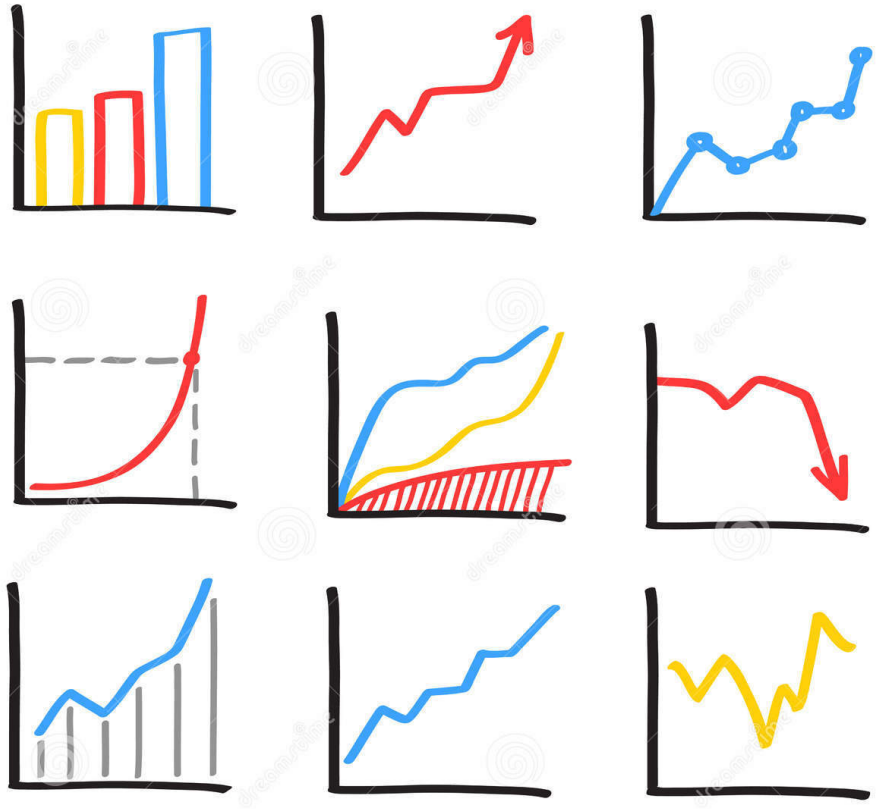
Statistica per l'economia

(9 CFU)



Lezione 7 e 8 aprile 2020

«Le relazioni statistiche. Indipendenza, connessione, associazione. La correlazione e la regressione»

Giuseppe Notarstefano

Le relazioni statistiche tra variabili statistiche: uno sguardo di insieme

- Una fonte molto interessante di informazioni statistiche è quella derivante dall'osservazione congiunta di due o più variabili: parleremo in tal senso di analisi *bivariata* o *multivariata*. Nel nostro corso ci limiteremo all'analisi bivariata.
- Cercheremo di strumenti e metodologie che ci consentano di individuare delle connessioni tra le diverse distribuzioni statistiche, tra i diversi modelli di variabilità espressi e dalla possibilità di spiegare (predire) una variabile attraverso un'altra ossia poter individuare un modello statistico.
- Tutto ciò si fonda sullo studio delle relazioni statistiche tra variabili (ponendo attenzione alle scale di misura) e si traduce attraverso alcuni importanti strumenti analitici che presenteremo di seguito tenendo conto che le relazioni statistiche possono essere:
  - ***Simmetriche*** quando la connessione è bidirezionale e la misura di essa ci informa su come e se sono connesse due diverse variabili/distribuzioni statistiche.
  - ***Asimmetriche*** quando la connessione va in una sola direzione e la misura ci informa su come una variabile/distribuzione «spiega» un'altra variabile/distribuzione.
- Un'altra importante classificazione riguarda la forma matematica di tale relazione che può essere:
  - Lineare
  - Non Lineare

Spiegare o predire una variabile attraverso un'altra variabile

- In diversi campi di applicazioni può essere interessante mettere in relazione i dati per cercare di spiegare o predire una variabile. Quando questa ha una natura statistica, tale relazione consiste nel trovare una relazione tra i due modelli di variabilità.
- Ad esempio sappiamo che in un territorio e in un periodo di tempo osserviamo che quando cresce il consumo delle famiglie cresce di solito anche il reddito disponibile ciò potrebbe voler dire che la variabilità dei consumi può essere spiegata dalla variabilità dei redditi.
- Quando focalizziamo la relazione in modo tale che una variabile viene considerata come indipendente (reddito) e l'altra invece dipende da questa (consumo) stiamo individuando un modello statistico che potrà servirci anche per studiare un modello economico (nel nostro caso il modello keynesiano del consumo).
- Chiameremo **variabile – risposta** la variabile dipendente e **variabile esplicativa** la variabile indipendente (possiamo avere anche più risposte e/o più esplicative nel caso dei modelli multivariati).
- Tali variabili possono essere sia di tipo quantitativo che qualitativo (o ordinale)

Le tabelle a doppia entrata o «cross-tabulazioni»

- Lo strumento analitiche che utilizziamo per studiare una distribuzione in modo congiunto è la tabella a doppia entrata, un oggetto matematico che sintetizza le informazioni nel seguente modo:

X \ Y	y_1	y_2	…	y_j	…	y_h	
x_1	n_{11}	n_{12}	…	n_{1j}		n_{1h}	**Somma per riga**
x_2	n_{21}	n_{22}		n_{2j}		n_{2h}	
…							
x_i	n_{i1}	n_{i2}		n_{ij}		n_{ih}	
…			…				
x_k	n_{k1}	n_{k2}	…	n_{kj}		n_{kh}	
	Somma per colonna						**Somma Totale (= numerosità)**

Le tabelle a doppia entrata: le distribuzioni marginali

- La somma per riga e la somma per colonna ci restituiscono le due distribuzioni di partenza (X e Y): tali distribuzioni sono dette «marginali».
- Nelle celle invece abbiamo la distribuzione congiunta delle due variabili X e Y e n_{ij} rappresenta la frequenza congiunta di X_i e Y_j.
- Dal punto di vista formale Le distribuzioni marginali vengono indicate come segue
  - $n_{i.}$ = marginale di riga che è relativa alla distribuzione di X;
  - $n_{.j}$ = marginale di colonna che è relativa alla distribuzione di Y
  - n.. = distribuzione congiunta di X e Y
  - Il punto indica che le modalità variano
- Le distribuzioni di frequenze congiunte e marginali possono essere espresse in termini relativi e assoluti.
- Le distribuzioni marginali ci offrono una rappresentazione delle singole variabili «indipendentemente» dalle altre.

Le tabelle a doppia entrata: le distribuzioni condizionate

- Osservare congiuntamente due distribuzioni di frequenza vuol dire anche capire come sono connessi due differenti modelli di variabilità: come varia la X al variare delle Y e viceversa.
- Per poter valutare questa variazione congiuntamente dobbiamo poter fissare «un punto di vista» ciò fissata una modalità della X o della Y osservare (che chiameremo **condizionante**) osserviamo la variabilità di Y o X (che chiameremo **condizionata**).
- In termini formali possiamo dire che, data una tabella a doppia entrata potremo ricavare le seguenti distribuzioni condizionate:
  - $X|yj \; \forall \; y = 1, 2, 3, \ldots j \ldots h$
  - $Y|xi \; \forall \; x = 1,2,3, \ldots i \ldots k$
- Le distribuzioni condizionate sono espresse solo in termini relativi.
- Il condizionamento è pertanto una procedura di «relativizzazione» rispetto ad uno dei possibili valori/modalità della variabile condizionante. Utilizziamo un valore di una variabili (es: la X) per «spiegare» la variazione della Y. In questo senso la condizionante si chiama anche «esplicativa» e la condizionata «risposta». La variabile condizionante sarà definita in seguito anche in altri modi: variabile indipendente, regressore o predittore , la variabile risposta invece verrà detta anche variabile dipendente.

Le tabelle a doppia entrata: quali informazioni?

- La tabella a doppia entrata è un modo per rappresentare congiuntamente due variabili di qualsiasi livello di misura (qualitative, quantitative o ordinali).
- Se la tabella a doppia entrata è costruita in riferimento a due variabili qualitative o ordinali, parleremo di «tabella di contingenza»).
- Le informazioni che ricaviamo dalla distribuzione congiunta ci permette di connettere due differenti modelli di variabilità, la distribuzione marginale invece osserva le distribuzioni una indipendentemente dall'altra e la distribuzione condizionata guarda la variabilità in relazione ad un valore fissato dell'altra.

Indipendenza statistica

- Due fenomeni (o meglio le loro misure statistiche) possono essere valutate come indipendenti se la variare dell'una, l'altra non varia.
- Tale nozione di indipendenza statistica è molto interessante e pertanto viene formalizzata nel modo seguente:
- X e Y sono statisticamente indipendenti

$$\text{se } X|y_j = Xi. \text{ e } Y|x_i = Y.j \ \forall \ i \text{ e } j$$

- Cioè se le distribuzioni condizionate per ciascuna modalità condizionante sono tra loro uguali e uguali alla marginale della condizionata.
- Da ciò si ricava che se X e Y sono statisticamente indipendenti, la frequenza congiunta (cella) sarà $n_{ij=} \frac{n_{i.} \times n_{.j}}{n_{..}}$ ovvero $f_{ij} = f_{i.} \times f_{.j}$
- L'indipendenza statistica è simmetrica: ossia se X è statisticamente indipendente da Y allora anche Y sarà statisticamente indipendente da X.

Modello di indipendenza e nozione di «contingenza»

- L'indipendenza statistica definisce anche un modello teorico di indipendenza (n_{ij}'), ossia la condizione teorica in cui le variabili sono considerate indipendenti.
- La contingenza è una misura che si ricava come differenza tra le frequenze effettivamente osservate e quelle «teoriche»:

$$c_{ij} = n_{ij} - n_{ij}' = n_{ij} - \frac{(n_{i.} \times n_{.j})}{N}$$

- È possibile calcolare un indice sintetico dalla somma per riga e per colonna delle contingenze di una tabella a doppia entrata per misurare la connessione tra le due variabili X e Y.

Indici di Associazione

- Tale misura è detta Chi quadrato (χ^2) proposta da Karl Pearson ed è particolarmente utilizzata per misurare l'associazione tra due caratteri in una tabella di contingenza

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

- Tale indice assume valore nullo (=0) nel caso in cui i due caratteri sono indipendenti ovvero non mostrano alcuna connessione o associazione, mentre man mano che cresce l'associazione l'indice assume valori sempre maggiore ma non ha un estremo superiore finito. Si dimostra infatti che il valore massimo del χ^2 è pari ad N per min[(h-1), (k-1)]. Proprio per ovviare tale limite sono stati proposti altre misure (pur essendo tale indice molto utilizzato):
  - L'indice $\phi^2 = \frac{\chi^2}{N} = \frac{\sum_{i=1}^{h}\sum_{j=1}^{k}\left(\frac{c_{ij}}{{n_{ij}}'}\right)^2}{\sum_{i=1}^{h}\sum_{j=1}^{k}{n_{ij}}'}{n_{ij}}'$ che può essere interpretata come una media ponderata dei quadrati delle contingenze relative laddove i pesi sono rappresentati dalle frequenze del modello di indipendenza ${n_{ij}}'$, tale indice ha un campo di variazione tra 0 e 1 (nel caso di tabelle di contingenza 2x2 dette anche «tetracoriche») o maggiore di 1 in tutti gli altri casi. Si dimostra che è possibile calcolare in modo ottimale il massimo di ϕ^2 =min[(h-1), (k-1)] che non risente da N. Tale valore massimo è utile per calcolare un altro indice.
  - L'indice V di Cramér dato dal valore normalizzato di ϕ^2 che varia tra 0 e 1 che si calcola come $(\phi^2/\min[(h-1), (k-1)])^{1/2}$
  - Gli indici χ^2 ϕ^2 e V sono simmetrici. È tuttavia possibile calcolare indici asimmetrici come il λ di Goodman e Kruskal che non tratteremo.

Lo «strano caso» delle tetracoriche

- Molte analisi di «riduzione» della complessità dei dati producono variabili dicotomiche o «dicotomizzate» (un carattere con due modalità, ad esempio presenza assenza di una determinata attitudine o proprietà) che studiate congiuntamente definiscono tabelle 2x2 o **tetracoriche**. Il loro trattamento ha sviluppato strumenti interessanti.

Y / X	y_1	y_2	Totale
x_1	a	b	a+b
x_2	c	d	c+d
totale	a+c	b+d	a+b+c+d

- I prodotti diagonali tra le celle vengono detti «prodotti incrociati» (eng. *cross products*).
  - Essi sono due:
  - axd lungo la diagonale principale
  - bxc lungo la diagonale secondaria
- Esiste **associazione massima** se il prodotto sulla diagonale secondaria è nullo
- Esiste **repulsione massima** (ossia associazione nulla) se il prodotto della diagonale principale è nullo
- Un indice sintetico è quello proposto da Yule detto Q (in onore di Quetelet) = $\frac{(a\times d)-(b\times c))}{(a\times d)+(b\times c)}$

Successi e insuccessi. Odds ed *Odd Ratio*

- È possibile che una tabella tetracorica rappresenti una relazione del tipo «successo/insuccesso»:

X \ Y	Successo	Insuccesso	
x_1	n_{11}	n_{12}	$n_{1\cdot}$
x_2	n_{21}	n_{22}	$n_{2\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	N

Una misura sintetica è rappresentata da Rapporto tra gli Odds, detto infatti **Odds Ratio**:

$$OR = \text{Odds}(Y|x_1) / \text{Odds}(Y|x_2) = \frac{\frac{n_{11}}{n_{12}}}{\frac{n_{21}}{n_{22}}} = \frac{n_{11} \times n_{22}}{n_{12} \times n_{12}}$$

Che viene chiamato anche rapporto dei prodotti incrociati (eng: cross-products ratio)

- Dato un fenomeno condizionante (ad esempio X) Il fenomeno condizionata potrà essere interpretato come rapporto tra successi e insuccessi e misurato dall'Odds come rapporto tra la frequenza dei successi (*casi favorevoli*) e degli insuccessi (*casi sfavorevoli*).
- Avremo così:

$$\text{Odds}(Y|x_1) = \frac{n_{11}}{n_{12}} \text{ e } \text{Odds}(Y|x_2) = \frac{n_{21}}{n_{22}}$$

- Gli Odds sono per costruzione sempre positivi (sono i rapporti tra frequenze!) e possono assumere valori maggiori o minori di 1:
  - > 1 Se i casi favorevoli sono superiori ai casi sfavorevoli
  - <1 se i casi favorevoli sono inferiori ai casi sfavorevoli
  - = 1 nel caso di indipendenza tra X e Y
- Ecco per gli Odds sono una misura del «rischio relativo » (eng: relative risk) di successo o insuccesso.

Ancora su Odds Ratio

- L'Odds ratio è interpretabile come una misura di associazione tra due variabili qualitative dicotomiche, in particolare di dimostra che l'indice Q di Yule può essere riletto in funzione del Rapporto OR:

$$Q = \frac{OR - 1}{OR + 1}$$

- Tale misura normalizzata si rivela interessante per cui se tende ad un 1 evidenza una relazione positiva tra le variabili (se tende a -1 la relazione sarà negativa), mentre se tende a 0 i due caratteri saranno statisticamente indipendenti.
- È possibile calcolare gli ODDS RATIO anche per tabelle di contingenza con variabili politomiche.

Le misure di concordanza

- Quando i caratteri della Tabella a doppia entrata sono ordinali, la misura di connessione che ci occorre deve poter valutare la concordanza o la discordanza delle posizioni definite dalla distribuzione di frequenza congiunta.
- Ci sarà **concordanza** se le modalità di ordine più elevato della X si associano più frequentemente a modalità di ordine elevato della Y mentre le modalità di di ordine basso della X si associano più frequentemente con le modalità di ordine basso della Y.
- Ci sarà **discordanza** se le modalità di ordine elevato della X si associano più frequentemente con le modalità di ordine basso della Y mentre le modalità di ordine basso della X si associano più frequentemente a modalità di ordine elevato della Y
- Le misure che presentiamo sono **simmetriche** e sono
  - L'indice Gamma di Goodman e Kruskal $\gamma = \frac{N_s - N_d}{N_s + N_d}$
  - L'indice Tau di Kendall $\tau = \frac{N_s - N_d}{\sqrt{(N_s + N_d + T_x)(N_s + N_d + T_y)}}$ esiste una versione dell'indice Tau $= \frac{2m(N_s - N_d)}{N^2(m-1)}$
- Dove N_s rappresentano il numero di coppie concordanti e N_d rappresentano il numero di coppie discordanti e T_x e T_y rappresentano il numero delle coppie che presentano rispettivamente uguale modalità rispetto alla X e alla Y.
- N è la numerosità ed *m* è il minimo tra il numero di righe k -1 e di colonne h-1della tabella

Confronti tra graduatorie e misure di cograduazione

- Quando i caratteri della Tabella a doppia entrata sono caratteri quantitativi ma che stiamo trattando come ordinali (graduatorie), parleremo di misure di cograduazione tra posizioni o «ranghi» all'interno della graduatoria.
- Una di queste misure è l'indice Rho di Spearman:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} {d_i}^2}{n(n^2 - 1)}$$

- d sono le differenze osservata tra posizioni (ranghi) per l'i-esima osservazione
- L'indice ρ varia tra -1 (massima discordanza e contrograduazione) e 1 (massima concordanza e cograduazione), è = 0 nel caso di indipendenza.