FONDAMENTI E METODI PER L'ANALISI EMPIRICA NELLE SCIENZE SOCIALI

Università degli Studi di Macerata
Dott. Mattia Tassinari

Elaborazione e analisi statistica dei dati (Cap. 9)

Il percorso in sintesi…

- La sociologia come scienza empirica
- Domanda di ricerca e revisione della letteratura
- Strategie/tipi di ricerca e fasi della ricerca
- Raccolta dei dati: campionamento e raccolta dati primari (questionario, intervista, focus group)
- Elaborazione ed analisi statistica dei dati
- Indicatori compositi (applicazioni)

La scienza è fatti di dati come una casa di pietre. Ma un ammasso di dati non è scienza più di quanto un mucchio di pietre sia una casa.

[Jules Henri Poincaré]

=> Necessità di 'sistemare' (elaborare ed analizzare) i dati al fine di renderli utili a rappresentare e sintetizzare un fenomeno di interesse per trarne indicazioni.

Statistica Descrittiva

- Si occupa della trattazione dei dati rilevati (popolazione o campione) su fenomeni misurabili allo scopo di **rappresentare** e **sintetizzare** i fenomeni di interesse

- Insieme di tecniche usate per descrivere le **caratteristiche di base dei dati** raccolti in un esperimento/studio

STATISTICA DESCRITTIVA

La **DISTRIBUZIONE** di una variabile Xi (es. voto esame dell'i-esimo studente) è il modo in cui le *modalità* (cioè i *valori,* per le variabili quantitative) si distribuiscono tra le unità statistiche (es., x1=18; x2=18; x3=24; …).

Xi = x1; x2 …… xn

Frequenza: è il numero delle unità statistiche su cui si presenta una particolare modalità (o valore)

=> distribuzione delle frequenze: esplicita quante volte (con quale frequenza) una determinata modalità/valore si presenta nel campione/popolazione

ANALISI UNIVARIATA e MULTIVARIATA

Le analisi **univariate** considerano una sola variabile rilevata sulle unità.

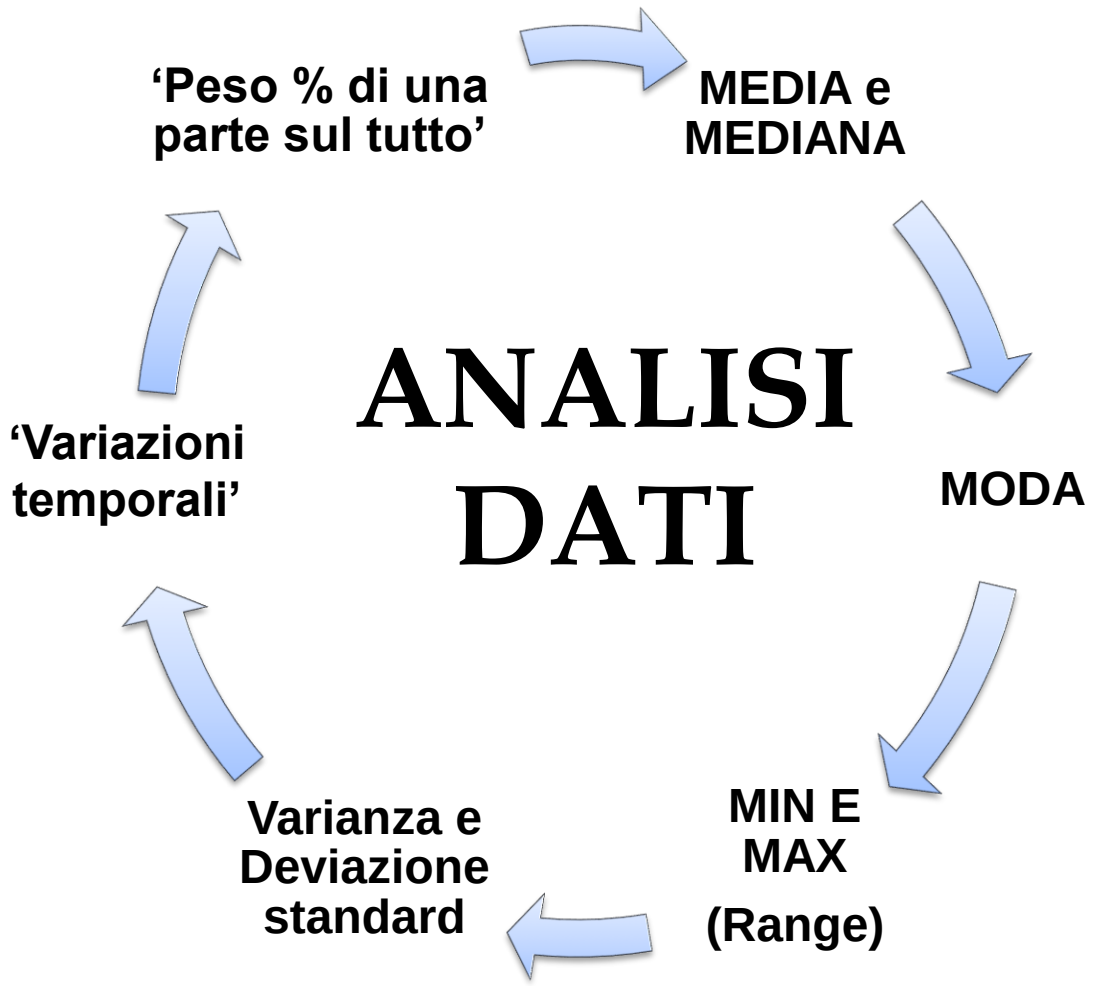
Nello studio congiunto di **due variabili** (analisi della relazione tra due variabili) si parla di analisi **bivariata**.

Lo studio congiunto di due o più variabili è detto analisi **multivariata** (ovviamente il multivariato include il bivariato).

Indici statistici di sintesi

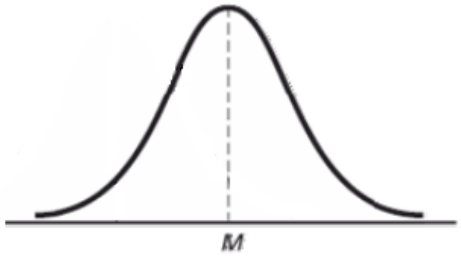
Per trarre delle indicazioni adeguate quando si considerano dati quantitativi, occorre che le caratteristiche principali delle osservazioni siano sintetizzate con opportune misure, dette **Indici statistici di sintesi**, e che tali misure siano adeguatamente analizzate e interpretate.

I principali Indici statistici ed elaborazioni in questa lezione …



Tipi di Indici: Misure di Tendenza Centrale

Nella maggior parte degli insiemi di dati, le osservazioni mostrano una tendenza a **raggrupparsi attorno a un valore centrale** (es. distribuzione *Normale*).



Risulta in genere quindi possibile selezionare un **valore tipico** per descrivere un intero insieme di dati.

Tale valore descrittivo è una **misura di posizione** o di **tendenza centrale**.

Tipi di **misure di tendenza centrale**:

- **Media**
- **Mediana**
- **Moda**

MEDIA

E' la misura di posizione più comune. Si calcola dividendo la somma dei valori osservati per il numero totale di osservazioni.

Es. Voto degli esami di uno studente:

18; 20; 20; 22; 25; 28; 30
MEDIA: 23,29

= MEDIA

MEDIA

Proprio perché il calcolo della media si basa su tutte le osservazioni, tale misura di posizione risulta influenzata da **valori estremi (outlier)**.

In presenza di valori estremi, la media aritmetica fornisce una **rappresentazione distorta** dei dati ed è pertanto opportuno in questi casi ricorrere ad altre misure di posizione.

LA MEDIANA

La **mediana** è il valore centrale in un insieme di dati ordinati dal valore più piccolo al più grande (cioè in ordine crescente).

La mediana

La mediana è l'osservazione che, nella serie ordinata dei dati, si lascia alla destra il 50% delle osservazioni e a sinistra il 50% delle osservazioni. Quindi, il 50% delle osservazioni risulteranno maggiori della mediana e il 50% risulteranno minori della mediana.

$$\text{Mediana} = \text{osservazione di posto } \frac{n+1}{2} \text{ nella serie ordinata} \qquad (3.2)$$

Voti esami: **18, 20, 20, 22, 25, 28, 30**

MEDIANA: 22

LA MEDIANA

Per trovare la posizione occupata dal valore mediano nella serie ordinata delle osservazioni si usa l'equazione secondo una delle due regole seguenti:

***REGOLA 1*.** Se l'ampiezza del campione è un numero **dispari**, la mediana coincide con il valore centrale, vale a dire con l'osservazione che occupa la posizione (N+1)/2 nella serie ordinata delle osservazioni.

REGOLA 2. Se l'ampiezza del campione è un numero **pari**, la mediana allora coincide con la media dei valori corrispondenti alle due osservazioni centrali.

La mediana non è influenzata dalle osservazioni estreme di un insieme di dati: nel caso di osservazioni estreme è quindi opportuno descrivere l'insieme di dati con la mediana piuttosto che con la media.

=MEDIANA

LA MODA

E' il valore più frequente in una distruzione

A differenza della media, la moda non è influenzata dagli outlier.

Tuttavia tale misura di posizione viene usata solo per scopi descrittivi, poiché è caratterizzata da maggiore variabilità rispetto alle altre misure di posizione (piccole variazioni in un insieme di dati possono far variare in modo consistente la moda).

Voti esami: **18, 20, 20, 22, 25, 28, 30**

MODA: 20

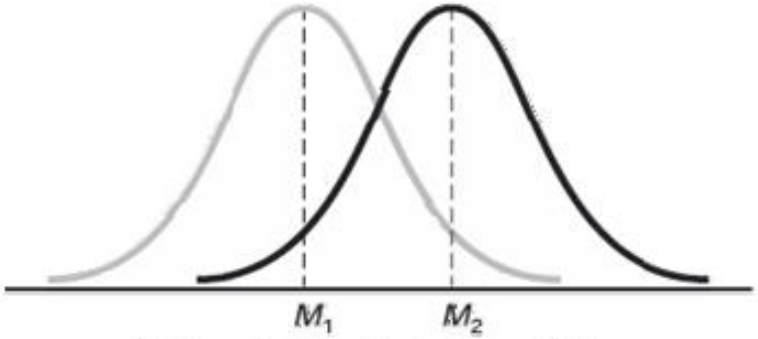
NOTA: un insieme di dati può non avere moda, se nessuno valore è "più tipico" degli altri.

=MODA

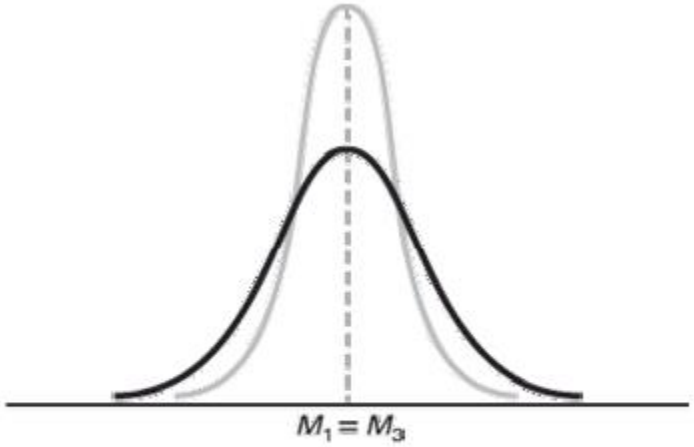
Misure di Dispersione (o di Variabilità)

Una seconda caratteristica importante di un insieme di dati è la **variabilità**. **La variabilità è la quantità di dispersione presente nei dati**.

Due insiemi di dati possono differire sia nella posizione che nella *variabilità*; oppure possono essere caratterizzati dalla stessa variabilità, ma da diversa misura di posizione; o ancora, possono essere dotati della stessa misura di posizione, ma differire notevolmente in termini di variabilità.



(*a*) Due distribuzioni simmetriche a forma campanulare che differiscono solo nella posizione



(*b*) Due distribuzioni simmetriche a forma campanulare che differiscono solo nella variabilità

Misure di dispersione: il Range

Il **range** (o **intervallo di variazione**) è la differenza tra il **valore** massimo e quello minimo in un insieme di dati (più l'**intervallo** è un numero alto, più i **valori** della serie sono lontani tra loro).

NOTA: un limite del range consiste nel fatto che non tiene conto di come i dati si distribuiscono effettivamente tra il valore più piccolo e quello più grande.

Per questo motivo, in presenza di osservazioni estreme, risulta una misura inadeguata della variabilità.

VOTI ESAMI:
18, 20, 20, 22, 25, 28, 30

MAX: 30
MIN: 18
Range = 12

=MAX
=MIN

Misure di dispersione: Varianza e Scarto Quadratico Medio (o Deviazione Standard)

Varianza e la sua radice quadrata, lo **scarto quadratico medio**, sintetizzano come le osservazioni si distribuiscono o si concentrano attorno alla loro media.

La varianza campionaria

La varianza campionaria è la somma dei quadrati delle differenze dalla media divisa per $(n - 1)$:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n - 1} \qquad (3.9)$$

dove

$\overline{X}$ = media aritmetica campionaria
n = ampiezza del campione
X_i = i–esima osservazione della variabile casuale X
$\sum_{i=1}^{n}(X_i - \overline{X})^2$ = somma dei quadrati delle differenze tra i valori X_i e $\overline{X}$

= VAR

Varianza: esempi

1° CASO: 18, 15, 20, 20, 30, 14, 12, 17

MEDIA = 18.25

Varianza = $(18-18.25)^2 + (15-18.25)^2 + (20-18.25)^2 + (20-18.25)^2 + (30-18.25)^2 + (14-18.25)^2 + (12-18.25)^2 + (17-18.25)^2 /7 =$ 30,5

2° CASO: 9, 10, 11, 10, 10, 13, 17, 30

MEDIA = 13.75; Varianza = 49,64

3° CASO: 31, 10, 15, 20, 24, 23, 12, 30

MEDIA = 20.63; Varianza = 61,70

Misure di dispersione: Scarto Quadratico Medio o Deviazione Standard (σ)

Lo scarto quadratico medio (o deviazione standard)

Lo scarto quadratico medio campionario (detto anche deviazione standard) è la radice quadrata della varianza campionaria:

$$S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}} \qquad (3.10)$$

NOTA: Una bassa Deviazione Standard indica che i valori sono vicini alla media. Una alta Deviazione Standard indica che i valori sono lontani dalla media.

= DEV.ST.POP

Deviazione Standard (σ): esempi

1° CASO: 18, 15, 20, 20, 30, 14, 12, 17

MEDIA = 18.25

σ = radice quadrata di $(18-18.25)^2 + (15-18.25)^2 + (20-18.25)^2 + (20-18.25)^2 + (30-18.25)^2 + (14-18.25)^2 + (12-18.25)^2 + (17-18.25)^2 / 7 =$ 5,2

2° CASO: 9, 10, 11, 10, 10, 13, 17, 30

MEDIA = 13.75; σ = 6.6

3° CASO: 31, 10, 15, 20, 24, 23, 12, 30

MEDIA = 20.63; σ = 7.3

NOTA: se le osservazioni sono tutte eguali (in modo che non vi è variabilità dei dati) il **Range**, la **Varianza** e lo **Scarto Quadratico Medio** sono tutti uguali a zero.

Varianza e Scarto Quadratico Medio: unità di misura

- La varianza possiede alcune importanti proprietà matematiche; tuttavia, la sua unità di misura coincide con il quadrato dell'unità di misura dei dati (es. euro al quadrato, metri al quadrato e così via).
- Mentre lo scarto quadratico medio è espresso nell'unità di misura originaria dei dati (es. euro o metri).

Misure di dispersione: il Coefficiente di variazione

A differenza delle altre misure di variabilità, il coefficiente di variazione è una misura relativa, espressa come una percentuale e non nell'unità di misura dei dati.

Il **coefficiente di variazione**, indicato con CV, misura la dispersione nell'insieme di dati relativamente alla media.

Viene utilizzato quando le distribuzioni hanno unità di misura differenti.

Il coefficiente di variazione

Il coefficiente di variazione è uguale allo scarto quadratico medio diviso per la media aritmetica, moltiplicato per 100%.

$$CV = \left(\frac{S}{|\bar{X}|}\right)100\% \qquad (3.11)$$

dove

S = scarto quadratico medio

$\bar{X}$ = valore assoluto della media aritmetica nell'insieme dei dati

= **Coefficiente di variazione** (Excel non fornisce una formula automatica, occorre scrivere la formula manualmente)

COEFFICIENTE DI VARIAZIONE

MEDIA: 18.25

σ = 5,2

CV = 28%

MEDIA 13.75

σ = 6.6

CV = 48%

MEDIA : 16.85

σ = 7.3

CV = 43%

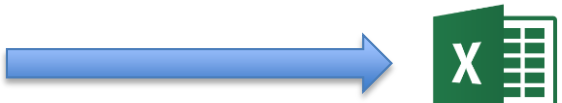
LE VARIAZIONI TEMPORALI (Tasso di crescita)

$$\frac{V2 - V1}{V1}$$

Valore assunto nel periodo finale meno il valore assunto nel periodo iniziale diviso il valore iniziale. L'indicatore si può esprimere in percentuale moltiplicando la frazione per 100.

UNA PARTE SUL TUTTO

- Definire il peso di un caso sull'intera distribuzione
- **Formula**: [(valore della singola parte/ valore totale della distribuzione)]* 100
- Tipica rappresentazione: Grafico a torta

Vedi esempio 

SULLE TESINE …

Iniziare ad elaborare e analizzare in Excel i dati raccolti per la tesina: **calcolare media, moda, min e max, deviazione standard, coefficiente di variazione, variazioni temporali, …** (la prossima lezione sarà dedicata in specifico alla costruzione di **indicatori compositi**).

RECAP: come strutturare la tesina (circa 5,000 parole)

1) Introduzione: presentare il **fenomeno di interesse** e la **domanda di ricerca** => presentare quale problematica si propone di affrontare la ricerca, evidenziando l'importanza del fenomeno, la necessità di studiarlo e il contributo originale offerto (perché il lettore dovrebbe essere interessato alla vostra ricerca rispetto ad altre presenti in letteratura).

2) Analisi della letteratura esistente: illustrare cosa è stato detto in precedenza sul fenomeno di interesse e con quale metodologia è stato studiato, descrivere quali sono le principali teorie di riferimento che lo interpretano, e indicare in quale prospettiva teorica si pone la ricerca e perché (quale teoria offre i criteri interpretativi migliori per rispondere alla domanda di ricerca), **citando in modo appropriato le fonti nel testo e in bibliografia**.

3) Materiali e metodi: descrivere la strategia (o tipo) di ricerca (es., descrittiva, esplicativa, valutativa; ricerca basata su esperimento, quantitativa/standard, qualitativa/non standard); specificare la fonte dei dati (es. indagine campionaria o sulla popolazione), la loro tipologia (qualitativa o quantitativa), gli strumenti di rilevazione (es. dati secondari tratti da …., questionario, intervista, focus group); descrivere la **matrice dei dati**, le **variabili selezionate** e perché le si è scelte; descrivere la metodologia di analisi dei dati (es., tecniche statistiche, econometriche, indicatori composti, ecc.).

4) Risultati: presentare le evidenze quantitative o qualitative che emergono dall'analisi dei dati empirici (es., **media, mediana, moda, min e max, deviazione standard, coefficiente di variazione, variazioni temporali, correlazioni, classifiche indicatori composti**), includendo **tabelle** e **grafici di sintesi**.

5) Commenti ai risultati/Discussione: **commentare i risultati ottenuti**, evidenziare la '**nuova conoscenza**' ottenuta con riferimento alla domanda di ricerca e a come tali risultati siano allineati o meno con la letteratura sul tema (es., nuovo modo di descrivere un dato fenomeno; conferma o meno a livello empirico delle ipotesi di partenza; …).

6) Conclusioni: richiamo del percorso di ricerca fatto e dei principali risultati; indicazioni normative sulle azioni da compiere in vista della soluzione delle problematiche sociali connesse al fenomeno studiato; limiti della ricerca e possibili sviluppi della linea di ricerca in futuro).

7) La bibliografia.