Lezione 5
Le misure di disuguaglianza

PROF. ROBERTO COSTA

SCIENZE DELL'EDUCAZIONE - STATISTICA SOCIALE (305SF)

Dove eravamo rimasti

Nella scorsa lezione abbiamo finito la parte sulle misure centrali e abbiamo iniziato a parlare delle misure di disuguaglianza.

Parlando di misure centrali abbiamo visto il criterio di Chisini:

$f(x_1, x_2 , \dots , x_n) = f(M,M, \dots,M)$

Abbiamo visto la sua applicazione nelle medie (aritmetica, geometrica, armonica e quadratica)

Abbiamo visto assieme come, da sole, le misure centrali ci aiutano a sintetizzare una distribuzione, ma non ci consentono di spiegare tutte le caratteristiche di una distribuzione.

Abbiamo visto misure di disuguaglianza, partendo dagli indici di variabilità di dispersione (range o campo di variazione) e rispetto ad un centro (varianza e scarto quadratico medio).

Dove eravamo rimasti

Facciamo un semplice esercizio su varianza e scarto quadratico medio, partendo dalla formula:

$\sigma^2 = \Sigma(x_i - m)^2/N$

Abbiamo tre osservazioni dei giorni attesi per ottenere un appuntamento per una visita specialistica:

10, 13, 16

Calcoliamo la media aritmetica:

M = (10+13+16)/3 = 39/3 = 13

Calcoliamo la varianza:

$\sigma^2 = \Sigma(x_i - m)^2/N = [(10\text{-}13)^2 + (13\text{-}13)^2 + (16\text{-}13)^2]/3 = (9 + 0 + 9)/3 = 18/3 = 6$

Lo scarto quadratico medio è pari alla radice quadrata della varianza.

Omogeneità e eterogeneità

Parlando di una distribuzione possiamo valutare come i valori rilevati si distribuiscano tra le modalità.

Distinguiamo quindi tra:

Massima omogeneità: quando tutte le osservazioni confluiscono in un'unica modalità.

Massima eterogeneità: quando le osservazioni si distribuiscono equamente tra tutte le modalità.

Ad es. intervisto gli studenti di un corso di studio e chiedo il genere.

Massima omogeneità: 10 interviste, 10 rispondenti di genere femminile e nessuno di genere maschile.

Massima eterogeneità: 5 rispondenti di genere femminile e 5 di genere maschile.

Omogeneità e eterogeneità

Nell'ambito delle scienze sociali, queste due condizioni estreme sono difficilmente riscontrabili, per cui può essere molto utile uno strumento che ci aiuti a misurare questo concetto.

Corrado Gini formulò **l'indice di eterogeneità**:

$$E = 1 - \sum_{i=1}^{k} f_i^2$$

Dove $f_i = \frac{n_i}{N}$

Il valore minimo sarà 0 quando una modalità ha frequenza relativa pari a 1 e tutte le altre modalità hanno frequenza relativa pari a 0.

Il valore massimo sarà pari a (k-1)/k dove k è il numero di modalità.

Omogeneità e eterogeneità

Riprendiamo l'esempio precedente, 10 interviste rileviamo la variabile genere che ha 2 modalità di risposta (k = 2). Nel primo caso 10 femmine (f =1) e 0 maschi (f = 0), nel secondo 5 femmine (f = 0,5) e 5 maschi (f = 0,5).

$E = 1 - \sum_{i=1}^{k} f_i^2$

Dove $f_i = \frac{n_i}{N}$

Nel primo caso avremo: E = 1 – (1^2 + 0^2), ovvero E = 1 – 1 quindi E = 0

Nel secondo caso:

E = 1 – ($0{,}5^2$ + $0{,}5^2$), ovvero E = 1 – 0,25 – 0,25 E = 0,5 che corrisponde al valore massimo

(2 – 1)/2 = 0,5

La rappresentazione grafica della variabilità

Utilizziamo il **grafico a scatola** (più comunemente chiamato box-plot) per rappresentare in un solo grafico diversi valori caratteristici di una variabile.

Proposto inizialmente da J. Tukey, viene utilizzato per farsi velocemente un'idea della distribuzione di una variabile.

Il box-plot

È composto dai seguenti elementi:

Una **linea continua** che rappresenta il valore della **mediana**.

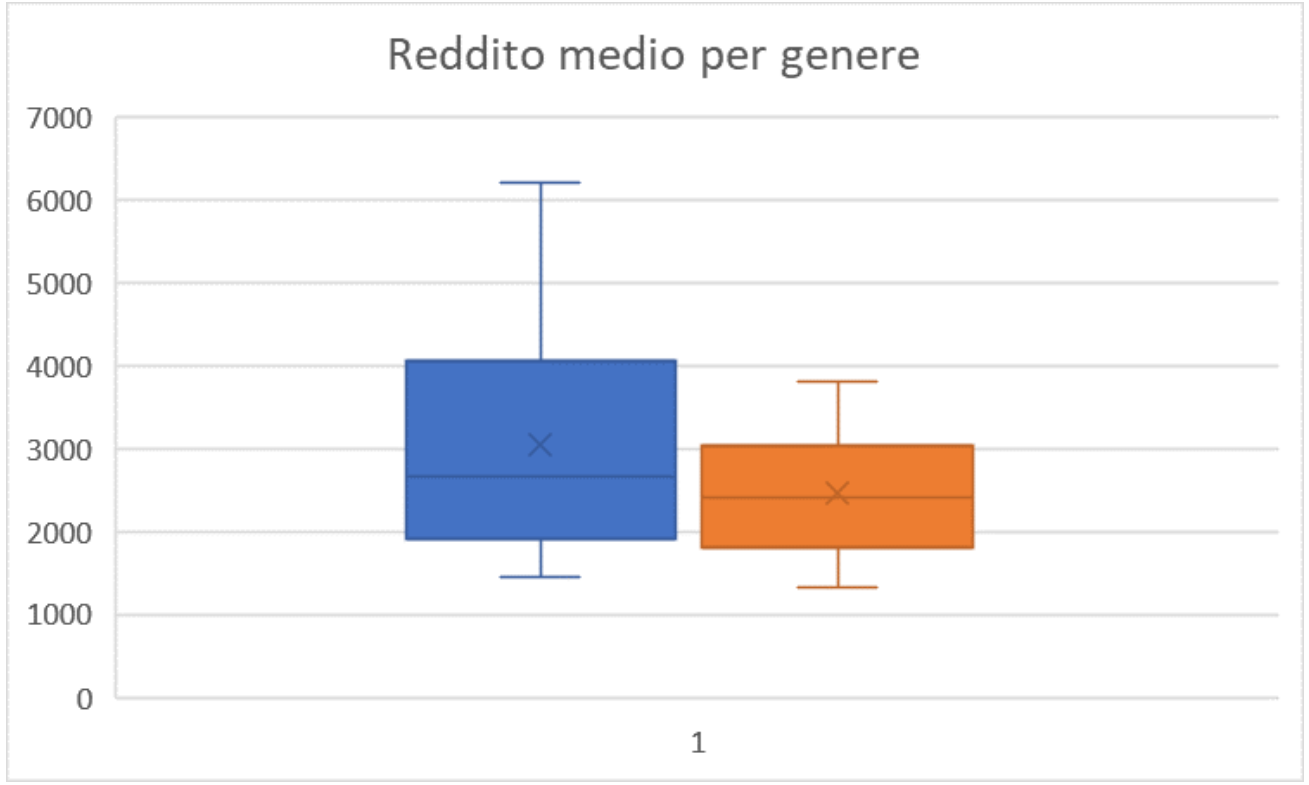
Un **box** che congiunge i valori che occupano il 1° e il 3° quartile, dove si colloca il 50% dei dati.

Due **whiskers** (baffetti) che congiungono i valori che non superano 1,5 volte l'estensione del box.

Eventuali **outliers** (valori estremi) che sforano i whiskers.

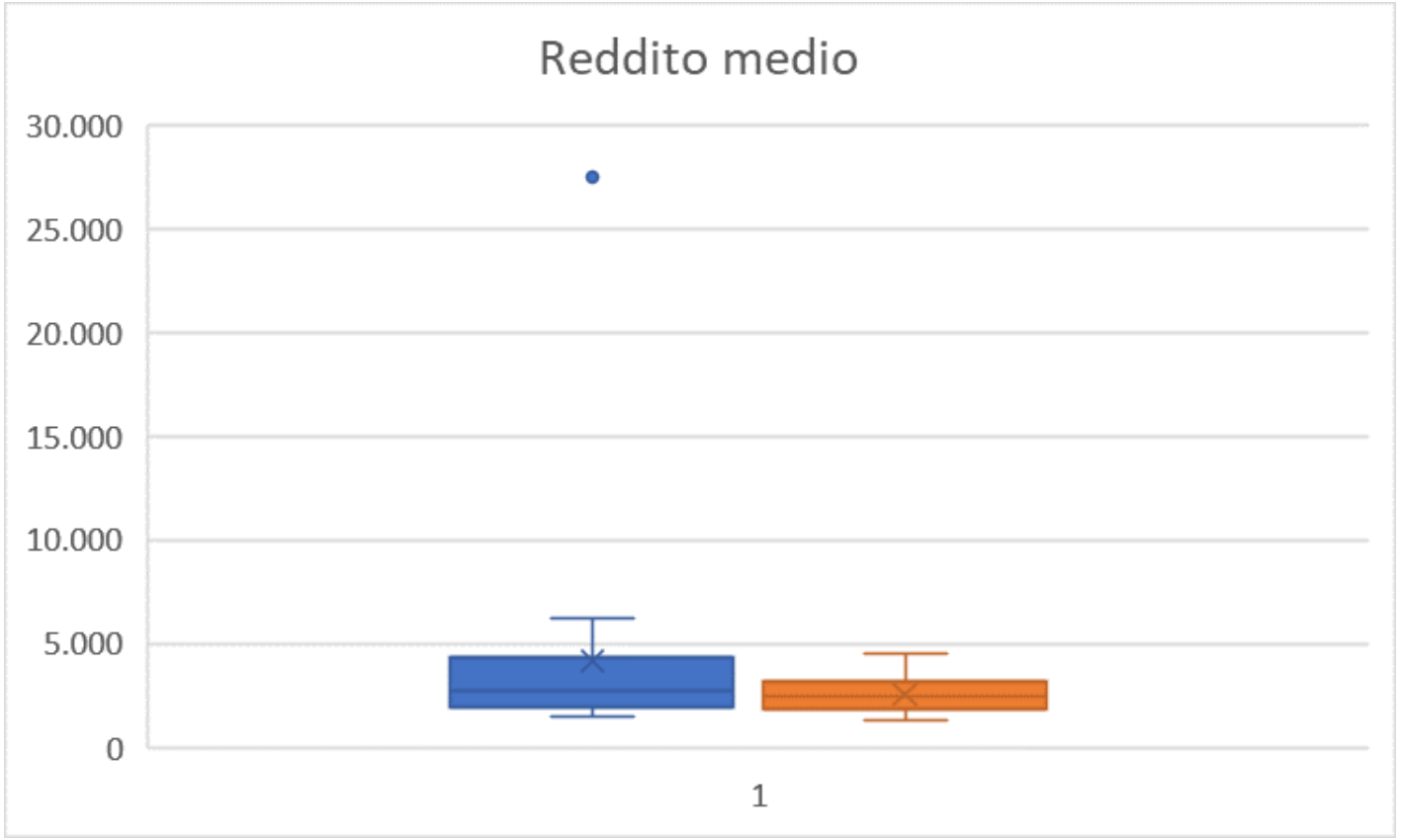
Il box-plot

Reddito medio	
maschi	femmine
1.450	1.320
1.540	1.465
1.650	1.510
1.720	1.600
1.865	1.745
1.950	1.875
2.100	2.000
2.250	2.180
2.350	2.270
2.520	2.320
2.650	2.400
2.850	2.550
2.950	2.635
3.150	2.700
3.540	2.850
3.825	2.975
4.300	3.100
4.450	3.300
5.200	3.450
5.450	3.600
6.200	3.800



Il box-plot

Reddito medio	
maschi	femmine
1.450	1.320
1.540	1.465
1.650	1.510
1.720	1.600
1.865	1.745
1.950	1.875
2.100	2.000
2.250	2.180
2.350	2.270
2.520	2.320
2.650	2.400
2.850	2.550
2.950	2.635
3.150	2.700
3.540	2.850
3.825	2.975
4.300	3.100
4.450	3.300
5.200	3.450
5.450	3.600
6.200	3.800
27.500	4.500



La forma di una distribuzione

Due variabili possono avere la stessa variabilità e lo stesso centro, ma essere molto diverse per il comportamento della distribuzione.

Si fa riferimento a due misure che analizzano la forma di una distribuzione:

L'asimmetria, che si verifica quando non è possibile individuare un asse verticale che divida la distribuzione in due parti specularmente uguali.

La curtosi, ovvero l'appiattimento di una distribuzione, che coincide con un peso più o meno accentuato delle sue code.

Asimmetria

La nozione di asimmetria ha senso se un carattere è almeno ordinabile.

Si misura confrontando 3 indici di posizione: media, moda e mediana.

Simmetria -> Moda = Mediana = Media

Asimmetria positiva -> Moda < Mediana < Media

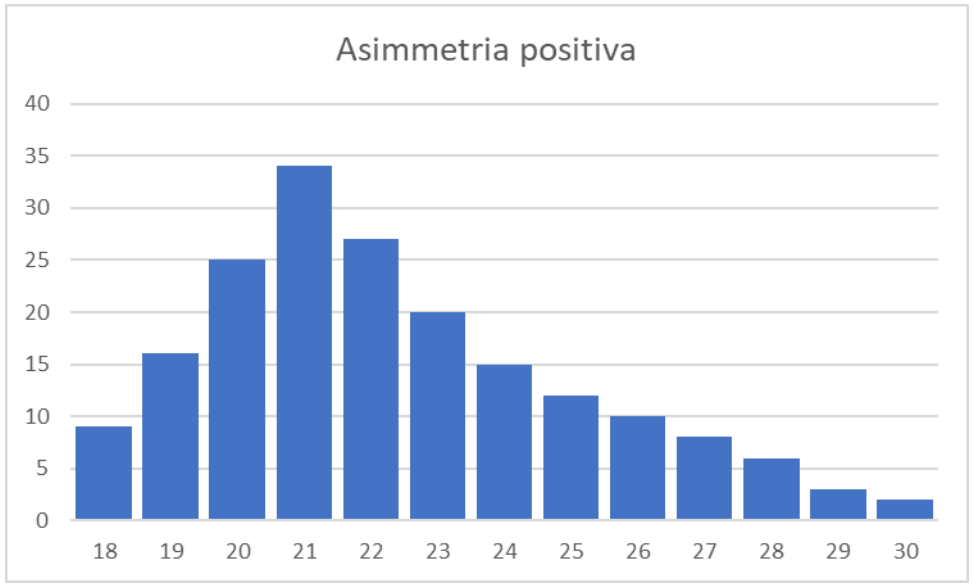
Asimmetria negativa -> Media < Mediana < Moda

Asimmetria

Vediamo due distribuzioni asimmetriche:

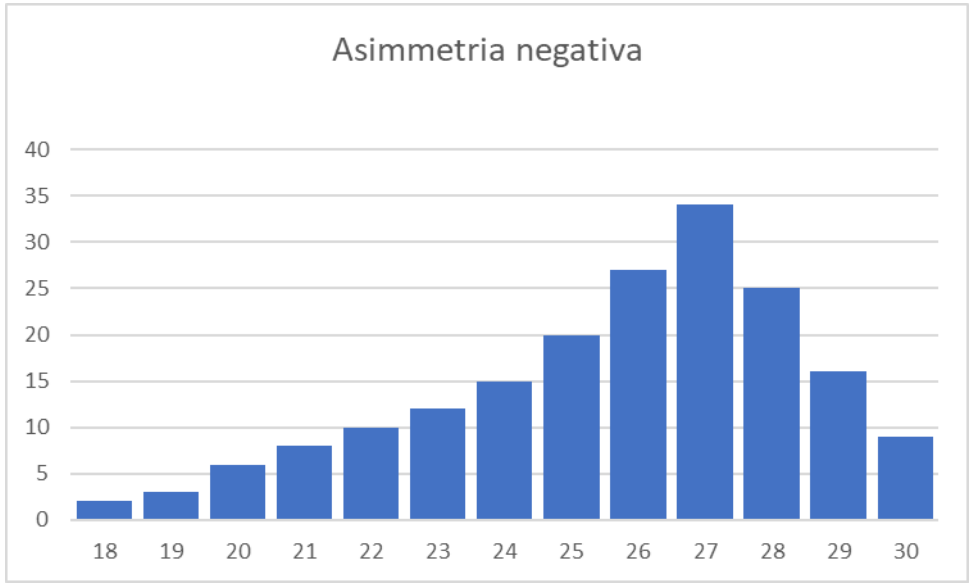
Media	22,4
Mediana	22
Moda	21

Moda < Mediana < Media



Media	25,6
Mediana	26
Moda	27

Media < Mediana < Moda



Indice di asimmetria

Si basa sul momento centrato della media aritmetica.

$$M_3 = \frac{\sum_{i=1}^{N}(x_i - m_x)^3}{N}$$

Elevando al cubo gli scarti dalla media, l'indice di asimmetria M_3 può assumere valori sia negativi che positivi.

Se $M_3 > 0$ si ha asimmetria positiva.

Se $M_3 < 0$ si ha asimmetria negativa.

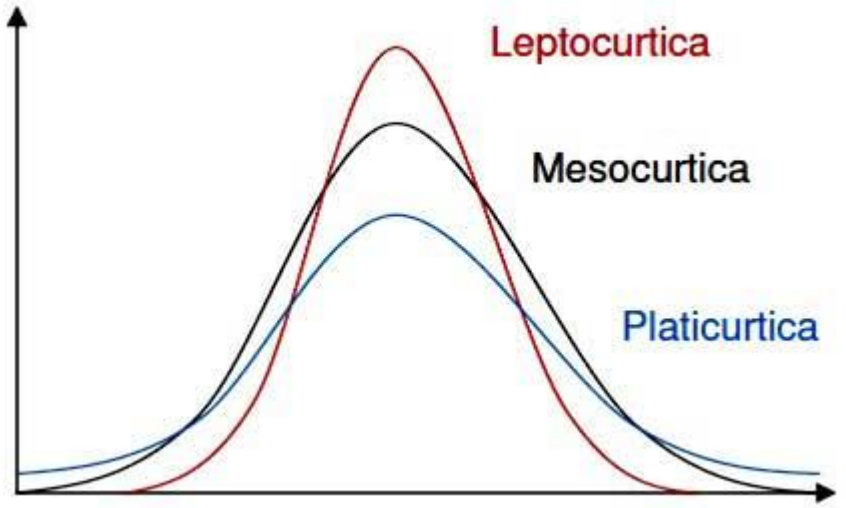
Appiattimento

Con il termine **appiattimento** si intende la forma più o meno appuntita di una distribuzione di dati, rispetto alla distribuzione normale. Di conseguenza esso indica il maggiore o minore peso dei valori posti agli estremi della distribuzione (code), rispetto a quelli della parte centrale.

Indice di curtosi di Pearson

Il rapporto tra la parte centrale della distribuzione e le code si definisce curtosi. Si può calcolare un indice con la media delle potenze quarte della variabile standardizzata.

$$B^2 = \frac{\sum_{i=1}^{N}((x_i - m_x)/\sigma)^4}{N}$$



Questo indice vale 3 quando la distribuzione ha due flessi equidistanti dal valore centrale (mesocurtica), è maggiore di 3 per distribuzioni più appuntite e minore di 3 per distribuzioni piuttosto piatte.

Concentrazione di una variabile trasferibile

Esiste una misura di disuguaglianza costruita per **variabili quantitative con proprietà trasferibili**.

Cosa significa proprietà trasferibili? Significa che sono cedibili ad altre unità.

Ad esempio sono **proprietà trasferibili** il reddito, i consumi, il numero di beni posseduti, ecc. che possono essere trasferiti da una persona all'altra.

Non sono proprietà trasferibili, invece, l'età, l'altezza, ecc.

Poter valutare come queste proprietà siano concentrate in poche unità dà una misura del livello di disuguaglianza.

Un caso classico riguarda la distribuzione del reddito.

Il rapporto di concentrazione di Gini

È il valore che misura tipicamente la disuguaglianza di proprietà trasferibili.

L'indice di concentrazione non valuta la ricchezza nel suo complesso, ma come questa si distribuisce all'interno di una popolazione, di un paese, ...

Il ragionamento che sta alla base di questo indice è molto semplice e parte dalla differenza tra come sarebbe distribuito il reddito in caso di equidistribuzione e la situazione concreta.

Ad esempio in caso di equidistribuzione il 5% di una popolazione dovrebbe detenere il 5% delle ricchezze, ma sappiamo bene che non è così.

Il rapporto di concentrazione di Gini

L'indice di Gini mette a confronto la proporzione dei casi sul numero di frequenze totale (p) e la proporzione della quantità posseduta sul totale (q).

$$R = 1 - \frac{\sum_{i=0}^{N-1} q_i}{\sum_{i=0}^{N-1} p_i}$$

è pari a **0** in presenza di **equidistribuzione** del reddito, cioè tutte le persone hanno la stessa ricchezza.

è pari a **1** in presenza di **massima concentrazione** del reddito, cioè solo una persona detiene tutto la ricchezza.

Cerchiamo di spiegare questa formula attraverso una rappresentazione grafica.

Ricordiamoci che dobbiamo iniziare dall'ordinare i dati dal valore più basso a quello più alto.

Nel caso del reddito, ordineremo dal reddito più basso a quello più elevato.

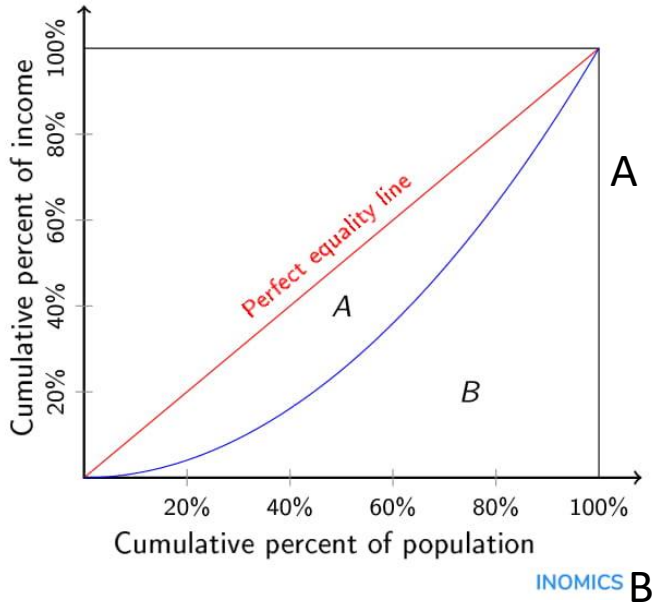
La curva di Lorenz

In caso di equidistribuzione le coppie di punti (p_i, q_i) si andrebbero ad allineare lungo la bisettrice OA (retta di equidistribuzione).

Se tutta la ricchezza fosse concentrata in una sola unità , la distribuzione sarebbe sempre pari a 0, ad eccezione dell'ultimo caso pari ad A (linea OBA).

Nei casi intermedi si darà luogo ad una spezzata sotto al segmento di equidistribuzione.

L'area compresa tra la spezzata di concentrazione e il segmento di equidistribuzione si chiama area di concentrazione.



A

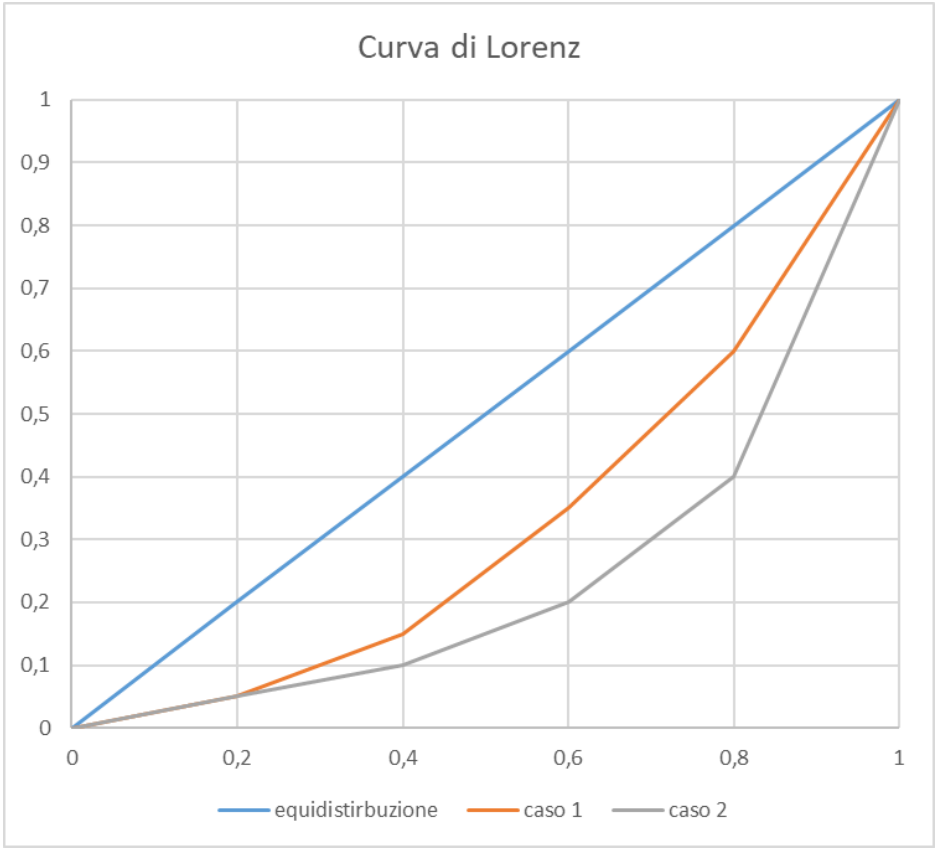
B

O

Facciamo un esempio

Unità	Equidistribuzione	Caso 1	Caso 2
1	0,20	0,05	0,05
2	0,20	0,10	0,05
3	0,20	0,20	0,10
4	0,20	0,25	0,20
5	0,20	0,40	0,60
Totale	1,00	1,00	1,00

Unità	Equidistribuzione p cumulata	Caso 1 - q cumulata	Caso 2 - q cumulata
1	0,20	0,05	0,05
2	0,40	0,15	0,10
3	0,60	0,35	0,20
4	0,80	0,60	0,40
5	1,00	1,00	1,00



Facciamo un esempio

R1 = 1 - (0,05+0,15+0,35+0,60)/(0,2+0,4+0,6+0,8)

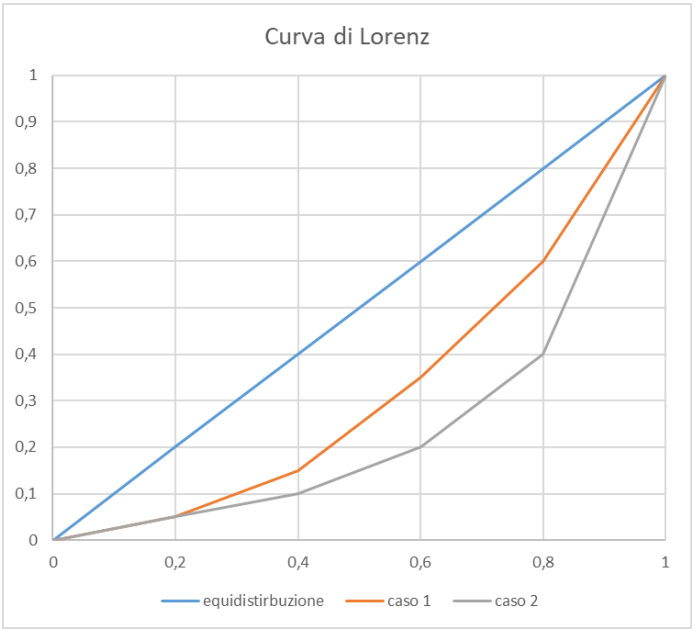
R1 = 1 – (1,15/2) = 1-0,575 = 0,425 (minore concentrazione)

R2 = 1 – (0,05+0,10+0,20+0,40)/(0,2+0,4+0,6+0,8)

R2 = 1 – (0,75/2) = 1-0,375 = 0,625 (maggiore concentrazione)

$$R = 1 - \frac{\sum_{i=0}^{N-1} q_i}{\sum_{i=0}^{N-1} p_i}$$

Unità	Equidistribuzione p cumulata	Caso 1 - q cumulata	Caso 2 - q cumulata
1	0,20	0,05	0,05
2	0,40	0,15	0,10
3	0,60	0,35	0,20
4	0,80	0,60	0,40
5	1,00	1,00	1,00



La standardizzazione

Nell'ambito della ricerca sociale spesso si necessita di confrontare i valori di due distribuzioni che non hanno la stessa unità di misura (pensiamo ad esempio al voto dell'esame di maturità che una volta era in sessantesimi e ora in centesimi)

È necessario riportare i differenti voti sulle stesse unità di misura.

Si applica la cosiddetta **standardizzazione**, ovvero trasformiamo i valori sottraendo da ogni valore la media e dividendo il risultato per la deviazione standard.

$$z_i = \frac{x_i - m_x}{\sigma}$$

La nuova variabile Z avrà media 0 e varianza 1, media, moda e mediana saranno coincidenti.

Confronti basati sui rapporti

Un modo semplice per confrontare i dati di una distribuzione consiste nei cosiddetti **rapporti statistici**.

Ad esempio per calcolare la frequenza relativa abbiamo messo a rapporto la frequenza di una modalità al totale della distribuzione.

Distinguiamo 4 tipologie di rapporti:

- **rapporto di composizione**,
- **rapporto di coesistenza**,
- **rapporto di derivazione**,
- **rapporto di densità**.

Rapporto di composizione

Si parla di **rapporto di composizione** o di **rapporto parte al tutto** quando dividiamo il valore di una frequenza di una modalità con il totale.

Ad esempio l'indice di vecchiaia si calcola dividendo il numero delle persone con 65 anni e oltre per il totale della popolazione residente.

Rapporto di coesistenza

Si parla di **rapporto di coesistenza** quando dividiamo il valore di una frequenza (o di una quantità) di una modalità per quello di un'altra modalità.

Ad esempio l'indice di dipendenza si calcola dividendo il numero delle persone con 65 anni e oltre e di persone con meno di 15 anni per il totale della popolazione attiva (tra 15 e 64 anni).

Rapporto di derivazione

Si parla di **rapporto di derivazione** quando dividiamo il valore di una frequenza (o di una quantità) di un fenomeno e quella di un altro che può essere considerato il suo presupposto logico.

Ad esempio posso mettere a rapporto il numero dei laureati e il numero degli iscritti di una facoltà.

Un altro rapporto di derivazione è il tasso di fecondità di una data popolazione in un determinato periodo (di solito, un anno). È il valore relativo di nati in quel periodo rispetto al numero delle donne in età feconda (fra i 15 e i 49 anni, secondo il parere di alcuni demografi, 15-44 anni secondo quello di altri).

Rapporto di densità

Si parla di **rapporto di densità** quando rapportiamo la dimensione globale di un fenomeno alla dimensione spaziale, temporale o caratterizzante cui esso fa riferimento.

Ad esempio la densità abitativa, ovvero il numero di abitanti e l'estensione del territorio in kmq, l'indice di affollamento di un'abitazione, ovvero il rapporto tra il numero di abitanti e il numero di stanze.

Coefficienti basati sulle differenze

Nella ricerca sociale ci troviamo frequentemente a mettere a confronto dei fenomeni in spazi o tempi differenti.

Pensiamo ad esempio quando confrontiamo la percentuale di laureati all'interno di una popolazione in tempi differenti, o quando analizziamo lo stesso fenomeno nello stesso tempo, ma su territori differenti.

Si parla in questo caso di:

- **Serie temporali**, la sequenza di valori che assume un determinato fenomeno nella stessa popolazione di riferimento in tempi diversi (ad es. % di disoccupati in Italia dal 2010 al 2021)
- **Serie territoriali**, la sequenza dei valori assunti da un determinato fenomeno nello stesso tempo, ma in aggregati territoriali diversi (ad es. % di disoccupati in Italia nel 2021 per regione o provincia)

Coefficienti basati sulle differenze

Se vogliamo studiare le variazioni di un fenomeno rilevato in situazioni temporali e/o territoriali differenti possiamo utilizzare le **variazioni assolute** o le **variazioni relative**.

Variazione assoluta = b – a

Variazione relativa = (b – a)/a

Variazione relativa percentuale = (b – a)/a * 100

Coefficienti basati sulle differenze

Dipendenti a tempo determinato/indeterminato (in migliaia)

	2° trim. 2021	2° trim. 2022
Tempo indeterminato	535	637
Tempo determinato	346	412

Fonte: Istat, Rilevazione continua sulle forze di lavoro

Variazione assoluta = b – a

Tempo indeterminato = 637 – 535 = 102

Tempo determinato = 412 – 346 = 66

Coefficienti basati sulle differenze

Dipendenti a tempo determinato/indeterminato (in migliaia)

	2° trim. 2021	2° trim. 2022
Tempo indeterminato	535	637
Tempo determinato	346	412

Fonte: Istat, Rilevazione continua sulle forze di lavoro

Variazione relativa = (b – a)/a

Tempo indeterminato = 102/535 =0,19

Tempo determinato = 66/346 = 0,19

Coefficienti basati sulle differenze

Dipendenti a tempo determinato/indeterminato (in migliaia)

	2° trim. 2021	2° trim. 2022
Tempo indeterminato	535	637
Tempo determinato	346	412

Fonte: Istat, Rilevazione continua sulle forze di lavoro

Variazione relativa percentuale = (b – a)/a * 100

Tempo indeterminato = 102/535 * 100 = 19,1%

Tempo determinato = 66/346 * 100 = 19,1%

I numeri indice

Per mettere a confronto l'andamento complessivo di una serie territoriale o temporale si ricorre ai numeri indice.

I numeri indice permettono di studiare l'intensità di un cambiamento di un fenomeno, nel tempo o nello spazio, facendo riferimento a un contesto chiamato base del numero indice, che solitamente assume valore 100.

In questo caso si parla di **numero indice a base fissa**.

I numeri indice

Guardiamo questa tabella, in particolare prendiamo i dati degli occupati dal 2° trim. 2019

	T2-2019	T3-2019	T4-2019	T1-2020	T2-2020	T3-2020	T4-2020	T1-2021	T2-2021	T3-2021	T4-2021	T1-2022	T2-2022
Condizione professionale europea													
forze lavoro	25 807	25 535	25 657	25 134	23 967	24 881	24 762	24 421	24 963	25 095	25 204	24 911	25 258
occupati	23 307	23 225	23 129	22 759	22 093	22 336	22 353	21 832	22 576	22 884	22 924	22 737	23 253
disoccupati	2 500	2 309	2 528	2 375	1 874	2 545	2 410	2 589	2 388	2 211	2 280	2 174	2 006
totale inattivi	25 755	25 975	25 846	26 376	27 549	26 585	26 641	26 961	26 357	26 216	26 006	26 305	25 917
forze lavoro potenziali	2 814	3 070	2 782	3 025	3 849	3 209	3 185	3 819	3 134	2 978	2 710	2 665	2 381
non cercano e non disponibili	22 941	22 904	23 063	23 350	23 701	23 376	23 456	23 142	23 223	23 238	23 296	23 640	23 537
totale	51 562	51 509	51 503	51 510	51 517	51 465	51 403	51 382	51 320	51 311	51 210	51 216	51 176

I numeri indice

Costruiamo i nostri numeri indice a base fissa dividendo i valori per il dato di riferimento, ovvero il 2° trimestre 2019.

	T2-2019	T3-2019	T4-2019	T1-2020	T2-2020	T3-2020	T4-2020	T1-2021	T2-2021	T3-2021	T4-2021	T1-2022	T2-2022
occupati	23.307	23.225	23.129	22.759	22.093	22.336	22.353	21.832	22.576	22.884	22.924	22.737	23.253
Numero indice	100,0	99,6	99,2	97,6	94,8	95,8	95,9	93,7	96,9	98,2	98,4	97,6	99,8

I numeri indice

Se invece rapportiamo ogni intensità alla precedente otteniamo un insieme di **numeri indice a base mobile**.

Il numero indice a base mobile informa sulle variazioni di un fenomeno da un periodo a quello successivo.

	T2-2019	T3-2019	T4-2019	T1-2020	T2-2020	T3-2020	T4-2020	T1-2021	T2-2021	T3-2021	T4-2021	T1-2022	T2-2022
occupati	23.307	23.225	23.129	22.759	22.093	22.336	22.353	21.832	22.576	22.884	22.924	22.737	23.253
Numero indice a base fissa	100,0	99,6	99,2	97,6	94,8	95,8	95,9	93,7	96,9	98,2	98,4	97,6	99,8
Numero indice a base mobile		99,6	99,6	98,4	97,1	101,1	100,1	97,7	103,4	101,4	100,2	99,2	102,3

Per chi è curioso

https://www.istat.it/it/benessere-e-sostenibilità

Sia per il Benessere equo e sostenibile che per i Sustainable Development Goals, potete trovare diversi indicatori.

Riuscite a trovare un esempio per le 4 tipologie di rapporti di cui abbiamo parlato nel corso della lezione?

- **rapporto di composizione**,
- **rapporto di coesistenza**,
- **rapporto di derivazione**,
- **rapporto di densità**.