

Statistica I

Unità F.2: mutua variabilità e concentrazione

Tommaso Rigon

Università Milano-Bicocca



Argomenti affrontati

- Differenza semplice media
- Mutua variabilità e caratteri trasferibili
- Rapporto di concentrazione di Gini
- Curva di Lorenz

Riferimenti al libro di testo

- §5.4 — §5.5

Descrizione del problema



- Siamo interessati a confrontare il **valore** in euro delle **squadre di calcio** di Serie A, inteso come la somma del valore di mercato dei singoli giocatori.
- Siamo interessati a quantificare la **disuguaglianza** tra i valori delle squadre.

I dati grezzi

- I dati provengono dal sito web <https://www.transfermarkt.it>
- I dati considerati fanno riferimento agli anni 2010 e 2021. Le squadre che popolano la Serie A sono 20, ma sono ovviamente diverse nei due anni considerati.
- **Nota.** Il numero di giocatori per squadra era mediamente diverso negli anni 2010 e 2021. Inoltre, il valore in euro risente dell'**inflazione**.

Valore della squadra in milioni di euro, anno 2010

[1]	390.15	361.35	332.43	257.00	207.73	168.70	142.50	129.15	116.25
[10]	112.55	103.43	87.80	74.13	68.70	64.90	63.45	55.20	49.10
[19]	47.18	35.50							

Valore della squadra in milioni di euro, anno 2021

[1]	602.90	525.90	518.55	476.70	429.25	415.35	311.30	253.80	216.90
[10]	185.93	147.18	145.50	120.80	109.35	103.00	96.60	77.65	74.70
[19]	68.50	36.15							

Alcuni indici descrittivi

- Per cominciare, descriviamo i dati utilizzando i concetti che abbiamo appreso finora.

	Anno 2010	Anno 2021
Media di valore	143.36	245.80
Mediana di valore	107.99	166.56
Scarto quadratico medio di valore	106.98	178.11

- Pertanto, i valori complessivi dei campionati 2010 e 2021 sono circa 2.87 e 4.92 miliardi di euro, rispettivamente.
- Inoltre, le varianze dei due campionati sono molto diverse.
- Possiamo quindi concludere che il campionato 2021 è caratterizzato da una maggiore disuguaglianza? **No**, perché i redditi complessivi nei due anni sono diversi.

La differenza semplice media

- Prima di proseguire con la nostra analisi, introduciamo un nuovo indice di variabilità.
- Differenza semplice media. La differenza semplice media dei dati x_1, \dots, x_n è

$$\Delta = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|.$$

- La differenza semplice media è quindi pari alla media delle distanze tra tutte le coppie di valori distinti.
- Questa definizione ricorda quella della **varianza**, in cui si è usato il valore assoluto al posto del quadrato.

Differenza semplice media: definizione alternativa

- La differenza semplice media può anche essere definita in una **maniera alternativa**, tramite una formula dal calcolo più agevole
- Differenza semplice media. La differenza semplice media dei dati x_1, \dots, x_n è

$$\Delta = \frac{4}{n(n-1)} \left(\sum_{i=1}^n i x_{(i)} \right) - 2 \bar{x} \frac{n+1}{n-1},$$

dove $x_{(1)}, \dots, x_{(n)}$ è il campione ordinato.

- Questa definizione è **più semplice da utilizzare** in pratica, anche se la sua interpretazione è **meno trasparente**.

Dimostrazione

- L'equivalenza tra le due definizioni, a meno del termine $n(n-1)$, si ottiene:

$$\begin{aligned}\sum_{i=1}^n \sum_{j=1}^n |x_i - x_j| &= 2 \sum_{i=1}^n \sum_{j=1}^i (x_{(i)} - x_{(j)}) = 2 \sum_{i=1}^n \sum_{j=1}^i x_{(i)} - 2 \sum_{i=1}^n \sum_{j=1}^i x_{(j)} \\ &= 2 \sum_{i=1}^n i x_{(i)} - 2 \{x_{(1)} + (x_{(1)} + x_{(2)}) \cdots + (x_{(1)} + \cdots + x_{(n)})\} \\ &= 2 \sum_{i=1}^n i x_{(i)} - 2 \{n x_{(1)} + (n-1)x_{(2)} + \cdots + 2x_{(n-1)} + x_{(n)}\} \\ &= 2 \sum_{i=1}^n i x_{(i)} - 2 \sum_{i=1}^n (n-i+1)x_{(i)} \\ &= 2 \sum_{i=1}^n i x_{(i)} + 2 \sum_{i=1}^n i x_{(i)} - 2(n+1) \sum_{i=1}^n x_{(i)} \\ &= 4 \sum_{i=1}^n i x_{(i)} - 2n(n+1)\bar{x}\end{aligned}$$

Proprietà della differenza semplice media

- **Proprietà.** La differenza semplice media è per costruzione sempre **maggiore o uguale a zero**, ovvero

$$\Delta \geq 0.$$

- Inoltre, la differenza semplice media è esattamente pari a zero solo se i dati sono uguali tra loro.

- Infatti ad esempio se

$$x_1 = x_2 = \dots = x_n = a,$$

dove $a \in \mathbb{R}$ è una costante qualsiasi, allora

$$\Delta = \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j| = \frac{1}{n(n-1)} (|a - a| + \dots + |a - a|) = 0.$$

- Si può dimostrare anche il viceversa: se $\Delta = 0$ allora necessariamente le osservazioni sono uguali tra loro.

Proprietà della differenza semplice media

- Proprietà. Se consideriamo i dati trasformati y_1, \dots, y_n , tali che

$$y_i = a + bx_i, \quad i = 1, \dots, n,$$

dove $a, b \in \mathbb{R}$ sono due numeri qualsiasi e siano Δ_x e Δ_y i rispettivi indici. Allora:

$$\Delta_y = |b|\Delta_x.$$

- La dimostrazione segue dalle proprietà delle sommatorie. Infatti:

$$\sum_{i=1}^n \sum_{j=1}^n |y_i - y_j| = \sum_{i=1}^n \sum_{j=1}^n |a + bx_i - a - bx_j| = |b| \sum_{i=1}^n \sum_{j=1}^n |x_i - x_j|.$$

Il risultato segue dividendo il tutto per $n(n-1)$.

- La differenza media semplice delle y_i pertanto non dipende dalla costante a , come accade per la varianza.

Commento al problema

	Anno 2010	Anno 2021
Somma di valore	2867.2	4916.01
Scarto quadratico medio di valore	106.98	178.11
Differenza semplice media di valore	117.68	206.27

- La differenza semplice media è un indice **complementare** alla varianza.
- Tuttavia, questo indice non consente di rispondere alla domanda originaria, ovvero misurare la **disuguaglianza** tra le disponibilità economiche tra le varie squadre.
- Sebbene la variabilità dell'anno 2021 sia maggiore di quella dell'anno 2010, questo è dovuto al fatto che il valore complessivo è differente nei due anni.
- Abbiamo pertanto bisogno di una sorta di indice **normalizzato**.

Caratteri trasferibili

- Una variabile si dice **trasferibile**, se è possibile immaginare il “trasferimento” di un certo ammontare da un’unità statistica ad un’altra.
- Sono esempi di variabili trasferibili il reddito, il consumo di un bene, i finanziamenti alle imprese, le nascite rispetto al comune di residenza della madre, etc.
- Consideriamo in questo contesto distribuzioni x_1, \dots, x_n a **valori positivi**.
- Il totale di un certo bene presente nella popolazione è pari alla somma

$$S = \sum_{i=1}^n x_i,$$

ovvero valore complessivo.

- Nel caso di variabili trasferibili è pertanto possibile la determinazione teorica della **variabilità minima** e **variabilità massima** all’interno di una prefissata distribuzione.

Minima e massima variabilità

Minima variabilità

- La minima variabilità (disuguaglianza) si osserva se le unità statistiche sono tutte uguali. Le unità statistiche sono equamente distribuite.
- Si osservi che in questo caso i dati sono nella forma:

$$x_1 = \dots = x_n = S/n.$$

Di conseguenza si avrà che $\Delta = 0$.

Massima variabilità

- La massima variabilità (disuguaglianza) si osserva se una singola unità statistica racchiude l'intero ammontare, mentre le rimanenti unità sono pari a 0.
- Si osservi che in questo caso i dati sono ad esempio nella forma:

$$x_1 = S, \quad x_2 = \dots = x_n = 0.$$

- La differenza media semplice dei dati, in condizione di massima variabilità è tale che

$$\Delta = \frac{4n}{n(n-1)}n\bar{x} - 2\bar{x}\frac{n+1}{(n-1)} = 2\bar{x}.$$

- Il teorema successivo chiarisce inoltre che $2\bar{x}$ è il massimo valore ottenibile.

Teorema

La differenza semplice dei dati x_1, \dots, x_n tali che $\sum_{i=1}^n x_i = S$ è tale che

$$\Delta \leq 2\bar{x},$$

ed è pari al valore massimo $\Delta = 2\bar{x}$ **se e solo se** una singola osservazione è pari a S .

- La dimostrazione del teorema verrà affrontata in un momento successivo.

Il rapporto di concentrazione di Gini

- Questi risultati portano alla definizione di un indice di variabilità **normalizzato**.
- Rapporto di concentrazione di Gini. L'indice di concentrazione di Gini della variabile trasferibile x_1, \dots, x_n è:

$$\mathcal{R} = \frac{\Delta}{(\text{massimo valore di } \Delta)} = \frac{\Delta}{2\bar{x}}.$$

- Per definizione, si ha che $0 \leq \mathcal{R} \leq 1$.
- L'interpretazione dell'indice è agevole: vale 0 in caso di **perfetta redistribuzione** del totale S , mentre vale 1 in condizione di **massima disuguaglianza**.

Proprietà del rapporto di concentrazione di Gini

- Proprietà. Se consideriamo i dati trasformati y_1, \dots, y_n , tali che

$$y_i = b x_i, \quad i = 1, \dots, n,$$

dove $b > 0$ è un numero positivo e siano \mathcal{R}_x e \mathcal{R}_y i rispettivi indici. Allora:

$$\mathcal{R}_y = \mathcal{R}_x.$$

- La dimostrazione segue dalle proprietà di Δ e della media aritmetica. Infatti:

$$\mathcal{R}_y = \frac{\Delta_y}{2\bar{y}} = \frac{|b|\Delta_x}{2b\bar{x}} = \frac{b\Delta_x}{2b\bar{x}} = \frac{\Delta_x}{2\bar{x}} = \mathcal{R}_x.$$

- Questa proprietà implica che una generica trasformazione di scala non modifica il valore di \mathcal{R} . In particolare, potremmo considerare equivalentemente i dati $y_i = x_i/S$.
- Poiché l'indice **non dipende dal totale** S , è possibile usare il rapporto di concentrazione di Gini per confrontare scenari in cui i valori complessivi differiscono.

Risultati e commento al problema

	Anno 2010	Anno 2021
Somma di valore	2867.2	4916.01
Media di valore	143.36	245.80
Differenza semplice media di valore	117.68	206.27
Rapporto di concentrazione (Gini) di valore	0.4105	0.4196

- L'indice di concentrazione di Gini è abbastanza elevato: questo indica una pronunciata disuguaglianza in entrambi gli anni.
- Complessivamente, i due campionati di Serie A sono **molto simili** in termini di disuguaglianza.

La curva di Lorenz

- La **curva di Lorenz** è uno strumento grafico con il quale possiamo analizzare la disuguaglianza di una distribuzione.
- Siano x_1, \dots, x_n dei dati a valori positivi. La curva di Lorenz è la **spezzata** che unisce le coppie di punti (p_i, q_i) , dove

$$p_i = \frac{i}{n}, \quad q_i = \frac{1}{S} \sum_{j=1}^i x_{(j)},$$

per $i = 1, \dots, n$ e dove si pone $p_0 = q_0 = 0$ per convenzione.

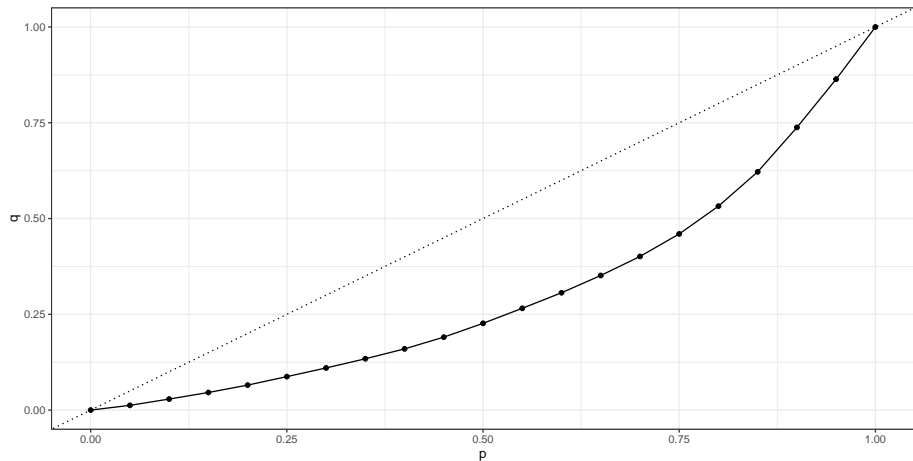
- In altri termini i punti della curva sono così composti:

$$p_i = (\text{Frazione cumulata dei primi } i \text{ individui}),$$

mentre si ha che

$$q_i = (\text{Frazione cumulata di } S \text{ posseduta dai primi } i \text{ individui}).$$

La curva di Lorenz



■ Curva di Lorenz calcolata con i dati del 2010.

La curva di Lorenz

- Se tutte le osservazioni $x_1 = \dots = x_n = S/n$ sono uguali, ovvero ne caso di **minima variabilità**, allora

$$p_i = \frac{i}{n}, \quad q_i = \frac{1}{S} \sum_{j=1}^i x_{(j)} = \frac{1}{S} \frac{S}{n} i = \frac{i}{n},$$

ovvero si ottiene $p_i = q_i$, per $i = 1, \dots, n$.

- In altri termini, quando la **disuguaglianza** (variabilità) è **nulla** allora la curva di Lorenz coincide con la **bisettrice**.

- Nel caso di **massima variabilità**, ovvero quando $x_{(1)} = \dots = x_{(n-1)} = 0$ e quando $x_{(n)} = S$, allora

$$p_i = \frac{i}{n}, \quad q_i = 0, \quad i = 0, \dots, n-1,$$

mentre $p_n = q_n = 1$.

- In altri termini quando la **disuguaglianza** (variabilità) è **massima** la curva di Lorenz coincide con una retta costante pari a 0, con un salto nell'ultimo punto.

La curva di Lorenz

- La **prima coppia di valori** della curva di Lorenz è

$$(p_1, q_1) = \left(\frac{1}{n}, \frac{x_{(1)}}{S} \right),$$

e rappresenta l'individuo più povero.

- **Esercizio - proprietà**. Per costruzione, le coordinate dei punti della curva di Lorenz sono tali che

$$q_i \leq p_i, \quad i = 1, \dots, n.$$

- Pertanto, le differenze $(p_i - q_i) \geq 0$ costituiscono misure dirette della **concentrazione**.
- In altri termini, la differenza $p_i - q_i$ misura in proporzione la quota di S che manca ai primi i individui per trovarsi in una posizione di equidistribuzione, ovvero $p_i = q_i$.

Rapporto di concentrazione di Gini

- Pertanto, un indice di concentrazione potrebbe basarsi sulle **differenze normalizzate**

$$\frac{p_i - q_i}{p_i}, \quad i = 1, \dots, n-1,$$

le quali sono tali che $0 \leq (p_i - q_i)/p_i \leq 1$.

- Una possibilità è considerare la **media aritmetica ponderata** di queste differenze, ottenendo il seguente indice di concentrazione.
- **Rapporto di concentrazione di Gini**. L'indice di concentrazione di Gini della variabile trasferibile x_1, \dots, x_n è:

$$\mathcal{R} = \frac{\sum_{i=1}^{n-1} p_i \left(\frac{p_i - q_i}{p_i} \right)}{\sum_{i=1}^{n-1} p_i} = \frac{\sum_{i=1}^{n-1} (p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = 1 - \frac{2}{n-1} \sum_{i=1}^{n-1} q_i,$$

sfruttando nell'ultimo passaggio il fatto che $\sum_{i=1}^{n-1} p_i = (n-1)/2$.

- Il rapporto di concentrazione di Gini è **standardizzato**, ovvero tale che $0 \leq \mathcal{R} \leq 1$, perchè è una media ponderata dei valori $(p_i - q_i)/q_i$, a loro volta compresi tra 0 e 1.

Rapporto di concentrazione di Gini

- Il rapporto di concentrazione di Gini \mathcal{R} è stato apparentemente definito due volte, tramite strumenti diversi.
- Sebbene questo non sia immediatamente ovvio, le due **definizioni coincidono**.

Teorema

L'indice di concentrazione di Gini della variabile trasferibile x_1, \dots, x_n è:

$$\mathcal{R} = \frac{\sum_{i=1}^{n-1} (p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{\Delta}{2\bar{x}}.$$

- Questa equivalenza è di grande importanza a fini **interpretativi**.
- Inoltre, poichè $\mathcal{R} \leq 1$, questa equivalenza **dimostra** anche la disequazione $\Delta \leq 2\bar{x}$.

Dimostrazione

- In primo luogo, si mostra che:

$$\begin{aligned}\sum_{i=1}^{n-1} q_i &= \frac{1}{S} \sum_{i=1}^{n-1} \sum_{j=1}^i x_{(j)} = \frac{1}{S} \{x_{(1)} + (x_{(1)} + x_{(2)}) \cdots + (x_{(1)} + \cdots + x_{(n-1)})\} \\ &= \frac{1}{S} \{(n-1)x_{(1)} + (n-2)x_{(2)} + \cdots + x_{(n-1)} + 0 \times x_{(n)}\} \\ &= \frac{1}{S} \sum_{i=1}^n (n-i)x_{(i)} = \frac{n}{S} \sum_{i=1}^n x_{(i)} - \frac{1}{S} \sum_{i=1}^n i x_{(i)} = n - \frac{1}{S} \sum_{i=1}^n i x_{(i)}.\end{aligned}$$

- Di conseguenza sostituendo e ricordando che $S = n\bar{x}$, si ottiene

$$\begin{aligned}\mathcal{R} &= 1 - \frac{2}{n-1} \sum_{i=1}^{n-1} q_i = 1 - \frac{2}{n-1} \left[n - \frac{1}{S} \sum_{i=1}^n i x_{(i)} \right] \\ &= \frac{1}{2\bar{x}} \left[\frac{4}{n(n-1)} \left(\sum_{i=1}^n i x_{(i)} \right) - 2\bar{x} \frac{n+1}{n-1} \right] = \frac{\Delta}{2\bar{x}}.\end{aligned}$$

Area sopra la curva di Lorenz

- Un modo per quantificare la disuguaglianza è calcolare l'**area** compresa tra la **curva di Lorenz** e la **bisettrice**.
- L'area tra la curva e la bisettrice è necessariamente compresa tra 0 (minima disuguaglianza) e $1/2$ (massima disuguaglianza).
- Di conseguenza è possibile considerare la quantità: $2 \times$ ("Area di concentrazione").

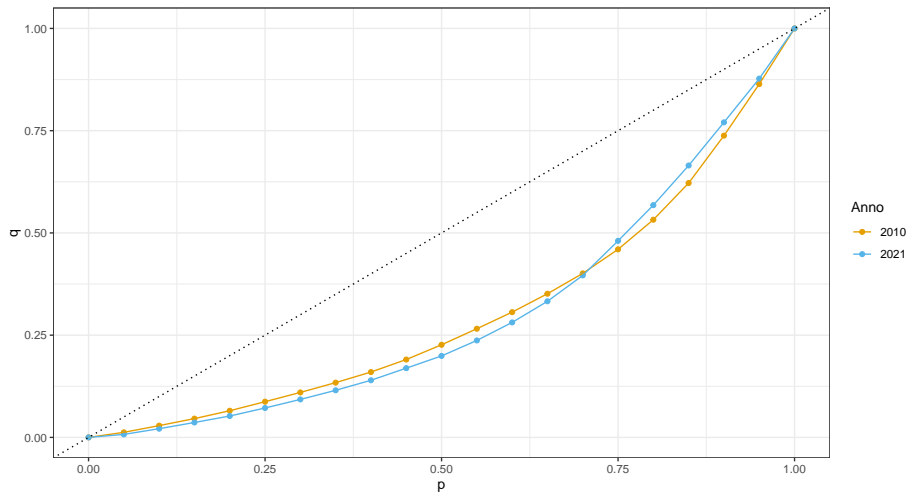
Teorema (senza dimostrazione)

L'indice di concentrazione di Gini della variabile trasferibile x_1, \dots, x_n è:

$$\mathcal{R} = \frac{n-1}{n} [2 \times (\text{"Area di concentrazione"})]$$

- Questo risultato fornisce un'ulteriore giustificazione ed interpretazione dell'indice di concentrazione di Gini.

Confronto tra le curve di Lorenz



■ Curva di Lorenz calcolata con i dati del **2010** e del **2021**.