LINEE GUIDA PER L'ANALISI DATI

QUALI ANALISI SCEGLIERE?

L'analisi dei dati rappresenta una fase fondamentale della ricerca e richiede un approccio progressivo e strutturato. Il primo passo consiste nel descrivere le informazioni raccolte, analizzando ciascuna variabile singolarmente. Questa fase, nota come analisi descrittiva, consente di ottenere una prima comprensione della distribuzione dei dati.

Nel caso di **variabili categoriali**, come il genere o il titolo di studio, è opportuno costruire tabelle di frequenza che consentano di osservare la distribuzione delle categorie nel campione.

Per le variabili **quantitative**, invece, come l'età o il peso, è necessario calcolare i principali indicatori statistici, tra cui la media, la deviazione standard, il valore minimo e massimo, al fine di comprendere la tendenza centrale e la variabilità dei dati.

A supporto di questa fase, è possibile utilizzare strumenti software, come ad esempio Excel o altri programmi di analisi statistica (es., SPSS, Jamovi, Jasp), che consentono di velocizzare il calcolo degli indici descrittivi e l'organizzazione dei dati.

Una volta descritte le variabili, è utile rappresentarle graficamente, in modo da rendere più immediata l'interpretazione dei risultati. Le **variabili categoriali** possono essere rappresentate attraverso **grafici a barre o a torta**, mentre per le **variabili continue** risultano più appropriati istogrammi e boxplot, che permettono di visualizzare la distribuzione e la dispersione dei dati.
Successivamente, l'attenzione si sposta sull'analisi delle relazioni tra variabili. Di seguito sono riportate alcune informazioni di base sulle principali analisi statistiche, così suddivise:

- A cosa serve?
- Quando si utilizza?
- Esempi
- Concetti statistici

1. ANALISI DESCRITTIVA DEI DATI

La statistica descrittiva e l'analisi esplorativa dei dati rappresentano il primo passaggio fondamentale di qualsiasi analisi statistica. Queste tecniche permettono di comprendere come i dati sono organizzati, individuare eventuali errori o valori anomali e ottenere una prima visione delle relazioni tra le variabili. In questa fase iniziale, l'obiettivo non è ancora quello di testare ipotesi, ma di "esplorare" i dati per capire cosa possono raccontare.

A cosa serve?

La statistica descrittiva consente di sintetizzare e rappresentare le caratteristiche principali dei dati, mentre l'analisi esplorativa ha una funzione più ampia: permette di individuare pattern, relazioni e anomalie all'interno del dataset. Insieme, queste tecniche aiutano a costruire una base solida per le analisi successive, facilitando la comprensione della struttura dei dati e orientando le scelte metodologiche.

Quando si utilizza?

Si utilizza sempre come prima fase dell'analisi dei dati, indipendentemente dal tipo di ricerca. In particolare:

- dopo la raccolta e organizzazione dei dati
- prima di effettuare analisi inferenziali o modelli statistici
- quando si vuole verificare la qualità dei dati (presenza di outlier o valori mancanti)

Questa fase include generalmente analisi univariate (una variabile alla volta) e, successivamente, analisi bivariate per esplorare le relazioni tra variabili.

Esempio

Si consideri un dataset relativo a studenti, contenente informazioni su età, genere, nazionalità e livello di benessere.

In una prima fase, si analizzano le singole variabili:

- Si calcola l'età media e la deviazione standard
- Si osserva la distribuzione del genere e della nazionalità attraverso frequenze
- Si calcola il livello di benessere medio e la deviazione standard

Es. Si consideri un campione di 100 studenti. Il 60% è di genere femminile e il 40% maschile. L'età media è pari a 17.8 anni (DS = 1.5), indicando che la maggior parte degli studenti si colloca tra 16 e 19 anni. L'85% è di nazionalità italiana.
Per quanto riguarda il benessere psicologico, la media è pari a 6.8 (DS = 1.2), suggerendo una distribuzione abbastanza omogenea senza forti variazioni tra i partecipanti.

Concetti statistici

- Controllare la normalità dei dati: asimmetria e curtosi
- Indici descrittivi: media, mediana, deviazione standard
- Outlier: valori anomali
- Valori mancanti: dati assenti che possono influenzare l'analisi
- Visualizzazione dei dati: grafici per interpretare distribuzioni

2. CORRELAZIONE BIVARIATA

A cosa serve?

La correlazione di Pearson è una misura statistica che consente di analizzare la relazione tra due variabili quantitative, permettendo di comprendere se e in che modo esse variano insieme. In particolare, la correlazione di Pearson è la più utilizzata e fornisce un coefficiente, indicato con *r*, che varia tra -1 e +1.
Questo coefficiente esprime sia la direzione sia la forza della relazione. Un valore positivo indica che le due variabili aumentano insieme (relazione diretta), mentre un valore negativo indica che al crescere di una variabile l'altra tende a diminuire (relazione inversa). Valori vicini a zero, invece, indicano assenza di relazione lineare.

Quando si utilizza?

La correlazione è particolarmente utile nelle fasi esplorative dell'analisi, in quanto consente di individuare associazioni tra variabili. Tuttavia, è fondamentale sottolineare che essa non permette di stabilire relazioni di causa-effetto, ma esclusivamente di evidenziare la presenza di un'associazione statistica. La correlazione di Pearson è appropriata quando le variabili sono quantitative, la relazione è lineare e non sono presenti *outlier* rilevanti. In caso contrario, può essere più opportuno utilizzare correlazioni non parametriche, come la correlazione di Spearman.

Esempio

Per comprendere meglio il suo utilizzo, si consideri un esempio concreto. Supponiamo di voler analizzare la relazione tra il numero di ore trascorse sui social media e il livello di benessere psicologico in un gruppo di studenti. Dopo aver raccolto i dati, si calcola il coefficiente di correlazione di Pearson e si ottiene un valore pari a $r = -0.45$. Questo risultato indica una correlazione negativa moderata: all'aumentare delle ore trascorse sui social, il livello di benessere tende a diminuire.
Per interpretare correttamente il risultato, è necessario considerare anche la significatività statistica. Se il p-value associato alla correlazione è inferiore a 0.05, si può affermare che la relazione osservata è statisticamente significativa e difficilmente attribuibile al caso.

Concetti statistici

- *r* varia tra -1 e +1
- $r > 0$: relazione positiva (le variabili aumentano insieme)
- $r < 0$: relazione negativa (una aumenta, l'altra diminuisce)
- $r \approx 0$: assenza di relazione lineare
- *p-value*: indica la significatività statistica della relazione

3. T-TEST

La distribuzione t di Student rappresenta il riferimento teorico alla base del t-test e consente di valutare se una differenza osservata tra gruppi o rispetto a un valore teorico è statisticamente significativa. Nella pratica della ricerca, tuttavia, l'interpretazione dei risultati non avviene più attraverso la consultazione manuale delle tavole, ma tramite software statistici che restituiscono direttamente gli indicatori necessari.

A cosa serve?

La distribuzione t serve a determinare la probabilità che una differenza osservata sia dovuta al caso. Nei software statistici, questa informazione viene restituita sotto forma di p-value, che consente di prendere decisioni sull'ipotesi nulla senza ricorrere alla tavola t.

Quando si utilizza?

Si utilizza nell'ambito dei t-test, che possono assumere diverse forme a seconda del disegno della ricerca:

- **t-test a un campione**: viene utilizzato quando si confronta la media di un gruppo con un valore teorico o di riferimento
- **t-test per campioni indipendenti**: si applica quando si confrontano due gruppi distinti tra loro (es. maschi vs femmine)
- **t-test per campioni appaiati o dipendenti**: si utilizza quando le osservazioni sono collegate, ad esempio misure effettuate sugli stessi soggetti in due momenti diversi (prima/dopo) oppure su coppie abbinate

Esempio

Si vuole verificare se esiste una differenza nel livello di benessere psicologico tra studenti maschi e femmine. Dopo aver eseguito un t-test per campioni indipendenti con un software statistico (ad esempio SPSS o Jamovi), si ottengono i seguenti risultati:

- Media maschi = 6.5
- Media femmine = 7.2
- t = -2.15
- p = 0.035

Il valore di p è inferiore a 0.05, pertanto si rifiuta l'ipotesi nulla e si conclude che esiste una differenza statisticamente significativa tra i due gruppi. In questo caso, le studentesse presentano un livello di benessere mediamente più elevato rispetto agli studenti maschi.

Concetti statistici

- Statistica t: misura della differenza tra gruppi in relazione alla variabilità dei dati
- p-value: probabilità che il risultato sia dovuto al caso

- Livello di significatività (α): soglia decisionale (es. 0.05)
- Gradi di libertà (gdl): influenzano il calcolo della distribuzione t

4. ANOVA (Analisi della Varianza)

L'ANOVA (Analisi della Varianza) è una tecnica statistica utilizzata per confrontare le medie di tre o più gruppi e verificare se esistono differenze significative tra di essi. Rispetto al t-test, che consente il confronto tra due gruppi, l'ANOVA permette di estendere l'analisi a situazioni più complesse, in cui si vogliono analizzare più condizioni o categorie contemporaneamente.

A cosa serve?

L'ANOVA consente di confrontare più gruppi contemporaneamente. L'analisi si basa sul confronto tra la variabilità tra i gruppi e la variabilità interna ai gruppi: se la variabilità tra i gruppi è sufficientemente grande rispetto a quella interna, si può concludere che esistono differenze significative.

Quando si utilizza?

Si utilizza quando:

- la variabile dipendente è numerica (continua)
- la variabile indipendente è categoriale con tre o più gruppi
- si vuole verificare l'effetto di uno o più fattori

Tipologie principali di ANOVA

- *ANOVA a una via (one-way ANOVA)*
  È la forma più semplice e viene utilizzata quando si analizza l'effetto di una sola variabile indipendente su una variabile dipendente.

  Esempio: Si confronta il benessere psicologico tra tre gruppi di età (14–16, 17–19, 20+). Una variabile indipendente (età), più gruppi.

- *ANOVA a due vie (two-way ANOVA)*
  Viene utilizzata quando si vogliono analizzare due variabili indipendenti contemporaneamente.

  Esempio:
  Si analizza il benessere in funzione di:

  - genere (maschi/femmine)
  - tipo di didattica (online/presenza)

- *ANOVA per misure ripetute (repeated measures ANOVA)*
  Si utilizza quando le stesse unità (stessi soggetti) vengono misurate più volte.

  Esempio:
  Si misura il livello di ansia:

- prima dell'intervento
- dopo l'intervento
- al follow-up

I dati sono dipendenti (stessi soggetti).

Attraverso l'ANOVA, il software calcola:

- il valore F
- il p-value

Se il p-value è inferiore a 0.05, si rifiuta l'ipotesi nulla e si conclude che esiste una differenza significativa tra almeno due gruppi. Tuttavia, l'ANOVA non indica quali gruppi differiscono: per questo è necessario effettuare test post-hoc (ad esempio Tukey) per individuare le differenze specifiche.

È inoltre importante verificare alcune assunzioni del modello, tra cui l'indipendenza delle osservazioni, la normalità dei dati (o dei residui) e l'omogeneità delle varianze tra i gruppi.

Tuttavia, un risultato statisticamente significativo non indica necessariamente un effetto rilevante dal punto di vista pratico; è quindi utile considerare anche misure di grandezza dell'effetto *(effect size)*.

Concetti statistici

- F: rapporto tra variabilità tra gruppi e variabilità interna
- H0: tutte le medie sono uguali
- p-value: indica la significatività del risultato
- Variabilità tra gruppi vs variabilità interna: base del test

5. CHI-QUADRATO E TABELLE DI CONTINGENZA

Il test del chi-quadrato è un'analisi statistica utilizzata per verificare se esiste una relazione tra due variabili categoriali. Si tratta di un test non parametrico, basato sul confronto tra le frequenze osservate nei dati e quelle attese nel caso in cui le variabili fossero indipendenti.

A cosa serve?

Il test del chi-quadrato consente di stabilire se due variabili categoriche sono associate oppure indipendenti. In particolare, confronta i dati osservati con quelli che ci si aspetterebbe se non ci fosse alcuna relazione: maggiore è la differenza tra questi valori, maggiore è la probabilità che esista un'associazione significativa tra le variabili.

Quando si utilizza?

Si utilizza quando:

- entrambe le variabili sono categoriali (nominali o ordinali)
- i dati sono organizzati in una tabella di contingenza
- si vuole verificare l'indipendenza tra le variabili

È particolarmente utile in studi sociali, educativi e psicologici, quando si analizzano frequenze o distribuzioni tra gruppi.

Esempio

Si consideri uno studio volto a verificare se esiste una relazione tra il tipo di scuola frequentata (liceo vs tecnico/professionale) e la partecipazione ad attività extracurriculari (sì/no).

I dati possono essere organizzati nella seguente tabella di contingenza 2×2:

Tipologia Scuola	**Partecipa (Sì)**	**Non partecipa (No)**	**Totale**
Liceo	33 (64.7%)	18 (35.3%)	51 (100%)
Tecnico/Professionale	17 (32.1%)	36 (67.9%)	53 (100%)
Totale	54 (51.9%)	50 (48.1%)	104 (100%)

Le percentuali sono calcolate per riga e mostrano che il 64,7% degli studenti del liceo partecipa ad attività extracurriculari, contro il 32,1% degli studenti degli istituti professionali.

Applicando il test del chi-quadrato, si ottiene:

- $\chi^2 = 11.086$
- $p < .001$

Esiste un’associazione significativa tra il tipo di scuola frequentata e la partecipazione alle attività extracurriculari ($\chi^2(1) = 11.086$, $p < .001$). In particolare, gli studenti del liceo partecipano in misura significativamente maggiore rispetto agli studenti degli istituti professionali.

Concetti statistici

- Frequenze osservate: dati reali raccolti
- Frequenze attese: valori teorici in caso di indipendenza
- χ^2 (chi-quadrato): misura della differenza tra osservato e atteso
- p-value: indica la significatività statistica
- Tabella di contingenza: struttura base dell’analisi

6. REGRESSIONE LINEARE

La regressione lineare è una tecnica statistica utilizzata per analizzare e modellare la relazione tra una variabile dipendente (Y) e una o più variabili indipendenti (X), con l'obiettivo di comprendere come varia la variabile di interesse al variare dei predittori. A differenza della correlazione, che si limita a misurare l'associazione tra variabili, la regressione consente di quantificare l'effetto di una variabile sull'altra e di costruire modelli predittivi.

A cosa serve?

La regressione lineare serve a stimare e interpretare la relazione tra variabili, permettendo di capire quanto una variabile indipendente influenzi la variabile dipendente. Inoltre, consente di prevedere il valore della variabile dipendente a partire dai valori delle variabili indipendenti, trasformando i dati in uno strumento utile per l'analisi e la previsione.

Quando si utilizza?

Si utilizza quando:

- La variabile dipendente è quantitativa (continua)
- Le variabili indipendenti possono essere una o più (modello semplice o multiplo)
- Si ipotizza una relazione lineare tra le variabili

È particolarmente utile quando si vuole passare da una semplice osservazione di relazione (correlazione) a una spiegazione più approfondita e predittiva del fenomeno.

Esempio

Si consideri uno studio volto ad analizzare in che modo il tempo dedicato allo studio influisca sul rendimento accademico degli studenti. Dopo aver raccolto i dati, si costruisce un modello di regressione lineare in cui:

- Y = voto medio
- X = ore di studio settimanali

Il software restituisce i seguenti risultati:

- $\beta = 0.40$
- $p = 0.01$
- $R^2 = 0.30$

Il coefficiente $\beta = 0.40$ indica che, in media, ogni ora aggiuntiva di studio è associata a un aumento di 0.40 punti nel voto medio. Il p-value inferiore a 0.05 indica che questo effetto è statisticamente significativo. Il valore $R^2 = 0.30$ suggerisce che il 30% della variabilità del rendimento accademico è spiegato dal numero di ore di studio.

Questo modello consente quindi non solo di descrivere la relazione tra le variabili, ma anche di prevedere il rendimento atteso in base al tempo di studio.

Concetti statistici

- Variabile dipendente (Y): variabile da spiegare
- Variabili indipendenti (X): predittori
- β (coefficiente di regressione): indica direzione e intensità dell'effetto
- Intercetta (β_0): valore di Y quando X = 0
- R^2: percentuale di varianza spiegata dal modello
- p-value: significatività statistica
- Termine di errore (ε): componente non spiegata dal modello

7. REGRESSIONE LOGISTICA BINARIA

La regressione logistica è una tecnica statistica utilizzata per analizzare la relazione tra una o più variabili indipendenti e una variabile dipendente categoriale, generalmente binaria. A differenza della regressione lineare, che prevede valori numerici continui, la regressione logistica stima la probabilità che si verifichi un determinato evento, espressa come valore compreso tra 0 e 1

A cosa serve?

La regressione logistica serve a modellare e interpretare la probabilità di un evento in funzione di una o più variabili predittive. Permette quindi di comprendere quali fattori influenzano il verificarsi di un determinato risultato e di classificare i casi in due categorie (ad esempio sì/no, presente/assente) .

Quando si utilizza?

Si utilizza quando:

- La variabile dipendente è binaria (es. successo/fallimento, sì/no)
- Si vogliono analizzare uno o più fattori predittivi (numerici o categoriali)
- L'obiettivo è stimare una probabilità o classificare i casi

È particolarmente utile negli studi predittivi e nei contesti in cui si analizzano fattori di rischio.

Esempio

Si consideri uno studio volto ad analizzare l'associazione tra vittimizzazione tra pari (bullismo) e la presenza di ansia. I risultati mostrano che essere vittima di bullismo è associato alla presenza di ansia ($OR = 4.57$; $p < .05$). Un $OR = 4.57$ significa che chi è vittima di bullismo ha un rischio più alto (circa 4–5 volte) di sviluppare ansia rispetto a chi non lo è. Poiché il p-value è inferiore a 0.05, questa associazione è statisticamente significativa.

Concetti statistici

- Probabilità (0–1): output del modello
- Odds: rapporto tra probabilità di evento e non evento
- p-value: significatività statistica