

Principi di Inferenza

Introduzione al corso

Laura D'Angelo
Università Milano Bicocca

Nome: Laura D'Angelo

Email: laura.dangelo@unimib.it

Pagina del corso: https://laura-dangelo.github.io/principi_inferenza2627

Ricevimento: in presenza o online, scrivere via mail per prenotarsi

11 Lezioni: Dal 23 settembre al 2 dicembre

- ▶ **Slide, note, ed esercizi**, che verranno rese disponibili sulla pagina del corso.
(Ringrazio Beatrice Franzolini per il materiale).
- ▶ **Testo di riferimento:**
S.M. Ross (2014). Introduzione alla Statistica. Apogeo Education, seconda edizione

- ▶ L'esame consiste in una prova scritta di due ore.
- ▶ Le domande nella prova sono sia esercizi sia domande aperte teoriche. Le domande di teoria servono a verificare la conoscenza dei principi di inferenza
- ▶ Gli esercizi sono analoghi a quelli che vedremo in classe.

Nel tempo, la statistica ha cambiato sembianze: oggi si chiama anche *Data Science*.
Non esiste una definizione univocamente accettata.

Alcune definizioni famose:

- ▶ “La statistica riguarda ciò che si può imparare dai dati”. *Anthony Davison*.
- ▶ “La statistica è essenzialmente lo studio dell'incertezza. Il ruolo degli statistici e' di assistere i lavoratori di altri campi, o i propri clienti, i quali incontrano l'incertezza nel loro lavoro.” *Dennis Lindley*.
- ▶ “La statistica e' la tecnologia necessaria per estrarre significato dai dati e per gestire l'incertezza”. *David Hand*.
- ▶ “La statistica e' la scienza delle decisioni in condizioni di incertezza”. *Domenico Piccolo*.

Che cos'è la statistica?

Idea chiave

La statistica è la **scienza** che trasforma i **dati** in **informazione** utile per prendere **decisioni** in presenza di **incertezza**.

- × Calcolare l'area di un triangolo noti base e altezza
- × Risolvere un'equazione di secondo grado
- × Verificare se una stringa è palindroma
- × Ordinare una lista di nomi alfabeticamente
- × Calcolare il percorso più breve tra due punti su una mappa
- × Simulare il moto di un pianeta con condizioni iniziali fissate

Problemi deterministici: nessuna incertezza, nessuna variabilità.

- ✓ Stimare l'altezza media della popolazione italiana da un campione
- ✓ Valutare se un nuovo farmaco è più efficace del trattamento standard
- ✓ Prevedere i risultati di un'elezione tramite sondaggi
- ✓ Analizzare la relazione tra ore di studio e voto agli esami
- ✓ Stimare la probabilità di default di un prestito
- ✓ Prevedere la domanda di energia elettrica domani
- ✓ Individuare gruppi di clienti con comportamenti simili
- ✓ Stimare l'impatto di una politica pubblica su occupazione e consumi

Problemi con variabilità e incertezza.

Sebbene oggi abbia un significato molto diverso, la parola statistica deriva dal vocabolo italiano *Stato*. Ghislini nel 1589 indica la statistica come “descrizione delle qualità che caratterizzano e degli elementi che compongono uno Stato”.

Già le antiche civiltà raccoglievano dati per amministrare e governare.

- ▶ **Egitto**: registri agricoli e censimenti per tassazione.
- ▶ **Babilonia**: tavolette d'argilla con dati su commerci e popolazione.
- ▶ **Cina antica**: censimenti periodici della popolazione (dal 2000 a.C.).
- ▶ **Roma**: censimenti per determinare imposte, leve militari, diritti civili.

Nel linguaggio moderno, chiameremmo questa disciplina “statistica ufficiale” oppure anche “statistica economica”.

Il ruolo della statistica cambiò notevolmente a partire dalla fine del 1800, anche grazie allo sviluppo del **calcolo delle probabilità** e del gioco d'azzardo.

La Royal Statistical Society (RSS) venne fondata nel 1834, mentre l'American Statistical Association (ASA) nel 1839. Il primo dipartimento di statistica al mondo venne fondato nel 1911 all'University College di Londra (UCL).

Nella prima metà del '900 avvenne la **prima rivoluzione della statistica**. Furono gettate le fondamenta della teoria della statistica moderna (tantissimi risultati di matematica statistica vengono derivati in questi anni).

A partire dagli anni '70 e '80 del secolo scorso è cominciata **una seconda rivoluzione**, grazie all'immensa potenza di calcolo e disponibilità di dati derivanti dallo sviluppo informatico.

Questa fase rivoluzionaria è ancora in atto.

- ▶ Esplosione dei dati digitali → big data.
- ▶ Computer moderni → più potenza di calcolo per analizzare i dati.
- ▶ Nascita della **data science**: fusione di statistica, informatica e machine learning.
- ▶ Algoritmi predittivi e intelligenza artificiale.
- ▶ Statistica rimane il fondamento per valutare incertezza, validità e affidabilità.
- ▶ Applicazioni in genomica, reti sociali, economia dei dati, medicina personalizzata.

Fasi di un'indagine/analisi statistica

- ① Definizione degli obiettivi / domanda a cui si vuole rispondere
- ② Raccolta dei dati
- ③ Pulizia dei dati
- ④ Descrizione e visualizzazione dei dati
- ⑤ Inferenza su quantità ignote
- ⑥ Interpretazione e comunicazione dei risultati

Statistica descrittiva

- ▶ Riassume e visualizza i dati osservati
- ▶ Misure: media, mediana, varianza, quantili
- ▶ Grafici: istogrammi, boxplot, scatterplot
- ▶ *Obiettivo*: **comprendere** e **rappresentare** le informazioni contenute nei dati a cui abbiamo accesso

Statistica inferenziale

- ▶ Generalizza da un campione alla popolazione
- ▶ Stima parametri e quantifica l'incertezza
- ▶ Test d'ipotesi, intervalli di confidenza, predizioni
- ▶ *Obiettivo*: **estendere oltre i dati osservati** e **decidere** sotto incertezza

La descrittiva riassume i dati osservati e li rappresenta; l'inferenziale consente di andare oltre i dati osservati.

Descrittiva e inferenziale non sono disgiunte ma complementari.

Il corso è organizzato in tre parti principali.

- ▶ Statistica descrittiva
- ▶ Probabilità
- ▶ Statistica inferenziale (o semplicemente “inferenza”)

Principi di Inferenza

Concetti preliminari e terminologia

Laura D'Angelo
Università Milano Bicocca

Definizione

Le **Popolazione** e' un insieme di elementi (ad esempio, persone fisiche, animali, squadre di calcio, utenti di un sito web), che è oggetto di studio statistico. Insieme di elementi che presentano (almeno) una caratteristica comune.

- ▶ La popolazione **dipende dalle domande / obiettivi di ricerca**.

Prima si chiarisce "che cosa voglio studiare?", poi si identifica la popolazione di riferimento.

Esempi:

- ▶ Voglio capire la soddisfazione dei consumatori relativa al prodotto X.
⇒ *Popolazione*: tutti i consumatori del prodotto X.
- ▶ Voglio stimare l'età media in Italia nel 2025.
⇒ *Popolazione*: tutti i residenti in Italia al 1° gennaio 2025.
- ▶ Può essere **reale** (persone residenti a Milano al 1/1/2025) o **virtuale** (tutti i pazienti che in futuro svilupperanno una certa patologia).
- ▶ Tale gruppo può essere **finito** o **infinito**.

Definizione

Un **campione** è un sottoinsieme della popolazione su cui si collezionano i dati e che viene usato per trarre conclusioni.

- ▶ La **numerosità campionaria** indica quante unità contiene.
- ▶ Il campione dovrebbe essere **rappresentativo**.
- ▶ Il concetto di *campionamento casuale* sarà cruciale per l'inferenza.

Esempio. Su 38.000 studenti iscritti all'Ateneo di Milano-Bicocca, si estraggono casualmente 500 per un'indagine sulla soddisfazione.

Definizione

L'**unità statistica** è l'elemento base che compone la popolazione e su cui vengono rilevate le variabili di interesse.

- ▶ Ogni unità fornisce una “riga” della matrice dei dati.
- ▶ Le unità possono essere persone, famiglie, scuole, aziende, prodotti, eventi, ecc.

Esempio. Nell'indagine sugli studenti di Bicocca:

- ▶ Popolazione = tutti gli studenti iscritti all'ateneo.
- ▶ Unità statistica = ciascuno studente.

Definizione

Una **variabile** è una caratteristica osservata sulla singola unità statistica.

Esempio. Nell'indagine sugli studenti di Bicocca:

- ▶ Popolazione = tutti gli studenti iscritti all'ateneo nel 2026.
- ▶ Unità statistica = ciascuno studente.
- ▶ Variabili osservate = età, media degli esami, frequenza alle lezioni, corso di studio, numero di esami superati, soddisfazione.

Definizione

Le **modalità** sono i valori distinti che la variabile può assumere.

Esempio. Nell'indagine sugli studenti di Bicocca:

- ▶ Variabile: numero di esami superati.
Modalità: 0, 1, 2, 3, ...
- ▶ Variabile: esito binario se lo studente ha usufruito di servizi di tutorato.
Modalità: "Sì" / "No".
- ▶ Variabile: livello di soddisfazione.
Modalità: "Per niente soddisfatto" / "Non soddisfatto" / "Soddisfatto" / "Molto soddisfatto".

Una variabile è detta **quantitativa** se le modalità sono numeri e rappresentano la “quantità” della variabile, mentre viene detta **qualitativa** in tutti gli altri casi.

Le variabili quantitative si distinguono in

- ① **Discrete**: valori interi (anni di età, numero di figli).
- ② **Continue**: intervallo reale (PIL, temperatura, altezza).

Le variabili qualitative si distinguono in

- ① **Nominali**: categorie senza ordine (religione, colore occhi).
- ② **Ordinali**: categorie con ordine (gradi di scuola / livelli di soddisfazione).

- ▶ Variabile quantitativa discreta: numero di tentativi a un esame.
- ▶ Variabile quantitativa continua: tempo necessario per arrivare in università.
- ▶ Variabile qualitativa nominale: tipo di scuola superiore frequentata.
- ▶ Variabile qualitativa ordinale: grado di interesse nel corso di statistica (scarso, medio, alto).

Studente	Tutorato	Esame superato	Media esami	Pendolare
Antonio	Si	Si	27	Si
Francesca	No	Si	26	No
Marco	Si	No	23	No
Giulia	No	No	23	Si
Marta	No	Si	26	Si
...	...	...	...	...
Matteo	Si	No	25	No

Domanda di ricerca: Il servizio di tutorato è utile per superare il corso di Statistica?

Popolazione di riferimento: tutti gli studenti iscritti al corso di Statistica.

Un **campione di studenti** è un sottoinsieme degli studenti iscritti al corso di Statistica, per cui sono state registrate le variabili “tutorato”, “esame superato”, “media esami” e “pendolare”.

Definizioni

Indagine censuaria: si osservano **tutte** le unità della popolazione.

Indagine campionaria: si osserva **un sottoinsieme** (campione) e si inferisce sulla popolazione.

Indagini campionarie

- ✓ Costi e tempi ridotti
- ✓ Maggior qualità dei dati
- ✗ *Errore di campionamento* (controllabile grazie alla probabilità)

Indagini censuarie

- ✗ Maggiori costi e tempi
- ✓ Copertura completa; nessun errore di campionamento
- ✗ Maggiori errori non campionari, come errori di misurazione, copertura, non risposta (molto difficili da controllare)