Statistica 1

Dati bivariati. I - Associazione

Alessandra Salvan e Laura Ventura

Dipartimento di Scienze Statistiche
Università di Padova
http://www.stat.unipd.it/

salvan@stat.unipd.it
ventura@stat.unipd.it

Dati e variabili

- Spesso l'informazione su aspetti di interesse della realtà è espressa in forma numerica (DATI).
- Esempi:
  - Per un certo insieme di neonati, si registrano il peso alla nascita e la durata della gravidanza.
  - Per un insieme di soggetti coinvolti in incidenti ciclistici, si rilevano il tipo di trauma e l'uso del casco.
  - In giorni consecutivi, si misurano i rendimenti di BTP e BUND
  - ...
  - Tutti questi esempi riguardano fenomeni NON DETERMINISTICI: non esattamente prevedibili.
  - Viceversa, sono fenomeni DETERMINISTICI quelli esattamente prevedibili. Ad esempio: la direzione del moto di una pallina lasciata cadere da una data altezza; il peso di un litro di acqua a una fissata temperatura,....

Dati e variabili

- C'è sotto qualcosa?
  Per un fenomeno non deterministico le quantità osservate differiscono da caso a caso in modo non prevedibile. C'è VARIABILITÀ. Interessa studiare metodi che permettano di individuare delle eventuali regolarità sottostanti.
- Le entità (individui, ore del giorno, ...) che vengono osservate nello studio sono dette UNITÀ STATISTICHE (casi).
- L'insieme di tutte le unità statistiche di interesse per lo studio è detto POPOLAZIONE di riferimento.
- Invece, un sottoinsieme di unità statistiche selezionate (spesso casualmente) da una popolazione è detto CAMPIONE. La dimensione del campione può variare da poche unità a molte migliaia di osservazioni.
- Una quantità di interesse nella popolazione è detta parametro, mentre la quantità calcolata sul campione è detta statistica.

DEF: Una VARIABILE (o CARATTERE) è una caratteristica di interesse rilevata sulle unità statistiche (ad esempio, età, peso, trattamento, ...).

Il termine 'variabile' evidenzia che la caratteristica di interesse può assumere una pluralità di valori. L'insieme dei valori possibili si può pensare noto, ma prima di fare l'osservazione su una unità statistica, non sappiamo quale valore si osserverà.

DEF: I valori distinti assunti da una variabile sono detti MODALITÀ della variabile. Le modalità si presumono note preliminarmente.

Una variabile può essere:

- QUALITATIVA o CATEGORIALE quando le sue modalità sono espresse in forma verbale (*sesso*, *livello di istruzione*, *trattamento*, ...).
- QUANTITATIVA (o NUMERICA) quando le modalità sono espresse da numeri (età, peso,...).

A sua volta una variabile qualitativa può essere:
– SCONNESSA o NOMINALE se non esiste nessun ordinamento tra le modalità.
Esempi:
la variabile *sesso* con modalità M e F;

la variabile *modo di somministrazione* con modalità ORALE, ENDOVENA, ...

– ORDINALE se è possibile individuare un ordinamento naturale delle modalità.
Esempi:
la variabile *livello di istruzione* con modalità ELEMENTARE, MEDIA INFERIORE, MEDIA SUPERIORE, ...;

la variabile *giudizio* con modalità INSUFFICIENTE, SUFFICIENTE, DISCRETO, OTTIMO.

• Se le modalità sono solo due si parla di variabili DICOTOMICHE o BINARIE (*sesso*, *presenza*, ...). A volte le due modalità sono espresse con valori numerici (0,1, oppure 1,2,...), ma il valore del numero non vuol dire assolutamente nulla!!

Una variabile quantitativa può essere:

– DISCRETA quando l'insieme delle modalità è finito o numerabile (stessa cardinalità dell'insieme dei naturali). Esempi:
- la variabile *numero di 'teste' in 10 lanci di una moneta*, con modalità 0,1, . . . , 10;
- le variabili *numero di sedute*, *numero di figli*, . . . con modalità 0, 1, 2, . . . ;

– CONTINUA quando l'insieme delle modalità è un intervallo, ossia un sottoinsieme, eventualmente illimitato, dei numeri reali. Esempi:
- la variabile *peso* (in kg) che ha come modalità possibili tutti i valori positivi,
- la variabile *dose* di un dato farmaco (in mg) con modalità da zero a 1000mg.
- eventuale suddivisione in classi.

Dati bivariati

- In molte situazioni interessa studiare se esiste una relazione tra due variabili misurate sulle stesse unità. Esempi:
  - *"Il fumo è in relazione con lo sviluppo di malattie respiratorie?"*
  - *"il voto di maturità è in relazione con la performance universitaria?"*
- Oppure si desidera prevedere il valore di una variabile conoscendo il valore di un'altra. Esempi:
  - *"conoscendo l'altezza del padre, è possibile prevedere l'altezza di un figlio?"*
  - *"conoscendo la durata della gravidanza, si può stimare il peso alla nascita?"*
- La statistica permette di rispondere a questo tipo di domande, con strumenti adatti alla natura delle variabili in esame.
- In questa unità si tratterà il caso di dati bivariati con due variabili qualitative, nella successiva si tratteranno le variabili quantitative.

Tabelle di contingenza

Tabelle di contingenza

- Il modo più comune per rappresentare sinteticamente i dati categoriali sono le Tabelle di contingenza (distribuzioni di frequenza doppie).
- Esse costituiscono l'organizzazione in formato tabulare delle frequenze per variabili qualitative bivariate.
- Le tabelle di contingenza possono essere anche uno strumento idoneo per indagare le relazioni esistenti tra le modalità di due caratteri quantitativi suddivisi in classi.

Rappresentazione generale di una tabella di contingenza

Distribuzione di frequenza doppia per X e Y:

			X			
Y	x_1	$\dots$	x_j	$\dots$	x_J	totale
y_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1J}	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iJ}	$n_{i.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_I	n_{I1}	$\dots$	n_{Ij}	$\dots$	n_{IJ}	$n_{I.}$
totale	$n_{.1}$	$\dots$	$n_{.j}$	$\dots$	$n_{.J}$	n

- n_{ij} $(i = 1, \dots, I, j = 1, \dots, J)$ distribuzione di frequenza congiunta
- $n_{i.}$ $(i = 1, \dots, I)$ distribuzione di frequenza marginale della Y
- $n_{.j}$ $(j = 1, \dots, J)$ distribuzione di frequenza marginale della X
- $n_{ij}/n_{i.}$ $(j = 1, \dots, J)$ distribuzione di frequenza condizionata della X data $Y = y_i$
- $n_{ij}/n_{.j}$ $(i = 1, \dots, I)$ distribuzione di frequenza condizionata della Y data $X = x_j$

ESEMPIO: Efficacia del casco protettivo

Nella tabella 2×2 che segue sono riportati i dati che illustrano i risultati di uno studio sull'efficacia dei caschi protettivi per bicicletta nella prevenzione dei traumi cranici (su $n = 793$ soggetti coinvolti in incidenti).

Trauma cranico	Casco SI	Casco NO	totale
SI	17	218	235
NO	130	428	558
totale	147	646	793

	X		
Y	x_1	x_2	totale
y_1	n_{11}	n_{12}	$n_{1.}$
y_2	n_{21}	n_{22}	$n_{2.}$
totale	$n_{.1}$	$n_{.2}$	n

PROBLEMA: Per esaminare l'efficacia del casco protettivo si vuole valutare se esiste un'associazione (relazione) tra traumi cranici (Y) ed uso del casco (X) tra i soggetti coinvolti in un incidente.

Date le due variabili categoriali, si vuole valutare se X e Y sono dipendenti.

L'indipendenza

- Nella tabella si possono considerare le distribuzioni condizionate di Y (Trauma Cranico) dato $X = x$ (Uso del Casco), nonché la distribuzione marginale di Y, considerate come distribuzioni di frequenza relativa, in modo da ovviare alle diverse numerosità.
- Una situazione estrema si ha quando le distribuzioni condizionate sono tutte uguali: in tale caso è inutile tenere sotto controllo X per evidenziare una fonte sistematica di variabilità dei valori di Y.
- Nell'esempio si ha:

Trauma cranico	Casco SI	Casco NO	totale
SI	$17/147 = 0.12$	$218/646 = 0.34$	$235/793 = 0.29$
NO	$130/147 = 0.88$	$428/646 = 0.66$	$558/793 = 0.71$
totale	1	1	1

da cui si nota che le distribuzioni non sono somiglianti.

Indipendenza

- Si parla di indipendenza statistica quando la conoscenza della modalità di una delle due variabili in esame non migliora la "previsione" della modalità dell'altra.
- Se le distribuzioni condizionate sono tutte somiglianti, allora Y è indipendente da X.
- Condizione necessaria e sufficiente affinché Y sia indipendente da X è che valga, per ogni $i = 1, \ldots, I$ e $j = 1, \ldots, J$, il seguente risultato.

Se X e Y sono indipendenti, la generica frequenza assoluta corrispondente alla i-esima modalità di X e alla j-esima modalità di Y deve essere uguale a

$$a_{ij} = \frac{n_{i.} n_{.j}}{n}$$

ossia le frequenze a_{ij} attese (teoriche) in ipotesi di indipendenza sono il prodotto tra totale della riga e totale della colonna diviso per n.

Dimostrazione

In base alla definizione, Y è indipendente da X se, per ogni j, le distribuzioni condizionate di Y dato $X = x_j$ sono tutte uguali, ossia si ha

$$\frac{n_{i1}}{n_{.1}} = \frac{n_{i2}}{n_{.2}} = \ldots = \frac{n_{ij}}{n_{.j}} = \ldots = \frac{n_{iJ}}{n_{.J}} = p_i^*$$

Ma allora

$$n_{i.} = \sum_{j=1}^{J} n_{ij} = \sum_{j=1}^{J} p_i^* n_{.j} = p_i^* \sum_{j=1}^{J} n_{.j} = n p_i^*$$

Pertanto deve valere l'identità

$$n_{i.} = n p_i^* = n \frac{n_{ij}}{n_{.j}}$$

da cui si ottiene

$$n_{ij} = \frac{n_{i.} n_{.j}}{n} = a_{ij}$$

nell'ipotesi di indipendenza.

Indice χ^2 di Pearson

La statistica χ^2 di Pearson è basata sul confronto tra le frequenze osservate e quelle attese in ipotesi di indipendenza.

La formula per il calcolo della statistica χ^2 è

$$\chi^2 = \sum \frac{(\text{osservate - attese})^2}{\text{attese}} = \sum \frac{(n_{ij} - a_{ij})^2}{a_{ij}}$$

Il calcolo è fatto confrontando le frequenze attese e quelle osservate per ogni cella della tabella, e poi i risultati sono sommati.

frequenze osservate

Trauma cranico	Casco SI	Casco NO	totale
SI	17	218	235
NO	130	428	558
totale	147	646	793

frequenze attese

Trauma cranico	Casco SI	Casco NO	totale
SI	43.56	191.44	235
NO	103.44	454.56	558
totale	147	646	793

Si trova $\chi^2 = 27.20$.

Interpretazione

Rimane da interpretare il valore calcolato per la statistica χ^2.

Per renderci conto se il valore trovato è "grande" o "piccolo" potrebbe essere utile sapere che

$$0 \leq \chi^2 \leq \max(\chi^2) = n \min\left((I-1),(J-1)\right)$$

Si ha
– $\chi^2 = 0$ nel caso di indipendenza tra X e Y ($n_{ij} = a_{ij}$)
– $\chi^2 = \max(\chi^2)$ nel caso di dipendenza perfetta tra X e Y (ad ogni modalità di X corrisponde sempre una sola modalità di Y)
– si avvicina sempre più a $\max(\chi^2)$ quanto più forte è il legame tra le due variabili studiate ($n_{ij} - a_{ij}$ grandi e quindi χ^2 grande)

Nell'esempio sull'efficacia del casco protettivo si ha:

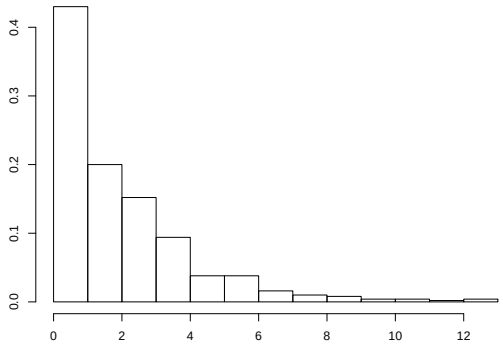
$$\chi^2 = 27.20 \quad n = 793 \quad I = J = 2$$

e $\max(\chi^2) = n \min((r-1),(c-1)) = 793 \min(1,1) = 793$. E quindi?

Interpretazione

- Se X e Y sono indipendenti, ci si aspetta un valore osservato della statistica χ^2 "piccolo".
- Viceversa, se X e Y sono dipendenti, ci si aspetta un valore osservato della statistica χ^2 "grande".
- Per interpretare il valore osservato della statistica χ^2 si può usare un riassunto "probabilistico" dell'evidenza contro l'ipotesi di indipendenza.
- Per capire bene questo serve il Calcolo delle Probabilità. Comunque, intuitivamente, pensiamo che la tabella sia ottenuta effettuando un campionamento casuale da una popolazione in cui c'è effettivamente indipendenza.
- Ipotizzando di ripetere il campionamento casuale molte volte, si calcola la proporzione di tabelle osservate che danno un valore della statistica χ^2 maggiore o uguale a quello osservato nei dati. Un valore piccolo di questa proporzione (p-value) indica che è difficile avere una tabella come quella osservata pescando da una popolazione dove c'è effettivamente indipendenza e dunque indica una evidenza contro l'ipotesi di indipendenza.

Interpretazione



La proporzione (frequenza relativa) di valori maggiori o uguali di 27.2 è praticamente zero.

Esempio

Nella tabella che segue viene mostrata una classificazione di $n = 141$ pesci predati e non predati (X) da parte di uccelli, secondo il livello di infestazione (Y) da parte di particolari vermi (trematodi).

	predati	non predati	totale
non infestati	1	49	50
poco infestati	10	35	45
molto infestati	37	9	46
totale	48	93	141

Essendo le due variabili qualitative, un indice appropriato per lo studio della relazione tra X e Y è la statistica χ^2 di Pearson.

Calcoliamo le frequenze attese nell'ipotesi di indipendenza:

	predati	non predati	totale
non infestati	17	33	50
poco infestati	15.3	29.7	45
molto infestati	15.7	30.3	46
totale	48	93	141

L'indice χ^2 di Pearson è allora:

$$\chi^2 = (1-17)^2/17 + \ldots + (9-30.3)^2/30.3 = 69.5$$

con $\max(\chi^2) = 141 \min((2-1),(3-1)) = 141$ e p-value$\doteq 0$.

Il valore trovato indica che i dati a disposizione evidenziano relazione tra pesci predati e non predati da parte di uccelli, secondo il livello di infestazione da parte di trematodi.

Esempio da prove Invalsi per la classe seconda superiore

Una scuola è costituita da due piani e i 900 alunni che la frequentano sono così distribuiti:

	biennio	triennio	totale
I piano	180	360	540
II piano	140	220	360
totale	320	580	900

Quali fra le seguenti affermazioni è falsa?

(A) Il 40% degli alunni della scuola si trova al II piano. (R. 360/900=0.4)
(B) I 2/3 degli alunni del I piano frequentano il triennio. (R. 360/540=0.667)
(C) Gli alunni del triennio costituiscono il 70% del totale. (R. 580/900=0.64)
(D) Il 20% degli alunni della scuola frequenta il biennio in un'aula del I piano. (R. 180/900=0.2)

Esempio da prove Invalsi per la classe seconda superiore

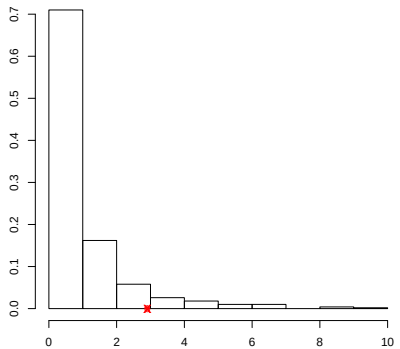
In più ... la tabella nell'ipotesi di indipendenza è:

	biennio	triennio	totale
I piano	192	348	540
II piano	128	232	360
totale	320	580	900

L'indice χ^2 di Pearson è allora:

$$\chi^2 = (180 - 192)^2/192 + \ldots + (220 - 232)^2/232 = 2.91$$

con $\max(\chi^2) = 900 \min((2-1),(2-1)) = 900$ e p-value$= 0.10$. Il valore trovato indica che i dati a disposizione non evidenziano una relazione tra X e Y.



Esercizi

(1) La seguente tabella mostra come 319 studenti universitari si distribuiscono sulla base delle due variabili $X =$ tipo di maturità e $Y =$ numero di esami superati durante il primo anno.

Maturità	Esami superati		
	0-1	2-5	> 5
classica	10	67	31
scientifica	4	52	36
altre	14	65	40

Si calcoli la statistica χ^2 di Pearson

(2) In un'indagine sulle preferenze alimentari si sono svolte 139 interviste e si è chiesto di indicare la preferenza tra tre alimenti liquidi (caffè-thè-succo) e tre alimenti solidi (biscotto-pane-brioche) da consumare a colazione. La tabella è tuttavia disponibile con alcuni dati mancanti (NA).

liquidi	solidi			
	biscotto	pane	brioche	tot
caffè	45	NA	5	58
thè	NA	NA	31	NA
succo	5	27	6	NA

1) Sapendo che 40 intervistati hanno risposto pane tra gli alimenti solidi, si completi la tabella.
2) Si calcoli la statistica χ^2 di Pearson.