POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

Corso di Statistica Sociale

CORSO DI LAUREA: SCIENZE DELL'EDUCAZIONE

DOCENTE: FRANCESCO SANTELLI

Prima di andare avanti…

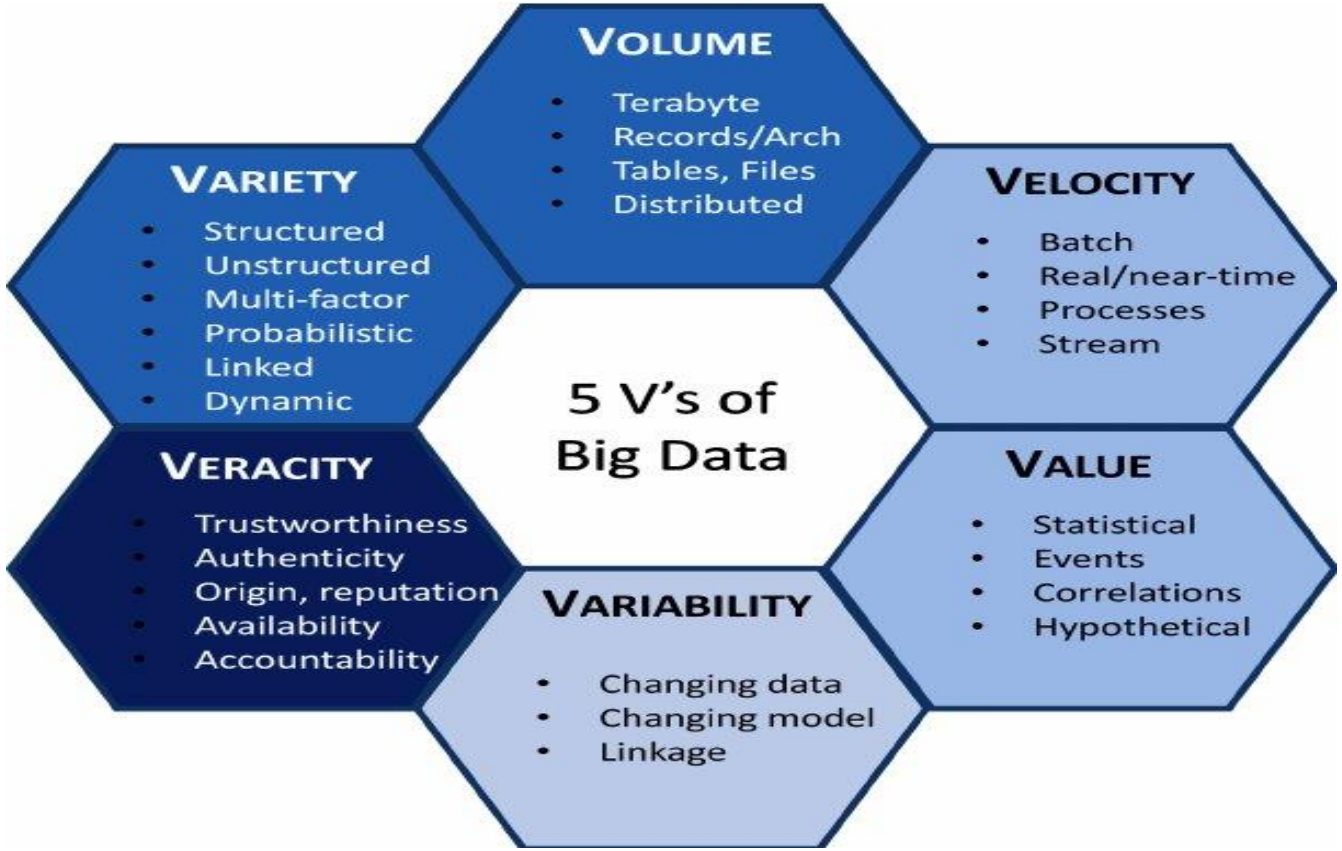
- Sulla prima lezione: dubbi? Perplessità? Curiosità?

- Avete avuto difficoltà nel creare la tabella di frequenza?

- Avete preso dimestichezza con il tremendo simbolo della sommatoria?

- Siete riusciti a scrivere in formula le «colonne» della tabella di frequenza?

Il nuovo paradigma statistico negli ultimi anni: i *Big Data*



Dati provenienti principalmente da **Social Media** e/o dispositivi tecnologici (**Internet of Things, IoT**)

Paradigma di ricerca diverso rispetto a ciò che abbiamo detto nella 1 lezione (popolazione, ricerca campionaria, rappresentatività ecc.)

Se cambiano tutte queste premessa, cambia il concetto stesso di statistica!

Data: ma quali? Diversi tipi di dati…

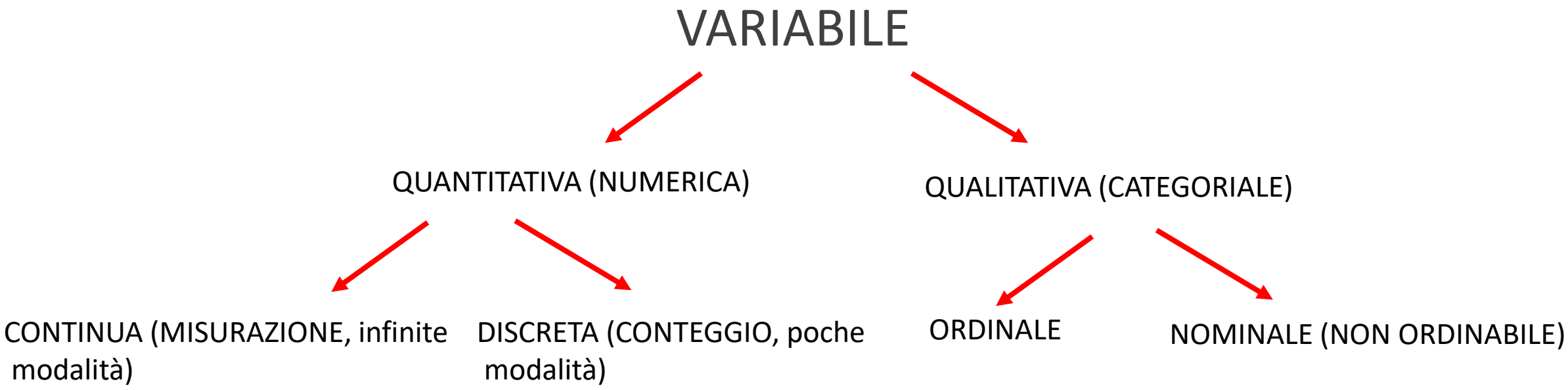
- Quando pensiamo al dato pensiamo al semplice numero rilevato per ogni individuo…ma non è sempre così!

- Ci sono dati **complessi:** dati ad intervallo, dati relazionali, dati simbolici ecc.

- A volte i dati sono **gerarchici**: studenti raggruppati in classi, raggruppate in scuole, raggruppate in città ecc.

- Noi ci limiteremo al **dato semplice** o **elementare**, ma anch'esso non è sempre un numero!

Differenti tipi di **variabili**



Esistono casi particolari di variabili che possono assumere solo due modalità (sano o malato, promosso o bocciato, italiano o straniero ecc.). Esse prendono il nome di **dicotomiche** o **binarie**

Un po’ di esempi…

- Altezza o peso: numerica continua (**misurazioni fisiche**)

- Reddito o risparmi: numerica continua (**economia**)

- Numero di volte che si è provato statistica prima di superarlo: numerica discreta (**conteggio o poche modalità**)

- Titolo di studio: qualitativa ordinale (**gerarchia**)

- Nazionalità: qualitativa nominale (**caratteristiche**)

Casi ambigui...

- Voto all'esame da 18 a 30 (Misurazione? Numerica? Conteggio? Poche o tante modalità?)

- Fasce di reddito (era numerica, ma ora?)

- Quante volte all'anno ti rechi al cinema (conteggio, in teoria discreta, ma ben 366 modalità!)

- Colore occhi/capelli (possiamo ipotizzare una classifica dal chiaro allo scuro ad esempio? E' ordinabile?)

La **natura** della variabile dipende quindi da *come è stata generata* ma anche da *come si presenta nei dati* e dall'*obiettivo* che ci si pone nell'analizzarli.

Sintetizzare i dati: *indici di tendenza centrale*

Con la tabella di frequenza abbiamo scritto in maniera intelligente i dati, ma non li abbiamo ancora sintetizzati...

- Perché sintetizzare? Io non voglio leggermi 1500 numeri nei quali mi perdo, ne voglio leggere **tre o quattro** che mi dicano tanto di tutti e 1500

- Come sintetizzare? Con quale criterio? Con quali calcoli?

- Istintivamente, noi cerchiamo come primo step un ***valore centrale*****.**

- I 3 principali modi (indici) di esprimere la tendenza centrale nei dati sono:

a) media b) mediana c) moda

Essi indicano in modo diverso attorno a quale valore sono concentrati i dati

La media: l'indice più diffuso

- Intuitivamente tutti noi sappiamo già cos'è: età media in una classe, reddito medio tra dei lavoratori, numero medio di articoli scientifici prodotti dai dipartimenti ecc.

E' infatti tra gli indici più utilizzati.

- 1) **Si può utilizzare solo per variabili numeriche (continue o discrete)**

Provate a fare la media del titolo di studio o della nazionalità…

- 2) **Tiene conto di tutte le osservazioni**
- 3) La sua formula è

$$\frac{1}{N}\sum_{i=1}^{n} x_i$$

In altre parole, prima sommiamo tutti gli elementi e poi dividiamo per quanti ne sono

Calcolo della media: esempio (1)

- Poniamo il caso di avere i dati delle regioni italiane relative alle prove Invalsi di Italiano delle regioni settentrionali:

V Aosta	201
Piemonte	203
Liguria	200
Lombardia	200
Bolzano	196
Trento	200
Veneto	200
Friuli-Ven.G.	202
Emilia-Romagna	198

Calcoliamo media Nord Italia: sommiamo valori e dividiamo per il numero di regioni

201 + 203 + 200…=1800 → 1800/N=1800/9=200

$$\bar{x} = \sum_{i=1}^{n} x_i = 200$$

Risultato finale cosa è? Che media è? Chi/cosa rappresenta? E' la media del Nord Italia?

Calcolo della media: esempio (2)

- Un ricercatore, più bravo di me e di voi, dice che in realtà la media del test Invalsi (prova di italiano) del Nord Italia è maggiore (seppur di poco) rispetto a quella da noi calcolata, considerando tutti gli studenti nel loro complesso:

$$\bar{x} = \sum_{i=1}^{n} x_i = 200 \qquad \bar{x} = 200{,}18$$

Come mai? Che cosa è successo? Abbiamo dimenticato qualcosa di utile…?

Pensiamo alle regioni, alcune sono piccole…altre grandi…

Calcolo della media: esempio (3). **La ponderazione**

Regione	***Voto***	***Popolazione in milioni***
V Aosta	201	0,1
Piemonte	203	4,4
Liguria	200	1,5
Lombardia	200	10
Bolzano	196	0,5
Trento	200	0,5
Veneto	200	4,9
Friuli-Ven.G.	202	1,2
Emilia-Romagna	198	4,4

Quando si calcola una media ponderata (o pesata) ogni unità statistica può avere peso differente, a seconda del volume, della popolazione, del reddito ecc.

In questo caso ponderiamo rispetto alla popolazione, se fosse stato uno studio sul numero di alberi magari avremmo ponderato sui km^2 di aree verdi ecc.

La mediana

La **mediana** è definita come: «modalità che si trova in posizione centrale di una serie ordinata di dati, e che quindi lascia alla propria sinistra il 50% delle osservazioni e alla propria destra il 50% delle osservazioni»

Facile su poche osservazioni. Poniamo il caso di osserva N=5 unità: 3 - 5 - 8 - 9 -10.

Sapreste indicare la modalità centrale?

Poniamo il caso di N pari, N=6. 3 - 5 - 8 - 9 -10 -12. E ora? → Si fa la media tra i due valori centrali!

Ma con N=100? N=15000? Aiutiamoci con una formula… $X_{\left(\frac{N+1}{2}\right)}$

Si legge: la modalità della variabile X alla posizione $\frac{N+1}{2}$ una volta ordinati i dati

La mediana: ordinare i dati e trovarla

Regione	*Voto*
V Aosta	201
Piemonte	203
Liguria	200
Lombardia	200
Bolzano	196
Trento	200
Veneto	200
Friuli-Ven.G.	202
Emilia-Romagna	198

1) Dati non ordinati…facciamolo! (se non lo farete molto probabilmente verrete bocciati)

Ricordiamoci che dobbiamo ordinare le modalità, non altro…

2) Utilizziamo la formula per capire a che posizione guardare

$$Mediana = X_{\left(\frac{N+1}{2}\right)} = X_{\left(\frac{9+1}{2}\right)} = X_5$$

Voto
196
198
200
200
200
200
201
202
203

Ora abbiamo capito che dobbiamo guardare la quinta posizione, partendo dal numero più piccolo

MEDIANA = 200

Meglio mediana o media?

Ci dicono cose diverse, spesso vengono utilizzate entrambe

A volte coincidono, a volte sono vicine, a volte sono molto distanti…

La media risponde alla domanda «se dovessi assegnare ad ogni individuo lo stesso valore per ottenere la stessa somma totale, quale valore dovrei assegnare?»

La mediana risponde alla domanda «se dovessi trovare il valore che si trova al centro e che divide in due parti di uguale frequenza i dati, a quale valore dovrei guardare?»

Valori anomali, o outlier

Sono valori decisamente diversi da tutti gli altri, o perché troppo grandi o perché troppo piccoli

Possono essere frutto di:

Errata digitazione o differente scala di misura (valori anomali erronei) Esempio: volevo scrivere 10 e ho scritto 100 digitando sul foglio excel, ho messo altezza in cm a tutti e ad una sola persona in metri ecc.

Valori di individui con caratteristiche particolari (valori anomali in senso stretto). Esempio: analisi sui redditi delle famiglie italiane e mi capita in un campione di 100 famiglie la famiglia Agnelli…

Se ci si accorge di valori anomali erronei tendenzialmente li si corregge/elimina, altrimenti valori anomali in senso stretto restano inclusi nell'analisi.

Si cerca di utilizzare indici in grado di «contenere» comunque questi outlier, in modo che non sballino tutte le analisi

Mediana più robusta di media!

Quando un indice è particolarmente in grado di «reggere» alla presenza di valori anomali, allora esso si dire robusto.

Poniamo il caso di avere 10 valori: 2-4-5-5-6-7-8-9-9-106

A quanto è uguale la media? → $\frac{1}{N}\sum_{i=1}^{n} x_i = 161/10 = 16{,}1$

Questo valore, 16.1, tende a non rappresentare né i 9 valori «normali» né il valore anomalo che è molto più elevato della media.

A quanto è uguale la mediana? $Mediana = X_{\left(\frac{N+1}{2}\right)} = X_{5,5} = \frac{6+7}{2} = 6{,}5$

DOMANDA

Quale dei due è stato influenzato dall'outlier 106? Che spiegazione ne diamo?

Un «*meme*» statistico…per ricordarvi della robustezza!

Moda

Quando qualcosa «va di moda», cosa intendiamo?

Probabilmente è un qualcosa (tendenza, atteggiamento ecc.) che segue la maggioranza…qualcosa di molto comune!

Nella tabella di frequenza è proprio la frequenza che indica quanto ogni modalità sia comune

La moda è infatti definita come: «La modalità a cui è associata la frequenza maggiore».

Sia essa frequenza assoluta o relativa non fa alcuna differenza

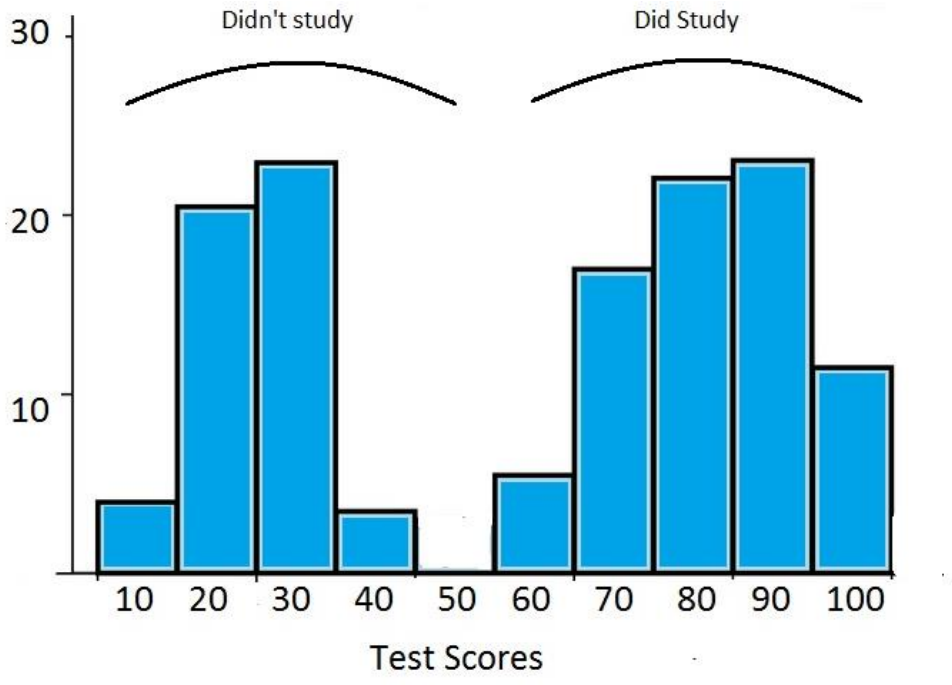
E se a due modalità è associata la stessa frequenza massima?

Allora in tal caso le mode sono due, e la distribuzione dei dati si dice bimodale

Esempio di distribuzione bimodale

Non hanno studiato punteggio medio di 25 circa

Questo è come se fosse il gruppo 1



Hanno studiato? Punteggio medio sugli 85

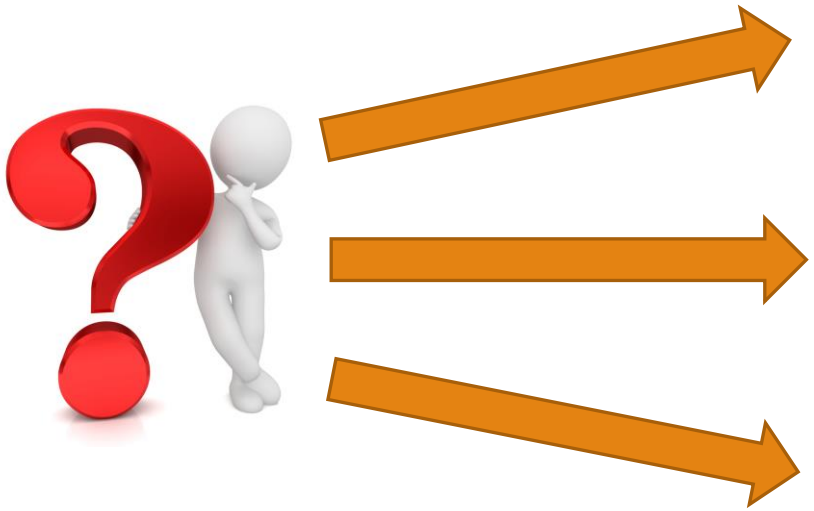
Questo è come se fosse il gruppo 2

Facendo la media di tutto otterremo circa 50, ma quanti studenti hanno preso effettivamente 50?

Un risultato simile lo otterremmo con la mediana…

In questa situazione non vogliamo andare al centro, perché il centro non rappresenta nessuno, ma piuttosto prendere le due mode: 30 e 90.

Media, moda o mediana? Perché non tutte e 3?



	Pro	Contro
MEDIA	Considera tutte le osservazioni (completezza)	Considera tutte le osservazioni (robustezza)
MEDIANA	1. Robustezza 2. Divide dati in due segmenti di pari frequenza	Tiene conto solo di 1 o 2 osservazioni centrali nel calcolo effettivo
MODA	1. Velocità 2. Indica dove si trova massima concentrazione	Nel caso di tante modalità perde di significato

Compito per casa!

Partendo da questi dati:

Voto	*CFU*
18	6
30	9
30	9
18	9
19	12
28	12

e utilizzando possibilmente sia Excel (con formule) sia i classici calcoli a mano:

1) Scrivere la tabella di frequenza

2) Trovare moda, media e mediana del voto. Commentare i risultati

3) Trovare la media PONDERATA. Come si è modificata rispetto alla media semplice? Commentare

4) Completare il seguente schema riassuntivo:

Indicando per quali tipo di variabili si può (ed è opportuno) calcolare i diversi indici di tendenza centrale

		Media	Mediana	Moda
Quantitativa	Continua			
	Discreta			
Qualitativa	Ordinale			
	Nominale			