot 2}, \ldots, f_{\cdot j}, \ldots, f_{\cdot s}$.

Per calcolare la *frequenza relativa congiunta della coppia* (x_i, y_j) si opera come segue:

$$f_{ij} = \frac{n_{ij}}{N}$$

Per calcolare la *frequenza relativa marginale di* x_i

$$f_{i\cdot} = \frac{n_{i\cdot}}{N} = \sum_{j=1}^{s} f_{ij}$$

ed, infine, per individuare la *frequenza relativa marginale di* y_i

$$f_{\cdot j} = \frac{n_{\cdot j}}{N} = \sum_{i=1}^{r} f_{ij}$$

Distribuzione di X condizionata alla modalità y_i di Y

Consideriamo adesso la distribuzione di X condizionata alla modalità y_j di Y (colonna j–esima della tabella). La distribuzione univariata si denota convenzionalmente con $X|Y = y_j$ o $X|y_j$

X|Y	freq. ass.	freq. rel.
x_1	n_{1j}	$\frac{n_{1j}}{n_{.j}}$
x_2	n_{2j}	$\frac{n_{2j}}{n_{.j}}$
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	$\frac{n_{ij}}{n_{.j}}$
$\vdots$	$\vdots$	$\vdots$
x_r	n_{rj}	$\frac{n_{rj}}{n_{.j}}$
Totale	$n_{.j}$	1

profili colonna: la tabella a doppia entrata contiene s diverse distribuzioni condizionate di X, una per ciascuna modalità di Y.

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	108	168	469
non coniugato	112	157	518	787
Totale	305	265	686	1256

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	0,63	0,41	0,24	
non coniugato	0,37	0,59	0,76	
Totale	1	1	1	

10.2.1 Distribuzione di Y condizionata alla modalità x_i di X

Consideriamo adesso la distribuzione di Y condizionata alla modalità x_i di X (riga i–esima della tabella). La distribuzione univariata: si denota con $Y|X = x_i$ o $Y|x_i$

Y|X	freq. ass.
y_1	n_{i1}
y_2	n_{i2}
$\vdots$	$\vdots$
y_i	n_{ij}
$\vdots$	$\vdots$
y_s	n_{is}
Totale	$n_{.j}$

$Y\|x_i$	freq. rel.
y_1	$\frac{n_{i1}}{n_{i.}}$
y_2	$\frac{n_{i2}}{n_{.i}}$
$\vdots$	$\vdots$
y_i	$\frac{n_{ij}}{n_{.i}}$
$\vdots$	$\vdots$
y_s	$\frac{n_{is}}{n_{.i}}$
Totale	1

profili riga: la tabella a doppia entrata contiene r diverse distribuzioni condizionate di Y (una per ciascuna modalità di X).

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	108	168	469
non coniugato	112	157	518	787

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	0,41	0,23	0,36	1
non coniugato	0,14	0,20	0,66	1

Occorre prestare molta attenzione a non confondere le frequenze relative congiunte con le frequenze relative condizionate

$$\frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} \quad e \quad \frac{n_{ij}}{n_{i.}} = \frac{f_{ij}}{f_{j.}}$$

Esempio

$$0,66 = \frac{518}{787} = \%(\text{Autobus}|\text{Nonconiugato})$$

mentre

$$0,76 = \frac{518}{686} = \%(\text{Nonconiugato}|\text{Autobus})$$

10.2.2 Medie e varianze marginali e condizionate

La distribuzione marginale e le distribuzioni condizionate, come abbiamo anticipato, sono tutte univariate, di conseguenza nel caso in cui X (o analogamente Y) sia una variabile quantitativa possiamo calcolare medie e varianze marginali e condizionate della nostra distribuzione bivariata.

Per calcolare le medie e le varianze marginali sarà rispettivamente:

$$\mu(X) = \frac{1}{N}\sum_{i=1}^{r} x_i \cdot n_i = \sum_{i=1}^{r} x_i \cdot f_i$$

$$\sigma^2(X) = \frac{1}{N}\sum_{i=1}^{r} x_i^2 \cdot n_i - [\mu(X)]^2$$

Se, invece, vogliamo calcolare medie e varianze condizionate sarà:

$$\mu(X|Y = y_j) = \frac{1}{n_{\cdot j}}\sum_{i=1}^{r} x_i \cdot n_{ij}$$

$$\sigma^2(X|Y = y_j) = \frac{1}{n_{\cdot j}}\sum_{i=1}^{r} x_i^2 \cdot n_{ij} - [\mu(X|Y = y_j)]^2$$

Per rendere più comprensibile quanto detto procediamo con un semplice esempio relativo al reddito mensile, espresso in migliaia di euro, di un collettivo di individui distinto per sesso:

[H]

Sesso	Reddito 1	2	3	Totale
M	14	12	14	40
F	11	18	11	40
Totale	25	30	25	80

[H]

Stato civile	Mezzo di trasporto Auto	Treno	Autobus	freq. rel. marg. di Stat.civ.
Coniugato	0,63	0,41	0,24	0,37
Non coniugato	0,37	0,59	0,76	0,63
Totale	1	1	1	

- $\mu(\text{Reddito}) = (1 \cdot 25 + 2 \cdot 30 + 3 \cdot 25)/80 = 2$
- $\sigma^2(\text{Reddito}) = (1^2 \cdot 25 + 2^2 \cdot 30 + 3^2 \cdot 25)/80 - 2^2 = 0,63$
- $\mu(\text{Reddito}|\text{Sesso} = M) = (1 \cdot 14 + 2 \cdot 12 + 3 \cdot 14)/40 = 2$
- $\mu(\text{Reddito}|\text{Sesso} = F) = (1 \cdot 11 + 2 \cdot 18 + 3 \cdot 11)/40 = 2$
- $\sigma^2(\text{Reddito}|\text{Sesso} = M) = (1^2 \cdot 14 + 2^2 \cdot 12 + 3^2 \cdot 14)/40 - 2^2 = 0,70$
- $\sigma^2(\text{Reddito}|\text{Sesso} = F) = (1^2 \cdot 11 + 2^2 \cdot 18 + 3^2 \cdot 11)/40 - 2^2 = 0,55$

10.2.3 Dipendenza e indipendenza tra due variabili

Il primo obiettivo legato all'osservazione congiunta di due caratteri, X e Y, osservati su uno stesso fenomeno è stabilire se vi sia o meno indipendenza tra i caratteri considerati. Riprendiamo l'esempio dell'azienda. Le distribuzioni di frequenza relativa della variabile "Stato civile" condizionate alle tre modalità della variabile "Mezzo di trasporto" sono:

Appare subito evidente che lo "Stato civile" non è indipendente dal "Mezzo di trasporto". Si noti, infatti, che in auto viaggia il 63% dei dipendenti coniugati, mentre se si considera chi viaggia in treno solo il 24% è coniugato (questo fa pensare che chi ha legami familiari preferisce spostarsi in maniera autonoma, ad esempio per non essere condizionato dagli orari dei mezzi pubblici). Il confronto tra le tre distribuzioni condizionate ha senso solo in termini di frequenze relative, non in termini di frequenze assolute poichè le marginali del "Mezzo di trasporto" sono diverse. Ad esempio, si commette un errore se si afferma che sono coniugate più persone che viaggiano in autobus rispetto alla quella che viaggiano in treno (168 contro 108), dato che 168 è il 24% dei dipendenti che viaggia in autobus, mentre 108 rappresenta ben il 41% dei dipendenti che viaggiano in treno.

Lo studio della dipendenza ci fornisce informazione aggiuntive, in quanto la conoscenza di X ci può fornire ulteriori informazioni riguardanti Y. Se la condizione di stato civile "coniugato" e la scelta del mezzo di trasporto fossero tra loro indipendenti, ci aspetteremmo di osservare delle distribuzioni di frequenza relativa condizionate fatte in questo modo:

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	freq. rel. marg. di Stat. Civ.
Coniugato	0,37	0,37	0,37	**0,37**
Non coniugato	0,63	0,63	0,63	**0,63**
Totale	1	1	1	

ossia:

1. tutte uguali tra loro;
2. uguali alla distribuzione marginale dello "stato civile", dato che questa non tiene conto della variabile "Mezzo di trasporto".

10.2.4 Indipendenza statistica in distribuzione tra variabili

Diciamo che X è *statisticamente indipendente* da Y se tutte le distribuzioni di frequenze relative di X condizionate alle modalità di Y coincidono con la distribuzione di frequenza relativa marginale di X, ossia se

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \qquad i = 1 \dots r;\ j = 1 \dots s$$

Questa condizione deve valere per ogni cella della tabella doppia.

Appare intuitivo comprendere che se X è indipendente da Y, allora Y è indipendente da X e viceversa.

Se X è indipendente da Y, allora

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \qquad i = 1 \dots r;\ j = 1 \dots s$$

ossia

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{N} \qquad i = 1 \dots r;\ j = 1 \dots s$$

ciò significa che le r distribuzioni di frequenza relativa di Y condizionate alle modalità di X sono tutte uguali alla distribuzione di frequenza relativa marginale di Y, quindi Y è statisticamente indipendente da X.

Analogamente, se Y è indipendente da X abbiamo che:

$$\frac{n_{ij}}{n_{i\cdot}} = \frac{n_{\cdot j}}{N} \qquad i = 1 \dots r;\ j = 1 \dots s$$

da cui

$$\frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}}{N} \qquad i = 1 \dots r;\ j = 1 \dots s$$

quindi le s distribuzioni di frequenza relativa di X condizionate alle modalità di Y sono tutte uguali alla distribuzione marginale di X. Pertanto X è statisticamente indipendente da Y.

In base a questa proposizione possiamo tranquillamente parlare di indipendenza di X e Y senza specificarne la *direzione*. Solitamente il concetto di indipendenza in distribuzione viene denotato attraverso il simbolo: $X \perp Y$.

In sintesi, X e Y sono indipendenti se:

- tutte le distribuzioni di frequenze relative condizionate di X|Y sono uguali alla marginale di X;
- tutte le distribuzioni di frequenze relative condizionate di Y|X sono uguali alla distribuzione di frequenza relativa marginale di Y;
- in corrispondenza di ciascuna cella della tabella doppia, la frequenza relativa congiunta è pari al prodotto delle rispettive frequenze relative marginali

$$f_{ij} = f_{i\cdot} \times f_{\cdot j} \qquad \forall i, j$$

Se ragioniamo ora in termini di frequenze assolute congiunte, $X \perp Y$ se e solo se

$$n_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{N} \qquad \forall i, j$$

ossia se, per ciascuna cella, n_{ij} è dato dal prodotto dei corrispondenti totali di riga $n_{i\cdot}$ e di colonna $n_{\cdot j}$, diviso per la numerosità campionaria N.

10.2.5 Dipendenza

Se X e Y non sono indipendenti, allora significa che vi è una qualche forma di dipendenza o associazione tra esse.

Diciamo che c'è *Massima Associazione* o *Dipendenza Perfetta* quando Y dipende perfettamente da X, quindi se in corrispondenza di ogni modalità di X vi è una sola modalità di Y (ossia se per ogni i vi è un solo j tale che $n_{ij} > 0$). Procediamo operando con un esempio:

Sesso	Occupati	Disoccupati
M	0	14
F	23	0

Tabella 10.12: Esempio di distribuzione bivariata

La dipendenza perfetta è asimmetrica: ossia se Y dipende perfettamente da X, X può non dipendere perfettamente da Y. Ci troviamo, invece, nel caso di interdipendenza perfetta, quando, considerando X e Y, ciascuna dipende perfettamente dall'altra.

10.3 Misura di associazione in una tabella a doppia entrata: l'indice Chi-quadrato

Se disponiamo di una serie di osservazioni per lòo studio di un fenomeno reale possiamo, quindi, raccoglierle in una tabella doppia, come abbiamo visto, e verificare se ci troviamo in una delle due situazioni estreme, ossia di perfetta indipendenza o perfetta associazione. Occorre, tuttavia, precisare che così come la dipendenza perfetta non è una condizione meno rara di quella di indipendenza perfetta. Questa condizione, infatti, si osserva esclusivamente quando tra le due variabili esiste una *relazione di tipo deterministico*, ossia quando una variabile è funzione dell'altra. In generale, il problema consiste nel valutare quanto è forte la dipendenza, o associazione, osservata.

Una possibile soluzione consiste nel valutare quanto ciò che abbiamo osservato si allontana da una situazione di indipendenza, ossia quanto la tabella doppia osservata si discosta dalla tabella

che avremmo dovuto osservare qualora X e Y fossero perfettamente indipendenti.

Infatti, nella situazione teorica di indipendenza, in ogni cella le frequenze assolute congiunte sarebbero date da

$$n^*_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{N}$$

e queste quantità sono direttamente confrontabili con le frequenze assolute congiunte effettivamente osservate n_{ij}.

In questo modo, tanto più le frequenze osservate n_{ij} si allontanano da quelle teoriche n^*_{ij} tanto maggiore sarà l'associazione tra le due variabili (vista come distanza dalla situazione di indipendenza).

Potremmo pensare di costruire le differenze

$$c_{ij} = n_{ij} - n^*_{ij}$$

e misurare l'associazione osservata attraverso

$$\sum_{i=1}^{r} \sum_{j=1}^{s} c_{ij}$$

Il problema è che questo indice è sempre uguale a 0, indipendentemente dalla tabella che abbiamo, dato che

$$\sum_{i=1}^{r} \sum_{j=1}^{s} c_{ij} = \sum_{i=1}^{r} \sum_{j=1}^{s} n_{ij} - \sum_{i=1}^{r} \sum_{j=1}^{s} n^*_{ij} = N - N = 0$$

Per ovviare a questo problema, in virtù del concetto di distanza tra n_{ij} ed n^*_{ij}, sono stati proposti alcuni indici basati sulle quantità $|c_{ij}|$ o c^2_{ij}, utilizzando l'indice chi-quadrato, come indicatore per misurare l'associazione. sarà quindi:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{c^2_{ij}}{n^*_{ij}} =$$

$$= \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

Il chi-quadrato possiede alcune caratteristiche che è importante menzionare: Il valore assunto dal è $\chi^2 \geqslant 0$ ed inoltre $\chi^2 = 0 \Leftrightarrow X \perp Y$. Il χ^2 è tanto più grande quanto più ci allontaniamo dal caso di indipendenza. Il chi-quadrato può essere calcolato anche attraverso la formula

$$\chi^2 = \left(\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right)$$

Il chi-quadrato è un indice di associazione simmetrico, in quanto rimane invariato se scambiamo il ruolo di X e Y poichè non tiene conto della direzione della dipendenza o, per dirlo con un linguaggio più informale, della relazione di causa-effetto.

10.3.1 Chi-quadrato normalizzato

Il valore dell'indice chi-quadrato dipende, oltre che dal livello di associazione presente tra le due variabili, anche dalla numerosità campionaria N e dal numero di modalità r ed s di X ed Y. Per facilitarne l'interpretazione, solitamente si ricorre ad una versione normalizzata dell'indice χ^2 che, essendo compresa tra 0 e 1, risulti più semplice da commentare. In particolare, è frequente l'uso dell'indice V di Cramer

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, s-1)}}$$

che varia tra 0, in caso di perfetta indipendenza, ed 1, qualora vi sia interdipendenza perfetta. Riprendendo l'esempio della nostra azienda.

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
coniugato	193	108	168	469
non coniugato	112	157	518	787
Totale	305	265	686	1256

Sotto l'ipotesi di indipendenza, la tabella delle frequenze teoriche $n_{ij}^* = n_{i\cdot} \cdot n_{\cdot j}/N$ è

	Mezzo di trasporto			
Stato civile	Auto	Treno	Autobus	Totale
Coniugato	$\frac{469 \cdot 305}{1256} = 113,9$	98,9	$\frac{469 \cdot 686}{1256} = 256,2$	469
Non Coniugato	$\frac{787 \cdot 305}{1256} = 191,1$	166,1	$\frac{787 \cdot 686}{1256} = 429,8$	787
Totale	305	265	686	1256

Il confronto tra frequenze teoriche e frequenze osservate ci indica, ad esempio, che senza la preferenza accordata al mezzo di trasporto "auto", un centinaio di dipendenti coniugati in più userebbe l'autobus.

$$\chi^2 = \frac{(193-113,9)^2}{113,9} + \frac{(108-98,9)^2}{98,9} + \cdots + \frac{(518-429,8)^2}{429,8} = 137,4$$

e pertanto

$$V = \sqrt{\frac{137,4}{1256 \cdot \min(3-1, 2-1)}} = 0,33$$

il che indica un certo grado di associazione tra il mezzo di trasporto utilizzato ed lo stato civile di coniugato.

10.4 Dipendenza di una variabile quantitativa da una qualitativa

Si può verificare la situazione in cui ci troviamo ad osservare una variabile quantitativa Y, classificata secondo le modalità di una variabile qualitativa X, e che il nostro scopo sia quello di analizzare il comportamento di quella quantitativa.

Più precisamente, si vuole verificare se l'analisi di Y può essere approfondita quando, invece di analizzare l'intero insieme delle sue osservazioni indistintamente, si prendono in considerazione queste suddivise nei gruppi identificati dalle modalità della variabile qualitativa. Ad esempio, la distribuzione del reddito pro capite (Y) per provincia italiana (X), oppure il peso (Y) per uomini/donne (X).

In questi contesti, i dati grezzi sono organizzati come segue:

	X			
x_1	x_2	x_3	...	x_r
y_{11}	y_{12}	y_{13}	...	y_{1r}
⋮	⋮	⋮	...	⋮
⋮	⋮	⋮	...	⋮
⋮	⋮	⋮	...	⋮
⋮	⋮	⋮	...	⋮
$y_{n_1,1}$	⋮	⋮	...	⋮
	⋮	$y_{n_3,3}$	...	⋮
	$y_{n_2,2}$		...	⋮
				$y_{n_r,r}$

dove ciascuna colonna ci dà la distribuzione di Y condizionata a ciascuna delle modalità di X : $Y|X = x_i$.

Alternativamente, i dati possono essere rappresentati attraverso una tabella a doppia entrata, dove le modalità della variabile quantitativa Y sono raggruppate in classi:

	Y				
X	$y_0 - y_1$	$y_2 - y_1$	...	$y_{s-1} - y_s$	Totale
x_1	n_{11}	n_{12}	...	n_{1s}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	...	n_{2s}	$n_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
x_r	n_{r1}	n_{r2}	...	n_{rs}	$n_{r\cdot}$
Totale	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot s}$	N

Per verificare quanto è utile la suddivisione in gruppi, bisogna sapere se queste distribuzioni condizionate sono simili oppure no. A riguardo una possibile soluzione consiste nel:

1. costruire le distribuzioni di frequenze relative condizionate $Y|X = x_i$;
2. rappresentarle graficamente nello stesso diagramma.

Ad esempio, possiamo rappresentare le distribuzioni condizionate $Y|X = x_i$ tramite istogrammi, e confrontarli tra loro: se sono tutti

uguali allora anche le distribuzioni condizionate sono uguali e vi è indipendenza tra Y e X.

10.4.1 Covarianza come misura della interdipendenza lineare

La covarianza misura l'interdipendenza lineare tra X e Y. La dipendenza sarà tanto maggiore in valore assoluto quanto maggiore è la tendenza dei punti (x_i, y_i) a disporsi lungo una retta. In particolare avremo tre possibili situazioni:

- covarianza positiva ⇒ retta crescente;
- covarianza negativa ⇒ retta decrescente;
- covarianza nulla ⇒ assenza di dipendenza lineare.

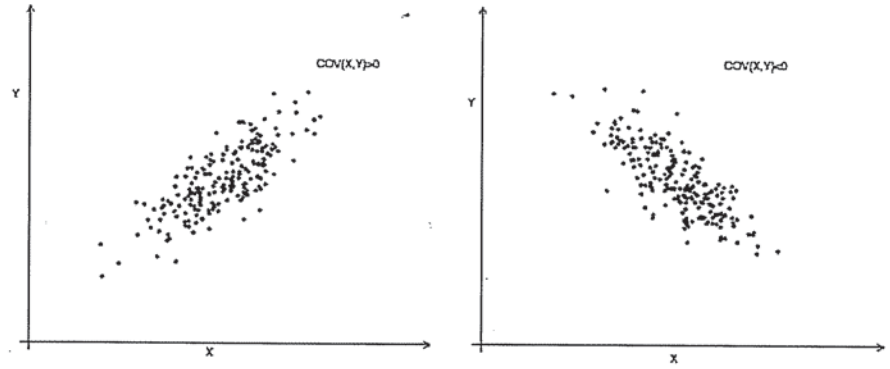


Figura 10.1: Misurazione della covarianza

10.4.2 Il coefficiente di correlazione lineare

L'ordine di grandezza della covarianza dipende dall'unità di misura con cui sono espresse le due variabili. Per cui la covarianza è una misura di interdipendenza assoluta. Non ha senso quindi confrontare covarianze calcolate su distribuzioni diverse. Tuttavia si dimostra che:

$$-\sqrt{V(X)V(Y)} \leqslant COV(X,Y) \leqslant \sqrt{V(X)V(Y)}$$

Allora, per ovviare al problema della scala di misura è opportuno trasformare la covarianza in un indice relativo

$$r = \frac{COV(X,Y)}{\sqrt{V(X)V(Y)}}$$

è chiamata *coefficiente di correlazione lineare*, il cui valore sarà $-1 \leqslant r \leqslant 1$. Per interpretare i valore di r diremo che:

- se $r = -1$ avremo una *dipendenza lineare perfetta* e i punti di disporranno perfettamente lungo una retta decrescente;
- se $r = +1$ avremo una *dipendenza lineare perfetta* e i punti si disporranno perfettamente lungo una rette crescente;
- $r = 0$ si verificherà, invece, solo se $COV(X;Y) = 0$, ossia nel caso in cui X e Y sono *incorrelate*. In questo caso ci esprimeremo dicendo che ci troviamo in assenza di dipendenza lineare.

Se $X \perp Y$, allora $COV(X,Y) = 0$ e $r = 0$; viceversa, se $COV(X,Y) = 0$ e $r = 0$, non necessariamente $X \perp Y$.

10.5 La regressione

Quando ci troviamo a comparare due variabili quantitative è interessante studiare il modo in cui l'una dipenda dall'altra. Diremo, quindi, che una variabile assumerà determinati valori in relazione ai valori assunti da un'altra variabile. Identifichiamo in questo modo una *variabile dipendente*, che convenzionalmente indichiamo con Y, ed osserviamo in che modo variano i suoi valori in funzione dei diversi valori che assume una *variabile indipendente* che indicheremo convenzionalmente con X. Se ad esempio volessimo esaminare la relazione tra "pressione arteriosa" ed "età" è intuitivo pensare di studiare la prima in funzione della seconda. Se, ancora, esaminassimo macchine con una diversa cilindrata di motore per valutare il numero di km percorribili in città con un litro di carburante, allora diremo che il numero dei km percorribili dipende dalla cilindrata del motore. In ogni caso l'obiettivo è quello di studiare il modo in cui la variabile

dipendente varia al variare della variabile dipendente, detta anche *esplicativa*.

La relazione di dipendenza di una variabile rispetto ad un'altra si esprime attraverso l'individuazione di una funzione matematica che esprima, appunto, i valori assunti dalla variabile dipendente come funzione dei valori assunti dalla variabile indipendente. In sintesi, occorre individuare una funzione $f(X)$ che approssimi l'andamento di Y.

Vediamo ora come occorre procedere. Per identificare $f(X)$, il primo passo consiste in una rappresentazione grafica delle osservazioni registrate, per mezzo di un diagramma di dispersione di Y rispetto a X. Il problema reale consiste nella necessità di individuare la "regola" che determina la variazione di una variabile al variare di un'altra, studiando la funzione matematica che possa descrivere tale rapporto di dipendenza. Possiamo pensare, quindi, che le nostre osservazioni sottendano ad un modello, esprimibile in termini di

$$Y = f(X) + \epsilon$$

dove $f(\cdot)$ rappresenta una funzione matematica che descrive la dipendenza di Y dalla variabile esplicativa X ed ϵ rappresenta la componente erratica, o stocastica, ossia parte delle oscillazioni di Y che non vengono non spiegate dalla funzione.

Il modello matematico più elementare per mezzo del quale si può descrivere la "regola" che sottende ad una serie di osservazioni di x_i e y_i è rappresentato dall'equazione di una generica retta

$$f(X) = a + bX$$

in questo caso ci esprimeremo in termini di *equazione della retta di regressione* o di *regressione lineare semplice*. Se usiamo una retta per descrivere l'andamento di Y in funzione di X, a ciascun valore osservato y_i della variabile Y, possiamo associare un valore teorico $\hat{y}_i$. Per cui diremo:

$$\hat{y}_i = a + bx_i$$

che è il valore previsto per Y in funzione dei parametri della retta di regressione.

Una volta costruito il diagramma di dispersione, in relazione ai valori delle nostre variabili, otterremo una *nuvola di punti* sulla quale possiamo tracciare differenti rette. Il problema è, quindi, quello di individuare quale retta approssima al meglio l'andamento della nostra nuvola di punti.

Per risolvere questo problema possiamo ricorrere al *metodo dei minimi quadrati* che consente, appunto, nell'individuare la retta che minimizza la distanza dei punti della "nuvola" di valori osservati y_i rispetto ai corrispondenti valori teorici y_i. Avrò quindi

$$Q = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{N} (y_i - a - bx_i)^2 \qquad (10.5.0.1)$$

dove b rappresenta il *coefficiente angolare* e a l'*intercetta*, chiamata anche ordinata all'origine. Pertanto il problema consiste nel calcolare i valori di a e b che minimizzano la distanza dei punti della "nuvola" dalla retta.

$$\min_{a,b} Q(a,b) = \min_{a,b} \sum_{i=1}^{N} (y_i \hat{-} y_i)^2 = \min_{a,b} \sum_{i=1}^{N} (y_i - a - bx_i)^2 \qquad (10.5.0.2)$$

Per trovare a e b che minimizzano la somma dei quadrati degli scarti deriviamo $Q(a,b)$ rispetto ad a e b e poniamo le derivate uguali a 0.

$$\frac{\partial Q(a,b)}{\partial a} = -2 \sum_{i=1}^{N} (y_i - a - bx_i) = 0 \qquad (10.5.0.3)$$

$$\frac{\partial Q(a,b)}{\partial b} = -2 \sum_{i=1}^{N} (y_i - a - bx_i)x_i = 0 \qquad (10.5.0.4)$$

Dalla prima equazione del sistema si deriva che

$$\sum_{i=1}^{N} (y_i - \hat{y}_i) = 0 \qquad (10.5.0.5)$$

ossia, complessivamente, valori teorici e valori osservati si eguagliano.

Inoltre, con alcuni passaggi, si deriva il sistema seguente

$$\begin{cases} \sum_{i=1}^{N} y_i - Na - b\sum_{i=1}^{N} x_i = 0 \\ \sum_{i=1}^{N} y_i x_i - a\sum_{i=1}^{N} x_i - b\sum_{i=1}^{N} x_1^2 = 0 \end{cases}$$

e risolvendolo rispetto ad a e b otteniamo i risultati finali

$$\hat{b} = \frac{Cov(X,Y)}{V(X)}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Si osservi che l'inclinazione della retta è positiva se la covarianza è positiva; è negativa se la covarianza è negativa. Inoltre a parità di $V(X)$, la retta sarà tanto più inclinata quanto maggiore è il valore, in termini assoluti, della covarianza.

10.5.1 Bontà di adattamento della retta di regressione alle osservazioni

Una volta individuata la retta di regressione, occorre valutare in quale misura la variabilità delle osservazioni sulla variabile dipendente viene catturata dalla retta di regressione. In altri termini, occorre stimare la bontà di adattamento della retta alle osservazioni.

In corrispondenza di ciascuna unità statistica, la differenza tra il valore osservato ed il valore teorico della retta di regressione viene definito residuo:

$e_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}_{xi}$ per $(i = 1, \ldots, N)$

La somma, e di conseguenza anche la media, dei residui è nulla, poiché

$$\sum_{i=1}^{N} e_i = \sum_{i=1}^{N} (y_i - \hat{y}_i) = 0$$

Allora una misura della bontà di adattamento della retta ai dati può essere fornita dalla varianza dei residui, che coincide con la media dei quadrati dei residui:

$$V(Res) = \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N} = \frac{\sum_{i=1}^{N} (y_i - \hat{a} - \hat{b}x_i)^2}{N}$$

Infatti, più la varianza dei residui è piccola, più i residui sono concentrati attorno a 0 e di conseguenza le osservazioni tendono a giacere sulla retta di regressione, ossia la retta "spiega" bene le variazioni della risposta. La varianza di Y è il valore massimo che può essere assunto dalla varianza dei residui:

$$0 \leqslant V(Res) \leqslant V(Y)$$

quindi, l'indice di bontà di adattamento più utilizzato è il *coefficiente di determinazione lineare* R^2:

$$R^2 = 1 - \frac{V(Res)}{V(Y)} \qquad (10.5.1.1)$$

dove $0 \leqslant R^2 \leqslant 1$

- $R^2 = 1$ se $V(Res) = 0$, tutti i residui sono nulli e le osservazioni giacciono perfettamente sulla retta;
- $R^2 = 0$ se $V(Res) = V(Y)$, ossia la retta di regressione non "spiega" per nulla la risposta.

R^2 coincide con il quadrato del coefficiente di correlazione lineare tra X e Y:

$$R^2 = \frac{Cov^2(X,Y)}{S^2(X)s^2(Y)} = r^2 \qquad (10.5.1.2)$$

Esempio

Dai dati pubblicati dall'Istat in occasione del 150° anniversario dell'Unità d'Italia, abbiamo estratto la serie storica degli studenti iscritti alla scuola secondaria ed all'università negli anni a partire dal 1941 fino al 2001 con cadenza decennale. I dati sono riportati nella tabella che segue:

Anni scolastici	t	Iscritti Univ.	Iscritti Sc. Sup.
1941 – 42	1	146	982
1951 – 52	2	227	1212
1961 – 62	3	288	2379
1971 – 72	4	760	4019
1981 – 82	5	1025	5302
1991 – 92	6	1475	5009
2001 – 02	7	1703	4378

Vogliamo valutare il trend di crescita degli iscritti per i due livelli scolastici e verificare se tra essi vi è correlazione. Per fare ciò impostiamo un'analisi della regressione stabilendo una relazione lineare tra anni scolastici e iscritti, ossia

$$y_i = a + bt_i + e_{t_i}$$

dove con y_t abbiamo indicato il numero degli iscritti all'università, con t la variabile indipendente tempo ($t = 1, 2, \ldots, 7$) corrispondente a ciascun anno scolastico, a e b i coefficienti del modello da stimare ed infine con $e_{t_i} = \hat{y}_i - y_i$ l'errore che si commette nell'interpolare i dati empirici y_i con il modello di regressione $y_i = a + bt_i$. Applicando il metodo dei minimi quadrati abbiamo detto che la stima di b è

$$\hat{b} = \frac{Cov(t,Y)}{V(t)}$$

mentre la stima di a è

$$a = \bar{y} - \hat{b}\bar{t}$$

Procedendo con i calcoli e ricordando che

$$Cov(t,Y) = \frac{1}{N}\sum_{i=1}^{N}(t_i - \bar{t})(y_i - \bar{y})$$

mentre

$$V(t) = \frac{1}{N}\sum_{i=1}^{N}(t_i - \bar{t})^2$$

possiamo facilmente costruire una tabella dei calcoli per gli iscritti all'università (10.2).

Essendo il periodo costituito da N=7 anni si hanno:

$$\bar{t} = 4$$

$$\bar{y} = 803,4$$

$$V(t) = \frac{1/N}{\sum_{i=1}^{N}}(t_i - \bar{t})^2 = \frac{1}{7}28 = 4$$

t	y_i	$t_i - \bar{t}$	$y_i - \bar{y}$	$(t_i - \bar{t})^2$	$(t_i - \bar{t})(y_i - \bar{y})$
1	146	-3	-657,4	9	1972,3
2	227	-2	-576,4	4	1152,9
3	288	-1	-515,4	1	515,4
4	760	0	-43,4	0	0,0
5	1025	1	221,6	1	221,6
6	1475	2	671,6	4	1343,1
7	1703	3	899,6	9	2698,7
Somma	5624	0	0	28,0	7904,0

Figura 10.2: Procedimento di calcolo per la regressione

$$\mathrm{Cov}(t, Y) = \frac{1/N}{\sum_{i=1}^{N}}(t_i - \bar{t})(y_i - \bar{y}) = \frac{1}{7}7904,0 = 1129,1$$

In definitiva

$$\hat{b} = \frac{\mathrm{Cov}(t, Y)}{V(t)} = 1129,1/4 = 282,3$$

In fine

$$a = \bar{y} - \hat{b}\bar{t} = 803,4 - 282,3(4) = -325,7$$

Quindi la retta di regressione può essere scritta

$$\hat{y}_i = -325,7 + 282,3t_i$$

Sostituendo a t_i i tempi 1,2,...,7 otteniamo i valori stimati degli iscritti all'università i cui valori sono riportati nella tabella (10.3).

Mentre se si vuole effettuare la previsione degli iscritti all'anno scolastico 2011/12 sia ha

$$\hat{y}_i = -325,7 + 282,3 \times 8 = 1932,7$$

Tuttavia essendo la variabile dipendente una variabile conteggio si approssima all'intero più vicino, quindi il numero di studenti universitari previsti per il 2011/12 è 1933.

Per ottenere la valutazione della bontà di adattamento calcoliamo la varianza dei residui

$$V(res) = \frac{\sum_{t=1}^{N}(y_i - \hat{y}_i)^2}{N} = 110179,4/7 = 15739,92$$

$\hat{y}_i$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
-43,4	35883,2	432212,3
238,9	140,6	332269,9
521,1	54355,6	265666,6
803,4	1886,0	1886,0
1085,7	3686,2	49093,9
1368,0	11449,0	451008,2
1650,3	2778,8	809228,8
	110179,4	2341365,7

Figura 10.3: Valori stimati degli iscritti all'università

e la varianza totale

$$V(Y) = \frac{\sum_{t=1}^{N}(y_i - \bar{y})^2}{N} = 2341365,7/7 = 334480,8$$

Infine

$$R^2 = 1 - \frac{V(res)}{V(Y)} = 1 - \frac{15739,92}{334480,8} = 0,953$$

che indica un ottimo accostamento tra modello di regressione e dati empirici; il grafico in figura 10.4 mette in chiara evidenza quanto appena sostenuto.

In modo analogo si possono ripetere le analisi di valutazione della retta di regressione per gli iscritti alle scuole secondarie che lasciamo per esercizio al lettore.

Proviamo ora ad analizzare la correlazione esistente tra numero di iscritti all'università e numero di iscritti alle scuole superiori.

Dalla teoria abbiamo detto che una misura della correlazione è data dal coefficiente

$$r = \frac{\mathrm{Cov}(t, Y)}{\sqrt{V(t)V(Y)}}$$

Dalle analisi precedenti abbiamo già calcolato la varianza degli iscritti all'università che risulta essere

$$V(Y) = 334480,8$$

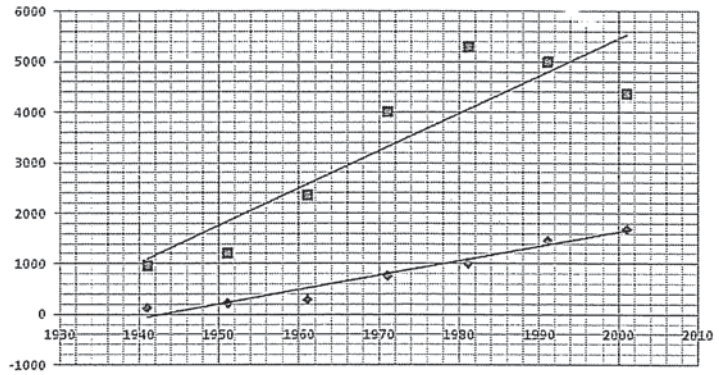


Figura 10.4: Grafico Regressione

ISCRITTI all'università Y	ISCRITTI scuole superiori X	$(x_i - \bar{x})^2$	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(x_i - \bar{x})$
146	982	5493666,3	-2343,9	-657,4	1540918,7
227	1212	4468392,0	-2113,9	-576,4	1218487,7
288	2379	896538,4	-946,9	-515,4	488037,2
760	4019	480447,0	693,1	-43,4	-30102,2
1025	5302	3905140.6	1976,1	221,6	437856,8
1475	5009	2832969,9	1683,1	671,6	1130350,7
1703	4378	1107004,6	1052.1	899,6	946477.7
5624	23281	19184158,9			5732026,4

Figura 10.5: Procedimento per il calcolo della correlazione

Calcoliamo le altre componenti dell'espressione.

La media degli iscritti alle scuole superiori è $\bar{x} = 3325,9$. Mentre la varianza degli iscritti alle scuole superiori è

$$V(X) = \frac{19184158,9}{7} = 2740594,1$$

La covarianza è

$$Cov(X,Y) = \frac{5732026,4}{7} = 818860,9$$

Infine il coefficiente di correlazione è

$$r = \frac{Cov(t,Y)}{\sqrt{V(t)V(Y)}} = \frac{818860,9}{\sqrt{2740594,1 \cdot 334480,8}} = 0,855$$

da cui si evince che tra gli iscritti all'università e gli iscritti alle scuole superiori c'è un elevata correlazione positiva.