****

**FACOLTA’ DI SCIENZE DELL’UOMO E DELLA SOCIETA’**

**Corso di laurea in Psicologia clinica**

**DISPENSA per il Corso di**

**Fondamenti epistemologici**

**e metodi della ricerca psicosociale**

**APPROCCI ALLA RICERCA EMPIRICA ED ELEMENTI DI STATISTICA PER LA RICERCA PSICOSOCIALE**

**Anno Accademico 2022/2023**

**Prof.ssa Maria Fobert Veutro**

Questa dispensa è articolata in due parti.

Nella prima parte anticipo in forma sintetica, e corredo con esempi, alcuni argomenti relativi ai **fondamenti epistemologici della ricerca socio-psicologica** (vedi cap. 4 del manuale adottato per il corso)**,** e illustro **le fasi di un disegno di ricerca standard, o quantitativa.**

La definizione di termini quali ‘epistemologia’, ‘gnoseologia’, ‘metodo’, ‘metodologia’ si trova nel primo capitolo del manuale.

La seconda parte della dispensa intende fornire a chi non li avesse studiati in precedenza **i primi rudimenti di statistica**, utili per comprendere ed assimilare sia gli argomenti relativi alla **raccolta delle informazioni** sia i riferimenti alle tecniche statistiche di **analisi dei dati** (vedi oltre: il disegno della ricerca standard), che compaiono nel manuale.

Si precisa che le tecniche di analisi presentate nel testo menzionato, e che verranno illustrate durante le lezioni, sono quelle più elementari. Esse sono sufficienti in quanto gli argomenti di Metodologia del corso sono dedicati più che all’analisi dei dati (che nella ricerca standard prevede l’uso di tecniche statistiche perlopiù complesse, oggetto di altri corsi), alle fondamentali e delicate **fasi che precedono l’analisi: la raccolta e l’organizzazione delle informazioni**; ai problemi e ai dubbi che queste pongono; alla scelta delle procedure adeguate per avviare e condurre correttamente una ricerca sociopsicologica.

Quindi gli argomenti illustrati nella seconda parte di questa dispensa anticipano ciò che nel disegno di una ricerca standard (si chiarirà successivamente cosa si intende) rappresenta una delle fasi finali (l’analisi dei dati). Essi però vengono presentati prima, e in forma semplificata, per far familiarizzare con (o ripassare) temi che, per la mia esperienza didattica, vengono considerati ostici (probabilmente perché richiedono il ricorso, sia pur minimo, alla matematica e alla logica) e rischiano così di ostacolare il processo di apprendimento. D’altra parte, la comprensione delle problematiche relative alle fasi di raccolta e di organizzazione delle informazioni è agevolata dalla conoscenza preliminare delle tecniche che potranno poi essere usate per analizzare i dati raccolti.

Le pagine riguardanti gli elementi di statistica sono frutto di sunti e di schematizzazioni di parti di due volumi - A. Marradi, *L’analisi monovariata*, 1993, FrancoAngeli, Milano, e P. Corbetta, *La ricerca sociale: metodologia e tecniche. Vol. IV: L’analisi dei dati*, 2003, Il Mulino, Bologna – e di mie integrazioni ed esemplificazioni. Si invitano gli studenti che desiderassero approfondire gli argomenti presentati a consultare i volumi menzionati.

**FONDAMENTI EPISTEMOLOGICI DELLA RICERCA PSICOSOCIALE**

(cap. 4 del manuale)

LO SVILUPPO DELLE RIFLESSIONI SUI MODI DI CONDURRE LA RICERCA SCIENTIFICA HA CONDOTTO A DUE APPROCCI DIVERSI:

* APPROCCIO **STANDARD**, DETTO ANCHE QUANTITATIVO
* APPROCCIO **NON**-**STANDARD**, DETTO ANCHE QUALITATIVO

LE DIFFERENZE TRA I DUE APPROCCI HANNO ORIGINE NEI **DIVERSI ASSUNTI** TEORICI ADOTTATI DAGLI SCIENZIATI E COMPORTANO L’USO DI **TECNICHE DI RILEVAZIONE E DI ANALISI DELLE INFORMAZIONI DEL TUTTO DIVERSE** (anche se sono frequenti ricerche che usano tecniche miste)

|  |  |
| --- | --- |
| **Standard***la scienza consiste nel formulare asserti su* ***relazioni fra proprietà*** *che* ***non dipendono da conoscenze e valutazioni personali del ricercatore*** | **non-standard***le scienze sociali non possono prescindere dalle* ***conoscenze*** |
|  | ***personali dello scienziato****;* |
| **metodo sperimentale***gli asserti impersonali riguardano* ***relazioni causali*** | **metodo dell’associazione***gli asserti impersonali possono riguardare solo* ***relazioni di associazione*** | *gli oggetti studiati sono* ***soggetti*** *e le loro* ***opinioni*** *sono* ***rilevanti*** |

* UN **ASSERTO** E’ UNA PROPOSIZIONE CHE STABILISCE UNA **RELAZIONE TRA LE PROPRIETA’** DEGLI OGGETTI DI STUDIO, CIOE’ A VARIAZIONI DEGLI STATI SU UNA PROPRIETA’ CORRISPONDONO VARIAZIONI DEGLI STATI SU UN’ALTRA PROPRIETA’ (O SU PIU’ PROPRIETA’)
* LE **PROPRIETA’** SONO CARATTERISTICHE DEGLI OGGETTI DI STUDIO. Ad esempio, se l’oggetto di studio è l’**individuo**, **proprietà** saranno: età, genere, titolo di studio, stato civile, reddito annuo, etc.
* Gli **STATI** della proprietà genere saranno: maschio, femmina; gli stati della proprietà titolo di studio sono: licenza elementare, media, diploma, laurea, titolo post-laurea, etc.
* GLI **ASSERTI** SONO DETTI **IMPERSONALI** QUANDO SONO FORMULATI SENZA IL CONTRIBUTO DELLE CONOSCENZE E DELLE VALUTAZIONI DEL RICERCATORE
* LE **RELAZIONI** SONO **CAUSALI** QUANDO LE VARIAZIONI DI UNA (O PIU’) PROPRIETA’ (VARIABILE DIPENDENTE) SONO CAUSATE CON CERTEZZA DALLE VARIAZIONI DI UN’ALTRA PROPRIETA’ (VARIABILE INDIPENDENTE)
* LE **RELAZIONI** SONO DI **ASSOCIAZIONE** QUANDO NON E’ POSSIBILE STABILIRE UN NESSO DI CAUSA ED EFFETTO FRA LE VARIAZIONI DI UNA PROPRIETA’ E QUELLA DI UN’ALTRA (O PIU’) **SENZA** LA VALUTAZIONE PERSONALE DEL RICERCATORE

**esempi di asserto:**

In Inghilterra i borghesi votano conservatore

L’uomo è razionale

I borghesi tendono a votare conservatore



I corpi si attraggono con forza direttamente proporzionale al prodotto delle loro masse e inversamente proporzionale al prodotto delle loro distanze

Nei comuni toscani, più alta è la percentuale dei commercianti, maggiore è la percentuale di voti ad AN

Gli uomini sono più intelligenti delle donne

L’uomo cerca la felicità

La propensione al suicidio è influenzata dal livello di anomia nella società



I giudici sono imparziali

Tutti i giovani umani di sesso maschile desiderano sessualmente la madre e quindi vedono il padre come fortunato rivale e lo odiano

**ASSUNTI DEL METODO SPERIMENTALE**

STUDIANDO I FENOMENI DELLA NATURA (NON UMANI) E’ POSSIBILE **SUDDIVIDERE** CON SICUREZZA L’INSIEME DELLE POSSIBILI **PROPRIETA’** COINVOLTE NELL’ESPERIMENTO **IN QUATTRO GRUPPI**:

1. UNA PROPRIETA’ (DETTA SPERIMENTALE O **DIPENDENTE**) DELLA QUALE SI OSSERVANO LE VARIAZIONI NELL’INTERVALLO DI TEMPO T
2. UNA O POCHE PROPRIETA’ (DETTE OPERATIVE O **INDIPENDENTI**) I CUI STATI SI FANNO VARIARE IN MODO CONTROLLATO NELLO STESSO INTERVALLO T
3. POCHE PROPRIETA’ CHE SI PREVEDE (IN BASE ALLE CONOSCENZE TEORICHE) POSSANO INFLUENZARE GLI STATI DELLA PROPRIETA’ SPERIMENTALE O DELLE PROPRIETA’ OPERATIVE, E CHE – IN QUEL PARTICOLARE ESPERIMENTO – SI TENGONO (NON: SI DANNO PER) **COSTANTI** NELL’INTERVALLO T
4. INFINITE PROPRIETA’ CHE SI DANNO (SEMPRE IN BASE ALLE CONOSCENZE TEORICHE) PER **IRRILEVANTI** IN QUELLA PARTICOLARE SITUAZIONE SPERIMENTALE

INOLTRE, NELLE SCIENZE NATURALI, GLI OGGETTI DI STUDIO SONO **FUNGIBILI** (si può dare per scontato che oggetti dello stesso tipo reagiscano sempre nello stesso modo nelle stesse condizioni). **CONSEGUENZA**: E’ POSSIBILE **ESTENDERE (GENERALIZZARE)** I RISULTATI DI UN ESPERIMENTO A TUTTI GLI OGGETTI DELLO STESSO TIPO

NELL’AMBITO DELL’**APPROCCIO STANDARD**, POICHE’ GLI ASSUNTI DEL METODO SPERIMENTALE **NON POSSONO ESSERE ADOTTATI PER LE SCIENZE UMANE perché le variabili non sono di solito manipolabili**: vedi assunti B e C, e **perché non si può assumere con certezza che molte variabili siano irrilevanti**: vedi assunto D. PER QUESTE SCIENZE (ECCETTO CAMPI DI RICERCA PARTICOLARI) SI USA IL METODO DELL’ASSOCIAZIONE

**ASSUNTI DEL METODO DELL’ASSOCIAZIONE**

* OGNI **CASO** (ad esempio un individuo) SI PUO’ SCOMPORRE PERFETTAMENTE NEI SUOI STATI SULLE DIVERSE PROPRIETA’ (ad esempio: maschio, 35 anni, diplomato, celibe, etc.)
* OGNI STATO SU UNA PROPRIETA’ (ad es.: diplomato), UNA VOLTA TRASFORMATO IN **DATO**, E’ DEL TUTTO SEPARATO DAL CASO (individuo) CHE LO DETIENE
* OGNI STATO SU UNA PROPRIETA’ E’ TOTALMENTE INDIPENDENTE DAGLI ALTRI STATI (es.: celibe, cattolico, egocentrico, etc.) DELLO STESSO CASO (individuo) SU TUTTE LE ALTRE PROPRIETA’
* OGNI STATO AL QUALE E’ STATO ASSEGNATO UN VALORE SIMBOLICO (di solito un numero), TRASFORMANDOLO IN UN **DATO**, E’ TRATTATO COME IDENTICO A QUALUNQUE ALTRO STATO CHE ABBIA RICEVUTO LO STESSO VALORE (ad es.: il codice 3 per “diplomato”)

LA RACCOLTA DEI DATI IN BASE A QUESTI **ASSUNTI** PERMETTE L’ANALISI DEGLI STESSI MEDIANTE LO STRUMENTO BASILARE DELLA RICERCA SOCIALE STANDARD: LA **MATRICE “CASI PER VARIABILI”** (vedi oltre).

LE PROPRIETA’ (genere, età, titolo di studio, reddito, religione, etc.) VENGONO TRASFORMATE IN **VARIABILI** MEDIANTE UNA **DEFINIZIONE OPERATIVA** CHE ASSEGNA VALORI SIMBOLICI (di solito numeri) AI VARI STATI SULLE PROPRIETA’, TRASFORMANDOLI IN **DATI**.

UNA VOLTA REGISTRATI I DATI NELLA MATRICE, SI PUO’ EFFETTUARE L’**ANALISI MONOVARIATA** ( ad esempio, quanti casi presenti nella matrice sono dello stesso genere, o della stessa fascia d’età, etc.) E, APPLICANDO LE TECNICHE STATISTICHE ADEGUATE ALLA NATURA DELLE VARIABILI, L’**ANALISI BIVARIATA** (ad esempio, quanti con lo stesso titolo di studio percepiscono un reddito di entità simile, o quanti dello stesso genere hanno una stessa appartenenza religiosa, etc.) E **MULTIVARIATA** (vedi oltre).

**L’APPROCCIO NON-STANDARD** (si usa solo nelle scienze umane)

**ORIGINI**

**Wilhelm Dilthey**, filosofo tedesco che alla fine dell’800 sosteneva che le **scienze umane** (*Geisteswissenschaften*, **scienze dello spirito**) non potevano perseguire modelli universali di spiegazione basati sulla causalità (cioè non potevano *erklären,* spiegare), ma dovevano cercare la **comprensione** (*verstehen*) delle motivazioni del soggetto attraverso una forma di empatia con esso, garantita dalla comune natura umana del ricercatore e del soggetto studiato.

* E’ un metodo particolarmente adatto per **esplorare fenomeni nuovi e/o sconosciuti, problemi delicati e/o incarnati da soggetti marginali, in stato di disagio**, etc.
* Per molti ricercatori però la scelta di questo approccio prescinde dall’oggetto della ricerca.

**CARATTERISTICHE**

a) orientamento a ridurre al minimo la separazione tra scienza e vita quotidiana; pertanto, la forte propensione ad abbandonare le torri d’avorio della scienza pura e **prendere contatto diretto con i soggetti** mentre svolgono le loro attività abituali, lasciando che esprimano **le loro visioni del mondo** con i loro termini e ponendoli al centro dei rapporti di ricerca; una grande attenzione per i problemi del **significato** e un’attitudine all’**ascolto** dei segnali che provengono dalla realtà invece di una rigida limitazione a modelli prestabiliti;

b) forte dipendenza dal **contesto**: la ricerca è confinata alla situazione specifica che studia;

c) preferenza per i problemi **micro** — che per alcuni ricercatori non esclude l’ambizione di affrontare su questa base anche i problemi macro;

d) orientamento marcatamente **idiografico**, descrittivo — che in molti autori non implica la rinuncia a formulare concetti di portata generale;

e) orientamento marcatamente **induttivo**: invece di “verificare ipotesi” pre-formulate, il ricercatore spalanca gli occhi sugli scenari e apre le orecchie a tutti i messaggi che la situazione gli presenta;

f) predilezione per la **comprensione globale di situazioni specifiche** più che per l’istituzione di relazioni causali lineari tra variabili — e ancor meno per la “verifica” di ipotesi sulle associazioni tra queste variabili;

1. la causazione, se contemplata, è considerata un processo che si ricostruisce attraverso **narrazioni** e non attraverso relazioni statistiche tra istantanee scaglionate nel tempo o tra variabili rilevate nello stesso momento (come avviene quando si ricorre alla matrice dei dati);

La mancanza di procedimenti saldamente stabiliti e generalmente accettati — di un “metodo scientifico” codificato e ritualizzato al quale attenersi — implica:

h) grande importanza delle **qualità e capacità del ricercatore e dei suoi collaboratori**, che nella ricerca di tipo standard sono spesso considerati meri esecutori. Esperienza in ricerche precedenti, conoscenza del segmento di realtà che si studia, flessibilità, creatività, intuizione, sono più importanti degli (eventuali) strumenti tecnici nel determinare la qualità della ricerca;

i) **difficoltà** nello sviluppare questo genere di ricerca senza cadere nel **banale**, nel **gratuito**, nell’**aneddotico**; e — d’altra parte — senza “diventare un nativo” (***go native***): alcuni ricercatori assimilano infatti così profondamente il punto di vista dei soggetti studiati che diventano o si sentono membri della loro comunità.

**TECNICHE** DI RACCOLTA E DI ANALISI DELLE INFORMAZIONI USATE NELLA

**RICERCA NON-STANDARD O QUALITATIVA**

***(tre azioni: osservare, interrogare, leggere)***

Queste tecniche vengono qui sommariamente presentate per fornirne una generica conoscenza (o ripasso) agli studenti che desiderassero approfondirle ma si precisa che esse **non verranno trattate in questo corso e non saranno oggetto d’esame**. Il corso invece è dedicato alle tecniche della ricerca standard o quantitativa.

**L’OSSERVAZIONE PARTECIPANTE (*osservare*)**

* Il ricercatore entra direttamente e per un periodo di tempo relativamente lungo nell’**ambiente naturale** di un determinato gruppo sociale, stabilisce **relazioni di interazione personale** con i suoi membri al fine di osservare le loro azioni e **capire le loro motivazioni** mediante un processo di **immedesimazione**.
* Può dichiarare apertamente che vuole entrare nel gruppo solo per studiarlo (**osservazione palese**), oppure può infiltrarsi nel gruppo fingendo di essere un membro come gli altri (**osservazione dissimulata**).
* Per entrare nella comunità spesso si serve di un **mediatore culturale**, cioè una persona che gode della fiducia della comunità e che è anche capace di comprendere le motivazioni e le necessità del ricercatore.
* Può servirsi anche di **informatori**, persone appartenenti alla comunità studiata che l’osservatore usa per acquisire informazioni ed interpretazioni dall’interno della cultura studiata.
* **QUANDO USARLA?** **Fenomeno nuovo**; **gruppi chiusi**, segreti, illegali; piccole società autonome, collocate in ambiti territorialmente definiti in un universo culturale chiuso (**studi di comunità**); subculture sviluppatesi all’interno di società complesse (**studi di subculture**).
* **COSA OSSERVARE?** **Contesto fisico** (spazi, luoghi), **contesto sociale** (soggetti coinvolti e loro funzioni e attività), **interazioni formali** (riti, procedure), **interazioni informali** (osservare anche fatti apparentemente banali), **interpretazioni degli attori sociali** (manifeste e implicite).
* **QUANDO E COSA REGISTRARE?** Stendere delle **note quotidiane** (la memoria è selettiva), che riportano: **descrizioni**, **interpretazioni del ricercatore** che comprendono riflessioni teoriche e reazioni emotive (servono a studiare la dinamica tra osservatore e osservato), **interpretazione dei soggetti studiati**. **Distinzione** tra descrizioni e interpretazioni. Le valutazioni vanno chiaramente attribuite a chi le ha espresse (evitare effetto alone). **Fedeltà**: riportare frasi e commenti letterali dei soggetti
* **COME ANALIZZARE LE INFORMAZIONI?** Far emergere **aspetti nuovi** dei quali magari i protagonisti non hanno consapevolezza. Individuare **classi**, **tipi**, **temi culturali** (valori che ispirano i comportamenti dei membri della comunità)
* **RAPPORTO DI RICERCA**: **stile narrativo**; descrizioni dettagliate; inserimento di **brani letterali**

**L’INTERVISTA NELLA RICERCA NON-STANDARD *(interrogare)***

* E’ uno strumento flessibile e adattabile che deve consentire all’intervistato di esprimere **il suo punto di vista con le sue parole**.
* Si usa anche nella ricerca standard per intervistare **soggetti di alto status**, **soggetti marginali** e per trattare **argomenti delicati**. In questi casi è inadatto il questionario standardizzato a risposte chiuse.
* Di solito l’**intervistatore** parla pochissimo, ha un ruolo **maieutico**. Richiede abilità, esperienza e capacità di **ascolto**. **Centralità dell’intervistato**.
* Far tematizzare l’intervistato. Invitarlo non ad esprimere opinioni ma a **raccontare episodi** specifici. Eventuali domande proiettive. **Probes**.
* **Campionamento** a scelta ragionata. Campionamento a valanga.

**TIPI DI INTERVISTA**

* **STRUTTURATA** (intermedia tra standard e non-standard): **domande uguali** e nello **stesso ordine** per tutti ma le **risposte** sono **aperte** (e poi possono essere “chiuse”). Si usa per non precludersi la possibilità di un’analisi standard.
* **SEMI-STRUTTURATA**: non ci sono domande predefinite ma una **traccia** di argomenti da trattare. I temi indicati nella traccia devono essere sottoposti a tutti ma l’ordine non è fisso. Tipi particolari di interviste semi-strutturate sono l’intervista **A TESTIMONI PRIVILEGIATI** e l’intervista **FOCALIZZATA** (quest’ultima è circoscritta ad un tema preciso).
* **NON STRUTTURATA** (**ERMENEUTICA**): **nessuna traccia**. Il ricercatore cerca di trattare i **temi** che gli interessano agevolando approfondimenti e arginando divagazioni eccessive. Non tutte le interviste toccano gli stessi argomenti. Importanti: empatia intervistatore-intervistato; capacità di elaborazione dell’intervistato.
* Un tipo estremo di intervista ermeneutica è la **STORIA DI VITA**, in cui nemmeno il tema è stabilito. Assume una forma cronologico-narrativa. L’intervistatore si lascia condurre dall’intervistato.
* **INTERVISTA DI GRUPPO**: si usa quando si vogliono studiare le dinamiche di un gruppo e anche per rilevare dichiarazioni che difficilmente vengono espressi in colloqui a due.
* **FOCUS GROUP**: il dibattito è focalizzato su un tema o un evento. L’intervistatore stimola e dirige la discussione ascoltando le diverse interpretazioni che emergono. Il conduttore deve impedire il monopolio della discussione da parte di alcuni.
* **ANALISI DELLE INTERVISTE**: si scompongono i testi e si ricostruiscono riunendo tutte le informazioni relative ad uno stesso tema. I criteri interpretativi (tematici e/o cronologici) devono essere espliciti e argomentati. Si possono creare classi e individuare tipi.
* **RAPPORTO DI RICERCA**: stile narrativo; si sviluppano le argomentazioni e, a loro sostegno e illustrazione, vengono riportati brani delle interviste.

**L’USO DEI DOCUMENTI *(leggere)***

Documento è qualsiasi materiale (non solo cartaceo) che fornisce informazioni su un fenomeno sociale. A differenza delle informazioni rilevate con l’osservazione o con le interviste, queste esistono **indipendentemente dall’intervento del ricercatore** e **non reagiscono all’interazione** con il ricercatore (eccetto le testimonianze orali). I documenti permettono di studiare anche il passato

**DOCUMENTI PERSONALI**

La storia usava e usa quelli prodotti da personalità di rilievo. Le scienze sociali anche quelli **prodotti dalla gente comune** che, non essendo finalizzati ad un uso pubblico, presentano la sfera intima ed una interpretazione personale

**AUTOBIOGRAFIE**. Resoconti scritti con visione diacronica. Rischi: immagine migliore o meno screditante; razionalizzazione a posteriori

**DIARI**. Di solito spontanei in quanto per uso personale; simultanei agli avvenimenti descritti. Rari; più comuni sono i resoconti giornalieri su richiesta del ricercatore (es.: bilanci del tempo)

**LETTERE**. Sono il prodotto dell’interazione tra due persone

**TESTIMONIANZE ORALI**. Si usano anche per lo studio delle società illetterate e per conoscere il punto di vista di persone analfabete, di solito anziane. Le storie di vita riguardano l’individuo. La storia orale riguarda la società: l’individuo racconta avvenimenti a cui ha preso parte o di cui è testimone oculare. Prima la storia si serviva solo di fonti scritte. La storia orale è nata in America: utile per conoscere il punto di vista di classi subalterne e minoranze etniche e linguistiche

**DOCUMENTI ISTITUZIONALI**

A **basso costo**. Svantaggi: **incompletezza dell’informazione**; **ufficialità della rappresentazione**

**MASS MEDIA**. Articoli, programmi, servizi, trasmissioni, inserti e spot pubblicitari, etc. Esempio: ricerca sulla soap opera, che rivela stereotipi sulla donna.

**NARRATIVA, TESTI PEDAGOGICI, RACCONTI DELLA CULTURA POPOLARE**. Consentono di rivivere gli avvenimenti grazie ad una esposizione empatica. Verismo di Zola, Maupassant, Flaubert, etc. Libri scolastici e fiabe: fonti sulla cultura e sull’ideologia di una società

**MATERIALE GIUDIZIARIO**. Sentenze, verbali di processi, testimonianze di pentiti, fascicoli relativi a detenuti, etc. Utili per studiare la devianza ma anche l’ideologia di uno Stato, mediante sentenze e commenti dei giudici

**DOCUMENTI DELLA POLITICA**. Manifesti, programmi, campagne elettorali, slogan, appelli, liste elettorali, liste votanti

**DOCUMENTI AZIENDALI E AMMINISTRATIVI**. Libri contabili, bilanci comunali, certificati medici, riviste professionali, temi di scuola, pagelle con giudizi, contratti, bollette, testamenti, orari ferroviari, biglietti di cinema e di teatro, elenchi telefonici, etc.

Fanno parte dei documenti anche le **TRACCE FISICHE**: tracce di erosione (usura di pavimenti o di libri) e tracce di accrescimento (rifiuti, polvere, scritte, graffiti)

**IL DISEGNO DELLA RICERCA STANDARD**

**(CON LA MATRICE DEI DATI)**

Si puntualizza che l’elenco delle fasi di una ricerca standard di seguito presentato è una schematizzazione e una visione d’insieme degli argomenti presenti nel manuale dal cap. V al cap. IX. **Pertanto si invitano gli studenti a** **collegare, sviluppandolo, ogni punto** di questo elenco agli argomenti del manuale e ad esercitarsi ad **esemplificare ogni punto in una possibile ricerca da loro ideata** (dall’individuazione del tema, dell’ambito spazio-temporale, dell’unità di analisi, al **disegno della mappa dei concetti**, alla scelta di definizioni operative e di indicatori adeguati, etc.), che potrà essere oggetto d’esame.

Spesso, ma non sempre, preliminarmente si analizza la letteratura sul tema della ricerca che si intende effettuare. Sulla base degli interessi cognitivi del ricercatore / del gruppo di ricerca / del committente, delle risorse, delle scadenze temporali, della disponibilità di collaboratori e del loro grado di addestramento, della disponibilità di supporti materiali (registratori, computers), etc.

1. si sceglie l’**AMBITO** SPAZIO-TEMPORALE. Nelle scienze umane i risultati di una ricerca non sono generalizzabili oltre l’ambito spazio-temporale entro il quale sono stati scelti i casi sui quali sono state raccolte le informazioni.
2. si sceglie l’**UNITÀ DI ANALISI,** cioè l’oggetto di studio.

 Risolviamo/approfondiamo il nostro problema cognitivo raccogliendo informazioni su Stati, su collegi elettorali, su quartieri, su piccole aziende, su individui adulti, su casalinghe, su ospiti di orfanatrofi, su gatti domestici?

1. si disegna la **MAPPA DEI CONCETTI**

cioè si cerca di individuare preliminarmente quali saranno le proprietà della nostra unità di analisi che potranno servire per risolvere/approfondire il nostro problema cognitivo. Cerchiamo di immaginare alcune delle plausibili relazioni tra loro, al fine di far emergere quanti più possibile dei concetti che ci serviranno prima di iniziare la raccolta delle informazioni invece di scoprirlo in fase di analisi dei dati, quando ormai le informazioni sono state raccolte ed è troppo tardi per rimediare ( punto k). **Si disegna materialmente! E’ opportuno esercitarsi a farlo su un foglio o una lavagna.**

1. si dà una **DEFINIZIONE OPERATIVA** dei concetti definibili direttamente (genere, età, titolo di studio etc.), e contestualmente ― anche alla luce dell’ambito e dell’unità scelta ― si sceglie lo **strumento** di raccolta delle informazioni (intervista con questionario standard, invio di questionario postale, intervista telefonica, ricorso a dati censuali o comunque di fonte pubblica, etc.).
2. si scelgono (più) **INDICATORI** per ciascun concetto della mappa non definibile direttamente**.** Un solo indicatore non restituisce mai la ricchezza semantica del concetto rappresentato. L’inadeguatezza degli indicatori è conseguenza di una insufficiente riflessione sulle varie dimensioni del concetto (alcuni aspetti saranno indicanti e altri aspetti saranno estranei, cioè rappresentano un altro concetto). **Attenzione a non confondere indicatori con indici!**
3. si scelgono i criteri di estrazione del **CAMPIONE** ― a meno che si intenda rilevare le informazioni su tutti gli esemplari dell’unità presenti nell’ambito spazio-temporale (*enumerazione completa*).
4. si eseguono le procedure previste dalla definizione operativa (operativizzazione) e si costruisce la **MATRICE** con i **DATI** raccolti.
5. si costruiscono degli **INDICI** per sintetizzare l’informazione raccolta attraverso gli indicatori (con criteri tipologici se gli indicatori sono variabili categoriali**;** mediante somme in qualche modo ponderate dei punteggi se gli indicatori sono variabili cardinali o quasi-cardinali). **Attenzione a non confondere indicatori con indici!**
6. si inizia l’analisi dei dati producendo le **DISTRIBUZIONI DI FREQUENZA** (**analisi monovariata**: vedi oltre) delle variabili che interessano.
7. si prosegue l’analisi trasformando in **modelli** alcuneipotesi circa le relazioni tra le variabili più importanti e applicando le procedure statistiche opportune.
8. si **RIFLETTE SUI RISULTATI** delle analisi, che di solito pongono nuovi interrogativi o suggeriscono ulteriori direttive di analisi. E’ in questa fase che ci si rende conto di non aver raccolto informazioni su alcune proprietà che invece potrebbero spiegare alcuni risultati che stiamo trovando, cioè dell’insufficienza della nostra mappa dei concetti (**** punto c).
9. si **PROSEGUE NELL’ANALISI** dei dati, approfondendola sulla base delle indicazioni raccolte nella fase iniziale.
10. si stende il **RAPPORTO DI RICERCA**.

**ELEMENTI DI STATISTICA PER LA RICERCA PSICOSOCIALE**

Preliminarmente introduco e definisco alcuni termini (che abbiamo già incontrato) la cui conoscenza è essenziale per comprendere gli argomenti metodologici del corso e la tecnica di analisi preliminare (e la più elementare) usata in ogni ricerca sociale standard: l’analisi monovariata (vedi oltre).

***Ambito spazio-temporale***. In una data ricerca si raccolgono informazioni su uno o più fenomeni entro certi limiti spaziali e temporali. Si può decidere, ad esempio, di condurre una ricerca sugli studenti iscritti all’Università di Catania nell’anno accademico 2015-16. Questi limiti racchiudono l’ambito di quella ricerca.

***Unità*** è il tipo di referente sul quale si raccolgono le informazioni. Nella ricerca sociale le unità più spesso usate sono l’individuo, la famiglia, l’impresa, il distretto elettorale, il comune, la provincia, lo Stato. Possono essere scelti come unità anche un intervallo temporale (l’anno, il mese) o la data, o anche un evento (ad esempio, la guerra, oppure la crisi economica).

***Popolazione*** è l’insieme (simbolo: **N**) degli esemplari dell’unità scelta esistenti entro l’ambito di una data ricerca. Se l’unità è il residente maggiorenne e l’ambito è il comune di Enna all’inizio del 2015, popolazione sarà l’insieme di tutti i residenti maggiorenni ad Enna a quella data.

***Campione***. Non sempre esistono le risorse economiche, le possibilità tecniche o l’interesse a raccogliere informazioni su tutti gli esemplari dell’unità scelta presenti entro l’ambito di una ricerca. In tal caso si sceglie un campione: con questo termine si designa un sotto-insieme degli esemplari dell’unità. Se il procedimento di scelta (*estrazione*) garantisce a tutti gli esemplari le stesse probabilità di essere scelti, il campione si dice *casuale*.

***Caso*** è ogni esemplare dell’unità sul quale si raccolgono effettivamente informazioni. Il totale dei casi si indica con il simbolo **n**. Si parla di ***enumerazione completa*** quando n = N; quando si effettua un ***campionamento***, n < N.

Naturalmente, **non si raccolgono informazioni su tutte le (innumerevoli) caratteristiche dei casi, ma solo su un limitato sottoinsieme**.

Nelle scienze sociali, le caratteristiche dei casi si dicono ***proprietà***, e ogni particolare modo di presentarsi di una proprietà si dice ***stato*** su quella proprietà. Una proprietà riferita ad un individuo è, facendo un esempio banale, il colore dei capelli. Gli stati sulla proprietà ‘colore dei capelli’ sono: castano, nero, rosso, biondo, etc. Un’altra proprietà (più interessante per le scienze umane) è, ad esempio, l’orientamento politico. In questo caso gli stati sono: destra, centro-destra, centro, etc. Una proprietà riferita ad una famiglia è, ad esempio, il numero dei componenti. Gli stati sono: 1, 2, 3, 4, etc. Se un comune ha un sindaco del PD e assessori del M5Stelle, quella specifica situazione è lo stato di quel comune sulla proprietà ‘tipo di giunta’.

Se decidiamo di raccogliere sistematicamente informazioni su, ad esempio, la proprietà ‘titolo di studio’ degli individui oggetto della nostra ricerca, dovremo dare una ***definizione operativa*** di quella proprietà, cioè redigere un elenco di stati che si ritengono significativamente distinti tra loro (di solito, per questa proprietà, sono: nessun titolo, elementari, medie inferiori, etc.), e che diventano così le ***modalità*** della variabile ‘titolo di studio’; assegnare a ciascuna di queste modalità un differente valore simbolico (di solito questi valori simbolici sono cifre ma talvolta è opportuno usare valori simbolici non numerici, ad esempio lettere). Poiché ogni stato dev’essere attribuito ad una ed una sola delle modalità previste, occorre stabilire regole per dirimere i casi dubbi (che nel caso di questo esempio non si pongono, ma in altri casi possono essere frequenti). Occorre poi assegnare al vettore (vedi oltre) che raccoglie e organizza queste informazioni una determinata posizione nella matrice dei dati (vedi oltre); ed infine elaborare delle procedure per individuare e correggere gli errori che si possono produrre nelle varie fasi del passaggio da una situazione reale a un dato su un vettore della matrice dei dati.

Dunque **la definizione operativa è quel complesso di regole che permette di rilevare empiricamente le informazioni che ci interessano. Essa trasforma gli stati in *dati* e le proprietà in *variabili*.**

***Variabile*** è quindi una proprietà di cui sia stata data una definizione operativa, permettendo così di trasformare una serie di situazioni reali (stati) in una serie di dati su un vettore.

***Vettore*** è una sequenza ordinata di informazioni relative allo stesso referente (il caso, o la proprietà, come vedremo tra poco). Nella ricerca sociale queste informazioni sono presentate in forma di valori simbolici (abitualmente cifre).

***Dato*** è il valore convenzionale assegnato allo stato di uno specifico caso su una specifica proprietà sulla base della sua attribuzione a una modalità della corrispondente variabile.

Due fasci di vettori paralleli, intersecandosi in modo che uno sia perpendicolare all’altro, costituiscono una ***matrice***. L’incrocio di ogni singolo vettore con un vettore perpendicolare costituisce una ***cella*** della matrice. Il numero delle celle è uguale al prodotto del numero dei due fasci di vettori paralleli che si incrociano. Se, ad esempio, un fascio di 10 vettori paralleli si incrocia perpendicolarmente con un fascio di 12 vettori paralleli, le celle della matrice che ne risulta saranno 10 x 12 = 120.

Nelle scienze umane i ***vettori-riga*** (detti così perché costituiscono le righe della matrice) hanno abitualmente per referente dei **casi**; i ***vettori-colonna*** (le colonne della matrice) hanno per referente delle **variabili**. Una matrice così costruita si chiama ***matrice dei dati*** perché le sue celle contengono, appunto, dei dati (esistono altri tipi di matrice; durante il corso le incontreremo). Due sono le condizioni necessarie affinché le informazioni afferenti ad un certo insieme di casi possano essere organizzate nella forma della matrice dei dati:

1. **L’unità d’analisi deve essere sempre la stessa**: per esempio deve trattarsi di informazioni raccolte tutte su individui, oppure tutte su comuni o su famiglie; non si può costruire una matrice dei dati in cui alcuni referenti delle righe sono individui, altre comuni, etc.
2. **Su tutti i casi studiati devono essere state rilevate le stesse informazioni**; nella matrice le righe hanno la stessa lunghezza e contengono le stesse variabili; non è possibile costruire una matrice dati se su un certo numero di casi sono state raccolte certe informazioni e su altri casi ne sono state raccolte delle altre.

Ecco un esempio di matrice dei dati, detta anche ***matrice casi per variabili***.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SOGGETTI | genere | età | titolo di studio | reddito annuo | appartenenza religiosa | etc. |
| Tizia |  |  |  |  |  |  |
| Caio |  |  |  |  |  |  |
| Sempronia |  |  |  |  |  |  |
| Filiberto |  |  |  |  |  |  |
| etc. |  |  |  |  |  |  |

Il ***tracciato-record*** (con ‘record’ in inglese si intende la riga della matrice) indica la posizione di ogni variabile nella matrice (per esempio informa che la variabile ‘genere’ si trova nella colonna 1 della matrice).

Il ***codice*** (o libro-codice; in inglese ***codebook***) riporta la sequenza delle variabili che formano i vettori-colonna della relativa matrice. Per ciascuna variabile categoriale e ordinale (vedi oltre) il codice riporta l’elenco delle modalità; ciascuna modalità è affiancata dal valore numerico che le è stato assegnato. Per ciascuna variabile cardinale (vedi oltre) il codice riporta il numero di cifre possibile, come l’età in questo esempio di stralcio di *codebook*.

|  |  |  |
| --- | --- | --- |
| COLONNA | VARIABILE | CODICI E MODALITA’ |
| 1-4 | N. caso | 1 - 1500 |
|  |  |  |
| 5 | Genere | 1. Maschio |
|  |  | 2. Femmina |
|  |  |  |
| 6-7 | Età | 18 - 70 |
|  |  |  |
| 8 | Titolo di studio | 0. Nessun titolo |
|  |  | 1. Elementare |
|  |  | 2. Licenza media |
|  |  | 3. Diploma |
|  |  | 4. Laurea |
|  |  | 9. Non risponde |
|  |  |  |
| 9 | Nell’ultimo anno lei è andato in chiesa? | 0. No, mai |
|  |  | 1. due-tre volte l’anno |
|  |  | 2. una volta al mese |
|  |  | 3. due-tre volte al mese |
|  |  | 4. una o più volte la settimana |
|  |  | 9. non risponde |
|  |  |  |

I criteri con cui si assegnano i codici alle modalità, quindi la definizione operativa, **cambiano a seconda della natura che attribuiamo alla corrispondente proprietà** (questo è un argomento fondamentale del corso e vi verrà dedicato ampio spazio – nel manuale è trattato nel capitolo 7), e quindi della maniera in cui ne rileviamo gli stati.

Le **proprietà** sono **di tipo diverso** (il quadro sinottico dei tipi di proprietà si trova nell’ultima pagina del cap. 7 del manuale). Una prima fondamentale distinzione è quella tra ***proprietà discrete*** e ***proprietà continue***.

Le prime hanno un numero finito di stati, chiaramente distinti (o distinguibili) l’uno dall’altro (esempi: nazionalità, titolo di studio, numero di componenti di una famiglia).

Le seconde hanno un numero infinito di stati impercettibilmente diversi l’uno dall’altro, come i punti di un *continuum* (esempi: età, reddito, distanza chilometrica dal luogo di lavoro, livello di xenofobia).

Le proprietà discrete possono essere: **categoriali** (es.: religione professata), **ordinali** (es.: gerarchia delle qualifiche lavorative), **cardinali** (es.: numero di figli).

Le proprietà continue possono essere: **misurabili** (es.: età), **non misurabili** (es.: un’opinione di favore/sfavore verso un oggetto).

A diversi tipi di proprietà corrispondono diversi tipi di variabile: dalle proprietà categoriali possono essere costruite **variabili categoriali**, da quelle ordinali **variabili ordinali**, e da quelle cardinali **variabili cardinali naturali** (derivate da conteggio).

Dalle proprietà continue misurabili possono essere costruite **variabili cardinali metriche** (derivate da misurazione), da quelle continue non misurabili **variabili quasi-cardinali** (molto importanti nelle scienze umane: se ne parlerà in seguito).

Di seguito illustro sinteticamente i diversi tipi di variabile.

Le ***variabili categoriali*** derivano da una operazione di ***classificazione*** (la definizione operativa di questo tipo di proprietà è una classificazione) degli stati della proprietà. I numeri assegnati alle modalità (ma potrebbero essere delle lettere o qualsiasi altro segno) non hanno alcun significato numerico, ma sono dei puri nomi. Anche se, prendendo ad esempio la variabile religione, assegniamo 1 alla modalità ‘cattolica’, 2 a ‘protestante’, 3 a ‘ebrea’, 4 a ‘musulmana’, etc., queste cifre sono del tutto arbitrarie e sono prive di significato numerico: se sulla variabile religione a un certo individuo è stato assegnato il valore 4, ciò non significa che quell’individuo ha una religione superiore a colui al quale è stato attribuito il valore 2. Le sole relazioni possibili fra le modalità di una variabile categoriale sono le relazioni di uguaglianza (due persone cattoliche sono uguali per religione) e di disuguaglianza (un cattolico ed un protestante sono diversi per religione). Una variabile categoriale avrà un numero finito (e di solito piuttosto limitato) di stati.

Le ***variabili ordinali*** nascono da una operazione di ***classificazione*** e poi di ***ordinamento*** degli stati sulla proprietà e i valori numerici attribuiti alle modalità devono sottostare al vincolo di ***monotonicità*** con gli stati: se lo stato B (ad esempio la licenza media) è considerato superiore allo stato A (la licenza elementare), il valore numerico assegnato alla modalità B dev’essere maggiore di quello assegnato alla modalità A. I valori numerici attribuiti alle modalità mantengono le proprietà ordinali dei numeri, nel senso che se attribuiamo 1 a ‘molto insoddisfatto’, 2 a ‘poco soddisfatto’, 3 a ‘soddisfatto’ e 4 a ‘molto soddisfatto’, possiamo dire che chi ha un punteggio 4 e più soddisfatto di chi ha un punteggio 2, **ma non possiamo dire che è soddisfatto il doppio**. Quindi, fra le modalità di una variabile ordinale è possibile instaurare, oltre alle relazioni di eguaglianza e disuguaglianza, anche le relazioni d’ordine, e cioè di maggiore e di minore. Però, nulla possiamo affermare sulla distanza esistente fra i valori e quindi su **quanto** un valore sia maggiore di un altro; **non possiamo assumere che gli intervalli fra un valore e il successivo siano uguali.**

Le ***variabili cardinali*** derivano da un’operazione di ***conteggio*** (possiamo contare il numero di figli delle famiglie oggetto del nostro studio: avremo una variabile cardinale naturale) oppure di ***misurazione*** (possiamo misurare il tempo impiegato da ciascun individuo del nostro campione per raggiungere il proprio luogo di lavoro o di studio: avremo una variabile cardinale metrica).

Quando esiste un’***unità di conto*** (cosiddetta ‘naturale’: due mele, tre alberi, dieci pietre, etc.) gli stati sulla proprietà vengono determinati contando oggetti o eventi (numero di esami sostenuti, numero di laureati nel comune, numero di elezioni tenute, etc.). I valori numerici nelle corrispondenti celle della matrice hanno pertanto tutte le caratteristiche dei numeri interi: tra le altre, il fatto che non esiste alcunché di intermedio fra due interi adiacenti.

Quando, invece, una proprietà è considerata continua (sono concepibili infiniti stati intermedi fra due stati comunque scelti; ad esempio: età, tempo impiegato per, distanza da) ed è stata stabilita un’***unità di misura*** (così come è stato fatto in congressi internazionali in cui si è deciso di adottare strumenti come il metro, l’orologio, ancora prima il calendario, etc.), si assegnano valori agli stati della proprietà mediante misurazione. In questi casi i valori delle variabili fruiscono di un pieno significato numerico. Nel caso della misurazione, l’esistenza di un’unità di misura fa sì che si conoscano le distanze tra i valori (fra 20 e 22 gradi c’è la stessa distanza che c’è fra 16 e 18 gradi, e tale distanza è la metà di quella esistente fra 40 e 44 gradi) perché **gli intervalli fra un valore e il successivo sono uguali**. Questo fatto permette di effettuare operazioni di addizione e di sottrazione fra i valori, di calcolare le distanze fra loro intercorrenti, e di applicare a queste differenze le quattro operazioni aritmetiche, aprendo la strada all’uso di tutte le principali procedure della statistica.

Nel caso del conteggio tutto ciò è ancora più legittimo in quanto esiste (in natura) uno “zero assoluto” (mentre le unità di misura sono convenzionali, e quindi anche lo zero – assenza di quella determinata caratteristica – è convenzionale).

Queste differenze fra le variabili, in particolare il fatto che su di esse non siano consentite le stesse operazioni, fanno sì che i dati dei tre tipi di variabili **debbano essere raccolti e analizzati con procedure del tutto diverse, fin dai livelli elementari**. Per esempio, nell’analisi monovariata (sarà definita fra poco), se la variabile è cardinale si può calcolare la media aritmetica, cosa che non si può fare se la variabile è categoriale (si può calcolare il reddito medio di un certo insieme di persone, ma non la religione media; si potrà individuare la religione più diffusa, ma è una cosa diversa).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **tipi di proprietà** | tipi di **definizione operativa** (trasforma la proprietà in variabile; cioè trasformagli stati in dati) | tipi di **variabile**(in sede di **raccolta**dei dati) | tipi di **variabile**(in sede di **analisi** mono e bivariata dei dati) | tipi di **variabile**(in sede di **analisi**multivariata dei dati) |
| considerate **discrete** | *categoriali*  | classificazione | categoriali | categoriali | categoriali |
| *ordinali* | classificazione | ordinali | ordinali | categoriali |
| *cardinali* | conteggio | cardinali naturali | cardinali | cardinali |
| considerate **continue** | *misurabili* | misurazione | cardinali metriche | cardinali | cardinali |
| ***non*** *misurabili* | *scaling* auto-ancorante | quasi-cardinali | cardinali | cardinali |
| *scaling* classico | ordinali | ordinali | categoriali |
| non rilevate direttamente |  | operazioni matematiche su variabili cardinali | cardinali derivate | cardinali | cardinali |

Il livello di analisi più elementare è quindi quello dell’***analisi monovariata***. Una tecnica di analisi si dice ‘monovariata’ se si occupa della distribuzione dei dati di un vettore fra le modalità della corrispondente variabile (detto sinteticamente: si occupa della distribuzione di una variabile), e delle principali caratteristiche di questa distribuzione. Si analizzano le variabili singolarmente prese, cioè ad una ad una senza porle in relazione tra loro. Analizzando la variabile “regione di residenza”, ad esempio, si può voler accertare quanti dei soggetti presenti nella matrice come casi risiedono in Lombardia, quanti in Sicilia, e così via; oppure: quanti hanno come titolo di studio la licenza elementare, quanti la licenza media, e così via; oppure: quanti sono femmine e quanti maschi; o ancora: qual è l’età media di quel gruppo di soggetti. La stessa analisi si può fare per tutte le variabili inserite nella matrice.

Una tecnica di analisi si dice ***bivariata*** se si occupa della distribuzione di due variabili congiuntamente considerate (distribuzione doppia o congiunta). Proseguendo con lo stesso esempio, si può accertare quanti di quei lombardi (di quei siciliani, etc.) sono laureati, quanti sono diplomati, e così via, e quindi stabilire se c’è una ***relazione***, nel nostro campione di individui, fra regione di residenza e titolo di studio. Oppure potremmo voler indagare sulla relazione fra regione di residenza e reddito annuo (ad esempio se i residenti nelle regioni settentrionali abbiano mediamente un reddito più alto di quello dei residenti nelle regioni meridionali); o ancora: su quella fra titolo di studio e reddito.

Quando le due variabili poste in relazione sono di tipo categoriale, per accertare la relazione si costruisce una *tabella di contingenza*, strumento tipico dell’analisi bivariata per variabili categoriali e ordinali (se ne parlerà durante il corso; capitolo 9 del manuale).

Se volessimo inserire una terza variabile – analisi trivariata – potremmo ad esempio accertare se c’è una relazione, e di che tipo, tra regione di residenza, genere, reddito.

E’ opportuno sottolineare che la relazione di cui si parla è di tipo statistico, cioè ***tendenziale***. Se affermiamo che c’è una relazione fra titolo di studio e reddito, nel senso che fra gli individui da noi analizzati al crescere del titolo di studio cresce anche il reddito, questa affermazione vale a livello medio: mediamente, nei casi studiati, le persone che hanno titolo di studio più elevato hanno anche reddito più elevato. Ciò non toglie che fra i nostri casi possano darsi anche soggetti con la terza elementare che guadagnano più di altri che sono laureati; ma, in generale, se accertiamo una relazione, ed è positiva, è più probabile che un individuo con laurea guadagni più di un individuo con titolo di studio inferiore.

La tecnica di analisi si dice ***multivariata*** se si occupa della distribuzione congiunta di tre o più variabili e serve a studiare le loro relazioni. Come è evidente, un modello che si limiti all’analisi bivariata può darci informazioni anche interessanti ma piuttosto riduttive del fenomeno che consideriamo. Con un’analisi multivariata potremmo sapere, ad esempio, se nel nostro campione esistono relazioni (e di che tipo) tra genere, titolo di studio, reddito, residenza, attività lavorativa. Quindi, quando la natura delle variabili lo consente, gli studiosi profittano delle sempre più potenti tecniche statistiche multivariate proposte negli ultimi decenni.

In ogni modo, l’analisi monovariata è un passaggio obbligato per prendere confidenza con i dati della ricerca e, come vedremo oltre, ha delle importanti funzioni.

Il modo più semplice ed efficace per rappresentare come **una** determinata variabile si distribuisce fra i casi studiati è quello di costruire una tabella chiamata ***distribuzione di frequenza***. E’ una rappresentazione nella quale ad ogni modalità della variabile viene associata la frequenza con la quale essa si presenta nei dati analizzati. In altre parole: il numero dei dati che presentano una determinata modalità su una variabile si dice frequenza di quella modalità, e l’elenco ordinato di tutte le modalità di una variabile, ciascuna con la sua frequenza, si dice distribuzione di frequenza di quella variabile.

Ecco un esempio di distribuzione di frequenza:

|  |
| --- |
| **Variabile Titolo di studio** |
| **Valore** | **Modalità** | **Frequenza** |
| 0 | nessuno | 15 |
| 1 | licenza elementare | 42 |
| 2 | licenza media | 191 |
| 3 | diploma | 165 |
| 4 | laurea | 45 |
| 9 | manca l’informazioneTotale | 22480 |

Quella presentata sopra è una distribuzione di ***frequenze assolute*** (o valori assoluti). Le chiamiamo ‘assolute’ in quanto abbiamo riportato accanto ad ogni valore della variabile il numero di casi che presentano quel valore, semplicemente contandoli. Tuttavia, se vogliamo confrontare la distribuzione di frequenza della variabile titolo di studio nel gruppo studiato con la distribuzione della stessa variabile in un altro gruppo, di diversa numerosità, non possiamo effettuare direttamente il confronto. Se il secondo gruppo è costituito da 1.200 persone, per poter effettuare il confronto al fine di sapere se il secondo gruppo presenta livelli di istruzione diversi dal primo, dobbiamo riferire le due distribuzioni ad un totale comune. Per annullare l’effetto della numerosità dei casi di una distribuzione di frequenza, possiamo semplicemente dividere ogni singola frequenza assoluta per il numero totale di casi della distribuzione. Il valore ottenuto si chiama ***proporzione***; la proporzione di casi in una data classe è data dal numero di casi della classe diviso il numero di casi totale (ad es.: il numero di soggetti in possesso di licenza elementare, 42, diviso il totale dei casi, 480. Il risultato è 0,08). Con questa operazione abbiamo relativizzato le frequenze ad un totale pari all’unità. Il modo più comune di relativizzare le frequenze è tuttavia quello di riportarle ad un totale pari a 100. Basta moltiplicare le proporzioni per 100 per ottenere le ***percentuali*** (nell’esempio precedente: il numero di soggetti in possesso di licenza elementare è l’8% del campione). Le frequenze espresse in percentuali si chiamano ***frequenze relative***.

Può essere rappresentata mediante una distribuzione di frequenza sia una variabile categoriale, sia una variabile ordinale (come quella dell’esempio presentato sopra), sia una variabile cardinale. In quest’ultimo caso, però, i valori della variabile cardinale vengono raggruppati in classi, aggregando i valori adiacenti.

Ecco un esempio di distribuzione di frequenza di una variabile cardinale:

|  |
| --- |
| **Variabile Età** |
| **Valore** | **Classi d’età** | **Frequenza** |
| 0 | 18-30 | 15 |
| 1 | 31-40 | 42 |
| 2 | 41-50 | 191 |
| 3 | 51-60 | 165 |
| 4 | 61-70 | 45 |
| 9 | manca l’informazioneTotale | 22480 |

Oltre che con una tabella, la distribuzione di una variabile può essere rappresentata integralmente anche con un grafico: vedremo oltre le rappresentazioni grafiche usate.

Informazioni ancora più sintetiche sulle variabili possono essere fornite calcolando e riportando ***valori caratteristici*** per l’analisi monovariata (vedremo fra qualche pagina quali sono i valori caratteristici adatti ai diversi tipi di variabile, come si calcolano e quali informazioni ci danno), e ***coefficienti*** per l’analisi bi- o multi-variata.

Il ricorso a valori caratteristici e/o coefficienti è necessario ogniqualvolta la relativa informazione dev’essere inserita in una formula matematica per calcolare altri valori caratteristici o coefficienti (di solito relativi a relazioni più complesse). Vi si ricorre, però, anche al di là di questa necessità, per risparmiare tempo e spazio tipografico: riportare media e varianza richiede assai meno tempo di quanto ne richieda disegnare un grafico e lo spazio occupato sarà di mezza riga anziché di almeno mezza pagina. Tuttavia, molti valori caratteristici e tutti i coefficienti hanno qualche significato solo per i pochi che conoscono la statistica; solo un esperto può capirne appieno il messaggio, valutando correttamente quello che dicono, notando quello che non dicono, tenendo presente quello che non possono dire. Invece, un grafico ben concepito può dare informazioni comprensibili a un insieme molto più ampio di persone. Ecco perché di solito nei mezzi di comunicazione di massa si esibiscono grafici (istogrammi, diagrammi a torta, etc.) e non valori caratteristici, meno che mai coefficienti.

Prima si diceva che l’analisi monovariata, oltre ad essere un passaggio verso forme più complesse di analisi, ha delle importanti funzioni.

La prima funzione è quella di ***controllare la plausibilità dei valori***. La illustro con alcuni esempi. Se la variabile analizzata è il genere, per le cui modalità sono stati previsti i valori 1 (femmina), 2 (maschio), 0 (informazione mancante), ogni altro valore (ad esempio: 3, 4, 27) che compaia nella distribuzione di frequenza (che non viene più costruita a mano contando tutti i valori riportati nella matrice ma automaticamente da un calcolatore elettronico) è dovuto ad un errore. Altri esempi: in una ricerca sui giovani in cui l’età sia stata registrata (com’è abituale) con i numeri interi corrispondenti agli anni, un dato avente valore 65 è certamente frutto di errore. Se l’unità è il comune, un dato avente valore 95 sulla variabile ‘percentuale di maschi sui residenti’ o sulla ‘percentuale di laureati sui residenti’ fa pensare ad un errore e deve essere controllato. In questi casi, esaminando la distribuzione di frequenza possiamo rilevare l’errore e porvi rimedio.

***Controlli di congruenza***. Si possono confrontare le distribuzioni di due variabili (**non è** analisi bivariata ma confronto di due analisi monovariate) per far emergere eventuali incongruenze. Se abbiamo chiesto agli intervistati se sono occupati o meno, e dalla distribuzione di frequenza di questa variabile appaiono 512 occupati, mentre poi dalla domanda sul tipo di lavoro effettuato risultano 550 risposte, è evidente che hanno dato la seconda risposta anche persone non occupate (probabilmente si tratta di pensionati che hanno riferito il loro precedente lavoro; però potrebbe anche darsi il caso che alcuni soggetti che lavorano saltuariamente non si considerano occupati ma poi riferiscono ugualmente la mansione che svolgono): l’ispezione delle distribuzioni di frequenza fa emergere l’errore e permette di correggerlo.

Altri controlli di congruenza si effettuano attraverso l’”***incrocio***” fra due variabili (in questo caso occorre costruire una *tabella di contingenza*, strumento tipico dell’analisi bivariata: se ne parlerà durante il corso; capitolo 9 del manuale): per esempio si può controllare che tutte le casalinghe siano donne, che chi va a messa tutte le domeniche si sia anche dichiarato praticante, che chi ha dichiarato il partito per il quale ha votato abbia anche detto di essere andato a votare, etc. Se qualcuno risulterà maschio e casalinga, oppure (perlomeno finora in Italia) femmina e sacerdote, o ancora sacerdote e sposato, avremo individuato delle combinazioni implausibili da correggere.

In questi casi, come in quello precedente, la correzione dell’errore può essere fatta sia risalendo alle informazioni originarie (per esempio ai questionari) per individuare il valore corretto da riportare nella matrice (se si tratta di errore di codifica); sia modificando i dati in modo che siano logicamente plausibili (se una persona risponde ad una serie di domande sui propri figli anche se in precedenza ha dichiarato di non averne, è presumibile che ci sia stato un errore nella registrazione della domanda iniziale); sia sostituendo il valore implausibile con il valore corrispondente a “dato mancante”.

Un’altra funzione dell’analisi monovariata è quella di ***consentire una valutazione critica del proprio lavoro***. Il lettore ha diritto, oltre che a ricevere risultati e commenti basati su dati depurati per quanto possibile da errori materiali, a prendere visione delle pezze d’appoggio, cioè delle fondamenta elementari dell’intero edificio analitico, per sofisticato che sia. In tal modo disporrà di maggiori elementi per recepire criticamente, anziché passivamente, le considerazioni e le conclusioni di chi ha analizzato i dati. Se pubblichiamo i risultati delle nostre analisi multivariate senza riportare e illustrare le distribuzioni delle variabili che abbiamo impiegato, priviamo il lettore di un elemento necessario a una valutazione critica del nostro lavoro. Ad esempio, **Almalaurea**, la più importante banca dati dei laureati in Italia, riporta nel suo sito tutte le distribuzioni di frequenza dei dati raccolti (basta andare nel sito per accertarsene).

Esaminando le distribuzioni di frequenza, dunque, il ricercatore può vedere come i principali fenomeni si presentano nel campione studiato e può rispondere ai primi rudimentali interrogativi: ad esempio, quanti sono i favorevoli a una certa iniziativa governativa, quanti ritengono che il problema più grave del paese sia la criminalità, quanti invece pensano che sia la disoccupazione, quanti sono gli operai nel campione, quanti i disoccupati, etc. Affinché sia del tutto chiaro, ribadisco ancora che questo tipo di analisi si limita a dirci come ogni variabile è distribuita fra i casi rilevati, senza dire alcunché sulle relazioni fra le variabili. Per esempio, ci dice che nel campione ci sono *tot* praticanti regolari, ma non ci dice se vanno a messa di più i giovani o gli anziani, di più i liberi professionisti degli operai, di più le donne degli uomini, etc.

Si diceva prima che informazioni più sintetiche sulle variabili possono essere fornite calcolando e riportando i ***valori caratteristici***: essi ci possono dire, ad esempio, qual è la religione più diffusa, qual è il reddito medio e se esso è distribuito in maniera equilibrata o disomogenea nel campione. Di tutte le caratteristiche di una distribuzione di frequenza, le più importanti sono due: la sua tendenza centrale e la sua variabilità. Diciamo subito che **i valori caratteristici di tendenza centrale e quelli di variabilità sono diversi a seconda del tipo di variabile** considerata: già da queste primissime operazioni di analisi dei dati lo/la studente/essa può constatare l’importanza della distinzione fra variabili categoriali, ordinali e cardinali.

I ***valori caratteristici di tendenza centrale*** ci dicono qual è, in una distribuzione di frequenza, quello che può essere considerato il baricentro dei suoi valori, vale a dire il valore che meglio di qualsiasi altro esprime la distribuzione quando si decidesse di sintetizzarla in un unico numero.

Questi valori ci segnalano quella che è la modalità centrale di una distribuzione ma nulla ci dicono del modo di collocarsi delle altre modalità attorno a questo centro. In particolare, nulla ci dicono sulla variabilità di questi valori: per esempio, due campioni possono presentare lo stesso reddito medio, ma avere distribuzioni completamente diverse (una situazione di sostanziale equilibrio fra i redditi ed un’altra, invece, di grandi disuguaglianze). Per questo motivo, per sintetizzare compiutamente una distribuzione, ai valori di tendenza centrale dobbiamo accompagnare i ***valori caratteristici di variabilità o di dispersione***.

Si è detto che i valori caratteristici di tendenza centrale e quelli di variabilità sono diversi a seconda del tipo di variabile considerata. Quindi li presentiamo distinti per tipo di variabile.

**Valori caratteristici per variabili categoriali**.

Per queste variabili, l’unico valore di tendenza centrale che è individuabile è la ***moda***. Essa è la modalità di una variabile che si presenta nella distribuzione con maggiore frequenza. Una distribuzione può anche essere chiamata ‘bimodale’, quando vi sono due modalità che si presentano con notevole frequenza, anche se una delle due ha una frequenza inferiore a quella dell’altra. Individuare la moda di una distribuzione di frequenza corrisponde a quello che facciamo quando, nel presentare i risultati di una certa domanda, diciamo, ad esempio, che la maggioranza ha risposto in un certo modo (per esempio, sulla domanda relativa a quali sono i problemi che più preoccupano il cittadino, la maggioranza ha risposto “la criminalità”). Questa informazione non dice molto sulla configurazione della distribuzione di frequenza perché non tiene conto della distribuzione degli altri valori. D’altra parte, non essendo possibile, come detto, effettuare operazioni sui valori della variabile categoriale, nulla si può dire sulla sua variabilità. Per questa ragione sovente è meglio presentare la distribuzione di frequenza o un grafico che la rappresenti.

Diciamo che una variabile categoriale ha una distribuzione massimamente *omogenea* quando tutti i casi si presentano con la stessa modalità: se, ad esempio, nel caso della variabile “religione” il 100% delle persone è cattolica. Viceversa, è massimamente eterogenea se i casi sono equidistribuiti fra le modalità: ad esempio, se la religione ha cinque modalità (cattolicesimo, protestantesimo, islam, religioni orientali, altro) e noi abbiamo il 20% in ciascuna di esse. Una distribuzione quindi è tanto più omogenea quanto più essa è concentrata su una (o poche) modalità o, per dirla in altro modo, quanto maggiore è lo squilibrio fra le modalità. La statistica propone diversi ***valori di equilibrio/squilibrio***. Tra qualche pagina, nello specchietto riassuntivo, presenteremo quello più semplice e intuitivo.

**Valori caratteristici per variabili ordinali**.

Se la variabile è ordinale, si può naturalmente individuare la ***moda*** della distribuzione. Ma poiché ora, diversamente dalla situazione precedente, i casi sono ordinabili, è possibile individuare anche un altro valore di tendenza centrale: la ***mediana***. In termini matematici, data una qualunque serie di cifre collocate in ordine di grandezza, la mediana è quella cifra che bipartisce la serie in modo da lasciare lo stesso numero di cifre dalle due parti.

Se però dobbiamo individuare **la mediana della distribuzione di frequenza di una variabile**, la procedura è lievemente più complessa. Immaginiamo di ordinare, sulla base di una variabile ordinale, tutti i casi di una distribuzione, dal minore al maggiore. Dopo averli ordinati, si prenda il caso che viene a trovarsi al centro della distribuzione. Se, per esempio, i casi sono 21, si tratterà dell’11° caso: la mediana della distribuzione è la modalità assunta dalla variabile su questo caso. Dunque la mediana di una variabile **non è** il caso ma è **la modalità del caso** che occupa il posto di mezzo nella distribuzione ordinata dei casi secondo quella variabile. E’ facile l’individuazione della mediana di una distribuzione di frequenza se si sono calcolate le percentuali. Sarà facilissimo se si è già costruito una tabella con le ***percentuali cumulate*** (se ne parlerà oltre): la mediana è quella categoria in cui rientra la percentuale 50.

La mediana può dare più informazioni della moda perché tiene conto anche dell’ordine delle categorie e, inoltre, è molto meno sensibile della media ai valori estremi di una distribuzione (in quanto è un valore posizionale, ossia non considera tutti i valori di una distribuzione) ma quest’ultimo pregio ha un’importanza limitata per l’analisi delle variabili ordinali perché i valori assegnati alle modalità sono quasi sempre assegnati dal ricercatore, che di solito non ha alcun motivo per introdurre valori estremi ai lati della distribuzione.

Mentre la mediana rileva la tendenza centrale di una distribuzione, altri valori caratteristici posizionali (i quantili) rilevano piuttosto la sua dispersione attorno a questa tendenza centrale: di questi, i più usati sono i due ***quartili*** (vedi oltre, nello specchietto riassuntivo). La differenza fra i valori dei due quartili (***differenza interquartile***) può essere usata per rilevare la dispersione della variabile. Altri valori caratteristici posizionali (***decili***, ***percentili***) sono poco utili per variabili ordinali, che di solito, come detto poc’anzi, hanno un numero piuttosto limitato di categorie.

I ***quantili***, come la mediana, sono detti ***valori di posizione***, nel senso che sono determinati tenendo conto sia delle frequenze sia dell’ordine delle categorie. In altre parole, derivano dalla posizione dei casi (sono le modalità o i valori corrispondenti a casi collocati in una determinata posizione) e sono determinati senza compiere alcuna manipolazione sui dati.

**Valori caratteristici per variabili cardinali**.

Le categorie di una variabile cardinale sono *mutuamente esclusive* (è uno dei requisiti dell’operazione intellettuale chiamata ‘***classificazione***’: se ne parlerà durante il corso; è illustrata nel capitolo 3 del manuale) come le categorie delle variabili categoriali e ordinali. Sono anche *ordinate*, come le categorie di una variabile ordinale. E, in più, abbiamo una precisa nozione della distanza che separa ciascuna categoria dalla precedente (sia che si tratti dell’unità di conto, che è *naturale*, sia che si tratti dell’unità di misura, che è *convenzionale*, cioè proposta e adottata da esseri umani: in questo caso dagli/dalle scienziati/e). Questa nozione ci permette di dare un’interpretazione cardinale alle cifre, che non solo etichettano ma *costituiscono* le modalità.

Tutti i valori caratteristici usati per gli altri due tipi di variabili (**moda**, **mediana**, **quantili**) hanno senso anche se riferiti a variabili cardinali, ma i valori caratteristici tipici di queste ultime sono ***sintetici***: essi sono infatti determinati sulla base delle frequenze di **tutte** le modalità (non così i valori caratteristici menzionati tre righe sopra) e del valore cardinale delle relative etichette numeriche.

Il più familiare dei valori sintetici è la ***media aritmetica*** che, come detto più volte, è calcolabile solo se la variabile è cardinale. Essa misura la tendenza centrale di una distribuzione.

Per quantificare la dispersione dei valori di una variabile cardinale sono stati proposti vari modi. Il più elementare è il ***campo di variazione*** (in inglese ***range***), cioè la differenza tra il valore massimo e il valore minimo della distribuzione. Tutti gli altri valori proposti specificamente per rilevare la variabilità delle distribuzioni delle variabili cardinali sono, invece, sintetici, ovvero sfruttano tutte le informazioni che abbiamo su di esse (cioè i valori numerici di ogni singola modalità). Essi sono quattro – ***devianza***, ***varianza***, ***scarto-tipo*** (***deviazione standard***) e ***coefficiente di variazione*** – e vengono illustrati nel relativo specchietto riassuntivo.

**Valori caratteristici delle distribuzioni di variabili categoriali**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Denominazione del valore caratteristico* | formula o metodo di calcolo | concetti base | pregi | difetti |
|  *moda* | è la categoria con la frequenza più alta (o percentuale più alta) | rileva la tendenza centrale della distribuzione individuando la categoria in cui i dati si concentrano | facilissima da individuare | si basa solo sulla frequenza modale |
| *Sq: misura di equilibrio/squilibrio tra le frequenze di tutte le categorie*  |  k Sq =∑ pj2 J=1dove pj= proporz. di casi che ricadono in quella categoria e k= n. di categorie  | si basa sulla probabilità che gli stati di due casi diversi siano stati assegnati alla stessa modalità. Essa è più alta quanto più è basso il n. delle categorie e quanto più sono squilibrate le frequenze a favore di una categoria. Somma le probabilità relative a tutte le varie categorie (pj2) | prende in considerazione le frequenze di tutte le categorie: utile, data l’importanza di una distribuzione equilibrata per l’analisi bi- e multivariata delle variabili categoriali | varia da 0 a 1 (ambedue = massimo squilibrio). Il valore corrispondente al minimo squilibrio varia a seconda del numero delle categorie: è uguale a 1/k |
| *Sq normalizzato* | Sqnorm =(k\*Sq-1) / (k-1) | si pone il valore di Sq in rapporto con i valori minimo e massimo che esso può assumere dato quel certo numero di categorie | prende in considerazione le frequenze di tutte le categorie. E’ un indice relativo: varia da 0 (minimo squilibrio) a 1 (massimo squilibrio) |  |
| *Eq: indice relativo di equilibrio* |  Eq =1-[(k\*Sq-1)/k-1)] | è l’inverso di Sq: deriva da questo calcolandone il complemento a 1 | stessi pregi di Sqnorm |  |

I valori caratteristici di equilibrio/squilibrio vengono presentati per conoscenza ma non saranno oggetto d’esame.forme di **rappresentazione grafica** delle distribuzioni di frequenza di **variabili categoriali**

 **pregi difetti**

|  |  |  |
| --- | --- | --- |
| **Rappresentazioni lineari** | * facili da disegnare e da interpretare
* facilitano la comparazione tra frequenze di categorie diverse
 | * introducono implicitamente ordine fra le categorie
* difficile valutare la quota di una categoria sul totale
 |
| * Istogramma
 | * facilita la comparazione tra le frequenze di categorie adiacenti
 | * suggerisce contiguità fra le categorie
 |
| * Diagramma a barre
 | * sottolinea l’autonomia semantica delle categorie
 | * difficilissima la percez. della % di una categoria sul totale
* difficilissima l’aggregazione “ad occhio” di categorie
 |
| **Rappresentazioni circolari** | * Suggeriscono di *meno* un ordine tra le categorie
 | * Laboriose da disegnare
* Interpretabilità non immediata
* Difficile la comparazione tra frequenze di categorie diverse
 |
| * Diagramma torta
 | * Facilita la percezione della quota di una categoria sul totale
* Facilita l’aggregazione ad occhio fra categorie adiacenti
 | * Laboriosissimo da disegnare
* Suggerisce contiguità fra le categorie
 |
| * Diagramma a raggi
 | * Sottolinea l’autonomia semantica
 | * Difficilissima l’aggregazione ad occhio di categorie adiacenti
* Difficile la comparazione tra frequenze di categorie diverse
 |

**Obiettivi** della rappresentazione **e forme preferibili**

sottolineare autonomia semantica delle categorie diagramma a raggi, diagramma a barre

evitare di introdurre ordine fra le categorie rappresentazioni circolari

facilitare comparazione tra frequenze di

categorie diverse rappresentazioni lineari

facilitare comparazione tra frequenze di

categorie adiacenti istogramma

facilitare percezione della quota di

una categoria sul totale diagramma a torta

facilitare aggregazione “ad occhio” di diagramma a torta

frequenze di categorie adiacenti

facilitare comprensione da parte di inesperti rappresentazioni lineari

disegnare rapidamente diagramma a barre e a raggi

## **Valori caratteristici delle distribuzioni di variabili ordinali**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Denominazione del valore caratteristico* | come si individua o si calcola | concetti base | pregi | difetti |
| *moda* | è la categoria con la frequenza più alta. (Una distrib. può essere uni/bi/tri-modale) | individua la categoria in cui i dati si concentrano | facilissima da individuare | non dà alcuna informazione sulle frequenze dalle altre categorie |
| *mediana* | è la categoria per la quale P < 50% e S < 50% dove P = tot. frequenze nelle categ. che la precedono e S = tot. categ. che la seguono  | oltre che delle frequenze, tiene conto dell’ordine delle categorie (valore posizionale). Se si calcolano le percentuali cumulate, la mediana è quella categoria in cui rientra la percentuale 50 | non è sensibile ai valori estremi (cosa che può anche essere un difetto, ma comunque è poco rilevante per l’analisi delle variabili ordinali) | essendo un valore posizionale (non tiene conto dell’ ampiezza degli intervalli) può dare informazioni ingannevoli(ma la cosa è poco rilevante per l’analisi delle var. ordinali) |
| *quartili**(decili, percentili)* | **1° quartile**: è la categoria per la quale P < 25% e S < 75%. **2° quartile**: coincide con la mediana.**3° quart**.: è la categoria per la quale P < 75% e S < 25%.  | rilevano la dispersione dei dati attorno alla mediana individuando specifici valori collocati in particolari punti della distribuzione (valori posizionali) | come per la mediana. |

## Per le variabili ordinali, distribuzioni di frequenza e ***istogrammi*** sono le forme più semplici di rappresentazione. Non sono usati i diagrammi a barre, a nastri, o a colonne perché barre, nastri e colonne appaiono separati, mentre le categorie ordinate sono quasi sempre opportunamente rappresentate come contigue. Non sono usate le forme circolari perché la loro principale funzione è proprio quella di evitare di suggerire un ordine fra le categorie quando tale ordine non c’è, mentre nel caso in esame l’ordine c’è, ed è opportuno sottolinearlo visivamente.

## Per questo motivo una forma altrettanto adeguata dell’istogramma è il ***diagramma a bandiera***, che è poi un istogramma ruotato di 90°, cioè composto di fasce orizzontali anziché di colonne. Il diagramma a bandiera è usato spesso per confrontare le distribuzioni della stessa variabile in due popolazioni o sottogruppi diversi (ad esempio, il partito votato dai maschi e quello votato dalle femmine). Peraltro, la rappresentazione che rispecchia meglio la natura di una proprietà con categorie ordinate è ***l’istogramma di composizione*** (in inglese ***mosaic bar***). Esso è costituito da una sola colonna, divisa in fasce di altezza proporzionale alle frequenze delle varie categorie.

## **Valori caratteristici delle distribuzioni di variabili cardinali**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| *Denominazione del valore caratteristico* | come si individua o si calcola |  concetti base | pregi | difetti |
| *moda* | è la categoria con la frequenza più alta.  | più della moda in sé, sono rilevanti i concetti di distribuz. unimodale, bimodale, trimodale.  |  | non dà alcuna informazione sulle frequenze delle altre categorie |
| *mediana, quartili, decili, percentili, minimo e massimo* | **mediana:**P < 50%, S < 50%**1° quartile:**P < 25%, S <75%**1° decile:**P < 10%, S <90%**1° percentile:**P < 100%, S <99% | tutti questi valori posizionali sono utili anche per l’analisi delle variabili cardinali. I perc. solo se le modalità e i dati sono molti. Per le var. quasi-cardinali, min e max non hanno interesse in quanto noti a priori e i decili lo hanno solo per le var. generate da tecniche basate sui segmenti. | i valori posiz. non tengono conto dell’ampiezza degli intervalli, e quindi non sono sensibili ai valori estremi (cosa che può essere un pregio o un difetto) |
| *campo di variazione* (range) | differenza tra il valore massimo e il valore minimo di una distribuzione | è il modo più elementare per quantificare la dispersione dei valori di una var. card. Sono state proposte (e usate) diverse varianti basate su questo stesso principio. |  |  |
| *media aritmetica* | ⎯ N X =∑ Xi / N i=1(dove Xi è il valore di un caso qualsiasi sulla var. X e N è il numero dei casi) | è un valore sintetico. In mancanza di ogni altra informazione, è la migliore previsione possibile sul valore di un qualsiasi Xi. La distanza di un valore dalla media aritmetica si dice *scarto*. La somma degli scarti è 0 (∑ Xi = 0) | considera le frequenze di tutte le modalità e il valore cardinale delle relative etichette numeriche | è molto sensibile ai valori estremi. In certi casi potrebbe essere opportuno escluderli dalla distribuzione e considerarli a parte |
| *devianza* | ∑ xi2dove xi = scarti dalla media aritmetica | come i tre che seguono, è un valore sintetico: dà un’ informazione sintetica sulla dispers. dei dati di una var. card. attorno alla loro media. È la somma dei quadrati degli scarti dalla media. Dipende, oltre che dalla dispersione dei dati, dal numero dei casi | come per la media aritmetica | si basa sul principio gaussiano (eliminare il segno elevando al quadrato gli scarti) che poggia sull’ assunto che una distribuzione abbia una forma campanulare |
| *varianza* | ∑ xi2 / NN è il numero dei casi | si divide per N la formula della devianza. Si usa per confrontare distribuz. con differente numero di casi. Come la devianza, è una grandezza quadratica ed è un quadrato ideale, media di tutti i quadrati costruiti sugli scarti | si neutralizza l’effetto del numero dei casi normalizzando la devianza rispetto ad essi | come per la devianza: l’elevazione al quadrato degli scarti conferisce un peso eccessivo ai valori estremi; grave quando la distribuz. è fort. asimmetrica  |
| *scarto-tipo* (deviazione standard) | S =√ ∑ xi2 / N  | estraendo la radice quadrata della varianza si ottiene la corrispondente grandezza lineare, cioè il lato del quadrato ideale.  | si usa quando serve una grandezza lineare; ad es. per confrontarla con un’altra grandezza lineare (come la media o gli scarti semplici) | come per la devianza e la varianza |
| *coefficiente di variazione* |   ⎯ V = sx /X | divide lo scarto-tipo per la media (che rappresenta la grandezza media dei punteggi) | si usa per confrontare le dispersioni di due var. card. che presentano, oltre ad un diverso numero di casi, medie non trascurabilmente diverse. Eliminando l’effetto dell’unità di misura o di conto si possono confrontare le dispersioni di qualsiasi variabile cardinale |  |

Quando i valori sono molti, una forma di rappresentazione usata abitualmente per le variabili cardinali è la ***curva di frequenza***. In ascissa si collocano i valori; in ordinata il numero dei dati che presentano il valore corrispondente in ascissa. Tra i punti si interpola una curva.

Quando i valori non sono molti, rappresentazioni corrette sono l’**istogramma** e il **diagramma a barre**. ***Forme di rappresentazione grafica della media e dello scarto-tipo*** di una variabile cardinale sono presentati nel manuale.

**ESERCIZI**

1. **Dalla matrice dei dati alla distribuzione di frequenza**

La matrice che segue riporta l'**età**, il **genere**, l'**area di provenienza** (NordOvest, NordEst, Centro, Sud), la **professione** (Operaio, Artigiano, Commerciante, Impiegato) e il **livello di istruzione** (Elementari, Medie, medie Superiori, Università) di un piccolo campione di 12 italiani.

|  |  |
| --- | --- |
| ind. |  età gen area prof istr |
| a)b)c)d)e)f)g)h)i)j)k)l) |  35 F S I S  45 F C I U  28 M S C M  17 F S O E  50 M NO A E  52 F S I S  46 F NE I U  39 F S O M  30 M NE O E  66 M C A U  20 F S C E  24 F S I M  |

Presenta le distribuzioni di frequenza delle 4 variabili categoriali

1. **Dalla matrice dei dati alla tabella di contingenza**

La matrice che segue riporta l'età, il genere, l'area di provenienza (NordOvest, NordEst, Centro, Sud), la professione (Operaio, Artigiano, Commerciante, Impiegato) e il livello di istruzione (Elementari, Medie,

Medie Superiori, Università) di un piccolo campione di 36 italiani.

|  |  |
| --- | --- |
| ind. |  età sex area prof istr |
| a)b)c)d)e)f)g)h)i)j)k)l)m)n)o)p)q)r)s)t)u)v)w)x)y)z)à)è)é)ì)ò)ù)&)ç)£)§) |  35 F S I S  45 F C I U  28 M S C M  17 F S O E  50 M NO A E  52 F S I S  46 F NE I U  39 F S O M  30 M NE O E  33 M C A U  20 F S C E  24 F S I M  57 M NO I U  42 F NO C M  40 M NE A E  23 F C I S  39 F S I U  19 F S I M 68 M S O E 25 F NO I U 35 F NE A E 44 M C C S 56 M C I U 21 F S C M 31 F S O M 43 F NO I M 27 M S I S 37 F C A S 48 F NO I U 51 F S C M 32 M NE I E 36 F S I S 34 F C O U 46 M NO A M 70 M C C E 53 F NE O E |

Costruisci la tabella di contingenza che mette in relazione le **due variabili** **categoriali** **genere** e **professione**. Conta e inserisci nella tabella le frequenze delle celle, i marginali di riga e i marginali di colonna (oltre al totale generale). Nell’esempio sotto (un’immaginaria tabella) puoi visualizzare come è costruita una tabella di contingenza.

In questo esempio di tabella di contingenza è riportata la ipotetica distribuzione congiunta di due variabili: **zona di appartenenza dei piccoli comuni (variabile X)** e **tipo di giunta (variabile Y)**. La categoria ‘altro’ della variabile ‘tipo di giunta’ comprende giunte di sinistra, giunte di destra, giunte di centro e miste (derivanti da liste civiche). 200 è il totale generale dei piccoli comuni.

|  |
| --- |
|   |
| X |  | centro-destra | centro-sinistra | altro | **marginali col.** |
| Nord | 15 |  8 | 54 | **77** |
| Centro | 17 | 45 | 11 | **73** |
| Sud e Isole | 23 |  7 | 20 | **50** |
| **margi-nali riga** | **55** | **60** | **85** | 200 |

1. **Valori caratteristici**
2. Individua la moda delle variabili categoriali della matrice precedente
3. Individua la mediana della variabile ordinale della matrice precedente
4. Calcola la media e lo scarto-tipo della variabile cardinale della matrice precedente
5. **Indici sommatori**

Cinque studenti universitari decidono di confrontare la propria propensione agli studi sociologici. Come indicatori prendono in considerazione i risultati a due esami del corso di laurea che frequentano. Questi indicatori, una volta operativizzati, danno luogo a quattro variabili: voto all’esame di Metodologia della ricerca (VOTO MET.); voto all’esame di Sociologia (VOTO SOC.); numero di volte in cui è stato rimandato/a all’esame di Metodologia (RIM. MET.); numero di volte in cui è stato rimandato/a all’esame di Sociologia.

Ecco la matrice dei dati; i dati relativi alla quarta variabile non vengono riportati in quanto nessuno dei cinque studenti è mai stato rimandato all’esame di Sociologia.

|  |  |  |  |
| --- | --- | --- | --- |
|  | VOTO MET. | VOTO SOC. | RIM. MET. |
| Luigi | 19 | 30 | 2 |
| Francesca | 25 | 20 | 0 |
| Alice | 30 | 26 | 1 |
| Giovanni | 20 | 26 | 3 |
| Alberto | 22 | 28 | 1 |

Costruisci l’**indice** che combina i dati di queste tre variabili, trascrivendo il punteggio di ciascuno studente sull’indice nelle celle del vettore disegnato sotto. Ecco i valori caratteristici che servono per i calcoli:

VOTO MET.: media = 23,2 scarto-tipo = 4

VOTO SOC.: media = 26 scarto-tipo = 3,3

RIM. MET.: media = 1,4 scarto-tipo = 1

|  |  |
| --- | --- |
|  | **indice di pro-****pensione agli****studi sociologici** |
| Luigi |  |
| Francesca |  |
| Alice |  |
| Giovanni |  |
| Alberto |  |

Cita la procedura che hai applicato e descrivi le operazioni che hai effettuato.