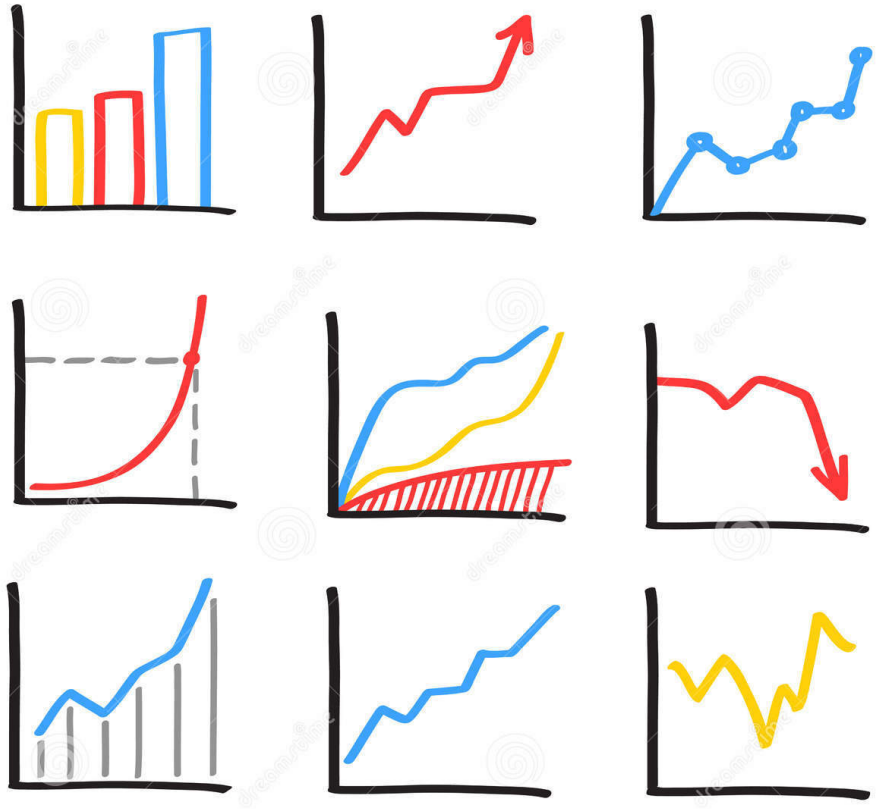
Statistica per l'economia

(9 CFU)



Lezione 24 – 25 marzo 2020

«Eterogeneità e Concentrazione e parametri di forma di una distribuzione»

Giuseppe Notarstefano

Misure di Mutua variabilità

- Gli indici di Mutua Variabilità effettuano confronti a coppie tra le diverse modalità assunte dalle unità del collettivo.
- Si tratta di misure della dissomiglianza o disuguaglianza che vengono calcolate osservando tutte le mutue differenze in valore assoluto

La differenza media semplice

$D = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$ che può assumere valori compresi tra 0 (nel caso di equidistribuzione e 2 volte la media nel caso di massima concentrazione

La differenza media semplice con ripetizione

$D_R = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|$ che può assumere valori compresi tra 0 (nel caso di equidistribuzione e 2 volte la media per (n-1)/n nel caso di massima concentrazione

I due indici sono legati dalla seguente relazione $D = D_R \, n/(n-1)$

Calcolo degli Indici D e D_R nel caso di dati raggruppati per frequenze

- Nel caso in cui i dati sono raggruppati per distribuzioni di frequenze, il calcolo degli indici sarà il seguente:

La differenza media semplice

$$D = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_i - x_j \right| n_i n_j$$

La differenza media semplice con ripetizione

$$D_R = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| x_i - x_j \right| n_i n_j$$

La concentrazione/1

- Una misura di variabilità utilizzata SOLO per i caratteri trasferibili (es: il Reddito o il numero di addetti) fu proposta da Corrado Gini – primo presidente dell'ISTAT ed esponente della scuola statistica italiana. Tale misura è il Rapporto di Concentrazione.
- Quando un carattere è trasferibile vuol dire che sono teoricamente possibili i trasferimenti del valore da un'unità ad un'altra e che la somma dei valori ha un significato preciso: essa rappresenta l'ammontare complessivo del carattere ($A=\sum_{i=1}^{n} x_i$).
- Se tale ammontare fosse distribuito «equamente» tra tutte le unità, di avrebbe pertanto una distribuzione uniforme e ciascuna unità possiederebbe una quota del carattere paria ad A/n uguale per tutti le n osservazioni. È facile osservare come tale quantità consiste anche nel valore medio della distribuzione infatti sarà che $\bar{X} = \frac{A}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$
- In tutti gli altri casi il carattere tende a concentrarsi inversamente ma non proporzionalmente alla variabilità della distribuzione sino al caso limite in cui tutto il carattere sia posseduto da un'unica unità e sia totalmente concentrato su tale unità mentre le altre non possiederebbero nessun valore ($x_i = 0 \ \forall \ i = 1, 2, 3, \dots n-1$). L'ennesima unità *n* possiederebbe pertanto l'intero ammontare complessivo $x_n = A = n\,\bar{X}$.

La concentrazione/2

- Una misura di variabilità utilizzata SOLO per i caratteri trasferibili (es: il Reddito o il numero di addetti) fu proposta da Corrado Gini – primo presidente dell'ISTAT ed esponente della scuola statistica italiana. Tale misura è il Rapporto di Concentrazione.
- Il Rapporto di concentrazione è pertanto una misura relativa compresa tra 0 (=concentrazione nulla ed equidistribuzione) e 1 (=massima concentrazione) e permette confronti tra diverse distribuzioni (ad esempio in tempi o spazi differenti).

La concentrazione/3

- Ci sono diversi modi per calcolare tale rapporto, confrontando ad esempio la distribuzione delle frequenze cumulata (F_i) e quelle del carattere cumulato (Q_i): nei casi intermedi (tra equidistribuzione e concentrazione massima) di concentrazione $F_i \geq Q_i$.
- L'indice sintetico deriva dall'osservazione delle differenze tra F_i e Q_i che vengono sommate e confrontate con il loro valore massimo ($\sum_{i=1}^{n-1} F_i$) per cui avremo che il rapporto di concentrazione R sarà:

$$R = \frac{\sum_{i=1}^{n-1}(F_i - Q_i)}{\sum_{i=1}^{n-1} F_i} = 1 - \frac{\sum_{i=1}^{n-1} Q_i}{\sum_{i=1}^{n-1} F_i}$$

- La concentrazione può anche essere calcolato come misura normalizzata della D in relazione al suo valore massimo

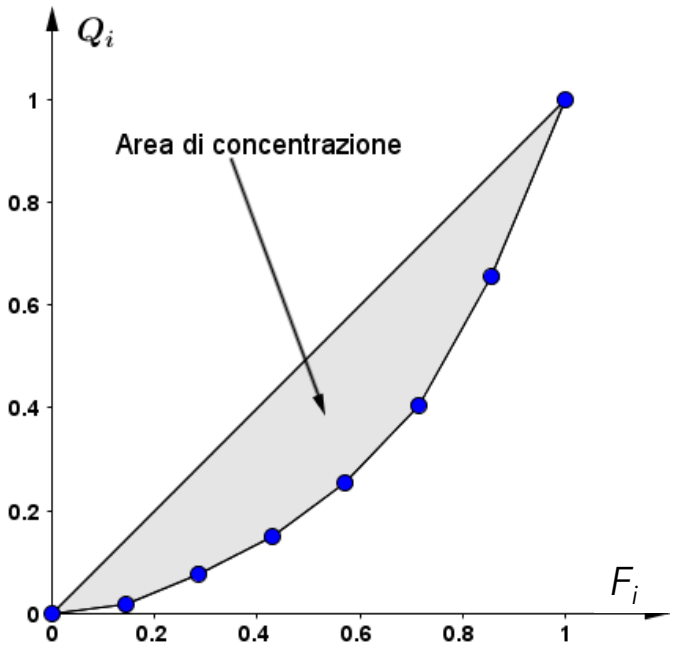
$$R = \frac{D}{2\bar{X}}$$

La curva di Lorenz

- Un modo molto efficace di rappresentare geometricamente e visualizzare graficamente la concentrazione attraverso la costruzione di una spezzata ottenuta dall'unione dei punti che hanno come coordinate i valori della frequenze cumulate (ascissa) e dell'ammontare cumulato (ordinata):

- Tale modello fu proposto da Otto Max Lorenz nel 1912 per analizzare la distribuzione dei redditi negli USA.

- Il segmento le cui coordinate sono (0-0; 1-1) rappresenta la retta di equidistribuzione e la spezzata ottenuta dalle coordinata (Fi; Qi) rappresenta i valori empirci osservati: la concentrazione viene calcolata come area compresa tra queste due curve,



Calcolo della Concentrazione nel caso di dati raggruppati per frequenze

- Utilizzando il suo modello Lorenz ha proposto un calcolo della concentrazione per dati raggruppati che può essere collegato al coefficiente di Gini, la cui formula è la seguente:

$$R = \left[1 - \sum_{j=1}^{k} \left(F_j - F_{j-1}\right)\left(Q_j + Q_{j-1}\right)\right] \frac{k}{k-1}$$

- Laddove k rappresenta il numero delle classi.

Valutare le informazioni statistiche in relazione alla loro variabilità: la normalizzazione e la standardizzazione

- Il coefficiente di variazione –misura relativa di variabilità – ci fornisce uno strumento per valutare la deviazione standard in percentuale rispetto alla media, ossia ci permette di dire in termini «neutri» (= punti percentuali) quale è la variabilità di un fenomeno in relazione alla sua scala di misura il cui ordine medio è viene misurato proprio dalla media (aritmetica)
- Abbiamo parlato talvolta di normalizzazione, con tale espressione intendiamo una trasformazione dei dati originari che tende ad eliminare gli effetti della scala originaria dei valori in modo da permettere le comparazioni su una scala (0,1), si può utilizzare il Min o il Max di una distribuzione sempre in relazione al range. Si preferisce in genere il Minimo perché rappresenta il valore più basso al di sotto del quale non c'è informazione e pertanto il valore trasformato sarà espresso in termini di differenza da dallo zero. Si usa il Massimo in genere per operare un «ribaltamento» ossia una interpretazione alla luce del valore oltre il quale non c'è informazione.

- Tra queste una delle più rilevanti è la standardizzazione di solito indicata con la lettera Z (eng: *Z-score*) che consiste in una trasformazione di scarto rispetto alla medie e relativizzazione rispetto alla propria deviazione standard:

$$z_i = \frac{x_{i-\overline{X_x}}}{\sigma_x}$$

Altre misure di forma di una distribuzione: l'Asimmetria (Skewness) e la Kurtosi

- Abbiamo spesso parlato di asimmetria, indicando la prevalenza di osservazioni su valori bassi (Asimmetria positiva – gobba a sinistra) o su valori alti (Asimmetria negativa – gobba a destra). possiamo introdurre desso una misura specifica della skewness proposta da Pearson:

$$S_{k1} = \frac{\bar{X} - M_o}{\sigma^2} \cong \frac{\bar{X} - M_e}{\sigma^2}$$

- Un'altra misura proposta da Fisher utilizza la variabile z-score o meglio la sua media cubica potenziata:

$$S_{k2} = \frac{\sum_{i=1}^{n} (x_i - \bar{X})^3}{n\sigma^3}$$

Il «peso» delle code: la Kurtosi

- In principio era l'ornitorinco (*platypus)*…!
- Pearson propone di chiamare kurtosi (o *dis-normalità*) la tendenza di una distribuzione di curvarsi
- Un effetto specifico connesso alla kurtosi è la leggerezza (o pesantezza) delle code ossia la rilevanza in termini di frequenze rilevate dei valori estremi
- È possibile calcolare la kurtosi con la formula proposta da Fisher:

$$\gamma = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^4}{n\sigma^4} - 3$$