UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Concetti e metodi per le analisi statistiche:

Precorso (a.a. 2025-2026)

Dipartimento di Scienze Politiche

Professoressa Associata: **Carella Miriam (Maria)**
Docente di *Elementi di statistica e demografia*

Dottor **Campaniello Luigi**
Laureato Magistrale in *Scienze dello Sviluppo e della Cooperazione Internazionale*

La Matematica
come linguaggio fondamentale

- La statistica utilizza il linguaggio e gli strumenti della matematica per dare senso ai dati
- I calcoli statistici si basano su operazioni semplici come addizioni, sottrazioni, moltiplicazioni, divisioni e radici quadrate
- Le formule statistiche sono scritte in un **codice di simboli** che comunica una sequenza specifica di operazioni da eseguire. Ogni simbolo rappresenta un'operazione o un elemento.

$$\bar{x} = \Sigma_{xi}/\mathrm{x_n} = (\mathrm{x_1}+\mathrm{x_2} + \ \dots + \ \mathrm{x_n})/\mathrm{x_n}$$

La Matematica
come linguaggio fondamentale

- La statistica utilizza il linguaggio e gli strumenti della matematica per dare senso ai dati
- I calcoli statistici si basano su operazioni semplici come addizioni, sottrazioni, moltiplicazioni, divisioni e radici quadrate
- Le formule statistiche sono scritte in un **codice di simboli** che comunica una sequenza specifica di operazioni da eseguire. Ogni simbolo rappresenta un'operazione o un elemento.

$$\bar{x} = \Sigma_{xi}/\mathrm{x_n} = (\mathrm{x_1}+\mathrm{x_2} + \ \dots + \ \mathrm{x_n})/\mathrm{x_n}$$

La Matematica
come linguaggio fondamentale

Legenda degli elementi:

$\bar{x}$ (x con una barra sopra) = La media aritmetica, cioè il valore medio che stiamo calcolando

$\mathbf{x_1}$, $\mathbf{x_2}$, $\mathbf{x_3}$... = I singoli valori del nostro dataset (il primo valore, il secondo valore, il terzo valore, e così via)

$\mathbf{x_n}$ = L'ultimo valore della serie (dove "n" indica la posizione dell'ultimo elemento)

n = Il numero totale di osservazioni/valori che abbiamo

Σ (sigma maiuscola) = Simbolo di sommatoria, significa *somma tutti i valori*

$\mathbf{x_i}$ = Un valore generico della serie (dove "i" può essere 1, 2, 3... fino a n)

Esempio pratico:

Se hai i voti: 7, 8, 6, 9, 5

$\mathbf{x_1} = 7$, $x_2 = 8$, $x_3 = 6$, $x_4 = 9$, $x_5 = 5$

n = 5 (abbiamo 5 voti)

$\bar{\mathbf{x}} = (7 + 8 + 6 + 9 + 5) / 5 = 35 / 5 = 7$

La Matematica
come linguaggio fondamentale

Teoria degli Insiemi: La Base per Definire gli Eventi

La probabilità (fondamento della statistica) si basa sulla teoria degli insiemi.

- **Insieme:** Una collezione di oggetti distinti (l'insieme dei numeri).
- **Elemento:** Ciascun oggetto all'interno di un insieme (1,2,3 …)
- **Spazio Campionario (S oppure Ω)**: È l'insieme di tutti i possibili risultati di un'osservazione.

Esempio: Se lanci una moneta, lo spazio campionario è S = {Testa, Croce}.

- **Evento (A)**: È un sottoinsieme dello spazio campionario. Rappresenta uno o più risultati specifici che ci interessano.

Esempio: Nel lancio di una moneta, l'evento "ottenere Testa" è ***A = {Testa}***.

- **Numero di Sottoinsiemi**: Per un insieme con n elementi (ad esempio, n lanci di una moneta), il numero totale di sottoinsiemi (eventi possibili; 2 ovvero testa, croce) è 2^n

Perché è essenziale?

È una competenza trasversale indispensabile per

- **Comprendere**
- **Analizzare**
- **Interpretare**

il mondo che ci circonda.

La statistica studia le tecniche per

- **Raccogliere**
- **Riassumere**
- **Presentare**
- **Elaborare**
- **Analizzare**

i dati necessari alla creazione di policy mirate alla risoluzione di problemi (sociopolitici).

Perché è essenziale?

La statistica si può applicare a:

a) Scienze politiche e sociali

- **Demografia**
  natalità, longevità, mortalità, migrazioni, ecc…
- **Politica**
  disuguaglianze, urbanizzazione, macro/micro economia, ecc…
- **Conflitti**
  sfollamenti, eserciti, armi, territori, ecc…

b) Sanità e Medicina (Epidemiologia)

- **Ospedalizzazione**
  quantità, tipo di ospedalizzazione, pazienti, malattie
- **Politiche pubbliche**
  Sistema sanitario, efficienza ospedaliera, accesso alle cure, sostegno pubblico, ecc…

Esempi

01 Guerra Civile in Myanmar

Obiettivo: Esaminare le attuali esigenze formative degli studenti di medicina del Myanmar nel contesto della guerra civile in corso.

Metodologia: analisi quantitative e qualitative

Popolazione e Campionamento: 298 studenti di medicina del Myanmar che avevano sospeso gli studi a causa della guerra civile. 10 di loro hanno condotto un'intervista.

Rilevazione Dati: questionario online per raccogliere i dati di base (demografici inclusi) ed elaborazione dell'intervista.

Analisi dei Dati: GraphPad V.5 per analisi quantitativa. Analisi tematico con metodo induttivo

Risultati Chiave: Molti studenti (89% quantitativo) hanno ritenuto utile continuare gli studi online. La pratica clinica, però, rimane l'ostacolo più importante.

Limitazioni: I dati raccolti riflettono la situazione degli studenti durante il conflitto, e le condizioni sul campo possono cambiare rapidamente.

Implicazioni e Raccomandazioni Future: indicazioni per la formazione degli studenti di medicina in Myanmar e in altre zone di guerra

Esempi

02 Cambiamento Climatico

Obiettivo: determinare l'impatto dell'urbanizzazione rapida sull'ambiente e sulla disuguaglianza sociale in Pakistan

Metodologia: studio quantitativo con disegno descrittivo

Popolazione e Campionamento: Residenti delle quattro città chiave del Pakistan: Karachi, Lahore, Multan e Quetta con campionamento casuale di 300 residenti

Rilevazione Dati: questionario strutturato con domande sia chiuse che aperte

Analisi dei Dati: analisi di correlazione, l'analisi di regressione e l'ANOVA (Analisi della Varianza)

Risultati Chiave: correlazioni positive significative con le disuguaglianze sociali, l'insicurezza abitativa e lo stress delle infrastrutture

Limitazioni: i dati auto-riferiti (self-reported data) potrebbero aver incluso errori dovuti a risposte socialmente desiderabili o a interpretazioni errate di alcune domande

Implicazioni e Raccomandazioni Future: Viene evidenziata l'importanza di implementare nuove tecnologie e pratiche più sostenibili

Esempi

03 Social Media

Obiettivo: misurare il livello di intenzione di acquisto degli utenti malesi della Generazione Z su TikTok e di esaminare l'influenza di variabili indipendenti

Metodologia: studio quantitativo con disegno descrittivo

Popolazione e Campionamento: 200 utenti TikTok della Generazione Z in Malesia, nati tra il 1997 e il 2012tramite il campionamento di convenienza

Rilevazione Dati: questionario di indagine strutturato sette sezioni e un totale di 36 domande

Analisi dei Dati: statistiche descrittive con SPSS. Analisi di correlazione di Sperman's rho e test di normalità di Kolmogorov-Smirnov

Risultati Chiave: impatto positivo significativo delle recensioni e dei commenti online sulle decisioni di acquisto

Limitazioni: il campionamento di convenienza limita la generalizzabilità dei risultati

Implicazioni e Raccomandazioni Future: esplorare l'efficacia nel guidare il successo dell'e-commerce

Contenuti

01 Cos’è la Statistica

02 Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

03 Fonti statistiche: fonti e dati

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

01

Cos’è la Statistica

Definizione e importanza nello studio dei fenomeni sociali.
Le fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati.

La **statistica** è una scienza che studia le tecniche per la **raccolta, la sintesi, la presentazione, l'elaborazione e l'analisi dei dati** (Simaremare et al., 2024).

Strumento essenziale per **dare senso** a grandi quantità di numeri o dati

a) Comprendere la realtà sociale:
Organizza, riassume, comunica dati.
b) Analizzare problemi sociali:
Aiuta nella ricerca qualitativa e quantitativa.
c) Validare e Valutare:
Permette di valutare l'affidabilità delle evidenze scientifiche.
d) Supportare le Decisioni:
Fornisce basi numeriche per giudizi informati (es. beneficio-rischio)

L'obiettivo principale nell'apprendere la statistica è sapere **quando applicare una particolare procedura e come interpretarne il risultato.**

[Casella & Berger, 2001]

02

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

La ricerca sociale si basa sulla raccolta, elaborazione, analisi e interpretazione dei dati.

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

1. Rilevazione dei dati (o raccolta)
2. Elaborazione e presentazione dei dati
3. Analisi dei dati
4. Interpretazione e formulazione dei risultati
5. Comunicazione delle conclusioni

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

1. Rilevazione (o raccolta) dei dati

Le fasi cruciali della rilevazione statistica includono:

1. Definizione della **popolazione** e **campionamento**

   a) La **popolazione** è l'intero insieme di individui o oggetti su cui si intendono trarre conclusioni.
   b) Il **campionamento** è il processo di selezione di un sottoinsieme (campione) dalla popolazione per la raccolta dati. Un campione ideale deve essere rappresentativo della popolazione di riferimento

Esempi:

Nello studio in Myanmar, la popolazione era composta dagli ***studenti di medicina del Myanmar*** *che avevano interrotto gli studi a causa della guerra civile*

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

1. Rilevazione (o raccolta) dei dati

Le fasi cruciali della rilevazione statistica includono:

2. Scelta della **tecnica di campionamento**

Il **campionamento casuale semplice** garantisce che ogni membro della popolazione abbia la stessa probabilità di essere incluso

Esistono disegni più complessi, come il campionamento casuale **stratificato** o a **cluster**.

Il campionamento di **convenienza** (o non probabilistico) implica una selezione basata sulla facilità di accesso ai partecipanti. Questo metodo può introdurre bias.

Esempi:

Lo studio "Israele vs Hamas" ha utilizzato un campionamento assimilabile a quello di convenienza, basato sull'accessibilità al team chirurgico

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

1. Rilevazione (o raccolta) dei dati

Le fasi cruciali della rilevazione statistica includono:

3. Identificazione e gestione del **bias**

Il bias è una distorsione sistematica che può rendere il campione non rappresentativo, ed è spesso difficile da misurare o eliminare

Bias principali:

a) **Bias di selezione:** Si verifica quando il metodo di campionamento favorisce sistematicamente certi individui.

b) **Bias di non risposta:** Quando i non rispondenti differiscono significativamente dai rispondenti nel campione selezionato.

c) **Bias di risposta:** Quando gli intervistati forniscono risposte imprecise o false (es. per difetti di memoria o sensibilità dell'argomento)

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

1. Rilevazione (o raccolta) dei dati

Le fasi cruciali della rilevazione statistica includono:

4. Definizione del **protocollo di misurazione**

È essenziale definire come, quando e con quale frequenza verranno effettuate le misurazioni :

Esempi:

Nello studio sui social media, è stato utilizzato un questionario strutturato distribuito tramite Google Forms via email, social media e WhatsApp.

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

2. Elaborazione e presentazione dei dati

1. Sintesi dei Dati:

I **dati grezzi** devono essere **organizzati** e **riassunti** per renderli gestibili e significativi.
Questa fase implica la computazione di statistiche riassuntive, come la media, la mediana e la deviazione standard.

- La **media** e la **mediana** indicano la tendenza centrale dei dati.
- La **deviazione standard** misura la variabilità o dispersione dei dati

Per i **dati categorici** (binari o multiplo), la sintesi avviene tramite **frequenze** e **percentuali**.

I **fogli di calcolo** servono ad organizzare le informazioni e analizzare le relazioni tra i dati raccolti.

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

2.1. Strumenti descrittivi

a) La **Media** è un **singolo valore** numerico che descrive sinteticamente un insieme di dati

Media aritmetica: somma di tutte le osservazioni divisa per il loro numero.

(1+5+2+4)/4=3

Proprietà e Funzioni Chiave

1. È l'unico numero per cui la somma algebrica delle differenze (o scarti) dei dati da quel numero è zero.

Le deviazioni sopra la media bilanciano esattamente quelle sotto la media.

2. Se ogni valore in un set di dati viene sostituito dalla media di quel set, la somma totale dei valori rimane invariata

Se ogni cittadino guadagnasse esattamente il reddito medio, il reddito totale della regione resterebbe invariato.

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

2.1. Strumenti descrittivi

3. La media calcolata da un **campione** è un ottimo stimatore della vera media della popolazione da cui il campione è tratto per la Legge dei Grandi Numeri.

In campionamenti casuali ripetuti, la media campionaria è più precisa rispetto alla mediana

4. La media è il valore che **minimizza la somma dei quadrati delle differenze tra i dati e il valore stesso.**

Questa proprietà è alla base di molti metodi statistici avanzati, come la regressione lineare, che mirano a minimizzare gli errori al quadrato.

6. La media è applicabile solo a dati numerici quantitativi, non a dati categorici (come genere, appartenenza politica) qualitativi

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

2.1. Strumenti descrittivi

6. È preferibile la **mediana** alla media quando vi è un'eccessiva differenza tra i valori estremi (outlier)

Se nel campione è presente una persona estremamente ricca, la media dei redditi può aumentare drasticamente, non riflettendo il reddito della maggior parte delle persone.

Ruolo nelle Scienze Politiche

• **Analisi di Sondaggi e Opinioni >** *Calcolare l'età media degli elettori di un partito, il punteggio medio attribuito a una politica su una scala di gradimento.*

• **Valutazione di Politiche Pubbliche >** *Determinare il reddito medio in una certa area dopo un programma di sviluppo (ma con cautela per gli outlier), o la spesa media pro capite per servizi sociali.*

• **Comprendere Fenomeni Sociali >** *Analizzare la durata media dei conflitti o l'incidenza media di eventi specifici, come nel caso delle esigenze educative degli studenti in Myanmar.*

.

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

2.1. Strumenti descrittivi

b) **La Mediana**

La mediana è una misura di **posizione centrale**. Rappresenta il valore centrale reale in un set di dati ordinato dal più piccolo al più grande

1 2 ***4*** *5 8 = 4* *1* ***2 (N) 4*** *5 = (2+4) / 2 = 3*

quando si analizzano variabili come i ***salari o i prezzi****, dove possono esserci pochi valori molto elevati che distorcerebbero la media, la mediana offre un'indicazione più rappresentativa dei "valori tipici«*

c) **La Moda**

La moda è il **valore che ricorre più frequentemente** in un set di dati.

Un set di dati può avere più di una moda (essere bimodale, multimodale) o non averne affatto.

1 3 **5 5 5** 14 **16 16** 20

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

2.1. Strumenti descrittivi

Misure di dispersione

- Quantificano quanto i dati in un insieme sono sparsi o distribuiti attorno al loro valore centrale
- Sono complementari alle misure di posizione centrale in quanto forniscono un quadro più completo della distribuzione dei dati

a) Varianza

Si calcola come la **media dei quadrati delle differenze di ciascun dato rispetto alla media aritmetica**

- Varianza elevata > valori sparpagliati rispetto alla media
- Varianza bassa > valori raggruppati vicino alla media.

b) Deviazione Standard

Radice quadrata della varianza. Rappresenta la distanza tipica di un'osservazione dalla media della distribuzione. Sconsigliata quando i valori estremi sono molto diversi tra loro

Cenni delle fasi di una ricerca sociale: dalla rilevazione all'interpretazione dei dati

3. Interpretazione e formulazione dei risultati

- I risultati statistici devono essere interpretati nel contesto della domanda di ricerca e trasformati in conclusioni significative
- **Significatività statistica** vs **Importanza scientifica**
- Le **rappresentazioni grafiche** sono un metodo di presentazione altamente efficace, preferite nella maggior parte dei casi per la loro capacità di rendere i risultati chiari e coinvolgenti

03

Fonti Statistiche

La ricerca sociale si basa sulla raccolta, elaborazione, analisi e interpretazione dei dati.

Fonti statistiche: fonti e dati

Le fonti statistiche si riferiscono ai luoghi o ai meccanismi da cui si ottengono i dati. I dati stessi sono i numeri o i punteggi ottenuti dalla ricerca attraverso l'osservazione e la misurazione. Le tipologie di dati sono:

Dati Quantitativi

In forma numerica (Quanto?)

a) Discreti

Assumono solo valori interi, pertanto senza la virgola

Esempio:

- *L'età di una persona: 10, 25, 37, ecc…*
- *ll numero di figli: 1, 2, 6, ecc…*

Il numero realmente quantificabile di qualcosa: ***quali sono i dati quantitativi discreti nei nostri casi studio?***

Fonti statistiche: fonti e dati

Le fonti statistiche si riferiscono ai luoghi o ai meccanismi da cui si ottengono i dati. I dati stessi sono i numeri o i punteggi ottenuti dalla ricerca attraverso l'osservazione e la misurazione. Le tipologie di dati sono:

Dati Quantitativi

In forma numerica (Quanto?)

b) Continui

Possono assumere qualsiasi valore all'interno di un intervallo specifico, inclusi infiniti valori decimali o frazionari. Sono generalmente il risultato di una misurazione

Esempio:

- *L'altezza non è limitata a valori interi; una persona può essere alta 1.75 metri, 1.755 metri, e così via. È il risultato di una misurazione che può variare infinitesimamente.*
- *Il peso può essere di 70 kg, 70.3 kg, 70.325 kg, ecc…*

la temperatura, il tempo, ecc… ***quali sono i dati quantitativi continui nei nostri casi studio?***

Fonti statistiche: fonti e dati

Le fonti statistiche si riferiscono ai luoghi o ai meccanismi da cui si ottengono i dati. I dati stessi sono i numeri o i punteggi ottenuti dalla ricerca attraverso l'osservazione e la misurazione. Le tipologie di dati sono:

Dati Qualitativi

- Descrivono **che tipo** o **quale categoria** catturando una caratteristica o un attributo
- Sono spesso espressi in **forma verbale** o **categorica**

Possono essere

a) Nominali

Non hanno un ordine naturale. A queste categorie vengono assegnate etichette ma non è possibile attribuire loro una grandezza.

la nazionalità (italiano, neozelandese, ecc…)

a.1) Dicotomiche

Sono un sottotipo di variabili nominali che presentano esattamente due categorie

Il sesso (maschio, femmina)

Fonti statistiche: fonti e dati

Le fonti statistiche si riferiscono ai luoghi o ai meccanismi da cui si ottengono i dati. I dati stessi sono i numeri o i punteggi ottenuti dalla ricerca attraverso l'osservazione e la misurazione. Le tipologie di dati sono:

Dati Qualitativi

- Descrivono **che tipo** o **quale categoria** catturando una caratteristica o un attributo
- Sono spesso espressi in **forma verbale** o **categorica**

c) Ordinali
Le categorie hanno un ordine naturale, ma l'entità della differenza tra le categorie non è necessariamente importante
La classificazione del dolore su una scala a 4 punti (nessuno, lieve, moderato, grave)

Per descrivere queste variabili, vengono comunemente utilizzate le **conte o le frequenze** degli individui in ogni categoria

La **moda**, ovvero l'osservazione più frequente, può essere riportata, specialmente per i dati nominali dove l'ordine non è rilevante

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Statistica e Dati:

Pre-corso introduttivo (Parte 2)

Fonti e Dati

Frequenze e Percentuali

I **dati categorici** sono spesso riassunti tramite

- **frequenze**
- **percentuali**

La percentuale di elettori che approvano una nuova legge (es. 60% approva), o la proporzione (quantitativo) di giovani che votano per un certo candidato (qualitativo).

Frequenze

Il numero di volte che un determinato evento, valore o categoria si verifica all'interno di un insieme di dati. Sono essenzialmente i conteggi grezzi delle osservazioni.

***Fi* = numero di volte che l'evento o categoria *i* si verifica**

Fonti e Dati

Frequenze e Percentuali

Se *n* è il totale delle osservazioni, la **frequenza relativa** *fi* del dato evento o categoria *i* è data da:

$$f_i = \frac{F_i}{n}$$

Dove:

- ***Fi*** è la frequenza assoluta (conteggio grezzo dell'evento)
- ***n*** è il totale degli eventi o osservazioni nell'insieme dati
- ***fi*** è la frequenza relativa, ossia la proporzione dell'evento rispetto al totale

Questa frequenza relativa può anche essere espressa in percentuale moltiplicando per 100:

$$\textbf{Frequenza percentuale} = \textbf{fi} \times 100 = \frac{F_i}{n} \times 100$$

Frequenza assoluta > conteggio grezzo

Frequenza relativa > frequenza assoluta / il totale degli elementi nel dataset.

Fonti e Dati

Frequenze e Percentuali

Caratteristiche della frequenza

Le frequenze sono utilizzate per descrivere dati qualitativi (categorici).

- Spesso visualizzate in **tabelle di frequenza** (1) o **grafici a barre** (2).

- La **moda** di un insieme di dati è l'osservazione più frequente.

la frequenza relativa a lungo termine di un risultato è interpretata come la sua probabilità; **risultati più probabili si verificano più frequentemente.**

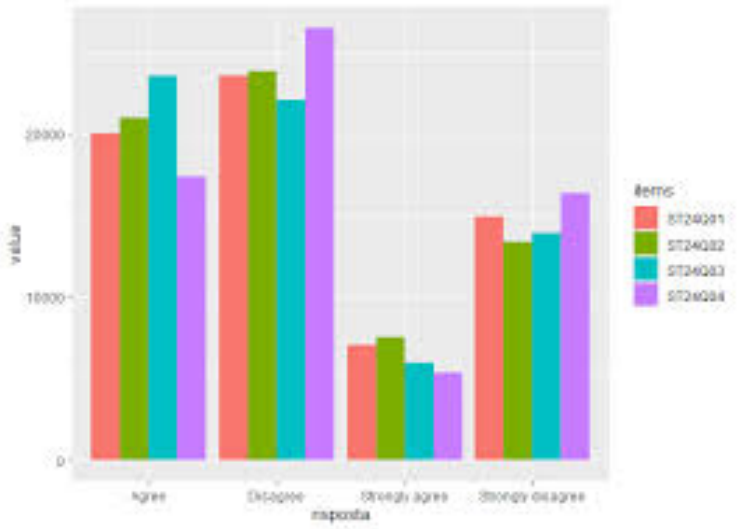
Fonti e Dati

Frequenze e Percentuali

(1) Tabella di frequenza

Altezza	Frequenza assoluta	Frequenza relativa
138	1	5%
140	3	15%
145	5	25%
148	3	15%
149	4	20%
151	3	15%
152	1	5%
TOTALE	**20**	**100%**

(2) Grafico a barre



Fonti e Dati

Frequenze e Percentuali

PROVIAMOCI !

1) Scegliere la/le variabile/i
2) Scegliere la frequenza assoluta

e poi ….

Fonti e Dati

Frequenze e Percentuali

Percentuali (e Proporzioni)

Proporzione
Numero decimale compreso tra 0 e 1 che indica una **frazione del totale**

Si calcola dividendo la frequenza di un evento per il numero totale di osservazioni. Le **proporzioni** e le **probabilità** sono **concetti identici**

$$p = \frac{F}{n}$$

Dove:
F è la frequenza dell'evento (numero di volte che l'evento si verifica)
n è il numero totale di osservazioni o casi

Percentuale

è un modo per esprimere una proporzione come una frazione di 100. Viene calcolata moltiplicando la proporzione per 100

$$P = p \cdot 100$$

Fonti e Dati

Frequenze e Percentuali

Le percentuali sono ampiamente utilizzate per riassumere i dati in vari contesti, come i **tassi**.

In un'analisi demografica, i dati possono essere presentati sia come **frequenza (n)** che come **percentuale (%)**

Analogie

- **Scopo comune**

servono a **riassumere e presentare i dati** in modo significativo

- **Relazione con il totale**

La proporzione è calcolata come **(frequenza / totale)** e la percentuale come **(frequenza / totale) * 100**.

Fonti e Dati

Frequenze e Percentuali

Differenze

- **Scala e formato**
  - **Frequenza** > conteggio assoluto (numero intero) di quante volte qualcosa si verifica.
  - **Proporzione** > valore decimale compreso tra 0 e 1.
  - **Percentuale** > valore compreso tra 0 e 100.

- **Interpretazione**
  - **Frequenza** > numero esatto di occorrenze. Non fornisce il contesto sulla sua importanza relativa rispetto all'insieme totale.
  - **Proporzione e Percentuale** > forniscono il contesto, indicando la dimensione relativa dell'occorrenza. Questo è cruciale per confrontare gruppi di diverse dimensioni o per valutare l'impatto di un valore.

- **Calcolo**
  - **Frequenza** > dato grezzo delle osservazioni.
  - **Proporzione** > dividendo la frequenza per il totale.
  - **Percentuale** > moltiplicando la proporzione per 100.

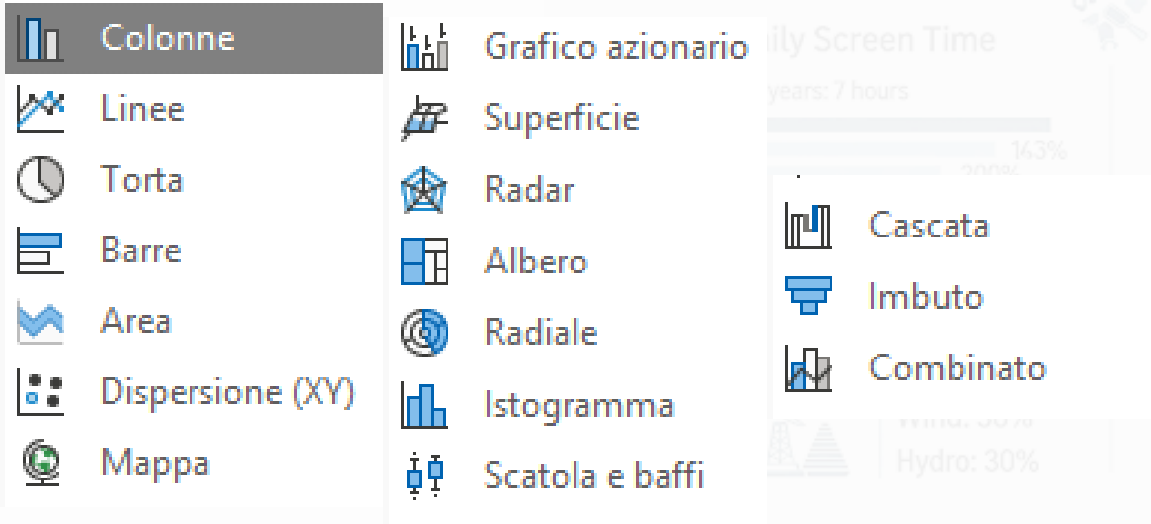
$$P_i = p_i \cdot 100 = F_i \cdot 100$$

Fonti e Dati

Grafici

c. Rappresentazione Grafica

I grafici sono strumenti preziosi per organizzare e visualizzare i dati e i loro modelli. **Danno vita ai dati**, rivelando informazioni che rimarrebbero nascoste o difficilmente interpretabili.



Fonti e Dati

Grafici

- Prendere decisioni immediate
- Aiutano a identificare le cause dei problemi stimolando l'implementazione di soluzioni.
- Possono fornire una base per predire il futuro, non limitandosi a spiegare il passato.

Fonti e Dati

Grafici

Esempio:

1. Popolazione italiana (dati non reali)

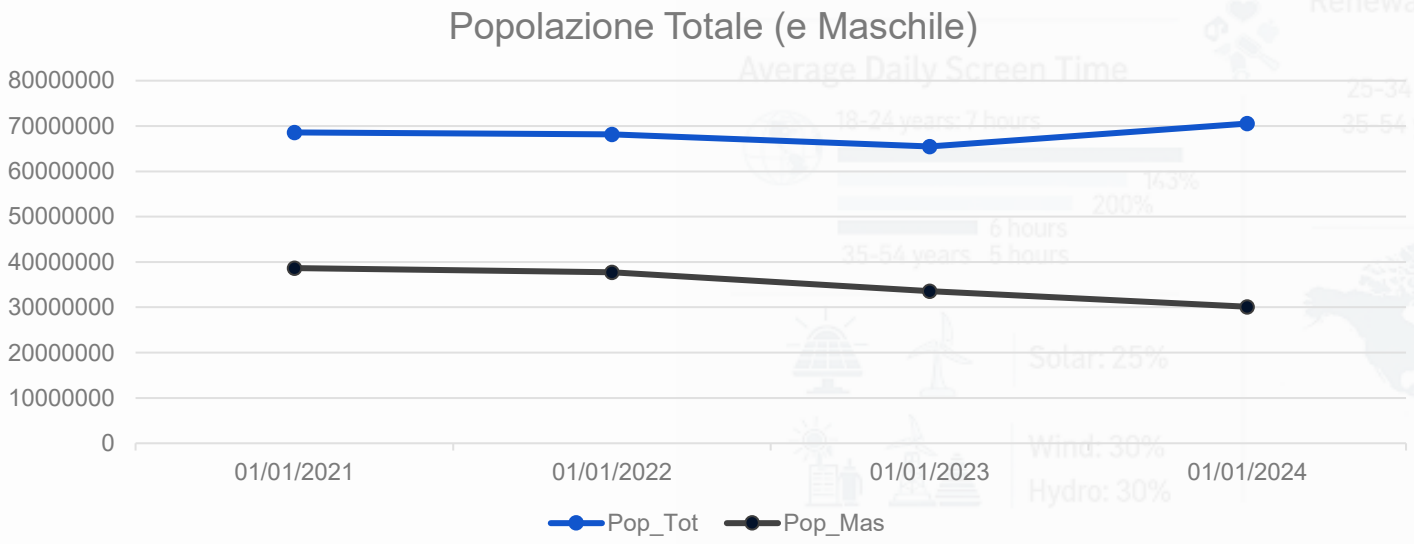
ANNO	2021	2022	2023	2024
Pop_Tot	68596324	68125364	65489325	63541289
Pop_Mas	38654782	37698421	36541274	36124589

Fonti e Dati

Proporzione e Percentuali

c. Rappresentazione Grafica

1. Popolazione italiana (dati non reali)



Fonti e Dati

Proporzione e Percentuali

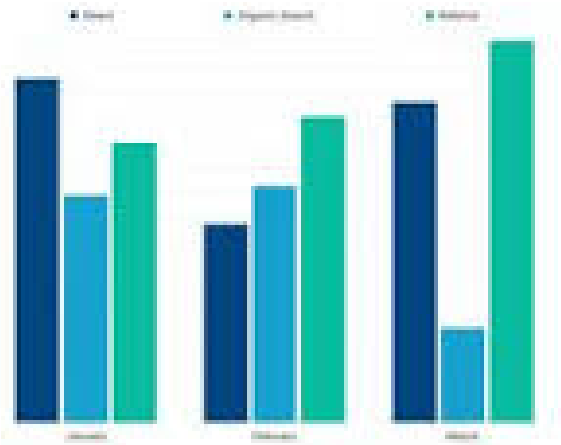
1. Grafico a Colonne

- Un **grafico a colonne** è essenzialmente un **grafico a barre ruotato di 90 gradi in senso antiorario**
- L'altezza corrisponde al **valore** (asse y)
- Tra le forme più comuni e **universali** per rappresentare i dati.
- Rappresentazione grafiche per le distribuzioni di frequenza
- Sono efficaci per confrontare valori tra **diverse categorie**.
  (possono mostrare i tassi di supporto per diversi partiti o candidati)

Fonti e Dati

Proporzione e Percentuali

1. Grafico a Colonne



Fonti e Dati

Grafici

2. Grafico a Linee

- Strumento di visualizzazione versatile e potente. Efficace per mostrare **tendenze e relazioni in dati continui**, spesso nel tempo

- Un grafico a linee rappresenta i punti dati collegandoli con segmenti di linea.
  - L'**altezza** di ogni punto corrisponde al suo valore
  - La **posizione** orizzontale rappresenta una variabile continua, come il tempo

Ideali per:

1. Mostrare come una variabile cambia nel tempo
2. Confrontare più serie di dati sulla stessa linea di base

Fonti e Dati

Grafici

2. Grafico a Linee

Vantaggi

- Eccellenti per capire rapidamente se i dati **aumentano, diminuiscono, rimangono stabili** o mostrano **schemi ciclici**.
- Mostrano i dati di cambiamento tra **due condizioni**

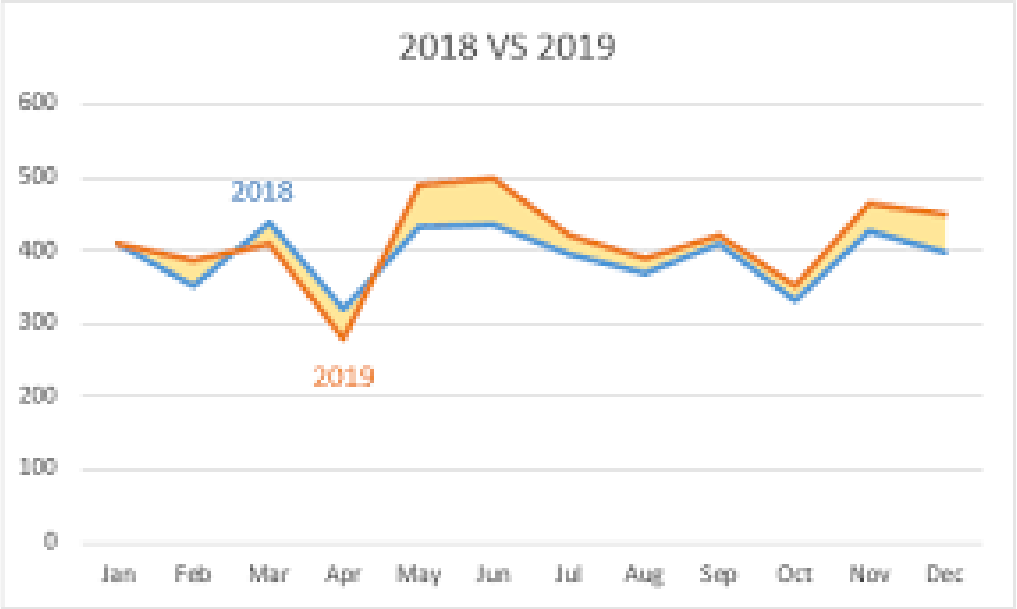
Per creare grafici a linee chiari e precisi, è consigliabile:

- Minimizzare il **Chartjunk** (eccessiva elaborazione artistica)
- Uso Appropriato dei **Tick Marks** (scelta dei limiti grafici)

Fonti e Dati

Grafici

2. **Grafico a Linee**



Fonti e Dati

Grafici

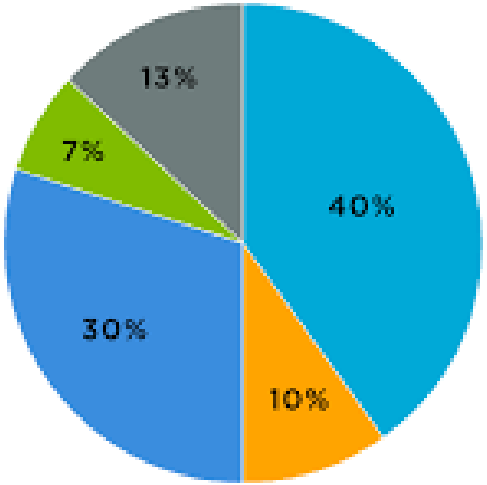
3. Grafico a Torta

- Un grafico a torta è utile quando gli utenti vogliono mostrare le **proporzioni relative di UN (E SOLO UNO) intero**
- Il suo scopo principale è confrontare le classi in termini relativi
- Notevole impatto visivo vs Efficacia visualizzazione dei dati (vedi diagramma a barre/colonne in pila)
- Può trasmettere solo una singola serie di dati

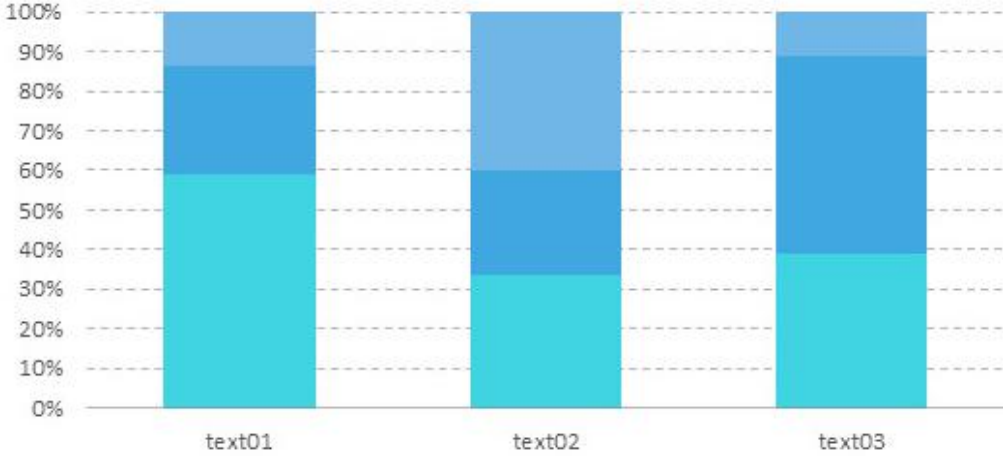
Fonti e Dati

Grafici

3. Grafico a Torta



A barre in pila 100%



Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

- Esplorano e tracciano le relazioni tra due variabili quantitative (dati bivariati).
- Osservare **la *forma, direzione* e *forza*** dell'associazione tra le variabili attraverso la **correlazione**
- Sono essenziali per verificare le ipotesi alla base di una particolare tecnica statistica e per valutarne la plausibilità del **modello**
- Permettono di osservare il modello generale dei dati e qualsiasi deviazione sorprendente da tale modello. Sono efficaci per identificare chiaramente i punti dati che sono "nettamente fuori linea" rispetto al resto dei dati, spesso indicando circostanze speciali.

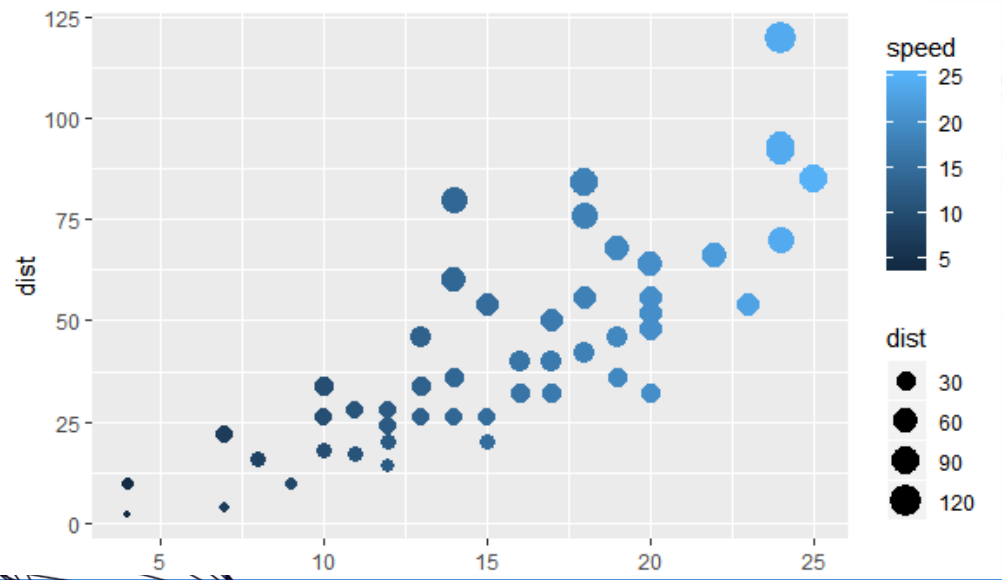
Limitazioni

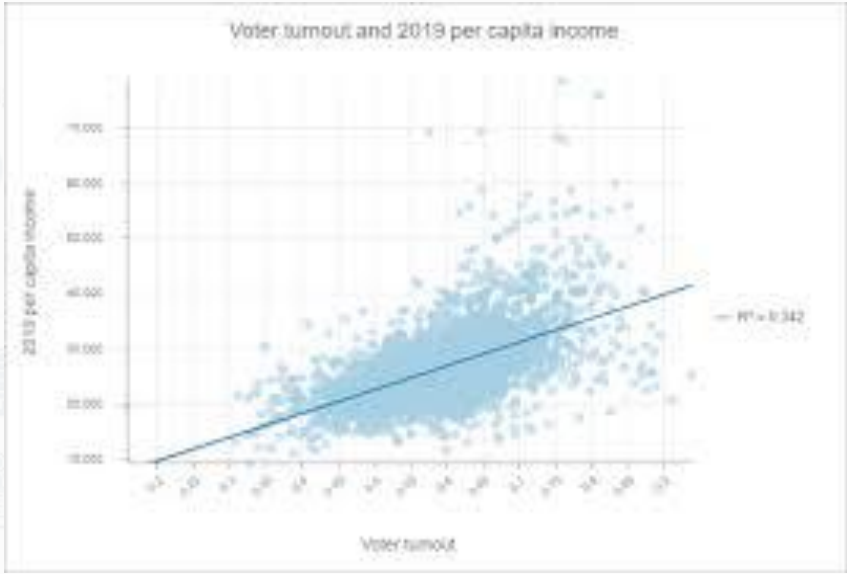
Possono diventare "sovraccarichi" e poco informativi quando utilizzati per insiemi di dati molto grandi, poiché l'eccessiva quantità di punti può renderli difficili da interpretare

Fonti e Dati

Grafici

4. **Grafico a Dispersione (XY)**





Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Forma

Si riferisce al modello generale che i punti dati disegnano sul grafico a dispersione. Osservando la disposizione dei punti, possiamo capire la natura fondamentale della relazione:

• **Linearità**: I punti tendono a raggrupparsi lungo una linea retta. Questo è il caso ideale per l'applicazione di tecniche come la regressione lineare.

• **Non Linearità**: I punti possono raggrupparsi attorno a una curva piuttosto che a una linea retta. Ad esempio, se i punti si addensano attorno a una parabola, ciò indica una forte relazione non lineare.

• **Assenza di Forma**: I punti possono apparire distribuiti in modo casuale, senza mostrare alcun modello evidente, suggerendo l'assenza di una relazione chiara tra le variabili.

Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Direzione

Indica come una variabile tende a cambiare in risposta ai cambiamenti dell'altra variabile. In un grafico a dispersione, questo si manifesta come una tendenza generale dei punti:

Direzione Positiva (Correlazione Positiva): I punti tendono ad andare verso l'alto e verso destra. Ciò significa che all'aumentare dei valori di una variabile, anche i valori dell'altra variabile tendono ad aumentare.

• **Direzione Negativa (Correlazione Negativa):** I punti tendono ad andare verso il basso e verso destra. Ciò significa che all'aumentare dei valori di una variabile, i valori dell'altra variabile tendono a diminuire.

• **Assenza di Direzione**: I punti non mostrano una tendenza chiara

Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Forza

si riferisce a quanto strettamente i punti dati si conformano al modello generale

• **Associazione Forte:** I punti sono molto vicini tra loro e si raggruppano strettamente attorno a una linea (o a una curva). La correlazione è una misura della forza della relazione matematica o il grado in cui due variabili continue sono linearmente correlate tra loro. Valori di correlazione vicini a 1 o -1 indicano una relazione lineare **forte.**

• **Associazione Debole:** I punti sono più dispersi e meno raggruppati attorno a una linea (o a una curva). Questo indica una relazione meno predittiva o più variabile. Un coefficiente di correlazione vicino a 0 suggerisce una relazione lineare debole

• **Assenza di Direzione**: I punti non mostrano una tendenza chiara

Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Coefficiente

Il coefficiente è tipicamente un fattore moltiplicativo numerico in un'espressione algebrica o una serie di termini. Nel contesto della statistica, un coefficiente è un valore numerico che descrive la relazione tra variabili in un modello.

Esempio

In un modello Y = b0 + b1X1 + b2X2, b1 è il coefficiente di regressione per X1, che indica quanto Y cambia quando X1 aumenta di un'unità, assumendo X2 costante.

Correlazione

Si riferisce a un'associazione o una relazione tra due o più cose. La correlazione non implica necessariamente una relazione di causa-effetto. Una correlazione elevata può esistere a causa di un terzo fattore (variabile "nascosta") o semplicemente per coincidenza.

Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Coefficiente di correlazione

È una misura adimensionale (senza unità di misura) e **standardizzata** per variare tra specifici limiti, rendendola utile per confrontare le relazioni tra diverse coppie di variabili. Il tipo più comune è il **coefficiente di correlazione prodotto-momento di Pearson**

$$\mathbf{r} = \frac{Cov(x,y)}{\sigma x \sigma y}$$

Dove:

$\mathrm{Cov}(X, Y)$ è la covarianza tra le due variabili X e Y

σ_X è la deviazione standard di X

σ_Y è la deviazione standard di Y

Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Modello

È una descrizione matematica di come vengono prodotti i dati o un quadro di riferimento per descrivere un processo o un sistema

I modelli servono a diversi scopi cruciali nell'analisi dei dati, nella ricerca e nello sviluppo:

- **Descrivere Processi**: Forniscono una descrizione matematica di come i dati sono prodotti o come un processo biologico o sociale si manifesta.

- **Spiegare e Prevedere:** Aiutano a capire le relazioni tra variabili e a prevedere esiti futuri.

- **Matematicità:** I modelli statistici sono spesso modelli matematici idealizzati per fenomeni reali. La loro esplorazione matematica delle proprietà e delle conseguenze è importante.

Fonti e Dati

Grafici

3. Grafico a Dispersione (XY)

Modello

Nonostante la loro utilità, i modelli presentano sfide:

• **Inadeguatezza alla Realtà**: Nessun modello può descrivere perfettamente il comportamento del mondo reale. Le sue ipotesi non sono mai pienamente soddisfatte.

• **Multicollinearità**: Una correlazione lineare elevata tra le variabili esplicative può influire sull'affidabilità delle stime dei coefficienti di regressione.

• **Rischio di Errori**: L'inadeguata acquisizione dei dati o la resistenza culturale possono ritardare l'applicazione efficace dei modelli.

Fonti e Dati

Grafici

5. Grafico a Mappa

- Visualizzare i dati su mappe o disegni per mostrare la distribuzione spaziale.
- Utilizzate per scopi specifici, pertanto non è un tipo grafico generico
- Tracciano e visualizzano dati in relazione a una posizione fisica o geografica. Mostrano la concentrazione o la distribuzione spaziale di fenomeni o eventi.
- Sono strumenti specifici per dati che hanno una **componente spaziale** chiara e possono essere mappati su una rappresentazione fisica
- Utilizzo di dati spaziali

Fonti e Dati

Grafici

4. Grafico a Mappa

L'efficacia di questi grafici è legata alla chiarezza e accuratezza del disegno o della mappa di base su cui i dati vengono proiettati.

Non propongono alternative dirette ai grafici a mappa

Dati Spaziali
Sono informazioni che contengono una **componente di localizzazione**

- Livello (1,2,3,4)
- Acronimi e sistemi
- Interazione con i dati in senso stretto
- Visualizzazione grafica (stock, ratio)
  Manualità con la standardizzazione

Fonti e Dati

Standardizzazione (o Normalizzazione)

È un processo che rende qualcosa **uniforme**, **confrontabile** o **conforme** a un determinato criterio o modello prestabilito.

In statistica, la standardizzazione è una tecnica fondamentale per trasformare i dati, rendendoli più gestibili, **comparabili** e idonei all'analisi.

• **Processo:** Tipicamente, questo processo implica sottrarre la media (o valore atteso) da ogni osservazione e dividere il risultato per la deviazione standard. Questa trasformazione produce una **variabile casuale standardizzata**

◦ **Creazione della Distribuzione Normale Standard:** convertire qualsiasi set di dati con distribuzione normale che ha una media di zero e una deviazione standard di uno (o varianza unitaria). Questa trasformazione consente di collocare i dati su una "Z Scale" comune.

Fonti e Dati

Grafici

Esempio (Z-score)
Immaginiamo di voler confrontare la performance di uno studente in due esami diversi:
Scenario:

- **Compito di Matematica liceale**:
  - Punteggio massimo: 10
    - Voto dello studente: 8
    - Media della classe: 6
    Deviazione standard della classe: 1,5

- **Esame di Storia universitario:**
  - Punteggio massimo: 30
    - Voto dello studente: 24
    - Media della classe: 20
    - Deviazione standard della classe: 3

Il processo di standardizzazione in statistica consiste nel trasformare un valore grezzo in uno Z-score, che indica quante deviazioni standard un'osservazione si discosta dalla media

Fonti e Dati

Grafici

Esempio (Z-score)
Calcolo Z-score (standardizzazione):

$$Z_{Matematica} = \frac{8-6}{1.5} = \frac{2}{1.5} = +1.33$$

$$Z_{Storia} = \frac{24-20}{3} = \frac{4}{3} = +1.33$$

Interpretazione:
•Il valore Z = +1.33 indica che lo studente ha superato la media della propria classe in entrambi gli esami di 1.33 deviazioni standard.

•Nonostante i punteggi grezzi e le scale (10 vs 30) siano diversi, i punteggi standardizzati ci mostrano che la performance relativa lo studente è stata identica in entrambi gli esami.

Importanza della standardizzazione:
•Permette il confronto diretto tra dati con scale diverse.
•Il punteggio Z indica la posizione di un valore rispetto alla media e alla variabilità del gruppo.

Fonti e Dati

Grafici

Esempio (Z-score)

Interpretazione:

•**Il valore Z** = +1.33 indica che lo studente ha superato la media della propria classe in entrambi gli esami di 1.33 deviazioni standard.

•Nonostante i punteggi grezzi e le scale (10 vs 30) siano diversi, i punteggi standardizzati ci mostrano che la performance relativa lo studente è stata identica in entrambi gli esami.

Importanza della standardizzazione:

•Permette il confronto diretto tra dati con scale diverse.

•Il punteggio Z indica la posizione di un valore rispetto alla media e alla variabilità del gruppo.