Mettiamo un po' d'ordine... e parliamo di dipendenza

Riassunto

Dopo aver fornito le definizioni di distribuzioni statistiche semplici e doppie, si passa allo studio della dipendenza tra caratteri associati, introducendo il chi-quadro per misurare la strettezza del legame associativo.
Parole chiave. *Distribuzioni statistiche, strettezza del legame associativo, dipendenza, chi-quadro.*

Summary

Concept of dependence and the chi-square index

Statistical distributions were defined, and after clarifying the concept of dependence, chi-square index was introduced to measure the strength of the association.
Key words. *Statistical distributions, strength of association, dependence, chi-square.*

Il punto centrale della ricerca scientifica è lo studio del legame associativo tra fenomeni. Infatti, è solo attraverso la sua conoscenza che si possono formulare previsioni.

Per restare nello spirito della rubrica, lo studio della dipendenza tra fenomeni (collettivi) sarà trattato esponendo i concetti con il minor apparato formale possibile, pur conservando un dignitoso rigore.

Limitandoci allo studio della relazione tra due fenomeni, l'argomento non è particolarmente complesso, ma è esteso. Si è deciso, pertanto, di trattarlo in due parti, cogliendo l'occasione in questa prima parte per dare definizioni rigorose di distribuzioni statistiche semplici e doppie e per introdurre un indice che misura la forza del legame associativo tra due caratteri. La seconda parte, più specificamente incentrata sull'analisi delle distribuzioni doppie, sarà esposta nel prossimo numero.

La Statistica si occupa di fenomeni collettivi, cioè di fenomeni la cui conoscenza richiede una massa di osservazioni di fenomeni elementari (sono fenomeni elementari quelli rilevati su ciascuna delle unità che costituiscono il collettivo). Quindi "collettivo" è un insieme di unità, come una popolazione, o anche un campione, o un gruppo.

Esempio 1. *Volendo studiare la sopravvivenza di pazienti affetti da uno stesso tipo di neoplasia ad uno stesso stadio, in cura negli ultimi 3 anni presso un certo centro oncologico, occorre rilevare la durata di sopravvivenza in tutti i pazienti che costituiscono tale collettivo.*

Non tutti i pazienti considerati hanno la stessa sopravvivenza: ciò che caratterizza il fenomeno collettivo è la **variabilità** ossia il fatto che il fenomeno oggetto di studio si manifesta in modo differente nelle diverse unità che costituiscono il collettivo.

In Statistica si chiama "carattere" il fenomeno oggetto di studio, purché immediatamente rilevabile nelle unità del collettivo. Ad esempio, sono caratteri il peso, l'età, il tipo di tumore (immediatamente rilevabili), ma non sono caratteri il potere, l'intelligenza, la qualità di vita che per essere rilevati sui singoli soggetti hanno bisogno di strumenti psicometrici che ne misurino il livello. Tali strumenti sono composti da item: le risposte a ciascun item, o ad una loro aggregazione, sono caratteri.

Il risultato della classificazione delle unità del collettivo secondo un certo carattere prende il nome di "distribuzione statistica semplice".

Esempio 2. *Il collettivo di cui all'esempio 1 sia composto da 8 pazienti, la cui sopravvivenza (in mesi) è 8, 12, 8, 8, 12, 22, 12, 12. Tali dati costituiscono la distribuzione semplice dei pazienti suddetti secondo la sopravvivenza. Se anziché di 8 pazienti, il collettivo fosse composto da 80 o, peggio, da 800 pazienti, con tale rappresentazione sarebbe difficile farsi un'idea della struttura della sopravvivenza. Si ricorre, allora, ad una distribuzione più compatta, introducendo la frequenza (assoluta), considerando ciascuna modalità del carattere (cioè i modi diversi con cui il carattere si manifesta nelle unità del collettivo) ed associandole il numero delle unità che la presentano:*

Sopravvivenza	n. pazienti
8	3
12	4
22	1
totale	8

Vi sono dunque due modi di rappresentare la distribuzione statistica:

1. *distribuzione per unità* (o *disaggregata*), ossia l'elenco di ciò che si è rilevato su ciascuna unità del collettivo (stringa dei dati nell'esempio precedente);
2. *distribuzione per modalità* (o *aggregata*), che è quella esposta in tabella.

Tali due rappresentazioni sono equivalenti, nel senso che data una di esse si può sempre passare all'altra, e,

quindi entrambe contengono la stessa quantità di informazione. Nell'es. 2, si è mostrato come passare da una distribuzione per unità ad una per modalità; il viceversa è banale: data la tabella dell'es. 2, basta scrivere 8 tre volte, 12 quattro volte, 22 una volta per tornare alla distribuzione per unità.

Osservazione. *Per chi ha (anche un po' di) familiarità con Excel, le colonne del foglio rappresentano altrettante distribuzioni semplici per unità, mentre ciascuna riga mostra ciò che si è rilevato su ogni unità rispetto a tutti i caratteri considerati. L'intero foglio di Excel è un prospetto rettangolare che ha tante righe quante sono le unità e tante colonne quanti sono i caratteri rilevati; esso prende il nome di "matrice dei dati".*

Distribuzioni doppie

Esempio 3. *Proseguendo l'es. 2, immaginiamo di rilevare anche il sesso degli 8 pazienti e, nell'ordine con cui è stata rilevata la sopravvivenza, gli attributi siano: F, F, F, M, M, F, F, M. Se gli attributi si riferiscono, nell'ordine, alle stesse unità in cui è stata rilevata la sopravvivenza, si ha una fondamentale informazione in più: l'associazione tra i due caratteri. Infatti dalla stringa di attributi (che è una distribuzione semplice per unità) si può passare alla corrispondente distribuzione per modalità:*

Sesso	n. pazienti
M	3
F	5
totale	8

ma, sapendo che la stringa degli attributi si riferisce nell'ordine alla sopravvivenza, siamo di fronte a due caratteri rilevati sulle stesse unità cioè a caratteri associati.

Abbiamo quindi costruito qualcosa di più di due distribuzioni semplici, cioè una distribuzione doppia che può rappresentarsi in forma di distribuzione doppia per unità (qui riportata per riga per motivi di spazio, mentre con Excel sarebbero due colonne):

Sopravvivenza	8	12	8	8	12	22	12	12
Sesso	F	F	F	M	M	F	F	M

o in forma di distribuzione per modalità; basta contare il numero di pazienti per ciascuna coppia di valori dei caratteri associati e riportare tale conteggio nella casella appropriata della tabella che segue:

Sopravv./Sesso →	M	F	totale
8	1	2	3
12	2	2	4
22	–	1	1
totale	3	5	8

che prende il nome di tabella doppia (o tabella a doppia entrata).

Quindi una distribuzione doppia è il risultato della classificazione delle unità del collettivo secondo due caratteri. La sua rappresentazione in forma di tabella doppia presenta ai margini le due distribuzioni semplici, che negli esempi 2 e 3 sono state viste separatamente (si invita il Lettore a rintracciarle nella tabella doppia); per questo motivo le distribuzioni (modalità e frequenze corrispondenti) che sono ai margini della tabella doppia sono dette **distribuzioni marginali**.

Oltre le distribuzioni marginali, nella tabella doppia appaiono anche altre distribuzioni semplici, dette **distribuzioni parziali (o condizionate o subordinate)**, che danno informazioni sull'associazione tra i caratteri.

Nella tabella doppia sopra riportata vi sono due distribuzioni parziali secondo la sopravvivenza, quella dei maschi e quella delle femmine. Inoltre, vi sono 3 distribuzioni parziali secondo il sesso, una per ciascuna modalità del carattere "sopravvivenza". Quindi, in totale, nella tabella doppia sono rappresentate 2 distribuzioni marginali e 5 distribuzioni parziali (2 secondo la sopravvivenza e 3 secondo il sesso), per un totale di 7 distribuzioni semplici (si invita il Lettore a rintracciarle nella tabella doppia).

Unico scopo di una distribuzione doppia è lo studio della relazione tra i caratteri associati, cioè rilevati sulle stesse unità del collettivo, (Sesso e Sopravvivenza, nell'esempio). In altre parole, l'unico obiettivo di una distribuzione doppia è valutare **se e come varia un carattere al variare dell'altro**; nell'esempio, se e come varia la sopravvivenza al variare del sesso.

Tabella di indipendenza

1. Premessa: distribuzioni simili

Due distribuzioni semplici si dicono uguali se hanno le stesse modalità e le frequenze assolute corrispondentemente uguali. Tale definizione è però troppo restrittiva per le applicazioni perché richiede alle due distribuzioni di avere lo stesso numero totale di unità (nell'es. 3, tale definizione sarebbe applicabile solo quando il numero dei maschi fosse uguale a quello delle femmine). Per rimuovere tale vincolo, si introduce il concetto di distribuzioni simili: due distribuzioni semplici si dicono simili se hanno le stesse modalità e le frequenze relative ordinatamente uguali (è questo il criterio di uguaglianza tra distribuzioni che interessa in Statistica). Come visto in precedenza in questa rubrica, le frequenze relative si ottengono dividendo le frequenze assolute (conteggi), per il loro totale; moltiplicando le frequenze relative per 100 si ottengono le ben note percentuali.

Osservazione. *Due distribuzioni simili hanno uguali tutti gli indici statistici di nostro interesse; ad esempio, hanno la stessa media, la stessa mediana, lo stesso scarto quadratico medio.*

2. Concetto di indipendenza

Nel linguaggio corrente, due fenomeni si dicono indipendenti se le variazioni dell'uno non influiscono sulle variazioni dell'altro.

Dati due caratteri associati (cioè rilevati sulle stesse unità del collettivo), X e Y, in Statistica si dice che X è indipendente da Y se, al variare di Y, X resta costante.

Consideriamo una tabella doppia, per fissare le idee, quella riportata nell'es. 3.

La sopravvivenza è indipendente dal sesso se, al variare del sesso, la sopravvivenza resta costante. "Al variare del sesso" vuol dire che il carattere "sesso" assume modalità M e modalità F. Come si è visto, per ogni modalità assunta dal carattere sesso, esiste una distribuzione parziale secondo il carattere "sopravvivenza". Allora la sopravvivenza è indipendente dal sesso se, al variare del sesso, le distribuzioni parziali secondo la sopravvivenza sono simili tra loro.

Esempio 4. *Riprendiamo la tabella doppia costruita nell'es. 3, riportando però le frequenze relative (per comodità di lettura, espresse in forma percentuale) nelle distribuzioni parziali e in quella marginale secondo la sopravvivenza:*

Sopravv./Sesso →	M	F	totale
8	33,3	40,0	37,5
12	66,7	40,0	50,0
22	–	20,0	12,5
totale	100	100	100

Come si può osservare, le distribuzioni parziali secondo la sopravvivenza non sono simili tra loro. Quindi, al variare del sesso, la sopravvivenza varia e, pertanto, nel collettivo considerato, tra sesso e sopravvivenza c'è una relazione: la sopravvivenza dipende dal sesso.

Osservazione. *In una tabella doppia, se il carattere X è indipendente dal carattere Y (ossia, X non varia al variare di Y), si può dimostrare che anche Y è indipendente da X (cioè Y non varia al variare di X) e, brevemente, si dice che X e Y sono indipendenti.*

Esempio 5. *Uno studio randomizzato a 3 bracci (T1, T2, T3) è stato condotto per valutare se i tre trattamenti hanno un differente impatto sulla qualità di vita (QoL), riassunta con uno score a 5 punti (crescenti in relazione alla QoL migliore). I risultati (dati non reali) sono rappresentati nella seguente tabella doppia:*

Tabella rilevata

QoL, score/tratt. →	T1	T2	T3	totale
0	9	16	15	40
1	30	10	20	60
2	20	50	30	100
3	50	40	30	120
4	11	44	25	80
totale	120	160	120	400

Accanto ad ogni tabella doppia rilevata, è possibile costruire una tabella di indipendenza in cui, cioè, i caratteri associati sono indipendenti.

La tabella di indipendenza ha le stesse distribuzioni marginali della tabella rilevata; cambiano solo le frequenze di associazione (quelle riportate nel corpo della tabella).

Si può dimostrare che le frequenze del corpo della tabella di indipendenza si ottengono moltiplicando le corrispondenti frequenze marginali e dividendo per il totale.

Tabella di indipendenza

QoL, score/tratt. →	T1	T2	T3	totale
0	12	16	12	40
1	18	24	18	60
2	30	40	30	100
3	36	48	36	120
4	24	32	24	80
totale	120	160	120	400

Consideriamo T1. Le frequenze del corpo della tabella di indipendenza sono state ottenute nel modo seguente (a partire dall'alto):

40 x 120/400 = 12;
60 x 120/400 = 18;
100 x 120/400 = 30;
120 x 120/400 = 36;
80 x 120/400 = 24, e così via per le altre colonne.

Calcolando sulla tabella di indipendenza le frequenze relative per ogni trattamento, si può verificare che sono corrispondentemente uguali e che, quindi, le tre distribuzioni parziali secondo lo score di QoL (una per ogni trattamento) sono simili tra loro a due a due.

Nell'esempio, la tabella rilevata non coincide con quella di indipendenza e, quindi, tra i caratteri associati c'è dipendenza (o connessione): al variare di un carattere (il trattamento), varia anche l'altro (lo score di QoL); in altre parole, lo score di QoL dipende dal trattamento (sarà compito dell'analisi della tabella doppia stabilire come varia lo score al variare del trattamento).

Osservazione. *Indipendenza è assenza di ogni legame associativo. Se si trova che la tabella rilevata coincide con quella di indipendenza, ogni analisi della dipendenza diventa inutile.*

Una misura della dipendenza

Quanto più la tabella rilevata è "vicina" a quella di indipendenza, tanto più debole è la dipendenza; viceversa, quanto maggiore è la "diversità" tra la tabella rilevata e quella di indipendenza, tanto più forte è il legame associativo tra i caratteri considerati.

Ricordando che la tabella rilevata ha le stesse distribuzioni marginali della tabella di indipendenza, una misura della strettezza del legame associativo non può che basarsi sulla diversità delle corrispondenti frequenze di associazione: quanto più queste sono diverse tra loro, tanto più la tabella rilevata sarà distante da quella di indipendenza e, quindi, tanto più forte sarà la connessione.

Esempio 6. *Si perviene alla costruzione del più usato indice di connessione confrontando casella per casella le due tabelle. Più precisamente, considerando le tabelle riportate nell'es. 5, a patire dalla casella in alto a sinistra, 9 – 12 misura la diversità tra le due frequenze.*

Per eliminare il segno, ne consideriamo il quadrato: $(9 - 12)^2$. Così facendo, però, abbiamo cambiato unità di misura; per restituire al confronto tra le due frequenze l'unità di misura originale (cioè per esprimerlo in forma di frequenza assoluta), ***si divide per la frequenza della tabella di indipendenza****: $(9 - 12)^2/12 = 0,75$.*

Analogamente per la seconda casella della stessa colonna, si ha: $(30 - 18)^2/18 = 8$; per la terza casella, $(20 - 30)^2/30 = 3,33$.

Si procede così per tutte le 15 caselle del corpo della tabella e si sommano i risultati (si invita il Lettore, per esercizio, a completare i calcoli; comunque, tale indice è calcolato da ogni package statistico).

L'indice così costruito prende il nome di ***chi-quadrato (o chi-quadro)*** *ed è un indice di connessione; esso, quindi,* ***misura la strettezza del legame associativo*** *tra i fenomeni considerati. Infatti,*

a. *chi-quadro vale 0 quando e solo quando la tabella rilevata coincide con quella di indipendenza (è intuitivo, ma se ne può dare dimostrazione);*
b. *chi-quadro cresce al crescere della connessione: è tanto maggiore quanto più sono grandi i suoi addendi, cioè quanto più le frequenze di associazione della tabella rilevata sono diverse da quelle della tabella di indipendenza.*

Se chi-quadro (che per costruzione non può assumere valori negativi) è maggiore di zero, tra i caratteri associati c'è dipendenza e, quindi, al variare di un carattere varia anche l'altro. Stabilire come varia un carattere al variare dell'altro è compito dell'analisi della distribuzione doppia, argomento che sarà trattato nel prossimo numero di CASCO.

Enzo Ballatori