Lezione 8
Analisi delle relazioni tra due caratteri

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Prima di cominciare

Appello anticipato 21/02/2023

Elaborati facoltativi

Contenuti: l'elaborato, in un numero contenuto di facciate (da 4 a 8), analizza un argomento (povertà, stili di vita, salute, istruzione, lavoro, relazioni sociali, ecc.) da un punto di vista statistico. È consigliabile il seguente approccio: ipotesi di partenza, dati a supporto dell'ipotesi, analisi dei dati, conclusioni.

Modalità: gli studenti potranno lavorare in piccoli gruppi (2/3 persone).

Supporto: una volta concluse le lezioni, sono a disposizione al giovedì dalle 14.30 alle 15.30 su appuntamento o via mail.

Consegna: entro venerdì 10/2/2022 via mail a roberto.costa@deams.units.it

Dove eravamo rimasti

Nella scorsa lezione abbiamo approfondito il tema delle relazioni tra due variabili.

Abbiamo visto come, a seconda del tipo di variabili utilizzate, abbiamo diversi strumenti per misurare la relazione tra due variabili.

Abbiamo imparato a calcolare la covarianza e l'indice di correlazione di Pearson, partendo dai microdati di una variabile quantitativa.

Ci sono dei dubbi?

Le relazioni tra due caratteri

Prendiamo in considerazione due variabili X e Y.

Come posso procedere nel caso di variabili qualitative, oppure in presenza di una tabella di contingenza (non dei microdati)?

Partiamo da un esempio: supponiamo di aver chiesto agli studenti di tre scuole medie quale percorso pensano di intraprendere dopo la licenza media.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	12	48	60
Scuola B	16	64	80
Scuola C	8	32	40
Totale	36	144	180

Indipendenza in distribuzione

Dobbiamo partire dalle frequenze condizionate.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,33	0,33	0,33
Scuola B	0,44	0,44	0,44
Scuola C	0,22	0,22	0,22
Totale	1,00	1,00	1,00

Distr. condiz. di X (scuola di orig.) rispetto a Y (scuola scelta) X|Y

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,20	0,80	1,00
Scuola B	0,20	0,80	1,00
Scuola C	0,20	0,80	1,00
Totale	0,20	0,80	1,00

Distr. condiz. di Y (scuola scelta) rispetto a X (scuola di orig.) Y|Y

Se vogliamo capire se due caratteri sono indipendenti devo analizzare le frequenze relative condizionate di X rispetto a Y e di Y rispetto a X.

Indipendenza in distribuzione

Nel nostro esempio possiamo vedere come le distribuzioni relative condizionate di X rispetto alle modalità di Y sono tutte uguali tra di loro e rispetto alla frequenza relativa marginale di X.

Anche le distribuzioni relative condizionate di Y rispetto alle modalità di X sono uguali tra di loro e uguali alla frequenza relativa marginale di Y.

Ci troviamo nel caso di **indipendenza statistica in distribuzione**.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,33	0,33	0,33
Scuola B	0,44	0,44	0,44
Scuola C	0,22	0,22	0,22
Totale	1,00	1,00	1,00

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Totale
Scuola A	0,20	0,80	1,00
Scuola B	0,20	0,80	1,00
Scuola C	0,20	0,80	1,00
Totale	0,20	0,80	1,00

Indipendenza statistica in distribuzione

Formalizziamo il ragionamento che abbiamo fatto prima.

X è statisticamente indipendente da Y se le *h* distribuzioni di frequenza relativa di X condizionate alle modalità di Y sono uguali alla frequenza relativa marginale di X:

$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$ per ogni i = 1, …, k e ogni j = 1, …, h

L'indipendenza è simmetrica, quindi

$\frac{n_{ij}}{n_{.j}} = \frac{n_{i.}}{N}$ per ogni i = 1, …, k e ogni j = 1, …, h

Possiamo parlare di indipendenza tra X e Y senza dover specificare una direzione.

Indipendenza statistica in distribuzione

In sintesi, X e Y sono indipendenti se le distribuzioni di frequenza relativa marginale di X|Y sono uguali alla distribuzione di frequenza relativa marginale di X e se le distribuzioni marginali di Y|X sono uguali alla distribuzione di frequenza relativa marginale di Y.

Partendo dalla definizione di indipendenza, dire che X e Y sono statisticamente indipendenti significa affermare che:

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

Ovvero ogni frequenza assoluta congiunta è pari al prodotto del totale della riga i per il totale della colonna j diviso per il numero totale di unità N.

La frequenza assoluta congiunta che otteniamo si chiama **frequenza teorica**, che otterremmo nel caso di indipendenza assoluta tra i caratteri X e Y.

n_{ij}^{*}= frequenza teorica = $\frac{n_{i.}n_{.j}}{N}$

Dipendenza perfetta in distribuzione

Vediamo ora un caso opposto di **dipendenza perfetta** in distribuzione di due variabili X e Y.

Partiamo da un esempio: supponiamo di aver chiesto agli studenti di una scuola da quale rione provengono.

Un carattere Y dipende perfettamente da X quando a ogni modalità di X è associata una sola modalità di Y, ovvero quando per ogni riga i c'è una sola colonna j dove $n_{ij} \neq 0$.

Scuola/rione	Rione 1	Rione 2	Totale
Scuola A	150		150
Scuola B		200	200
Scuola C	100		100
Totale	250	200	450

Dipendenza perfetta in distribuzione

Vediamo ora un caso opposto di dipendenza perfetta in distribuzione di due variabili X e Y.

La relazione è unidirezionale.

In questo caso per ogni riga c'è un solo valore di $X \neq 0$.

Non è vero il contrario: per ogni colonna non c'è un solo valore di $Y \neq 0$.

Scuola/rione	Rione 1	Rione 2	Totale
Scuola A	1	0	1
Scuola B	0	1	1
Scuola C	1	0	1

Interdipendenza perfetta in distribuzione

Si parla di **interdipendenza perfetta** tra due caratteri X e Y quando a ogni modalità di X è associata una sola modalità di Y e, allo stesso tempo, a ogni modalità di Y è associata una sola modalità di X.

Questo significa che per ogni riga i c'è solo una colonna j dove $n_{ij} \neq 0$ e viceversa.

Possiamo trovare l'interdipendenza perfetta solo nel caso di una tabella quadrata, ovvero con lo stesso numero di righe e colonne.

Scuola/rione	Rione 1	Rione 2	Rione 3	Totale
Scuola A	150			150
Scuola B		200		200
Scuola C			100	100
Totale	150	200	100	450

Dipendenza e indipendenza in distribuzione

Pe riassumere, abbiamo visto tre diversi casi che possiamo verificare in una tabella a doppia entrata:

1. Indipendenza in distribuzione

$$\frac{n_{ij}}{n_{i.}} = \frac{n_{.j}}{N}$$ per ogni I = 1, …, k e ogni j = 1, …, h

2. Dipendenza perfetta in distribuzione
3. Interdipendenza perfetta in distribuzione

Misure di dipendenza

Come possiamo facilmente immaginare, nelle scienze sociali è pressoché impossibile trovarsi nelle situazioni di indipendenza o dipendenza perfetta.

Ci troveremo in situazioni di connessione intermedia, che dovremo misurare con appositi indici.

Il nostro punto di partenza sarà lo scarto tra il valore osservato e il valore teorico ($n_{ij}^{*}=\frac{n_{i.}n_{.j}}{N}$).

Chiameremo **contingenze** gli scarti tra valori osservati e teorici:

$$c_{ij}=n_{ij}-n_{ij}^{*}$$

Misure di dipendenza

Potrebbe funzionare la somma degli scarti tra valori teorici e osservati?

$\sum_{i=1}^{k}\sum_{j=1}^{h} c_{ij}=\sum_{i=1}^{k}\sum_{j=1}^{h} n_{ij} - \sum_{i=1}^{k}\sum_{j=1}^{h} n^{*}_{ij}$

Purtroppo no, perché la somma delle frequenze osservate è sempre pari a N, così come la somma delle frequenze teoriche.

$\sum_{i=1}^{k}\sum_{j=1}^{h} c_{ij}$= N – N = 0

Esercitiamoci

Abbiamo la seguente tabella con le frequenze relative della scuola media di origine e della scuola superiore scelta.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A. Manzoni	10	50	5	65
Scuola B. Croce	15	65	10	90
Scuola C. Levi	10	30	5	45
Totale	35	145	20	200

Esercitiamoci

Calcoliamo le contingenze, passo dopo passo.

PASSO 1: Partiamo dal calcolo dei valori teorici $n^*_{ij} = \frac{n_{i.} n_{.j}}{N}$

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	11,38	47,13	6,50	65
Scuola B	15,75	65,25	9,00	90
Scuola C	7,88	32,63	4,50	45
Totale	35	145	20	200

Indice di associazione χ^2 di Pearson

Karl Pearson costruisce un indice, noto come **indice di associazione del χ^2 di Pearson**, facendo ricorso ai quadrati delle contingenze, divise per la frequenza teorica.

La somma di questi valori è l'indice di associazione χ^2.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{c_{ij}^2}{n_{ij}^*}$$

Le proprietà dell'indice di associazione χ^2 di Pearson

L'indice di associazione del χ^2 di Pearson, gode delle seguenti proprietà:

- L'indice χ^2 è simmetrico. Non tiene conto della dipendenza (causa-effetto) e rimane invariato se scambiamo il ruolo di X e Y.
- È sempre non negativo $\chi^2 \geq 0$
- Assume valore 0 nel caso di indipendenza tra X e Y (associazione nulla)
- Assume valori prossimo allo 0 in caso di bassa associazione
- L'indice è tanto più grande quanto più ci si allontana dal caso di indipendenza
- A parità di associazione l'indice aumenta al crescere di N

L'ultima proprietà di χ^2 ne rappresenta anche il suo limite, poiché l'indice cresce, anche se l'associazione non cambia, se aumenta il collettivo osservato.

Esercitiamoci

n_{ij}

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	10	50	5	65
Scuola B	15	65	10	90
Scuola C	10	30	5	45
Totale	35	145	20	200

n_{ij}^{*}

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	11,38	47,13	6,50	65
Scuola B	15,75	65,25	9,00	90
Scuola C	7,88	32,63	4,50	45
Totale	35	145	20	200

PASSO 2: Calcoliamo le contingenze $c_{ij}=n_{ij}-n_{ij}^{*}$ e vediamo che la somma delle contingenze è uguale a 0.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro
Scuola A	-1,38	2,88	-1,50
Scuola B	-0,75	-0,25	1,00
Scuola C	2,13	-2,63	0,50

$$c_{ij}=n_{ij}-n_{ij}^{*}$$

Esercitiamoci

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro
Scuola A	-1,38	2,88	-1,50
Scuola B	-0,75	-0,25	1
Scuola C	2,13	-2,63	0,50

Calcoliamo le contingenze, passo dopo passo.

PASSO 3: Calcoliamo l'indice $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{c_{ij}^2}{n_{ij}^*}$

c_{ij}

c_{ij}^2

$\frac{c_{ij}^2}{n_{ij}^*}$

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	1,89	8,27	2,25	12,41
Scuola B	0,56	0,06	1,00	1,63
Scuola C	4,52	6,89	0,25	11,66
Totale	6,97	15,22	3,50	25,69

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	0,17	0,18	0,35	0,69
Scuola B	0,04	0,00	0,11	0,15
Scuola C	0,57	0,21	0,06	0,84
Totale	0,78	0,39	0,51	1,68

Esercitiamoci

L'indice $\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{c_{ij}^2}{n_{ij}^*}$ ovvero χ^2 è la sommatoria di tutte le contingenze al quadrato divise per le frequenze relative teoriche.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	0,17	0,18	0,35	0,69
Scuola B	0,04	0,00	0,11	0,15
Scuola C	0,57	0,21	0,06	0,84
Totale	0,78	0,39	0,51	1,68

Contingenza quadratica media

Dal momento che l'indice χ^2 aumenta al crescere del collettivo osservato, è opportuno utilizzare un indice che non dipenda da N.

Sempre Karl Pearson ha proposto un altro indice Φ^2 (Phi quadro), o **contingenza quadratica media**.

L'indice è: $\Phi^2 = \frac{\chi^2}{N}$

Che può essere calcolato anche nel seguente modo, che non richiede il calcolo delle frequenze teoriche.

$$\Phi^2 = \sum_{i=1}^{k} \sum_{j=1}^{h} \frac{n_{ij}^2}{n_{i.} n_{.j}} - 1$$

Le proprietà dell'indice Φ^2 (Phi quadro)

Le proprietà dell'indice Φ^2 sono le seguenti:

1. È un indice di dipendenza simmetrico
2. È sempre non negativo: $\Phi^2 \geq 0$
3. Assume valore 0 nel caso di indipendenza tra X e Y (associazione nulla)
4. Il valore massimo che può assumere è pari al valore più piccolo tra il numero di righe della tabella -1 (k – 1) e il numero di colonne della tabella – 1 (h – 1).

$$\max \Phi^2 = \min [(k - 1); (h - 1)]$$

5. Di conseguenza assume valore pari a 1 nel caso in cui il numero di righe o il numero di colonne sia pari a 2, altrimenti assumerà valore maggiore di 1.

Esercitiamoci

Calcoliamo l'indice $\Phi^2 = \frac{\chi^2}{N}$.

χ^2 = 1,68 N = 200

$\Phi^2 = \frac{1,68}{200}$ = 0,008

Il valore è prossimo allo 0, quindi siamo in presenza di bassa associazione.

Scuola di orig./scuola scelta	Liceo	Istituto tecnico	Altro	Totale
Scuola A	0,17	0,18	0,35	0,69
Scuola B	0,04	0,00	0,11	0,15
Scuola C	0,57	0,21	0,06	0,84
Totale	0,78	0,39	0,51	1,68

L’indice V di Cramer

Abbiamo visto che la contingenza quadratica media ha come minimo 0 (associazione nulla) e come massimo il valore più piccolo tra il numero di righe – 1 e il numero di colonne – 1.

$$\max \Phi^2 = \min [(k - 1); (h - 1)]$$

Se vogliamo ottenere un indice che vari tra 0 e 1, dovremo rapportare il valore di Φ^2 con il suo massimo.

L’indice normalizzato (che varia tra 0 e 1) più utilizzato è **l’indice V di Cramer** che si ottiene dalla radice quadrata del rapporto tra Φ^2 e il suo valore massimo.

$$V = \sqrt{\frac{\Phi^2}{\min [(k - 1); (h - 1)]}}$$

Esercitiamoci

Calcoliamo l'indice $V = \sqrt{\frac{\Phi^2}{\min [(k-1); (h-1)]}}$.

$\Phi^2 = \frac{1,68}{200}$ = 0,008

min [(k - 1); (h - 1)]

k - 1 = 3 – 1 = 2 (numero di colonne)

h - 1 = 3 – 1 = 2 (numero di righe)

min [(k - 1); (h - 1)] = 2

V = $\sqrt{\frac{0,008}{2}}$ = 0,065

Il valore è prossimo allo 0, quindi siamo in presenza di bassa associazione.

Come sarà l'esame

Domande legate alla conoscenza teorica:

1 - I **microdati** sono:

A. l'informazione trattata ad un livello minimo, dati grezzi
B. Una sintesi di dati aggregati
C. Informazioni aggiuntive che aiutano a contestualizzare i dati prodotti

2 - Una **popolazione o collettivo di stato** è:

A. Una popolazione definibile precisando un unico istante di tempo
B. Una popolazione definibile in un intervallo di tempo

Come sarà l'esame

Domande legate all'applicazione delle conoscenze teoriche:

3 - Il numero di **abitanti di un comune al 1.1.2022** è un esempio di:

A. Popolazione o collettivo di stato
B. Popolazione o collettivo di movimento

4 - Il numero di **laureati in Scienze dell'educazione nell'anno accademico 2021/22** è un esempio di:

A. Popolazione o collettivo di stato
B. Popolazione o collettivo di movimento

Come sarà l'esame

Domande legate alla conoscenza teorica:

5 - Si definisce **popolazione empirica**:

A. Se tutte le unità che la compongono possono entrare a far parte di un campione
B. Se alcune delle sue unità non possono essere effettivamente osservate

Domande legate all'applicazione delle conoscenze teoriche

6 - Gli **abbonati ad una rivista online** ad una certa data un esempio di:

A. Popolazione empirica
B. Popolazione teorica

Come sarà l'esame

Domande legate alla conoscenza teorica:

7 - Si definisce **rapporto di composizione**:

A. Il rapporto tra l'ammontare di una modalità e l'ammontare complessivo
B. Il rapporto tra l'ammontare di una modalità e quello di un'altra modalità della stessa variabile
C. Il rapporto tra l'ammontare di un fenomeno e quella di un altro che può essere considerato il suo presupposto logico

Domande legate all'applicazione delle conoscenze teoriche

8 - Il **tasso di attività** (rapporto percentuale tra le forze di lavoro e la corrispondente popolazione di riferimento) è un esempio di:

A. Rapporto di derivazione
B. Rapporto di coesistenza
C. Rapporto di composizione

Come sarà l’esame - soluzioni

1 - I **microdati** sono:

l’informazione trattata ad un livello minimo, dati grezzi

2 - Una **popolazione o collettivo di stato** è:

Una popolazione definibile precisando un unico istante di tempo

3 - Il numero di **abitanti di un comune al 1.1.2022** è un esempio di:

Popolazione o collettivo di stato

4 - Il numero di **laureati in Scienze dell’educazione nell’anno accademico 2021/22** è un esempio di:

Popolazione o collettivo di movimento

Come sarà l’esame - soluzioni

5 - Si definisce **popolazione empirica**:

Se tutte le unità che la compongono possono entrare a far parte di un campione

6 - Gli **abbonati ad una rivista online** ad una certa data un esempio di:

Popolazione empirica

7 - Si definisce **rapporto di composizione**:

Il rapporto tra l’ammontare di una modalità e l’ammontare complessivo

8 - Il **tasso di attività** (rapporto percentuale tra le forze di lavoro e la corrispondente popolazione di riferimento) è un esempio di:

Rapporto di composizione

Esercitiamoci

Studente	Inglese	Sociologia	Storia soc.	Geografia	Pedagogia
Voti agli esami	27	28	30	25	30

Calcoliamo i seguenti valori centrali:

media ___

mediana ___

media aritmetica ___

Esercitiamoci

Studente	Inglese	Sociologia	Storia soc.	Geografia	Pedagogia
Voti agli esami	27	28	30	25	30

Calcoliamo i seguenti valori di disuguaglianza:

range ___

scarto quadratico medio ___

varianza ___