POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

Corso di Statistica Sociale

CORSO DI LAUREA: SCIENZE DELL'EDUCAZIONE

DOCENTE: FRANCESCO SANTELLI

Bivariata: secondo caso

- Cosa abbiamo visto nella scorsa lezione? Come si analizza la relazione tra due variabili **quantitative**
- Altri casi? Due variabili qualitative oppure una qualitativa ed una quantitativa
- ***Nel caso di cui variabili qualitative, non si può parlare di:***
  - 1) Covarianza
  - 2) Correlazione
  - 3) Diagramma a dispersione
- Questi concetti non sono applicabili nel caso di variabili qualitative…
  - Come valutiamo allora la relazione che c'è ?

Due variabili qualitative che si incrociano..

		Fumo	
		Sì	No
	M	25	35
Genere	F	20	30

La tabella generata si chiama ***tabella di contingenza***

Essa incrocia le modalità di 2 variabili, formando delle celle che sono dei veri e propri incroci, cioè individui che presentano contemporaneamente ***la modalità in riga e quella in colonna***

Quindi, non si lavora più sui singoli valori numerici delle variabili (che infatti non sono numeri…) ma sulle frequenze che si incrociano

La tabella di contingenza in dettaglio

Variabile Fumo

Modalità fumo: si o no?

Ogni cella, ***frequenza congiunta***

Variabile Genere

		Fumo		
		Sì	No	
	M	25	35	60
Genere	F	20	30	50
		45	65	

Modalità genere: M o F?

Ogni cella esterna
frequenza marginale
(o di riga o di colonna)

La somma delle frequenze marginali, o di riga o di colonna, darà sempre il totale, che qui è uguale a 110

Quindi, N=110

Perché si costruisce?

- Vogliamo valutare la relazione (presente o meno) tra due variabili qualitative
- Vogliamo comunque vedere la ***distribuzione congiunta*** delle due insieme
- Congiunta, cioè ***insieme***: frequenze congiunte nelle celle frutto di incroci
- Vogliamo valutare la ***distribuzione condizionata*** di una rispetto all'altra
- Condizionata, ***cioè una data l'altra***: una sola colonna o una sola riga
- La relazione tra due variabili qualitative prende il nome di associazione, che è l'equivalente di quello che prima era la correlazione per variabili quantitative

Esempi di tabelle di contingenza

	Soda	Coffee	Tea	Water	Total
20-29	10	8	5	2	25
30-39	11	9	2	3	25
40-49	8	9	1	7	25
50-59	9	8	3	5	25
Total	38	34	11	17	100

Health status	Gender	Test result		
		Positive	Negative	Equivocal
Diseased	Male	a	b	c
	Female	d	e	f
Healthy	Male	g	h	i
	Female	j	k	l

Here, one variable depicts health status, another depicts gender, and the third depicts the outcome in the form a test result

Anche variabili numeriche discretizzate (età)
Possono essere usate in tabelle di contingenza

Oltre che due variabili, è possibile addirittura incrociarne ben 3!!
La tabella si complica ma fino a 3 variabili è ancora leggibile

Esercizietto 1

		Religioso			
		No	Poco	Molto	
	Giovane	20	15	10	?
Età	Adulto	?	20	20	60
	Anziano	5	5	?	?
		?	?	45	?

1) Completare la tabella di contingenza, inserendo tutto ciò che manca al posto dei punti interrogativi
2) Individuare:

a) La percentuale di anziani che sono molto religiosi
b) La percentuali di poco religiosi in generale
c) La percentuale, sul totale, di giovani non religiosi

La tabella teorica di indipendenza

		Fumo		
		Sì	No	
	M	25	35	60
Genere	F	20	30	50
		45	65	110

		Fumo		
		Sì	No	
	M	24,54545	35,45455	60
Genere	F	20,45455	29,54545	50
		45	65	110

Cosa hanno in comune le tue tabelle, quella a sinistra che è la tabella di contingenza osservata e quella a destra che è La tabella teorica nel caso di indipendenza?

1) L'impostazione: le modalità e le variabili
2) I marginali di riga e di colonna
3) La numerosità totale: 110

Cosa cambia?? Le celle interne, cioè le frequenze congiunte!!

Calcolo delle frequenze teoriche di indipendenza

Ogni cella è il risultato di: (marginale di riga*marginale di colonna)/totale

Qui dobbiamo calcolare 4 celle teoriche di indipendenza, uguali a:

Sì-M= 60*45/110

No-M= 60*65/110

		Fumo		
		Sì	No	
	M	24,54545	35,45455	60
Genere	F	20,45455	29,54545	50
		45	65	110

Sì-F= 50*45/110

No-F= 50*65/110

Calcolo differenze

Per valutare quanto le nostre due variabili, FUMO e GENERE siano associate, occorre confrontare le frequenze «vere» con quelle «teoriche di indipendenza»:

Vere	Teoriche	Differenza	Quadrati	Quadrati diviso teorico
25	24,55	0,45	0,21	0,008
20	20,45	-0,45	0,21	0,010
35	35,45	-0,45	0,21	0,006
30	29,55	0,45	0,21	0,007
				0,031

Le differenze sommano a zero. Tali differenze sono la «distanza» di ogni frequenza osservata (cella) rispetto Alla frequenza teorica di indipendenza. Si fanno i quadrati e si divide per quella teorica. Si somma tutto e si ottiene un Indice di questa differenza globale tra le due tabelle. Si chiama: ***indice chi quadrato***

Si scrive così: χ^2 Il valore soglia da guardare è 3,84: se la somma dei quadrati diviso il teorico supera questo Valore, allora le due variabili sono fortemente associate; altrimenti, no. In questo caso…no!

Esercizietto 2

		Studenti Lavoratori		
		Sì	No	
	M	10	20	30
Genere	F	10	5	15
		20	25	45
Vere	Teoriche	Differenza	Quadrati	Quadrati diviso teorico

1) Completare la tabella teorica Di indipendenza
2) Completare la tabella per arrivare al calcolo dell'indice «chi-quadrato»
3) Data la soglia di prima che è 3,84…
Cosa affermiamo sull'associazione di queste 2 variabili?

Ultimo caso…

Una variabile qualitativa ed una quantitativa..

Una variabile qualitativa definisce i gruppi, e si suppone che quella quantitativa sia influenza dalla variabile qualitativa

Check grafico: ad esempio, boxplot divisi per i diversi gruppi, ma ne esistono altri diecimila! Istogrammi Per i diversi gruppi, diagrammi a dispersione con colori diversi a seconda del gruppo ecc.

Indici non li vedremo (per vostra fortuna)

La variabile qualitativa si dice indipendente, e quella quantitativa si dice dipendente (perché dipende dai gruppi definiti dalla qualitativa!!)

Mentre prima (correlazione e associazione) le due variabili venivano poste sullo stesso piano, qui

Invece una viene chiaramente «prima» e una «dopo».

I box plot così scaturiti si chiamano ***boxplot condizionati***

Ultimo esercizietto del corso!!

Ore di studio	Genere
12	F
5	M
0	F
6	F
20	M
50	M
3	M
5	F
7	M
10	M
12	F
13	F
24	M
13	F
15	M
20	F

1) Costruire boxplot AFFIANCATI delle ore di studio per genere

Quindi ottenendo un boxplot per i maschi ed uno per le femmine

2) Individuare e commentare la presenza di eventuali valori anomali (outlier)

3) Confrontare le due distribuzione: quale genere è piu studioso?

Quale più variabile?

4) Costruire un box-plot unico di tutte le osservazioni!

Per la prossima volta…

1) Proponete argomenti extra da analizzare… Considerate i vostri lavori di gruppo e cosa può servirvi

Proposte:

A) Mediana dati in classe

B) Analisi di regressione («simile» alla correlazione)

C) Test statistici (Test su valori medi, Test Chi-Quadrato ecc.)

D) Nuove rappresentazioni grafiche?

E) Analisi bivariate per qualitativa/quantitativa: eta quadro, ANOVA

2) Intanto, Link per alcuni esercizi: http://local.disia.unifi.it/gmm/eco/esame/esercizi_elem.pdf

3) Prossima lezione: Vedremo excel e alcuni di questi argomenti da voi proposti

4) Nona lezione: ricapitolazione generale argomenti «base» del corso

5) Decima e ultima lezione: presentazione lavori di gruppo e valutazione dei lavori. Sarete anche voi a valutare i lavori degli altri gruppi!