



Approfondimenti di statistica

1. Variabili qualitative e variabili quantitative

Nell'ambito di una ricerca statistica le unità statistiche non vengono osservate nella loro globalità ma solo per alcune delle caratteristiche o proprietà che risultano rilevanti per lo scopo della ricerca. Se si considerano, ad esempio, gli studenti di una regione frequentanti la scuola secondaria di secondo grado per ognuno di essi si può determinare: l'età, l'anno di nascita, il luogo di nascita, la professione dei genitori, la statura, la classe frequentata, la lingua straniera studiata, il numero di fratelli, ecc.

Il valore o l'attributo (se il fenomeno è qualitativo) con cui il carattere si manifesta sull'unità statistica è detto modalità. Per esempio il carattere "lingua straniera studiata" si presenta sull'unità *i*-esima osservata, alunna Anna Massa, con modalità "francese".

I caratteri osservati possono essere classificati in qualitativi e quantitativi: la distinzione dipende dalle operazioni che permettono di determinare le modalità del carattere stesso. Le operazioni canoniche con le quali si attribuisce un valore a una modalità sono le seguenti: assegnazione, conteggio e misurazione. L'operazione di assegnazione determina caratteri o variabili dette *qualitative* o *mutabili*, le operazioni di conteggio e misurazione determinano caratteri o variabili dette *quantitative*.

L'operazione di assegnazione può essere distinta in classificazione e ordinamento

a seconda che le categorie o attributi presentino oppure no una relazione d'ordine. L'operazione di misurazione presuppone l'esistenza di un'unità di misura alla quale rapportare l'ammontare di carattere posseduto da ciascun caso. L'operazione di conteggio consiste nell'enumerazione degli oggetti posseduti o con i quali si è in relazione; questa operazione è possibile quando la proprietà è pensabile come quantità discreta. Quindi, a seconda delle operazioni compiute, vengono generate diversi tipi di variabili. La costruzione delle variabili è il passaggio indispensabile per poter applicare la statistica.

Lo studio e la definizione delle variabili come qualitative e quantitative è oggetto della teoria della misurazione. La teoria della misurazione che si è maggiormente diffusa è dovuta a Stevens (1946). Quest'ultimo classifica i dati secondo quattro livelli di scala: nominale, ordinale, di intervalli e di rapporti. I quattro tipi di scala stanno tra loro in una precisa gerarchia: la scala nominale rappresenta il livello di misurazione più basso, sono poche le operazioni matematiche possibili a questo livello di scala; la scala di rapporti è invece il livello più alto, la variabile gode di tutte le proprietà matematiche dei numeri reali e di conseguenza rende più ampio il ventaglio di operazioni matematiche che possono essere compiute.

Scala nominale

Si ha una scala nominale quando tra le modalità della variabile è possibile stabi-

lire solo una relazione di uguaglianza o di disuguaglianza. In questo caso i numeri che rappresentano le modalità della variabile hanno unicamente la funzione di simboli, di etichette, e potrebbero essere sostituiti con qualsiasi altro simbolo numerico e non.

I dati espressi su scala nominale vengono anche indicati come variabili categoriali o variabili qualitative sconnesse o mutabili (esempio: professione dei genitori).

Scala ordinale

Si ha una scala ordinale quando tra le modalità della variabile è possibile stabilire oltre a una relazione di uguaglianza o di disuguaglianza anche una relazione di maggiore o minore. I dati espressi su scala ordinale vengono anche indicati come variabili ordinali, o variabili qualitative (esempio: classe frequentata).

Scala di intervalli

Con il livello di scala di intervalli entriamo nel mondo della misurazione in senso stretto: in questo caso oltre a ordinare i soggetti in relazione al fatto che possiedono in misura maggiore o minore una certa caratteristica possiamo dire quanto sono differenti tra di loro. Un esempio tipico di scala di intervalli sono i termometri.

Scala di rapporti

Nelle scale di rapporti esiste uno zero assoluto, esso non è convenzionale e coincide con l'assenza di carattere. I dati espressi su scala di intervalli e di rapporti vengono anche indicati come variabili cardinali o variabili quantitative. I fenomeni espressi su scala di rapporti (esempio: la statura, il reddito) consentono ogni tipo di elaborazione matematica.

Lo studio e la definizione delle variabili come qualitative e quantitative è oggetto della teoria della misurazione. La teoria della misurazione che si è maggiormente diffusa è dovuta a Stevens (1946)



2. Operatori monovariati

Per analizzare i dati raccolti è necessario che essi siano riassunti in modo da poter essere elaborati. Un primo andamento del carattere osservato si ottiene dall'analisi dei dati disposti in tabelle o distribuzioni di frequenze. Queste ultime possono essere utilmente rappresentate anche tramite grafici con lo scopo di visualizzare in modo più chiaro l'andamento del fenomeno. La scelta del tipo di rappresentazione grafica dipende dalla natura del carattere osservato. Se il carattere è qualitativo il grafico ha una finalità descrittiva, se il carattere è quantitativo la finalità può essere anche interpretativa (ad esempio l'analisi di una serie storica). Dopo l'analisi dell'andamento grafico del carattere, per avere tutte le informazioni possibili, si passa all'elaborazione vera e propria dei dati raccolti.

Molto spesso, e per le variabili quantitative ciò accade il più delle volte, le informazioni relative a una distribuzione vengono riassunte in un singolo numero o statistica. Quest'ultima è il risultato di una formula applicata ai dati: l'operatore statistico. Una delle caratteristiche della statistica è che deve essere appropriata per il livello di scala della variabile considerata, quindi l'operatore statistico sarà scelto tenendo conto del livello di scala della variabile oltre che dell'interesse del ricercatore. Analizzando una sola variabile possiamo distinguere tra operatori che valutano la tendenza centrale e altri che valutano la variabilità. Con il primo si intende un valore caratteristico di una distribuzione che in alcuni casi è un centro geometrico, con il secondo si intende un valore che riassume l'andamento della distribuzione dei casi all'interno dei valori o delle modalità che la variabile può assumere.

2.1 Operatori di tendenza centrale

Prima di trattare singolarmente le principali misure di tendenza centrale, si

Tabella 1		Tipo di variabili		
		Qualitative		Quantitative
		categoriale	ordinale	cardinale
Statistica	moda	si	si	si
	mediana	no	si	si
	media	no	no	si

considera, attraverso uno schema, l'idoneità di ciascuna di esse per il livello di scala della variabile considerata (**tabella 1**).

Si vede che qualsiasi statistica applicabile a un livello di scala inferiore può essere applicata anche ai livelli superiori; in questo caso, però, si accetta di trascurare l'informazione aggiuntiva che si avrebbe utilizzando l'operatore appropriato per quel determinato livello di scala. Quando la variabile è quantitativa possiamo domandarci quale tra gli operatori di tendenza centrale fornisce più informazioni sulla distribuzione.

La moda

La moda è la modalità più ricorrente, cioè il valore della variabile a cui corrisponde la frequenza più elevata. Essa è adeguata a rappresentare la distribuzione di una variabile categoriale.

La mediana

La mediana è la modalità a cui appartiene il caso che divide esattamente a metà la distribuzione. Se n è un numero dispari la mediana è univoca; se n è un numero pari abbiamo due casi mediani rispettivamente il $(n/2)$ -esimo e $(n/2+1)$ -esimo soggetto. Se i due valori centrali appartengono a due modalità differenti si parlerà di modalità mediane anziché di mediana. Questo operatore può essere utilizzato in presenza di variabili ordinali in quanto la sua applicazione presuppone un ordinamento delle modalità del carattere.

La media

Quando la variabile è almeno a livello di scala intervalli è sempre possibile calcolare moda e mediana, tuttavia la capacità informativa di entrambe è inferiore a quella della media. Questa misura prende, infatti, in considerazione tutti i valori esatti della distribuzione. Tra le medie quella maggiormente utilizzata è la media aritmetica, tuttavia esistono anche altri tipi di media: la geometrica, l'armonica e la quadratica. In generale date le modalità x_1, x_2, \dots, x_k di una variabile quantitativa la media è quel valore numerico M tale che, applicando l'operazione anziché ad x_1, x_2, \dots, x_n , a n valori uguali a M , il risultato dell'operazione è lo stesso. Per ogni operazione matematica si ha una media diversa. Così se l'operazione è l'addizione \bar{M} è la media aritmetica

$$\bar{M} = \frac{1}{n} \sum_{i=1}^n x_i;$$

se è la moltiplicazione M_g è la media geometrica

$$M_g = \sqrt[n]{\prod_{i=1}^n x_i};$$

se è la somma degli inversi M_a è la media armonica

$$M_a = \frac{n}{\sum_{i=1}^n \frac{1}{x_i}};$$

se è la somma dei quadrati M_q è la media quadratica

$$M_q = \frac{1}{n} \sum_{i=1}^n x_i^q.$$

Se la media è calcolata sulla distribuzione unitaria delle modalità della variabile si parla di media semplice, se è calcolata sulla distribuzione di frequenza si parla di media ponderata. In fase di elaborazione dei dati ci si chiede quale tra tutte le medie viste è più adatta a descrivere il problema analizzato. In generale non esistono criteri di scelta; oltre che tener conto della natura del fenomeno oggetto di misura e dello scopo della ricerca, si deve tener conto di come è stato ottenuto l'ammontare che si vuole equiripartire con la media. Se all'ammontare del carattere si perviene con l'addizione la media è quella aritmetica, se invece si perviene con il prodotto la media è quella geometrica. A volte nella scelta della media si tiene conto anche delle proprietà che essa possiede. Se, per esempio, data una distribuzione di misurazioni di una stessa grandezza, la media deve fornire una stima della misura della grandezza stessa, allora in questo caso si tiene conto degli errori accidentali di cui sono affette le varie misurazioni e si cerca una media che renda nulla la somma degli scarti positivi e negativi. La media più indicata è quella aritmetica, in quanto una delle sue proprietà è proprio quella che la somma algebrica degli scarti di tutti i valori di x_i dalla loro media aritmetica è nulla,

$$\sum_{i=1}^n (x_i - \bar{x}) = 0.$$

2.2 Operatori di dispersione

La dispersione assume nomi differenti a seconda del livello di scala della variabile. Così per le variabili categoriali si parla di *eterogeneità* e *omogeneità* di una distribuzione; per le variabili ordinali si parla di *variabilità non metrica* e si riserva il ter-

mine di *variabilità metrica* alle sole variabili cardinali.

Mutabilità

Per le variabili a livello nominale la dispersione di una variabile può essere quantificata attraverso l'eterogeneità o il suo complemento, l'omogeneità.

In generale, data una variabile a k modalità, la massima omogeneità si ha nel caso in cui una sola modalità ha frequenza assoluta pari ad n ; la massima eterogeneità si ottiene quando ciascuna modalità ha la stessa frequenza, pari a n/k . Solitamente una distribuzione presenta un grado di omogeneità o eterogeneità intermedio. Esistono diversi indici di omogeneità e di eterogeneità: essi si basano sulle frequenze relative.

Variabilità non metrica

Per sfruttare la maggior capacità informativa delle variabili a livello di scala ordinale, rispetto alle variabili categoriali si deve ricorrere a un operatore che raggiunga il suo massimo quando le osservazioni sono equidistribuite nelle due modalità estreme, e il suo minimo quando le osservazioni sono equamente ripartite all'interno di tutte le modalità.

Variabilità metrica

Quando le variabili sono cardinali possiamo distinguere almeno tre famiglie di

operatori: misure di variabilità globale intrinseche, intervalli di variazione e scarti da un valore centrale.

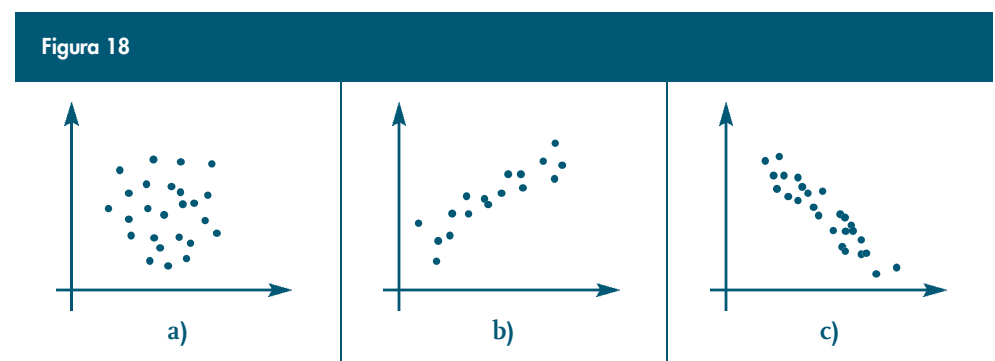
Gli indici di *variabilità globale* danno una misura di quanto differiscono tra loro tutti i termini della distribuzione. Essi prendono in considerazione le differenze tra tutte le coppie di valori della variabile.

Gli *intervalli di variazione* sono indici che quantificano la variabilità misurando la diversità tra due particolari termini della distribuzione. Il più semplice e immediato intervallo di variazione è dato dalla differenza tra il valore massimo e il valore minimo della distribuzione. Esso viene denominato in modi differenti: campo di variazione, gamma o, in inglese, *range* $W = x_{\min} - x_{\max}$.

Le misure di variabilità metrica più interessanti e utilizzate sono gli indici basati sugli *scarti dalla media*. All'interno della famiglia degli scarti dalla media rientrano quegli indici che misurano la diversità tra ciascun termine della distribuzione e il valore centrale della distribuzione. Essi sono i più utilizzati quando le variabili sono cardinali, in quanto più informativi (prendono in considerazione tutti i valori della distribuzione e il valore centrale). Un indice tra i più utilizzati è la *varianza*:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}.$$

La varianza ha lo svantaggio di essere una grandezza quadratica e quindi non





direttamente confrontabile con altri valori della distribuzione come la media. Per passare dalla varianza a una misura espressa nella stessa unità di misura della variabile di partenza è sufficiente estrarne la radice quadrata. L'indice così ottenuto viene detto *deviazione standard* (s), o *scarto quadratico medio*:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}}$$

Nella formula della varianza e della deviazione standard la somma degli scarti al quadrato viene divisa per il numero di osservazioni in modo da rendere le misure indipendenti dall'ampiezza del collettivo sul quale vengono calcolate. Per poter confrontare la variabilità di distribuzioni espresse con diverse unità di misura si può ricorrere al *coefficiente di variazione*:

$$Cv = \frac{s}{\bar{x}}$$

Il coefficiente di variazione, essendo il rapporto tra due grandezze espresse nella stessa unità di misura, è un numero puro e permette confronti diretti tra qualsiasi distribuzione.

3. Operatori bivariati

La rappresentazione dei dati in tabelle di contingenza, e più in generale tutta l'analisi bivariata, ha come scopo principale lo studio della relazione tra variabili. Dal punto di vista semantico la relazione tra due variabili può essere analizzata tenendo conto dell'intensità della relazione, della forma della relazione e della determinazione o variabilità della variabile dipendente riprodotta dalla variabile indipendente.

In termini generali, date due variabili possiamo affermare che tra esse esiste una relazione se i valori su una variabile variano in modo sistematico con i valori sull'altra.

Più precisamente parliamo di *connessione* intendendo la misura della forza (intensità) della relazione tra due variabili o specularmente dell'indipendenza tra due variabili.

Di una relazione tra variabili possiamo calcolare il grado di *concordanza* (o discordanza). Consideriamo le variabili ordinali e cardinali: la concordanza, oltre che sull'intensità, ci informa sulla direzione della variazione di una variabile al variare dell'altra. Se al crescere di V_1 cresce anche V_2 la misura avrà segno positivo. Viceversa, se al crescere di V_1 V_2 decresce, la misura avrà segno negativo. Data una connessione o una concordanza è possibile calcolare la *determinazione*: misura che ci dice, in valori percentuali, quanta variabilità o mutabilità è 'spiegata' o prevista, o riprodotta, da una variabile sull'altra. Una misura di determinazione valuta quanto i valori di una variabile sono riproducibili a partire dai valori assunti dall'altra variabile.

La tabella di contingenza fornisce una prima indicazione dell'eventuale presenza o assenza di relazione. Quando le variabili sono cardinali o ordinali (con un numero sufficientemente ampio di modalità distinte) una simile opportunità è offerta dalla loro rappresentazione grafica (vedi [figura 18](#)).

La figura (a) rappresenta la situazione di assenza di relazione, le figure (b) e (c) rappresentano i casi di relazione molto forte in un caso positiva, nell'altro negativa. I grafici b) e c) raffigurano, inoltre, esempi di relazioni lineari, ossia approssimabili attraverso l'equazione di una retta che metta in corrispondenza i valori di X e di Y nel seguente modo: $y = a * x + b$. In altri casi la relazione sussiste ma è approssimabile da una equazione più complessa rispetto a quella di una retta; si tratta infatti di una relazione non lineare che può assumere diverse forme.

4.1 Operatori di connessione

In generale diciamo che gli operatori di connessione restituiscono uno scalare

sempre positivo; essi assumono valore zero in assenza di connessione e valore uno, o maggiore di uno, nel caso ci sia connessione tra le due variabili.

Gli operatori di connessione si applicano principalmente alle variabili categoriali e in misura minore alle variabili ordinali. Quando le variabili si trovano a un livello di scala più elevato si ricorre agli operatori di concordanza, più informativi, in quanto in grado di quantificare non soltanto la forza della relazione, ma anche la direzione.

Tra gli operatori di connessione quello più diffuso è il X^2 quadrato

$$X^2 = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Esso è costruito sullo scarto al quadrato tra i valori osservati e i valori teorici, questi ultimi calcolati ipotizzando una situazione di indipendenza tra le variabili.

Il X^2 quadrato assume come valore minimo lo zero, mentre il valore massimo dipende dal numero dei casi e dai gradi di libertà associati alla tabella di contingenza. In generale possiamo definire i *gradi di libertà* come il numero di valori indipendenti necessari per definire lo stato di un sistema. Ora, si può dimostrare che il valore massimo del X^2 quadrato corrisponde al valore minimo tra le seguenti espressioni: $n(k-1)$, $n(h-1)$. Come si può notare il valore massimo dipende strettamente dall'ampiezza del collettivo sul quale viene calcolato e dal numero di righe e colonne della tabella. Questo rende impossibile confrontare direttamente il X^2 quadrato ottenuto su tavole di diverso formato e su collettivi diversi.

4.2 Operatori di concordanza

Gli operatori di concordanza si caratterizzano per la presenza di un punto neutro, lo zero, che segnala l'assenza di una relazione *monotona* tra le due variabili e

due poli, uno negativo e uno positivo, che segnalano rispettivamente una discordanza e una concordanza.

Tra le misure di concordanza quelle più diffuse, rispettivamente per le variabili ordinali e cardinali, sono il coefficiente r_s di Spearman e il coefficiente di correlazione r di Bravais-Pearson.

Di gran lunga più utilizzato è il coefficiente di correlazione r di Bravais-Pearson ottenuto dividendo la covarianza per il suo valore massimo:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

L' r ha così la comoda proprietà di assumere come valori massimo e minimo rispettivamente $+1$ e -1 , indipendentemente dall'unità di misura in cui sono espresse le variabili, e valore 0 in assenza di concordanza. Quando r assume valore 1 le due variabili, X_i e Y_i , sono in una precisa relazione funzionale, una relazione lineare.

Il coefficiente b di regressione

Un altro indice molto usato è il coefficiente di regressione. Esso è strettamente connesso alle misure fin qui presentate, in particolare esso si può ottenere come rapporto tra la covarianza (s_{xy}) e la varianza della variabile dipendente (s_y o s_x). Se indichiamo con X_i la variabile indipendente e con Y_i la variabile dipendente otteniamo

$$b_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Esso ci informa su quanto aumenta la variabile dipendente al variare di un'unità della variabile indipendente.

4.3 Operatori di determinazione

Gli operatori di determinazione possono assumere soltanto valori positivi e hanno come valore massimo 1 essendo delle quote di variabilità riprodotta.

Date due variabili X_i e Y_i attraverso gli operatori di determinazione possiamo valutare la forza della loro relazione in termini di variabilità o mutabilità della Y_i – variabile assunta come dipendente – riprodotta dalla X_i , assunta come variabile indipendente. Tanto maggiore è il grado di determinazione tanto più è forte la relazione. In altre parole la determinazione valuta la prevedibilità dei valori di Y_i a partire dalla conoscenza dei valori assunti dalla variabile X_i .

Quando le variabili sono cardinali la determinazione viene valutata attraverso il coefficiente R^2 .

Il coefficiente R^2 è dato dalla seguente formula:

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

Lo scarto di y_i dalla media della distribuzione può essere scomposto nella somma di due distanze:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

Se eleviamo al quadrato i due termini dell'equazione e sommiamo per tutti i casi, attraverso semplici passaggi arriviamo a:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

Il termine a sinistra viene detto *devianza totale* e non è altro che la devianza della variabile Y ; l'ultimo termine a destra è la devianza calcolata non sui valori puntuali y , ma sulle loro stime, e viene detta

devianza riprodotta. Il primo termine a destra del segno uguale rappresenta la devianza residuale, non riprodotta. Più la relazione tra le variabili è stretta più il valore stimato (\hat{y}_i) sarà vicino a quello effettivo (y_i) e di conseguenza più ridotta sarà la devianza residua.

Un altro indice molto usato quando una delle due variabili è cardinale è l'indice η^2 . In questi casi l'interesse è prevedere i valori della variabile cardinale a partire dai valori di una variabile categoriale o ordinale.

Bibliografia

Blalock H. M. Jr.

1984 *Statistica per la ricerca sociale*, Il Mulino, Bologna.

Landenna G.

1984 *Fondamenti di statistica descrittiva*, Il Mulino, Bologna.

Leti G.

1983 *Statistica descrittiva*, Il Mulino, Bologna.

Marradi A.

1995 *L'analisi monovariata*, Franco Angeli, Milano.

Frosini B.V., Montinaro M., Nicoloni G.

1996 *Complementi ed esercizi di statistica*, Tirrenia Stampatori, Torino.

Testa S., Albano R.

1999 *Statistica descrittiva per la ricerca psicologica e sociale*, Trauben, Torino.

Albano R., Massa A., Testa S.

1999 *Statistica inferenziale per la ricerca psicologica e sociale*, Trauben, Torino.

Intranet del MIUR, area Supporto alle decisioni.