Riassunto tecniche quantitative d'analisi

Tecniche quantitative di analisi (Università degli Studi di Milano-Bicocca)

Capitolo 1

RILEVAZIONI DELLE INFORMAZIONI

Fase che precede l'analisi statistica:

- Raccolta dati
- Organizzazione dati in matrice dati

Il ricercatore si pone interrogativi e cerca di formulare risposte basandole per quanto più possibile su riscontri con la realtà.
Spesso si avvale dell'uso di tecniche = complesso più o meno codificato di norme e modi di procedere, riconosciuto da una collettività trasmesso o trasmissibile per apprendimento, elaborato allo scopo di svolgere una data attività manuale o intellettuale ricorrente.
Le **tecniche** sono caratterizzate da:

- codificazione
- condivisione da parte di una comunità
- possibilità di insegnarne il funzionamento ad altri
- carattere ricorrente dei problemi

Tra la fase di rilevazione e la fase di analisi e quindi tra le **tecniche di rilevazione e quelle di analisi**
Fase di rilevazione → dopo aver definito gli interrogativi di ricerca si sceglie il luogo, il periodo e l'oggetto individuando gli specifici oggetti sui quali si indagano le caratteristiche.
Fase di analisi → elaborazione dei dati al fine di acquisire elementi conoscitivi intorno alla realtà.
La **statistica** si situa in questa fase infatti essa è costituita da UN INSIEME DI TECNICHE PER L'ANALISI DEI DATI.
La scelta delle tecniche di rilevazione e di analisi da adottare dipende dalla natura del problema cognitivo affrontato.
In ogni indagine, tecniche di rilevazione e tecniche di analisi sottostanno ad un ordine cronologico, di fatti la rilevazione precede l'analisi.
La scelta di una determinata tecnica di rilevazione: determinate tecniche d'analisi suppongono che le informazioni siano rilevate in un certo modo, determinate tecniche di rilevazione pregiudicano la possibilità di ricorrere ad alcune tecniche di analisi.
Le tecniche di analisi statistica dei dati presuppongono infatti che le informazioni da sottoporre ad analisi siano state registrate e organizzare in una matrice dati
In entrambi in casi si parla di STATISTICA DESCRITTIVA
La rilevazione strutturata si incentra su 3 elementi:

- Proprietà → caratteristica che è possibile attribuire ad un determinato tipo di oggetto o unità di analisi
- Stati → diversi modi in cui quella caratteristica si può manifestare
- Unità → referenti sui quali si rilevano informazioni

Il percorso tipico di un'indagine sociologica consiste in un itinerario ciclico che inizia dalla teoria, attraversa le fasi di rilevazione delle informazioni di analisi dei dati e ritorna alla teoria.
Le unità sulle quali il ricercatore vuole rilevare informazioni e di conseguenza le proprietà dipendono dall'itinerario dei suoi interrogativi di ricerca → ***DISEGNO DI RICERCA.***
Questo deve essere tradotto in termini empirici, le proprietà in variabili, le unità in casi e gli stati in dati o valori.
UNITA' D'ANALISI → tipo di oggetto sul quale si vogliono rilevare informazioni.
Possono essere individui singoli o anche gruppi strutturati di individui come aggregati territoriali, o testi scritti o altri prodotti culturali, luoghi, eventi e situazioni → è chiaro quindi che non c'è limite in ciò che può essere definito come unità di una ricerca empirica.
Alla scelta dell'unità si accompagna l'ambito spazio-temporale che determinano la popolazione di riferimento (U).
Quando l'unità corrisponde ad un prodotto naturale o testi la popolazione di riferimento si chiama **CORPUS.**
L'ambito spazio temporale definisce i confini entro i quali sono generalizzabili i risultati della ricerca.
Un'importante distinzione va **tracciata tra unità di rilevazione ed unità di analisi**, è possibile che una ricerca preveda che le informazioni vengano rilevate presso un tipo di unità e poi riferite con un altro tipo di unità.

Ogni volta che i due tipi di unità non coincidono il ricercatore deve adoperarsi al fine di evitare le particolari caratteristiche percezioni e aspettative degli oggetti osservati vengano attribuite indebitamente ad oggetti di altro tipo.
Definita la popolazione quindi il ricercatore deve decidere se rilevare le informazioni da **tutti gli oggetti** che le appartengono oppure solo da **sottoinsiemi** → di solito per comodità si ricorre al campionamento.
Queste decisioni determinano presso quali elementi della popolazione di riferimento verranno effettivamente rilevate le informazioni.
I casi → sono gli esemplari di una data unità di analisi che vengono effettivamente inclusi nell'indagine
Per tradurre una proprietà in termini empirici occorre darne **una DEFINIZIONE OPERATIVA** ossia stabilire alcune procedure per rilevare gli stati delle proprietà sui casi e registrarli in forma simbolica al fine di sottoporli ad analisi.
Il passaggio successivo consiste nella applicazione delle sopracitate procedure ai casi studiati: si tratta **dell'OPERATIVIZZAZIONE IN SENSO STRETTO.**
La definizione operativa viene fatta a tavolino mentre l'operativizzazione è la sua applicazione pratica

- → con la formulazione di una definizione operativa e la sua applicazione, una proprietà viene trasformata in **VARIABILE** e i suoi stati vengono trasformati in **MODALITA' della variabile stessa**

Ad ognuna di essa viene applicato un differente **valore simbolico**.
L'operativizzazione nelle scienze sociali non è differente da quella nelle scienze naturali:
non sussiste un rapporto di corrispondenza biunivoca fra proprietà e variabile, in quanto una proprietà può essere operativizzata in modi diversi → ***la decisione di come operativizzare una proprietà è affidata al ricercatore al quale è solo chiesto di***:

- esplicitare
- giustificare le scelte

La definizione operativa oltre ad avere una caratteristica di arbitrarietà rappresenta anche il criterio di oggettività della ricerca sociale.
Infatti secondo Marradi solo dopo che si è stabilita una catena di operazioni attraverso le quali lo stato di una serie di soggetti sulle proprietà X, Y e Z viene rilevato, classificato e registrato abbiamo compiuto un passo per ridurre l'opinabilità delle nostre affermazioni.
Concetti più specifici → indicatori, dai quali è più facile partire per costruire una definizione operativa.
Nel passaggio dalle proprietà e dall'unità di analisi alle variabili ai casi, il ricercatore deve sempre essere consapevole del ruolo ricoperto dall' **ERRORE DI RILEVAZIONE**.
Questo errore corrisponde al divario tra i dati registrati e la realtà che si vuole indagare
L'errore è composto da due componenti:

- errore sistematico
- errore accidentale

valore osservato = stato effettivo + errore sistematico + errore accidentale
per cui ….
Errore di rilevazione = valore registrato – stato effettivo = errore sistematico + errore accidentale

Errore sistematico →è costante nel senso che si presenta in tutte le rilevazioni, il suo valore medio sul totale dei casi non è mai 0 ma assume un valore o positivo o negativo quindi tende o a sovrastimare o sottostimare lo stato effettivo.
Errore accidentale → varia da rilevazione a rilevazione, varia in ipotetiche repliche della stessa rilevazione sullo stesso individuo, e varia da individuo a individuo.
Si tratta di oscillazioni che tendono ad annullarsi a vicenda
L'errore sistematico è dunque una parte di errore comune a tutte le applicazioni di una determinata rilevazione, l'errore accidentale è una parte di errore specifica di ogni singola rilevazione.
Gli errori possono sorgere principalmente in due fasi:

1. Fase teorica o di indicazione
2. Fase empirica o di operativizzazione

In entrambi le fasi infatti gli indicatori può essere stato scelto male, una definizione operativa può essere applicata male:

- **L'errore nella fase di indicazione** è un errore di tipo sistematico, in questo caso l'indicatore copre malamente la proprietà generale che interessa e quindi si dice che c'è un difetto nel rapporto di indicazione.
- **L'errore nella fase di operativizzazione** può essere sia sistematico sia accidentale, in particolare possiamo distinguerne 3 fasi:

• la fase di selezione dei casi
• la fase di rilevazione delle informazioni
• la fase dei trattamenti dei dati

- **Gli errori di selezione** sono dovuti al fatto che si cerca di rilevare informazioni non presso l'intera popolazione di riferimento bensì su un campione

Vi sono 3 tipi di errore di selezione:

• errore di copertura
• errore di campionamento
• errore di non risposta

- **Gli errori di osservazione** possono essere addebitati a 4 fonti**:**

1- all'intervistatore

2- all'intervistato

3- errori dovuti allo strumento

4- errori dovuti al modo di somministrazione

- **Gli errori del trattamento** dei dati si verificano dopo che le informazioni sono state rilevate e consistono in errori di codifica, di trascrizione, di memorizzazione su supporto informatico e di elaborazioni.

La varietà della natura e delle fonti di errori, questo modo articolato di vedere l'errore viene chiamato **approccio ALL'ERRORE COMPLESSIVO,** questo non è misurabile, in quanto troppe componenti sfuggono al controllo del ricercatore.

Per VARIABILE si intende una PROPRIETA' OPERATIVIZZATA:

Le variabili costituiscono l'elemento centrale dell'analisi empirica: esse sono i termini essenziali, gli elementi fondamentali, il vocabolario delle scienze sociali.

Un modo importante per classificare le variabili riguarda il tipo di operazioni logiche e matematiche alle quali i loro valori possono essere sottoposti, in quanto stabilisce la legittimità di determinate procedure di analisi statistica.

In prima battuta si possono distinguere **le variabili di tre tipi: nominali, ordinali e cardinali**

- **VARIABILI NOMINALI**→ risultano dall'operativizzazione di una proprietà che assume stati discreti non ordinabili, la natura discreta significa che esistono confini molto netti fra essi, tali per cui non è possibile immaginare stati intermedi.
  L'operativizzazione che permette di passare dzlla proprietà alla variabile in questo caso si basa sulla **CLASSIFICAZIONE.**

La classificazione consiste nell'individuazione dell'insieme di classi che corrispondono agli stati che una proprietà può assumere.

Queste classi, che poi vengono fatte corrispondere alle modalità della variabile devono presentare 2 requisiti:

1. esaustività
2. mutua esclusività

Alle modalità delle variabili viene associato un valore che serve ad identificare le modalità e differenziarla dalle altre. Un caso particolare delle variabili nominali è quello in cui le modalità **sono 2 →dicotomiche**

- se la proprietà da operativizzare presenta stati discreti ordinabili, la risultante è la **VARIABILE ORDINALI**.

Una variabile ordinale non è nota la distanza che intercorre fra le diverse modalità

La definizione operativa in questo caso si basa sull'assegnazione a modalità ordinate o semplicemente sull'ordinamento il quale tiene conto, oltre ai requisiti menzionati per classificazione, dell'ordine sottesi agli stati della proprietà, per questo vengono quasi sempre usati serie di numeri naturali.
Le variabili possono essere ordinali per due motivi:

a. Perché derivano da proprietà originariamente costruite da stati discreti.
b. Perché derivano da proprietà continue, che sono state registrate su una sequenza solo ordinale per difetto di strumenti di misurazione.

- **VARIABILI CARDINALI** → sono caratterizzate dall'assegnazione di valori numerici che hanno pieno significato numerico, i numeri possiedono non solo le caratteristiche ordinali dei numeri ma anche quelle cardinali.

Dato il carattere cardinale dei valori, fra le modalità di una variabile di questo tipo non solo si potranno stabilire relazioni di eguaglianza e di diversità e non solo relazione d'ordine, si potranno effettuare anche tutte le operazioni aritmetiche sui valori**.**
Le variabili cardinali possono essere create a partire da due tipi di finizione operativa: **la misura e il conteggio.**

1. **La misurazione suppone le seguenti condizioni:**
   la proprietà da misurare è continua, cioè può assumere infiniti stati intermedi in un dato intervallo fra due stati qualsiasi, e la seconda è che la comunità scientifica ha elaborato e accettato un'unità di misura prestabilita che funge da grandezza di riferimento con la quale si può confrontare la grandezza da misurare
2. **Il conteggio sta al centro della definizione operativa quando:**

- La proprietà da registrare è discreta, assume cioè stati finiti, non frazionabili

- La proprietà è concepibile come il possesso o la relazione con un determinato numero di elementi, in questo caso l'operativizzazione consiste semplicemente nel contare gli elementi posseduti dal caso o con i quali quest'ultimo è in relazione.

Le variabili cardinali basate sulla misurazione sono rare nelle scienze umane, tali variabili sono tutte derivate da proprietà tipiche delle scienze naturali.
Esse tuttavia non riescono a passare dalle condizioni di proprietà continue a quelle di variabili cardinali per un difetto nella fase di operativizzazione, in particolare per la difficoltà, forse l'impossibilità, di ideare un'unità di misura applicabile agli atteggiamenti umani.
L'obiettivo delle tecniche con le scale autoancoranti o le scale di collocazione è quello di avvicinarsi a delle misurazioni nel senso proprio, dando luogo a variabili nelle quali la distanza tra due valori si nota → per questo diciamo che le **tecniche di scaling danno vita a variabili di tipo quasi-cardinali**.
Si noti la **CUMULATIVITA**' delle caratteristiche dei tre principali tipi di variabili presentati: si tratta di livelli dove ognuno include gli attributi dei livelli inferiori.
Fra i valori delle variabili nominali si possono solo instaurare relazioni di eguaglianza e diversità, fra quelle variabili ordinali si possono stabilire anche quelle di ordinamento, e fra i valori delle variabili cardinali, oltre alle relazioni menzionate, si possono instaurare quelle legate alla conoscenza della distanza tra i valori.
Queste differenze formali tra le variabili, in particolare il fatto che su di esse non siano consentite operazioni, fanno si che i tre tipi di variabili debbano essere analizzati con procedure diverse dai livelli più elementari.
Va aggiunto che le tecniche di analisi dei dati che la statistica ha sviluppato sono destinati o a variabili nominali o cardinali, mentre sono rare quelle finalizzate esplicitamente a quelle ordinali.
Quindi una variabile ordinale può sempre essere trattata come se fosse una variabile nominale, trascurando il fatto che le sue categorie sono ordinate.
Naturalmente se si procede in questo modo si perdono informazioni.
Infatti le tecniche di analisi messe a punto per le variabili cardinali sono assai più numerose, solide, più semplici e permettono analisi assai più sofisticate delle tecniche messe a punto per le variabili nominali.
Quando si trovasse di fronte a una variabile congiuntamente ordinale, il ricercatore dovrà, quasi sempre, sceglie se trattarla con le tecniche delle variabili nominali o con quelle delle variabili cardinali: sappia che se opta per la seconda soluzione egli dovrà interpretare i sui risultati con estrema cautela.

Le variabili dicotomiche possono essere trattate statisticamente con gli strumenti propri delle variabili cardinali e questo perché non si ponga il problema delle distanze che separano i valori.
A causa di questa preziosa caratteristica, talvolta il ricercatore dicotomizza variabili a più categorie aggregando modalità dal significato prossimo.

MATRICE DATI

Si organizza il materiale grezzo in una forma tale da poter essere analizzato con gli strumenti di analisi statistica.
In generale questo processo di organizzazione del materiale empirico consiste nella trasformazione in una matrice di valori, la cosi detta matrice dati.
La matrice dati consiste in *un insieme rettangolare di numeri*,dove in riga abbiamo i casi e in colonna le variabili, da ogni cella derivante dall'incrocio abbiamo un dato, ossia il valore registrato per una particolare variabile per un particolare caso.
Due sono le condizioni necessarie perché le informazioni afferenti a un certo insieme di casi possono essere organizzati nella forma di matrice dati:

- L'unità d'analisi deve essere sempre la stessa
- Su tutti i casi devono essere state rilevate le stesse informazioni

L'operazione di inserimento del materiale empirico grezzo in una matrice dati viene chiamata **codifica** e avviene con l'ausilio del codice.
Il **codice** è un documento che indica la posizione di ogni variabile nella matrice dati e assegna ad ogni modalità di ogni variabile un valore numerico.
Nella pratica della ricerca sociale, molto spesso il codice è incorporato nel questionario spesso → accanto ad ogni domanda si riporta la posizione della variabile generata dalla domanda stessa sulla riga, e ogni risposta è contrassegnata da un valore.
Ogni riga della matrice dati corrisponde a un caso→ leggendo una riga si sa come quell'individuo ha risposto alle domande.
Ogni colonna della matrice corrisponde ad una variabile → leggendo una colonna si conosce l'insieme delle risposte date a quella domanda da tutti gli intervistati.
L'insieme dei dati a partire dal pacco di questionari da luogo ad una matrice rettangolare di numeri
La matrice dati memorizza su un supporto informatico che viene chiamato con il termine file.
Operazione di codifica ossia di trasformazione delle modalità delle variabili in valori.
Che noti che fra le risposte codificate è stata prevista la modalità non risposta (vedi esempio pg37)
L'operazione può avvenire in maniera assai semplice digitando i valori della matrice sulla tastiera di un computer, creando in questo modo un file in formato cosidetto Ascii, oppure si può utilizzare un foglio elettronico o un data-base, soluzione preferibile poiché avita alcuni possibili errori di registrazione.
Ci sono procedure automatizzate di immissione dati come nel caso delle tecniche cati (interviste telefoniche assistite al cumputer) o capi (interviste faccia a faccia assistite al computer) nelle quali i questionario viene letto dall'intervistatore direttamente dal video di un computer e la risposta viene immediatamente digitata sulla tastiera e memorizzata nella matrice dati senza la mediazione di supporti cartacei.
Assieme alla matrice, vanno anche fornite al programma di elaborazione le istituzioni di definizione delle variabili, che permettono al programma stesso di leggere la matrice dati → a questo punto la matrice dati risulta trasformata nel cosidetto sistem file ed è pronta per l'analisi statistica

FONTI STATISTICHE

I dati che gli servono sono già stati rilevati, nella maggioranza dei casi da enti pubblici.
L'attività dell'amministrazione pubblica genera dati sia per effetto della normale attività amministrativa, sia per mezzo di rilevazione aventi un esplicito fine conoscitivo.
In tutti quei casi il dato statistico è il sottoprodotto di un atto amministrativo, possiamo parlare di **RILEVAZIONE INDIRETTA**.
Altre volte invece la produzione del dato avviene per **rilevazione DIRETTA** → le informazioni vengono espressamente raccolte al fine di conoscere un determinato fenomeno sociale.
L'unità d'analisi non è costituita dall'individuo ma dal territorio.
I dati vengono ricondotti e resi accessibili soltanto a livello aggregato.
Il ricercatore può consultare rappresentazioni tabulari di distribuzione di frequenza ma non la matrice dati originaria a partire da quelle distribuzioni sono state prodotte.

In questo caso il ricercatore può compiere l'analisi secondaria e applicare tutte le più comuni tecniche di analisi statistica dei dati.
Le opportunità di analisi secondaria stanno aumentando grazie ai processi tecnologici.
Naturalmente avvalersi delle opportunità offerte dalle fonti statistiche non è privo di limiti infatti:

- Il fatto di doversi servire di dati già esistenti e raccolti con finalità diverse da quelle implicate dall'interrogativo di ricerca crea sovente situazioni nelle quali la natura dei dati non soddisfa le esigenze del ricercatore che deve accettare anche le definizioni operative adottate dagli enti produttori.
- È facile che le fonti statistiche non esprimano le informazioni in formati immediatamente fruibili per gli specifici intenti del ricercatore, non riportino determinate tabulazioni incrociate, non disarticolino le informazioni in base alle variabili interessate.
- Nei paesi sviluppati esiste un elevato numero di enti produttori di dati, il che favorisce ridondanze e sovrapposizioni nelle informazioni. Ne consegue che il ricercatore deve essere particolarmente attento non solo quando combina informazioni provenienti da fonti diverse ma anche quando decide da quale attingere tra le più fonti che offrono lo stesso prodotto.
- Le statistiche ufficiali spesso riguardano esclusivamente proprietà fattuali, riferite a informazioni oggettive e comportamentali, con esclusione delle opinioni, motivazioni e atteggiamenti.

Nel 1989 l'apparato della statistica sociale in Italia ha subito una trasformazione, con l'istituzione del sistema statistico nazionale avente l'obbiettivo di coordinare tutte le competenze e le attività di raccolta dati nei vari organismi centrali e periferici della pubblica amministrazione.
In questo quadro **l'istituto nazionale di statistica (ISTAT**) è diventate un organo del SISTAN.
L'Istat coordina la raccolta di informazioni in molti enti pubblici e ospita nelle sue pubblicazioni i dati statistici più rilevanti prodotti da questi enti.
Il bollettino mensile di statistica che ha la funzione di pubblicare nei rispettivi annuari.
L'Istat conduce numerosi censimenti con cadenza decennale → vedi pagina 41.

Capitolo 2

ANALISI MONOVARIATA

Le tecniche monovariate hanno come punto di partenza la distribuzione di frequenza in cui ogni modalità della variabile viene associata la frequenza con cui essa si presenta nella matrice.
Al fine di descrivere le distribuzioni di frequenze, le tecniche di analisi statistica fanno ampio uso di forme di rappresentazioni tabulare e grafica, nonché di valori caratteristici che danno informazioni sintetiche su alcune caratteristiche della distribuzione.

1. Rappresentazioni tabulari di distribuzioni di frequenza

La forma più diffusa di rappresentazioni di **distribuzione di frequenza è la tabella che presenta 2 colonne**, la prima presenta le modalità della variabile sotto esame, mentre nella seconda accanto a ciascuna modalità il numero di volte che il dato corrispondente compare nella corrispondente colonna della matrice dati.
Per cogliere meglio l'incidenza delle singole modalità rispetto alla distribuzione complessiva e rispetto alle altre modalità, si ricorre alle **FREQUENZE RELATIVE** che annullano l'effetto delle numerosità dei casi.
Un primo tipo di frequenza relativa è una proporzione, che si ottiene dividendo ogni singola frequenza assoluta per il numero totale dei casi della distribuzione.
La somma delle proporzioni di tutte le modalità è sempre uguale a 1, di conseguenza tutte le frequenze si collocano entro un campo di variazione che va da 0 a 1, li che agevola il confronto fra frequenze di modalità diverse.
Quando si quantificano le differenze fra due percentuali, occorre fare attenzione ad essere concettualmente e terminologicamente precisi.
Se le variabili di cui si vuole rappresentare la distribuzione è di tipo ordinale o cardinale, è possibile avvantaggiarsi della relazione d'ordine sottesa alle sue categorie per calcolare anche un altro tipo di frequenza.

La **frequenza cumulata** di una categoria corrisponde al numero di casi che appartengono a quella categoria o a una categoria precedente.
La frequenza retro-cumulata di una categoria corrispondente al numero di casi che appartengono a quella categoria o a una categoria successiva di norma il ricercatore dovrebbe attenersi a un criterio di parsimoniosità nella presentazione dei risultati.
E pertanto limitarsi a presentare un solo tipo di frequenza, gli converrà quindi presentare solo le frequenze percentuali, accompagnate però dall'indicazione della base del calcolo delle percentuali, ossia il numero complessivo di casi in valore assoluto.
La specificazione del numero dei casi sui quali le percentuali sono state calcolate ha due finalità:

- Comunicare lo spessore empirico dei risultati
- Permettere di risalire comunque alle frequenze assolute

Se la variabile è ordinale o cardinale, in genere sarà opportuno rispettare la relazione d'ordine sottesa alle sue modalità.
Nel caso di una **distribuzione riferita a una variabile nominale** il ricercatore ha maggiore libertà.
Quando una **variabile è cardinale** è possibile che le modalità siano assai numerose, in questo caso una tabulazione che riportasse ogni modalità e la sua frequenza non realizzerebbe l'obiettivo di rappresentare in modi sintetico la distribuzione.
Quando si riportano le **frequenze percentuali** occorre evitare di specificare un numero eccessivo di valori decimali, se i dati sono stati rilevati da un'inchiesta campionaria è opportuno riportare al massimo un solo valore decimale (regola dell'arrotondamento).
Lo 0 è un numero avente dignità pari a quella di tutti gli altri numeri.

2. Dati errati e mancanti

Pulizia preventiva → è facile che le attività di rilevazione dati diano luogo ad errori e incongruenze nella matrice dati, i quali vanno eliminati prima di effettuare operazioni di analisi statistica vera e proprio.
In primo luogo occorre controllare che tutti i dati riportati in una determinata colonna della matrice dati siano plausibili, appartengano cioè al ventaglio di valore previsti dal codice per la corrispondente variabile.
Un controllo più articolato consiste nel confrontare le distribuzioni di variabili fra loro concatenate per far emergere eventuali incongruenze.
L'ispirazione delle distribuzioni di frequenza fa emergere l'errore e perfette di correggerlo.
Altri controlli di congruenza si effettuano per mezzo di una tabulazione incrociata di due variabili.
In secondo luogo occorre accettarsi che la matrice dati non presenti dei buchi.
Si tratta del problema dei dati mancanti:
In alcune situazioni è possibile esaminare le modalità assunte dal caso su altre variabili e infierire da queste lo stato non rilevato.
La soluzione più comune è quella di prevedere sin dalla fase della definizione operativa una categoria residuale cui assegnare tutti i casi di cui non è possibile rilevato lo stato, e includerla nel codice per la variabile in questione. Anzi in alcune situazioni conviene prevedere e tenere distanti più categorie mancanti.
I dati mancanti comportano sempre una complicazione nell'analisi dei dati.
Se essi derivano da errore di codifica o dalla non applicabilità della definizione operativa, conviene escluderli dall'analisi.

3. Rappresentazioni pratiche di distribuzioni di frequenza

Le distribuzioni di frequenza possono essere rappresentate anche in forma grafica, tali rappresentazioni non forniscono a rigore informazioni aggiuntive rispetto alle formule tabulari ,a sono di grande efficacia comunicativa in quanto non chiedono al lettore alcuna competenza numerica.
Nell'ambito dell'analisi monovariata **le rappresentazioni grafiche si basano su un semplice principio**: le dimensioni dei segni corrispondenti alle diverse modalità di una variabile sono direttamente proporzionali alle rispettive frequenze di tali modalità.

tipi di rappresentazioni grafiche

RAPPRESENTAZIONI LINEARI = vengono sviluppate lungo **due dimensioni spaziali**: le modalità delle variabili vengono disposte lungo una dimensione e le frequenze vengono rappresentate lungo l'altra tracciando, in corrispondenza di ciascuna modalità, un segno di lunghezza proporzionale alla corrispondente frequenza.
Queste da luogo a **due rappresentazioni differenti**:

A →le modalità vengono disposte lungo la dimensione orizzontale o lungo quelle verticali, e di conseguenza i segni vengono sviluppati verticalmente e orizzontalmente? La soluzione adottata è indifferente e di solito va presa esclusivamente in funzione dello spazio a disposizione per il grafico.
B →Di solito si ricorre a rettangoli, di base eguale e di lunghezza proporzionale alle frequenze, nel qual caso si ha un diagramma a colonne o un diagramma a nastri.
Questi possono essere ricondotti entrambi sotto l'etichetta DIAGRAMMI A BARRE.
C →Se il grafico riporta frequenze percentuali, ad esempio può essere disegnato in modo tale da riportare tutte le frequenze fra lo 0 e il 100 % oppure fra lo 0 e la frequenza più elevata fra quelle registrate oppure soltanto le frequenze comprese tra quella più bassa e la più elevata.
Vi sono due tipi di ***RAPPRESENTAZIONI CIRCOLARI***
Nel diagramma a settori circolari il numero complessivi di casi viene fatto corrispondere all'area di un cerchio, la quale viene suddivisa in un numero di settore pari al numero di modalità.
Ogni settore ha una superficie proporzionale alla frequenza delle modalità corrispondente.
Nel DIAGRAMMA A RAGGERA viene fatto partire da un unico punto un numero di raggi pari al numero di modalità.
I raggi, che sono disposti ad intervalli regolari, hanno una lunghezza proporzionale alla frequenza della modalità corrispondente.
Le rappresentazioni lineari suggeriscono implicitamente che esista un ordine fra le categorie, che quando tale ordine non esiste, questa percezione è meno marcata nelle rappresentazioni circolari.
Inoltre le rappresentazioni lineari agevolano il confronto visivo fra le due modalità sul totale dei casi, le rappresentazioni circolari, al contrario, facilitano la percezione dell'incidenza di una modalità sul totale, ma rendono più arduo il confronto fra due modalità, i grafici circolari vanno evitati quando la variabile presenta un elevato numero di modalità.
A prescindere dal tipo di rappresentazione grafica prescelta, **vi sono anche altre decisioni da prendere**:

a- I motivi che militino a favore dell'uso delle frequenze percentuali nelle rappresentazioni tabulari rimangino validi anche per quelle grafiche. Naturalmente è importante che il tipo di frequenza usato risultati evidente a chi legge il grafico.
b- Come nelle rappresentazioni tabulari, l'evidenziazione della presenza di dati mancanti è senz'altro consigliabile se permette di interpretare o inquadrare in maniera più efficace la distribuzione fra le categorie sostantive.

I grafici fino ad ora esaminati sono usati per qualsiasi variabile, tuttavia, per le variabili ordinali e specie per quelle cardinali, si gode, come nelle altre tecniche di analisi, di qualche operazione aggiuntiva.
Per le variabili cardinali si può ricorrere all'**istogramma** il quale consiste in un diagramma a colonne congiunte, nel quale le basi dei rettangoli sono proporzionati all'ampiezza delle modalità, ed è l'area dei rettangoli ad essere proporzionale alla frequenza.
Se le modalità di una variabile cardinale sono state aggregate in classi di diversa ampiezza, le basi dei rettangoli sono di lunghezza diversa e occorrerà costruire rettangoli aventi altezza proporzionale al rapporto fra frequenza e ampiezza della classe
Se la variabile è ordinale o cardinale, è possibile rappresentarne la distribuzione con un **istogramma di composizione**.
Il grafico è costituito da un rettangolo diviso in fasce di lunghezza proporzionale alle frequenze delle corrispondenti modalità.
Inoltre l'istogramma di composizione, come anche la **spezzata a gradini** sottolinea la natura cumulativa delle frequenze riferite a una variabile ordinale.
Quando le modalità sono particolarmente numerose e la variabile è cardinale, anziché disegnare tanti istogrammi conviene rappresentare ogni frequenza con un punto collocato all'estremità dell'istogramma e congiungere questi punti con segmenti.
Nella collocazione dei punti occorre rispettare la natura cardinale delle modalità e quindi rendere le distanze fra i punti, rilevate lungo la dimensione orizzontale, proporzionali alle distanze fra i valori delle variabili.
Si può rappresentare graficamente anche una distribuzione cumulata di frequenza di una variabile cardinale ricorrendo ad una sorta di poligono di frequenza, ossia **l'OGIVA** in cui i punti corrispondenti alle varie

modalità siano collocati a una distanza dall'origine che sia proporzionale non alla frequenza ma alla frequenza cumulata.
I diversi softwer oggi permettono di creare grafici anche molto articolati con uno sforzo minimo.
Naturalmente le rappresentazioni possono combinare in maniera diversa gli elementi grafici esposti o avvalersi o inventare altri elementi ancora.

4. **Tendenza centrale**

La distribuzione di frequenza è una descrizione relativamente articolata di una variabile, che specifica esattamente quanti casi ricadono in ciascuna categoria della variabile.
Di tutte le ***caratteristiche di una distribuzione di frequenza*** le più importanti sono due:

1. **La TENDENZA CENTRALE** = di una distribuzione è la modalità della relativa variabile verso la quale i casi tendono a gravitare ossia il baricentro della distribuzione.

Quella più elementare è la **moda**, che è la modalità di una variabile alla quale è associata la maggiore frequenza, si tratta quindi di un valore molto povero dal punto di vista informativo, la moda è l'unico valore caratteristico che rileva la tendenza centrale nelle variabili nominali.
Le distribuzioni possono presentare anche sotto-mode, ossia modalità diverse dalla moda che presentano comunque frequenze relativamente elevate, altri baricentri dalla distribuzione, meno importanti del primo ma comunque degli di attenzione.
Per le variabili ordinali è possibile rilevare anche un altro valore caratteristico: la **mediana**.
La mediana di una variabile è dunque la modalità del caso che occupa il posto in mezzo nella distribuzione ordinata dei casi secondo quella variabile.
La determinazione della mediana è molto facile se si consulta una tabulazione che riporta le frequenze cumulate oppure nell'ambito delle rappresentazioni grafiche, un istogramma di composizione o una spezzata a gradini, la mediana infatti corrisponde alla modalità in corrispondenza delle quali le frequenze cumulate superano la soglia del 50%.
La **media aritmetica** è il valore caratteristico più rilevante e noto fra quelli che rilevano la tendenza centrale delle variabili cardinali, ed è data dalla somma dei valori assunti dalla variabile su tutti i casi:

X è la variabile / X- è la media di X / N numero totale dei casi

Naturalmente ha senso avvalersi della **media soltanto se la variabile è cardinale**, in quanto il calcolo richiede che i valori vengano sommati e poi divisi per il numero dei casi: operazioni che si possono effettuare solo se i valori hanno un pieno significato numerico.
Se una distribuzione di frequenza riguarda una variabile cardinale, è possibile determinare moda, mediana e media.
Questi tre però coincidono molto raramente, e in genere conviene avvalersi della media in quanto ogni singolo dato della distribuzione contribuisce a determinare il suo valore.
La mediana è infatti meno sensibile della media ai valori estremi.
Se si vuole eliminare l'effetto distorcente di valori estremi si può prima eliminare un numero predeterminato di dati collocati ai due estremi della distribuzione e poi calcolare la media sui dati rimanenti.
Quando ci si avvale di fonti statistiche ufficiali di solito non si ha accesso a una matrice dati e occorre accontentarsi delle sole rappresentazioni di distribuzioni di frequenze che vengono pubblicate.
Può accedere di aver accesso a una distribuzione riferita a una variabile cardinale in cui le frequenze sono riferite a classi di valori anziché ai singoli valori.

5. **Variabilità**

I valori caratteristici che rilevano la tendenza centrale segnalano il baricentro di una distribuzione di frequenza, ma nulla ci dicono del modo di collocarsi delle altre modalità attorno a questo centro di gravità.
Quindi per descrivere più compiutamente una distribuzione, oltre alla tendenza centrale **occorre anche rilevare la sua variabilità**.

Peraltro questa si manifesta in maniera diversa a seconda del tipo di variabile preso in esame, ed esiste a questo proposito una grande varietà terminologica: si parla talvolta di squilibrio, eterogeneità, dispersione e diversità.
Una variabile nominale presenta una distribuzione caratterizzata da scarsa variabilità quando tutti i casi si addensano nella sua categoria modale, la variabilità minima si ha quando il 100% dei casi assume la medesima modalità, e in questo caso si parla di massima ***omogeneità***.
La distribuzione è massimamente ***eterogenea*** quando i casi sono equi - distribuiti fra le modalità, ossia quando ogni modalità raccoglie esattamente lo stesso numero di casi.

INDICE DI OMOGENEITA' (O)

È un *Indice dato dalla somma dei quadrati delle frequenze proporzionali*.
Il **valore assunto da questo indice dipende da due fattori**: **a)** è tanto più elevato quanto più i casi si raccolgono in poche modalità e **b)** quanto minore è il numero delle modalità.
Il valore massimo è sempre uguale a 1
Assume il minimo valore quando tutte le frequenze sono eguali fra loro e quindi **eguali a 1/k**
Di solito tuttavia è più utile avere un indice di omogeneità che non dipende dal numero di modalità, che permetta di confrontare la variabilità di distribuzione riferite a variabili con un diverso numero di categorie.
In particolare, è utile avere un cosiddetto indice relativo i cui valori presentino un campo di variazione che va da 0 a 1.
Ha un campo di variazione = 1/K :1

INDICE DI OMOGENEITA' RELATIVA (Orel)

Dove O è l'indice di omogeneità assoluta (E=1-O) e K è il numero di modalità
Se i casi vengono divisi in 4 parti di numerosità eguale, le modalità che segnano i confini fra i 4 quarti sono detti **quartili**.
Il primo è il valore al di sotto del quale si trova il 25% dei casi e al di sopra del quale si trova il 75%, il secondo quartile coincide con la mediana, il terzo quartile ha il 75% dei casi sotto si se e il 25% sopra di sé.
Se la distribuzione è molto dispersa, anche il 50% centrale dei casi si distribuirà su un arco piuttosto ampio di valori, e la differenza fra l'1 e 3 quartile sarà elevata, la differenza fra i valori assunti dai due quartili- ossia la differenza interquartile.
Può dunque essere usata per rilevare la dispersione della distribuzione.

Q= Q3-Q1

Se le variabili di cui si analizzano le distribuzioni sono cardinali, una prima idea della loro variabilità si può avere esaminando il loro ***campo di variazione***, ossia la differenza che intercorre fra il valore minimo e il valore massimo.
La media aritmetica ha una caratteristica moto importante di cui ci si avvale per la rilevazione della variabilità delle distribuzioni delle variabili cardinali.
Per ogni valore della distribuzione si può calcolare lo scarto della medi.
Una distribuzione di una variabile cardinale è tanto più dispersa quanto più i suoi casi presentano valori che sono distanti dalla media, ossia quanto più sono grandi gli scarti della media.
Un modo apparentemente semplice e diretto di rilevare la dispersione di una variabile cardinale è calcolare la media aritmetica dell'insieme degli scarti.
Somma degli scarti dei singoli valori della media è sempre eguale a 0
Formula:

Si può aggirare tale problema se degli scarti della media si considera non valore algebrico ma il valore assoluto.
Se si sommano i valori assoluti degli scarti della media e si divide tale somma per il numero dei casi: si ottiene lo **scostamento semplice medio:**

Lo scostamento semplice medio ha un significato immediato e intuitivamente fondato, ma non viene normalmente usato per rilevare la variabilità di distribuzioni riferite a variabili cardinali.
Infatti il problema viene aggirato elevandoli al quadrato.
Si tratta di una soluzione altrettanto efficacie per annullare il segno negativo degli scarti presentati dai valori inferiori alla media.
Inoltre l'elevazione al quadrato degli scarti conferisce anche un maggior peso agli scarti più consistenti.
Se si sommano gli scarti elevati al quadrato, si divide la somma per il numero di casi e poi si estrae la radice quadrata del risultato si ottiene la **DEVIAZONE STANDARD**

Formula:

Il quadrato della deviazione standard è la VARIANZA
La varianza non viene usata nell'ambito dell'analisi monovariata, anche perché si tratta di una cosiddetta grandezza quadratica, che a differenza della deviazione standard, non può essere messa in relazione con grandezze con la media aritmetica.
Tuttavia la varianza presenta alcune caratteristiche matematiche la rendono utile nell'analisi delle relazioni fra variabili.
E' possibile che le due distribuzioni si riferiscano a variabili basate su diverse unità di conto o di misura, e anche qualora le due variabili siano espresse nelle unità di conto/ misura può darsi che i valori di una delle distribuzioni siano di grandezza sensibilmente diversa rispetto ai valori dell'altra distribuzione.
Se si vogliono confrontare fra di loro le variabilità di distribuzioni aventi medie molto diverse, conviene ricorrere a un valore caratteristico che tenga conto della media.
Il **COEFFICIENTE DI VARIAZIONE** fa ciò dividendo la deviazione standard per la media:

Formula

Un'efficace rappresentazione grafica della distribuzione di una variabile cardinale, che veicola in forma compatta informazioni riguardanti diversi valori caratteristici riferiti sia alla tendenza centrale che alla variabilità, è il BOXPLOT.
Quanto più il rettangolo è alto, tanto più la distribuzione è spersa.
La riga orizzontale collocata all'interno del rettangolo designa la media, e l'asterisco corrisponde alla posizione della mediana.
Le asticelle che si estendono al di fuori del rettangolo arriva fino al valore minimo e massimo della distribuzione, e quindi l'altezza complessiva del diagramma corrisponde al campo di variazione.

6. Concentrazione

Le variabili nominali sono massimamente disperse quando la distribuzione è equilibrata, ossia ogni categoria raccoglie lo stesso numero di casi.
Invece una variabile ordinale o cardinale che presenti una distribuzione equilibrata non è affatto massimamente dispersa: la massima dispersione si ottiene quando tutti i casi si dividono, equamente, nelle due categorie.
Questo diverso modo di manifestarsi della variabilità deriva dal fatto che le modalità di una variabile nominale non stanno in alcuna relazione d'ordine, per cui per esse non ha senso parlare di categorie estreme.
Quando la variabile è cardinale e consiste in quantità possedute dei casi di una ricerca, allora può interessare stabilire in che misura tali quantità siano concentrate o al contrario equi - distribuite fra i casi.
La variabile presenta al contrario una concentrazione massima se l'ammontare complessivo A è tutto attribuito a un solo caso.
La concentrazione è un modo particolare di guardare alla variabilità, e ha senso parlarne solo quando la variabile cardinale è interpretabile come quantità o ammontari posseduti dai casi ed è possibile immaginare di trasferire le quantità da un caso all'altro.

Il rapporto di concentrazione di Gini, il più noto; e altri valori caratteristici simili sono usati per studiare le disuguaglianze nella distribuzione della ricchezza, ma possono essere applicati ad altre situazioni di concentrazione/ disuguaglianza.
IL RAPPORTO DI CONCENTRAZIONE DI GINI → si calcola nel seguente modo: si ordinano gli N casi in ordine crescente di valore sulla variabile in esame, e poi si calcolano le proporzioni cumulate dei soggetti e dei redditi. Queste proporzioni vengono designate rispettivamente da pi e qi.
Se non c'è equi - distribuzione, tutti i valori qi sono inferiori ai corrispondenti valori pi.
Si calcola ne seguente modo:

7. Serie territoriali e storiche

Un caso particolarmente importante si da quando **l'unità d'analisi è costituita da un aggregato territoriale**.
Per le unità d'analisi di questo genere ci si avvale spesso di dati tratti da fonti statistiche ufficiali.
Il fatto di avvalersi di fonti ufficiali significa che la fase di raccolta delle informazioni spesso si esaurisce nell'individuazione delle fonti, nella scelta dei dati, nella loro acquisizione e nel loro adattamento ai propri scopi.
Molte variabili riferite ad aggregati territoriali sono di **tipo cardinale.**
Naturalmente, ad aggregati territoriali possono anche essere associate variabili nominali od ordinali.
Le variabili riferite ad aggregati territoriali territoriali e basati su conteggi o misurazioni possono essere estremamente fuorvianti se i loro valori non vengono messi in rapporto con la dimensione della popolazione di tali aggregati.
I dati riferiti ad aggregati territoriali si presentano **spesso in due forme** al fine di studiare l'andamento di un fenomeno nel tempo e nello spazio.
Una serie territoriale è una sequenza di valori assunti da una variabile nello stesso momento in diversi aggregati territoriali.
Una serie storica riporta in sequenza i valori assunti da una variabile su uno stesso aggregato territoriale in tempo diversi.
Per rappresentare graficamente una serie storica, si colloca sull'asse orizzontale del grafico la variabile temporale e sull'asse verticale i valori assunti dalla variabile in esame.
In corrispondenza di ogni periodo la variabile assume un solo valore, e quindi la serie storica si rappresenta come una ***serie di punti uniti da una spezzata***.
Questa congiunzione dei punti è legittima in quanto la variabile temporale ha natura cardinale.
Per rappresentare graficamente una serie territoriale, si ricorre invece a un diagramma a colonne o a una figura analoga, in quanto la variabile regione ha natura nominale.
Una rappresentazione grafica di grande efficacia comunicativa per le serie territoriali è rappresentata dai ***cartogrammi,*** i quali, raffigurando la distribuzione geografica del fenomeno studiato, mettono in luce associazione territoriale che non sarebbero altrettanto percepibili mediante la lettura di una tabella.
Di fronte a una serie storica o territoriale, o anche a una differenza fra due misure, ci si chiede come poter valutare le variazioni di un fenomeno rilevato in situazioni, temporalmente o territorialmente, diverse.
Se indichiamo con [**a]** e [**b]** le due grandezze, possiamo calcolare fra di esse sia **le VARIAZIONI ASSOLUTE che quella RELATIVA**, dove la seconda viene relativizzata dividendola per quello dei due turni che si assume a riferimento.
La variazione è di solito espressa in forma di percentuale.
Nell'analisi delle variazioni relative è quindi bene guardare sempre con attenzione alla base di partenza e diffidare i resoconti in cui vengono presentate solo le variazioni relative senza specificare il termine di riferimento.
Quando si esaminano valori percentuali è importante distinguere fra variazione percentuale e variazione di punti percentuali.
Per esprimere le variazioni nel tempo il ricercatore può estendere il ragionamento sotteso alla distribuzione tra variazione assoluta e relativa e avvalersi dei **numeri indice**.

I numeri indici sono sequenze di valori, assunti da una stessa variabile ma riferiti a periodi diversi, che sono stati relativizzati rispetto a un valore della sequenza convenzionalmente preso come riferimento e posto eguale a 100.
I numeri indice, tuttavia, sono utili non tanto per calcolare la variazione relativa di una rilevazione rispetto un'altra, quanto ***per mettere in luce le variazioni di una serie temporale rispetto a un periodo assunto come base di riferimento***.
La stessa procedura può essere applicata anche alle serie territoriali, dove si assume come base di riferimento il valore assunto sulle variabili dall'insieme degli aggregati territoriali.
I valori assunti numeri indice non dipendono dall'unità di misura o di conto in cui sono espressi, essi sono numeri puri e permettono il confronto fra le variabili più disparate.
I numeri indice permettono quindi di confrontare due serie temporali o territoriali riferite a variabili anche molto diverse fra loro.

Capitolo 3

LA TRASFORMAZIONE DEI DATI

FOCUS: Come i dati possono essere sottoposti ad alcuni tipi di trasformazione al fine di rendere più fruibili le informazioni incorporate nella matrice dei dati.
Il momento conclusivo della rilevazione delle informazioni è rappresentato dalla **MATRICE DATI**.
La matrice dati si può modificare:il ricercatore si trova spesso in una situazione in cui è opportuno estrarre nuove informazioni dalla matrice dati. Mediante un'accorta trasformazione dei dati è possibile generare nuove variabili, che rendono più evidenti alcune informazioni che altrimenti sono più difficili da scorgere (rimane esclusa la possibilità di generare nuovi CASI a partire da quelli esistenti).
Matrice dati modificabile→

- Casi
- Variabili

- Ricodifica
- Aggregazione di valori

La definizione operativa di una variabile può prevedere anche un numero molto elevato di modalità.
Si pensi alla proprietà " popolazione residente "da operativizzare in un'indagine su tutti i comuni italiani.
Se la definizione operativa che crea la variabile" popolazione residente "prevede la registrazione del numero preciso di residenti, i relativi dati rischiano di essere sin troppo carichi di informazione.
Una rappresentazione tabulare o grafica che riproducesse la distribuzione di frequenza con precisione sarebbe troppo articolata per svolgere la sua funzione precipua, ossia offrire una rappresentazione sintetica della distribuzione. Per ovviare a questo problema è ricercatore può adottare due strategie diverse:

1) applicare sin dall'inizio una definizione operativa con relativamente poche modalità. Questa è la fase di aggregazione di valori con poche modalità.

2) la seconda alternativa può essere la ricodifica ovvero la creazione di una nuova variabile che contiene dati generati a partire dalla variabile già esistente.

Le modalità della prima variabile vengono aggregati in un numero minore di intervalli cui corrispondono alle modalità della seconda variabile.
Il codice viene modificato di conseguenza e finisce per elencare una nuova variabile assieme alle corrispondenti modalità.
La nuova variabile, che può essere creata in qualsiasi momento anche dopo anni dalla costruzione della matrice dati, non aggiunge nuove informazioni alla matrice dati, semplicemente rende le informazioni presenti fruibili in maniera diversa.
Si chiama ricodifica perché comporta l'assegnazione di nuovi valori, o sia codici, alle modalità di una variabile.
Ma quale delle due strategie conviene adottare?
La seconda strategia da al ricercatore maggiori possibilità di scelta in quanto può servirsi dell'una o dell'altra variabile: se vuole esprimere la tendenza centrale con la media o la dispersione con la deviazione standard se avverrà della prima variabile, se vuole descrivere la distribuzione con una tabella un diagramma a barre fari corsa la seconda.
Se invece adotta la prima strategia che prevede la definizione operativa riduttiva sin dall'inizio, per rappresentare la tendenza centrale dovrà accontentarsi della mediana o della moda o della media stimata.

Molte variabili cardinali presentano un elevato numero di modalità e di solito non si è interessati a conoscere con precisione la frequenza associata a ciascuna modalità piuttosto si interessate alla sintesi della rappresentazione della distribuzione.
L'opportunità di aggregare può manifestarsi anche in relazione a variabili nominali o comunque con poche modalità: se alcune modalità presentano una frequenza molto bassa può essere opportuno raggrupparle in una sola modalità.
Sul piano pratico conviene considerare l'aggregazione di valori come un modo per creare variabili aggiuntive.
L'analisi statistica dei dati può richiedere un'operazione di ricodifica anche in altre situazioni quando ad esempio il ricercatore vuole trattare una variabile ordinale come se fosse cardinale.

2. OMOGENEIZZAZIONE DEI CAMPI DI VARIAZIONE

La normalizzazione consiste nella trasformazione di un insieme di valori numerici al fine di collocarli in un sistema di riferimento che ne facilità l'interpretazione.
La trasformazione delle frequenze assolute in frequenze percentuali colloca tali frequenze in un sistema in cui si sa priori che valori variano da zero un massimo di 100 il che gli permette di interpretare in termini relativamente univoci l'incidenza di una determinata modalità in una distribuzione di frequenza.
Allo stesso modo infatti la trasformazione di una frequenza assoluta in una proporzione colloca quest'ultima in un sistema di riferimento in cui valori variano da 0 a 1.
Una forma semplice di normalizzazione consiste nel mettere in relazione i valori di una variabile cardinale con il valore più basso e il valore più alto che si possano assumere su quella variabile.
ESEMPIO: Nel sistema universitario italiano le votazioni conseguibili in sede di esame di laurea possono variare fra un minimo di 66 è un massimo di 110, questo intervallo costituisce il sistema di riferimento per la normalizzazione.
Un qualsiasi voto di laurea ad esempio può essere trasformato in un valore normalizzato applicando la seguente equazione:

$$Ni = (Xi - Xmin) / (Xmax - X min)$$

Ni= dato normalizzato
Xi= dato da normalizzare
Xmin=valore minimo possibile per la variabile in questione.
Xmax=valore massimo.

Il ricercatore può anche preferire un campo di variazione dei valori della nuova variabile diverso da zero e uno. In questo caso è sufficiente adottare la seguente funzione di trasformazione:

$$Ni = K * [(Xi - Xmin) / (Xmax - X min)]$$

In cui K è uguale al nuovo Massimo che il ricercatore desidera assegnare alla scala dei valori possibili. Se K viene posto = 100, il risultatante sistema di riferimento prevede valori che variano fra un minimo di 0 e un massimo di 100.
La trasformazione del campo di variazione può essere utile qualora il ricercatore voglia confrontare i dati appartenenti a due o più variabili che presentano intervalli di variazione diverse.
É sufficiente applicare la funzione di trasformazione a tutte le variabili in modo tale da ricondurle tutte ad uno stesso sistema di riferimento.

Trasformazione del campo di variazione --> normalizzazione --> confrontare

Ma... La forma appena descritta di normalizzazione, che potremmo chiamare "assoluta "non tiene in alcun conto dell'effettiva distribuzione di frequenza dei dati da normalizzare: l'effettivo campo di variazione determinato dal valore minimo e dal valore massimo teoricamente possibili, non dal minimo ed al massimo effettivamente rilevati, per cui tale campo può essere anche sensibilmente più ristretto di 0-k.
Se si vuole sfruttare appieno l'intervallo di variazione desiderato si può operare una normalizzazione relativa, dove nelle formule precedentemente illustrate Xmin è uguale al valore più basso effettivamente rilevato e x-max è uguale al valore più elevato effettivamente rilevato.
La normalizzazione relativa ha il vantaggio di incorporare alcune informazioni empiriche circa la dispersione dei dati, tale per cui si ha la certezza che sulla nuova variabile almeno un caso assumerà il valore 0 e almeno un caso valore K.

3. STANDARDIZZAZIONE

La normalizzazione relativa costituisce il risultato è così poco soddisfacenti che so ancora già distribuzione empirica relativamente debole: si basa soltanto sul valore minimo e su quello massimo, senza considerare come gli altri valori si distribuiscono entro con l'intervallo.
La standardizzazione è una procedura di normalizzazione particolarmente utile nell'analisi di variabili cardinali.
Essa trasforma i dati originari in punti standard che non risentano né dell'unità di misura né della variabile, né della tendenza centrale della distribuzione né della sua dispersione.
La standardizzazione consiste in una doppia trasformazione: prima si normalizzano i dati rispetto alla loro media poi si normalizzano i risultati scarti rispetto alla deviazione standard.
La prima normalizzazione consiste nella trasformazione di ogni valore nel suo scarto dalla media.
La seconda normalizzazione consiste nella compressione o nella dilatazione della distribuzione dei punteggi, a seconda della sua dispersione.
In particolare oggi scarto ottenuto con la prima normalizzazione viene diviso per la deviazione standard.
Abbiamo dunque fatto due operazioni sui punteggi originari: abbiamo loro sottratto la media della distribuzione e abbiamo diviso questa differenza per la deviazione standard della distribuzione:

$$Zi= (Xi - X)/ S$$

L' unità di misura di una distribuzione standardizzata si chiama ***punto standard.***
Il punto standard della deviazione standard della distribuzione originaria.
La nuova variabile Z presenta alcune caratteristiche: la media = 0 e la deviazione standard è uguale a 1.
Tutti i dati a appartenente distribuzioni standardizzate sono confrontabili fra di loro, in quanto hanno la stessa unità di misura quindi il punto standard, la stessa tendenza centrale e la stessa di dispersione.
La standardizzazione permette di confrontare dati appartenenti a variabili basate su una stessa definizione operativa ma con distribuzioni empiriche diverse.
I confronti basati sulla standardizzazione si servono esclusivamente di informazioni già contenute nella matrice dati, non servono informazioni esterne.

4. TRATTAMENTO DEGLI STILI DI RISPOSTA

Le **proprietà** che non sono state ideate con unità di misura, sono proprietà che hanno per oggetto delle grandezze psicologiche e che vengono concettualizzate come atteggiamenti e opinioni di individui.
Spesso per operativizzare queste proprietà si ricorre a ***tecniche di scaling.***
Uno dei motivi per cui le **variabili** create in questo modo si dicono **"quasi cardinali"** è che le rilevazioni non vengono espresse in unità di misura vere e proprio.
La fedeltà dei dati raccolti con tecniche di scaling può essere inficiata della presenza di cosiddetti ***stili di risposta***, cioè dalla propensione degli intervistati ad assegnare punteggi in base a fattori che non hanno a che fare con il loro grado di favore per gli elementi sottoposti a valutazione.
Lo stile più noto è l'acquiescenza, ossia la tendenza ad essere favorevoli, ovvero assegnare punteggi elevati ea prescindere dall'oggetto.
Dalla matrice dati emergono diversi stili di risposta.
La media dei punteggi dati da un singolo individuo agli elementi della batteria rileva la tendenza centrale, la deviazione standard rileva la dispersione dei punteggi.
Con queste due informazioni si può applicare una **procedura chiamata deflazione.**
Se calcoliamo la media dei punteggi questi ultimi possono essere normalizzati.
Questa trasformazione fa sì che l'insieme dei punteggi espressi da uno stesso caso presenti in una media pari a zero.
A questo punto -> dividiamo ogni punteggio normalizzato per la deviazione standard dell'insieme dei punteggi espressi da ogni soggetto, introduciamo una seconda normalizzazione che tiene conto di quanto ogni punteggio si allontana dalla media dei punteggi espressi dal soggetto.
I punteggi così deflazionati sembrano ancora più fedeli.
La DEFLAZIONE, così è una doppia normalizzazione dei punteggi.
La deflazione ricorda la standardizzazione ma

Sono uguali per:

1- prevedono la sottrazione di un dato da una media.
2- la sua divisione per una deviazione standard.

3- infatti i punteggi deflazionati di uno stesso caso presentano una media eguale a 0 e una deviazione standard pari a 1.

Si differisce per:

A) la standardizzazione opera per colonna e trasforma tutti i dati di una colonna.

B) la deflazione opera invece per riga e trasforma soltanto i dati riferiti ad una stessa batteria.

La media e la deviazione standard che vengono inclusi nell' algoritmo di deflazione ai riferiscono non ai dati di tutti i casi su un' unica variabile, bensì ai dati di un caso su un insieme di più variabili.

5. INDICI E TIPOLOGIE

Molte ***proprietà*** sono complesse generali e che per operativizzarle si individuano proprietà più specifiche denominate ***INDICATORI*** a partire dalle quali è più agevole costruire una ***definizione operativa***.

Tuttavia ciascuna di queste definizioni operative da luogo una variabile distinta.

Di norma si desidera in tale situazione, ricondurre **l'insieme di variabili** in un unica variabile, quindi di tornare alla proprietà generale dopo che questa è stata frazionata semplificata in un certo numero di proprietà più semplice.

La nuova variabile di sintesi viene in genere chiamata ***INDICE.***

In generale l'indice é una variabile costruita combinando altre variabili preesistenti, che riassume le informazioni veicolate da queste ultime.

Le procedure per la costruzione di un indice variano a seconda della natura delle variabili da combinare: cardinali o nominali.

1) costruzione di un indice con variabili CARDINALI

Dalle le variabili che vengono registrate nella matrice dati è possibile creare una nuova variabile calcolando per ogni caso la somma dei punteggi conseguiti su ciascuna delle 6 variabili preesistenti.

Ii= V1+V2...vi

I è uguale al punteggio attribuito sull'indice, v1 è il dato registrato per la prima variabile.

In primo luogo il ricercatore deve prendere la decisione di trattare tutte le variabili come se fossero cardinali.

In secondo luogo va notato che le variabili non hanno tutte le stesso orientamento semantico dunque occorre invertire i punteggi, e quindi occorre effettuare un'inversione di polarità semantica, oppure alternativamente attribuire un segno negativo ad alcune variabili.

In terzo luogo bisogna chiedersi se le variabili non condividono lo stesso campo di variazione e non sono espresse con la stessa unità di misura occorre omogeneizzare i loro campi di variazione.

Questa uniformità può essere ottenuta operando ricodifiche.

Come quarta fase bisogna chiedersi se le variabili hanno la stessa validità. la validità di un indicatore dipende dalla sua strettezza del rapporto semantico che legga ciascuno dei concetti specifici alla proprietà generale che gli interessa.

Se non è così allora conviene attribuire pesi differenziati alle variabili.

2) costruzione di un indice a partire da variabili NOMINALI.

Se si vuole costruire un indice a partire da variabili nominali e le operazioni di combinazione delle variabili devono essere di tipo logico anziché matematico.

A partire da due variabili già presenti in matrice dati si incrociano, quindi le possibili combinazioni sarebbero la moltiplicazione tra le due variabili.

Le possibili combinazioni sarebbero ancora più numerose, infatti questa combinazione da luogo un cosiddetto ***spazio degli attributi*** e il ricercatore spesso si prefigge di ridurlo per renderlo meno complesse più maneggevole.

Come? Ad esempio aggregando in un unico tipo le categorie.

Gli schemi di classificazione che risultano dalla combinazione di due o più schemi preesistenti si chiamano anche tipologie e questo termine viene preferito a di indice quando le variabili combinate sono nominali.

6. RAPPORTI STATISTICI E ALTRE FORME DI COMBINAZIONI FRA VARIABILI.

Le variabili cardinali derivano dalla combinazione di altre variabili cardinali, si chiamano ***variabili derivate.***

Il ricorso a variabili derivate è particolarmente diffuso quando l'unità di analisi è un aggregato territoriale.

In genere quando ci si trova in situazioni in cui valori assunti da una variabile risentono della diversa diminuzione degli aggregati occorre relativizzare tali valori in funzione delle rispettive basi di riferimento calcolando il **rapporto statistico**.

Tali rapporti permettono di effettuare comparazioni un tempo nello spazio fra situazioni diverse.
Ne esistono vari tipi:

1) I rapporti di composizione

Mettono in relazione la parte al tutto, consiste nel mettere relazioni una parte di un fenomeno al fenomeno stesso nella sua totalità.

2) Un rapporto di coesistenza

É un rapporto fra due parti di un insieme, o sia il rapporto fra la frequenza di una modalità è la frequenza di un'altra.

3) Un rapporto di derivazione

Corrisponde al rapporto fra la misura di un fenomeno è quella di un altro che può essere considerato un suo presupposto necessario.

4)Rapporti medi

Le due grandezze messe in relazione a tengono due fenomeni diversi. I rapporti Medi sono una sorta di categoria residua che raccoglie rapporti che non ricadono nei casi precedenti.
Nelle rilevazioni basate sulla somministrazione di un questionario strutturato spesso ci si avvale di domande multirisposta.
Spesso la trasformazione di una variabile nominale insieme di variabili dicotomiche viene effettuata in vista dell'analisi multivariata, infatti le risultanti variabili dette anche **dummy** sono dicotomiche e quindi trattabili come se fossero cardinali mentre la variabile originale non lo è.

Capitolo 4

ANALISI BIVARIATA

1. UNA VARIABILE NON BASTA

Un ricercatore raccoglie informazioni su un ampio ventaglio di proprietà delle unità che fanno parte del suo campione, e c'ho per **motivi economici che scientifici**.
I motivi economici hanno a che fare con le risorse impiegate mentre i motivi scientifici spingono a raccogliere informazioni su molte proprietà, ossia prevedere molte variabili nella matrice dati.
Per quanto l'oggetto cognitivo di una ricerca possa sembrare semplice, spesso è possibile enucleare diverse proprietà da operativizzare al fine di descriverlo adeguatamente in termini empirici.
Dunque converrà rilevare informazioni anche su altri proprietà.
In altre parole per giungere ad una vera conoscenza non superficiale dell'oggetto cognitivo di un'indagine, il ricercatore deve raccogliere informazioni e analizzare dati riguardo un ampio numero di caratteristiche che non hanno alcun legame semantico con l'oggetto cognitivo.

2. IPOTESI

Un' ipotesi di solito viene espressa mediante una proposizione che mette in relazione due o più variabili.
Va sottolineato che le ipotesi spesso non forniscono indicazioni precise circa il modo in cui le variabili devono essere definite operativamente.
Naturalmente questa ambiguità può essere circoscritto a formulando le ipotesi in maniera più puntuale.
Il controllo empirico di un'ipotesi implica la specificazione anche di altri suoi elementi chiave.
Le ipotesi possono essere formulate sia prima sia dopo la raccolta delle informazioni che verranno sottoposte ad analisi.
Questa distinzione temporale importante: se una determinata variabile non compare nella matrice dati, è evidente che essa non può essere inserita in alcuni protesi.
Le ipotesi possono anche essere più articolate e chiamare in causa anche più variabili.
L'analisi delle relazioni fra le due variabili è soltanto una fase intermedia in vista di analisi più complesse ed esaurienti.

3. DISTRIBUZIONE DI FREQUENZA CONGIUNTA

L'analisi statistica delle relazioni fra due variabili si basa sull'esame delle ***distribuzioni di frequenza congiunte***.
Una distribuzione congiunta è semplicemente l'incrocio di due o più distribuzioni di frequenza semplici quindi monovariate.
La distribuzione congiunta si distingue da una semplice distribuzione monovariata in quanto l'assegnazione di ogni caso ad una cella di frequenza tiene conto dei valori assunti su più variabili.

La distribuzione congiunta si è compatibile con le due distribuzioni ma lo sono anche tante altre distribuzioni.
Le cosiddette **frequenze marginali** Ossia le frequenze che compaiono nella riga totale, nella colonna totale devono essere identiche a quelle delle distribuzioni monovariate ,per il resto una volta che è stato rispettato questo vincolo la distribuzione congiunta può assumere qualsiasi conformazione.
Per ora è sufficiente sottolineare che la conoscenza delle distribuzioni monovariata e di due variabili ci permette di dire molto poco sulla distribuzione congiunta delle stesse due variabili.

4. FORMA, FORZA E DIREZIONE

Il controllo empirico di un'ipotesi consiste nell'esame di una distribuzione congiunta.
Con le tecniche di analisi bivariata, si cerca di individuare quali siano la forma, la forza e la direzione delle relazioni fa due variabili.

FORMA:

Con forma si intende le possibili forme della relazione fra le due variabili in questione.
In alcuni casi l'esame di una relazione permette di individuare neanche il segno, il quale è una componente della sua forma.
Perché sia possibile **parlare di segno** occorre che le modalità delle variabili messe in relazione siano ordinate lungo da qualche dimensione.
In altre parole entrambe le variabili devono essere ordinali o cardinali.

FORZA:

L'accertamento nella forma di una relazione non esaurisce certo la conoscenza che possiamo estrarre dalla matrice dati.
Le distribuzioni congiunte manifestano altrettante relazioni bivariate che condividono la stessa forma.
Tuttavia la forza delle relazioni varia sensibilmente nella distribuzioni.
In poche parole la forza è conoscere lo stato di un caso su una variabile la quale permette di prevedere il suo stato sull'altra variabile soltanto con un margine di errore relativamente ampio.
Così come possibile immaginare relazioni aventi la stessa forma ma forza diversa, sono possibili relazioni avente la stessa forza ma diverse forme.

DIREZIONE:

L'esistenza di una relazione fra due variabili, la sua forza, la sua forma e il suo eventuale segno sono caratteristiche che possono essere accertate mediante l'uso di appropriate tecniche di analisi statistica alle quali sono in ultima analisi riconducibili a il canone delle variazioni concomitanti ->

"qualunque fenomeno che va in un modo qualsiasi ogni volta che un altro fenomeno varia in qualche modo particolare è una causa un effetto di quel fenomeno o è connesso a quel fenomeno mediante qualche fatto di causazione".

Dalla citazione si evince il **PRINCIPIO DI CAUSAZIONE**, infatti sapere se Intercorre un nesso di causalità fra le variabili e quale variabile influenza le altre è detto altrimenti ***direzione causale della relazione***.
La spiegazione causale è un obiettivo centrale nella ricerca sociale.
Il rapporto causa-effetto ci si riferisce all'esistenza di un nesso fra eventi tale per cui la manifestazione di un determinato evento è la conseguenza diretta e necessaria della manifestazione di un altro evento.
Per poter qualificare come causale una relazione fra due eventi occorre anche spiegare perché un evento è implicato dall'altro, ossia specificare quale sia il meccanismo causale, attraverso quali processi la supposta causa produce il supposto effetto.
Quasi sempre le relazioni fra variabili hanno un carattere solo tendenziali.
L'individuazione di una direzione causale è resa difficile dal fatto che molte relazioni sono bidirezionali, ossia le variabili si influenzano reciprocamente.
Peraltro fra relazioni bidirezionali è possibile distinguere **relazioni simmetriche** (in cui le relazioni fra variabili si influenzano reciprocamente)con le **asimmetriche**(in cui l'influenza esercitata da una variabile sull'altra è comunque Maggiore dell'influenza che subisce).
Spesso le relazioni bidirezionali asimmetriche vengono trattate come se fossero unidirezionali.
Spesso è agevole individuare l'esistenza di una relazione e persino stabilirne la direzione causale, ma non è per questo facile ricostruire il meccanismo causale. Infine occorre tenere conto che qualsiasi variabile dipendente dipende da molti elementi al di là di quelli che ricercatore in grado di inserire in una relazione fra variabili.

Quando si parla di un rapporto causa-effetto si farà spesso riferimento alla forma, forza e direzione di una relazione fra variabili.

5. VARIABILE INDIPENDENTE E DIPENDENTE

L'attribuzione di una direzione causale implica l'attribuzione degli aggettivi dipendente e indipendente alle variabili messe in relazione: la variabile indipendente influisce sulla variabile dipendente senza esserne a sua volta influenzata.
Lo status dipendente o indipendente di una variabile può cambiare a seconda della variabile con cui viene messa relazione.

X ->VARIABILE INDIPENDENTE
Y-> VARIABILE DIPENDENTE
=Relazione bivariata

L'individuazione della direzione di una relazione poggia sulle conoscenze del contesto sulle competenze interpretative possedute da ricercatore.
Ci sono alcune proprietà che vengono operativizzate in quasi tutte le ricerche e che corrispondono a caratteristiche di norme mutabili: sesso o genere, anno di nascita, luogo di nascita, identità etnica.. e le risultanti variabili non possono che essere indipendenti rispetto ad altre variabili.
In analisi bivariata la scelta delle tecniche da adottare dipende dal tipo di variabili (nominali cardinali,ordinali)analizzate.
Inoltre va ricordato che le variabili ordinali occupano una posizione intermedia rispetto alle variabili nominali da una parte e quelle cardinali dall'altra.
In sede di analisi e multivariata le variabili ordinali vengono trattate come se fossero cardinali o nominali.
Di solito le variabili ordinali vengono trattate alla stregua di quelle nominali e i due tipi di variabili vengono ricondotti sotto le espressioni ***variabili categoriali***.
Dunque la scelta delle tecniche di analisi dipende dalla natura cardinale o categoriale delle variabili.

Capitolo 5

ANALISI BIVARIATA: QUANDO LE VARIABILI SONO CATEGORIALI (TABULAZIONE INCROCIATA)

1. Tabelle a doppia entrata

Dopo aver trattato le analisi delle variabili, è giunto il momento di passare al cuore dell'analisi dei dati: lo studio delle relazioni fra variabili. Le variabili considerate siano solo due (analisi bivariata) e siano entrambe categoriali.
Per fare ciò organizziamo i dati in una **tabella** che chiamiamo a **DOPPIA ENTRATA** (detta anche incrocio o tabulazione incrociata), nella quale collochiamo in riga una variabile(**variabile di riga**), in colonna l'altra (**variabile di colonna)** e *nelle celle*, definite dall'incrocio fra le righe e le colonne, il numero di casi che presentano le corrispondenti modalità delle due variabili (**frequenze**).
Talvolta alla tabella così espressa vengono anche aggiunti i totali di riga e di colonna delle frequenze, che chiamiamo frequenze marginali, o più brevemente marginali e che corrispondono alla frequenze delle variabili singolarmente prese, come nell' analisi monovariata.
Su queste frequenze assolute si possono effettuare tra diversi tipi di percentualizzazione:
- percentuali di riga
- percentuali di colonna
- percentuali sul totale (percentualizzare tutte le frequenze di cella sul totale generale).
La tabella sulle frequenze assolute e sulle percentuali sul totale non sono utili, rimangono le percentuali per riga e le percentuali per Colonna, una delle due serve per analizzare la relazione fra le due variabili, l'altra va scartata...***come rientrare su questa scelta?***
La scelta della percentuale sbagliata può portare il ricercatore completamente fuori strada.
La percentuale sbagliata infatti invece di pareggiare le base di riferimento riporta le differenze nella popolazione: non funziona quindi più come percentuale e se viene utilizzata per analizzare la relazione fra due variabili può condurre a gravi errori.
Per scegliere fra percentuali di riga e percentuali di colonna mi sono dei ***criteri guida***:
1) si sceglie la percentuale di colonna quando si vuole analizzare l'influenza che la variabile poste in colonna a sulla variabile posta in riga.

2)si sceglie la percentuale di riga quando si vuole analizzare l'influenza che la variabile posta in riga ha sulla variabile posta in colonna.
Detto diversamente: si definisce qual è la variabile indipendente e si percentualizzata all'interno delle sue modalità.
Innanzitutto il principio della percentualizzazione all'interno delle modalità della variabile indipendente rimane il criterio guida quando il nostro obiettivo è quello di **studiare la relazione causale esistente fra variabile dipendente e indipendente**, ma le altre situazioni può essere utile calcolare le percentuali nell'altra direzione.
In questo modo otteniamo quello che potremmo chiamare ***un profilo***, questo modo del confronto fra i profili delle modalità della variabile dipendente e il profilo dell'intera popolazione per stabilire la relazione tra le variabili della tabella è più laborioso e meno diretto del precedente.
Tuttavia in certi casi esso rappresenta l'unica via disposizione e cioè quando l'indagine condotta su tutti i casi ma solo su un sottoinsieme della popolazione.
Ma può aver senso calcolare percentuali sia per riga sia per Colonna quando non è possibile individuare con chiarezza una variabile dipendente o indipendente in quanto la direzione causale può andare in un verso come nell'altro.
Va chiarito che quando sono chiaramente individuabili le variabili indipendente dipendente e l'obiettivo è quello di controllare indirettamente l'esistenza di un nesso causale fra le due variabile, il criterio da assumere quello delle percentuali entro le modalità della variabile indipendente.

• Come si presentano le tabelle?

Una tabella efficiente, completa e adeguata:
-Va riportata solo la percentualizzazione che serve, quella di riga oppure quella di colonna.
-non è necessario presentare le frequenze assolute.
- è utile che ogni riga o colonna percentuale finisca col totale 100, che agevola la lettura della tabella.
- È indispensabile riportare per ogni colonna (riga)la base delle percentuali (N) sulle quali Esse sono state calcolate.
-E' assai imprudente calcolare commentare percentuali su base inferiore a 50 casi, questa limitazione può comportare interventi di aggregazione sulle modalità delle variabili indipendenti.
Infatti sono le modalità della variabile indipendente a fungere da base per la percentuale.
-le tabelle devono essere sempre intestate punto e infatti importante è che la tabella sia autoesplicativa che essa cioè contenga tutte le informazioni necessarie per la comprensione.
Va ricordato che quando le variabili derivano da questionari, è indispensabile che il lettore sia messo al corrente delle esatta formulazione della domanda: riportare il testo della domanda nell'intestazione della tabella.
- Infine va ricordato che la somma di percentuali e legittima se i valori sommati appartengono alla stessa distribuzione percentuale, ma errata se le percentuali sommate appartengono a due diverse distribuzioni.

• Come si interpretano le tabelle?

Nell'interpretazione e commento delle tabelle, si suggerisce di selezionare la modalità più significative della variabile dipendente e centrare su di questa è l'analisi.
Si consiglia anche di non dare troppo rileva differenze percentuali esigue: in linea molto generale nel coso di inchiesta campionaria, differenza tra percentuali perché sia degna di nota deve essere superiore ai 5 punti percentuali.
Dopo aver fatto la effettuato l'aggregazione,calcolato le percentuali, otteniamo una tabella che quella che dovremmo andare interpretare.
Il commento di una tabella deve rispondere all'interrogativo: c'è una relazione fra le due variabili?
Un commento non deve essere semplicemente descrittivo, questi commenti non servono praticamente a niente in quanto non fanno un confronto fra le due variabili.
Un commento deve invece prendere una modalità significativa della variabile dipendente vedere come la sua percentuale varia tra i varia.
Es:" possiamo affermare che l' insoddisfazione cresce col crescere della grandezza delle citta".
Se la variabile dipendente ordina le virgole spesso di grande utilità aggregare le modalità estreme e contigue della variabile dipendente.
Questo modo di procedere comporta di solito è una notevole pulizia della relazione.

Un modo abbastanza utile per interpretare la tabella consiste nel calcolare la differenza fra due modalità di risposta, oppure fra le risposte positive quelle negative.
Questa differenza viene chiamata ***indice di differenza percentuale***, e permette di leggere i dati tenendo conto simultaneamente dall'andamento di più modalità nella variabile dipendente.
Facendo così eliminiamo l'effetto della categoria neutra intermedia che spesso ha un andamento fluttuante e disturba l' analisi.
L'andamento della relazione emerge con chiarezza, il segno dell'indice conferisce ancora maggiore visibilità all'andamento e facilita il commento.

2. Tabelle a doppia entrata particolari

• tavole di mobilità sociale

All'interno delle tabulazioni incrociate, un caso di particolare interesse è dato dalle **tavole di mobilità sociale,** nelle quali su una dimensione si colloca la classe sociale dei soggetti studiati e sull'altra quella dei loro padri.
La sua struttura è quella di una normale tabella doppia entrata: essa tuttavia presenta delle particolarità che derivano dalle molteplici linee di lettura che offre.
Va segnalato il **significato che assumono le celle**: poiché le due variabili hanno le stesse modalità, nelle celle sulla diagonale si trovano i soggetti cosiddetti immobili, mentre nelle celle fuori dalla diagonale si collocano i soggetti mobili: se la variabile classe sociale una variabile ordinale nel triangolo sopra la diagonale abbiamo i soggetti che hanno sperimentato un processo di mobilità ascendente è nel triangolo sotto la diagonale coloro che hanno invece sperimentato un processo di mobilità sociale discendente.
Diversamente da quanto accade con le normali tabella doppia entrata, in questo caso **tutte e tre le forme di percentualizzazione assumono un precisa significato.**
Le percentuali entro le modalità della variabile indipendente ci dicono qual è l'influenza della classe sociale di partenza su quella d'arrivo, le percentuali per riga ci danno informazioni sulle origini sociali dei ceti attuali.
Mentre la terza percentualizzazione quella sul totale ci dà informazioni sul processo generale di mobilità sociale.
Sommando le percentuali delle Celle sulla diagonale del triangolo superiore della matrice del triangolo inferiore otteniamo rispettivamente i tassi di immobilità sociale, mobilità ascendente e di mobilità discendente.

• tabulazione incrociata nel caso di più di due variabili

Data la complessità di una tabella, per poterla raffigurare abbiamo presentato solo una modalità della variabile dipendente.
Naturalmente non c'è più in questa un totale che fa 100.
Ancora per semplificare la tabella e soprattutto per rendere più il numero delle basi delle percentuali le due variabili indipendente sono state dicotomizzate.

3. Rappresentazioni grafiche della relazione fra due variabili

È utile rappresentare **graficamente la relazione fra due variabili nominali**.
Per fare ciò si utilizzano i **diagrammi a barre oppure le spezzate**.
Si riporta in un piano cartesiano, sull'asse orizzontale e le modalità della variabile indipendente e sull'asse verticale le frequenze percentuali relative alla modalità della variabile dipendente che abbiamo scelto come più rilevante.
-Se la variabile indipendente è nominale si può utilizzare solo il diagramma colonne.
- Se la variabile è ordinale oppure cardinale è raggruppata per classi possiamo anche rappresentare la relazione della tabella tramite una spezzata.

4. Misure di forza della relazione

Ma esiste un modo meno impressionistico più aggettivo per misurare la forza della relazione fra due variabili nominali ordinali?

Distingueremo la trattazione a seconda che la relazione sia fra variabili nominali, per le quali parleremo di ***misure di associazione,*** o fra variabili ordinali e va in questo si parlerà di ***misure di cograduazione***.

- **frequenze osservate e frequenze attese**

Quale forma super ebbe la tabella in caso di indipendenza fra le variabili? esattamente calcolare la differenza fra le frequenze attese sotto l'ipotesi di indipendenza e le frequenze osservate effettivamente nei dati.

Quando abbiamo indipendenza? quando le percentuali di variabile dipendente e variabile indipendente solo uguali in tutte le categorie, e quindi sono uguali a quelle sul totale della popolazione.

$$F_e \text{ (frequenza attesa di ogni cella)} = \frac{marginale\ X\ marginale}{totale\ della\ tabella}$$

In questo modo abbiamo calcolato le frequenze attese, si tratta a questo punto di calcolare la ***differenza fra la tabella delle frequenze osservate e quella delle frequenze attese sotto le ipotesi di indipendenza***.

Questa differenza viene sintetizzata in un' unica misura → **chi-quadrato** (X^2)

Per ogni cella facciamo la differenza fra frequenza osservata e frequenza attesa, eleviamo al quadrato e la dividiamo per la frequenza attesa.

Infine sommiamo per tutte le celle questi valori, sintetizzando così in un unico numero le differenze fra le celle punto il calcolo di chi quadrato viene effettuato sulle frequenze assolute, non sulle percentuali.

$$X^2=\sum \frac{(f0-fe)^2}{fe}$$

Il valore di chi quadrato assumere il valore zero nel caso limite di indipendenza perfetta nei dati, mentre sarà tanto più elevato quanto maggiore è la distanza fra frequenza osservate e frequenze attese, cioè tanto più le frequenze osservate si allontanano dalle ipotesi di indipendenza.

Il valore della chi quadrato non può essere utilizzato direttamente come misura della forza della relazione fra due variabili nominali perché dipende dalla numerosità dei dati della tabella.

Se il campione raddoppia, il valore del chi quadrato raddoppia, se triplica anche il chi quadrato triplica.

Lo statistico svedese **Harald Cramèr** ha proposto il seguente **indice V (V di Cramèr**); che assume valori fra 0 (indipendenza) e 1 (relazione perfetta).

$$V= \frac{\sqrt{X^2}}{\sqrt{N(k-1)}}$$

- X^2 è normalizzato
- Dove K è uguale al numero di modalità della variabile con il minore numero di modalità.

- **Misure di associazione fra variabili dicotomiche**

Se le due variabili messe relazioni sono entrambi dicotomiche, allora V coincide con ***il coefficiente di correlazione r di Pearson***, che è una misura da utilizzare quando entrambe le variabili sono cardinali, che si può calcolare anche quando le variabili sono dicotomiche.

In queste tabelle 2x2, in cui V= r abbiamo:

$$X^2= \frac{N(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

Di cui

$$X^2= \sqrt{\frac{X^2}{N}} = \frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = r$$

- Dove a,b,c,d sono le frequenze della tabella.

Nel caso della tabella 2x2 una misura semplicissima di associazione è costituita da ***d (differenza fra le proporzioni).***

Questo indice può variare fra -1 e +1 (-1<d<1), col valore 0 che sta a significare assenza di relazione fra le due variabili.

L' indice ha un' immediata interpretazione: il suo valore è pari al coefficiente di regressione tra le due variabili quando esse sono espresse nei valori 0 e 1.

• Misure di associazione basate sulla riduzione proporzionale dell' errore

Se due variabili X e Y sono indipendenti, il fatto di conoscere per una certa unità d' analisi il valore assunto su di essa da X non ci aiuta a predire il valore assunto sulla stessa da Y; all' opposto, se sono perfettamente associate, il fatto di conoscere X ci permette di predire senza errore il valore di Y.
La misura di associazione corrisponde alla proporzione di riduzione degli ***errori di previsione (Pre):***

Misura di associazione Pre= $\frac{(E1-E2)}{E1}$

E1= numero di errori nel prevedere Y che si commettono senza conoscere X.
E2= numero di errori che si commettono conoscendo X.
Sono state proposte diverse misure di associazione per variabili nominali basate su questo criterio.
Le più note sono:

- (Lettera greca lambda)
- (Lettera greca tau)

Queste misure assumono valore differente a seconda di quale delle due variabili viene scelta come variabile dipendente.
Per questo motivo vengono chiamate misure di associazione asimmetriche.

• Misure di cograduazione

Se la relazione fa una variabile nominale è un ordinare, useremo le stesse misure di associazione che abbiamo appena presentato per il caso di due variabili nominali.
Se le variabili sono entrambe ordinali, possiamo sempre impiegare le nostre misure di associazione viste per le variabili nominali, ma possiamo anche utilizzarle delle nuove.
In questo caso parliamo di ***misure di cograduazione***: quando la relazione è tra variabili ordinali essa assume anche un segno, una relazione si dice positiva se a valori alti di una variabile tendono a corrispondere valori alti dell'altra, si dice negativa se a valori alti dell'una tendono a corrispondere valori bassi dall'altra.
Nel caso di una tabella doppia entrata fra variabili ordinali, la presenza di una relazione fa sì che le frequenze si addensi ma lungo una diagonale:
- diagonale principale in caso di relazione positiva.
- diagonale secondaria in caso di relazione negativa.

Sono state proposte **diverse misure di cograduazione**: e se cioè si basano su un confronto fra i valori assunti dalle variabili X e Y su tutte le possibili coppie di casi.
Una coppia di casi è detta concordante se su un caso di valore di X e di Y sono entrambi maggiori o minori dei valori delle stesse variabili sull'altro caso.
Una coppia è detta discordante se una variabile assume su un caso un valore maggiore mentre l'altro un valore minore rispetto ai valori assunti sull'altro caso.
Se ci sono tante coppie concordante quante discordanti allora non c'è cograduazione punto sulla base di questo meccanismo sono state proposte diverse misure di cograduazione:

$$lambda = \frac{C-D}{C+D}$$

- C= n di coppie concordanti
- D=n di coppie discordanti
- Gamma assume valore +1 in caso di perfetta relazione positiva, mentre -1 in caso di perfetta relazione negativa e 0 in assenza di relazione.

Prima di concludere segnaliamo l'esistenza di ***un'altra famiglia di misure di cograduazione*****: le graduatorie**, derivanti da un processo di ordinamento delle unità di analisi in sequenza ordinata.
Si tratta pertanto di variabili ordinali con molte modalità.

- **Scarso uso delle misure di associazione e di cograduazione**

Le misure di forza delle relazioni tra variabili nominali e ordinali non sono molto utilizzate nella ricerca sociale per almeno tre motivi:
1) la loro pluralità non ne ha facilitato la diffusione e popolarità.
2) quando le variabili sono nominali può avere poco senso calcolare un'unica misura sintetica di associazione, In quanto può darsi che la forza della relazione sia dovuta il comportamento specifico di alcune modalità.
3) queste misure sono di difficile interpretazione.

- **Rapporti di probabilità**

Mentre chiamavamo proporzione il rapporto fra la parte e il tutto, chiamiamo questo rapporto fra la frequenza di una categoria e la frequenza della categoria alternativa **rapporto di probabilità** (odds)[14] e lo indicheremo con la **lettera greca omega**.

$$\omega = \frac{f1}{f2} = \frac{P1}{1-P1}$$

Il passaggio dalla proporzione al rapporto di probabilità e viceversa è molto semplice:

il rapporto di probabilità assume valore 1 quando le due categorie della variabile hanno lo stesso peso; ha come valore minimo il valore zero ma non ho un limite superiore.
Abbiamo visto che la tecnica più comunemente utilizzata per studiare la relazione fra le variabili fa ricorso alle proporzioni in particolare alle ***proporzioni condizionate***, ora invece delle proporzioni condizionate e cioè dei rapporti fra frequenze parziali totali per le due categorie d'istruzione si considerino i rapporti di ***probabilità condizionati*** e cioè **rapporti favorevoli/contrari** sempre per le due categorie di istruzione.
Abbiamo *tanti rapporti di probabilità condizionati per la variabile Y* (atteggiamento verso la pena di morte) quante sono le categorie della variabile X (istruzione) nel nostro caso , Essendo due le categorie della variabile X abbiamo due rapporti di probabilità, uno per i meno istruiti a uno e uno per i più istruiti omega due.
Questo confronto può essere formalizzato dal rapporto fra gli **odds condizionati** che in inglese si chiamano **odds ratio** che traduciamo con ***rapporto di associazione***.

Il valore ottenuto può essere interpretato nel modo che segue:
Posto uguale ad uno il rapporto favorevoli/contrari fra più istruiti, esso assume il valore 3.33 fra i meno istruiti, come a dire che passando dei più istruiti i miei istruiti il rapporto tra favorevoli e contrari passa ad oltre il triplo.
il rapporto di associazione fra 2 variabili può assumere valore compreso fra 0 e +∞, passando per il valore 1, il quale si verifica in caso di indipendenza fra due variabili, più il valore è lontano da 1, maggiore è la forza della relazione.
Valori superiori ad 1 stanno a significare **un'associazione positiva** fra le variabili mentre **valori inferiori ad 1** stanno a significare **un'associazione negativa**.
Per associazione positiva intendiamo che i soggetti della categoria X1 hanno probabilità di collocarsi nella categoria Y1maggiore di quanto sia la probabilità dei soggetti della categoria X2 e viceversa.
Quindi, invertendo l'ordine delle righe o delle colonne, si ottiene un valore del rapporto di associazione che è l'inverso del valore originario.
Mentre il valore del rapporto resta immutato se si cambia l'orientamento della tabella si scambiano cioè le righe quelle colonne.
Si noti che il valore del rapporto di associazione non risente della dimensione del campione nei cambi se entrambe le frequenze di una riga di una colonna sono moltiplicati per una costante.
Questa stabilità costituisce una caratteristica importante di questa misura e mostra come essa sia capace di cogliere la struttura delle relazioni fra le due variabili senza risentire delle variazioni campionarie.

Questa misura presuppone una logica dicotomica in quanto il rapporto di probabilità è il rapporto tra la frequenza di appartenenza ad una data categoria e la frequenza di non appartenenza. Che poi questa non appartenenza si articoli in un'altra o in più alternative è irrilevante questo **quando si tratta di una sola variabile**.
Quanto alla relazione fra due variabili cioè rapporti di probabilità condizionati e la variabile dipendente che deve essere vista in un'ottica dicotomica ; la variabile indipendente può avere anche più di due categorie in questo caso si calcoleranno tanti rapporti di probabilità condizionati quante sono le modalità della variabile indipendente.
Da ciò si deduce come i rapporti di probabilità condizionati non sono che un modo alternativo per guardare alla relazione fra due variabili alternativo a quello che si serve invece delle proporzioni condizionate.
Possiamo anche trasformare i rapporti di probabilità in proporzioni:

$$p=¿ \quad \omega/(1+\omega)$$

Mentre i rapporti di associazione possono essere calcolati solo su tabelle 2x2.

CAPITOLO 6

ANALISI BIVARIATA: QUANDO LA VARIABILE DIPENDENTE E' CARDINALE (REGRESSIONE SEMPLICE)

1. REGRESSIONE LINEARE SEMPLICE

La principale tecnica utilizzata dei ricercatori per effettuare questo tipo di analisi della ***regressione lineare semplice***.
Consideriamo il caso in cui anche la variabile indipendente è di tipo cardinale.
L'obiettivo conoscitivo fondamentale nelle scienze sociali: stabilire in **quale misura la variabile indipendente influisce su quella dipendent**e o più **tecnicamente rilevare l'intensità dell'effetto esercitato dalla variabile indipendente su quella dipendente**.
L'obbiettivo di questa analisi sarà quello di determinare la forma,strettezza della relazione e l'intensità dell'effetto esercitato.
Quando si analizza la relazione fra due variabili cardinali, il primo passo consiste nel **rappresentare graficamente tale relazione mediante il cosiddetto diagramma di dispersione**.
Quest'ultimo è un semplice piano cartesiano che ordina I valori della variabile indipendente X lungo lasse orizzontale e valori della variabile dipendente Y lungo l'asse verticale.
Ogni osservazione viene collocata all'interno del piano cartesiano nel punto in cui i suoi valori di X e Y si intersecano.
L'insieme dei punti così tracciati illustra visivamente il modo in cui le due variabili covariano cioè variano insieme.
Il diagramma di dispersione ci consente dunque di desumere la ***forma della relazione***, ma non si dice nulla di preciso sull'intensità dell'effetto causale.
Ogni equazione è definita dalla sua forma funzionale e dei valori assunti i suoi parametri.
La forma funzionale più semplice e più comunemente usata nelle scienze sociali e quella lineare, che corrisponde alla seguente equazione: Y uguale alfa più beta X → $Y=\alpha+\beta X$
l'equazione lineare afferma che il valore della v.d. Y è uguale al parametro α più [+] il valore assunto dalla v.ind. X moltiplicato [x] per il parametro β.
Il risultato finale di quest'operazione che la relazione fra X e Y può essere espressa mediante una semplice linea retta la cui distanza dell'asse orizzontale è determinata dal valore assunto dal parametro alfa e la sua inclinazione è determinata dal valore assunto dal parametro beta.
Più precisamente in un'equazione lineare il **parametro alfa** noto come **intercetta** esprime il valore assunto da Y quando X è uguale a zero mentre il **parametro beta** chiamato **gradiente /recettore** ci dice di quanto varia il valore di Y per ogni variazione unitaria di X.
In questi termini il valore assunto dal **parametro beta** rappresenta la quantità di maggior interesse per il ricercatore in quanto esprime ***l'intensità dell'effetto esercitato dalla variabile indipendente su quella dipendente***; questo effetto è costante qualunque sia il valore di X.

L'equazione lineare costituisce la base della regressione lineare semplice cioè della tecnica che come abbiamo accennato all'inizio del capitolo è comunemente utilizzata dagli scienziati sociali per analizzare le relazioni fra coppie di variabili cardinali.
Conoscendo il valore di X è sempre possibile predire con precisione il valore di Y.
Lo scopo della riflessione lineare semplice è proprio questo stimare i valori dei parametri dell'equazione lineare alfa e beta corrispondente alla retta che meglio di ogni altra approssima la covariazione osservata fra la variabile indipendente e quella dipendente tale retta assume la seguente forma matematica:
Yi (con capellino sopra) uguale alfa più beta Xi

$$\hat{Y}_i = \alpha + \beta X_i$$

A Y è stato aggiunto un accento circonflesso che indica che i valori della variabile dipendente definiti dall'equazione lineare non solo quelli osservati ma bensì quelli predetti sulla base dei parametri alfa e beta stimati.
Se vogliamo esprimere in forma matematica valore osservati di Y allora abbiamo aggiungere all'equazione lineare detta anche equazione predittiva o modello di regressione lineare un ulteriore elemento come segue:

$$\hat{Y}_i = \alpha + \beta X_i + \varepsilon_i$$

Il nuovo termine rappresenta i cosiddetti ***errori di predizione*** cioè esprime, per ciascun caso i, la differenza fra il valore osservato di Y è quello predetto dal modello di regressione lineare.
Gli errori di predizione sono anche chiamati ***residui*** perché corrispondere a quella parte del valore di Y e chi va oltre la relazione lineare rappresentata dall'equazione predittiva, cioè quella parte del valore di Y che non può essere spiegata dall'effetto di X.
Ciò significa che il termine ε esprime l'influenza esercitata su Y da tutti i fattori casuali che non solo presi esplicitamente in considerazione dal modello di regressione lineare prescelto.
Tali fattori possono essere classificati **in tre categorie**

1. in primo luogo non è detto che la relazione fra X e Y si è perfettamente lineare: qua e là Possono esserci delle non linearità che vengono assorbite dagli errori di predizione.
2. In secondo luogo, il modello di regressione lineare semplice esprime valori di Y come funzione di un'unica variabile indipendente X senza tenere conto del fatto che tali valori possono essere influenzati in modo significativo anche da altre variabili; l'effetto esercitato su Y da tutte queste altre variabili un incluse nel modello contribuisce a determinare gli errori di predizione.
3. Infine il comportamento umano è caratterizzato da una certa dose di casualità di cui nessun modello di regressione potrebbe mai rendere conto e fa sì che il valore di Y non sia mai perfettamente prevedibile; gli errori di predizione rappresentano anche questa casualità intrinseca nella manifestazione di Y.

Lo scopo della regressione ideale semplicemente quello di stimare i valori dei parametri α e β corrispondente alla retta che approssima meglio di ogni altra la covariazione osservata fra X e Y, ciò equivale che la migliore retta di regressione è quella che minimizza la differenza fra i valori osservati di Y e quelli predetti dal modello cioè che minimizza gli errori di predizione.
La migliore retta di regressione è quella che minimizza la somma degli errori di predizione al quadrato cioè che rende minima la quantità.

I valori dei parametri alfa e beta soddisfano questo criterio perciò detti ***stime dei minimi quadrati***.
Come dobbiamo interpretare i valori delle variabili dipendenti privati il modello di digestione lineare?
Come spesso accade statistica anche in questi casi ci viene in aiuto il concetto di media.

Nella nostra discussione precedente abbiamo sottolineato che lo scostamento fra i valori osservati e i valori predetti di Y riflette l'azione di una serie di fattori che non vengono tenuti in esplicita considerazione del modello di regressione neanche prescelto.
In alcuni casi questi fattori fanno sì che valore servati siano maggiori di quelli predetti mentre in altri casi si verifica esattamente il contrario. Se le reazioni frei zitti non effettivamente lineare, nel complesso i casi del primo tipo tenderanno essere controbilanciati dai casi del secondo tipo, così che in media i valori di Y osservati in corrispondenza di ogni dato livello di X approssimeranno il valore di Y predetto per quel livello di X.
Dunque i valori di Y i predetti dal modello di regressione lineare prescelto possono essere interpretati come stime del valore medio della variabile dipendente osservato in corrispondenza di ciascun livello della variabile indipendente.
Questa affermazione lascia ancora spazio di almeno due obiezioni.
-Innanzitutto in alcuni casi i valori medi osservati di Y si discostano dai corrispondenti valori predetti di misura rilevante.
-In secondo luogo l'immagine della relazione fra voto di laurea e livello di reddito che si ricava dall'andamento dei valori medi della variabile dipendente non è affatto lineare, talvolta reddito medio osservato aumenta all'aumentare del voto ma in altri casi diminuisce.
(Secondo esempio riportato a pp: 153)
Possiamo fare due osservazioni:

- ***In primo luogo***: bisogna tenere conto del fatto che nel nostro esempio i valori medi osservati di Y sono stati calcolati sulla base di pochi casi e di conseguenza risultano piuttosto instabili;
- ***In secondo luogo***: è importante sottolineare che l'obiettivo di ogni modello di regressione non è quello di riprodurre esattamente la relazione osservata fra due variabili ma di evidenziare le caratteristiche salienti in modo tale da offrirne una rappresentazione parsimoniosa e intelligibile.

Possiamo rivedere leggermente la nostra definizione precedente e dire che i valori di Y predetti dal modello di regressione lineare prescelto possono essere interpretati come stime dei valori medi dizione che si osserverebbe in corrispondenza di diversi livelli di X se relazione fra le due variabili fosse perfettamente lineare.
In quest'ottica il valore assunto dal parametro beta ci dice di quanto varia in media il valore di Y per ogni variazione unitaria di X assumendo che la relazione tra le due variabili sia perfettamente lineare.
L'intensità dell'effetto come abbiamo già detto espresso dal parametro beta, però i ricercatori non si accontentano del valore assunto dal parametro beta, ma vogliono anche ***stabilire la strettezza della relazione fra X e Y cioè la misura in cui la retta di regressione approssima la covariazione osservata fra la variabile indipendente e quella dipendente***.
Tale approssimazione è tanto maggiore quanto minore è la differenza fra i valori di Y predetti dalla retta di regressione e valore di Y effettivamente osservati cioè quanto minore è la somma degli errori di predizione al quadrato.
Rilevare la strettezza della relazione fra X e Y dunque ***equivale a calcolare il potere predittivo della retta di regressione stimata*** cioè stabilire con quale precisione la conoscenza dei valori di X ci consente di indovinare i valori di Y.
Una misura di potere predittivo poco utilizzata molto efficace è il cosiddetto ***errore standard della regressione*** la cui formula è:

$$\sigma(\varepsilon) = \frac{somma\,degli\,errori\,di\,predizione\,al\,quadrato}{parametri\,stimati}$$

Ovvero…

Questa misura equivale alla radice quadrata della somma degli errori di predizione al quadrato divisa per il numero di casi -2.
L'errore standard della regressione può essere interpretato come una misura dell'errore di predizione medio.

Dunque quanto è maggiore il valore assunto da questa misura tanto minore è il potere predittivo della retta di regressione.
Una misura di potere predittivo ben più nota e ampiamente utilizzata nelle scienze sociali è il coefficiente di determinazione, indicato dal simbolo **R alla seconda** (R^2)
Per valutare il grado di precisione predittivo ovvero il potere predittivo della semplice media possiamo **calcolare la somma delle differenze al quadrato fra i valori osservati di Y e quelli predetti dalla media** cioè:

Dove Y con il trattino sopra denota la media osservata di Y.
Quanto maggiore è il valore di questa somma tanto maggiore è l'errore di predizione complessivo e quindi tanto minore è il potere predittivo della media.
A questo punto disponiamo di due misure dell'errore di predizione quello che si riferisce la predizione basata solo sulla media di Ye di quello che si riferisce la predizione basata sulla retta di regressione.

R^2=

Dunque il coefficiente di determinazione è una misura relativa del potere predittivo la retta di regressione.
Si tratta cioè di una misura del tipo Pre, come quelle esaminate nel capitolo cinque, in quanto **esprime la riduzione percentuale dell'errore di predizione iniziale** che si ottiene prendendo in considerazione i valori di X.
In termini del tutto equivalenti si può anche dire che il **coefficiente di determinazione rappresenta la percentuale di variazione di Y spiegata dalla variabile indipendente**.
R^2 può assumere valori compresi fra **zero** (Che equivale al caso in cui X non esercita alcun influenza su Y e quindi non può contribuire a predirne valori) **e uno** che equivale al caso in cui tutti i valori osservati di Y sono perfettamente predetti della retta di regressione.
Nonostante la popolarità il coefficiente di determinazione è una misura spesso sopravvalutata e talvolta utilizzata in modo inappropriato.
In primo luogo serve ambiguamente lo scopo per il quale è stato originariamente concepito, cioè misurarne il potere predittivo della retta di regressione.
Questa affermazione si basa sul fatto **che il valore assunto da R alla seconda dipende in modo sostanziale e non solo dalla somma degli errori di predizione al quadrato, ma anche dalla varianza della variabile indipendente**.
Precisamente, a parità di ogni altra condizione al valore di R alla seconda è tanto più elevato quanto maggiore è la varianza di X.
Al contrario dell'errore standard della regressione non è influenzata in alcun modo dalle caratteristiche della distribuzione di X ma dipende solo dalla somma degli errori di predizione quadrato.
Se il modello di regressione [A] da luogo ad un valore di $\sigma(\varepsilon)$ è minore di quello prodotto dal modello B possiamo essere certi che il potere predittivo del modello a e superiore a quello del modello B.

Invece se il valore R alla seconda associato il modello di regressione [C] risulta maggiore di quella di quella associato al modello [D] non significa necessariamente che il potere predittivo del modello e superiore [A] quello del modello [D].
Infatti se nel modello [D]la varianza dice molto minore che nel modello potrebbe anche essere che in realtà il potere predittivo del modello C è inferiore a quello del modello D.
Spesso il valore R alla seconda noto come coefficiente di correlazione lineare di Pearson viene interpretato non solo come misura dell'astrattezza della relazione fra X e Y ma anche come misura dell'intensità dell'effetto esercitato da X su Y.
Questa interpretazione è **assolutamente scorretta** in quanto implica l'equiparazione di due concetti analiticamente e sostanzialmente distinti.
La strettezza della relazione fra X e Y non è altro che la capacità della retta di regressione di approssimare geometricamente i valori osservati di Y.
Tale capacità non ha nulla a che fare con l'intensità dell'effetto esercitato da X su Y ma espressa esclusivamente dal parametro beta.
In altri termini possiamo dire i casi in cui esso esercita un intenso effetto su Y ma la retta di regressione possiede un basso valore predittivo, i casi in cui X esercita un effetto molto debole su Y ma la retta di regressione possiede un potere predittivo elevato.
L'incongruità del coefficiente di determinazione emerge pienamente quando lo si esprime utilizzando la seguente formula:

R^2=

Dove **Var(X)** denota la varianza della variabile indipendente.
Come si può constatare il valore assunto da R^2dipende contemporaneamente **da tre elementi** :

- L'intensità dell'effetto esercitato da X su Y, rappresentata dal parametro beta.
- Il potere predittivo della retta di regressione rappresentato dall'errore standard della regressione $\sigma(\varepsilon)$.
- La varianza di X.

  Due più valori di R alla seconda uguali possono derivare da combinazioni anche molto diverse di questi tre elementi.
  Significa che il coefficiente di determinazione è intrinsecamente ambiguo e scarsamente informativo.

Concludendo quando si valutano i risultati di un modello di regressione lineare semplice bisogna distinguere nettamente fra l'intensità dell'effetto esercitato da Y; la strettezza della relazione esistente fra X e Y.
La rilevanza sostanziale della prima misura evidente e l'abbiamo sottolineato più volte nel caso della nostra discussione.
E' invece dubbio che sia utile conoscere l'assetto della relazione fra X e Y.
È opportuno tenere conto dei limiti intrinseci al coefficiente di determinazione e optare per errore standard della regressione.

3.CASI ANOMALI E CASI INFLUENTI

In questo momento ci occuperemo dei ***casi anomali e casi influenti***.
Nel contesto della regressione lineare semplice **un caso anomalo o outlier**, È un'osservazione in corrispondenza della quale la variabile dipendente assume un valore atipico dato il valore assunto dalla variabile indipendente.
Un caso anomalo di per sé non rappresenta un problema per la regressione lineare, lo diventa solo

quando il suo valore di X eccentrico cioè si discosta dal valore medio X (con il trattino sopra) in misura apprezzabile.
In questa situazione **il caso anomalo viene definito caso influente** in quanto la sua presenza influisce in modo significativo sui risultati della regressione specificatamente sulle stime dei parametri alfa e beta.
(eventualmente vedere grafici pp: 160)
Per ciascun valore di X ci sono alcuni casi che attirano la retta di regressione verso un altro mentre vene sono altri che la tirano verso il basso.
Se la nuvola di punti analizzata una forma regolare queste due forze di attrazione tendono a controbilanciarsi e di conseguenza la rete di questione tende a situarsi al centro della nuvola di punti.
Al contrario se in corrispondenza di un dato valore di X e la distribuzione di Y risulta simmetrica allora in quel punto una delle due forze di attrazione prevarrà sull'altra attirando nella propria direzione cioè un alto o basso la retta di regressione.
Questo spostamento della retta di regressione della sua sede naturale sarà tanto più marcato quanto maggiore è il grado combinato di anomalia ed eccentricità dei casi responsabile della simmetria.

Con una formula breve possiamo allora **dire che l'influenza esercitata da ogni determinato caso sulla retta di regressione è uguale al prodotto del suo grado di anomalia per il suo grado di eccentricità**: ***influenza = anomalia di Y/X per eccentricità di X .***

Tutti i casi, esercitano una certa influenza sulla retta di regressione; solo alcuni si meritano lo status di caso influente.
Ma come? Effettuare un'ispezione visuale del diagramma di dispersione.
È opportuno avvalersi anche di alcuni indici numerici appositamente costruiti.
In questa sede menzioneremo brevemente quattro di questi indici

1- Hat value [$\mathbf{h_i}$]

Esprime la misura in cui il valore X assunto dal caso i è eccentrico rispetto alla media X (con il trattino sopra).

2- Residuo studentizzato [$\mathbf{E^*_i}$]

Il secondo a sua volta e ottenuto applicando una particolare standardizzazione al residuo εi in quanto tale [E] esprime il grado di anomalia del valore condizionato di Y assunto dal caso i.

3- Indice di Cook [$\mathbf{D_i}$]

Combina la misura di eccentricità del primo quella misura di anomalia del secondo e pertanto esprime il grado complessivo di influenza esercitata dal casuali sulla retta di regressione.

4- Indice beta [$\mathbf{DFBETAS_i}$]

Esprime in forma standardizzata l'influenza esercitata dal caso i suoi suo parametro beta, ciò significa che la presenza del caso i accresce il valore di beta; quando invece assume un valore negativo allora la diminuisce.
Nella ricerca di casi influenti è consigliabile considerare congiuntamente valori assunti da tutti e quattro questi indici in quanto ognuno di essi offre informazioni parzialmente diverse.
(es, pp 164-165)
Se la diagnosi del problema è piuttosto chiara altrettanto non si può dire della terapia consigliata.
In primo luogo bisogna cercare di capire **l'origine dell'anomalia osservata**; se ci troviamo di fronte una semplice errore di registrazione dei dati il problema,si risolve facilmente apportando le opportune correzioni e stimando nuovamente la retta di regressione.
Se invece l'anomalia reale allora evidente che la relazione fra X e Y presenta delle peculiarità che il modello di regressione lineare semplice non è in grado di spiegare.

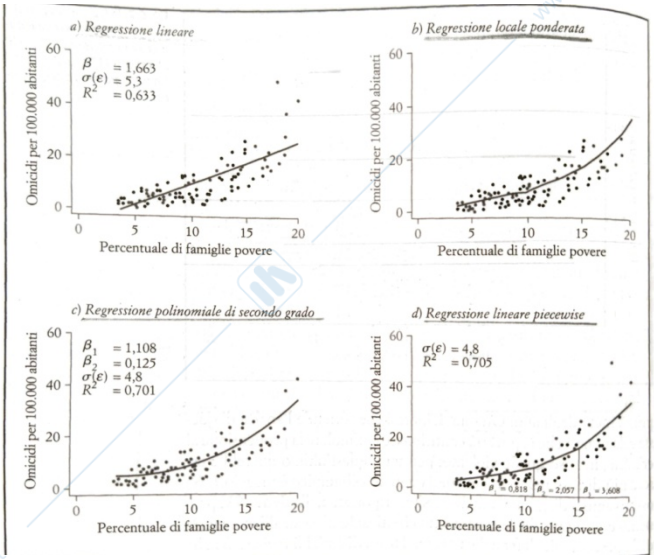
Tali peculiarità possono avere un carattere sistematico oppure costituire delle eccezioni isolate. **Nel primo caso** è necessario riformulare il modello aggiungendo altre variabili indipendenti o modificando la forma funzionale dell'effetto esercitato da X su Y.
Nel secondo caso invece si possono seguire ***due strategie alternative***.
La prima consiste nell'analizzare la reazione di interesse omettendo le osservazioni problematiche. L'eliminazione dell'analisi dei casi influenti tutta via quasi sempre sconsigliabile soprattutto perché comporta una perdita di informazioni tanto più cospicua quanto minore è il numero di casi presi in esame.
In alternativa è preferibile ricorrere a tecniche di stima della retta di regressione che tengono opportunamente conto dell'eventuale presenza di casi influenti.
Una di queste tecniche nota come ***regressione robusta***, consiste nell'assegnare a ciascun caso un peso indirettamente proporzionale al suo grado di influenza e nel calcolare la retta di regressione dando maggiore importanza ai casi più pesanti.
In questo modo si ottiene una retta di regressione che dipende molto dei casi poco influenti e poco dei casi molto influenti il cui effetto distorcente viene così neutralizzato o comunque ampiamente ridotto.

4.

Talvolta la relazione fra X e Y presenta dei caratteri sistematici che il modello di regressione lineare non riesce a spiegare.
Una delle possibili cause di questo anomalia è che la **relazione oggetto di analisi è intrinsecamente non lineare** e pertanto non può essere rappresentato in modo appropriato mediante una semplice linea retta.
Per analizzare le reazioni bivariate bisogna fare ricorso a tecniche particolari alcune delle quali sono semplici estensioni della regressione lineare nelle pagine precedenti.
Tre di queste tecniche sono:

1- regressione locale ponderata
2- regressione polinomiale
3- regressione lineare piecewise

Per rilevare l'esistenza di eventuali non linearità nella relazione fra due variabili cardinali**; il primo passo da compiere** consiste nell'ispezione visuale del diagramma di dispersione corrispondente.
es. pp: 167-168.



La retta di regressione offre una stima scorretta dell'intensità dell'effetto esercitato da X su Y, tale intensità non essendo costante per tutti i valori di X non può essere espressa da un unico parametro β.

Per ovviare a questa inadeguatezza della normale regressione lineare bisogna ricorrere ad alcune tecniche specificatamente dedicate all'analisi delle relazioni Bivariate non lineari.
Alcune di queste tecniche sono **di tipo non parametrico** cioè hanno un carattere prevalentemente esplorativo e si limitano a produrre una rappresentazione grafica della relazione oggetto di analisi.
Altre tecniche invece **sono parametriche** cioè sono finalizzate a stimare i valori di alcuni parametri che consentono di quantificare l'intensità dell'effetto esercitato da X su Y.
La tecnica della regressione locali ponderata (1) prevede che il valore atteso di Y corrispondente ogni possibile valore di $\mathbf{X^c}$ venga stimato applicando il suddetto valore di X e i suoi vicini una regressione lineare ponderata.
I vicini di un dato valore X^c sono definiti come gli SxN casi osservati che assumono i valori di X meno distanti da X^c , Dove S denota la proporzione di casi considerata.
Il valore del parametro S dipende dal grado di dettaglio con il quale si vuole rappresentare la relazione fra X e Y.
Quando **S grande** allora i valori di Y e predetti per i diversi valori di X^c, danno luogo a una curva **smussata** cioè relativamente insensibile alle non linearità locali.
Al contrario quando **S piccolo** la curva che si ottiene tende a riprodurre tutte le non linearità locali e quindi adesso però una forma **frastagliata**.
Alcune di queste tecniche parametriche sono mere estensioni della regressione lineare semplice e quindi sono relativamente facili da utilizzare.
La prima è la cosiddetta **regressione polinomiale (2)** che consiste nell'applicare i dati una normale regressione lineare in cui la variabile indipendente sia stata preventivamente trasformata in un polinomio di grado K.
Nelle scienze sociali la regressione polinomiale utilizzata è quella di secondo grado la cui equazione predittiva assume la seguente forma:

$$\hat{Y}_i = \alpha + \beta_1 X_1 + \beta_2 X^2_i$$

Secondo questa equazione il valore predetto di Y è uguale alla somma di tre termini: il parametro alfa, il valore naturale di X moltiplicato per il parametro beta uno, il valore di X al quadrato moltiplicato per il parametro beta due.
Tale equazione corrisponde una relazione fra X e Y di forma quadratica cioè caratterizzata da un unico punto di curvatura; quando la covariazione fra X e Y assume invece una forma cubica cioè presenta 2 punti di curvatura allora è opportuno fare ricorso alla regressione polinomiale di terzo grado la quale equazione predittiva è:

$$\hat{Y}_i = \alpha + \beta_1 X_1 + \beta_2 X^2_i + \beta_3 X^3_i$$

Nella normale regressione lineare l'effetto di X su Y è espresso unicamente dal valore assunto dal parametro β:

$$\frac{\Delta \hat{Y}}{\Delta X} = \beta$$

Ovvero ogni volta che X varia di unità, il valore predetto di Y varia di β unità.

Nella regressione polinomiale di secondo grado questa semplice relazione non vale perché la variabile indipendente entra nell'equazione predittiva in duplice forma: al naturale e al quadrato.di conseguenza il calcolo dell'effetto esercitato da Y diventa più complicato e assume la seguente forma:

$$\frac{\Delta \hat{Y}}{\Delta X} = \beta_1 + \beta_2 + 2\beta_2 X$$

Ovvero ogni volta che X varia di unità il valore predetto di Y varia di un ammontare che dipende sia dei valori stimati dei parametri **beta uno** e **beta due** sia dal valore di partenza di X.
Un'altra estensione della regressione lineare semplice che permette di analizzare in modo appropriato le relazioni bivariate non lineari è **la regressione lineare piecewise (3)**.
L'idea che sta alla base di questa tecnica è piuttosto semplice: l'effetto esercitato da Y varia in funzione della regione della variabile X preso in considerazione; tuttavia all'interno di ognuna di queste regioni l'effetto dice su Y rimane costante.
Dunque per applicare la regressione lineare piecewise bisogna innanzitutto suddividere la gamma dei valori osservati di X in una serie di K regioni distinte fra loro ma internamente omogenei; il valore di X che separa due regioni contigue è detto ***nodo***.
Una volta determinati i ***K -1 nodi*** che delimitano le diverse regioni di X è necessario creare **k regressori** ognuno dei quali rappresenta una data regione ed è ottenuto trasformando in modo opportuno la variabile indipendente.
Questo tipo di regressione (piecewise) offre un innegabile vantaggio i valori stimati e sui parametri esprimono in modo diretto l'effetto variabile esercitato da istruzione.
Per contro la regressione polinomiale hai il pregio di evitare ricercatore l'onere di stabilire le regioni in cui si articola la relazione fra X e Y.

5.

Fino a questo momento abbiamo preso in esame solo relazioni fra coppie di variabili cardinali.
In altri casi vi è *l'esigenza di analizzare il modo in cui la variabile dipendente cardinale è influenzata dalla variabile indipendente di tipo categoriale cioè nominale ordinale*.
PROBLEMA: A prima vista alcune variabili operazioni potrebbero sembrare impraticabili in quanto la variabile indipendente si articola in una coppia di categorie il cui valore non è quantificabile cioè non può essere espresso da numeri che possiedono proprietà matematiche.
In realtà anche l'informazione contenuta nelle variabili qualitative come il genere può essere espresso in termini quantitativi e quindi utilizzata nella regressione lineare.
Il modo per effettuare questa traduzione è piuttosto semplice e si basa sull'idea di presenza-assenza delle modalità in cui si articola la variabile categoriale di interesse.
Ad esempio quando il soggetto analizzato è un uomo allora possiamo dire che in questo caso la modalità maschio è presente mentre la modalità femmina è assente.
Queste **informazioni possono essere espresse in termini quantitativi mediante due regressioni:**
Il primo repressore X^M: ha lo scopo di rappresentare la modalità maschio e supera valori uno in tutti i casi in cui tale modalità è presente il valore zero in tutti i casi in cui la modalità estinte.
Il secondo regressore X^F: ha lo scopo di rappresentare la modalità femmina e assumerà valore 1 in tutti i casi in cui tale modalità è presente, e il valore 0 in tutti i casi in cui essa non è presente quindi assente.
I regressori di questo tipo sono detti ***regressori indicatori*** proprio perché hanno la funzione di indicare se una determinata modalità di una data variabile categoriale è presenti in un caso oppure non lo è.
Avendo tradotto la variabile qualitativa genere in due regressori , siamo in grado di usare la regressione lineare per misurare l'effetto esercitato dal genere sul reddito all'interno del nostro gruppo di laureati.

Prima di fare tutto ciò però è molto importante che le informazioni contenute nei repressori $\mathbf{X^M}$ e $\mathbf{X^F}$ non sono indipendenti.
Conoscendo il valore assunto di uno dei due regressori siamo in grado di stabilire con precisione il valore assunto dall'altro.
Ciò significa che tutte le informazioni di cui abbiamo bisogno per misurare l'effetto esercitato dal genere sul reddito è contenuta in uno solo dei due aggressori.
La categoria esclusa viene chiamata ***categoria di riferimento*** e svolge un ruolo essenziale nell'interpretazione dei risultati della regressione.
Dal punto di vista matematico la scelta della categoria da escludere è rilevante.
Nella nostra analisi assumeremo una categoria di riferimento della variabile "genere" la modalità "maschio" e pertanto rileveremo effetti esercitati dal genere sul reddito per mezzo del regressore $\mathbf{X^F}$.
Il modello di regressione appropriato per quest'analisi assume dunque la seguente forma:

$$\hat{Y}_i = \alpha + \beta X^F_i$$

Come si può osservare queste equazioni del tutto identico a quello utilizzato nel paragrafo uno per stimare l'effetto lineare.
Sul piano formale anche l'interpretazione dei parametri è identica.
Dunque analizzare l'effetto esercitato dalla variabile indipendente categoriale sulla variabile dipendente cardinale mediante la regressione lineare equivale a misurare le differenze osservate fra le diverse categorie di X in termini di valori medi di Y.
Quando gli interrogativi di ricerca si complicano e l'analisi assume un carattere multivariato allora la regressione lineare diventa l'unico strumento in grado di misurare in modo appropriato gli effetti esercitati da una più variabili categoriali sulla variabile cardinale gli interessi.
E'utile considerare le situazioni in cui la variabile indipendente politomica cioè si articola in tre o più categorie.
Il procedimento da seguire in questi casi sostanzialmente identico a quello illustrato sopra.
es. pp: 177-178.
Il fatto che i valori assunti dei parametri beta uno e beta due debbano essere interpretati relativamente alla categoria di riferimento non impedisce di fare confronti anche fra le altre categorie della variabile indipendente.

Capitolo 7

ANALISI MULTIVARIATA

Obbiettivo: sottolineare i limiti dell'analisi bivariata e di introdurre la logica dell'analisi multivariata, cioè delle analisi in cui la variabile indipendente è espressa come funzione di due o più variabili esplicative.
1.
L'analisi bivariata in linea di principio ha un solo obiettivo: stabilire se le due variabili oggetto di studio compaiono in modo sistematico.
Nella maggior parte dei casi però constatare l'esistenza di una relazione fra due variabili non è sufficiente: scienziati sociali interessa soprattutto stabilire se la covariazione osservata è il frutto di un rapporto di causa effetto fra le due variabili implicate cioè se i valori assunti dall'uno dipendono dei valori assunti dall'altra.
In tutte le nostre analisi abbiamo implicitamente assunto che X fosse causa di Y questo assunto però richiede di essere giustificato sia sul piano teorico sia sul piano statistico.
A livello teorico è necessario rendere espliciti i meccanismi in base ai quali si ritiene che i valori assunti da Y possono essere influenzati dei valori assunti da X.

A livello statistico si è autorizzati a postulare l'esistenza di una relazione causale fra X e Y solo quando vengono soddisfatte **due condizioni**:
1- X e Y devono che variare in modo sistematico
2- la covariazione osservata fra X e Y deve essere spuria, cioè deve manifestarsi anche quando si tiene sotto controllo l'azione esercitata da altre variabili.
La seconda condizione sottolinea un fatto fondamentale ai fini della nostra discussione: **se si vuole stimare l'effetto causale esercitato da X su Y non ci si può limitare ad un'analisi bivariata della loro relazione** ma bisogna tenere conto del ruolo che altre variabili potrebbero svolgere nel plasmare la relazione stessa.
In altre parole **per studiare in termini causali delle relazioni** fra variabili è necessario passare **dall'analisi bivariata all'analisi multivariata**.
In caso contrario si corre il rischio di farsi ingannare dall'apparenza affermando l'esistenza di una relazione causale che in realtà è assente o addirittura presente una forma opposta a quella messa in luce dall'analisi bivariata.
Paradosso di Simpson: la forma della covariazione fra X e Y messo in luce da analisi Bivariata risulta invertita quando introduciamo nell'analisi uno o più variabili supplementari in funzione di "controllo".
2.
La funzione primaria dell'analisi multivariata è quella di stimare il vero effetto causale esercitato da Y cioè di stabilire la misura in cui valori assunti da Y dipendono dei valori assunti da X.
Abbiamo anche detto che questo obiettivo viene conseguito tenendo sotto controllo, cioè neutralizzando, gli effetti distorcenti esercitati da uno o più variabili supplementari che proprio per questo motivo sono definiti ***variabili di controllo***.
Problema: come si scelgono le variabili di controllo devo [illegible]?
Per rispondere è utile riprendere il concetto di direzione
introdurre il ***concetto di ordine causale fra più variabili***..

X→Y

Questo semplice schema afferma che la variabile X precede la variabile Y e l'ordine causale pertanto X può esercitare effetti su Y , ma Y può esercitare effetti su X.

Tutte le variabili diversi da X e Y possono essere classificate in base alla posizione da essi occupati all'interno **dell'ordine causale** elementare definito dalla variabile indipendente e quella dipendente.
Tale classificazione comprende **tre categorie principali**:
1- variabili antecedenti: sono quelle che nell'ordine causale precedono sia X che Y
2- variabili intervenienti: sono quelle che nell'ordine causale precedono Y ma seguono X
3- variabili susseguenti: sono quelle che nell'ordine causale seguono sia X che Y.
Esiste poi un **quarto tipo di variabili** che chiameremo **4-variabili concomitanti** che pur risultando genuinamente correlate con X non sono chiaramente identificabili come cause né come effetti della stessa X.
A fini pratici queste variabili possono essere trattati alla stessa stregua di quelle antecedenti e quindi essere considerate come variabili che nell'ordine causale precedono sia X che Y.
Non tutte le variabili diverse da X e Y però sono ugualmente adatte a svolgere il ruolo di variabili di controllo.
INFATTI →affinché **una data variabile** che indicheremo con il simbolo generico **Z** possa produrre effetti distorcenti sulla relazione fra X e Y osservata a livello bivariato è necessario che essa eserciti un effetto casuale sia sulla variabile indipendente che su quella dipendente.

Questa condizione si può verificare solo quando X precede sia X che Y nell'ordine causale.
Queste considerazioni ci portano a formulare un **principio fondamentale dell'analisi multivariata**:
Per misurare correttamente l'effetto causale esercitato da X su Y **bisogna**:

1- Includere nell'analisi tutte le variabili antecedenti concomitanti che esercitano un effetto causale sia su X che su Y.
2- Escludere dall'analisi tutte le variabili intervenienti e susseguenti.

La logica che sta alla base della seconda clausola: l'esclusione dell'analisi delle variabili susseguenti dovrebbe essere evidente, non ha senso tenere sotto controllo l'azione svolta da una variabile che in virtù della posizione che occupata nell'ordine causale non può esercitare alcun effetto su X e Y, le ragioni dell'esclusione delle variabili intervenienti invece sono più articolate verranno chiarite dopo.
es pp: 186-187

3.

TAB. 7.3. Relazione fra livello di istruzione (*Z*) e *a*) senso di appartenenza alla Chiesa cattolica (*X*); *b*) atteggiamento verso la pena di morte (*Y*)

a) Relazione fra Z e X

Senso di appartenenza alla Chiesa cattolica	Livello di istruzione: Medio-basso	Alto
Basso o nullo	13,2	70,8
Medio o alto	86,8	29,2
Totale (N)	100,0 (760)	100,0 (240)
Effetto bivariato	$d_{ZX} = -0{,}576$	

b) Relazione fra Z e Y

Atteggiamento verso la pena di morte	Livello di istruzione: Medio-basso	Alto
Contrario	22,4	75,0
Favorevole	77,6	25,0
Totale (N)	100,0 (760)	100,0 (240)
Effetto bivariato	$d_{ZY} = -0{,}526$	

Questi dati mettono in luce un fatto importante: X e Y hanno un caso in cui cioè sono entrambe influenzate negativamente indipendentemente l'uno dall'altra dal livello di istruzione.
Ciò significa che quando livello di istruzione aumenta la probabilità di essere cattolico la probabilità di essere favorevole alla pena di morte diminuiscono contemporaneamente dando così l'impressione a livello bivariato di essere correlate positivamente.
Esattamente a livello bivariato succede che la correlazione positiva fra X e Y riprodotto artificialmente da Z si somma al vero effetto causale esercitato dalla variabile indipendente su quella dipendente dando così luogo all'effetto di variato osservato.
In formula:
Effetto bivariato= effetto causale + effetto spurio
Ovvero:
Effetto causale= effetto bivariato - effetto spurio.
Per rimuovere l'effetto spuria dell'effetto più variato e ottenere così una stima corretta dell'effetto casuale è necessario tenere costante il valore delle variabili di controllo.

Il coefficiente $d_{XY/Z=1}$ esprime l'effetto esercitato da X su Y quando la variabile Z assume valore uno cioè quando ad esempio livello di istruzione è medio basso.

Similmente il coefficiente $d_{XY/Z=2}$ esprime l'effetto esercitato da Y quando la variabile Z assume valore due cioè quando ad esempio il valore distruzione è alto.
Entrambi questi effetti sono detti ***effetti condizionati di X su Y*** in quanto sono misurati a parità di Z cioè tenendo costante il valore della variabile di controllo.

il coefficiente $d_{XY/Z}$ esprime ***l'effetto netto*** di X su Y o più precisamente l'effetto medio esercitato da X su Y al netto degli effetti esercitati dalla variabile Z.
Se vi sono buone ragioni per ritenere che la variabile Z sia l'unico fattore in grado di alterare le relazioni Bivariate fra X e Y allora il coefficiente di $d_{XY/Z}$ può essere interpretato come misura corretta cioè priva di ogni componente spuria dell'effetto causale esercitato da X su Y.

In questo caso la differenza fra l'effetto di variato dice su Y rappresentato dal coefficiente dXY il corrispondente effetto causale è rappresentato dal coefficiente di XY. Z esprime l'ammontare dell'effetto spurio attribuibile a Z:

$$d_{XYZ} = d_{XY} - d_{XY.Z}$$

Gli esempi riportati sul libro rappresenta un ***effetto bivariato parzialmente spurio*** : questo caso si verifica quando X esercita un dato effetto causale su Y, ma a livello bivariato, tale effetto risulta alterato dall'azione esercitata da una o più variabili supplementari.
L'alterazione in questione può verificarsi in diversi modi: talvolta si limita ad aumentare e diminuire l'intensità dell'effetto reale mantenendone però invariato il segno; altre volte si esprime in modo più subdolo, invertendo il segno stesso dell'effetto oggetto di studio o facendolo addirittura sparire.
Accanto a quello parzialmente spuri, esistono anche ***gli effetti bivariati completamente spuri***.
In questo caso si verifica quando non esiste alcuna relazione causale fra X e Y ma a livello Bivariato l'azione esercitata da uno o più variabili supplementari crea l' illusione che tale relazione esista realmente.

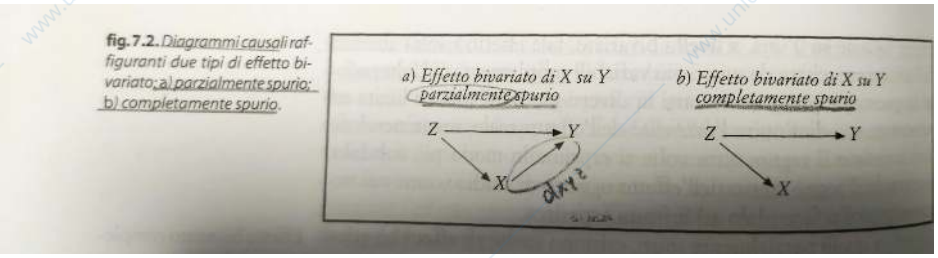


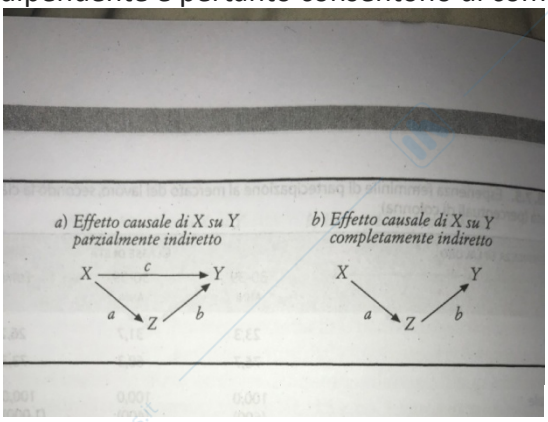
fig. 7.2. Diagrammi causali raffiguranti due tipi di effetto bivariato: a) parzialmente spurio; b) completamente spurio.

Diagrammi causali che illustrano graficamente i due casi discussi sul libro.
Nei diagrammi di questo tipo la posizione occupata da ciascuna variabile lungo la dimensione orizzontale sinistra destra riflette la sua collocazione all'interno dell'ordine causale; inoltre le frecce sono utilizzate per indicare l'esistenza di relazioni causali nette fra coppie di variabili.
Come si può vedere in entrambi i casi Z allo status di variabile antecedente esercitano effetto causale sia su X che su Y.
Ciò in cui i due diagrammi differiscono è la presenza della freccia che simboleggia l'effetto causale esercitato da X su Y: nel caso dell'effetto bivariato parzialmente spurio tale freccia è presente mentre nel caso dell'effetto bivariato completamente spurio è assente.

4.

Nell'analisi multivariata non serve solo a neutralizzare gli effetti spuri provocati da alcune variabili sulle relazioni causali di interesse.
Talvolta le variabili di controllo sono utilizzate per comprendere meglio la natura di questa relazione cioè per mettere in luce la catena di cause ed effetti che nel suo insieme produce l'effetto causale complessivo esercitato dalla variabile indipendente su quella dipendente.
Per comprendere questa funzione dell'analisi multivariata si consideri questo esempio: supponiamo di essere interessati a stimare il cambiamento nel tempo della partecipazione femminile al mercato del lavoro, per ottenere questa stima decidiamo di porre a confronto le esperienze lavorative di due gruppi di donne: quelle nate fra il 91 e quelle tra il 70 e quelli nati fra il 1941 e tra il 1950.
Esiste una relazione negativa: all'aumentare dell'età diminuisce la probabilità di avere avuto un'esperienza di lavoro retribuito.
Precisamente il valore assoluto del coefficiente d_{XY} indica che il passaggio dalla generazione più giovane quella più anziana determina una diminuzione delle probabilità di partecipazione al mercato del lavoro pari a 8,4 punti percentuali.

Il valore assunto dal coefficiente d_{XY} rappresenta una stima dell'effetto più variato esercitato dalla variabile indipendente su quella dipendente.
Per semplificare la nostra discussione supponiamo che tale effetto sia privo di qualsiasi componente spuria.
Ne consegue che il valore assunto da d_{XY} può essere visto come una stima corretta dell'effetto causale esercitato dall'età sulle probabilità di partecipare alla forza lavoro.
Il fatto che l'effetto sia negativo indica che con il tempo le opportunità delle donne di partecipazione al mercato del lavoro sono aumentate avvicinandosi sempre di più quello degli uomini.
Questo cambiamento può essere interpretato in diversi modi: da un lato si può ritenere che l'aumento della partecipazione femminile al mercato del lavoro sia il frutto di un vero e proprio rinnovamento generazionale cioè sia una conseguenza del declino dei modelli tradizionali di rapporto fra sessi.
In alternativa si può ipotizzare che il mutamento osservato sia il prodotto di un semplice effetto di composizione cioè il fatto che nel corso del tempo all'interno della popolazione femminile è aumentata la diffusione di caratteristiche che di per sé aumentano la propensione a partecipare al mercato del lavoro.
Queste caratteristiche possono essere viste come ***variabili intervenienti*** cioè come variabili che mediano in tutto o in parte l'effetto causale esercitato dalla variabile indipendente su quella dipendente e pertanto consentono di comprenderne meglio la natura.



C= effetto causale diretto
Come si può vedere, in entrambi casi la variabile indipendente X che esercita un effetto causale sulle variabile di controllo **Z**; Quest'ultimo a sua volta esercita un effetto causale sulla variabile dipendente Y.
Il prodotto di questi due effetti causali (axb) rappresenta il cosiddetto **effetto causale indiretto di Y** cioè *l'influenza che la variabile indipendente esercita sulla variabile dipendente per mezzo della variabile di controllo*.
In un certo senso Z funziona come cinghia di trasmissione: X fa muovere Z il cui movimento fa muovere Y pertanto ogni variazione di X produce seppur indirettamente una variazione in Y.
Nel primo caso (a) il legame diretto fra i X e Y non è assente: ciò significa che l'influenza esercitata dalla variabile indipendente su quella dipendente interamente mediato dalle azioni svolte dalla variabile di controllo.
Pertanto:

effetto causale totale = a x B = effetto indiretto.

Al contrario **nel secondo caso** (b) il legame diretto fra X e Y non è presente.
In questo caso dunque l'effetto causale totale esercitato da X su Y può essere visto come la somma di due aspetti parziali: l'effetto indiretto effetto diretto.

In formula:

effetto causale totale uguale = a x B + c = effetto indiretto + effetto diretto.

Ora per comprendere meglio introduciamo nell'esempio il **livello di istruzione**.

Il nostro obiettivo è quello di verificare in quale misura l'effetto causale esercitato dall'età sulla partecipazione femminile al mercato del lavoro può essere spiegato dal ruolo di mediazione svolto dal livello di istruzione.

Formula:

effetto causale totale = effetto indiretto + effetto diretto

Dunque l'effetto causale totale esercitato dall'età sulla probabilità delle donne di partecipare al mercato del lavoro è attribuibile per circa due terzi all'effetto di composizione legato livello distruzione e per rimanere terzo ad un genuino effetto generazionale.

5.

In alcuni casi l'effetto causale esercitato dalla variabile indipendente X sulla variabile dipendente Y si manifesta in modi diversi a seconda del valore assunto dalla variabile di controllo Z.

Quando ciò si verifica siamo in presenza di quello che viene tecnicamente definito un ***effetto di interazione***.

(es completo pp: 193)

Dopo aver spiegato tutto l'esempio chiaramente ci troviamo di fronte un effetto di interazione: l'effetto causale esercitato dal genere sullo status occupazionale non è costante bensì dipende dal valore assunto dalla variabile livello distruzione ad esempio.

In questo caso dunque non ha senso esprimere l'influenza totale di X su Y mediante il coefficiente sintetico $\mathbf{d_{XYZ}}$; piuttosto **è opportuno considerare separatamente i valori assunti dei singoli effetti condizionati, rappresentati dei coefficienti $d_{XY/Z=1}$ e $d_{XY/Z=2}$.**

6.

Al termine di questa nostra discussione può essere utile riepilogare gli elementi essenziali del linguaggio degli effetti che costituisce la base dell'analisi multivariata.

Il punto di partenza è **l'effetto di X e Y osservato a livello bivariato**.

Introducendo delle analisi **tutte le variabili antecedenti rilevanti** otteniamo una stima corretta del vero effetto causale esercitato da X su Y e per differenza una stima dell'effetto spuri attribuibile alle variabili antecedenti.

Se poi introduciamo nell'analisi anche **tutte le variabili intervenienti rilevanti** otteniamo una stima corretta dell'effetto diretto di X su Y e per differenza una stima del dell'effetto indiretto attribuibile alla mediazione delle variabili intervenienti.

Le varie tecniche di analisi multivariata disponibili tutta via sono perfettamente in grado in misura maggiore o minore di affrontare tutte queste situazioni anche quelle più complicate.

I programmi di analisi statistica attualmente disponibili associati all'elevata potenza di calcolo garantita dei computer e dell'ultima generazione consente al ricercatore di stimare con semplicità e rapidità a tutti gli effetti che usa degli interessi anche quando il sistema di variabili analizzato esteso e intricato.

È opportuno però sottolineare un punto essenziale: la relativa facilità con la quale è possibile utilizzare in pratica le tecniche di analisi multivariata deve essere vista come garanzia di ottenere sempre comunque risposte precise e media dei propri interrogativi di ricerca.

Se questi ultimi sono formulati in modo ambiguo e poco chiaro i risultati dell'analisi multivariate per quanto oggettivi e corretti sul piano statistico possono suggerire interpretazioni fuorvianti.

Lo stesso problema si verifica se gli interrogativi di ricerca pur formulati chiaramente vengono tradotti nel linguaggio statistico in modo impreciso scorretto, chiarezza concettuale precisione tecnica e dunque sono i due ingredienti essenziali per ottenere delle buone analisi multivariata.