POLO UNIVERSITARIO
FONDAZIONE
PORTOGRUARO CAMPUS

Corso di Statistica Sociale

CORSO DI LAUREA: SCIENZE DELL'EDUCAZIONE

DOCENTE: FRANCESCO SANTELLI

Prima di andare avanti…

- - Sulla seconda lezione: dubbi? Perplessità? Curiosità?
- - Come va con Excel? Avete svolto i punti dell'esercizio della volta scorsa?
- - Discutiamo della tabella indici-tipi di variabili
- - Avete avuto difficoltà nel calcolo a mano?

Oltre la mediana… (1)

Richiamino:

-La mediana è la modalità che occupa la posizione centrale. ***Si può calcolare per..***

-Divide in due la distribuzione dei dati ***ordinati***: una metà sopra la mediana e una metà sotto

-Formula: $X_{\left(\frac{N+1}{2}\right)}$

- La mediana è anche definita come il 50esimo percentile
- La mediana è anche definita come il secondo quartile

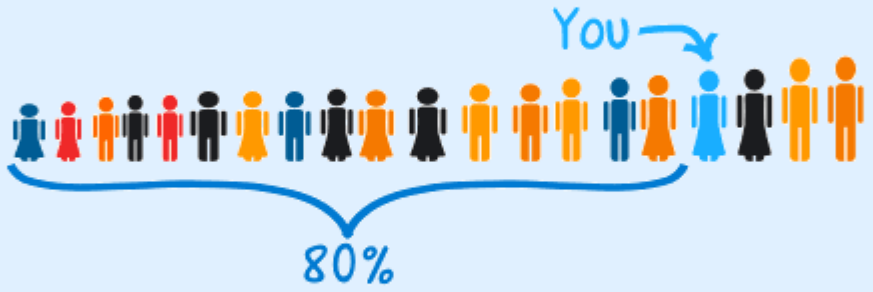
Oltre la mediana… (2)

Ma che sono i percentili e i quartili!?

Un X Percentile è la modalità che lascia alla propria sinistra (prima di sé, al di sotto di sé, a sinistra di sé ecc.) una percentuale X dei dati

Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:



That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Oltre la mediana… (3)

Ma che sono i percentili e i quartili!?

Un quartile è la modalità che lascia alla propria sinistra (prima di sé, al di sotto di sé, a sinistra di sé ecc.) una percentuale dei dati pari a **25%, 50 %** o **75%.**

Si chiama quartile proprio perché divide in 4 parti di pari frequenza la distribuzione.

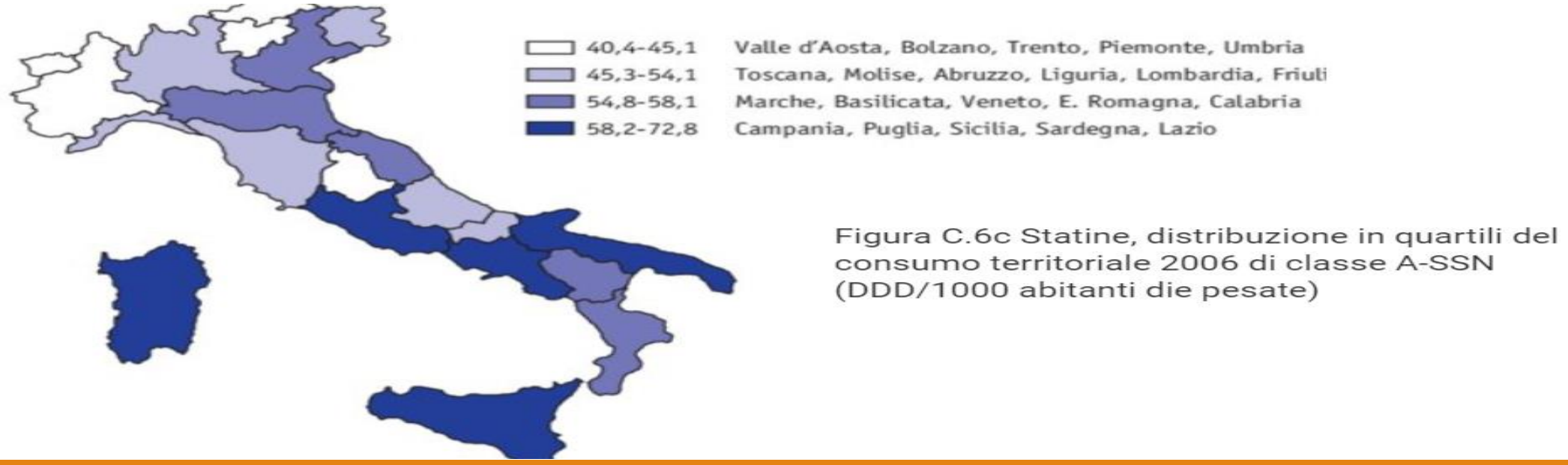
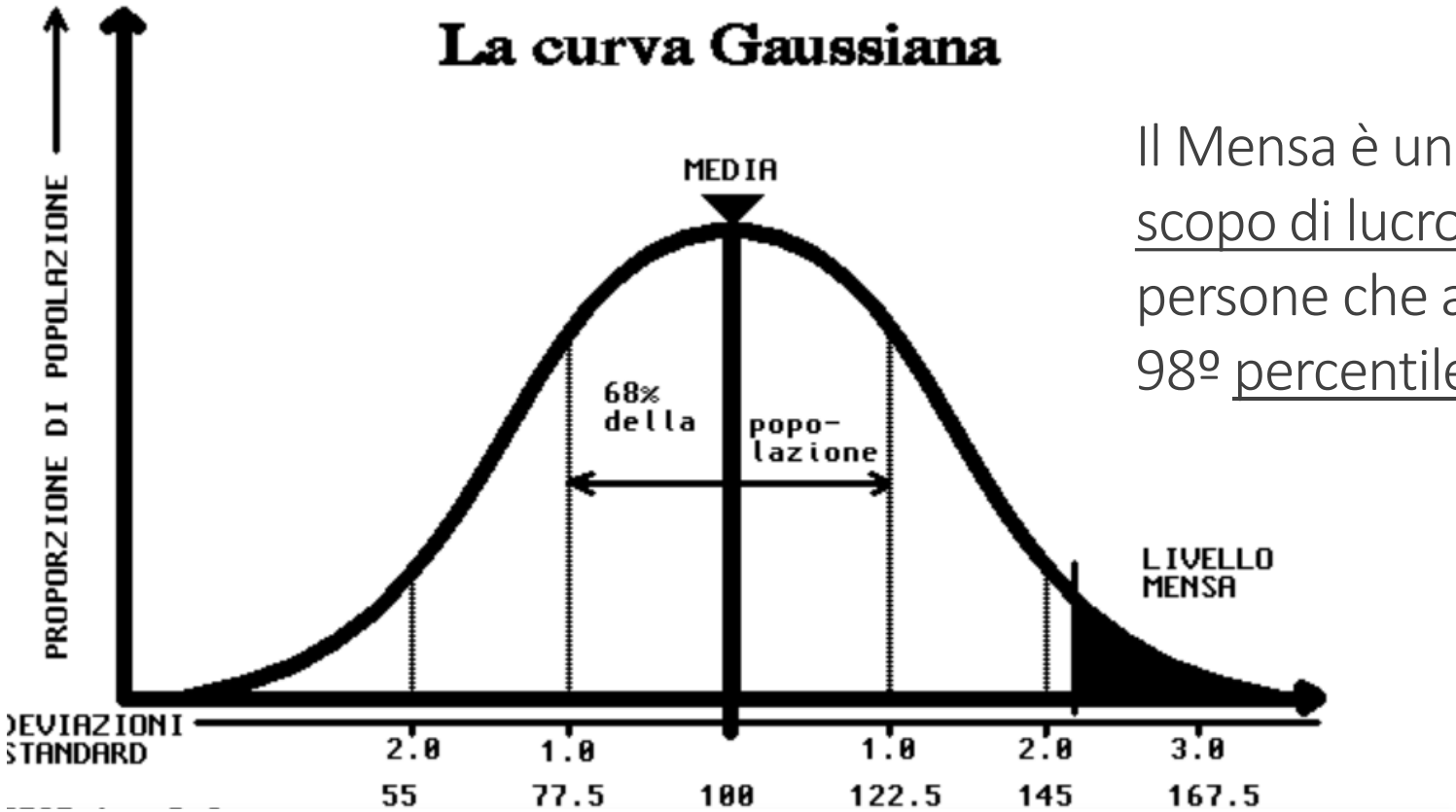


Figura C.6c Statine, distribuzione in quartili del consumo territoriale 2006 di classe A-SSN (DDD/1000 abitanti die pesate)

Un esempio di utilizzo dei percentili



Il Mensa è un'associazione internazionale senza scopo di lucro di cui possono essere membri le persone che abbiano raggiunto o superato il 98º percentile del QI (quoziente d'intelligenza).

Soddisfiamo ora il vostro amore per le formule! Troviamo le posizioni dei quartili

La formula base di tutti i quartili è: $X_{\left(\frac{N+1}{4}\right)}$ poi di volta in volta si moltiplica per 1 (Q_1), 2(Q_2) e 3 (Q_3)

Primo quartile → Q_1 → $X_{\left(\frac{N+1}{4}*1\right)}$

Secondo quartile → Q_2 → Mediana → $X_{\left(\frac{N+1}{\cancel{4}}*\cancel{2}\right)} = X_{\left(\frac{N+1}{2}\right)}$

Terzo quartile → Q_3 → $X_{\left(\frac{N+1}{4}*3\right)}$

Soddisfiamo ora il vostro amore per le formule! Capiamo i percentili

- Per i percentili, il calcolo per trovare la posizione è ancora più semplice: va trovata la modalità che si trova alla posizione:

$$X_{((N+1)*p)}$$

Per il corrispettivo percentile *p* che stiamo cercando espresso in decimali.

Ad esempio, se stiamo cercando il 65° percentile, la posizione a cui guardare sarà:

$$X_{((N+1)*0{,}65)}$$

Esercizietto (1) (a mano o Excel)

Regione	***N° Atenei***
Abruzzo	3
Basilicata	1
Calabria	3
Campania	6
EmiliaR.	4
FVG	2
Lazio	6
Liguria	1
Lombardia	8
Marche	4
Molise	1
Piemonte	3
Puglia	4
Sardegna	2
Sicilia	3
Toscana	7
TAAD.	1
Umbria	2
V Aosta	0
Veneto	4

Dati i seguenti dati sulle 20 regioni italiane e il loro numero di atenei, si calcolino:

1) La mediana
2) I due quartili rimanenti, il primo Q_1 e il terzo Q_3

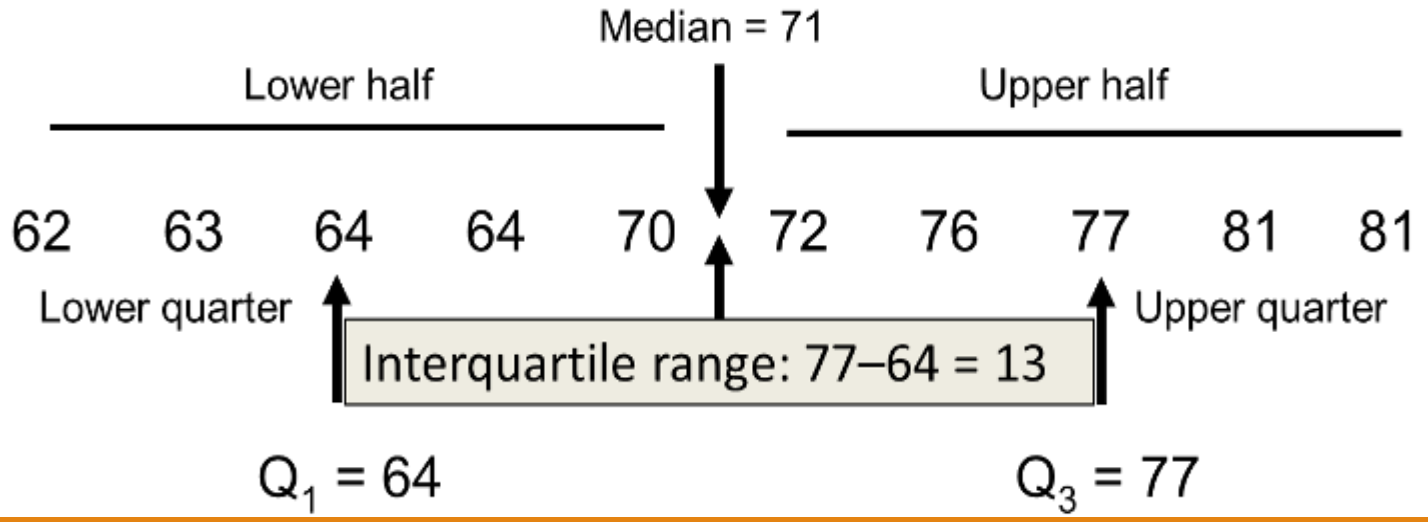
3) La percentuale delle 20 regioni italiane comprese tra il primo quartile e la mediana
4) Dove si trova il 90° percentile (0,90).
5) Quante regioni si trovano al di sotto del 10° percentile? (0,10).

Il range interquartile (IQR)

Serve a capire la massa centrale dei dati (il 50% piu normale, che si comporta con tendenze medie, che non si discosta dalla massa ecc.) tra quali valori è compreso, cioè in quale range.

E' semplicemente la differenza tra il terzo quartile Q_3 e il primo quartile Q_1

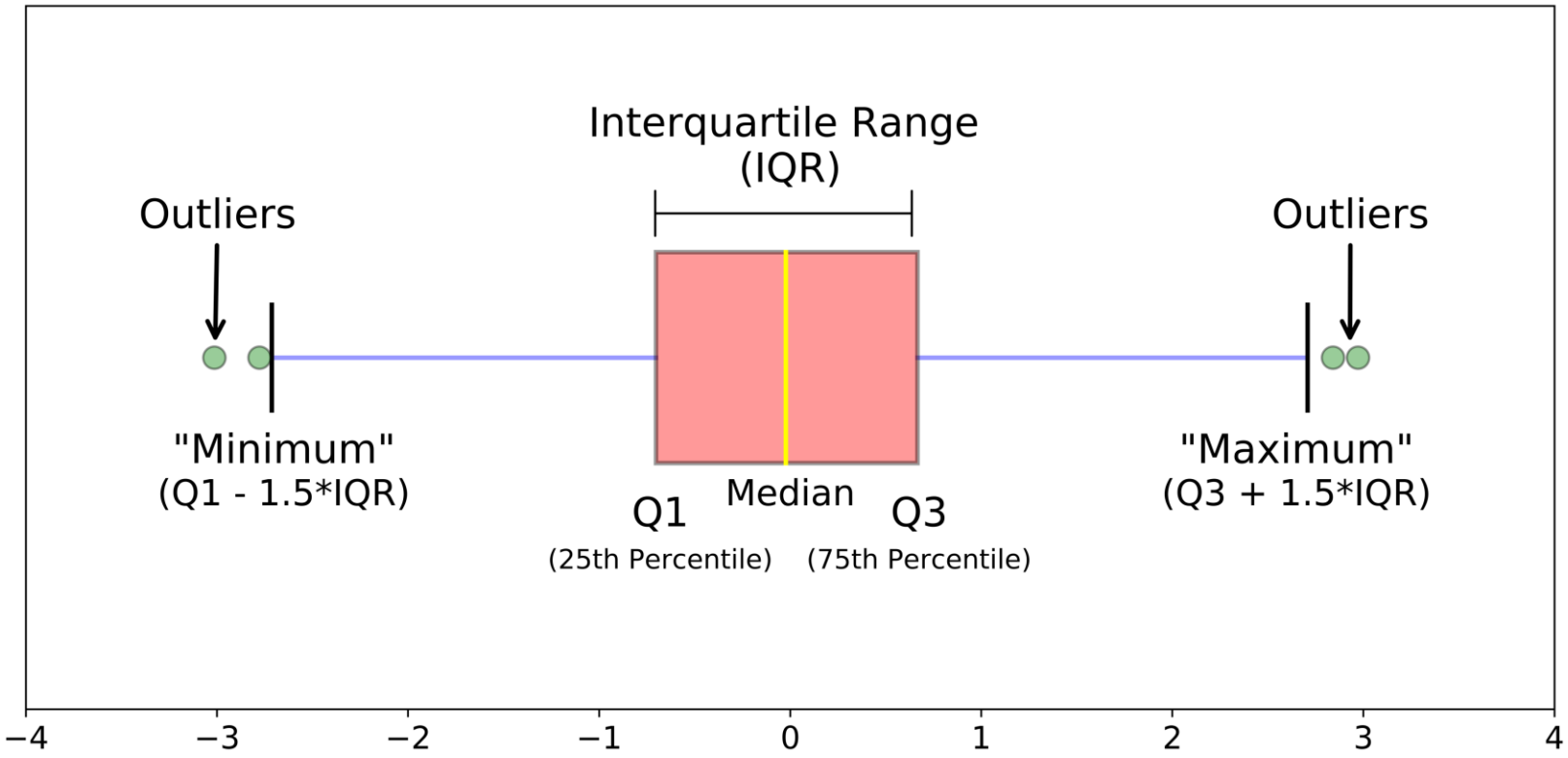
$$IQR = Q_3 - Q_1$$



Rappresentazioni grafiche: BOX-PLOT

Uno dei piu adatti a rappresentare dati continui e utilizza tutte le quantità che abbiamo già calcolato!

Ci servono:
1) Mediana
2) Quartili
3) IQR



Unica nuova quantità, per disegnare i «baffi»
1) IQR*1,5

Tutto ciò esterno ai «baffi» si considera un outlier

Esercizietto (2) (a mano o Excel)

Regione	*N° Atenei*
Abruzzo	3
Basilicata	1
Calabria	3
Campania	6
EmiliaR.	4
FVG	2
Lazio	6
Liguria	1
Lombardia	8
Marche	4
Molise	1
Piemonte	3
Puglia	4
Sardegna	2
Sicilia	3
Toscana	7
TAAD.	1
Umbria	2
V Aosta	0
Veneto	4

Torniamo ai dati di prima sulle regioni e sugli atenei:

1) Si calcolino tutti gli elementi necessari al boxplot
2) Si valuti la presenza o meno di outliers
3) Si ipotizzi che il software di analisi dei dati abbia modificato il valore della valle d'aosta da 0 a 99. Come cambierebbe il boxplot?
4) Si calcolino la media e la moda della distribuzione

Quante rappresentazioni grafiche esistono?

Sono praticamente infinite!! Noi ne vedremo solo alcune, le principali

Si pongono obiettivi diversi e sono strutturata per variabili diverse

Alcune lavorano con alcuni indici, altre con curve, altre con mappe, altre con frequenze ecc.

Spesso non sono intercambiabili: le informazioni che ricaviamo da un tipo di grafico non lo ricaveremo da un altro!

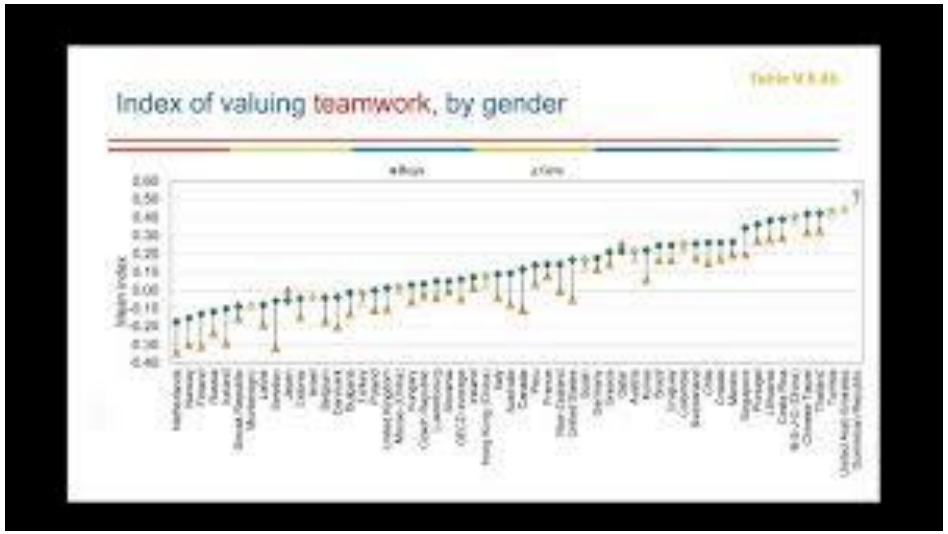
Meglio utilizzare indici? Meglio tabelle? Meglio grafici?

Dipende dai dati e dagli obiettivi, solitamente non si discute lo stesso fenomeno utilizzando tutti gli strumenti disponibili, altrimenti si è ridondanti!

Alcuni esempi (1)

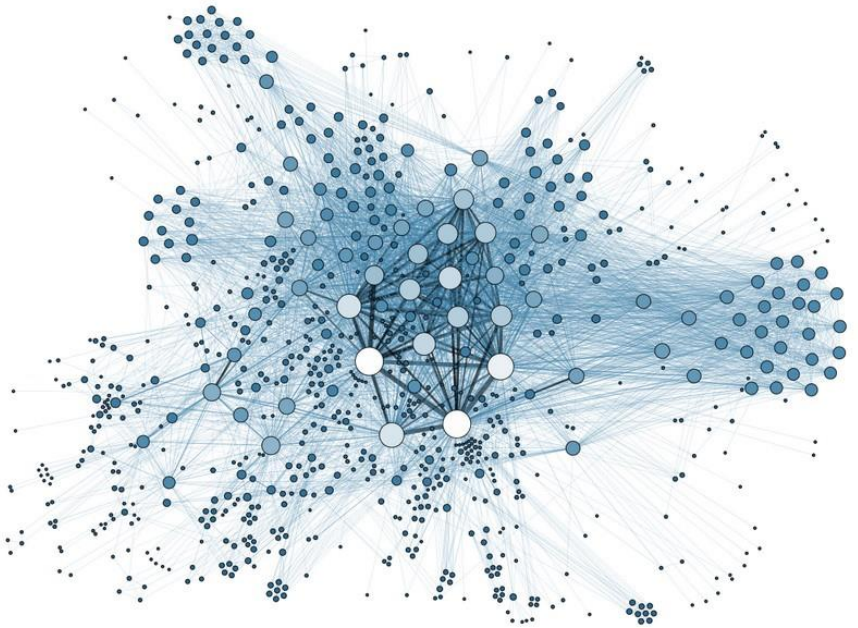
Linee, barre,
Frecce, puntini,
Torte, fette
Curve, rette
Piramidi, box-plot..

Alcuni esempi (2)



Serie storica divisa per genere

Analisi di rete:
Ogni individuo un pallino (nodo), ogni Legame-relazione una linea.



Wordcloud: parole più frequenti
Scritte più in grande

I 4 grafici che vedremo

Il boxplot è usato per rappresentare dati numerici (variabili continue ma anche discrete).

Ma il tipo di rappresentazione grafica dipende dalla natura della variabile (o delle variabili).

Altre rappresentazioni molto utilizzate (che vedremo più nel dettaglio) sono:

1) Diagrammi a torta
2) Diagrammi a barre

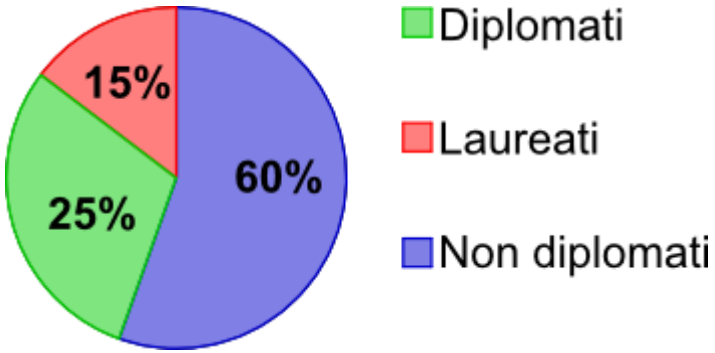
→ Qualitative o discrete, poche modalità.
Non vanno bene per continue

3) Istogrammi → Una variabile continua (come il box-plot) e divise in classi.
Non vanno bene per qualitative o discrete con poche modalità.

4) Grafico Radar → Più continue legate allo stesso macro-concetto (scale di soddisfazione Da 1 a 5).

Diagramma aTorta

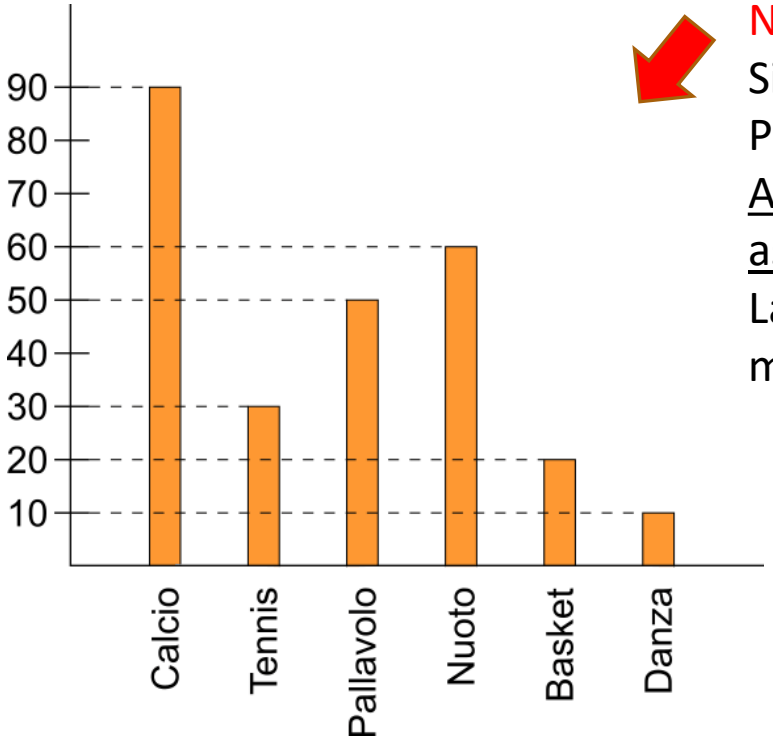
Rappresentazione grafica per variabili con poche modalità
Ogni fetta della torta è tanto grande quanto tanto grande è la frequenza
Solitamente si usano frequenze percentuali



Si può utilizzare per qualitative ma
Anche per numeriche discrete

Se è stato costruito bene, la somma delle fette deve fare….

Barre semplici



Sport preferito da un numero di 260 studenti

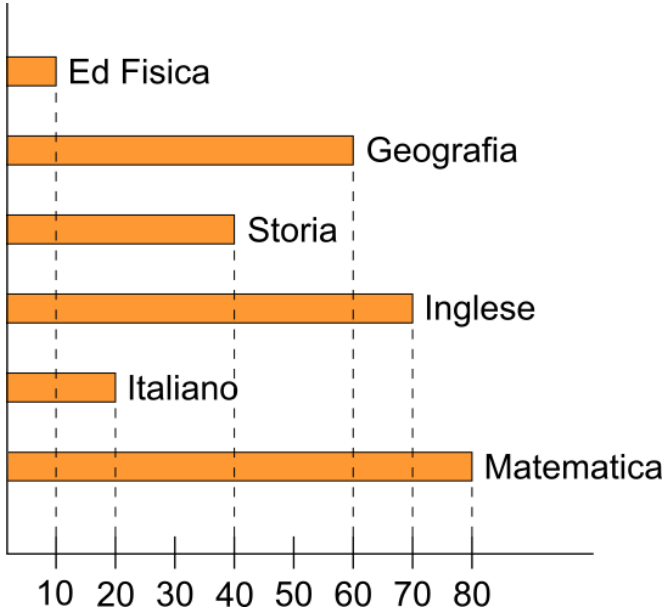
N=260

Si usano frequenze assolute

Poche modalità di solito su asse X

<u>Altezza barra proporzionale a frequenza assoluta</u>

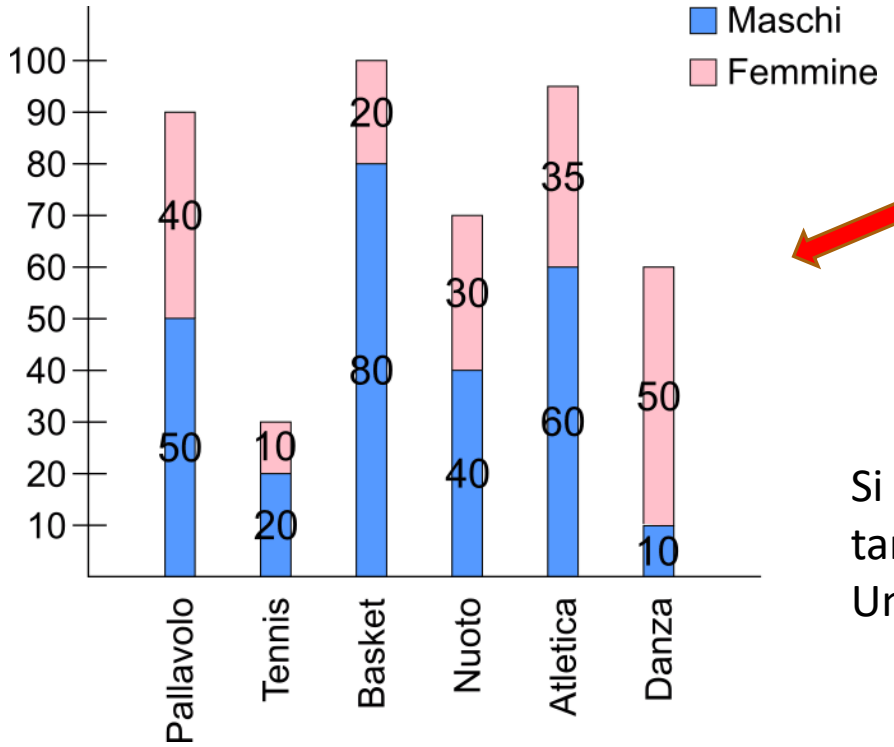
Larghezza barra solitamente uguale per tutte le modalità e non ha alcun significato statistico



nel caso di barre orizzontali e non verticali?

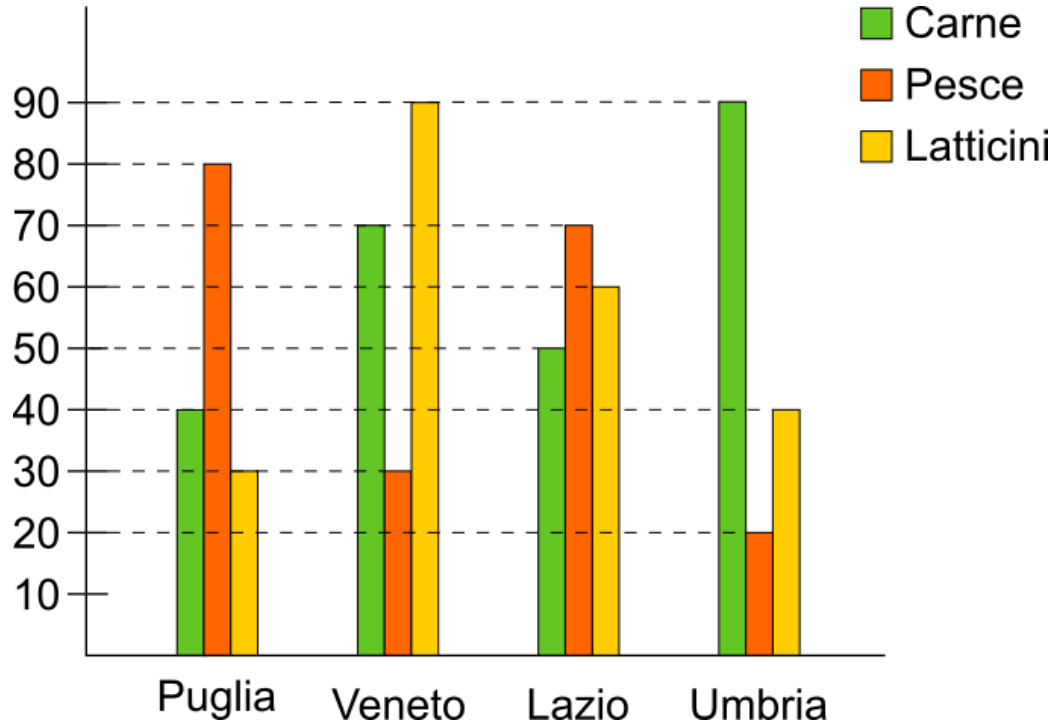
INTERPRETAZIONE RESTA IDENTICA

Barre sovrapposte o affiancate



Si «sovrappone» su
Ogni barra un'altra
variabile come il genere.
Si ottengono
Due piccole barre
sovrapposte

Si mettono una vicina all'altra,
tante piccole barre utilizzando
Un'altra variabile.



Esercizietto (3): da finire a casa

Regione	*N° Atenei*	*Area*	*Sport piu diffuso dopo il calcio*	*N° Atleti prof. (migliaia)*
Abruzzo	3	Centro	Tennis	3,5
Basilicata	1	Sud	Badminton	1
Calabria	3	Sud	Tennis	2,2
Campania	6	Sud	Basket	8
EmiliaR.	4	Nord	Tennis	8
FVG	2	Nord	Basket	2
Lazio	6	Centro	Tennis	8,5
Liguria	1	Nord	Vela	3
Lombardia	8	Nord	Basket	12
Marche	4	Centro	Pallavolo	4
Molise	1	Sud	Penthatlon	1
Piemonte	3	Nord	Tennis	6
Puglia	4	Sud	Tennis	4
Sardegna	2	Sud	Pallavolo	2
Sicilia	3	Sud	Tennis	3
Toscana	7	Centro	Tennis	7
TAAD.	1	Nord	Invernali	3
Umbria	2	Centro	Tennis	2
V Aosta	0	Nord	Invernali	0,1
Veneto	4	Nord	Pallavolo	6

1. Calcolare boxplot per n° atleti per regione
2. Confrontare gli outliers individuati qui con quelli degli atenei
3. Costruire due grafici a barre semplici, uno orizzontale e uno verticale: uno per l'area geografica e uno per lo sport più diffuso solo il calcio
4. Costruire un diagramma a torta per lo sport più diffuso solo il calcio
5. Provare a costruire un grafico a barre sovrapposte che abbia senso con questi dati…