Lezione 6
Analisi delle relazioni tra due caratteri

POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo finito la parte sulle misure di disuguaglianza.

Abbiamo visto come possiamo studiare la forma di una distribuzione attraverso la simmetria e la curtosi, abbiamo imparato a misurare omogeneità, disomogeneità e concentrazione ad esempio attraverso gli indici costruiti da Corrado Gini.

Ci sono dei dubbi?

Prima di partire con un argomento nuovo, vediamo i «compiti per casa».

Esempi di indicatori

Situazione economica della famiglia > rapporto di composizione,

Famiglie che dichiarano che la propria situazione economica è peggiorata o molto peggiorata rispetto all'anno precedente sul totale.

Disuguaglianza del reddito netto -> rapporto di coesistenza,

Rapporto fra il reddito equivalente totale ricevuto dal 20% della popolazione con il più alto reddito e quello ricevuto dal 20% della popolazione con il più basso reddito.

Affollamento degli istituti di pena -> rapporto di derivazione,

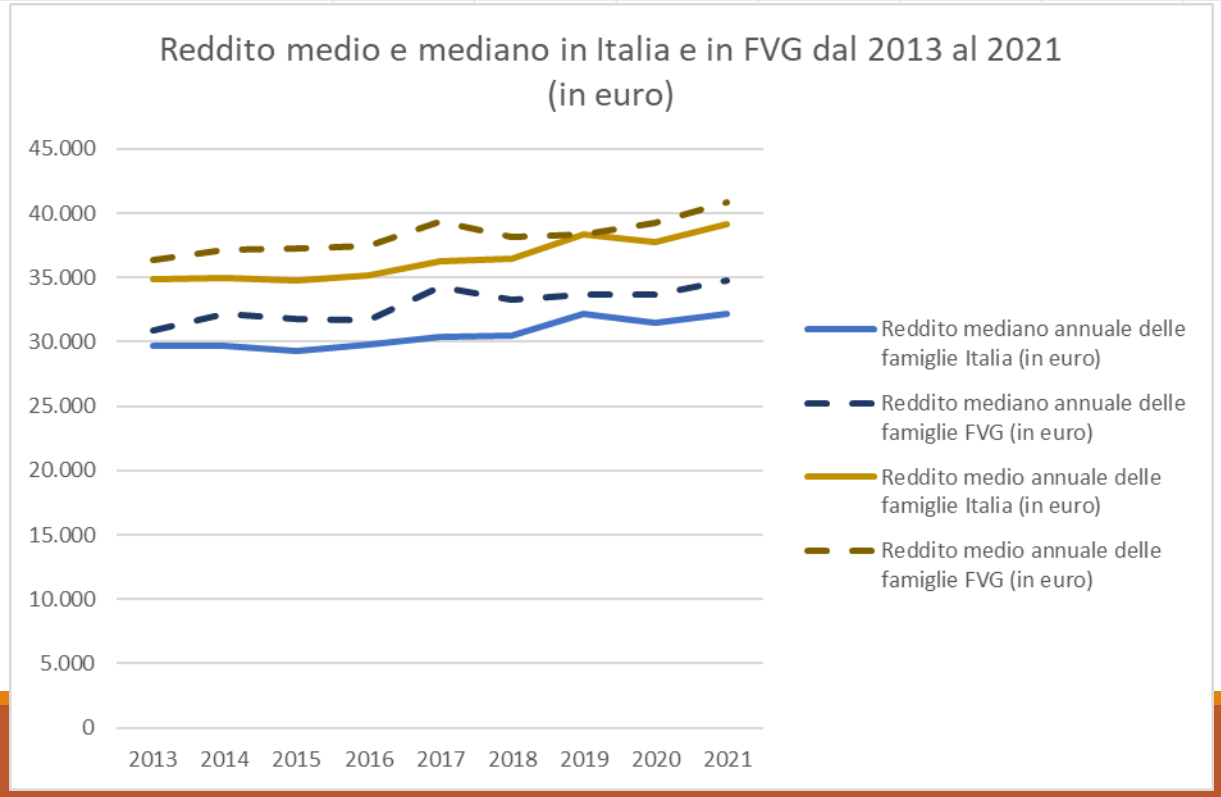
Percentuale di detenuti presenti in istituti di detenzione sul totale dei posti disponibili definiti dalla capienza regolamentare.

Pressione delle attività estrattive -> rapporto di densità.

Volume di risorse minerali non energetiche estratte (metri cubi) per km^2.

Parlando di reddito…

	2013	2014	2015	2016	2017	2018	2019	2020	2021
Reddito mediano annuale delle famiglie Italia (in euro)	29.712	29.694	29.273	29.778	30.352	30.482	32.208	31.495	32.168
Reddito mediano annuale delle famiglie FVG (in euro)	30.921	32.225	31.801	31.692	34.283	33.250	33.650	33.716	34.780
Reddito medio annuale delle famiglie Italia (in euro)	34.866	35.017	34.743	35.204	36.293	36.416	38.319	37.786	39.144
Reddito medio annuale delle famiglie FVG (in euro)	36.405	37.136	37.291	37.483	39.307	38.128	38.321	39.267	40.832



Divario di genere nelle retribuzioni

La volta scorsa abbiamo fatto un esempio in cui abbiamo tirato in ballo la differenza di genere nelle retribuzioni.

I dati presentati erano inventati, ma non sono così distanti dalla realtà.

A questo proposito vi segnalo una pubblicazione dell'Istat, non recentissima, ma certamente molto interessante:

https://www.istat.it/donne-uomini/index.html?lang=it

Secondo voi l'Italia come si colloca in confronto agli altri Paesi dell'UE?

Quanto siamo alla pari?

Divario retributivo di genere per Stato UE*

[differenza tra la retribuzione oraria media lorda dei lavoratori dipendenti di sesso maschile e femminile come % della retribuzione lorda maschile**]

** I dati risalgono al 2020, ad eccezione di quelli di Grecia e Irlanda che sono del 2018*

*** Dati calcolati prendendo come riferimento le imprese con 10 o più dipendenti, ad eccezione della Cechia: dati per le imprese con 1 o più dipendenti.*

Fonte: Eurostat (SDG_05_20) (2022)

Gli indici di variabilità

Nelle scorse lezioni non abbiamo parlato di un indice relativo, il **coefficiente di variazione**.

Varianza e scarto quadratico medio sono indici assoluti, per cui non possono essere utilizzati per confrontare due distribuzioni diverse.

Nel caso di distribuzioni in cui tutti i valori, o almeno la media della distribuzione, sono positivi, si può calcolare il coefficiente di variazione, semplicemente suddividendo la deviazione standard per la media.

$$C_v = \frac{\sigma}{m_x}$$

Analisi delle relazioni tra due caratteri

Fino ad oggi abbiamo visto le operazioni che possiamo fare sulle variabili prese singolarmente.

Abbiamo imparato a calcolare degli indici di tendenza centrale e degli indici di variabilità, che ci hanno aiutato a capire meglio le caratteristiche di una variabile.

Da oggi cominciamo un percorso che ci porterà a studiare congiuntamente due variabili rilevate sullo stesso collettivo.

In termini statistici questo significa individuare e misurare statisticamente il **legame** tra due variabili.

Quando studiamo la relazione tra due variabili, parliamo di **analisi bivariata**.

Analisi delle relazioni tra due caratteri

Nell'ambito delle scienze sociali spesso ci troviamo dover verificare delle ipotesi.

Ad esempio:

«Gli studenti che frequentano regolarmente le lezioni ottengono migliori risultati agli esami».

«Il numero di figli avuti da una donna diminuisce al crescere del suo titolo di studio».

Entrambe le ipotesi mettono in relazione due variabili:

- **frequenza alle lezioni** e **esito degli esami** e
- **livello di istruzione** e **numero di figli**.

Rappresentazione congiunta di due fenomeni

Nella scorse lezioni siamo partiti dalla **distribuzione unitaria semplice**, ovvero dall'elencazione delle modalità di una variabile che si presentano in una matrice dei dati (ad esempio: voti all'esame di statistica sociale).

Si parla di **distribuzione unitaria multipla** quando l'elencazione viene fatta su più variabili.

Le distribuzioni unitarie, non ci aiutano però a cogliere le caratteristiche di un fenomeno, per cui siamo passati alla distribuzione di frequenza per avere una rappresentazione più efficace e sintetica.

Distribuzione unitaria multipla

Ecco un esempio di distribuzione unitaria multipla.

Troviamo un elenco di unità statistiche (da 1 a 14) e i valori che vengono assunti con riferimento alle variabili genere ed età.

Studente	Genere	Età
1	M	21
2	F	20
3	M	22
4	F	19
5	F	20
6	F	20
7	M	19
8	F	19
9	F	21
10	M	20
11	F	22
12	M	19
13	F	21
14	M	21

La distribuzione di frequenza multipla

La distribuzione di frequenza multipla ci consente di rappresentare contemporaneamente due variabili in una tabella, per ogni coppia di modalità (x_i, y_j) di due variabili (X e Y), la **frequenza congiunta** n_{ij}, ovvero il numero di unità che possiedono contemporaneamente la modalità x_i della variabile X e la modalità y_j della variabile Y.

Qual è il numero totale di coppie possibili?

È il prodotto tra le modalità (o le classi) delle due variabili, indipendentemente dalla tipologia di variabili (qualitative o quantitative).

Se la variabile X ha k modalità e la variabile Y ha h modalità, il numero complessivo di coppie di modalità a cui associare una frequenza sarà pari a **k X h**.

Distribuzione di frequenza doppia

Torniamo alla distribuzione unitaria vista prima, dove abbiamo riportato due variabili (il genere e l'età) di un gruppo di studenti.

Le variabili hanno rispettivamente 2 e 4 modalità, quindi il numero complessivo di classi che mettono assieme genere ed età sono 2 X 4, ovvero 8.

(M; 19) (F; 19)

(M; 20) (F; 20)

(M; 21) (F; 21)

(M; 22) (F; 22)

Studente	Genere	Età
1	M	21
2	F	20
3	M	22
4	F	19
5	F	20
6	F	20
7	M	19
8	F	19
9	F	21
10	M	20
11	F	22
12	M	19
13	F	21
14	M	21

Distribuzione di frequenza doppia

Elenco le 8 possibili coppie di modalità e le corrispondenti frequenze assolute:

(x_i, y_j)	n_{ij}
(M; 19)	2
(M; 20)	1
(M; 21)	2
(M; 22)	1
(F; 19)	2
(F; 20)	3
(F; 21)	2
(F; 22)	1
Totale	14

Genere/età	19	20	21	22	Totale
F	2	3	2	1	8
M	2	1	2	1	6
Totale	4	4	4	2	14

Tabella di contingenza

La **tabella di contingenza** (o **tabella a doppia entrata**) riporta le frequenze congiunte delle modalità delle due variabili X e Y.

La tabella di contingenza comprende la frequenza assoluta congiunta delle due variabili, i totali di riga e di colonna e il totale delle unità osservate.

Genere/età	19	20	21	22	Totale riga
F	2	3	2	1	8
M	2	1	2	1	6
Totale colonna	4	4	4	2	14

La tabella di contingenza

I totali di riga e di colonna vengono detti anche frequenze marginali e corrispondono alle frequenze assolute delle singole variabili.

Genere/età	19	20	21	22	Totale riga
F	2	3	2	1	8
M	2	1	2	1	6
Totale colonna	4	4	4	2	14

Genere	Freq. assoluta
F	8
M	6
Totale	14

Età	Freq. assoluta
19	4
20	4
21	4
22	2
Totale	14

La tabella di contingenza

Una riga o una colonna di una tabella di contingenza rappresentano la **distribuzione condizionata**, ovvero la distribuzione alla presenza di una condizione.

Ad esempio la prima riga della tabella di contingenza rappresenta la distribuzione dell'età condizionata al fatto di essere femmina.

Genere/età	19	20	21	22	Totale riga
F	2	3	2	1	8
M	2	1	2	1	6
Totale colonna	4	4	4	2	14

La tabella di contingenza

Come si indica la frequenza condizionata di X a una modalità della variabile Y?

$$X|Y = y_j$$

Si leggerà che la distribuzione della variabile X è condizionata alla modalità j-esima della variabile Y.

Nell'esempio qui sotto leggeremo che la distribuzione della variabile età è condizionata alla modalità femmina della variabile genere.

Genere/età	19	20	21	22	Totale riga
F	2	3	2	1	8
M	2	1	2	1	6
Totale colonna	4	4	4	2	14

Rappresentazione in simboli della distribuzione doppia

Rappresentiamo una distribuzione doppia di frequenze assolute.

Abbiamo due variabili X e Y, rispettivamente con k e h modalità.

- La modalità generica di X è x_i (con i = 1, …, k)
- La modalità generica di Y è y_j (con j = 1, …, h)

La frequenza assoluta congiunta sarà n_{ij} e rappresenta il numero di unità che posseggono congiuntamente la modalità i-esima della variabile X e la modalità j-esima della variabile Y.

I totali di riga e di colonna si indicano rispettivamente con $n_{i.}$ e $n_{.j}$

Rappresentazione in simboli della distribuzione doppia

X/Y	y_1	y_2	…	y_j	…	y_h	Marginali di riga
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1h}	$n_{1.}$
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2h}	$n_{2.}$
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ih}	$n_{i.}$
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{kh}	$n_{k.}$
Marginali di colonna	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.h}$	N

La frequenza relativa

Si può rappresentare, come abbiamo fatto per la frequenza doppia assoluta, la distribuzione di frequenza doppia relativa.

La frequenza doppia relativa per una coppia generica di punti (x_i, y_j) sarà sarà f_{ij} dove:

$f_{ij} = \frac{n_{ij}}{N}$ con $\sum_{i=1}^{k} \sum_{j=1}^{h} f_{ij}$

Le frequenze relative marginali

Nelle tre tabelle che trovate qui a fianco troviamo:

Tabella di frequenza con frequenze congiunte relative

Distribuzione dell'età condizionata al genere

Distribuzione del genere condizionato all'età

Genere/età	19	20	21	22	Totale riga
F	0,14	0,21	0,14	0,07	0,57
M	0,14	0,07	0,14	0,07	0,43
Totale colonna	0,29	0,29	0,29	0,14	1,00

Genere/età	19	20	21	22	Totale riga
F	0,25	0,38	0,25	0,13	1,00
M	0,33	0,17	0,33	0,17	1,00
Totale colonna	0,29	0,29	0,29	0,14	1,00

Genere/età	19	20	21	22	Totale riga
F	0,50	0,75	0,50	0,50	0,57
M	0,50	0,25	0,50	0,50	0,43
Totale colonna	1,00	1,00	1,00	1,00	1,00

Valori caratteristici della distribuzione doppia

Possiamo calcolare tutti gli indici di sintesi e di variabilità (ovviamente a seconda del tipo di variabile) per le due distribuzioni marginali di una distribuzione doppia di frequenza.

Se X è una variabile quantitativa posso calcolare la **media marginale** e la **varianza marginale** di X.

Per le distribuzioni condizionate si parla più correttamente di **media condizionata** e **varianza condizionata**.

Valori caratteristici della distribuzione doppia

Nel caso in cui la variabile X sia quantitativa, possiamo calcolare la media e la varianza della distribuzione marginale.

M(x) = $\frac{1}{N}\sum_{i=1}^{k} x_i\, n_{i.} = \sum_{i=1}^{k} x_i f_{i.}$

Var(x) = $\frac{1}{N}\sum_{i=1}^{k} x_i^2 n_{i.} - \ M^2(X)$

Lo stesso possiamo fare per Y, purché sia quantitativa.

M(y) = $\frac{1}{N}\sum_{j=1}^{h} y_j\, n_{.j} = \sum_{j=1}^{h} y_j f_{.j}$

Var(y) = $\frac{1}{N}\sum_{j=1}^{h} y_j^2 n_{.j} - \ M^2(Y)$

Sintetizziamo una distribuzione doppia

La distribuzione doppia di frequenza può essere sintetizzata nel suo complesso.

Se X e Y sono variabili qualitative ordinate possiamo trovare, come valore di sintesi un punto mediano, ovvero la coppia (Me(X), Me(Y)), ovvero le due mediane delle distribuzioni marginali di X e Y.

Se X e Y sono variabili quantitative possiamo trovare, come valore di sintesi un punto medio, ovvero la coppia (M(X), M(Y)), ovvero le due medie aritmetiche delle distribuzioni marginali di X e Y.

In questo caso posso anche calcolare una misura della variabilità della distribuzione doppia, ovvero la coppia (Var(X), Var(Y)), ovvero le due varianze delle distribuzioni marginali di X e Y.

Adesso esercitiamoci

Esercitiamoci

Chiedo a 5 bambini quanti libri hanno letto per svago nell'ultimo anno.

Queste sono le risposte ottenute:

Bambino	1	2	3	4	5
n_i	7	12	9	15	5

Proviamo a calcolare un indice di tendenza centrale e un paio di indici di variabilità

Esercitiamoci

Bambino	1	2	3	4	5
n_i	7	12	9	15	5

Calcoliamo la media aritmetica:

$M(X)=\frac{1}{N}\sum_{i=1}^{n} x_i n_i$

$M(X)=\frac{(7+12+9+15+5)}{5} = \frac{48}{5} = 9{,}6$

Esercitiamoci

Bambino	1	2	3	4	5
n_i	7	12	9	15	5

Calcoliamo il campo di variazione:

range $= x_{max}$-x_{min}

range = 15 – 5 = 10

Esercitiamoci

Bambino	1	2	3	4	5
n_i	7	12	9	15	5

Calcoliamo la varianza:

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{n}(x_i - m_x)^2$$

$$\mathrm{Var}(X) = \frac{1}{N}\sum_{i=1}^{n}(x_i - 9{,}6)^2$$

$$\mathrm{Var}(X) = [(7 - 9{,}6)^2 + (12 - 9{,}6)^2 + (9 - 9{,}6)^2 + (15 - 9{,}6)^2 + (5 - 9{,}6)^2]/5$$

$$\mathrm{Var}(X) = (6{,}76 + 5{,}76 + 0{,}36 + 29{,}16 + 21{,}16)/5$$

$$\mathrm{Var}(X) = 63{,}20/5 = 12{,}64$$

Esercitiamoci

Bambino	1	2	3	4	5
n_i	7	12	9	15	5

Se conosciamo la varianza (σ^2) possiamo anche calcolare lo scarto quadratico medio (σ)

$\sigma = \sqrt{12{,}64}$ = 3,55

Infine possiamo calcolare il coefficiente di variazione che abbiamo visto oggi:

$C_v = \frac{\sigma}{m_x}$

$C_v = \frac{3{,}55}{9{,}6}$ = 0,37