Fondamenti di Analisi Dati

Emanuele Galiano

Anno Accademico 2025/2026

Indice

Capitolo 1

Concetti chiave dell'analisi dei dati

1.1 Concetti fondamentali

1.1.1 Data

Un dato è un insieme di valori raccolti riguardanti un fenomeno, un evento o un'entità specifica. I dati possono essere numerici, testuali, visivi o di altro tipo e sono fondamentali per l'analisi statistica e il machine learning.

Definizione 1.1

1.1.2 Popolazione

La popolazione è l'insieme completo di tutte le osservazioni o unità di interesse in uno studio statistico. Può essere costituita da persone, oggetti, eventi o qualsiasi altra entità che si desidera analizzare.

Definizione 1.2

La popolazione può essere denotata dal simbolo Ω.

1.1.3 Osservazioni

Un'osservazione è un'istanza specifica di dati raccolti su un'entità o un evento. In un dataset, ogni riga rappresenta un'osservazione, che include tutti i valori delle variabili misurate per quell'istanza.

Definizione 1.3

Dato un insieme popolazione Ω, le osservazioni si indicheranno con:

$$\omega \in \Omega$$

1.1.4 Campione

Un campione è un sottoinsieme rappresentativo della popolazione, selezionato per l'analisi statistica. I campioni sono utilizzati per fare inferenze sulla popolazione più ampia, poiché spesso è impraticabile o impossibile raccogliere dati su ogni membro della popolazione.

Definizione 1.4

Si parla di un campione come la tupla:

$$\{\omega^{(1)}, \omega^{(2)}, ..., \omega^{(n)}\} \subseteq \Omega$$

Dove ogni $\omega^{(i)}$ è un'osservazione appartenente alla popolazione Ω e n è la dimensione del campione.

1.1.5 Variabile

Una variabile, o feature, è una caratteristica misurabile o osservabile di un'osservazione. Quindi è una funzione che associa ad ogni osservazione un valore in un certo dominio.

Definizione 1.5

Definiamo la variabile X come una funzione:

$$X : \Omega \to S$$

Dove S è l'insieme dei possibili valori che la variabile può assumere. Quindi possiamo ridefinire l'osservazione come:

$$\omega \to x$$

Ovvero, l'osservazione ω viene mappata al valore x tramite la variabile X. Esistono diverse tipologie di variabili:

- **Quantitative:** assumono valori numerici e permettono operazioni aritmetiche.
- **Qualitative:** assumono valori categorici e non permettono operazioni aritmetiche.
- **Discrete:** assumono un numero finito o numerabile di valori.
- **Continue:** assumono un numero infinito di valori all'interno di un intervallo.
- **Scalare:** assumono un singolo valore per ogni osservazione.
- **Multi-dimensionali:** assumono più valori per ogni osservazione.

1.1.6 Scale di misurazione

Le scale di misurazione sono sistemi utilizzati per classificare e quantificare le variabili in base alle loro caratteristiche.

Definizione 1.6

Esistono diverse scale di misurazione:

- **Nominale:** classifica le osservazioni in categorie senza un ordine specifico.
- **Ordinale:** classifica le osservazioni in categorie con un ordine specifico, ma senza una distanza definita tra le categorie.
- **Intervallo:** misura le differenze tra le osservazioni, ma non ha un punto zero assoluto.
- **Ratio:** misura le differenze tra le osservazioni e ha un punto zero assoluto.

1.2 Organizzazione dei dati

1.2.1 Datasets

Un dataset è una raccolta strutturata di dati organizzati in righe e colonne, dove ogni riga rappresenta un'osservazione e ogni colonna rappresenta una variabile.

Definizione 1.7

1.2.2 Design matrix

Una design matrix è una rappresentazione tabellare dei dati in cui le righe corrispondono alle osservazioni e le colonne corrispondono alle variabili (o feature).

Definizione 1.8

Da questo possiamo notare che:

- Le Osservazioni sono rappresentate dalle righe del dataset.
- Le Variabili (o feature) sono rappresentate dalle colonne del dataset, catturate grazie alle osservazioni.
- I campioni e la popolazione, nel dataset, possono essere visti come una sottotabella del completo.

Valori nulli. In alcuni casi, i dati raccolti possono essere incompleti, con alcune osservazioni che mancano di valori per determinate variabili. Questi valori mancanti sono spesso indicati come "NA" (Not Available) o "null". Potrebbe essere dato da un errore durante la raccolta dati, ma in generale la gestione dei dati mancanti è importante nell'analisi dei dati, poiché può influenzare i risultati delle analisi statistiche e/o dei modelli di machine learning.

Outlier. In un dataset, un outlier è un'osservazione che si discosta significativamente dalle altre osservazioni, un valore "fuori scala". Gli outlier possono essere il risultato di errori di misurazione, errori di inserimento dati o possono rappresentare fenomeni rari ma validi.

1.3 Collezione dei dati

I dati possono essere ottenuti attraverso diverse fonti e metodi. Si parla del primo step per procedere all'analisi dei dati, in quanto la qualità e la rilevanza dei dati raccolti influenzano direttamente i risultati dell'analisi.

1.3.1 Surveys

I surveys sono strumenti di raccolta dati che consistono in questionari o interviste progettati per ottenere informazioni specifiche da un gruppo di persone.

Definizione 1.9

I surveys possono essere somministrati in vari modi, tra cui questionari cartacei, interviste telefoniche, sondaggi online o interviste faccia a faccia. La progettazione di un survey efficace richiede attenzione alla formulazione delle domande, alla selezione del campione e alla modalità di somministrazione per garantire la raccolta di dati accurati e rappresentativi.

Il problema dei survey è che spesso le risposte possono essere influenzate da bias (come il bias di desiderabilità sociale, dove i partecipanti rispondono in modo da apparire più favorevoli agli occhi degli altri, piuttosto che fornire risposte oneste), ma in generale non si parla di dati *altamente affidabili*.

1.3.2 Esperimenti

Gli esperimenti sono studi controllati in cui i ricercatori manipolano una o più variabili indipendenti per osservare l'effetto su una o più variabili dipendenti.

Definizione 1.10

Gli esperimenti possono essere condotti in laboratorio o sul campo e richiedono un'attenta progettazione per garantire che i risultati siano validi e affidabili. Gli esperimenti spesso includono gruppi di controllo e randomizzazione per minimizzare i bias e isolare gli effetti delle variabili manipolate.

Uno dei tipi di esperimenti più comuni è la tipologia RCT (Randomized Controlled Trial), in cui i partecipanti sono assegnati casualmente a gruppi sperimentali o di controllo.

Gli esperimenti hanno un difetto: sono costosi e richiedono tempo per essere condotti, ma in generale si parla di dati *molto affidabili*.

1.3.3 Dati osservabili

> *I dati osservabili sono dati raccolti attraverso l'osservazione diretta di fenomeni, eventi o comportamenti senza manipolazione o intervento da parte del ricercatore.*
>
> Definizione 1.11

Questi eventi sono utili quando gli esperimenti sono impraticabili, non etici oppure vanno fuori budget.

I dati osservabili hanno un difetto: possono essere influenzati da fattori esterni non controllati, ma in generale si parla di dati *affidabili*.

Fonti online

Nella data science moderna, una fonte sempre più comune di dati è rappresentata dalle fonti online. Questi dati possono essere raccolti da siti web, social media, database pubblici e altre piattaforme digitali.

1.4 Data Wrangling

> *Il data wrangling, o data munging, è il processo di pulizia, trasformazione e organizzazione dei dati grezzi in un formato utilizzabile per l'analisi.*
>
> Definizione 1.12

Come dicevamo nella sezione 1.2.2, i dati grezzi non sono mai perfettamente puliti: spesso contengono errori, valori mancanti, outlier e formati incoerenti che devono essere affrontati prima di procedere con l'analisi.

1.4.1 Gestire i valori nulli o mancanti

I valori nulli o mancanti possono essere gestiti in diversi modi, tra cui:

- Rimozione delle osservazioni con valori mancanti.
- Imputazione dei valori mancanti utilizzando la media, la mediana o la moda delle altre osservazioni.
- Utilizzo di modelli predittivi per stimare i valori mancanti basati su altre variabili.

1.4.2 Conversione dei dati

Alcune volte i dati vengono raccolti male, allora è necessario convertirli in formati più utili per l'analisi.

1.4.3 Rinominazione e riformattazione dei dati

Rinominare e formattare i dati in modo coerente può facilitare l'analisi e la comprensione del dataset.

1.4.4 Creazione di nuove feature

A volte è utile creare nuove feature basate su quelle esistenti per migliorare l'analisi.

1.4.5 Filtraggio e selezione dei dati

Filtrare e selezionare i dati rilevanti per l'analisi può migliorare l'efficienza e la precisione dei risultati.

1.5 Formato dei dati

I dati possono essere organizzati in formati differenti in base alle esisgenze dell'analisi.

1.5.1 Formato largo

Quando si parla di formato largo (wide format), i dati sono organizzati in modo tale che:

- Ogni variabile è rappresentata da una colonna separata.
- Ogni osservazione è rappresentata da una riga separata.

I vantaggi sono la facile lettura e la struttura intuitiva, ma può essere inefficiente per dataset con molte variabili o osservazioni. Si usa infatti, nell'analisi esplorativa dei dati e nella visualizzazione.

1.5.2 Formato lungo

Nel caso del formato lungo (long format), i dati sono organizzati in modo tale che:

- Le variabili sono rappresentate in una colonna separata.
- Le osservazioni sono rappresentate in più righe, con ogni riga che rappresenta una combinazione di osservazione e variabile.

I vantaggi sono l'efficienza nella memorizzazione e la facilità di manipolazione dei dati, ma può essere più difficile da leggere e interpretare. Si usa infatti, spesso nei modelli statistici e nelle analisi di serie temporali.

1.5.3 Altre tecniche di Wrangling

Esistono altre tecniche di data wrangling che possono essere utilizzate per preparare i dati per l'analisi, tra cui:

- Conversione dell'unità di misura
- Normalizzazione dei valori
- Aggregazione dei dati
- Riduzione della dimensionalità
- Rimuovere duplicati
- Estrarre informazione e parsing di stringhe

- Ricampionamento dei dati tra formati lunghi e larghi.

1.6 Il flusso di lavoro dell'Analisi dei dati

La definizione che viene data, all'analisi dei dati, è:

> *L'analisi dei dati è il processo di ispezione, pulizia, trasformazione e modellazione dei dati con l'obiettivo di scoprire informazioni utili, trarre conclusioni e supportare il processo decisionale.*
>
> Definizione 1.13

1.6.1 Passaggi fondamentali

Il flusso di lavoro tipico per l'analisi dei dati include i seguenti passaggi fondamentali:

Ispezione o Data exploration: Esplorare i dati per comprendere la loro struttura, qualità e caratteristiche principali.

Pulizia dei dati: Rimuovere o correggere errori, valori mancanti e outlier nei dati.

Trasformazione dei dati: Modificare i dati per renderli più adatti all'analisi, ad esempio creando nuove feature o convertendo i dati in formati diversi.

Modellazione dei dati: Applicare tecniche statistiche o di machine learning per analizzare i dati e trarre conclusioni.

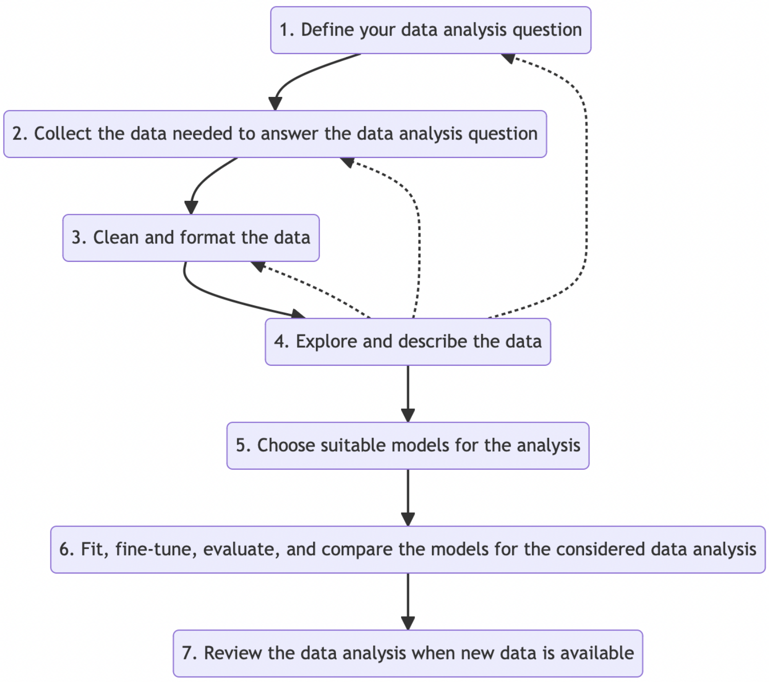

Figura 1.1: Flusso di lavoro dell'analisi dei dati.

Generalmente esiste un workflow iterativo e, ormai, standardizzato per l'analisi dei dati, che può essere riassunto nei seguenti passaggi:

1. Definire l'obiettivo dell'analisi.
2. Raccogliere i dati rilevanti.
3. Pulire e preparare i dati. Gestire i valori mancanti, rimuovere outlier, creare nuove feature, ecc.
4. Esplorare i dati. Utilizzare tecniche di visualizzazione e statistiche descrittive per comprendere la distribuzione dei dati, identificare pattern interessanti e relazioni tra le variabili.
5. Model lifecycle. Applicare un modello statistico o di machine learning per analizzare i dati.
6. Revisionare e aggiornare l'analisi. Basandosi sui risultati ottenuti, si potrebbe rivedere l'analisi, apportare modifiche ai dati o al modello e ripetere il processo se necessario.

Sebbene questo sembri un modello molto lineare e sequenziale, spesso non lo è. Infatti, si può rappresentare il flusso di lavoro dell'analisi dei dati come un ciclo iterativo, in cui si torna indietro a fasi precedenti in base ai risultati ottenuti e alle nuove informazioni scoperte durante l'analisi (come si vede in figura 1.1).

Capitolo 2

Visualizzazione e descrizione dei dati

Il problema principale dei grandi dataset di dati è che spesso sono difficili da interpretare direttamente. Per questo motivo, è necessario andare più a fondo per scoprire le strutture e i pattern che sono nascosti e non visibili semplicemente da una rapida stampa dei dati grezzi. La **statistica** aiuta in questo, in quanto fornisce uno strumento per riassumere e descrivere i dati in modo significativo.

2.1 Frequenze assolute

Un primo modo di descrivere i dati è quello di calcolare le **frequenze assolute**.

> *Siano $a_1, \dots, a_n$ i valori distinti di una variabile che stiamo considerando. La frequenza assoluta n_i allora si piuò definire come il numero di volte che il valore a_i appare nel dataset.*
>
> Definizione 2.1

Si noti che la somma delle frequenze assolute è uguale al numero N totale di osservazioni nel dataset:

$$\sum_i n_i = N$$

Nel nostro esempio del negozio, se consideriamo la variabile `Gender`, possiamo calcolare le frequenze assolute dei valori distinti:

- Maschio: 40
- Femmina: 60

Per rappresentare le frequenze assolute, possiamo utilizzare un **grafico a barre**. Un grafico a barre mostra le categorie della variabile sull'asse delle ascisse e le frequenze assolute sull'asse delle ordinate. Ogni barra rappresenta una categoria distinta, e l'altezza della barra corrisponde alla frequenza assoluta di quella categoria.

2.2 Frequenze relative

Il problema delle frequenze assolute è che non tengono conto della dimensione totale del dataset. Per questo motivo, spesso è più utile calcolare le **frequenze relative**.

La frequenza relativa f_i di un valore a_j si definisce come il rapporto tra la frequenza assoluta n_i e il numero totale di osservazioni N:

$$f_i = \frac{n_i}{N}, \qquad \forall j \in 1, ..., n$$

Definizione 2.2

Ovviamente, da questo seguono due cose:

1. Se $n_j \leq n \Rightarrow f_j \leq 1 \forall j$
2. La somma delle frequenze relative è uguale a 1:

$$\sum_j f_j = \sum_j \frac{n_j}{N} = \frac{1}{N} \sum_j n_j = \frac{N}{N} = 1$$

Anche qui è molto utile confrontare tutti gli elementi con un grafico a barre, l'unica cosa che cambia è che le frequenze hanno una scala diversa.

2.2.1 Grafici a barre

Un altro modo per contare il numero di elementi di una determinata classe è utilizzare un **grafico a barre**. In questo tipo di grafico, ogni barra rappresenta una categoria distinta della variabile e l'altezza della barra rappresenta la frequenza assoluta o relativa di quella categoria. I grafici a barre sono particolarmente utili per confrontare le frequenze tra diverse categorie.

Esiste una variante, il **grafico a barre impilato** in cui le barre sono suddivise in segmenti che rappresentano sottocategorie. Questo tipo di grafico è utile per visualizzare la composizione delle categorie principali.

2.2.2 Grafici a torta

Un altro tipo di visualizzazione molto comune per le frequenze relative è il **grafico a torta** (o *pie chart*). In questo tipo di grafico, un cerchio è suddiviso in spicchi, ciascuno dei quali rappresenta una categoria distinta della variabile. L'area di ogni spicchio è proporzionale alla frequenza relativa della categoria corrispondente. I grafici a torta sono utili per mostrare la proporzione delle categorie rispetto al totale.

Questi grafici funzionano bene quando ci sono poche classi, ma molto male negli altri casi. Non sono infatti adatti a mostrare differenze tra categorie simili.

2.3 ECDF - Empirical Cumulative Distribution Function

Un altro modo per visualizzare la distribuzione di una variabile è utilizzare la **funzione di distribuzione cumulativa empirica** (ECDF).

> *La ECDF di una variabile X è una funzione che, per ogni valore x, restituisce la proporzione di osservazioni nel dataset che sono minori o uguali a x.*
>
> Definizione 2.3

Matematicamente, la ECDF si può definire come:

$$F(x) = \frac{1}{N} \sum_{i=1}^{N} I(X_i \leq x)$$

dove I è la funzione indicatrice che vale 1 se la condizione è vera e 0 altrimenti.

La ECDF è utile perché fornisce una rappresentazione completa della distribuzione dei dati e consente di confrontare facilmente diverse distribuzioni.

2.4 Istogrammi

Un altro strumento utile per la visualizzazione dei dati continui è l'**istogramma**. Un istogramma è un grafico che rappresenta la distribuzione di una variabile continua suddividendo l'intervallo dei valori in **classi** (o **bin**) e contando il numero di osservazioni che cadono in ciascuna classe.

Per ogni istogramma però, si deve scegliere il numero di classi e la loro ampiezza. Questo numero può essere ottenuto in modo arbitrario, oppure utilizzando delle euristiche come la **regola di Sturges**:

$$k = \lceil \log_2(N) + 1 \rceil$$

dove k è il numero di classi e N è il numero totale di osservazioni nel dataset.

2.5 Stima di densità

Un problema degli istogrammi è che la loro forma dipende fortemente dalla scelta del numero di classi e dalla loro ampiezza. Si può utilizzare la **stima di densità** per ottenere una rappresentazione più fluida della distribuzione dei dati continui.

La stima di densità è una tecnica che cerca di risolvere questo problema ottenendo un'approssimazione continua della distribuzione dei dati.

2.6 Statistiche di sommario

Per descrivere i dati in modo più sintetico, si possono calcolare alcune **statistiche di sommario** che riassumono le caratteristiche principali della distribuzione dei dati.

2.6.1 Dimensione

La dimensione di un campione univariato $\{x_i\}_i^D$ è il numero di valori che contiene:

$$|\{x_i\}_i^D| = D$$

Definizione 2.4

Le varie dimensioni possono variare, in quanto ogni colonna può contenere valori nulli.

2.6.2 Misure di tendenza centrale

Le misure di tendenza centrale sono statistiche che descrivono il centro della distribuzione dei dati.

Media

La media di un campione univariato $\{x_i\}_i^D$ è definita come:

$$\bar{x} = \frac{1}{D} \sum_{i=1}^{D} x_i$$

Definizione 2.5

Da notare che la media è un valore facile da interpretare, ma non sempre affidabile in quanto è molto sensibile ai valori anomali (outlier).

Quantili

Per definire la mediana, è necessario prima definire altri valori statistici chiamati **quantili**.

Un quantile di ordine α di un campione univariato $\{x_i\}_i^D$ è il valore q_α tale che una frazione α delle osservazioni è minore o uguale a q_α:

$$q_\alpha = \inf\{x | F(x) \geq \alpha\}$$

dove $F(x)$ è la funzione di distribuzione cumulativa empirica (ECDF) del campione.

Definizione 2.6

Percentili. Dalla misura dei quantili, possiamo derivare i **percentili**, che sono quantili di ordine α espressi come percentuali. Ad esempio, il 25° percentile (o primo quartile) è il valore sotto il quale si trova il 25% delle osservazioni.

Quartili. I **quartili** sono quantili che dividono il dataset in quattro parti uguali. Il primo quartile (Q1) è il 25° percentile, il secondo quartile (Q2) è la mediana (50° percentile), e il terzo quartile (Q3) è il 75° percentile.

Mediana

Una volta definiti i quantili, è facile definire la mediana.

La mediana di un campione univariato $\{x_i\}_i^D$ è il quantile di ordine 0.5:

$$mediana = q_{0.5}$$

Definizione 2.7

Un'altro modo di definire la mediana, non utilizzando i quantili, per un certo campione ordinato $x_1 \leq x_2 \leq ... \leq x_n$ è:

$$\hat{x}_{0.5} = \begin{cases} x_{n+1}/2 & \text{se } n \text{ è dispari.} \\ \frac{x_{n/2}+x_{n/2+1}}{2} & \text{se } n \text{ è pari.} \end{cases}$$

Si nota che la mediana è meno sensibile ai valori anomali rispetto alla media, rendendola una misura più robusta della tendenza centrale in presenza di outlier.

Moda

La moda $\overline{x}_M$ di un campione univariato $\{x_i\}_i^D$ è il valore che appare con la massima frequenza nel campione.

Definizione 2.8

Formalizzando la definizione, si può scrivere che:

$$\overline{x}_M = a_j \Leftrightarrow n_j = \max\{n_1, ..., n_k\}$$

dove a_j sono i valori distinti della variabile e n_j le loro frequenze assolute.

2.7 Misure di dispersione

Per misure di dispersione si intendono delle statistiche che descrivono quanto i dati sono sparsi o concentrati attorno a una misura di tendenza centrale.

2.7.1 Minimo, Massimo

Il minimo x_{min} e il massimo x_{max} di un campione univariato $\{x_i\}_i^D$ sono definiti come:

$$x_{min} = \min\{x_i | i = 1, ..., D\}$$

$$x_{max} = \max\{x_i | i = 1, ..., D\}$$

Definizione 2.9

2.7.2 Intervallo

Dalla definizione di minimo e massimo, si può definire una misura di dispersione chiamata **intervallo**.

L'intervallo R di un campione univariato $\{x_i\}_i^D$ è definito come la differenza tra il massimo e il minimo:

$$R = x_{max} - x_{min}$$

Definizione 2.10

L'intervallo ha un problema, non è una misura molto robusta di dispersione in quanto dipende solo dai valori estremi del dataset, che possono essere influenzati da outlier.

2.7.3 Intervallo interquantile (IQR)

L'intervallo interquantile IQR di un campione univariato $\{x_i\}_i^D$ è definito come la differenza tra il terzo quartile $Q3$ e il primo quartile $Q1$:

$$IQR = Q3 - Q1$$

Definizione 2.11

Quindi, possiamo dedurre che la mediana si trova al centro dell'IQR. L'IQR è una misura più robusta di dispersione rispetto all'intervallo, in quanto non è influenzata dai valori estremi del dataset.

2.7.4 Varianza

La varianza di un campione univariato $\{x_i\}_i^D$ è definita come:

$$s^2 = \frac{1}{D-1}\sum_{i=1}^{D}(x_i - \bar{x})^2$$

dove $\bar{x}$ è la media del campione.

Definizione 2.12

La varianza misura la dispersione dei dati attorno alla media. Un valore di varianza più alto indica che i dati sono più sparsi, mentre un valore più basso indica che i dati sono più concentrati attorno alla media.

Il problema di questa misura di dispersione, è che è espressa nelle unità al quadrato della variabile originale, rendendo difficile l'interpretazione diretta. Inoltre è molto sensibile agli outlier.

2.7.5 Deviazione standard

La deviazione standard di un campione univariato $\{x_i\}_i^D$ è definita come la radice quadrata della varianza:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{D-1}\sum_{i=1}^{D}(x_i - \bar{x})^2}$$

Definizione 2.13

Si preferisce alla varianza, la deviazione standard perché risolve il problema delle unità al quadrato, essendo espressa nelle stesse unità della variabile originale. Tuttavia, rimane sensibile agli outlier.

2.8 Normalizzazione

Il problema di utilizzare queste statistiche di sommario è che non sono comparabili tra variabili con scale diverse. Per risolvere questo problema, si può utilizzare la **normalizzazione** dei dati. Questo problema si risolve con la normalizzazione e ne esistono diverse tipologie.

2.8.1 Min-Max Scaling

La normalizzazione Min-Max di un campione univariato $\{x_i\}_i^D$ è definita come:

$$x_{norm} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

dove x_{min} e x_{max} sono il minimo e il massimo del campione.

Definizione 2.14

Questa tecnica trasforma i dati in un intervallo compreso tra 0 e 1, risolvendo il problema della scala dei dati. Tuttavia è sempre sensibile agli outliers.

2.8.2 Normalizzazione tra -1 e 1

La normalizzazione tra -1 e 1 di un campione univariato $\{x_i\}_i^D$ è definita come:

$$x_{norm} = \frac{x_{\max} + x_{\min}}{2} + \frac{x_i - \frac{x_{\max}+x_{\min}}{2}}{\frac{x_{\max}-x_{\min}}{2}}$$

Definizione 2.15

Questa tecnica trasforma i dati in un intervallo compreso tra -1 e 1, risolvendo il problema della scala dei dati. Tuttavia, anche questa tecnica è sensibile agli outliers.

2.8.3 Standardizzazione (Z-score)

Molto spesso è utile normalizzare i dati in modo che abbiano media 0 e deviazione standard 1. Questo processo è chiamato **standardizzazione** o **Z-score normalization**.

La standardizzazione (Z-score) di un campione univariato $\{x_i\}_i^D$ *è definita come:*

$$z_i = \frac{x_i - \bar{x}}{s}$$

dove $\bar{x}$ *è la media del campione e* s *è la deviazione standard del campione.*

Definizione 2.16

Questa tecnica trasforma i dati in modo che abbiano media 0 e deviazione standard 1, rendendo le variabili comparabili tra loro. Inoltre, la standardizzazione è meno sensibile agli outliers rispetto alle altre tecniche di normalizzazione.

2.9 Indicatori di forma

Oltre alle misure di tendenza centrale e di dispersione, esistono anche degli indicatori di forma che descrivono la forma della distribuzione dei dati.

2.9.1 Asimmetria

L'asimmetria (skewness) di un campione univariato $\{x_i\}_i^D$ *è definita come:*

$$\frac{1}{D}\sum_{i=1}^{D}\left(\frac{x_i - \bar{x}}{s}\right)^3$$

dove $\bar{x}$ *è la media del campione e* s *è la deviazione standard del campione.*

Definizione 2.17

L'asimmetria misura la simmetria della distribuzione dei dati attorno alla media. Un valore di asimmetria positivo indica una distribuzione asimmetrica a destra, mentre un valore negativo indica una distribuzione asimmetrica a sinistra. Quanto un valore è più vicino allo zero, tanto più la distribuzione è simmetrica.

2.9.2 Curtosi

La curtosi (kurtosis) di un campione univariato $\{x_i\}_i^D$ *è definita come:*

$$\frac{1}{D}\sum_{i=1}^{D}\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$$

dove $\bar{x}$ *è la media del campione e* s *è la deviazione standard del campione.*

Definizione 2.18

La curtosi misura la "punta" della distribuzione dei dati. Un valore di curtosi positivo indica una distribuzione "leptocurtica" (più appuntita), mentre un valore negativo indica una distribuzione "platicurtica" (più piatta). Quanto il valore è più vicino allo zero, tanto più la distribuzione è simile a una distribuzione normale.

2.10 Statistiche descrittive

Tutte queste misure di tendenza centrale, dispersione e forma possono essere riassunte in una tabella chiamata **statistica descrittiva**. Questa tabella fornisce una panoramica completa delle caratteristiche principali della distribuzione dei dati e consente di confrontare facilmente diverse variabili.

Un grafico molto utile per rappresentare la statistica descrittiva è il **box plot** (o *diagramma a scatola*). In questo tipo di grafico, una scatola rappresenta l'intervallo interquantile (IQR) della variabile, con una linea all'interno della scatola che rappresenta la mediana. Le "whiskers" (linee esterne) si estendono fino ai valori minimo e massimo, escludendo gli outlier che sono rappresentati come punti separati. Il box plot è utile per visualizzare la distribuzione dei dati, identificare outlier e confrontare diverse variabili in modo rapido ed efficace.

Capitolo 3

Probabilità nell'analisi dei dati

Durante l'analisi dei dati, spesso ci si imbatte in situazioni in cui è necessario comprendere e quantificare l'incertezza associata ai dati raccolti. La probabilità fornisce un quadro teorico per affrontare queste situazioni, permettendo di fare inferenze basate su campioni di dati.

3.1 Esperimenti casuali

3.1.1 Spazio degli eventi

*Si definisce **spazio degli eventi** l'insieme di tutti i possibili risultati di un esperimento casuale:*

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$$

Definizione 3.1

Dallo spazio degli eventi[1], possiamo dare una seconda definizione di **evento semplice**.

*Un **evento semplice** è un sottoinsieme dello spazio degli eventi che contiene un solo elemento:*

$$A = \{\omega_i\}$$

Definizione 3.2

Un esempio di un evento semplice è l'estrazione di un asso da un mazzo di carte standard.

Molto spesso però, il risultato atteso è più grande di un singolo evento semplice. In questo caso, definiamo un **evento**.

*Un **evento** è un sottoinsieme dello spazio degli eventi che contiene più di un elemento:*

$$A = \{\omega_i, \omega_j, \ldots\} \subseteq \Omega$$

Definizione 3.3

[1] Notare che lo spazio degli eventi è denominato dallo Ω come la popolazione (sotto-sezione 1.1.2), non è un errore ma una convenzione (non sono la stessa cosa)

Spesso si denota $\overline{A} = \Omega \setminus A$ per indicare l'evento complementare di A. Da questa definizione possiamo trarre due semplici conclusioni:

- Lo spazio degli eventi Ω è un evento, in quanto contiene tutti gli eventi semplici. In particolare viene chiamato **evento certo**, in quanto sempre vero.
- L'insieme vuoto $\emptyset$ è un evento che non può mai avvenire, chiamato **evento impossibile**.

3.2 Variabili Aleatorie

Una ***variabile aleatoria*** *è una funzione che associa ad ogni evento semplice un numero reale:*

$$X : \Omega \to E$$

Dove E è uno spazio misurabile.

Definizione 3.4

3.3 Definizione dei dati

Nel contesto dell'analisi dei dati, i ***dati*** *sono considerati come realizzazioni di variabili aleatorie. Ogni osservazione raccolta durante un esperimento casuale può essere vista come un'istanza di una variabile aleatoria.*

Definizione 3.5

3.4 Probabilità

Poiché i risultati di una variabile aleatoria sono legati a eventi stocastici e non deterministici, è necessario definire una misura che quantifichi la probabilità di accadimento di tali eventi.

Una ***misura di probabilità*** *è una funzione P che assegna a ogni evento $A \subseteq \Omega$ un numero reale compreso tra 0 e 1.*

Definizione 3.6

3.4.1 Assiomi

La misura di probabilità, generalmente soddisfa alcuni assiomi:

- $P(\Omega) = 1$: La probabilità dell'evento certo è 1.
- $P(\emptyset) = 0$: La probabilità dell'evento impossibile è 0.
- Se $A_1, A_2, \ldots, A_n$ sono eventi mutuamente esclusivi, allora:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i)$$

3.4.2 Proprietà

Dalle definizioni e dagli assiomi sopra riportati, possiamo derivare alcune proprietà utili della misura di probabilità:

- **Complementarità**: $P(\overline{A}) = 1 - P(A)$
- **Monotonicità**: Se $A \subseteq B$, allora $P(A) \leq P(B)$
- **Additività**: Per due eventi qualsiasi A e B:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3.4.3 Probabilità di Laplace

In alcuni casi particolari, come negli esperimenti con esiti equiprobabili, possiamo utilizzare la definizione classica di probabilità proposta da Laplace.

*La **probabilità di Laplace** di un evento A è definita come il rapporto tra il numero di esiti favorevoli all'evento A e il numero totale di esiti possibili:*

$$P(A) = \frac{\textit{numero di esiti favorevoli ad } A}{\textit{numero totale di esiti possibili}}$$

Definizione 3.7

3.5 Stima di probabilità dalle osservazioni

Nell'analisi dei dati, spesso non conosciamo la distribuzione di probabilità sottostante. In questi casi, possiamo stimare la probabilità di un evento basandoci sulle osservazioni raccolte.

3.5.1 Approccio frequentista

L'approccio frequentista stima la probabilità di un evento A come il rapporto tra il numero di volte in cui si verifica l'evento e il numero totale di osservazioni effettuate:

$$P(A) \approx \frac{\text{numero di volte che si verifica } A}{\text{numero totale di osservazioni}}$$

Il problema di questo approccio è che la stima può essere imprecisa se il numero di osservazioni è limitato o se l'evento è raro.

3.5.2 Approccio bayesiano

L'approccio bayesiano utilizza il teorema di Bayes per aggiornare la stima della probabilità di un evento A.

Teorema di Bayes. Siano A e B due eventi con $P(B) > 0$. Allora la probabilità condizionata di A dato B è data da:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In questo contesto, $P(A)$ rappresenta la probabilità a priori dell'evento A, mentre $P(A|B)$ è la probabilità a posteriori di A dopo aver osservato l'evento B.

Questo approccio 'incorpora la probabilità a priori e aggiorna la stima in base alle nuove evidenze, risultando particolarmente utile in situazioni con dati limitati o in presenza di eventi rari. In questi contesti, infatti, vediamo la probabilità come una stima di *incertezza* piuttosto che una frequenza oggettiva.

Un esempio può essere la diagnosi medica, dove la probabilità di una malattia può essere aggiornata in base ai sintomi osservati e ai risultati dei test diagnostici.

3.6 Probabilità congiunta

In molti casi, è utile considerare la probabilità di due o più eventi che si verificano simultaneamente. Questa è nota come **probabilità congiunta**.

*La **probabilità congiunta** di due eventi A e B è definita come la probabilità che entrambi gli eventi si verifichino.*

Definizione 3.8

Un esempio comune è la probabilità che un paziente abbia sia la febbre che un'infezione batterica.

N.B. La probabilità congiunta non è una probabilità condizionata.

3.6.1 Regola della somma

Un modo di calcolare la probabilità congiunta è attraverso le tabelle di contingenza, che mostrano la frequenza con cui si verificano combinazioni di eventi.

Un altro modo è sfruttare un approccio frequentista, contando le occorrenze congiunte degli eventi nei dati osservati.

Si fa questo considerando una tabella di contingenza, ovvero una tabella che riporta le frequenze con cui si verificano le combinazioni di due o più variabili categoriali. Ad esempio, supponiamo di avere due eventi A e B con due possibili esiti ciascuno (vero/falso). La tabella di contingenza potrebbe essere strutturata come segue:

oprule	$Y = y_1$	$Y = y_2$	$\dots$	$Y = y_l$	Total
$X = x_1$	n_{11}	n_{12}	$\dots$	n_{1l}	n_{1+}
$X = x_2$	n_{21}	n_{22}	$\dots$	n_{2l}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X = x_k$	n_{k1}	n_{k2}	$\dots$	n_{kl}	n_{k+}
Total	n_{+1}	n_{+2}	$\dots$	n_{+l}	n

Tabella 3.1: Esempio generale di tabella di contingenza

Dove:

- n_{ij} rappresenta il numero di osservazioni in cui l'evento $X = x_i$ e l'evento $Y = y_j$ si verificano contemporaneamente.
- n_{i+} è il totale delle osservazioni per l'evento $X = x_i$.
- n_{+j} è il totale delle osservazioni per l'evento $Y = y_j$.
- n è il numero totale di osservazioni.

Da questo, possiamo calcolare la probabilità congiunta $P(X = x_i, Y = y_j)$ come:

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{n}$$

Dalla tabella 3.1, la probabilità marginale è definita come:

$$P(X = x_i) = \frac{n_{i+}}{n}$$

$$P(Y = y_j) = \frac{n_{+j}}{n}$$

Definizione 3.9

Da qui, possiamo definire la probabilità marginale di un evento X come la somma delle probabilità congiunte su tutti i possibili esiti dell'altro evento Y:

$$P(X = x_i) = \frac{n_{i+}}{n} = \frac{\sum_j n_{ij}}{n} = \sum_j \frac{n_{ij}}{n} = \sum_j P(X = x_i, Y = y_j)$$

Questo risultato è conosciuto come **regola della somma**, che ci permette di calcolare la probabilità marginale di un evento sommando le probabilità congiunte con tutti gli esiti dell'altro evento (marginalizzazione).

3.7 Probabilità condizionata

In molte situazioni, è utile conoscere la probabilità di un evento dato che un altro evento si è verificato. Questa è nota come **probabilità condizionata**.

La ***probabilità condizionata*** *di un evento A dato che un evento B si è verificato è definita come:*

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

Definizione 3.10

Possiamo utilizzare di nuovo la tabella di contingenza per calcolare la probabilità condizionata. Ad esempio, la probabilità condizionata di $X = x_i$ dato $Y = y_j$ è:

$$P(X = x_i|Y = y_j) = \frac{\#\text{casi dove } X = x_i \text{ e } Y = y_j}{\#\text{casi dove } Y = y_j} = \frac{n_{ij}}{n} \cdot \frac{n}{n_{+j}} = \frac{n_{ij}}{n_{+j}} = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

N.B. La probabilità condizionata è definita solo se $P(Y) > 0$, anche perché non si può definire la probabilità di un evento dato che un evento impossibile si è verificato: sarebbe un controsenso.

3.7.1 Regola del prodotto

Un modo alternativo di calcolare la probabilità congiunta è attraverso la regola del prodotto, che sfrutta la probabilità condizionata.

Possiamo dire, che la probabilità condizionata è:

$$P(X = x, Y = y) = \frac{P(X = x|Y = y)}{P(Y = y)}$$

Da qui segue:

$$P(X = x, Y = y) = P(X = x|Y = y) \cdot P(Y = y)$$

La probabilità del prodotto ci permette di calcolare la probabilità congiunta di due eventi moltiplicando la probabilità condizionata di uno degli eventi per la probabilità dell'altro evento.

3.8 Regola della catena per le probabilità condizionate

Quando si lavora con più variabili, la regola del prodotto può essere estesa per calcolare la probabilità condizionata di una variabile dato un insieme di altre variabili. Questa estensione è nota come **regola della catena**.

Sapendo che:

$$P(X,Y,Z) = P(X|Y,Z) \cdot P(Y,Z)$$

Possiamo espandere ulteriormente $P(Y,Z)$ usando la regola del prodotto:

$$P(Y,Z) = P(Y|Z) \cdot P(Z)$$

Sostituendo questa espressione nella precedente, otteniamo:

$$P(X,Y,Z) = P(X|Y,Z) \cdot P(Y|Z) \cdot P(Z)$$

Questa è la regola della catena per tre variabili. In generale, per n variabili $X_1, X_2, \ldots, X_n$, la regola della catena si estende come segue:

$$P(X_1, X_2, \ldots, X_n) = P(X_1) \prod_{i=2}^{n} P(X_i | X_1, X_2, \ldots, X_{i-1})$$

3.9 Indipendenza

Due variabili aleatorie X e Y sono dette **indipendenti** se la conoscenza del valore di una non fornisce alcuna informazione sul valore dell'altra. In termini di probabilità, questo si traduce nella seguente condizione:

$$P(X|Y) = P(X) \cdot P(Y)$$

L'indipendenza si denota come:

$$X \perp Y$$

Se due variabili sono indipendenti, allora la probabilità condizionata di una variabile dato l'altra è semplicemente la probabilità marginale della prima variabile:

$$P(X|Y) = \frac{P(X, Y)}{P(Y)} = \frac{P(X) \cdot P(Y)}{P(Y)} = P(X)$$

3.9.1 Indipendenza condizionata

Due variabili aleatorie X e Y sono dette indipendenti condizionalmente rispetto a una terza variabile Z se, dato il valore di Z, la conoscenza del valore di X non fornisce alcuna informazione sul valore di Y, e viceversa. In termini di probabilità, questo si traduce nella seguente condizione:

$$P(X, Y|Z) = P(X|Z) \cdot P(Y|Z)$$

E quindi, si può denotare come:

$$X \perp Y | Z$$

Capitolo 4

Associazioni di variabili

Spesso, visualizzare statistiche su un campione di dati univariato non è sufficiente. Ad esempio, potremmo voler esaminare la relazione tra due variabili quantitative, come l'altezza e il peso di un gruppo di persone, si parla quindi di come due variabili sono associate tra loro.

> *Si parla di associazione tra due variabile se, osservando i valori assunti da una variabile, è possibile fare delle previsioni sui valori assunti dall'altra variabile.*
>
> Definizione 4.1

4.1 Misure di associazioni tra variabili discrete

4.1.1 Indipendenza

Prima di parlare di misure di associazione tra variabili discrete, è importante definire quando due variabili discrete sono **indipendenti**. Si parla di indipendenza, quando la conoscenza di una variabile non fornisce alcuna informazione sull'altra variabile.

La tabella di contingenza (tabella 3.1) può spiegare bene il concetto di indipendenza tra due variabili discrete: tutti quei valori, che sono marginali (cioè che riguardano una sola variabile, indicati con n_{i+} oppure n_{+j}) non forniscono alcuna informazione sui valori congiunti (cioè che riguardano entrambe le variabili, indicati con n_{ij}). Ipotizziamo quindi, di non conoscere i valori delle frequenze congiunte f_{ij}, ma solo le frequenze marginali $f_{i\cdot}$ e $f_{\cdot j}$:

oprule	$Y = y_1$	$Y = y_2$	$\dots$	$Y = y_l$	Total
$X = x_1$	—	—	$\dots$	—	n_{1+}
$X = x_2$	—	—	$\dots$	—	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$X = x_k$	—	—	$\dots$	—	n_{k+}
Total	n_{+1}	n_{+2}	$\dots$	n_{+l}	n

Tabella 4.1: Tabella di contingenza (solo marginali)

Se dovessimo ricostruire i valori mancanti e le variabili sono indipendenti, potremmo dire:

$$P(X = x_i, Y = y_i) = P(X = x_i)P(Y = y_i)$$

Ricordando che:

$$P(X = x_i) = \frac{n_{i+}}{n} \quad , \quad P(Y = y_i) = \frac{n_{+i}}{n} \quad , \quad P(X = x_i, Y = y_i) = \frac{n_{ij}}{n}$$

Possiamo ricostruire le frequenze congiunte mancanti come:

$$\hat{n}_{ij} = n \cdot P(X = x_i, Y = y_i) = n \cdot P(X = x_i)P(Y = y_i) = n \cdot \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}$$

Da cui possiamo scrivere:

$$\hat{n}_{ij} = P(X = x_i)n_{+j} = P(Y = y_j)n_{i+}$$

N.B.: utilizziamo $\hat{n}_{ij}$ (e non n_{ij})per indicare che si tratta di una stima della frequenza congiunta, calcolata assumendo l'indipendenza tra le variabili.

Questo ci dice, che date due variabili indipendenti, la frequenza congiunta stimata può essere calcolata a partire dalle frequenze marginali.

4.1.2 Statistica di Pearson

La statistica di Pearson, chiamata χ^2, misura la discrepanza tra le frequenze congiunte osservate e quelle attese sotto l'ipotesi di indipendenza. Segue la formula:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

Dove k, l sono rispettivamente il numero di categorie delle variabili X e Y.

Questa statistica calcola il quadrato delle differenze (discrepanza) e lo normalizza rispetto alle frequenze attese, in modo da tenere conto delle diverse dimensioni delle categorie. Un valore elevato di χ^2 indica una maggiore discrepanza tra le frequenze osservate e quelle attese, suggerendo una possibile associazione tra le variabili. Al contrario, un valore basso suggerisce che le variabili sono probabilmente indipendenti.

Questa statistica tuttavia, ha dei problemi. Il valore atteso per ogni cella deve essere sufficientemente grande (almeno 5) per garantire l'accuratezza dell'approssimazione della distribuzione χ^2. Inoltre, il valore di χ^2 dipende dalla dimensione del campione: campioni più grandi tendono a produrre valori più elevati di χ^2, anche per associazioni deboli.

4.1.3 Statistica di Cramér

Per superare i limiti della statistica di Pearson, si può utilizzare la statistica di Cramér, che normalizza il valore di χ^2 per tener conto della dimensione del campione e del numero di categorie delle variabili. La formula è la seguente:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, l-1)}}$$

Dove n è la dimensione del campione, e k, l sono rispettivamente il numero di categorie delle variabili X e Y.

Questa statistica ritorna un valore tra 0 e 1, dove 0 indica nessuna associazione tra le variabili e 1 indica un'associazione perfetta. Essendo normalizzata, la statistica di Cramér è meno influenzata dalla dimensione del campione rispetto a χ^2, rendendola una misura più robusta dell'associazione tra variabili discrete.

4.1.4 Rischio relativo

Il rischio relativo (RR) è una misura utilizzata per quantificare l'associazione tra due variabili discrete, spesso in studi epidemiologici. Esso confronta la probabilità di un evento (ad esempio, una malattia) tra due gruppi distinti (ad esempio, esposti e non esposti a un fattore di rischio). La formula per calcolare il rischio relativo è la seguente:

$$RR = \frac{P(E|A)}{P(E|\neg A)}$$

Dove $P(E|A)$ è la probabilità dell'evento E dato l'esposizione A, e $P(E|\neg A)$ è la probabilità dell'evento E dato la non esposizione $\neg A$.

Per fare un esempio, consideriamo uno studio che esamina l'associazione tra il fumo (esposizione) e lo sviluppo di una malattia polmonare (evento). Supponiamo di avere i seguenti dati:

- Numero di fumatori che sviluppano la malattia: a
- Numero di fumatori che non sviluppano la malattia: b
- Numero di non fumatori che sviluppano la malattia: c
- Numero di non fumatori che non sviluppano la malattia: d

La probabilità di sviluppare la malattia tra i fumatori è:

$$P(E|A) = \frac{a}{a+b}$$

La probabilità di sviluppare la malattia tra i non fumatori è:

$$P(E|\neg A) = \frac{c}{c+d}$$

Pertanto, il rischio relativo è dato da:

$$RR = \frac{P(E|A)}{P(E|\neg A)} = \frac{a/(a+b)}{c/(c+d)}$$

Un valore di RR maggiore di 1 indica che l'esposizione è associata a un aumento del rischio dell'evento, mentre un valore inferiore a 1 indica una riduzione del rischio. Un valore di RR uguale a 1 suggerisce che non vi è alcuna associazione tra l'esposizione e l'evento.

4.1.5 Odds Ratio

L'Odds Ratio (OR) è una misura di associazione utilizzata per quantificare la forza della relazione tra due variabili discrete, spesso in studi caso-controllo. L'OR confronta le probabilità (odds) di un evento tra due gruppi distinti. La formula per calcolare l'Odds Ratio è la seguente:

$$OR = \frac{P(E|A)/P(\neg E|A)}{P(E|\neg A)/P(\neg E|\neg A)}$$

Per fare lo stesso esempio precedente, consideriamo i seguenti dati:

- Numero di fumatori che sviluppano la malattia: a
- Numero di fumatori che non sviluppano la malattia: b
- Numero di non fumatori che sviluppano la malattia: c
- Numero di non fumatori che non sviluppano la malattia: d

La probabilità di sviluppare la malattia tra i fumatori è:

$$P(E|A) = \frac{a}{a+b}$$

La probabilità di non sviluppare la malattia tra i fumatori è:

$$P(\neg E|A) = \frac{b}{a+b}$$

La probabilità di sviluppare la malattia tra i non fumatori è:

$$P(E|\neg A) = \frac{c}{c+d}$$

Quindi l'Odds Ratio è dato da:

$$OR = \frac{(a/b)}{(c/d)} = \frac{a \cdot d}{b \cdot c}$$

Un valore di OR maggiore di 1 indica che l'esposizione è associata a un aumento delle probabilità dell'evento, mentre un valore inferiore a 1 indica una riduzione delle probabilità. Un valore di OR uguale a 1 suggerisce che non vi è alcuna associazione tra l'esposizione e l'evento.

4.2 Misure di associazioni tra variabili continue

Il problema delle variabili continue, rispetto a quelle discrete, è che non possiamo costruire una tabella di contingenza. Per questo motivo, utilizziamo misure di associazione basate sulla covarianza e sulla correlazione.

4.2.1 Visualizzazione grafica dell'associazione

Esistono modi grafici di visualizzare una possibile associazione tra due variabili continue, come lo **scatter plot** (sezione **??**) che mostra i punti dati in un piano cartesiano, permettendo di osservare visivamente la relazione tra le due variabili. Si può creare una matrice di scatter plot per visualizzare le relazioni tra più variabili continue contemporaneamente, chiamata **scatter matrix** (sotto-sezione **??**).

Un altro modo è utilizzare gli **hexbin plots** (sezione **??**), ovvero una forma di istogramma a due dimensioni. In questo tipo di grafico, lo spazio bidimensionale viene suddiviso in esagoni (hexbin) e il colore di ciascun esagono rappresenta la densità dei punti dati in quella regione. Questo è particolarmente utile per visualizzare grandi quantità di dati, poiché riduce il sovraffollamento e rende più facile identificare le aree con alta concentrazione di punti.

Un ulteriore modo per visualizzare l'associazione tra due variabili continue è attraverso i **grafici di densità e di contorno** (sezione **??**). Questi grafici mostrano la distribuzione congiunta delle due variabili, evidenziando le aree di maggiore densità dei dati. I grafici di contorno, in particolare, utilizzano linee per collegare punti con la stessa densità, facilitando l'identificazione di pattern e relazioni tra le variabili.

4.2.2 Covarianza

La varianza misura la dispersione di una singola variabile rispetto alla sua media. La covarianza estende questo concetto a due variabili.

La covarianza misura come due variabili X, Y variano insieme rispetto alle loro medie $\bar{x}, \bar{y}$. Si definisce come:

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

Definizione 4.2

Esiste una versione campionaria della covarianza, che utilizza $N-1$ al denominatore invece di N, per correggere il bias nella stima della covarianza dalla popolazione (molto importante per piccoli campioni):

$$Cov_{campione}(X,Y) = \frac{1}{N-1}\sum_{i=1}^{N}(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$$

Notiamo che, in base al valore risultante da $(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})$, possiamo avere tre casi:

- Se il valore è positivo, significa che entrambe le variabili si discostano dalla loro media nello stesso verso (entrambe sopra o entrambe sotto la media), contribuendo positivamente alla covarianza.

- Se il valore è negativo, significa che le variabili si discostano dalla loro media in direzioni opposte (una sopra e l'altra sotto la media), contribuendo negativamente alla covarianza.
- Quanto il valore è più vicino allo zero, tanto meno le due variabili sono associate tra loro.

In un grafico cartesiano, i punti dati possono essere distribuiti in quattro quadranti:

- Primo quadrante (entrambe le variabili sopra la media): contribuisce positivamente alla covarianza.
- Secondo quadrante (una variabile sopra la media e l'altra sotto): contribuisce negativamente alla covarianza.
- Terzo quadrante (entrambe le variabili sotto la media): contribuisce positivamente alla covarianza.
- Quarto quadrante (una variabile sotto la media e l'altra sopra): contribuisce negativamente alla covarianza.

Si noti che, la covarianza di due variabili uguali, è uguale alla varianza:

$$Cov(X, X) = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \bar{x})^2 = var$$

Il problema principale della covarianza è che il suo valore dipende dalle unità di misura delle variabili. Ad esempio, se una variabile è misurata in metri e l'altra in chilogrammi, la covarianza avrà unità di misura miste (metri · chilogrammi), rendendo difficile l'interpretazione del valore.

4.2.3 Coefficiente di correlazione di Pearson

Per superare il problema delle unità di misura nella covarianza, si utilizza il coefficiente di correlazione di Pearson, che normalizza la covarianza dividendo per il prodotto delle deviazioni standard delle due variabili. La formula è la seguente:

$$p(X, Y) = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} = \frac{1}{N} \sum_{i=1}^{N} \frac{(x^{(i)} - \bar{x})}{\sigma_X} \cdot \frac{(y^{(i)} - \bar{y})}{\sigma_Y} = Cov(z(X), z(Y))$$

Dove σ_X e σ_Y sono le deviazioni standard delle variabili X e Y rispettivamente.

Il coefficiente di correlazione di Pearson varia tra -1 e 1, poiché è stato utilizzata la standardizzazione (z-scoring) z per eliminare le unità di misura:

$$z(X) = \frac{X - \bar{X}}{\sigma_X}$$

Un valore di 1 indica una correlazione positiva perfetta, $p(x, x)$, mentre un valore di -1 indica una correlazione negativa perfetta, $p(x, -x)$. Quanto un valore è più vicino a 0, tanto meno le due variabili sono associate tra loro. Si può anche interpretare il segno: un valore positivo indica che le variabili tendono a variare nella stessa direzione, mentre un valore negativo indica che tendono a variare in direzioni opposte.

Coefficiente di correlazione di Pearson per variabili discrete. Anche per le variabili discrete, si può calcolare il coefficiente di correlazione di Pearson, trattando le categorie come valori numerici, ma non ha senso farlo sulle variabili nominali (senza ordine intrinseco). Per le variabili ordinali, invece, si può assegnare un ordine numerico alle categorie e calcolare il coefficiente di correlazione di Pearson sui ranghi delle categorie.

Si noti tuttavia, che se calcolato nel caso specifico di una variabile continua e una dicotomica (0/1), il coefficiente di correlazione di Pearson è equivalente a un valore chiamato **punto-biseriale**.

Alla luce di tutto questo, il principale problema di questo coefficiente di correlazione, è che misura solo relazioni lineari tra le variabili. Se la relazione tra le variabili è non lineare, il coefficiente di correlazione di Pearson potrebbe non catturare adeguatamente l'associazione tra di esse.

4.2.4 Coefficiente di correlazione di Spearman

Il coefficiente di correlazione di Spearman è una misura non parametrica della correlazione tra due variabili. A differenza del coefficiente di Pearson, che si basa sui valori effettivi delle variabili, il coefficiente di Spearman si basa sui ranghi dei valori. Per rango si intende la posizione di un valore all'interno di un insieme ordinato di valori e l'intuizione alla base di questa misura è che, se due variabili sono correlate in modo monotono (cioè una variabile tende ad aumentare quando l'altra aumenta, indipendentemente dalla forma della relazione), allora i loro ranghi dovrebbero essere correlati. La formula per calcolare il coefficiente di correlazione di Spearman, siano $R(x^{(i)}), R(y^{(i)})$ i ranghi associati alle variabili X, Y, è la seguente:

$$R = 1 - \frac{6\sum_{i=1}^{N}(R(x^{(i)}) - R(y^{(i)}))^2}{N(N^2-1)}$$

Dove N è il numero di coppie di dati.

Questo risultato è normalizzato in $[-1, 1]$ e può essere interpretato in modo simile al coefficiente di Pearson: un valore di 1 indica una correlazione positiva perfetta tra i ranghi, -1 indica una correlazione negativa perfetta, e 0 indica nessuna correlazione tra i ranghi.

4.2.5 Coefficiente di correlazione di Kendall

Il coefficiente di correlazione di Kendall, noto anche come τ di Kendall, è un'altra misura non parametrica della correlazione tra due variabili basata sui ranghi. A differenza del coefficiente di Spearman, che si basa sulle differenze tra i ranghi, il coefficiente di Kendall si basa sul concetto di coppie concordanti e discordanti. Una coppia di osservazioni $(x^{(i)}, y^{(i)})$ e $(x^{(j)}, y^{(j)})$ è considerata concordante se l'ordine dei ranghi è lo stesso per entrambe le variabili:

$$(x^{(i)} - x^{(j)})(y^{(i)} - y^{(j)}) > 0$$

Al contrario, è considerata discordante se l'ordine dei ranghi è opposto:

$$(x^{(i)} - x^{(j)})(y^{(i)} - y^{(j)}) < 0$$

Il coefficiente τ di Kendall è calcolato come:

$$\tau = \frac{(C - D)}{\frac{1}{2}N(N-1)}$$

Dove C è il numero di coppie concordanti, D è il numero di coppie discordanti, e N è il numero totale di osservazioni.

Questo risultato è normalizzato in $[-1, 1]$ e può essere interpretato in modo simile al coefficiente di Pearson: un valore di 1 indica una correlazione positiva perfetta tra i ranghi, -1 indica una correlazione negativa perfetta, e 0 indica nessuna correlazione tra i ranghi.

Matrice di correlazione. Una matrice di correlazione è una tabella che mostra i coefficienti di correlazione tra più variabili (sezione **??**). Ogni cella della matrice rappresenta il coefficiente di correlazione tra due variabili specifiche. La matrice è simmetrica, poiché il coefficiente di correlazione tra la variabile X e la variabile Y è lo stesso del coefficiente tra Y e X. La diagonale principale della matrice contiene sempre il valore 1, poiché ogni variabile è perfettamente correlata con se stessa.

Da questa matrice, può nascere una heatmap (sezione **??**): una rappresentazione grafica della matrice di correlazione, in cui i valori dei coefficienti di correlazione sono rappresentati da colori. Le heatmap facilitano l'identificazione visiva delle relazioni tra le variabili, evidenziando rapidamente quali coppie di variabili sono fortemente correlate (positive o negative) e quali non lo sono.

4.3 Consigli utili su come scegliere la misura di associazione

Di seguito alcuni suggerimenti pratici, organizzati per tipo di variabili e obiettivo dell'analisi.

Discrete - Discrete: se entrambe le variabili sono categoriali usare tabelle di contingenza, statistica di Pearson (χ^2) per testare indipendenza, e Cramér's V per ottenere una misura normalizzata dell'intensità dell'associazione. Per studi caso-controllo o epidemiologici valutare anche Odds Ratio o Rischio Relativo.

Continuous - Continuous: valutare prima la relazione visiva con uno scatter plot. Usare il coefficiente di Pearson per relazioni lineari (assunzione: approssimativamente normale e assenza di outlier influenti). Se la relazione è monotona ma non lineare usare Spearman o Kendall (ranghi) che sono più robusti agli outlier.

Ordinal - Ordinal: i coefficienti basati sui ranghi (Spearman, Kendall) sono i più appropriati perché rispettano l'ordine senza imporre distanze metriche arbitrarie.

Più variabili / esplorazione: utilizzare matrici di correlazione (per variabili continue), heatmap, scatter-matrix e analisi di regressione (lineare o non lineare) per stimare l'effetto condizionato di una variabile su un'altra mentre si controlla per confondenti.

Capitolo 5

Distribuzione dei dati

Il problema dei dati è che non sempre la singola probabilità di un evento è sufficiente a descrivere il fenomeno che stiamo studiando. Spesso infatti è necessario considerare l'insieme delle possibili occorrenze di un evento.

5.1 Distribuzione di probabilità

La distribuzione di probabilità di una variabile casuale descrive come le probabilità sono distribuite tra i possibili valori che la variabile può assumere.

Definizione 5.1

Nel caso di variabili casuali discrete, la distribuzione di probabilità viene denominata **funzione di massa di probabilità** (pmf, *probability mass function*), mentre nel caso di variabili casuali continue viene chiamata **funzione di densità di probabilità** (pdf, *probability density function*).

Una distribuzione di probabilità caratterizza completamente una variabile casuale, in quanto fornisce tutte le informazioni necessarie per calcolare le probabilità di qualsiasi evento associato a quella variabile. Denotiamo con $P(X)$ la distribuzione di probabilità della variabile casuale X e possiamo scrivere che "X **segue la distribuzione** $P(X)$" come:

$$X \sim P(X)$$

5.2 Distribuzioni discrete

Una distribuzione di probabilità discreta è utilizzata per variabili casuali che possono assumere solo un numero finito o numerabile di valori.

5.2.1 Funzione di massa di probabilità (PMF)

La **funzione di massa di probabilità** (PMF) è la funzione che descrive la probabilità di ciascun valore possibile.

Una PMF sulla variabile casuale X è definita come:

$$P : \Omega \to [0, 1]$$

La PMF deve soddisfare una proprietà:

$$\sum_{x \in \Omega} P(x) = 1$$

Ipotizzando di simulare il lancio di un dado a sei facce, la PMF associata a questo esperimento è:

$$P(X = x) = \begin{cases} \frac{1}{6} & \text{se } x \in \{1, 2, 3, 4, 5, 6\} \\ 0 & \text{altrimenti} \end{cases}$$

Questo significa che ogni faccia del dado ha una probabilità di $\frac{1}{6}$ di essere estratta, e la somma delle probabilità di tutte le facce è uguale a 1, come richiesto dalla proprietà della PMF.

5.2.2 Funzione di distribuzione cumulativa (CDF)

Se consideriamo il caso dove la variabile casuale X può essere ordinata (come nel caso dei numeri interi), possiamo definire la **funzione di distribuzione cumulativa** (CDF, *cumulative distribution function*) come:

$$F(x) = P(X \leq x) = \sum_{y \leq x} P(X = y)$$

La CDF fornisce la probabilità cumulativa che la variabile casuale X assuma un valore minore o uguale a x (per esempio, tutte le persone con altezza minore o uguale a 170 cm).

5.3 Distribuzioni continue

Una distribuzione di probabilità continua è utilizzata per variabili casuali che possono assumere un infinito numero di valori all'interno di un intervallo. Tuttavia, per poter capire il concetto di distribuzione continua, bisogna prima introdurre il concetto di funzione di densità di probabilità (PDF).

5.3.1 Funzione di densità di probabilità (PDF)

La **funzione di densità di probabilità** (PDF) è la funzione che descrive la densità di probabilità in ogni punto dell'intervallo. Una PDF sulla variabile casuale X è definita come:

$$f : \Omega \to [0, 1]$$

E deve soddisfare la seguente proprietà:

$$\int f(x)\, dx = 1$$

(Questa condizione è equivalente alla somma delle probabilità nella distribuzione discreta).

Si può utilizzare la PDF per calcolare la probabilità che la variabile casuale X assuma un valore all'interno di un intervallo specifico $[a, b]$:

$$P(a \leq X \leq b) = \int_a^b f(x)\, dx$$

La pdf descrive la densità di probabilità in ogni punto, ma non fornisce direttamente la probabilità di un singolo punto, poiché in una distribuzione continua la probabilità di un singolo punto è sempre zero: si parla infatti di densità di probabilità e non di probabilità puntuale. Tanto è più grande il valore della pdf in un punto, tanto più alta è la probabilità che la variabile casuale assuma valori vicini a quel punto.

PDF uniforme. Un esempio di PDF è la distribuzione uniforme continua su un intervallo $[a, b]$, dove la PDF è definita come:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{se } a \leq x \leq b \\ 0 & \text{altrimenti} \end{cases}$$

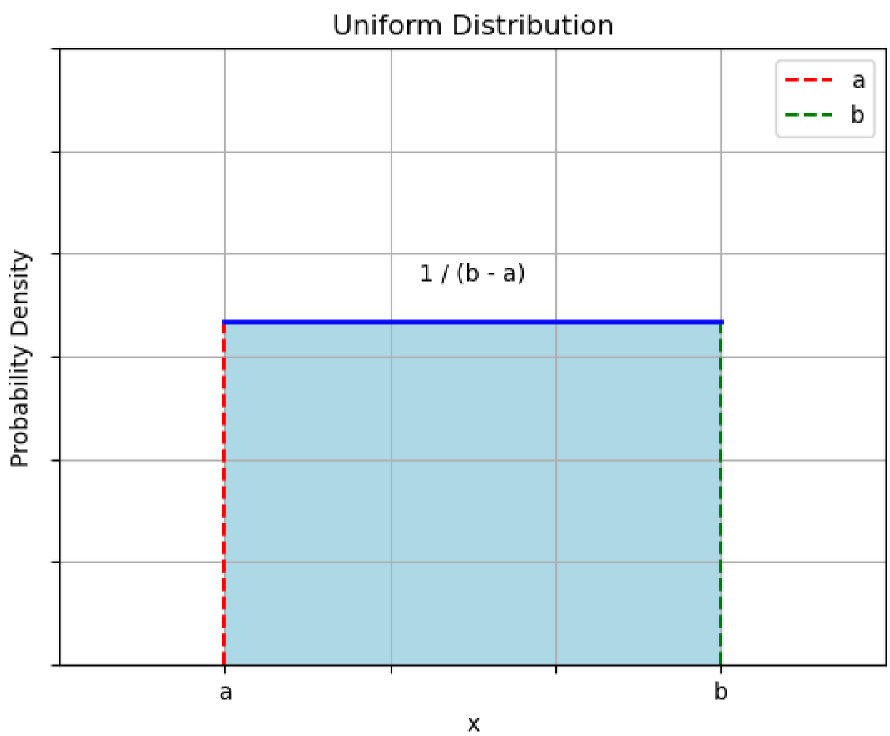

Figura 5.1: Esempio di PDF uniforme continua su un intervallo $[a, b]$. La funzione è costante all'interno dell'intervallo e zero al di fuori.

Approssimare la PDF con istogrammi. In pratica, quando si lavora con dati continui, spesso si cerca di approssimare la PDF utilizzando istogrammi (sezione **??**).

Per un certo bin b_j sia h_j l'altezza dell'istogramma e sia w_j la larghezza del bin. Allora possiamo dire, per un istogramma normalizzato[1], che l'altezza del bin è:

$$h_j = \frac{c_j}{n}$$

dove c_j è il conteggio degli elementi nel bin b_j e n è il numero totale di elementi.
Da questo, possiamo approssimare l'area dell'istogramma nel bin b_j come:

$$A_j = h_j \cdot w_j = \frac{c_j}{n} \cdot w_j$$

Calcolata quest'area per un j-esimo bin, possiamo calcolare l'area dell'istogramma per intero come:

$$\int_{-\infty}^{\infty} H(X)\,dx = \sum_j A_j = \sum_j \frac{c_j}{n} \cdot w_j = \sum_j w_j \neq 1$$

Per soddisfare la proprietà di una PDF, dobbiamo normalizzare l'istogramma in modo che l'area totale sia pari a 1. Per fare ciò, possiamo dividere l'altezza di ogni bin per l'area totale dell'istogramma:

$$h'_j = \frac{c_j}{n \cdot \sum_j w_j}$$

e da qui riscrivere l'area del bin normalizzato come:

$$\int_{-\infty}^{\infty} H'(X)\,dx = \sum_j A'_j = \sum_j h'_j \cdot w_j = \sum_j \frac{c_j}{n \cdot \sum_j w_j} \cdot w_j = 1$$

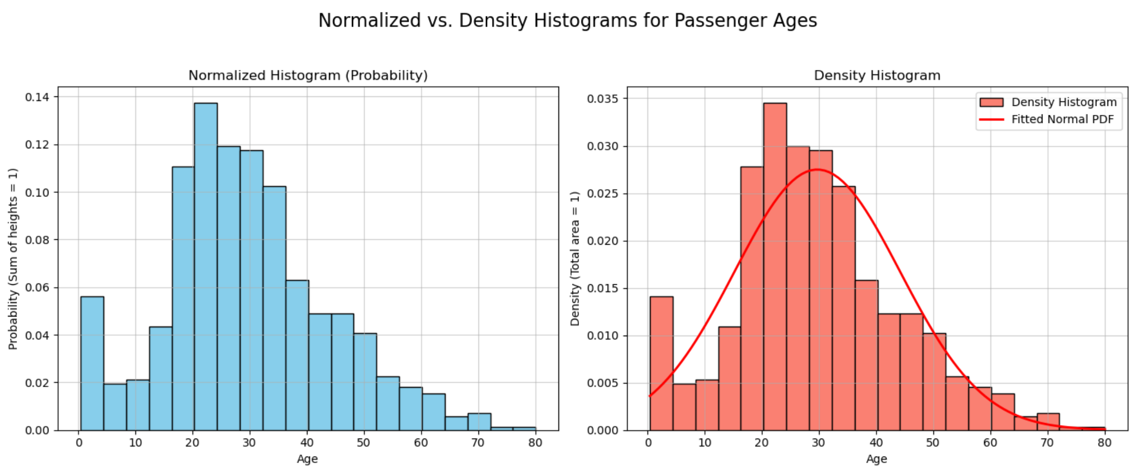

Figura 5.2: Confronto tra due rappresentazioni dell'istogramma delle età dei passeggeri. A sinistra: istogramma normalizzato, in cui l'altezza di ciascun bin è pari alla frequenza relativa (la somma delle altezze dei bin è 1); l'asse y è interpretato come probabilità. A destra: istogramma di densità, in cui l'area totale sotto le barre è 1 e quindi le altezze approssimano la funzione di densità di probabilità continua. In rosso è mostrata la PDF normale stimata sui dati.

[1] Un istogramma è detto normalizzato quando l'area totale sotto l'istogramma è pari a 1. Ovviamente non soddisfano i requisiti di una PDF, ma possono comunque fornire un'approssimazione utile.

5.3.2 Funzione di distribuzione cumulativa (CDF)

Ricordando che la CDF è definita come la probabilità cumulativa che la variabile casuale X assuma un valore minore o uguale a x, possiamo estendere questa definizione anche alle distribuzioni continue:

$$F(x) = P(X \leq x)$$

La CDF fornisce la probabilità cumulativa che la variabile casuale X assuma un valore minore o uguale a x. In una distribuzione continua, la CDF è ottenuta integrando la PDF:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Dove $f(t)$ è la PDF della variabile casuale X.

Possiamo trarre diverse conclusioni da qui:

- La CDF è una funzione non decrescente, poiché le probabilità cumulative non possono diminuire al crescere di x.
- Conoscendo la CDF, possiamo ottenere la PDF derivando la CDF:

$$f(x) = \frac{d}{dx} F(x)$$

- La CDF di una distribuzione continua è continua, a differenza della CDF di una distribuzione discreta che può presentare salti.

5.4 Distribuzioni di probabilità comuni

Ci sono diverse distribuzioni di probabilità comuni che vengono spesso utilizzate in statistica e machine learning. Molto spesso si utilizzano perché si nota che la distribuzione che stiamo studiando si avvicina a una di queste distribuzioni standard e in quel caso si possono sfruttare le loro proprietà per fare inferenza statistica.

5.4.1 Distribuzione uniforme discreta

La distribuzione uniforme discreta è una distribuzione di probabilità in cui tutti i valori possibili di una variabile casuale discreta hanno la stessa probabilità di verificarsi al variare di un parametro k:

$$P(X = a_i) = \frac{1}{k}$$

Dove $\Omega = \{a_1, ..., a_k\}$.

5.4.2 Distribuzione di Bernoulli

La distribuzione di Bernoulli è una distribuzione di probabilità discreta che descrive un esperimento con due possibili esiti: successo e fallimento. La distribuzione viene definito come un singolo parametro ϕ, che rappresenta la probabilità di successo. Quindi possiamo formulare la distribuzione di Bernoulli come:

$$P(X = x) = \begin{cases} \phi & \text{se } x = 1 \\ 1 - \phi & \text{se } x = 0 \end{cases}$$

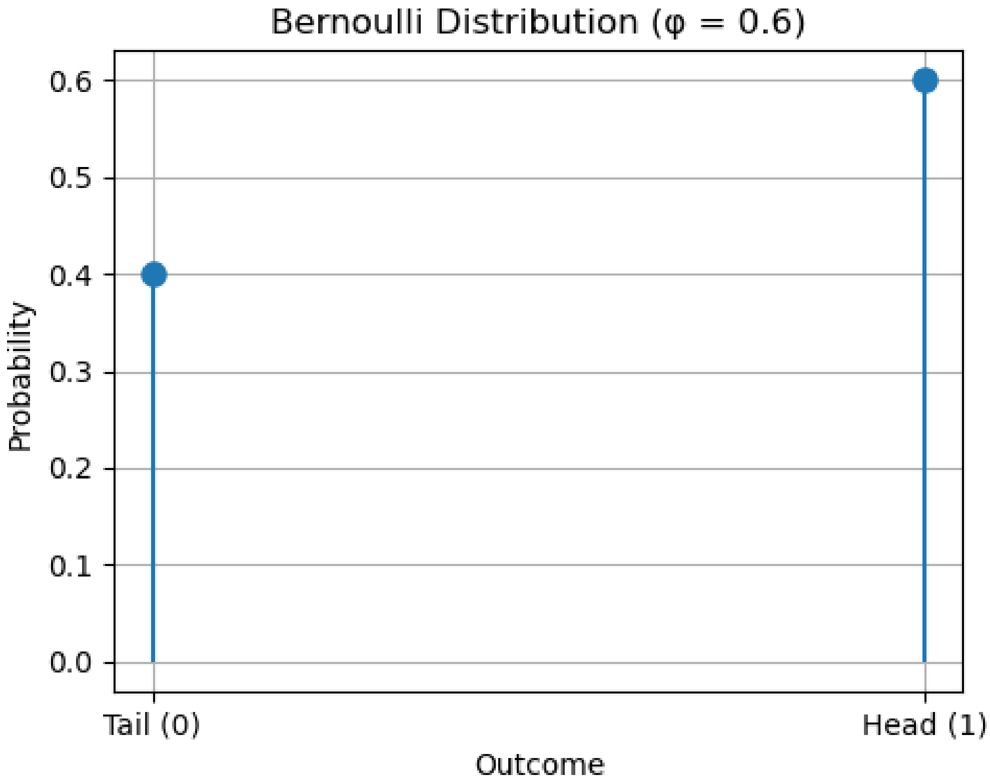

Figura 5.3: Esempio di distribuzione di Bernoulli su una moneta truccata con parametro $\phi = 0.6$. La probabilità di successo (ovvero fare testa) è 0.6, mentre la probabilità di fallimento (fare croce) è 0.4

5.4.3 Distribuzione binomiale

La distribuzione binomiale è una distribuzione di probabilità discreta (PMF) su numeri naturali con parametri n e p. Essa modella la probabilità di ottenere k successi in una sequenza di n esperimenti indipendenti che seguono una distribuzione di Bernoulli con parametro p.
La funzione di massa di probabilità è data da:

$$P(k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Dove:

- k è il numero di successi

- n è il numero di prove indipendenti
- p è la probabilità di successo in una singola prova

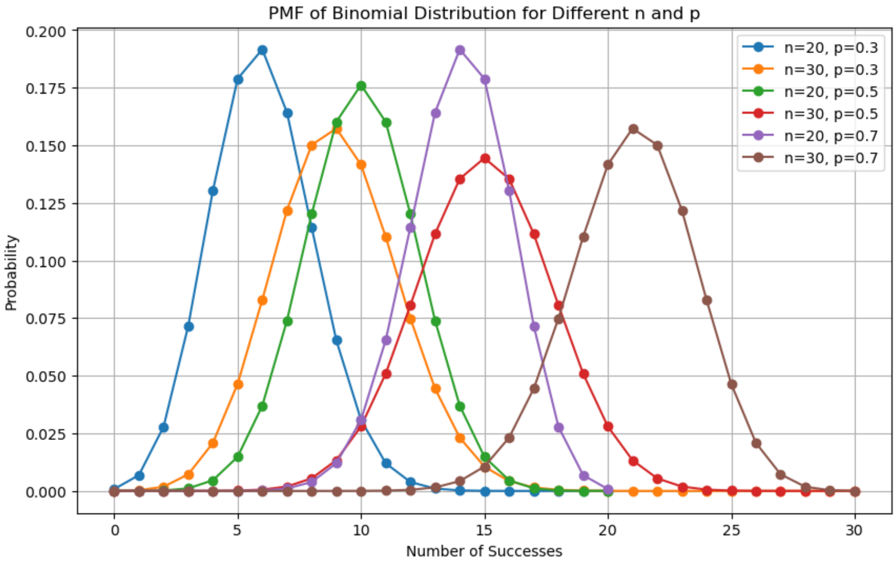

Figura 5.4: Funzioni di massa di probabilità (PMF) della distribuzione binomiale per diversi valori di n (numero di prove) e p (probabilità di successo in ciascuna prova). Ogni curva mostra la probabilità di ottenere k successi su n prove indipendenti: si osserva che il picco della distribuzione si sposta attorno al valore atteso np e che, al crescere di n, la distribuzione tende a diventare più concentrata e più simile a una forma "gaussiana".

Lancio di una moneta. Un esempio comune di distribuzione binomiale è il lancio di una moneta più volte, dove si vuole calcolare la probabilità di ottenere un certo numero di teste (successi) in un numero totale di lanci (prove):

$$P(k) = \binom{n}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{n-k} = \binom{n}{k} \left(\frac{1}{2}\right)^n$$

Dove n è il numero totale di lanci e k è il numero di teste ottenute. Se prendessimo $n = 3$ numero di lanci, $k = 3$ numero di successi (teste) e $p = 0.5$ probabilità di successo in ogni lancio, avremmo:

$$P(3) = \binom{3}{3}(0.5)^3(1 - 0.5)^{3-3} = 1 \cdot (0.5)^3 \cdot 1 = 0.125$$

5.4.4 Distribuzione categorica

La distribuzione categorica è una generalizzazione della distribuzione di Bernoulli per variabili casuali che possono assumere k categorie distinto per un certo k finito.

La distribuzione categorica è definita da un vettore di probabilità $\boldsymbol{p} \in [0,1]^k$, dove ogni p_i ci dà la probabilità di essere nell'i-esima categoria, e deve soddisfare la condizione:

$$\sum_{i=1}^{k} p_i = 1$$

La forma analitica è data da:

$$P(x = i) = p_i \quad \text{per } i = 1, 2, \ldots, k$$

Esempio: lancio di un dado truccato. Consideriamo un singolo lancio di un dado a sei facce truccato. Gli esiti possibili sono:

$$X \in \{1, 2, 3, 4, 5, 6\}.$$

Se il dado è truccato in modo che le probabilità siano:

- $P(X = 1) = 0.10$
- $P(X = 2) = 0.15$
- $P(X = 3) = 0.20$
- $P(X = 4) = 0.20$
- $P(X = 5) = 0.10$
- $P(X = 6) = 0.25$

Allora l'esito segue una distribuzione categorica con $k = 6$ e probabilità specificate $\{p_1, p_2, \ldots, p_6\} = \{0.1,\ 0.15,\ 0.2,\ 0.2,\ 0.1,\ 0.25\}$.

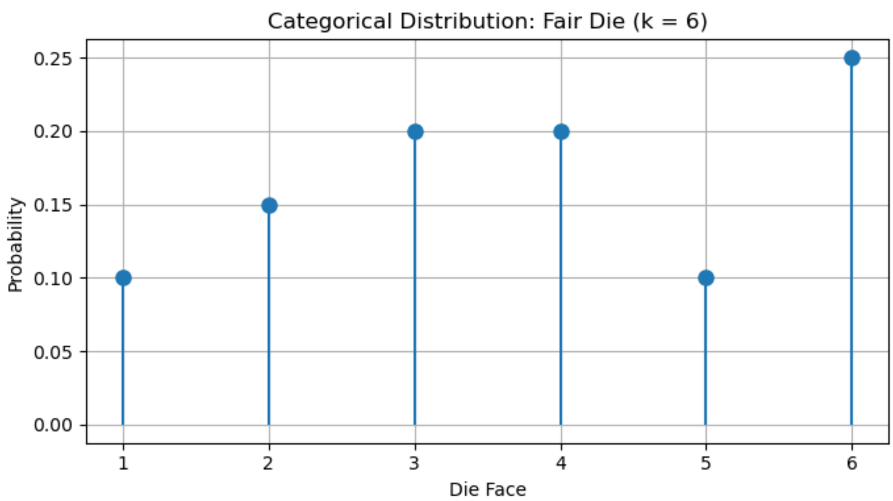

Figura 5.5: Esempio di distribuzione categorica per il lancio di un dado truccato. Le altezze delle barre rappresentano le probabilità associate a ciascuna faccia del dado.

5.4.5 Distribuzione multinomiale

La distribuzione multinomiale è una generalizzazione della distribuzione binomiale per esperimenti con più di due possibili esiti. Viene utilizzata per modellare il numero di occorrenze di ciascuna categoria in una serie di prove indipendenti. In particolare, la distribuzione multinomiale descrive la probabilità di ottenere esattamente $(n_1, \ldots, n_k)$ occorrenze per ciascuna delle k categorie in una sequenza di n espertimenti indipendenti che seguono una distribuzione categorica con probabilità $(p_1, \ldots, p_k)$.

Possiamo definire i parametri della distribuzione multinomiale come:

- n: numero totale di prove indipendenti
- k: numero di categorie
- p_i: probabilità di successo per la categoria i (con $i = 1, 2, \ldots, k$)[2]

Da questo, la funzione di massa di probabilità (PMF) della distribuzione multinomiale è data da:

$$P(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1!n_2!\ldots n_k!} p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

Lancio di un dado più volte. Un esempio comune di distribuzione multinomiale è il lancio di un dado più volte, dove si vuole calcolare la probabilità di ottenere un certo numero di occorrenze per ciascuna faccia del dado in un numero totale di lanci. Considerato un dado a sei facce, qual è la probabilità di ottenere:

- 3 volte il numero 1, $n_1 = 3$
- 2 volte il numero 2, $n_2 = 2$
- 4 volte il numero 3, $n_3 = 4$
- 5 volte il numero 4, $n_4 = 5$
- 0 volte il numero 5, $n_5 = 0$
- 1 volta il numero 6, $n_6 = 1$

Sapendo che la probabilità di ottenere ciascuna faccia del dado è $p_i = \frac{1}{6}$ per ogni $i = 1, 2, \ldots, 6$. Abbiamo $k = 6$ categorie e $n = 15$ lanci, quindi possiamo calcolare la probabilità come:

$$P(3, 2, 4, 5, 0, 1) = \frac{15!}{3!2!4!5!0!1!} \left(\frac{1}{6}\right)^3 \left(\frac{1}{6}\right)^2 \left(\frac{1}{6}\right)^4 \left(\frac{1}{6}\right)^5 \left(\frac{1}{6}\right)^0 \left(\frac{1}{6}\right)^1 = 8.04 \times 10^{-5}$$

[2] La somma totale di ogni p_i deve fare 1.

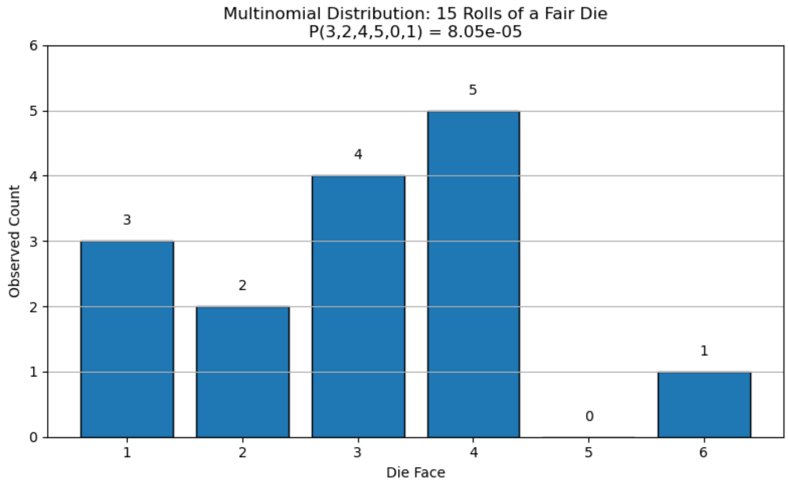

Figura 5.6: Esempio di distribuzione multinomiale per il lancio di un dado a sei facce 15 volte. Le altezze delle barre rappresentano le probabilità associate a ciascuna combinazione di occorrenze delle facce del dado.

5.4.6 Distribuzione Gaussiana (Normale)

La distribuzione Gaussiana, o distribuzione normale, è una delle distribuzioni di probabilità più importanti e ampiamente utilizzate in statistica e machine learning. È molto importante perché, è una funzione di densità molto comune che descrive molti fenomeni naturali e processi casuali.

La distribuzione Gaussiana è caratterizzata da due parametri principali:

- La media $\mu \in \mathbb{R}$ che rappresenta il centro della distribuzione.
- La deviazione standard $\sigma > 0$ che misura la dispersione dei dati intorno alla media.

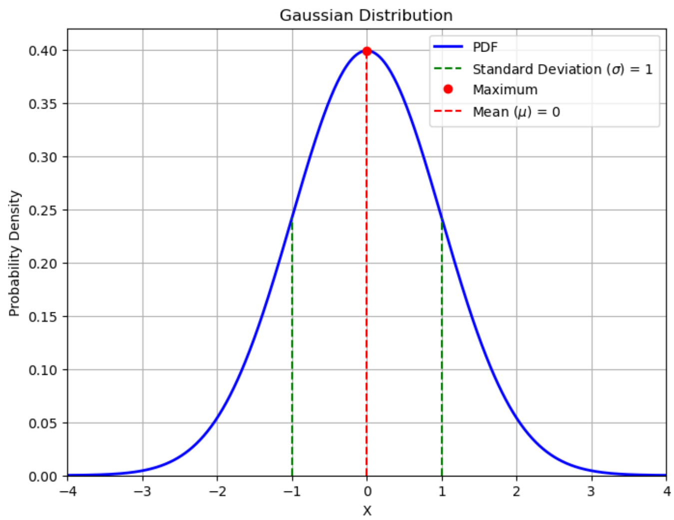

Figura 5.7: Esempio di distribuzione Gaussiana (Normale) con media $\mu = 0$ e deviazione standard $\sigma = 1$. La curva a campana rappresenta la funzione di densità di probabilità (PDF) della distribuzione.

La formula analitica della distribuzione normale è data da:

$$N(x, \mu, \sigma) = \sqrt{\frac{1}{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Questa è la **funzione di densità di probabilità** (PDF) della normale. Ogni pezzo della formula ha un significato preciso:

- Il termine μ è la media. Controlla *dove* è centrata la campana: il punto in cui la curva raggiunge il valore massimo è proprio $x = \mu$. Spostare μ verso destra o verso sinistra sposta tutta la distribuzione senza cambiarne la forma.
- Il termine σ è la deviazione standard. Controlla *quanto è larga* o *stretta* la campana. Se σ è piccolo, la distribuzione è molto concentrata attorno alla media (picco alto e stretto). Se σ è grande, la distribuzione è più sparsa (curva più bassa e larga). La varianza della distribuzione è σ^2.
- Il fattore $\frac{1}{\sigma\sqrt{2\pi}}$, è un fattore di normalizzazione. Serve a garantire che l'area totale sotto la curva valga 1, cioè che la funzione sia davvero una densità di probabilità valida:
$$\int_{-\infty}^{+\infty} N(x, \mu, \sigma)\, dx = 1.$$
Senza questo termine la forma sarebbe "a campana", ma non rappresenterebbe una distribuzione di probabilità corretta.
- Il termine esponenziale $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ determina la forma a campana. Il numeratore $(x-\mu)^2$ misura quanto x è lontano dalla media: più x è lontano, più questo quadrato cresce, e quindi più l'esponenziale crolla verso 0. Il denominatore $2\sigma^2$ controlla quanto velocemente crolla: con una σ grande, si scende più lentamente (code più larghe); con una σ piccola, si scende molto in fretta.

Possiamo notare una cosa molto interessante della distribuzione normale: calcolando l'intervallo di valori che si trovano entro una certa distanza dalla media, possiamo osservare che:

- Circa il 68% dei valori si trova entro una deviazione standard dalla media $(\mu-\sigma, \mu+\sigma)$.
- Circa il 95% dei valori si trova entro due deviazioni standard dalla media $(\mu - 2\sigma, \mu + 2\sigma)$.
- Circa il 99.7% dei valori si trova entro tre deviazioni standard dalla media $(\mu - 3\sigma, \mu + 3\sigma)$.

Questa cosa è utile per notare che, se una certa distribuzione segue la normale, allora possiamo fare delle stime sulla probabilità che un certo valore cada entro un certo intervallo dalla media. In pratica, la PDF ci dice quanto è "densa" la probabilità attorno a ciascun punto x, mentre l'area sotto la curva tra due punti a e b ci dice la probabilità che la variabile aleatoria cada proprio tra a e b. Questo è esattamente ciò che rende la normale così comoda per stimare intervalli di confidenza, fare assunzioni sui dati e modellare rumore nei modelli di machine learning.

5.4.7 Teorema del limite centrale

Il teorema del limite centrale è uno dei risultati più importanti in statistica.

> *Il teorema del limite centrale afferma che la distribuzione della somma (o della media) di un gran numero di variabili casuali $\{X_i\}_{i=1}^n$ indipendenti e identicamente distribuite tende a una distribuzione normale per $n \to \infty$, indipendentemente dalla distribuzione originale delle variabili casuali X_i.*
>
> Definizione 5.2

Questo risultato è fondamentale perché giustifica la diffusione della distribuzione normale in molti contesti statistici e applicazioni pratiche. In particolare, il teorema del limite centrale consente di utilizzare la distribuzione normale come approssimazione per la distribuzione di campioni di grandi dimensioni, anche quando la distribuzione originale non è normale.

Esempio. Consideriamo un dado equo a sei facce, in cui ogni faccia da 1 a 6 ha la stessa probabilità $\frac{1}{6}$. Se lanciamo il dado una sola volta, il risultato è chiaramente discreto (1, 2, 3, 4, 5 o 6) e la distribuzione non assomiglia affatto a una Gaussiana.

Adesso però facciamo così: lanciamo il dado n volte, calcoliamo la *media* dei risultati ottenuti, e ripetiamo questo esperimento molte volte. Per $n = 1$ le medie possibili coincidono con i valori del dado, quindi la distribuzione è piatta e discreta. Per $n = 2, 5, 10$ la distribuzione delle medie campionarie comincia a diventare più "a campana". Per $n = 50$ e soprattutto $n = 5000$, le medie campionarie si concentrano attorno a circa 3.5 (che è il valore atteso di un dado equo) e la loro distribuzione è ormai molto simile a una distribuzione normale.

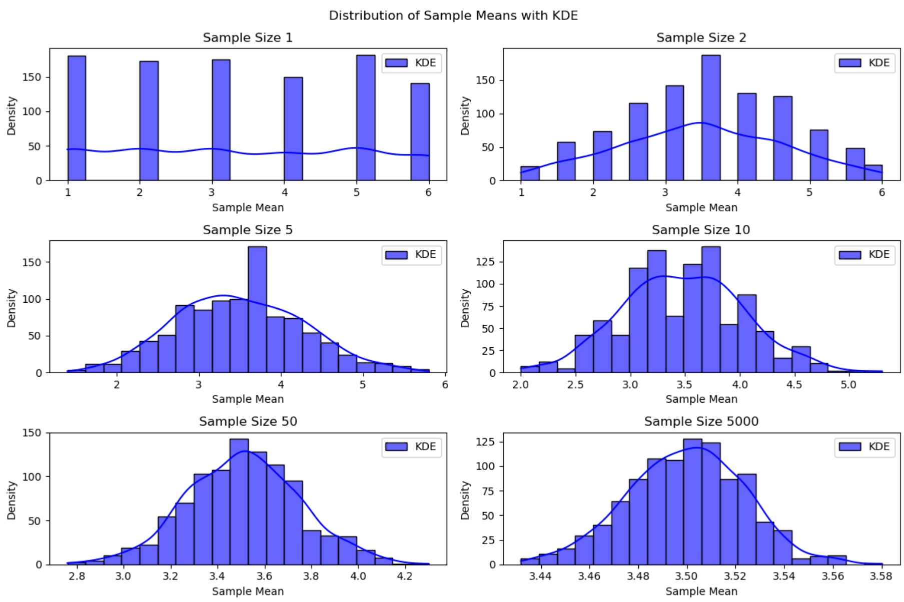

Figura 5.8: Esempio del teorema del limite centrale con il lancio di un dado equo. La distribuzione delle medie campionarie tende a una distribuzione normale.

5.4.8 Distribuzione Gaussiana Multivariata

La distribuzione Gaussiana multivariata è un'estensione della distribuzione normale a più dimensioni. Viene utilizzata per modellare vettori di variabili casuali continue che possono essere correlate tra loro. La distribuzione è caratterizzata da due parametri principali:

- Il vettore di medie $\boldsymbol{\mu} \in \mathbb{R}^d$, che rappresenta il centro della distribuzione.
- La matrice di covarianza $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$, che misura la dispersione e la correlazione tra le variabili.

La formulazione analitica della distribuzione Gaussiana multivariata è data da:

$$N(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{1}{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

A due dimensioni, ovviamente, il vettore μ rappresenta il centro della distribuzione nel piano XY, mentre la matrice di covarianza Σ determina la forma e l'orientamento delle ellissi di densità di probabilità.
Spesso è difficile visualizzare la gaussiana multivariata in un grafico 3D, ma possiamo rappresentarla tramite curve di livello (linee di uguale densità) che mostrano come la densità di probabilità varia nello spazio bidimensionale.
Come si può notare nella figura 5.9, la distribuzione Gaussiana multivariata con media $(0, 0)$ e correlazione positiva tra le due variabili mostra un picco al centro (la media) e le linee di livello formano ellissi orientate, indicando sia la varianza lungo le direzioni principali sia la dipendenza lineare tra le due componenti.
Sotto la superficie 3D, le curve di livello rappresentano le linee di uguale densità sul piano XY. L'orientamento e la forma di queste ellissi sono determinate dalla matrice di covarianza $\boldsymbol{\Sigma}$.

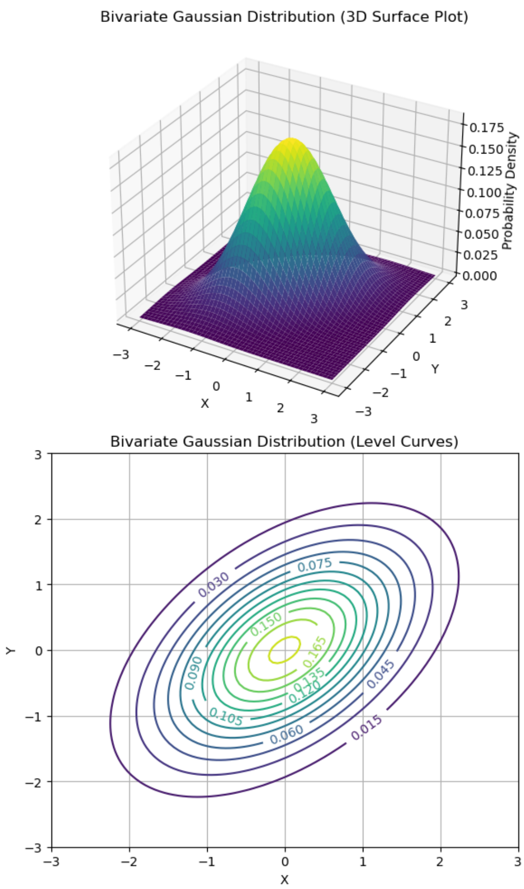

Figura 5.9: Distribuzione Gaussiana bivariata con media $(0, 0)$. La superficie 3D mostra la densità congiunta, mentre le curve di livello (ellissi) mostrano linee di uguale densità sul piano XY. L'orientamento delle ellissi è determinato dalla matrice di covarianza Σ.

Effetti di Σ. In una distribuzione normale a singola dimensione, l'effetto della dispersione è controllato dalla deviazione standard σ. Nella distribuzione Gaussiana multivariata, la matrice di covarianza $\mathbf{\Sigma}$ svolge un ruolo simile, ma in modo più complesso, in quanto determina **forma, orientamento e scala** della distribuzione nello spazio multidimensionale:

- Quando Σ è una matrice diagonale[3], le variabili sono indipendenti tra loro (in quanto $\sigma_{xy} = \sigma_{yx} = 0$), e la distribuzione appare come un "ellissoide" allineato con gli assi.
- Se la varianza lungo le assi è diversa (cioè gli elementi diagonali di Σ sono diversi), la distribuzione sarà più "allungata" in alcune direzioni rispetto ad altre.
- Gli elementi fuori diagonale di Σ rappresentano la covarianza tra le variabili. Se questi valori sono positivi, indica una correlazione positiva (le variabili tendono a crescere insieme); se sono negativi, indica una correlazione negativa (una variabile tende a crescere mentre l'altra diminuisce). Questo si riflette nell'orientamento delle ellissi di densità di probabilità.

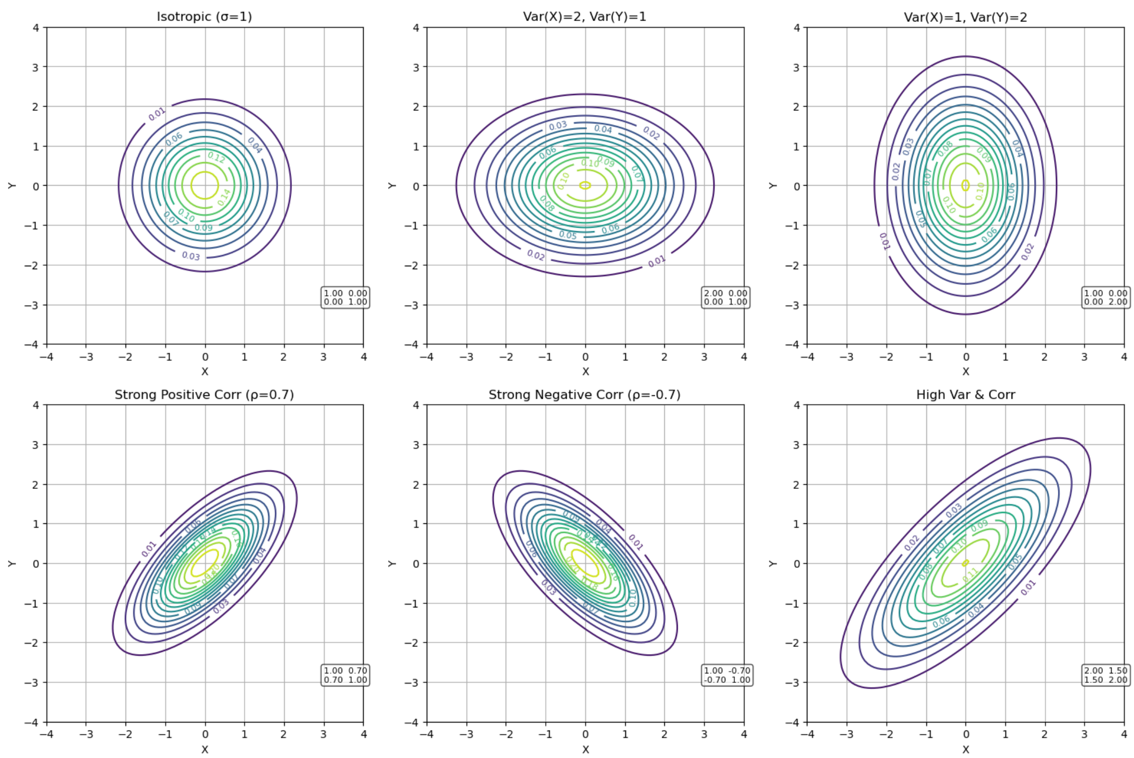

Figura 5.10: Effetto della matrice di covarianza Σ sulla distribuzione Gaussiana bivariata con media $(0, 0)$. Ogni pannello mostra le curve di livello (isodensità), che evidenziano forma, orientamento e scala della distribuzione. Nella riga superiore: caso isotropico (varianze uguali su X e Y), poi varianza di X maggiore di quella di Y, poi varianza di Y maggiore di quella di X; in questi casi Σ è diagonale, quindi non c'è covarianza e le ellissi sono allineate con gli assi. Nella riga inferiore: presenza di termini fuori diagonale in Σ (covarianza non nulla), che introduce correlazione tra le variabili. Con correlazione positiva le ellissi ruotano lungo la diagonale crescente, con correlazione negativa lungo la diagonale decrescente. Questo mostra che Σ controlla non solo quanto è larga la distribuzione in ciascuna direzione, ma anche come è orientata nello spazio.

[3]Una matrice diagonale è una matrice quadrata in cui tutti gli elementi al di fuori della diagonale principale sono zero.

Stime dei parametri nella distribuzione Gaussiana multivariata. Per stimare i parametri della distribuzione Gaussiana multivariata, ovvero il vettore di medie $\boldsymbol{\mu}$ e la matrice di covarianza $\boldsymbol{\Sigma}$ possono essere ottenuti grazie alla massima verosimiglianza (MLE, *maximum likelihood estimation*) sui dati osservati.
Nel caso **univariato**:

- La stima della media μ è data dalla media campionaria:

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

- La stima della varianza σ^2 è data dalla varianza campionaria:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2$$

Dove x_i sono i dati osservati e n è il numero totale di osservazioni.
Nel caso **multivariato**:

- La stima del vettore di medie $\boldsymbol{\mu}$ è data dalla media campionaria vettoriale:

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$$

- La stima della matrice di covarianza $\boldsymbol{\Sigma}$ è data dalla matrice di covarianza in relazione alla variabile casuale X:

$$\hat{\boldsymbol{\Sigma}} = Cov(X) \Rightarrow \hat{\boldsymbol{\Sigma}}_{ij} = Cov(X_i, X_j)$$

5.5 Descrivere una distribuzione di probabilità

Esistono diverse misure per descrivere una distribuzione di probabilità, che si basano su concetti statistici come la media, la varianza e la forma della distribuzione stessa.

5.5.1 Aspettativa (media)

L'aspettativa, o valore atteso, di una variabile casuale X è una misura della tendenza centrale della distribuzione di probabilità. É simile alla media aritmetica, perché quando computiamo una media non facciamo altro che sommare tutti i valori di un insieme e dividere per il numero totale di valori. L'aspettativa fa la stessa cosa, ma tiene conto delle probabilità associate a ciascun valore.

In particolare, possiamo definire per una variabile casuale discreta l'aspettativa come se fosse una media pesata:

$$E[X]_{X\sim P} = \sum_{x\in\Omega} x \cdot P(x)$$

Nel caso di variabili continue, l'aspettativa è definita come:

$$E[X]_{X\sim P} = \int_{x\in\Omega} x \cdot f(x)\, dx$$

Dove $f(x)$ è la funzione di densità di probabilità (PDF) della variabile casuale X.

5.5.2 Varianza e deviazione standard

La varianza è una misura della dispersione dei valori di una variabile casuale rispetto alla sua media. Indica quanto i valori si discostano in media dall'aspettativa. La varianza di una variabile casuale X è definita come:

$$Var_{X\sim P}[X] = E\left[(X - E_{X\sim P}[X])^2\right]$$

Si utilizza spesso, però, la deviazione standard, che è la radice quadrata della varianza in quanto è più interpretabile:

$$\sigma = \sqrt{Var_{X\sim P}[X]}$$

5.5.3 Covarianza

La covarianza è una misura della relazione lineare tra due variabili casuali X e Y. Indica come le due variabili variano insieme. La covarianza è definita come:

$$Cov_{X\sim P_X, Y\sim P_Y}[X, Y] = E\left[(X - E_{X\sim P_X}[X])(Y - E_{Y\sim P_Y}[Y])\right]$$

Si possono distinguere i termini:

- $E[X], E[Y]$ sono le aspettative (medie) delle variabili casuali X e Y.
- $(X - E_{X\sim P_X}[X])$ e $(Y - E_{Y\sim P_Y}[Y])$ sono le deviazioni delle variabili casuali X e Y rispetto alle loro aspettative.
- $(X - E[X])(Y - E[Y])$ rappresenta il prodotto delle deviazioni, che indica come le due variabili variano insieme.

Possiamo distinguere tre casi:

- Se la covarianza è positiva, significa che quando una variabile aumenta, l'altra tende ad aumentare anch'essa.
- Se la covarianza è negativa, significa che quando una variabile aumenta, l'altra tende a diminuire.
- Tanto la covarianza è vicina a zero, tanto meno le due variabili sono correlate linearmente.

Se X è una variabile multidimensionale con d componenti, la covarianza può essere rappresentata come una matrice di covarianza $\Sigma \in \mathbb{R}^{d\times d}$, dove ogni elemento Σ_{ij} rappresenta la covarianza tra la i-esima e la j-esima componente di X:

$$\Sigma_{ij} = Cov[X_i, X_j]$$

5.5.4 Entropia

L'entropia è una misura della quantità di incertezza o imprevedibilità associata a una distribuzione di probabilità. In altre parole, l'entropia quantifica quanto "disordinata" o "casuale" è una variabile casuale.

Self-information. Prima di definire l'entropia, è utile introdurre il concetto di **self-information** (o informazione auto-riferita) di un evento x, che misura la quantità di informazione associata al verificarsi di quell'evento. La self-information è definita come:

$$I(x) = -\log P(x)$$

Il logaritmo viene solitamente calcolato in base 2 (bit) o in base e (nat).

Questa definizione ha senso perché eventi con bassa probabilità (cioè eventi rari) forniscono più informazione quando si verificano, mentre eventi con alta probabilità (eventi comuni) forniscono meno informazione. Possiamo anche notare che:

- Il logaritmo rende la self-information additiva per eventi indipendenti. Se due eventi x e y sono indipendenti, allora:

$$I(X = x, Y = y) = -\log[P(X = x) \cdot P(Y = y)] = I(x) + I(y)$$

- Inoltre il logaritmo negativo riflette il fatto che eventi più probabili (con $P(x)$ vicino a 1) hanno self-information più bassa, mentre eventi meno probabili (con $P(x)$ vicino a 0) hanno self-information più alta.

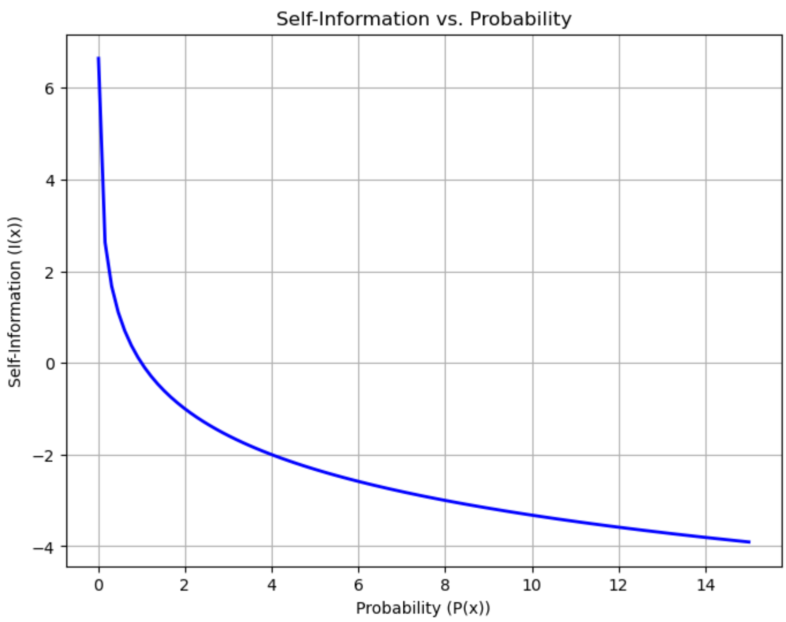

Figura 5.11: Esempio di self-information $I(x) = -\log P(x)$ per diversi valori di probabilità $P(x)$. Si nota che eventi con bassa probabilità (vicino a 0) hanno alta self-information, mentre eventi con alta probabilità (vicino a 1) hanno bassa self-information.

Entropia di una distribuzione. Per una certa variabile casuale X con distribuzione di probabilità $P(X)$, l'entropia $H(X)$ è definita come l'aspettativa della self-information:

- Per una variabile casuale discreta:

$$H(X) = E_{X \sim P}[I(X)] = -\sum_{x \in \Omega} P(x) \log P(x)$$

- Per una variabile casuale continua:

$$H(X) = E_{X \sim P}[I(X)] = -\int_{x \in \Omega} f(x) \log f(x)\, dx$$

Entropia di una variabile di Bernoulli Un esempio semplice è l'entropia di una variabile casuale di Bernoulli con parametro ϕ (probabilità di successo):

$$H(X) = -\left[\phi \log \phi + (1-\phi) \log(1-\phi)\right]$$

Si può notare che l'entropia è massima quando $\phi = 0.5$ (massima incertezza) e minima quando $\phi = 0$ o $\phi = 1$ (nessuna incertezza), come si vede nell'immagine 5.12.

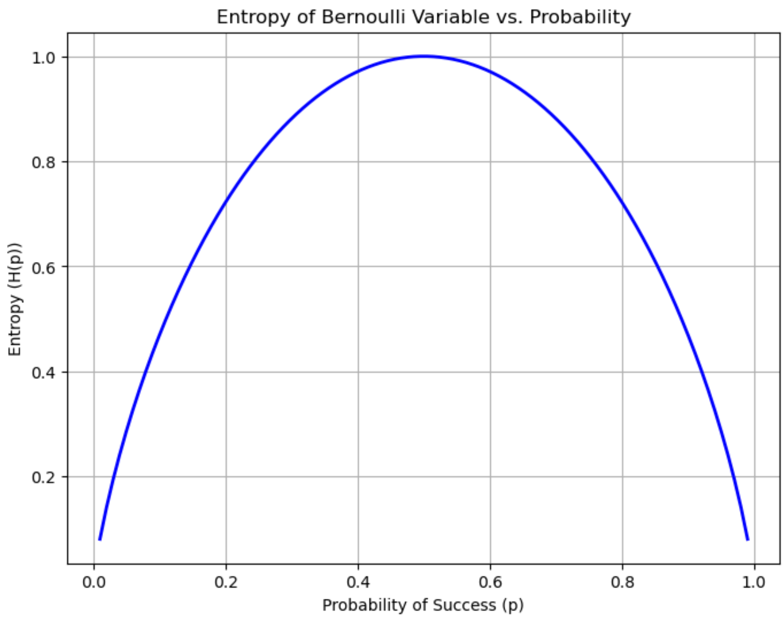

Figura 5.12: Entropia $H(X)$ di una variabile casuale di Bernoulli in funzione del parametro ϕ. L'entropia è massima a $\phi = 0.5$ e minima a $\phi = 0$ o $\phi = 1$.

5.5.5 Standardizzazione

La standardizzazione è una tecnica utilizzata per trasformare una variabile casuale in modo che abbia media zero e deviazione standard uno. Questo processo è utile per confrontare variabili con scale diverse o per preparare i dati per algoritmi di machine learning che sono sensibili alla scala delle caratteristiche. La standardizzazione di una variabile casuale X

con media μ e deviazione standard σ è data dalla formula:

$$Z = \frac{X - \mu_X}{\sigma_X} = \frac{X - E[X]}{\sqrt{Var[X]}}$$

Dove Z è la variabile standardizzata. Dopo la standardizzazione, Z avrà una media di 0 e una deviazione standard di 1. Questa tecnica è anche chiamata z-scoring.

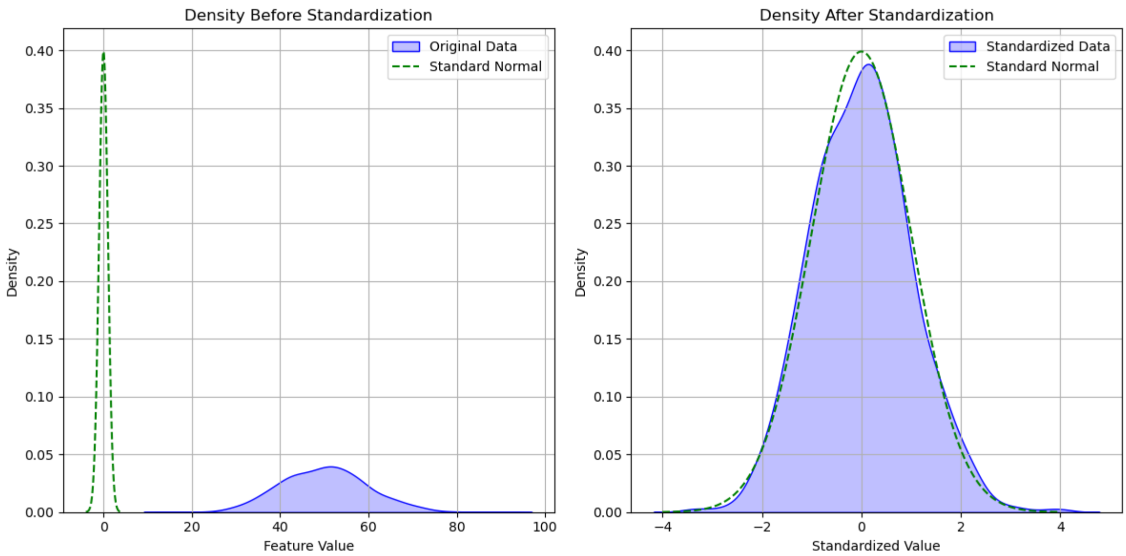

Figura 5.13: Effetto della standardizzazione su una variabile casuale. A sinistra: distribuzione originale della variabile (in blu), confrontata con una normale standard (media 0, deviazione standard 1) mostrata in verde tratteggiato: le scale sono diverse, quindi le curve non sono confrontabili direttamente. A destra: la stessa variabile dopo standardizzazione $Z = \frac{X-\mu_X}{\sigma_X}$; ora i dati trasformati hanno media 0 e deviazione standard 1 e risultano allineati alla distribuzione normale standard.

Come si evince dalla figura 5.13, la standardizzazione consente di confrontare direttamente la distribuzione della variabile casuale con una distribuzione normale standard, facilitando l'analisi statistica.

Capitolo 6

Inferenza Statistica

Molto spesso, in statistica, ci si trova nella situazione di dover prendere decisioni o fare previsioni basate su dati campionari. L'inferenza statistica fornisce gli strumenti necessari per trarre conclusioni riguardo a una popolazione più ampia a partire da un campione limitato di dati.

6.1 Campionamento

Il campionamento è il processo di selezione di un sottoinsieme di individui, oggetti o osservazioni da una popolazione più grande. Quando scarichiamo un dataset, il campionamento è stato già effettuato per noi, mentre se collezionassimo i dati dovremmo campionare dalla popolazione per intero.

6.1.1 Campionamento casuale semplice

Il modo più facile di selezionare un campione dalla popolazione è in **modo casuale**. Questo tipo di campionamento è detto **campionamento casuale semplice** (simple random sampling). Si fanno due assunzioni principali:

- Ogni elemento ha la stessa probabilità di essere selezionato. (selezione equi-probabile).
- La selezione di un elemento non influenza la selezione di un altro elemento (selezione indipendente).

Grazie a questo approccio garantiamo che, per un grande un numero di campioni, le proprietà del campione riflettano quelle della popolazione.

Un problema di questo tipo di campionamento è che, in pratica, è difficile da realizzare. Inoltre non è sempre detto che le assunzioni di selezione equi-probabile e indipendente siano soddisfatte: Ipotizziamo di chiedere agli abitanti di una città se sono soddisfatti dei servizi pubblici, ma lo facciamo solo in centro città. In questo caso, la selezione non è equi-probabile, in quanto gli abitanti delle periferie non hanno la stessa probabilità di essere selezionati rispetto a quelli del centro città.

Inoltre, in questo tipo di campionamento è molto importante il numero di campioni che si estraggono: più campioni si estraggono, più le proprietà del campione tenderanno a riflettere quelle della popolazione.

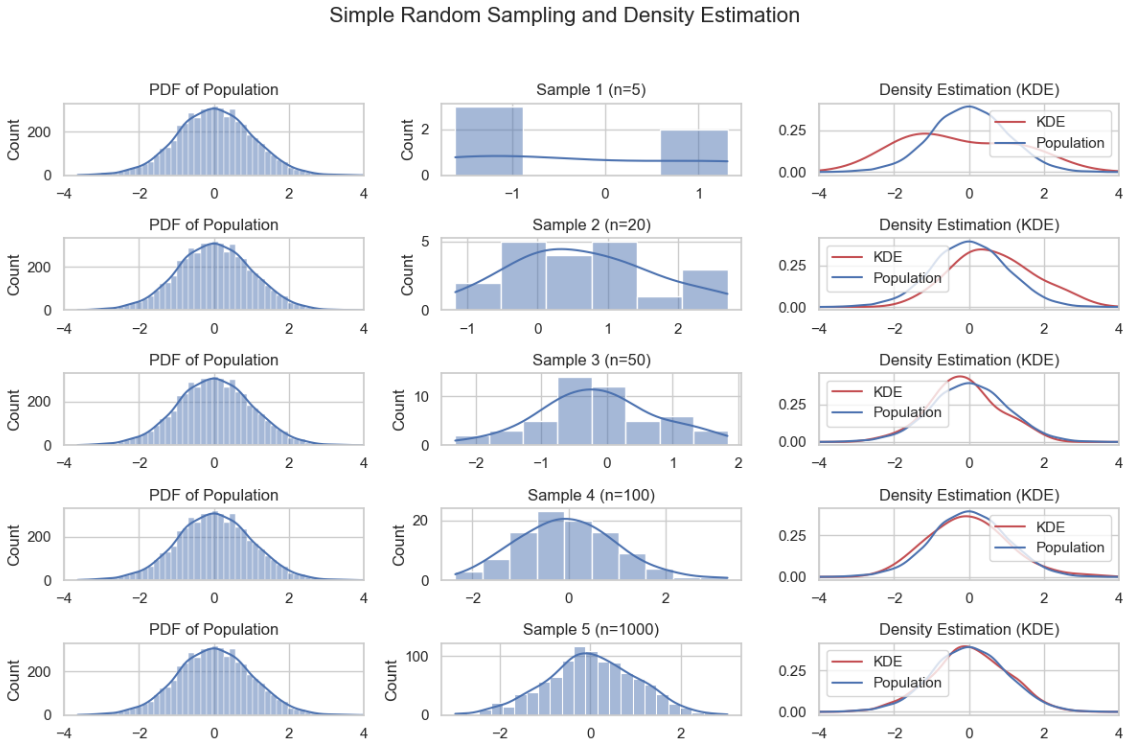

Figura 6.1: Illustrazione del campionamento casuale semplice da una popolazione normale. Ogni riga corrisponde a un diverso numero di osservazioni campionate dalla stessa popolazione ($n = 5, 20, 50, 100, 1000$). Nella prima colonna è mostrata la distribuzione della popolazione (istogramma ad alta risoluzione con la sua densità teorica). Nella seconda colonna è mostrato l'istogramma del singolo campione estratto a quella dimensione n, con una stima di densità sovrapposta. Nella terza colonna è mostrata la stima di densità (KDE, in rosso) del campione confrontata con la densità della popolazione reale (in blu). All'aumentare della dimensione del campione, l'istogramma e la densità stimata del campione diventano via via più simili alla distribuzione originale della popolazione: questo evidenzia che campioni più grandi approssimano meglio la popolazione.

6.1.2 Campionamento stratificato

Uno dei problemi riscontrabili durante il campionamento è che la popolazione è **eterogenea**, ovvero è composta da sottogruppi con caratteristiche diverse. In questi casi, il campionamento casuale semplice potrebbe non essere rappresentativo della popolazione intera, poiché alcuni sottogruppi potrebbero essere sottorappresentati o sovrarappresentati nel campione.

Per risolvere questa problematica si usa il **campionamento stratificato** (stratified sampling). In questo approccio, la popolazione viene suddivisa in sottogruppi omogenei chiamati **strati** (strata) basati su caratteristiche rilevanti (ad esempio età, genere, reddito).

Successivamente, si esegue un campionamento casuale semplice all'interno di ciascuno strato.

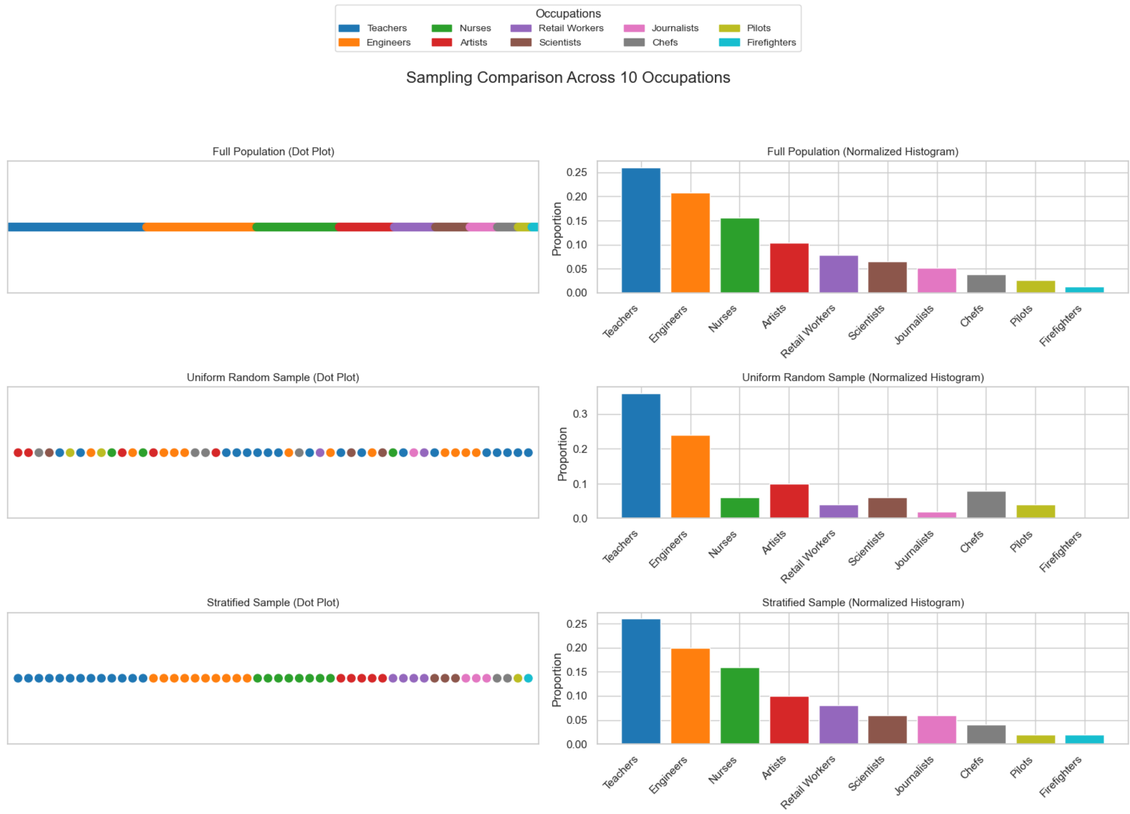

Figura 6.2: Confronto tra diversi metodi di campionamento su una popolazione suddivisa in 10 professioni. Riga superiore: popolazione completa, mostrata sia come dot plot (a sinistra, ogni punto è un individuo colorato per professione) sia come istogramma normalizzato (a destra, proporzioni reali di ciascuna professione nella popolazione). Riga centrale: campione ottenuto tramite campionamento casuale semplice (uniform random sample). Le proporzioni osservate nel campione possono discostarsi da quelle reali della popolazione, specialmente per le categorie meno frequenti. Riga inferiore: campione stratificato (stratified sample), in cui si forza la presenza di ogni categoria in proporzione alla popolazione. In questo caso, l'istogramma delle proporzioni nel campione è molto più fedele alla distribuzione originale.

Come si vede nella figura 6.2, il campionamento stratificato garantisce che ogni strato sia rappresentato nel campione in proporzione alla sua presenza nella popolazione, migliorando così la rappresentatività del campione sulla popolazione complessiva.

6.2 Campionare la distribuzione della media

Uno degli obiettivi principali dell'inferenza statistica è stimare parametri della popolazione, come la media o la varianza, a partire dai dati campionari. Un concetto fondamentale in questo contesto è la **distribuzione campionaria** (sampling distribution) di una statistica, che descrive come quella statistica varia da un campione all'altro. In particolare, la distribuzione campionaria della media campionaria è di grande interesse. La media campionaria è la media calcolata su un campione estratto dalla popolazione. Questo viene fatto perché spesso non si ha accesso all'intera popolazione, ma solo a un campione di essa.

Consideriamo questo esempio: un panificio pacchi di biscotti da 1kg ciascuno. Ogni pacco ha un peso che varia leggermente a causa delle variazioni nel processo di produzione. Supponiamo di prendere un campione randomico di $n = 1000$ pacchi e misurare il loro peso medio:

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

Troviamo tuttavia, che il peso medio è uguale a 1000.2g, leggermente superiore al peso nominale di 1000g. Ci rendiamo conto che se ripetessimo l'estrazione di un campione di 1000 pacchi e calcolassimo nuovamente la media, otterremmo un valore leggermente diverso. Questo accade perché ogni campione è diverso e quindi la media campionaria varia da campione a campione.

Questo fenomeno ripetuto è descritto dalla **distribuzione campionaria della media** (sampling distribution of the sample mean). La distribuzione campionaria della media descrive come la media campionaria varia quando si estraggono ripetutamente campioni dalla popolazione. Trattando ogni pacco come una variabile casuale X_i che ha $E[X_i] = \mu$ e $Var(X_i) = \sigma^2$, la media campionaria $\bar{X}$ è anch'essa una variabile casuale con:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Dal limite del teorema centrale (sotto-sezione 5.4.7) si sa che, per campioni sufficientemente grandi, la distribuzione campionaria della media tende a una distribuzione normale con media μ e varianza $\frac{\sigma^2}{n}$. Quindi possiamo scrivere:

$$\begin{aligned}
\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu \\
\text{Var}[\bar{X}] &= \text{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[X_i] = \frac{\sigma^2}{n} \\
\text{Std}[\bar{X}] &= \sqrt{\text{Var}[\bar{X}]} = \frac{\sigma}{\sqrt{n}}
\end{aligned}$$

Possiamo notare diverse cose:

- La media della distribuzione campionaria della media tende alla media della popolazione μ. Questo significa che la media campionaria è uno stimatore non distorto della media della popolazione.
- La deviazione standard della distribuzione campionaria della media quantifica la precisione della stima della media campionaria, in quanto un campione più grande (maggiore n) riduce la variabilità della media campionaria intorno alla media della popolazione.

Tuttavia persiste un problema: nella maggior parte dei casi, non conosciamo la varianza della popolazione σ^2.

6.2.1 Errore standard

Per risolvere il problema dela varianza sconosciuta, si può stimare la varianza della popolazione utilizzando quella che si chiama "varianza campionaria" (sample variance):

$$s_{n-1} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Da questa misura, possiamo definire l'errore standard (standard error) della media campionaria come:

$$SE_{\bar{X}} = \frac{s_{n-1}}{\sqrt{n}}$$

L'errore standard fornisce una stima della variabilità della media campionaria intorno alla media della popolazione. È molto simile alla deviazione standard della distribuzione campionaria della media, ma utilizza la varianza stimata dal campione invece della varianza reale della popolazione.

Notiamo anche una cosa: ridurre l'errore standard è molto costoso, in quanto per dimezzare l'errore standard bisogna quadruplicare la dimensione del campione n perché l'errore standard decresce con la radice quadrata di n.

Questo può essere descritto bene in una figura del genere (figura 6.3):

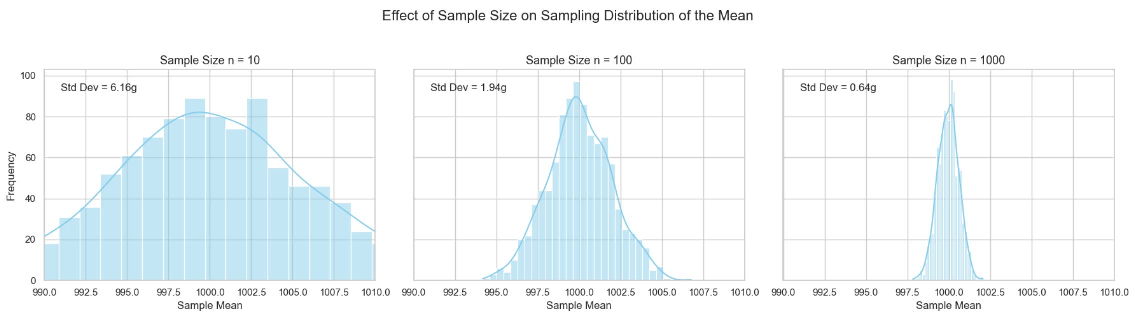

Figura 6.3: Distribuzione delle medie campionarie per diversi valori di n (10, 100, 1000) ottenute simulando il campionamento ripetuto dalla stessa popolazione. Ogni pannello mostra l'istogramma delle medie dei campioni e una stima di densità sovrapposta. All'aumentare della dimensione del campione, la distribuzione delle medie diventa più stretta attorno al valore medio della popolazione e la deviazione standard della media campionaria (errore standard) diminuisce in accordo con $\text{Std}[\bar{X}] = \sigma/\sqrt{n}$.

6.2.2 Distribuzione t-Student

Per risolvere il problema della stima su una piccola dimensione del campione, si può utilizzare la distribuzione t di Student (Student's t-distribution). Questa distribuzione è simile alla distribuzione normale, ma ha code più pesanti, il che significa che c'è una maggiore probabilità di osservare valori estremi. Viene definita come:

$$t_{n-1} = \frac{\bar{X} - \mu}{SE_{\bar{X}}}$$

Dove $n-1$ sono i gradi di libertà[1] (degrees of freedom). Al crescere di n la distribuzione t di Student si avvicina sempre più alla distribuzione normale (figura 6.4).

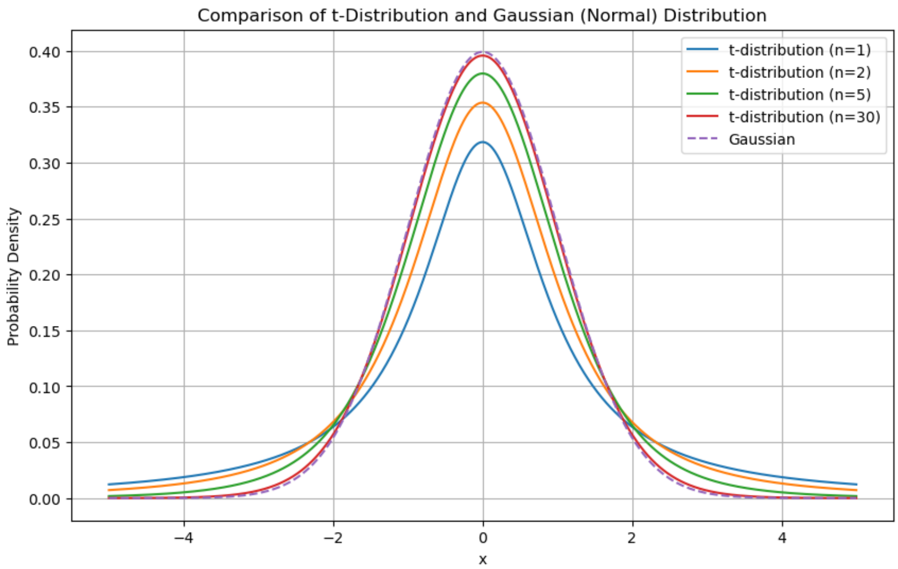

Figura 6.4: Confronto tra la distribuzione t di Student e la distribuzione normale standard. Ogni curva t corrisponde a un diverso numero di gradi di libertà ($n = 1, 2, 5, 30$). Per valori piccoli di n la distribuzione t ha code più pesanti (maggiore probabilità di valori estremi) e un picco più basso rispetto alla Gaussiana. All'aumentare di n, la distribuzione t si avvicina alla distribuzione normale standard, fino a diventare praticamente indistinguibile per n grandi.

6.2.3 Intervallo di confidenza

Nel nostro esempio del panificio, se dovessimo riportare il peso medio dei pacchi di biscotti, potremmo voler includere una misura della nostra incertezza riguardo a questa stima. Un modo comune per farlo è attraverso un **intervallo di confidenza** (confidence interval). Un intervallo di confidenza fornisce un range di valori all'interno del quale ci aspettiamo che il vero parametro della popolazione (in questo caso, la media del peso dei pacchi) cada con una certa probabilità (ad esempio, il 95%).

Ricordando che la media campionaria $\bar{X}$ segue una distribuzione normale centrata sulla media della popolazione μ e ricordando che per una distribuzione normale il 68% dei valori

[1] I gradi di libertà rappresentano il numero di valori indipendenti che possono variare in un'analisi statistica. Nel caso della distribuzione t di Student, i gradi di libertà sono pari a $n-1$ perché stiamo stimando la media della popolazione a partire da un campione di dimensione n.

è compreso in $\mu \pm \sigma$ possiamo scrivere:

$$P(\mu - \sigma \leq \bar{X} \leq \mu + \sigma) = 0.68$$

E ovviamente segue che:

$$\bar{x} \in [\mu - \sigma, \mu + \sigma] \Leftrightarrow \mu \in [\bar{x} - \sigma, \bar{x} + \sigma]$$

E quindi possiamo stimare un intervallo di confidenza al 68% per la media della popolazione come:

$$P(\bar{x} - \sigma \leq \mu \leq \bar{x} + \sigma) = 0.68$$

Questo risultato può anche essere visto nel grafico della figura 6.5.

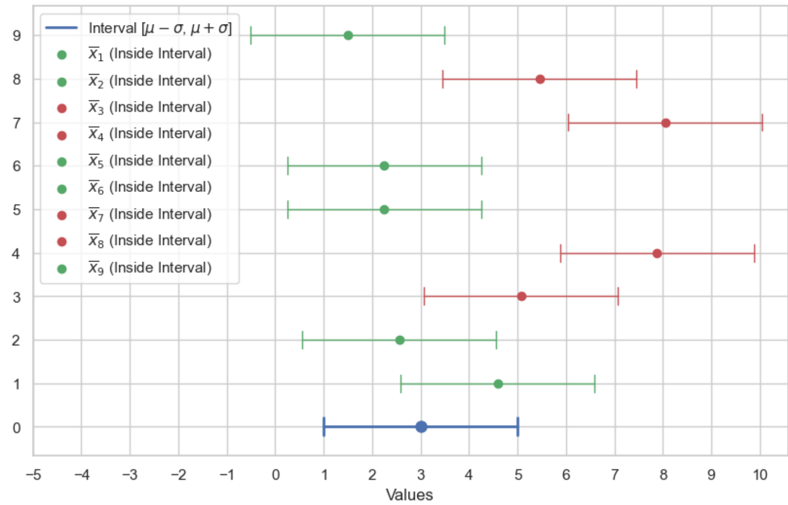

Figura 6.5: Visualizzazione dell'intervallo $[\mu - \sigma,\ \mu + \sigma]$ (barra orizzontale blu) e delle medie campionarie $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_9$ ottenute da campioni diversi. Per ciascun campione è mostrato il suo intervallo $\bar{X}_i \pm \sigma$ (barra orizzontale). I punti verdi indicano le medie campionarie i cui intervalli contengono la media reale μ, mentre i punti rossi indicano quelle che non la contengono. L'idea è che, ripetendo il campionamento, la maggior parte degli intervalli stimati copre il valore vero del parametro, ma non tutti.

Questo è un risultato molto potente, in quanto ci permette di quantificare l'incertezza associata alla nostra stima della media della popolazione. Tuttavia, è importante notare che l'intervallo di confidenza dipende dalla dimensione del campione e dalla variabilità dei dati: campioni più grandi e dati meno variabili portano a intervalli di confidenza più stretti, indicando una maggiore precisione nella stima della media della popolazione.

Chiamiamo quindi, in questo contesto, l'intervallo $[\bar{x} - \sigma, \bar{x} + \sigma]$ di confidenza al 68%.

Il problema però rimmane che molto spesso non abbiamo a disposizione la deviazione standard della popolazione σ, ma solo quella del campione s_{n-1}, da cui ricaviamo lo standard error e lo sostituiamo nell'intervallo di confidenza:

$$[\bar{x} - SE_{\bar{X}}, \bar{x} + SE_{\bar{X}}]$$

Generalizzazione per altri livelli di confidenza. In generale, si può generalizzare questa formulazione per un certa percentuale di confidenza p chiamata **livello di confidenza** (confidence level). Da questo, formuliamo la probabilità come:

$$p = P(\bar{x} - \beta\sigma \leq \mu \leq \bar{x} + \beta\sigma)$$

Che porta a:

$$[\bar{x} - \beta \cdot SE_{\bar{X}}, \bar{x} + \beta \cdot SE_{\bar{X}}]$$

Livello di significatività. Possiamo scrivere una formulazione alternativa basata su un parametro α definito come **livello di significatività** (significance level). In questo contesto α rappresenta la probabilità che l'intervallo di confidenza non contenga il vero parametro della popolazione. Quindi, se vogliamo un intervallo di confidenza del 95%, il livello di significatività sarà $\alpha = 0.05$. Quindi:

- Con probabilità $1 - \alpha$, l'intervallo di confidenza contiene il vero parametro della popolazione, quindi "cattura" μ.
- Con probabilità α, l'intervallo di confidenza non contiene il vero parametro della popolazione, quindi "manca" μ.

Se scegliessimo $\alpha = 0.05$, avremmo un intervallo di confidenza del 95% dato da:

$$[\bar{x} - \beta \cdot SE_{\bar{X}}, \bar{x} + \beta \cdot SE_{\bar{X}}]$$

Dove β è scelto in modo tale che l'area sotto la curva della distribuzione normale tra $-\beta$ e $+\beta$ sia pari a $1 - \alpha$. Per $\alpha = 0.05$, β è approssimativamente uguale a 1.96.

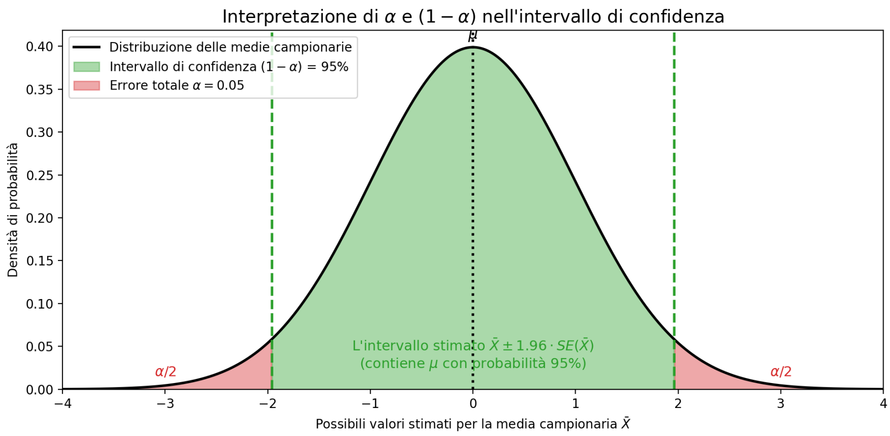

Figura 6.6: Interpretazione dell'intervallo di confidenza al 95%. L'area verde rappresenta l'intervallo di confidenza $(1 - \alpha) = 95\%$, che corrisponde ai valori della media campionaria $\bar{X}$ che, una volta stimati sul campione, "catturano" la vera media μ. Le aree rosse nelle code (ciascuna di area $\alpha/2$) rappresentano i casi in cui l'intervallo stimato non contiene la media reale: la probabilità complessiva di errore è $\alpha = 0.05$. Le linee tratteggiate verdi indicano i limiti $\bar{X} \pm 1.96 \cdot SE(\bar{X})$, mentre la linea tratteggiata nera indica la vera media μ.

6.3 Bootstrapping

Il bootstrapping è una tecnica di inferenza statistica che consente di stimare la distribuzione di una statistica campionaria quando è piccola e non segue una distribuzione normale. Questa tecnica si basa sul concetto di campionamento con reinserimento dal campione originale per creare nuovi campioni chiamati **campioni bootstrap**.

L'idea alla basa del bootstrapping è semplice:

1. Si parte con il campione di dimensione n.
2. Si creano B nuovi campioni bootstrap, ciascuno di dimensione n, estraendo casualmente con reinserimento dal campione originale. Il nuovo campione avrà dimensione originale ma alcuni elementi potrebbero essere ripetuti, mentre altri potrebbero non essere selezionati.
3. Si calcola la statistica di interesse (ad esempio, la media, la mediana, la varianza) per ciascun campione bootstrap.
4. Si ripetono gli step 2 e 3 per un numero elevato di volte per ottenere una distribuzione della statistica di interesse.
5. Si ricostruisce la distribuzione della statistica di interesse dai valori calcolati sui campioni bootstrap.

6.4 Stimatori

Nell'inferenza statistica, uno degli obiettivi principali è stimare i parametri della popolazione a partire dai dati campionari. Per fare questo, si introducono gli **stimatori** (estimators), che sono funzioni dei dati campionari utilizzate per stimare i parametri della popolazione. Un esempio è la media, che è uno stimatore della media della popolazione.

6.4.1 Stimatore del bias

Sia X una variabile casuale e siano $x = (x_1, x_2, \ldots, x_n)$ un campione di dimensione n estratto da X. Sia $T(X)$ uno stimatore della quantità di popolazione ϕ:

$$T(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Poiché il campione cambia, anche il valore di $T(X)$ cambia. Quindi $T(X)$ è anch'esso una variabile casuale con una sua distribuzione.

Possiamo quindi definire il **bias** (bias) dello stimatore $T(X)$ come la differenza tra il valore atteso dello stimatore e il vero valore del parametro della popolazione:

$$\text{Bias}(T) = \mathbb{E}[T(X)] - \phi$$

Che è una misura di quanto lo stimatore si discosta in media dal vero parametro della popolazione. Se il bias è zero, lo stimatore è detto **non distorto** (unbiased), altrimenti è **distorto** (biased).

6.4.2 Stimatore della varianza

Varianza distorta. Se volessimo riprendere l'esempio dei biscotti, potremmo usare la varianza campionaria per stimare la varianza della popolazione:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Il problema di questa stima è che è uno stimatore distorto della varianza della popolazione, in quanto il valore atteso è:

$$\mathbb{E}[s_n^2] = \frac{n-1}{n}\sigma^2$$

che è sempre minore della varianza reale della popolazione σ^2:

$$\frac{n-1}{n} < 1 \quad \Rightarrow \quad \mathbb{E}[s_n^2] < \sigma^2$$

Quindi lo stimatore s_n^2 è distorto, con un bias negativo e va a sottostimare la varianza della popolazione. Per risolvere si può usare lo stimatore della varianza non distorto:

$$s_{n-1}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

risolvendo il problema del bias perché ha valore atteso $\mathbb{E}[s_{n-1}^2] = \sigma^2$.

6.4.3 Varianza di uno stimatore

Grazie alla varianza possiamo misurare la precisione di uno stimatore. La **varianza di uno stimatore** è definita come:

$$Var(T(X)) = \mathbb{E}[(T(X) - \mathbb{E}[T(X)])^2]$$

Anche qui una bassa misura di varianza indica che lo stimatore è preciso, mentre una varianza alta indica che lo stimatore è meno preciso.

6.4.4 Bias-Varianza Tradeoff

Nella scelta di uno stimatore, spesso si deve affrontare un compromesso tra bias e varianza, noto come **bias-variance tradeoff**. Uno stimatore con un bias basso potrebbe avere una varianza elevata, mentre uno stimatore con una varianza bassa potrebbe avere un bias elevato. Si possono distinguere quattro casi principali, (illustrati nella figura 6.7):

- **Basso bias, bassa varianza:** Le stime sono vicine tra loro e vicine al valore vero. Lo stimatore è sia accurato che stabile (caso ideale).
- **Basso bias, alta varianza:** Le stime sono in media corrette (attorno al valore vero), ma molto disperse. Lo stimatore è accurato in media, ma instabile.
- **Alto bias, bassa varianza:** Le stime sono tutte raggruppate, ma lontane dal valore vero. Lo stimatore è sistematicamente sbilanciato, ma coerente.
- **Alto bias, alta varianza:** Le stime sono lontane dal valore vero e molto disperse. È il caso peggiore: impreciso e inaccurato.

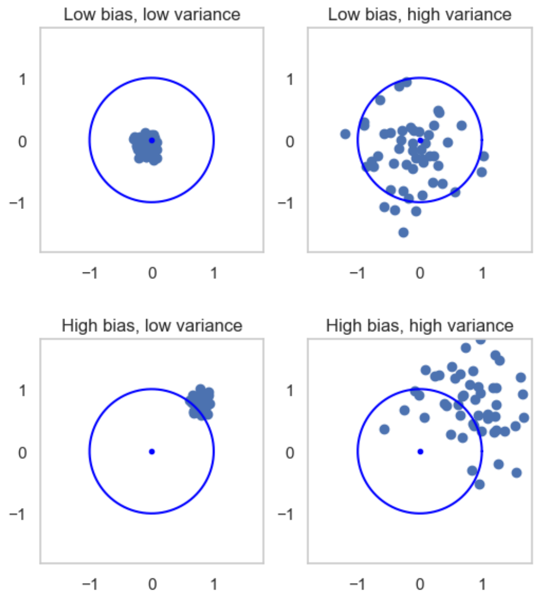

Figura 6.7: Illustrazione concettuale di bias e varianza con l'analogia del bersaglio. Ogni pannello dell'immagine mostra una serie di stime (punti blu) rispetto al valore vero (centro del bersaglio). Il modo in cui i punti si distribuiscono rispetto al centro riflette combinazioni diverse di bias (quanto siamo lontani dal valore vero) e varianza (quanto le stime sono stabili tra loro).

6.5 Test statistici

I test statistici sono procedure utilizzate per prendere decisioni riguardo a una popolazione basandosi su dati campionari. Questi test permettono di valutare ipotesi specifiche riguardo a parametri della popolazione, come la media o la varianza, e di determinare se le osservazioni campionarie forniscono prove sufficienti per accettare o rifiutare tali ipotesi.

6.5.1 Test di ipotesi

Gli intervalli di confidenza e gli stimatori sono strettamente legati ai **test di ipotesi** (hypothesis testing), che sono procedure statistiche utilizzate per prendere decisioni riguardo a una popolazione basandosi su dati campionari. La differenza sta nel fatto che mentre gli intervalli di confidenza forniscono un range di valori plausibili per un parametro della popolazione, i test di ipotesi valutano la validità di una specifica affermazione riguardo a quel parametro.

Si definisce:

- **Ipotesi nulla** (null hypothesis, H_0): è l'ipotesi di base che si vuole testare. Spesso rappresenta uno stato di "nessun effetto" o "nessuna differenza".
- **Ipotesi alternativa** (alternative hypothesis, H_a): è l'ipotesi che si vuole sostenere se i dati forniscono prove sufficienti contro l'ipotesi nulla.

L'ipotesi nulla è l'ipotesi che stiamo cercando di smentire[2] e usiamo l'ipotesi alternativa come supporto per la nostra affermazione.

Prima di procedere con il testo, si sceglie un **livello di significatività** α, che rappresenta la probabilità di rifiutare l'ipotesi nulla H_0 quando essa è vera. Dopo ci chiediamo qual è la probabilità di osservare i dati campionari (o qualcosa di più estremo) sotto l'assunzione che l'ipotesi nulla sia vera. Questa probabilità è chiamata **valore p** (p-value).

Per fare questo, utilizziamo la statistica di test usando la distribuzione t di Student (sezione 6.2.2) perché ci dice come si comporta la media campionaria rispetto alla media della popolazione sotto l'ipotesi nulla.

Dopo aver calcolato il valore p, lo confrontiamo con il livello di significatività α:

- Se il valore p è minore o uguale a α, rifiutiamo l'ipotesi nulla H_0 a favore dell'ipotesi alternativa H_a.
- Se il valore p è maggiore di α, non rifiutiamo l'ipotesi nulla H_0.

Esempio: una moneta è truccata? Immaginiamo di avere una moneta e di voler capire se è equa (ossia, se la probabilità di ottenere testa è uguale a quella di ottenere croce). Formuliamo le ipotesi:

- Ipotesi nulla H_0: la moneta è equa, quindi $p = 0.5$.
- Ipotesi alternativa H_a: la moneta è truccata, quindi $p \neq 0.5$.

Facciamo l'esperimento: lanciamo la moneta 10 volte e otteniamo 9 teste e 1 croce. Fissiamo un livello di significatività $\alpha = 0.05$, ovvero siamo disposti ad accettare un 5% di rischio di rifiutare l'ipotesi nulla quando essa è vera (ovvero dire "la moneta è truccata" quando in realtà è equa).

Calcoliamo il valore p (che indica la probabilità di ottenere 9 o più teste in 10 lanci se la moneta fosse equa):

$$\text{Valore p} = P(X \geq 9) = P(X = 9) + P(X = 10) = \binom{10}{9}(0.5)^{10} + \binom{10}{10}(0.5)^{10} \approxeq 1.6\%$$

Con una moneta onesta ($P(Testa) = P(Croce) = 0.5$), la probabilità di ottenere 9 o più teste in 10 lanci è:

$$\binom{10}{9}(0.5)^{10} \cdot (0.5)^{10} + \binom{10}{10}(0.5)^{10} = \frac{10}{1024} + \frac{1}{1024} = \frac{11}{1024} \approx 1.07\%$$

Ma siccome stiamo testando $p \neq 0.5$, dobbiamo considerare anche l'altra coda della distribuzione (ovvero ottenere 1 o meno teste in 10 lanci):

$$\binom{10}{0}(0.5)^{10} + \binom{10}{1}(0.5)^{10} = \frac{1}{1024} + \frac{10}{1024} = \frac{11}{1024} \approx 1.07\%$$

[2] Un po' come avviene con le dimostrazioni per assurdo.

Quindi se la moneta fosse onesta, un risultato così estremo capitrebbe solo nel:

$$p - value \approx 2 \cdot 1.07\% = 2.14\%$$

Adesso confrontiamo il p-value con α:

$$2.14\% < 5\%$$

Poiché il p-value è minore di α, rifiutiamo l'ipotesi nulla H_0 e concludiamo che c'è evidenza sufficiente per suggerire che la moneta è truccata (ipotesi alternativa H_a).

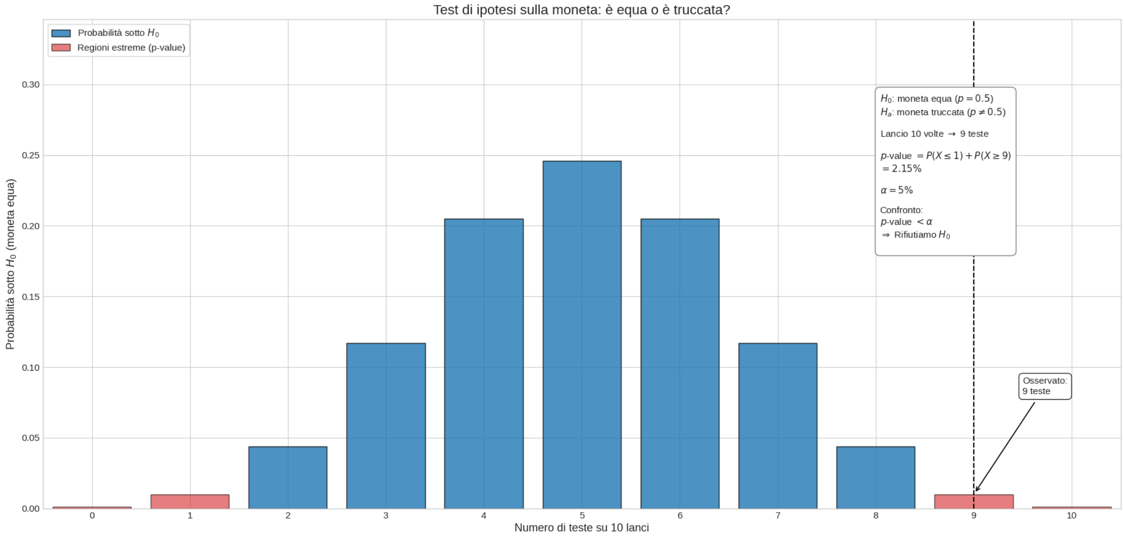

Figura 6.8: Esempio di test di ipotesi su una moneta. Le barre mostrano la probabilità di ottenere k teste su 10 lanci se la moneta fosse equa ($H_0 : p = 0.5$). Le barre rosse evidenziano gli esiti "estremi" ($k \leq 1$ o $k \geq 9$) che contribuiscono al valore p in un test bilaterale. Nell'esperimento osserviamo 9 teste (linea tratteggiata): questo caso cade in zona estrema. Il valore $p \approx 2.15\%$ è minore di $\alpha = 5\%$, quindi rifiutiamo H_0 e concludiamo che ci sono evidenze per dire che la moneta è truccata ($H_a : p \neq 0.5$).

Tipi di errori nei test statistici. Il test di ipotesi è una tipologia di test che può portare a due tipi di errori: rigettare l'ipotesi nulla oppure non rigettarla e questa è una tipologia di classificazione binaria.

Come in tutte le classificazioni binarie, non si ha mai la probabilità del 100% di prendere la decisione giusta. Si possono quindi commettere due tipi di errori:

- **Errore di tipo I** (Type I error): si verifica quando si rifiuta l'ipotesi nulla H_0 quando essa è vera. La probabilità di commettere un errore di tipo I è pari al livello di significatività α.
- **Errore di tipo II** (Type II error): si verifica quando non si rifiuta l'ipotesi nulla H_0 quando essa è falsa. La probabilità di commettere un errore di tipo II è indicata con β.

Da questo, si può costruire una tabella di contingenza che riassume i possibili esiti di una classificazione binaria:

	Ipotesi nulla vera	**Ipotesi nulla falsa**
Rifiuto	Errore di tipo I (α)	Vero positivo
Non rifiuto	Vero negativo	Errore di tipo II (β)

Tabella 6.1: Tabella di contingenza per i test di ipotesi.

6.5.2 T-test a un campione

Il t-test a un campione (one-sample t-test) è un test statistico utilizzato per determinare se la media di un campione differisce significativamente da un valore specifico della popolazione. Questo test è particolarmente utile quando la varianza della popolazione è sconosciuta e il campione è di dimensioni ridotte.

6.5.3 T-test a due campioni

Il t-test a due campioni (two-sample t-test) è un test statistico utilizzato per confrontare le medie di due gruppi indipendenti e determinare se esistono differenze significative tra di esse. Questo test è utile quando si vuole valutare l'effetto di un trattamento o di una condizione su due gruppi distinti.

6.5.4 Test χ^2 per indipendenza

Il test χ^2 per indipendenza è un test statistico utilizzato per determinare se esiste una relazione significativa tra due variabili categoriali. Questo test confronta le frequenze osservate in un campione con le frequenze attese se le due variabili fossero indipendenti. Utilizza il test di ipotesi, formulando come ipotesi nulla H_0 l'indipendenza tra le due variabili e come ipotesi alternativa H_a la dipendenza tra di esse.

È spesso accompagnato dalla statistica di Cramér, che misura la forza dell'associazione tra le due variabili categoriali.

6.5.5 Test χ^2 di bontà di adattamento

Il test χ^2 di bontà di adattamento è un test statistico utilizzato per determinare se un insieme di dati osservati si discosta significativamente da un modello teorico atteso. Questo test è spesso utilizzato per verificare se una distribuzione di frequenze osservate si adatta a una distribuzione attesa, come la distribuzione uniforme o la distribuzione normale. L'ipotesi nulla H_0 in questo caso afferma che non ci sono differenze significative tra le frequenze osservate e quelle attese, mentre l'ipotesi alternativa H_a suggerisce che ci sono differenze significative.

6.5.6 Test di correlazione di Pearson

Il test di correlazione di Pearson è un test statistico utilizzato per misurare la forza e la direzione della relazione lineare tra due variabili continue. Il coefficiente di correlazione di Pearson, denotato come r, varia tra -1 e 1, dove valori vicini a 1 indicano una forte correlazione positiva, valori vicini a -1 indicano una forte correlazione negativa, e valori vicini a 0 indicano nessuna correlazione lineare (sotto-sezione 4.2.3). L'ipotesi nulla H_0 afferma che non esiste una correlazione significativa tra le due variabili, mentre l'ipotesi alternativa H_a suggerisce che esiste una correlazione significativa.

6.5.7 Test di correlazione di Spearman

Il test di correlazione di Spearman, d'altra parte, è un test non parametrico utilizzato per misurare la forza e la direzione della relazione monotona tra due variabili ordinali o continue. Il coefficiente di correlazione di Spearman, denotato come ρ (rho), varia anch'esso tra -1 e 1, con interpretazioni simili a quelle del coefficiente di Pearson (sotto-sezione 4.2.4). L'ipotesi nulla H_0 afferma che non esiste una correlazione monotona significativa tra le due variabili, mentre l'ipotesi alternativa H_a suggerisce che esiste una correlazione monotona significativa.

6.6 Valutare quando un campione è distribuito normalmente

Per valutare se un campione di dati segue una distribuzione normale, si possono utilizzare diversi metodi statistici e grafici.

6.6.1 Grafici Q-Q

I grafici Q-Q (quantile-quantile) sono strumenti grafici per confrontare la distribuzione di un campione con una distribuzione teorica (ad esempio una normale). Si mettono sull'asse orizzontale i quantili teorici e sull'asse verticale i quantili osservati nel campione.

- Se i punti stanno vicino a una linea retta, il campione segue bene la distribuzione teorica.
- Se i punti si allontanano dalla linea in modo sistematico, i dati non seguono quella distribuzione.

Un esempio di Q-Q plot rispetto alla normale è mostrato in figura 6.9.

All'inizio interpretare un Q-Q plot non è sempre immediato. Per questo spesso si confrontano diversi casi tipici, così da riconoscere pattern ricorrenti (figura 6.10).

6.6.2 Test di normalità di Shapiro-Wilk

Il test di Shapiro-Wilk è un test statistico utilizzato per valutare se un campione di dati segue una distribuzione normale. È usato principalmente quando si ha un campione di dimensioni ridotte (tipicamente $n \leq 2000$). Il test funziona calcolando una statistica W che confronta l'ordine dei dati osservati con l'ordine atteso se i dati fossero normalmente

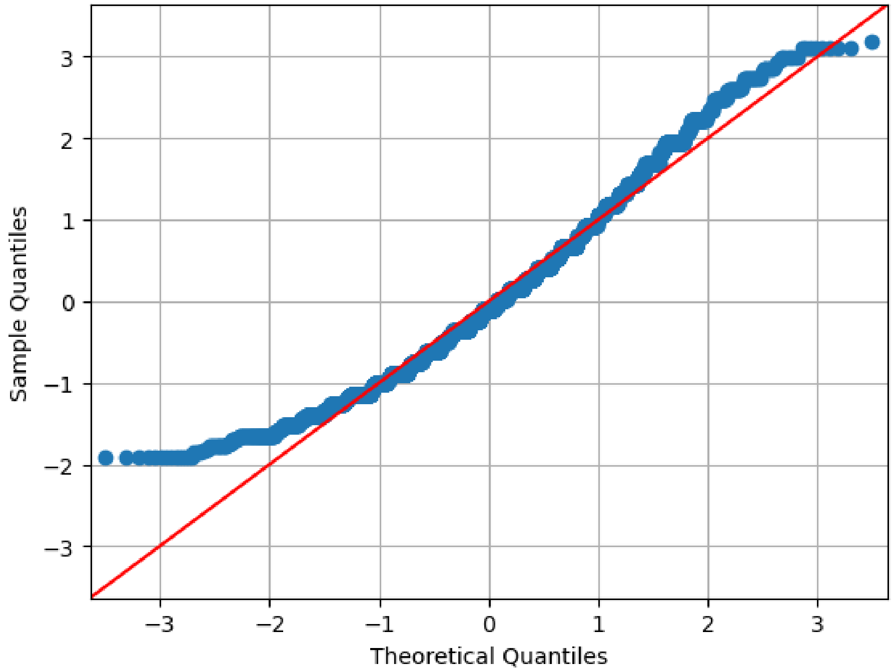

Figura 6.9: Q-Q plot rispetto alla normale: più i punti seguono la linea rossa, più i dati possono essere considerati circa normali.

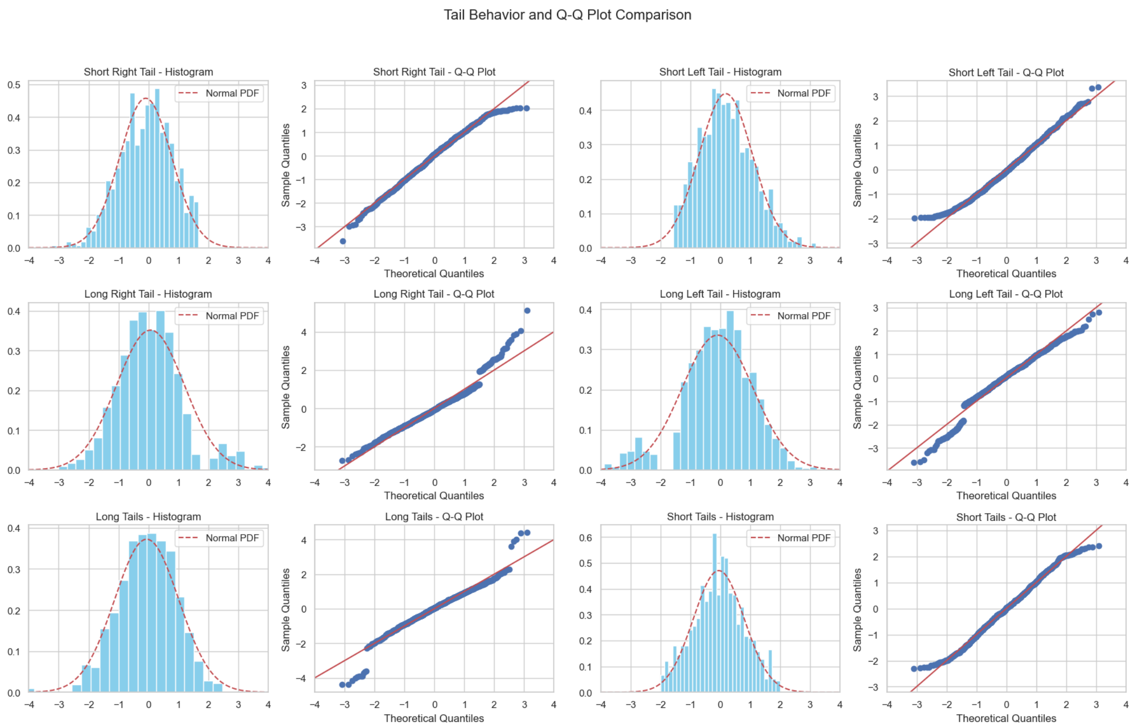

Figura 6.10: Relazione tra forma delle code e Q-Q plot. Code troppo leggere (*short tails*) piegano verso l'interno; code pesanti (*long tails*) si incurvano verso l'esterno; una coda destra lunga fa salire la parte destra del Q-Q plot sopra la linea; una coda sinistra lunga fa scendere la parte sinistra sotto la linea.

distribuiti. Se il valore di W è significativamente basso, si rifiuta l'ipotesi nulla H_0 che i dati seguono una distribuzione normale.

6.6.3 Test K^2 di D'Agostino

Per campioni grandi ($n \geq 50$) si può utilizzare il test K^2 di D'Agostino, che valuta la normalità basandosi su due misure: la skewness (asimmetria) e la kurtosis (appiattimento). Il test calcola una statistica K^2 combinando queste due misure e confronta il risultato con una distribuzione χ^2. Se il valore di K^2 è significativamente alto, si rifiuta l'ipotesi nulla H_0 che i dati seguono una distribuzione normale.

Capitolo 7

Analisi predittiva

L'analisi predittiva è una branca dell'analisi dei dati che utilizza tecniche statistiche, di machine learning e di data mining per prevedere eventi futuri basandosi su dati storici.

I principali obiettivi sono:

- Fare inferenza sulle relazioni tra variabili.
- Costruire modelli che possono essere usati per fare predizioni su nuovi dati.

7.1 Modello

Prima di continuare, è necessario dare la definizione di "modello".

> *Un modello è una rappresentazione semplificata di un sistema complesso, che cattura le caratteristiche essenziali del sistema per permettere l'analisi e la previsione del suo comportamento.*
>
> Definizione 7.1

Si può vedere una analogia di un modello come una cartina geografica: non rappresenta ogni dettaglio del territorio, ma fornisce informazioni sufficienti per orientarsi e pianificare un percorso. Ma questo porta a una conclusione, quella cartina è tecnicamente **sbagliata**, poiché non rappresenta a pieno il territorio.
Lo statistico George Box ha detto:

> "All models are wrong, but some are useful."

Ovvero, tutti i modelli sono sbagliati in quanto semplificazioni della realtà, ma alcuni possono essere utili per fare previsioni accurate e prendere decisioni informate. Per esempio, utilizzando un modello che predice il BMI (Body Mass Index) di una popolazione possiamo fare delle previsioni solo su una parte della popolazione, ma non su ogni singolo individuo (come si vede in figura 7.1).

7.1.1 Modelli predittivi

Quando facciamo analisi predittiva possiamo identificare:

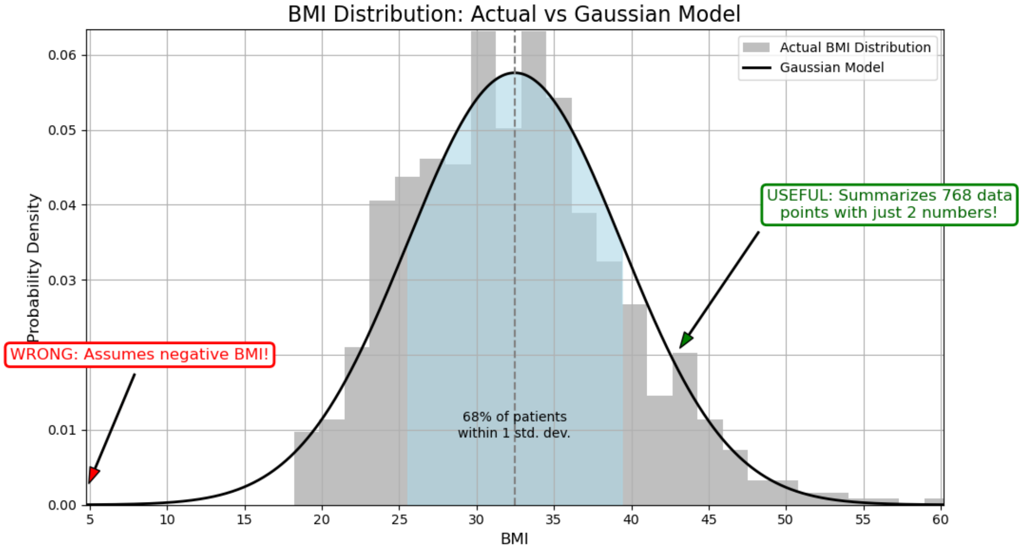

Figura 7.1: Distribuzione del BMI in una popolazione. Istogramma reale composto da $n = 768$ individui (barre grigie) a confronto con un modello gaussiano (linea nera). La Gaussiana riassume con media e deviazione una parte della popolazione in modo corretto, ma assegna probabilità anche a BMI negativi.

- Una variabile Y detta **Variabile dipendente** o *variabile di risposta*, che rappresenta l'output che vogliamo prevedere.
- Un vettore di variabili $X = (X_1, X_2, \ldots, X_p)$ dette **Variabili indipendenti** o *predittori*, che rappresentano gli input utilizzati per fare la previsione.

Possiamo esprimere il modello predittivo come una funzione matematica, con una forma generale:

$$Y = f(X) + \epsilon$$

dove f è il modello che descrive la relazione tra le variabili indipendenti e la variabile dipendente, ed ϵ è un termine di errore che rappresenta la variabilità non spiegata dal modello.

7.2 Predizione vs Spiegazione

Un aspetto importante dell'analisi predittiva è la distinzione tra **predizione** e **spiegazione**.

7.2.1 Predizione

La predizione si concentra sulla capacità del modello di fare previsioni accurate su nuovi dati, indipendentemente dal fatto che il modello sia interpretabile o meno.

Definizione 7.2

Quindi risponde a una semplice domanda: "Cosa avverrà?". Nel caso particolare della predizione, l'accuratezza del modello è la metrica più importante.

Per fare un esempio, si pensi a un modello per prevedere il prezzo delle case basandosi su caratteristiche come la posizione, la dimensione e il numero di stanze. Anche se la rete neurale può essere complessa e difficile da interpretare, se riesce a fare previsioni accurate sui prezzi delle case, allora è considerata un buon modello predittivo.

7.2.2 Spiegazione

La spiegazione si concentra sulla comprensione delle relazioni tra le variabili e sull'interpretabilità del modello.

Definizione 7.3

In questo caso il modello deve essere al 100% interpretabile, rispondendo alla domanda: "Perché avverrà?". Qui l'accuratezza del modello è meno importante rispetto alla capacità di spiegare i fenomeni osservati.

Per fare un esempio, si consideri un modello per analizzare come una malattia e da quali fattori essa dipende (età, stile di vita, genetica, ecc.). In questo caso, un modello semplice è preferibile, anche se meno accurato, perché permette ai medici di comprendere i fattori di rischio e di prendere decisioni informate sui trattamenti.

7.2.3 Compromesso tra Predizione e Spiegazione

Spesso esiste un compromesso tra predizione e spiegazione, in quanto i due obiettivi non sono spesso mutuamente esclusivi. Un modello potente infatti, se il problema lo permette, deve sia fare inferenza che predizione.

Si pensi a un dataset di rischio del diabete di tipo 1: un modello potrebbe essere utilizzato per prevedere la probabilità che un individuo sviluppi la malattia (predizione), ma potrebbe anche essere utile per identificare i fattori di rischio associati alla malattia (spiegazione).

7.3 Statistica vs Machine Learning

Un altro aspetto importante dell'analisi predittiva è la distinzione tra **statistica** e **machine learning**.

7.3.1 Approccio statistico

L'approccio statistico è concentrato sul capire il modello e sull'inferenza. Si comporta come in figura 7.2, come una "glass box", dove il funzionamento interno del modello è trasparente e interpretabile. Le sue metodologie includono:

- Affidarsi alle assunzioni fatte sui dati (distribuzioni, linearità, indipendenza, ecc.).
- Utilizzare molto il testo di ipotesi, con p-values e intervalli di confidenza.
- Prediligere modelli semplici e interpretabili.

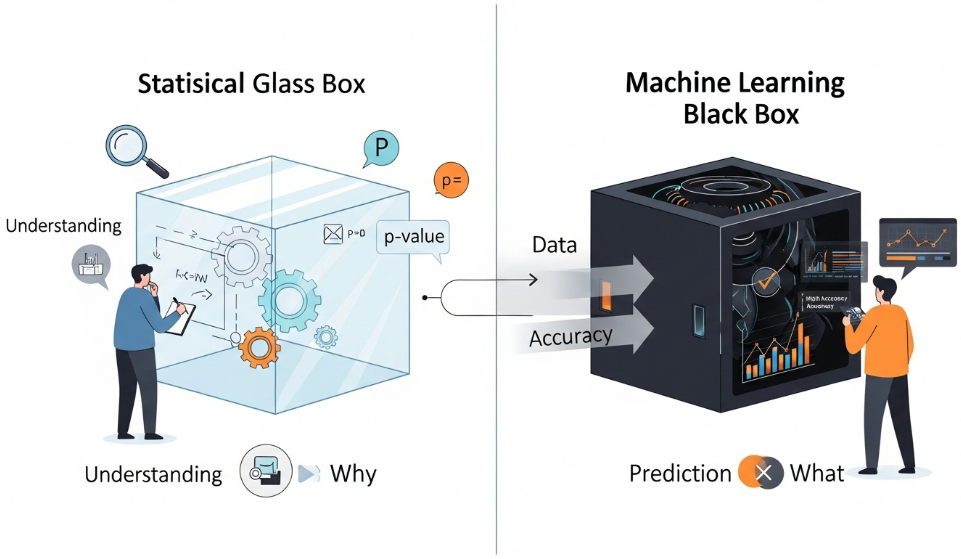

Figura 7.2: Confronto tra *glass box* statistica e *black box* di ML: la statistica privilegia interpretabilità e spiegazione del *perché*, mentre il machine learning privilegia accuratezza predittiva sul *che cosa* a partire dai dati, spesso con modelli opachi.

7.3.2 Approccio di Machine Learning

L'approccio di Machine Learning ha un focus primario sulla predizione accurata. Si comporta come in figura 7.2, come una "black box", dove il funzionamento interno del modello può essere complesso e difficile da interpretare. Le sue metodologie includono:

- Le performance del modello si vedono sui dati, senza fare molte assunzioni a priori.
- Si affida allo split dei dati in training, validation e test set per simulare la generalizzazione.
- Non vengono misurate le performance con p-values, ma con metriche di accuratezza predittiva sul test-set.
- L'intepretabilità è una cosa in più, non un requisito.

7.3.3 Trade-Off di Complessità-Interpretabilità

Quando si lavora su modelli predittivi, statistici o di machine learning, spesso si deve affrontare un trade-off tra complessità e interpetabilità del modello:

Modelli semplici : un modello semplice (come la regressione lineare) è facile da interpretare e spiegare, ma potrebbe non catturare tutte le complessità dei dati, portando a una minore accuratezza predittiva. Come già detto prima, è esattamente la prerogativa della statistica.

Modelli complessi : un modello complesso (come le reti neurali profonde) può catturare meglio le complessità dei dati e fornire previsioni più accurate, ma spesso è difficile da interpretare e spiegare. Questo è il punto di forza del machine learning.

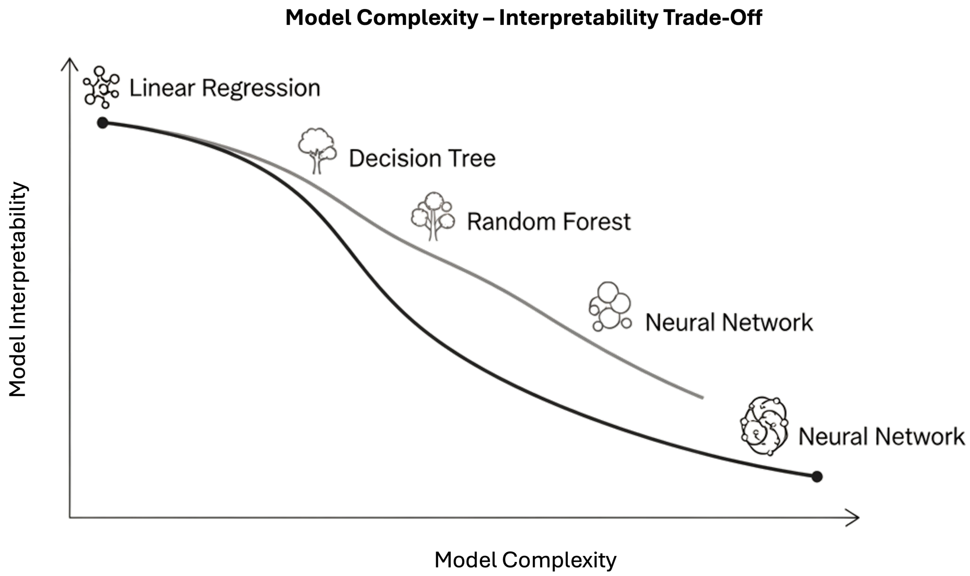

Figura 7.3: Accuratezza vs complessità del modello: dalla Regressione Lineare a Decision Tree e Random Forest fino alle Reti Neurali, l'errore tende a diminuire man mano che cresce la complessità (ma aumenta il costo/opaquezza del modello).

7.4 Tipologie di problema

Una volta stabilito se vogliamo capire i dati o fare predizione, dobbiamo identificare il tipo di problema che stiamo affrontando. I principali tipi di problemi nel machine learning sono:

- **Supervised learning**: il modello viene addestrato su un dataset etichettato, dove ogni esempio di input ha una corrispondente etichetta di output. L'obiettivo è imparare una funzione che mappa gli input agli output corretti.
- **Unsupervised learning**: il modello viene addestrato su un dataset non etichettato, dove non ci sono etichette di output. L'obiettivo è trovare strutture nascoste nei dati, come cluster o pattern.

7.4.1 Regressione

La regressione è un tipo di problema di supervised learning in cui l'obiettivo è prevedere un valore y continuo. Dobbiamo trovare quindi il miglior "fit" di una funzione all'interno dei nostri dati, per descrivere a pieno la relazione tra le variabili indipendenti e la variabile dipendente.

Per fare un esempio legato alla figura 7.4, si consideri un dataset che contiene informazioni sulle case, come la dimensione in metri quadri, il numero di stanze e il prezzo di vendita. L'obiettivo è prevedere il prezzo di una casa basandosi sulle sue caratteristiche. Un modello di regressione potrebbe essere utilizzato per trovare la relazione tra la dimensione della casa e il suo prezzo, permettendo di fare previsioni sui prezzi delle case in base alle loro dimensioni.

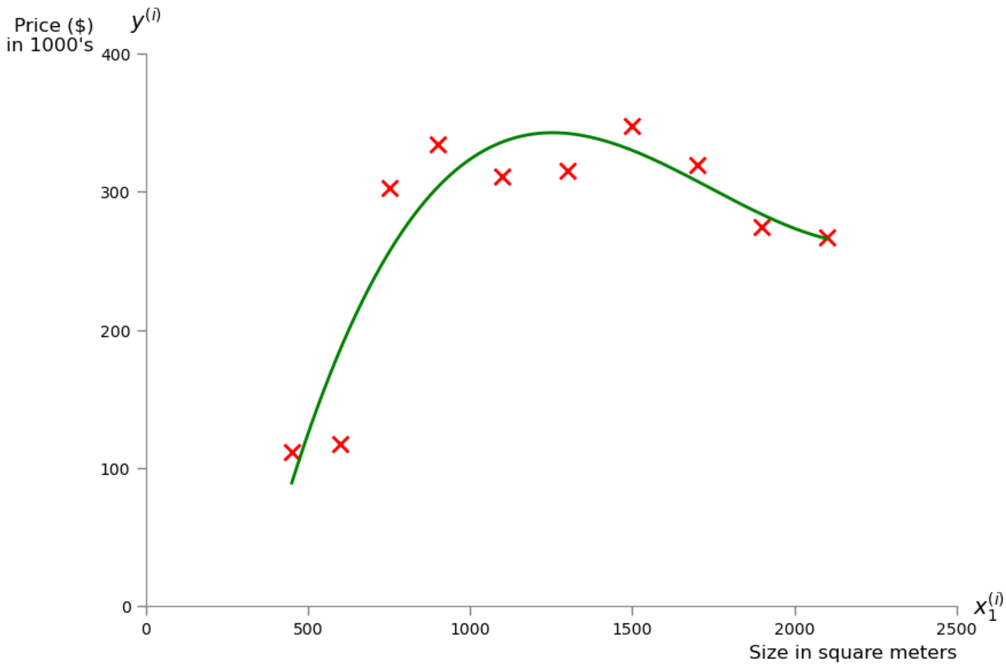

Figura 7.4: Esempio di problema di regressione: prevedere il prezzo di una casa in base ai metri quadri.

7.4.2 Classificazione

La classificazione è un altro tipo di problema di supervised learning in cui l'obiettivo è prevedere una categoria o classe discreta y(etichetta). In questo caso, durante il training, il modello impara a mappare gli input alle classi corrette.

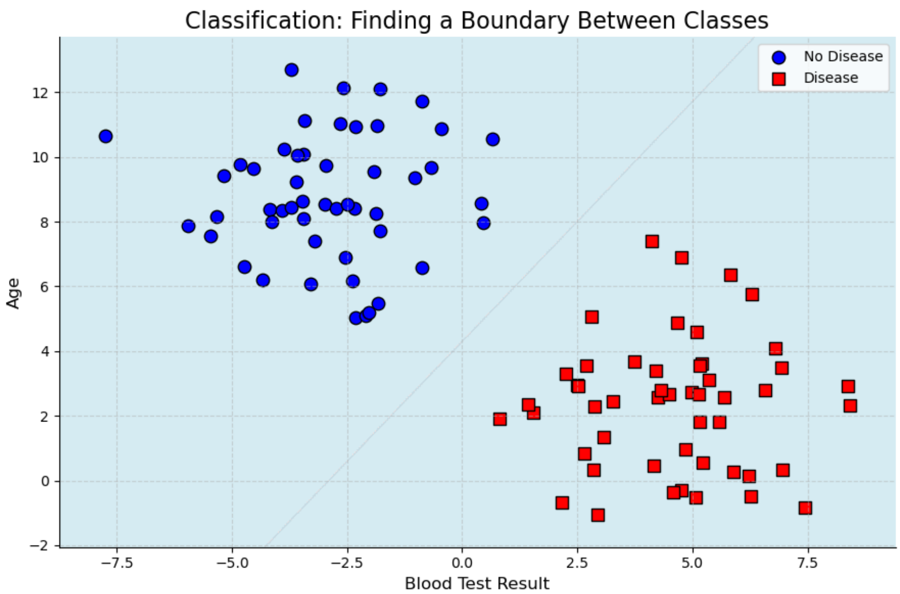

Figura 7.5: Esempio di problema di classificazione: capire se una malattia è presente oppure no in base all'età e a un'analisi del sangue.

Continuando l'esempio in figura 7.5, si consideri un dataset che contiene informazioni sui pazienti, come l'età, il sesso e i risultati di un'analisi del sangue. L'obiettivo è prevedere se un paziente ha una certa malattia (ad esempio, diabete) basandosi sulle sue caratteristiche. Nel caso particolare dell'immagine, notiamo che può essere separata da una retta per risolvere il problema.

7.4.3 Clustering

Il clustering è un problema di unsupervised learning in cui l'obiettivo è raggruppare i dati in cluster basandosi sulla somiglianza tra gli esempi. In questo caso, il modello cerca di identificare strutture nascoste nei dati senza l'uso di etichette di output.

Figura 7.6: Esempio di problema di clustering: raggruppare i clienti in segmenti basandosi sui loro comportamenti di acquisto.

Il modello, unsupervised, viene addestrato ad identificare gruppi con caratteristiche simili. Nella figura 7.6 si considera un esempio dove si vogliono raggruppare quei gruppi di persone in base all'età e a una feature chiamata "Spending Score", che indica quanto una persona spende in un negozio.

7.5 Modelli parametrici vs Modelli non parametrici

I modelli possono differirsi anche in base a come rappresentano la funzione $f(X)$ che mappa gli input agli output. Possiamo distinguere tra modelli parametrici, che assumono una forma funzionale specifica, e modelli non parametrici, che non fanno assunzioni rigide sulla forma della funzione.

7.5.1 Modelli parametrici

Nello specifico, un modello parametrico è caratterizzato da un numero fisso di parametri che definiscono la funzione $f(X)$. Questi modelli sono spesso più semplici e veloci da addestrare, ma possono essere limitati nella loro capacità di catturare la complessità dei dati. Qui il compito principale è quello di modificare i valori dei parametri per adattare il modello ai dati di training.

Un esempio comune di modello parametrico è la regressione lineare, dove la funzione $f(X)$ è una combinazione lineare delle variabili indipendenti:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

dove $\beta_0, \beta_1, \ldots, \beta_p$ sono i parametri del modello.

Ha come pro il fatto che è semplice, veloce e richiede meno dati. Tuttavia questi modelli sono "biased": le assunzioni iniziali possono portare a errori sistematici se i dati non seguono quelle assunzioni, inoltre sono più propensi all'*underfitting*[1].

7.5.2 Modelli non parametrici

Un modello non parametrico, invece, non assume una forma funzionale specifica per la funzione $f(X)$. Questi modelli sono più flessibili e possono adattarsi meglio alla complessità dei dati, ma possono essere più lenti da addestrare e richiedere più dati. Qui il compito principale è quello di memorizzare i dati di training e utilizzare questi dati per fare predizioni sui nuovi input.

Un esempio comune di modello non parametrico è il k-Nearest Neighbors (k-NN), dove la predizione per un nuovo input viene fatta basandosi sui k esempi più vicini nel dataset di training.

Ha come pro il fatto che è flessibile e può catturare relazioni complesse nei dati. Tuttavia questi modelli sono "low-bias": non fanno assunzioni rigide sui dati, ma possono essere più propensi all'*overfitting*[2].

7.6 Learning

Alla base dei problemi di Machine Learning c'è il concetto di **learning**, ovvero l'apprendimento di una funzione $f(X)$ dai dati.

[1] L'underfitting è un fenomeno causato da un modello troppo "semplice", che non riesce a catturare la complessità dei dati.

[2] L'overfitting è un fenomeno causato da un modello troppo "complesso", che si adatta troppo bene ai dati di training, ma non generalizza bene su nuovi dati.

7.6.1 Definizione formale

Siano $\mathcal{X}$ e $\mathcal{Y}$ gli spazi, rispettivamente, di input e di output al modello. Possono essere visti come spazi di variabili aleatorie, in particolare una variabile X e una Y con osservazioni $x \in \mathcal{X}, y \in \mathcal{Y}$. L'obiettivo è trovare una funzione che risponde alla nostra ipotesi $h \in \mathcal{H}$, dove $\mathcal{H}$ è lo spazio delle ipotesi (funzioni candidate), ovvero una funzione:

$$h : \mathcal{X} \to \mathcal{Y}$$

che approssima la relazione tra X e Y. Possiamo scrivere una funzione che lo fa non in modo preciso, ma approssimando a una predizione $\hat{y} \approx y \in \mathcal{Y}$:

$$\hat{y} = h(x)$$

Definizione 7.4

Esempio 1: non parametrico. Per fare un esempio, immaginiamo $\mathcal{X} = \mathbb{R}^m$ e $\mathcal{Y} = \{0, 1\}$, ovvero uno spazio di input a m variabili che rappresentano lo spazio dei risultati di un'analisi del sangue e uno spazio di output binario che rappresenta la presenza o meno di una malattia. L'obiettivo è trovare una funzione h che mappa i risultati dell'analisi del sangue alla presenza o meno della malattia.

Esempio 2: parametrico. In un altro esempio, immaginiamo di dover classificare le email in spam e non: in questo caso possiamo definire $\mathcal{X}$ come lo spazio delle email rappresentate da vettori di caratteristiche (come la frequenza di certe parole chiave) e $\mathcal{Y} = \{0, 1\}$ come lo spazio delle etichette (spam o non spam). L'obiettivo è trovare una funzione h che mappa le caratteristiche delle email alle etichette corrette:

$$h(x) = \begin{cases} 1 & f(x) > \theta \\ 0 & f(x) \leq \theta \end{cases}$$

Dove θ è esattamente il nostro parametro, in questo caso una *soglia*.

7.6.2 Il processo di Learning

Per trovare la funzione h che meglio approssima la relazione tra X e Y, dobbiamo avere un modo per valutare quanto bene una funzione h si adatta ai dati. Questo viene fatto utilizzando una funzione di perdita (loss function) $L(y, \hat{y})$, che misura l'errore tra il valore reale y e la predizione $\hat{y}$ fatta dal modello.

In realtà vorremmo calcolare l'errore atteso (o rischio) del modello su tutta la distribuzione dei dati, ma non avendo accesso a questa distribuzione, possiamo solo stimare l'errore empirico $R(h)$ definita sulla nostra ipotesi h^*[3] come **l'errore atteso** sotto la distribuzione

[3] Indichiamo la nostra ipotesi con un asterisco h^* perché rappresenta la migliore funzione di approssimazione possibile dati i dati a disposizione.

di dati $P(X,Y)$:

$$R(h) = \mathbb{E}_{(X,Y)\sim P}[L(Y, h(X))]$$

In questo caso, si definisce **obiettivo del learning statistico** come la risoluzione del problema di ottimizzazione seguente:

$$h^* = \arg\min_{h\in\mathcal{H}} R(h)$$

7.6.3 ERM: Empirical Risk Minimization

Esiste un problema: non possiamo calcolare la funzione di rischio $R(h)$ perché non conosciamo la vera distribuzione $P(X,Y)$. Tuttavia, possiamo stimare il rischio empirico $R_{emp}(h)$ utilizzando un dataset di training[4] di N esempi che sono un campione rappresentativo della popolazione. Si definisce TR training set, come l'insieme delle coppie:

$$\mathrm{TR} = \{(x_i, y_i)\}_{i=1}^N$$

Grazie a questo e alla **legge dei grandi numeri**[5], possiamo stimare il rischio empirico come:

$$R_{\mathrm{emp}}(h) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, h(x_i))$$

Quindi, possiamo risolvere il problema di ottimizzazione empirica:

$$h_{\mathrm{emp}} = \arg\min_{h\in\mathcal{H}} R_{\mathrm{emp}}(h)$$

Questo approccio è noto come **Empirical Risk Minimization** (ERM), ovvero la minimizzazione del rischio empirico.

7.7 Capacità del modello

Un aspetto cruciale nell'analisi predittiva è la capacità del modello, ovvero la sua capacità di adattarsi ai dati di training e di generalizzare a nuovi dati. Possiamo definire formalmente la capacità come una relazione allo spazio e alla "ricchezza" (intesa come complessità) dello spazio delle ipotesi $\mathcal{H}$ dal quale il modello può scegliere la funzione h. Esistono modelli:

A bassa capacità : modelli con uno spazio delle ipotesi limitato, che possono adattarsi solo a funzioni semplici. Questi modelli sono meno propensi all'overfitting, ma possono soffrire di underfitting. Un esempio è una classificazione lineare con una retta: sicuramente andrà bene per dati linearmente separabili, ma nei dati più complessi non riuscirà a catturare le relazioni tra le variabili.

Ad alta capacità : modelli con uno spazio delle ipotesi ampio, che possono adattarsi a funzioni complesse. Questi modelli sono più propensi all'overfitting, ma possono

[4] Il dataset di training è un insieme di dati utilizzati per addestrare il modello.

[5] La legge dei grandi numeri afferma che, al crescere del numero di osservazioni, la media campionaria converge alla media della popolazione.

catturare meglio le relazioni nei dati. Un esempio è una classificazione non lineare con un polinomio di grado 10: sicuramente riuscirà a catturare le relazioni nei dati complessi, ma rischia di adattarsi troppo ai dati di training e di non generalizzare bene su nuovi dati.

7.7.1 Misurare la capacità del modello

Per misurare la capacità de modello si deve misurare quanto bene il modello generalizza sui nuovi dati. Questa metrica è solitamente una funzione che valuta l'accuratezza del modello su un dataset di test separato dal training set. Un modo comune, nel caso della regressione, è utilizzare l'errore quadratico medio (Mean Squared Error, MSE):

$$R_{\text{emp}}(h) = \frac{1}{M}\sum_{j=1}^{M}(y_j - h(x_j))^2$$

Generalmente però, questa metrica, non basta. Perché si potrebbe pensare di utilizzare come mappatura di una regressione il valore del training set:

$$\hat{h}(x) = y \qquad (x, y) \in \text{TR}$$

ottenendo un errore di 0 sul training set, ma un errore altissimo sul test set. Per questo motivo si usano tecniche di validazione incrociata (cross-validation) per stimare la capacità del modello in modo più robusto.

Questo porta alla definizione di due concetti già visti in breve, **underfitting** e **overfitting**:

Underfitting Si verifica quando un modello è troppo semplice per catturare le relazioni nei dati, portando a prestazioni scadenti sia sul training set che sul test set. Per esempio, dei dati che seguono una distribuzione quadratica vengono approssimati con una retta (7.7, sinistra).

Overfitting Si verifica quando un modello è troppo complesso e si adatta troppo ai dati di training, catturando il rumore invece della vera relazione tra le variabili. Questo porta a buone prestazioni sul training set ma scarse sul test set. Per esempio, dei dati che seguono una distribuzione quadratica vengono approssimati con un polinomio di grado 5 (7.7, destra).

7.7.2 Bias e Varianza

Per valutare le prestazioni di un modello possiamo usare i valori del bias e della varianza per fare delle stime su come il modello si comporta sui dati.

Bias. Il bias è **l'errore sistematico** che il modello commette sui dati. Un modello con alto bias tende a sottostimare la complessità del problema, portando a errori elevati sia sul training set che sul test set. Questo fenomeno è noto come **underfitting**. Nell'immagine 7.7, il grafico a sinistra mostra un esempio di underfitting, dove il modello lineare non riesce

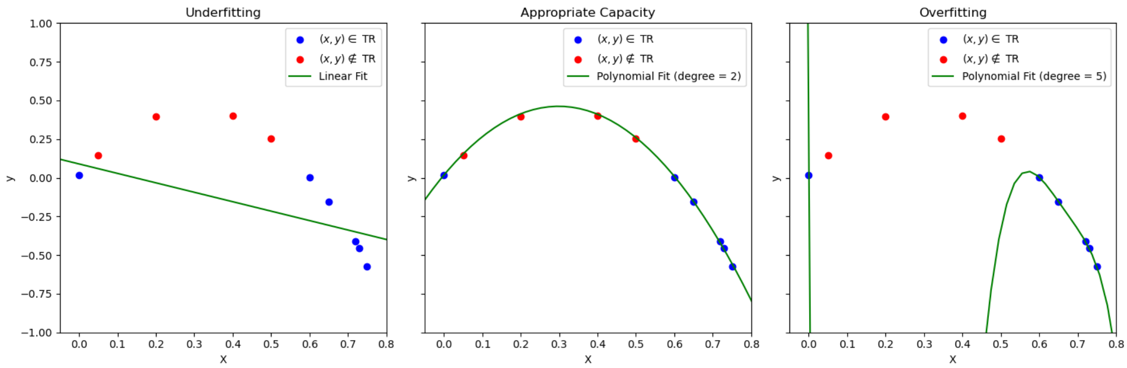

Figura 7.7: Underfitting—capacità adeguata—overfitting in regressione polinomiale: punti blu = dati di training (TR), punti rossi = fuori da TR; la linea verde mostra il fit del modello: lineare (sinistra), polinomio di grado 2 (centro) e polinomio di grado 5 (destra).

a catturare la relazione tra le variabili e commette un errore sistematico sia sul training set (punti blu) che sul test set (punti rossi).

Varianza. La varianza rappresenta, invece, la sensibilità del modello alle variazioni nei dati di training. Un modello con alta varianza tende a sovradattarsi ai dati di training, catturando il rumore invece della vera relazione tra le variabili. Questo porta a errori bassi sul training set ma elevati sul test set, fenomeno noto come **overfitting**. Nell'immagine 7.7, il grafico a destra mostra un esempio di overfitting, dove il modello polinomiale è troppo complesso e si adatta troppo strettamente ai dati di training, risultando in prestazioni scadenti sui dati di test[6].

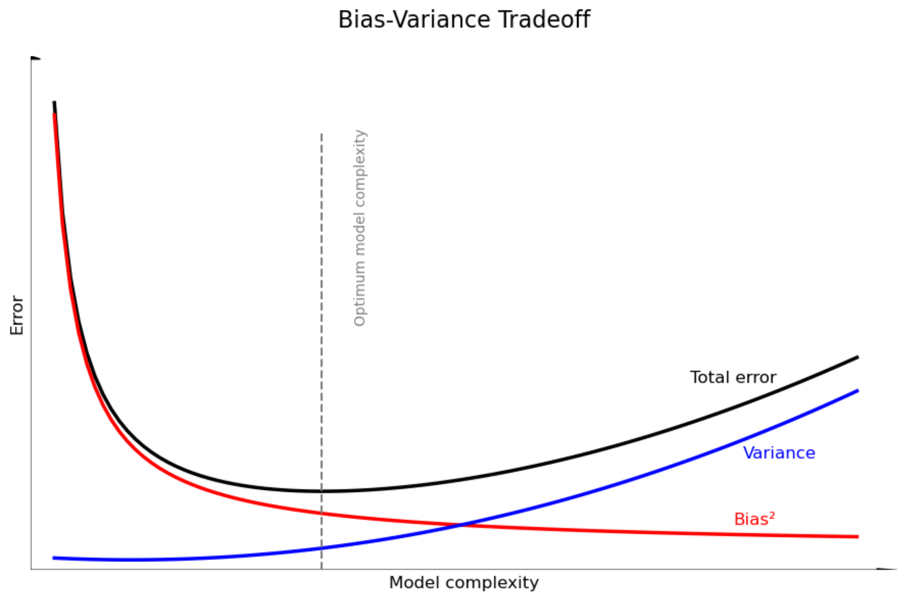

Figura 7.8: Trade-off tra bias e varianza in funzione della complessità del modello.

[6]Si noti che basterebbe rimuovere un punto del training set per far cambiare completamente il modello, in quanto altamente sensibile ai dati di input.

7.7.3 Parametri vs Iperparametri

Per controllare la capacità del modello, possiamo agire su due tipi di parametri:

Parametri I parametri sono i valori che il modello impara durante il processo di training. Questi parametri definiscono la funzione h che mappa gli input agli output. Per esempio, in una regressione lineare, i parametri sono i coefficienti della retta.

Iperparametri Gli iperparametri sono i valori che vengono impostati prima del processo di training e controllano il comportamento del modello. Questi iperparametri influenzano la capacità del modello e il modo in cui viene addestrato. Sono coefficienti che hanno fattori esterni, come il tasso di apprendimento, la profondità di un albero decisionale o il grado del polinomio nella funzione di regressione.

7.8 Selezione del modello

Il problema a questo punto diventa: come facciamo a far sì che l'algoritmo di learning (ERM) trovi i migliori parametri per il modello fornito un set di iperparametri.

7.8.1 Approccio 1: selezione statistica

Quando il nostro obiettivo principale è l'*inferenza* (il modello glass-box), selezionamo un modello in base a quanto spiega bene i dati in favore della semplicità. Questo approccio utilizza l'intero dataset **in una volta**, perché non cerchiamo di "predire" il futuro, ma di trovare la migliore spiegazione per i dati che osserviamo.

Si usano misure statistiche che bilanciano bontà di fit e complessità del modello, come:

- **p-values**: Facciamo un test sul valore di ogni variabile e potremmo rimuovere quelle con un p-value alto (ovvero quelle che creano rumore).
- $\mathbf{R^2}$: Questa misura indica la proporzione di varianza nella variabile dipendente che è spiegata dalle variabili indipendenti nel modello. Un valore di R^2 più alto indica un modello che spiega meglio i dati.

7.8.2 Approccio 2: selezione predittiva

Quando il nostro obiettivo principale è la *predizione* (il modello black-box), selezioniamo un modello in base a quanto bene predice nuovi dati, bilanciando accuratezza e complessità. Questo approccio utilizza tecniche di validazione incrociata per stimare la capacità del modello di generalizzare a nuovi dati.

Esiste infatti una "golden rule" (reegola d'oro) nei modelli predittivi:

> "The performance of a model on the data it was trained on is *irrelevant*. The only measure that matters is its performance on new, unseen data."

Ovvero, è inutile valutare un modello in base a quanto bene si adatta ai dati di training: l'unica metrica che conta è quanto bene si comporta su nuovi dati mai visti prima.

Per valutare le prestazioni dobbiamo ricordare che il rischio non è calcolabile, quindi usiamo il rischio empirico sul test set. Generalmente si utilizzano *misure di performance* oppure *misura di errori* (loss functon):

- **Loss Function**: la loss function è molto utile nei casi di training, in quanto misura l'errore tra la predizione del modello e il valore reale. Per esempio, nella regressione si può usare l'errore quadratico medio (MSE), mentre nella classificazione si può usare la log-loss.
- **Metriche di performance**: le metriche di performance sono utilizzate per valutare le prestazioni del modello su un dataset di test.

Una misura di performance p misura l'insieme delle verità Y e delle predizioni corrispondenti $\hat{Y}$ per un certo dataset:

$$p : \mathcal{Y}^N \times \mathcal{Y}^N \to \mathbb{R}$$

La differenza con il rischio empirico è che questa misura valuta direttamente le prestazioni del modello, senza passare per una funzione di perdita.

7.8.3 Validazione Holdout

La validazione holdout (o single split) è una tecnica semplice per stimare la capacità di generalizzazione di un modello. Consiste nel dividere il dataset in due parti: un training set e un test set. Il modello viene addestrato sul training set e poi valutato sul test set.

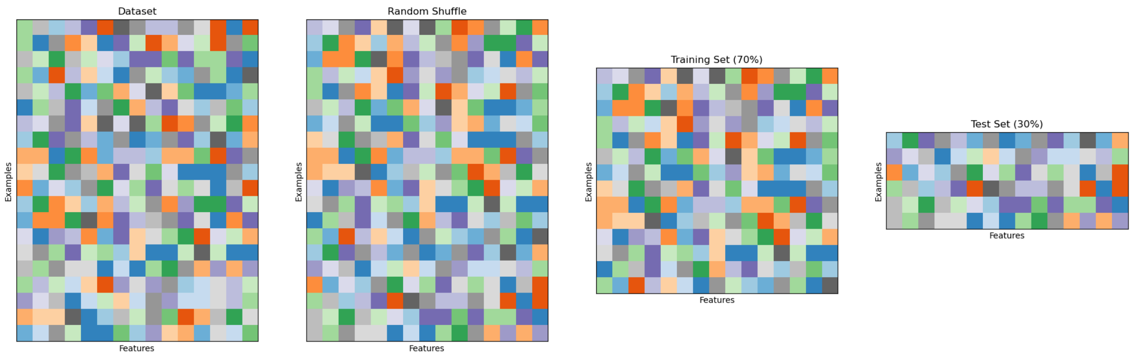

Figura 7.9: Validazione Holdout: il dataset viene prima mescolato, poi diviso in training set (70%) e test set (30%). Il modello viene addestrato sul training set e valutato sul test set per stimare la capacità di generalizzazione.

Questo paradigma è utile per modelli semplici e dataset grandi, in quanto un dataset più piccolo restituirebbe con qualità inferiore le stime delle prestazioni del modello.

7.8.4 K-Fold Cross-Validation

La K-Fold Cross-Validation risolve il problema di dataset piccoli: divide il dataset in diversi sottogruppi (folds) e utilizza ogni sottogruppo sia come test che come training set in diverse iterazioni. In particolare, il dataset viene diviso in K folds di dimensioni approssimativamente uguali. In ogni iterazione, uno dei folds viene utilizzato come test set,

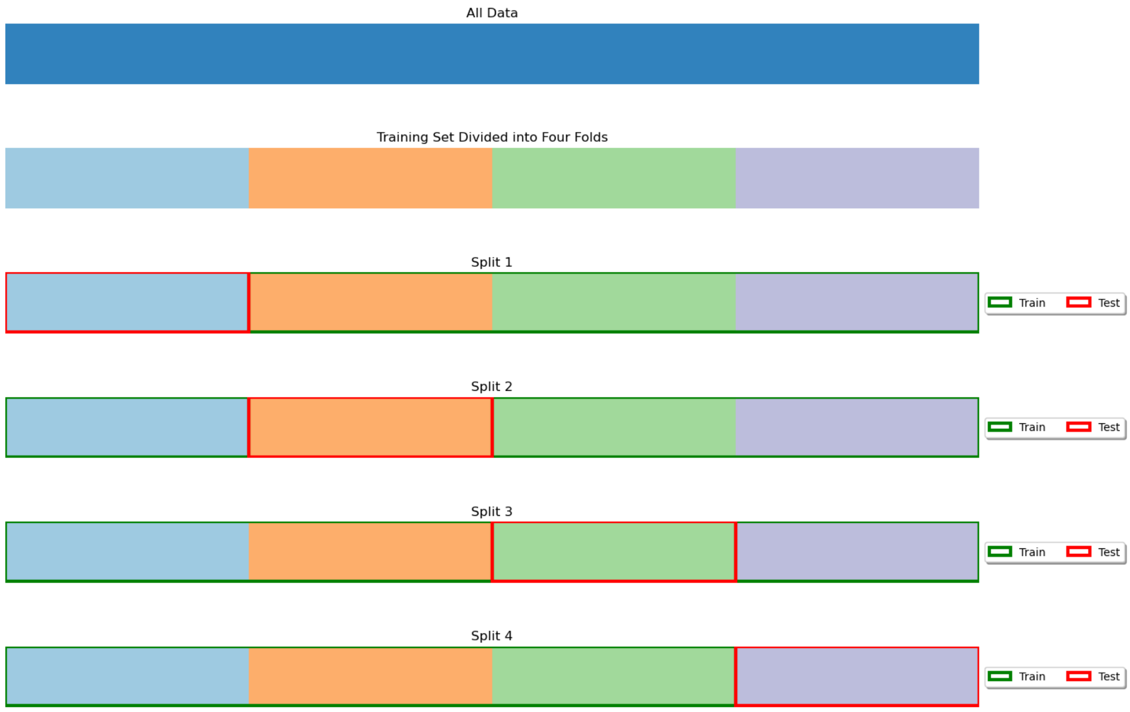

Figura 7.10: K-Fold Cross-Validation: il dataset viene diviso in K folds. In ogni iterazione, un fold viene utilizzato come test set e gli altri $K-1$ folds come training set. Questo processo viene ripetuto K volte, e le prestazioni del modello vengono mediate su tutte le iterazioni per ottenere una stima più robusta della capacità di generalizzazione.

mentre gli altri $K-1$ folds vengono utilizzati come training set. Questo processo viene ripetuto K volte, in modo che ogni sottogruppo venga utilizzato come test set una volta.

Alla fine del processo, le prestazioni del modello vengono mediate su tutte le iterazioni per ottenere una stima più robusta della capacità di generalizzazione del modello. Questo metodo è particolarmente utile quando si lavora con dataset di dimensioni limitate, in quanto consente di utilizzare tutti i dati disponibili sia per l'addestramento che per la valutazione del modello.

Rimangono due problemi:

- É una tecnica computazionalmente costosa, in quanto se il training di un modello richiede 1 giorno, questo paradigma ne fa impiegare 4.
- Gli iperparametri non vengono ottimizzati durante il processo di training.

7.8.5 Leave-One-Out Cross-Validation (LOOCV)

La Leave-One-Out Cross Validation (LOOCV) è una variante estrema della K-Fold Cross-Validation, in cui il numero di folds K è uguale al numero di esempi nel dataset. In altre parole, in ogni iterazione, un singolo esempio viene utilizzato come test set, mentre tutti gli altri esempi vengono utilizzati come training set. Questo processo viene ripetuto per ogni esempio nel dataset.

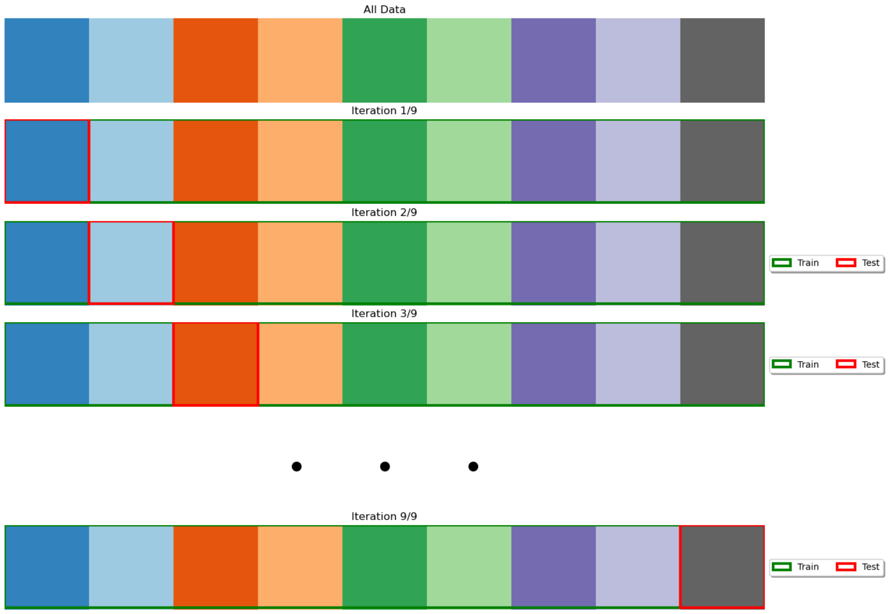

Figura 7.11: Leave-One-Out Cross-Validation (LOOCV): in ogni iterazione, un singolo esempio viene utilizzato come test set, mentre tutti gli altri esempi vengono utilizzati come training set. Questo processo viene ripetuto per ogni esempio nel dataset.

Questo approccio è molto utile quando si lavora con dataset molto piccoli, in quanto consente di utilizzare quasi tutti i dati disponibili per l'addestramento del modello in ogni iterazione. Tuttavia, la LOOCV soffre degli stessi problemi della K-Fold Cross Validation.

7.8.6 Ottimizzazione degli iperparametri

Il problema degli iperaparametri, con questi paradigmi, comunque persiste.

Grid Search. Alcuni algoritmi suggeriscono di utilizzare una **grid search**: un tipo di ricerca esaustiva sugli iperparametri, in cui si definisce una griglia di valori possibili per ogni iperparametro e si valuta il modello per ogni combinazione di valori nella griglia. Questo approccio risolve indubbiamente l'ottimizzazione degli iperparametri, ma può essere computazionalmente costoso, specialmente con un gran numero di iperparametri e valori da esplorare.

Validation set. Un altro approccio è quello di utilizzare un **validation set**: un sottoinsieme del dataset separato dal training set e dal test set, utilizzato per ottimizzare gli iperparametri. Per esempio, in una Validazione Holdout, si potrebbe dividere il dataset in tre parti: training set, validation set e test set. Il modello viene addestrato sul training set, gli iperparametri vengono ottimizzati sul validation set e infine le prestazioni del modello vengono valutate sul test set.

Ci sono tuttavia, combinazioni migliori. Si potrebbe effettuare, per esempio, una K-Fold Cross Validation dove si divide inizialmente il dataset in training/validation set e test set: il training/validation set dopo è sottoposto a una variante di K-Fold Cross Validation per

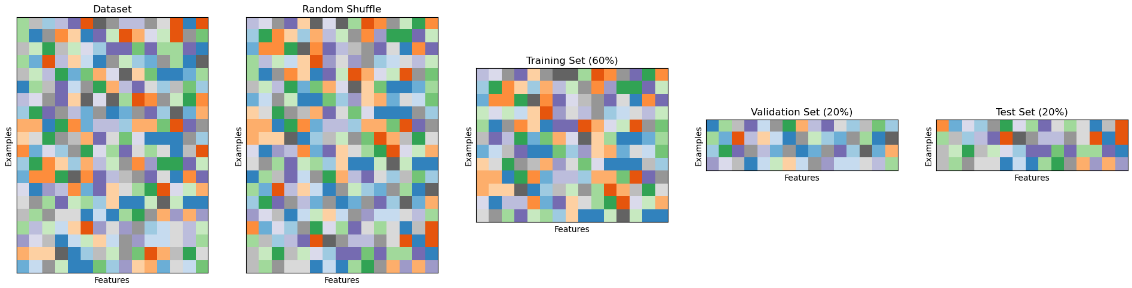

Figura 7.12: Utilizzo di un validation set per l'ottimizzazione degli iperparametri: il dataset viene diviso in training set (60%), validation set (20%) e test set (20%). Il modello viene addestrato sul training set, gli iperparametri vengono ottimizzati sul validation set e infine le prestazioni del modello vengono valutate sul test set.

ottimizzare gli iperparametri, mentre il test set viene usato solo alla fine per valutare le prestazioni finali del modello.

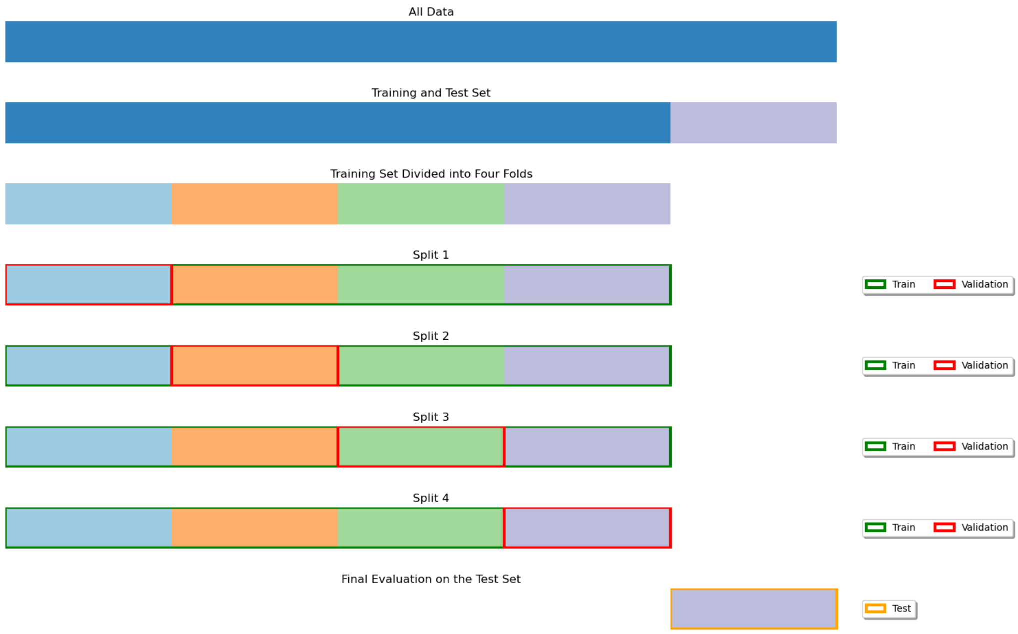

Figura 7.13: K-Fold Cross Validation con ottimizzazione degli iperparametri: il dataset viene diviso in training/validation set e test set. Il training/validation set viene sottoposto a K-Fold Cross Validation per ottimizzare gli iperparametri, mentre il test set viene utilizzato solo alla fine per valutare le prestazioni finali del modello.

Capitolo 8

Regressione lineare

La regressione lineare è una tecnica utilizzata per modellare la relazione tra una variabile dipendente e una o più variabili indipendenti. In particolare si rileva utile in tutti quei casi dove si vuole studiare l'effetto di una o più variabili esplicative su una variabile di interesse, oppure si vogliono fare previsioni sui valori futuri della variabile di interesse basandosi sui valori delle variabili esplicative.

8.1 Formalizzazione della regressione

La regressione lineare mira a studiare l'associazione *tra una variabile dipendente Y e una o più variabili indipendenti $X_1, X_2, \ldots, X_p$ definendo un modello matematico f tale che:*

$$Y = f(X) + \epsilon$$

Dove ϵ rappresenta l'errore o il **rumore** *nel modello. L'obiettivo della regressione è stimare la funzione f in modo da minimizzare l'errore tra i valori osservati di Y e i valori predetti dal modello.*

Definizione 8.1

Si parla di *rumore* con il valore di ϵ per indicare tutte quelle variabili che influenzano Y ma che non sono state incluse nel modello. Anche perché basti pensare al fatto che f non è un modello deterministico ma probabilistico, quindi non è possibile prevedere con certezza il valore di Y dato X.

8.2 Regressione lineare semplice

Si parla di **regressione lineare semplice** quando si ha una sola variabile indipendente X. In questo caso, il modello di regressione lineare può essere espresso come:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Esempio. Ipotizziamo di voler prendere in considerazione un dataset che contiene informazioni sul numero di ore di studio e i voti ottenuti dagli studenti in un esame. Vogliamo capire se esiste una relazione tra il numero di ore di studio (variabile indipendente X) e il voto ottenuto (variabile dipendente Y), con un modello circa così:

$$\text{voto} = \beta_0 + \beta_1 \cdot \text{ore\_studio} + \epsilon$$

8.2.1 Analogia geometrica con una retta

In realtà la forma del regressore lineare è, in questo caso, quella di una retta:

$$y = mx + q \longrightarrow Y = \beta_1 X + \beta_0$$

Dove m rappresenta la pendenza della retta (coefficiente angolare) e q rappresenta l'intercetta con l'asse delle ordinate (coefficiente lineare).

Si può facilmente notare, ricordando un po' di algebra (figura 8.1), che:

- Al variare del parametro β_1 si forma un fascio di rette proprio[1] passanti per il punto $(0, \beta_0)$, ovvero l'intercetta con l'asse delle ordinate (a sinistra in figura 8.1).
- Al variare del parametro β_0 si forma un fascio di rette improprio[2] con pendenza β_1 (a destra in figura 8.1).

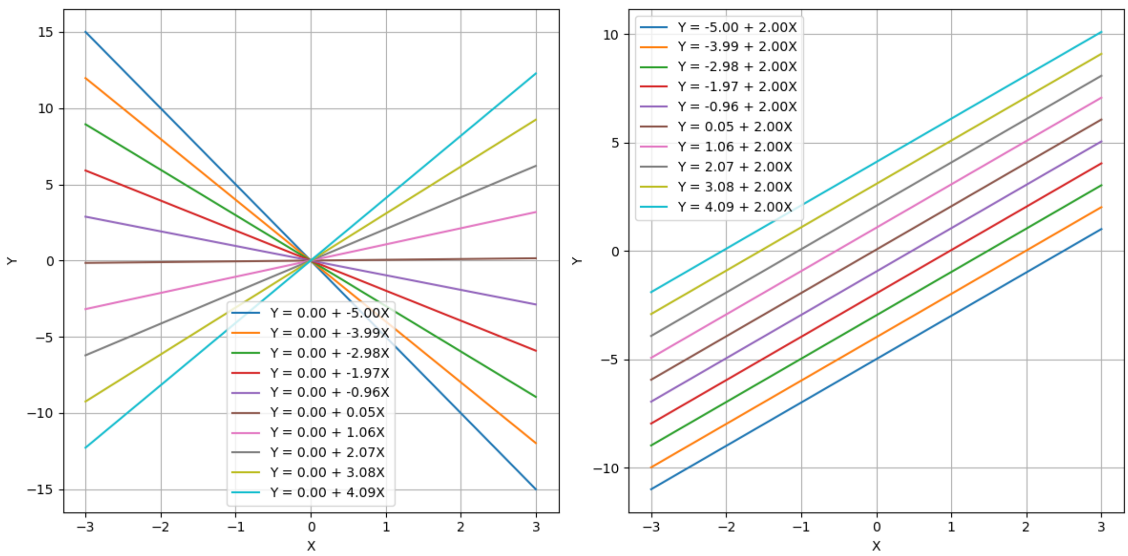

Figura 8.1: Esempio di rette di regressione al variare del coefficiente: a sinistra varia β_1 mantenendo fisso β_0, a destra varia β_0 mantenendo fisso β_1.

[1] Un fascio di rette proprio è un insieme di rette che passano tutte per uno stesso punto.
[2] Un fascio di rette improprio è un insieme di rette parallele tra loro.

8.2.2 OLS: Ordinary Least Squares

Per stimare i coefficienti β_0 e β_1 del modello "ottimale" bisogna definire cos'è un modello *ottimale*. Si può dire ottimale un modello se predice bene la variabile Y dalla variabile X. Per misurare questo, partiamo con il definire il nostro insieme di osservazioni:

$$\{(x_i, y_i)\}_{i=1}^n$$

Dove n è il numero di osservazioni, x_i è il valore della variabile indipendente per l'osservazione i e y_i è il valore della variabile dipendente per l'osservazione i.
Sia:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

il valore predetto dal modello per l'osservazione i, dove $\hat{\beta}_0$ e $\hat{\beta}_1$ sono le stime dei coefficienti. Per ogni punto dati (x_i, y_i), definiamo la deviazione della predizione dal valore corretto y_i come:

$$e_i = y_i - \hat{y}_i$$

Questa quantità è chiamata **residuo** o **errore di predizione**. Ovviamente questo valore può essere positivo o negativo, a seconda che la predizione sia inferiore o superiore al valore reale. Da questo valore definiamo la **RSS: Residual Sum of Squares** (Somma dei quadrati dei residui) come:

$$RSS = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Con questo valore possiamo avere una misura di un modello ottimale: un modello è tanto più ottimale quanto più piccolo è il valore di RSS. Quindi l'obiettivo diventa minimizzare il valore di RSS trovando i valori ottimali di $\hat{\beta}_0$ e $\hat{\beta}_1$:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg\min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i))^2$$

che viene chiamata **loss (o cost) function**, ovvero la funzione che misura l'errore del modello ed è quella da minimizzare.

Per minimizzare questa funzione si può usare il calcolo differenziale ricordando che per trovare i punti di minimo di una funzione basta trovare i punti in cui la derivata prima è nulla[3]. Calcoliamo le derivate parziali[4] della funzione di loss rispetto a β_0 e β_1 e le poniamo uguali a zero:

$$\frac{\partial RSS}{\partial \beta_0} = 0, \qquad \frac{\partial RSS}{\partial \beta_1} = 0$$

Risolvendo il sistema di equazioni si ottengono le seguenti formule per i coefficienti

[3]Quando la derivata prima è nulla, la crescenza della curva cambia e quindi si ha un punto di massimo o minimo locale.

[4]I parametri sono due, non si può derivare rispetto a entrambi contemporaneamente, quindi si calcolano le derivate parziali.

stimati:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 \ = \bar{y} - \hat{\beta}_1 \, \bar{x}$$

Dove $\bar{x}$ e $\bar{y}$ sono le medie dei valori di X e Y rispettivamente:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Varianza dalla popolazione ideale. Considerando una popolazione ideale dove:

$$Y = 2x + 1$$

ci si aspetta che i campioni estratti da questa popolazione abbiano coefficienti $\hat{\beta}_0 \approx 1$ e $\hat{\beta}_1 \approx 2$, ma a causa del rumore e della variabilità dei dati, i valori stimati possono differire leggermente da questi valori ideali.

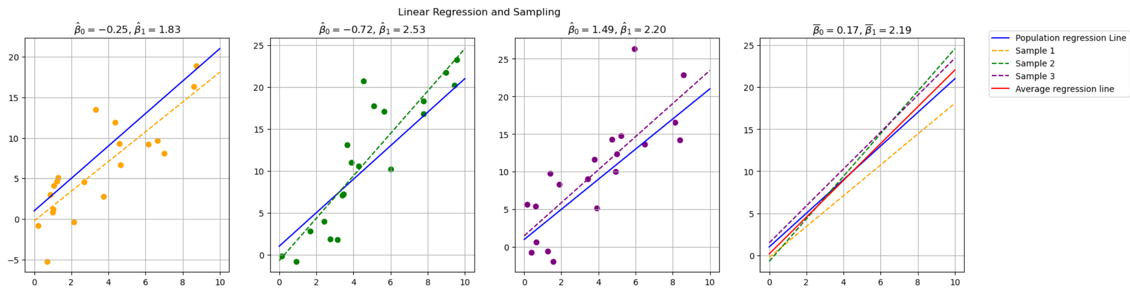

Figura 8.2: Esempio di regressione lineare e variabilità campionaria. Ogni pannello mostra un campione diverso estratto dalla stessa popolazione con la rispettiva retta di regressione stimata $(\hat{\beta}_0, \hat{\beta}_1)$. La linea blu rappresenta la retta di regressione della popolazione, mentre l'ultimo grafico illustra la media delle rette di regressione campionarie (in rosso), evidenziando come, in media, essa coincida con la vera retta di regressione della popolazione.

Dalla figura 8.2 si può notare come, nonostante la variabilità dei campioni, la media delle rette di regressione campionarie (in rosso nell'ultimo grafico) tende a coincidere con la vera retta di regressione della popolazione (in blu), indicando che il metodo OLS potrebbe avere una varianza bassa, ma rimane comunque uno stimatore[5] *unbiased*.

8.2.3 Intervalli di confidenza per i coefficienti

Poiché i coefficienti stimati $\hat{\beta}_0$ e $\hat{\beta}_1$ sono calcolati su un campione di dati, possonoe ssere visti come *stimatori* dei veri coefficienti della popolazione. Da qui possiamo calcolare gli **intervalli di confidenza** e gli **standard errors** (SE) per questi coefficienti, che ci danno un'idea della precisione delle nostre stime.

[5] Ricordiamo che uno stimatore è qualcosa che ci fornisce una stima dei parametri della popolazione basata sui dati campionari.

Esempio. Prendiamo per esempio un dataset così formato:

	displacement	cylinders	horsepower	weight	acceleration	model_year	origin	mpg
0	307.0	8	130.0	3504	12.0	70	1	18.0
1	350.0	8	165.0	3693	11.5	70	1	15.0
2	318.0	8	150.0	3436	11.0	70	1	18.0
3	304.0	8	150.0	3433	12.0	70	1	16.0
4	302.0	8	140.0	3449	10.5	70	1	17.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
393	140.0	4	86.0	2790	15.6	82	1	27.0
394	97.0	4	52.0	2130	24.6	82	2	44.0
395	135.0	4	84.0	2295	11.6	82	1	32.0
396	120.0	4	79.0	2625	18.6	82	1	28.0
397	119.0	4	82.0	2720	19.4	82	1	31.0
298 rows × 8 columns								

Tabella 8.1: Esempio di dataset con caratteristiche delle automobili

Da qui creiamo il modello:

$$\text{mpg} \approx= \beta_0 + \text{horsepower} \cdot \beta_1 + \epsilon$$

E otteniamo i seguenti risultati:

COEFFICIENT	**SE**	**INTERVALLI DI CONFIDENZA**
$\hat{\beta}_0 = 39.94$	0.717	[38.53, 41.35]
$\hat{\beta}_1 = -0.1578$	0.006	[-0.17, -0.15]

Tabella 8.2: Stima dello standard error e degli intervalli di confidenza per i coefficienti della regressione

Da gli intervalli di confidenza nella tabella 8.2 si può notare come il coefficiente $\hat{\beta}_1$ abbia un intervallo di confidenza che non include lo zero, suggerendo che esiste una relazione significativa tra la variabile indipendente (horsepower) e la variabile dipendente (mpg). Inoltre, lo standard error relativamente basso indica che la stima del coefficiente è precisa.

Un tipo di plot a cui si può fare riferimento a uno scatter plot con la retta di regressione e gli intervalli di confidenza intorno alla retta stessa, come mostrato in figura 8.3 (più approfondito nella sotto-sezione **??**).

8.2.4 Test statistici per la significatività dei coefficienti

Nell'esempio di prima abbiamo considerato i coefficienti come stimatori della popolazione. Quindi possiamo eseguire dei test statistici per verificare se questi coefficienti sono significativamente diversi da zero. Questo perché per $\beta_1 = 0$ non esiste correlazione tra le variabili X e Y e il regressore diventa:

$$Y = \beta_0 + \epsilon$$

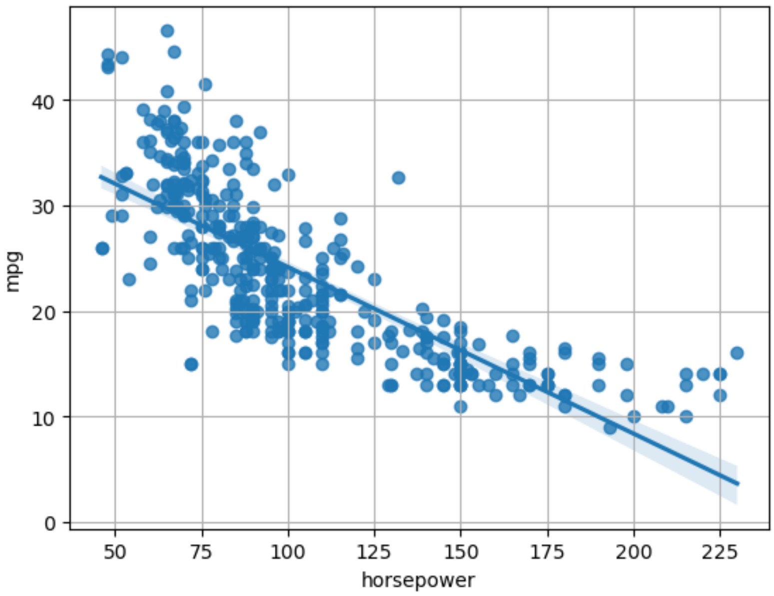

Figura 8.3: Scatter plot della relazione tra *horsepower* e *mpg* con retta di regressione stimata e relativo intervallo di confidenza.

che non dipende da X. Eseguiamo dunque un test di ipotesi, definendo l'*ipotesi nulla*:

$$H_0 : \text{Non c'è associazione tra } X \text{ e } Y \Leftrightarrow \beta_1 = 0$$

e l'*ipotesi alternativa*:

$$H_a : \text{C'è associazione tra } X \text{ e } Y \Leftrightarrow \beta_1 \neq 0$$

Per testare queste ipotesi, si può utilizzare il test t-Student, calcolando il valore t come:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

Dove $SE(\hat{\beta}_1)$ è lo standard error del coefficiente stimato. Confrontando il valore t calcolato con la distribuzione t-Student con $n-2$ gradi di libertà (dove n è il numero di osservazioni), si può determinare il valore p associato. Se il p-value è inferiore a una soglia di significatività predefinita (ad esempio, 0.05), si rifiuta l'ipotesi nulla, suggerendo che esiste una relazione significativa tra X e Y.

Si può eseguire lo stesso test anche per l'intercetta β_0, anche se il modello non cambierà molto perché non è un coefficiente di una variabile indipendente. Si possono comunque aggiornare i risultati:

8.3 Valutazione del modello di regressione

Dopo aver stimato i coefficienti del modello di regressione, è importante valutare la bontà del modello. I test statistici ci dicono se una relazione statistica significativa esiste, ma non

COEFFICIENT	SE	t	P > |t|	INTERVALLI DI CONFIDENZA
$\hat{\beta}_0 = 39.94$	0.717	55.66	0.000	[38.53, 41.35]
$\hat{\beta}_1 = -0.1578$	0.006	-26.30	0.000	[-0.17, -0.15]

Tabella 8.3: Stima dello standard error e degli intervalli di confidenza per i coefficienti della regressione

ci dicono quanto bene il modello si **adatta** ai dati o quanto sono **accurate** le predizioni del modello, per le quali si usano metriche diverse:

Metriche per la bontà del modello. Queste metriche misurano quanto bene il modello si adatta ai dati osservati. Un esempio comune è il coefficiente di determinazione R^2, che indica la proporzione della varianza nella variabile dipendente che è spiegata dalle variabili indipendenti nel modello. Un valore di R^2 vicino a 1 indica un buon adattamento del modello ai dati.

Metriche per l'accuratezza delle predizioni. Queste metriche valutano quanto accurate sono le predizioni del modello sui dati nuovi o non visti. Esempi comuni includono l'Errore Quadratico Medio (MSE) e l'Errore Assoluto Medio (MAE). Queste metriche misurano la differenza tra i valori predetti dal modello e i valori effettivi osservati. Un valore più basso di MSE o MAE indica una maggiore accuratezza delle predizioni del modello.

Residui e RSS. Tutte le metriche di regressione sono costruite sui **residui** (sotto-sezione 8.2.2), che rappresentano la differenza tra i valori osservati e i valori predetti dal modello. La somma dei quadrati dei residui (RSS) è una misura comune della bontà del modello, che quantifica la quantità totale di errore nelle predizioni del modello. Un valore più basso di RSS indica un miglior adattamento del modello ai dati.

8.3.1 Metriche per la bontà del modello

Generalmente il nostro obiettivo in queste metriche è capire[6] quanto il modello spiega i dati osservati, ovvero quanto della variabilità totale della variabile dipendente Y è spiegata dalle variabili indipendenti X.

RSE: Residual Standard Error. Il Residual Standard Error (RSE) è una misura della dispersione dei residui intorno alla retta di regressione stimata. Si calcola come la radice quadrata della somma dei quadrati dei residui divisa per i gradi di libertà:

$$RSE = \sqrt{\frac{RSS}{n-p-1}} = \sqrt{\frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

[6]Nel caso della statistica, capire spesso significa *inferire*: ovvero stimare parametri e testare ipotesi

dove n è il numero di osservazioni e p è il numero di variabili indipendenti nel modello, ovvero i "gradi di libertà" del modello, nel costro caso di regressione semplice $p = 1 \Rightarrow$ il denominatore diventa $n - 2$.

Poiché l'RSE è una misura della dispersione dei residui, un valore più basso di RSE indica che i residui sono più vicini alla retta di regressione stimata, suggerendo un miglior adattamento del modello ai dati. Inoltre, l'RSE è espresso nelle stesse unità della variabile dipendente Y, rendendolo interpretabile nel contesto del problema di regressione. Ha tuttavia un difetto: non è normalizzato, quindi non permette di confrontare modelli con variabili dipendenti diverse e per capire se è buono bisogna confrontarlo con la scala dei valori di Y.

Statistica di $\mathbf{R^2}$. La statistica di R^2 (coefficiente di determinazione) è una misura assoluta della bontà del modello di regressione. Essa quantifica la proporzione della varianza nella variabile dipendente Y che è spiegata dalle variabili indipendenti X nel modello. Per calcolare, dobbiamo prima calcolare la somma totale dei quadrati (TSS), che rappresenta la variabilità totale nella variabile dipendente:

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Dove $\bar{y}$ è la media dei valori osservati di Y. A questo punto, possiamo calcolare R^2 come:

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{TSS - RSS}{TSS}$$

Il valore di R^2 varia tra 0 e 1, dove un valore di 0 indica che il modello non spiega alcuna variabilità nella variabile dipendente, mentre un valore di 1 indica che il modello spiega tutta la variabilità. In generale, un valore più alto di R^2 indica un miglior adattamento del modello ai dati. Tuttavia, è importante notare che un alto valore di R^2 non implica necessariamente che il modello sia appropriato o che le variabili indipendenti siano causalmente correlate alla variabile dipendente, perché questa misura spiega la correlazione ma non la causalità.

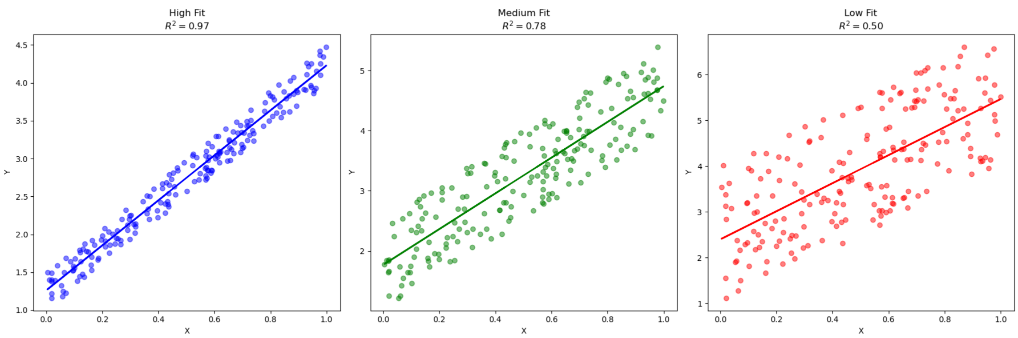

Figura 8.4: Confronto tra tre modelli di regressione lineare con livelli diversi di varianza del rumore: *High Fit* con $R^2 = 0.97$, *Medium Fit* con $R^2 = 0.78$ e *Low Fit* con $R^2 = 0.50$. All'aumentare della variabilità dei dati attorno alla retta di regressione, la bontà dell'adattamento diminuisce.

8.3.2 Grafici di diagnostica

Le metriche sono unicamente numeriche e quantificano l'errore, senza però spiegare *perché* l'errore esiste. Anche qui, si utilizzano i residui per fare dei grafici di diagnostica che aiutano a capire se il modello è adatto o meno ai dati.

Grafico dei residui vs valori predetti. Un grafico comune è il grafico dei residui contro i valori predetti. In questo grafico, i residui e_i sono tracciati sull'asse delle ordinate contro i valori predetti $\hat{y}_i$ sull'asse delle ascisse. Questo grafico aiuta a identificare eventuali pattern nei residui che potrebbero indicare problemi con il modello di regressione. Esistono tre casi possibili:

- I residui sono distribuiti casualmente intorno allo zero, senza pattern evidente. Questo indica che il modello di regressione è appropriato per i dati.
- Si forma una "U" o una ⋒ nei residui, suggerendo che il modello di regressione lineare non cattura una relazione non lineare tra X e Y. In questo caso, potrebbe essere necessario considerare un modello di regressione non lineare.
- La dispersione dei residui aumenta o diminuisce con i valori predetti, indicando eteroschedasticità[7]. Questo suggerisce che la variabilità dei residui non è costante e potrebbe richiedere una trasformazione delle variabili o l'uso di un modello di regressione più robusto.

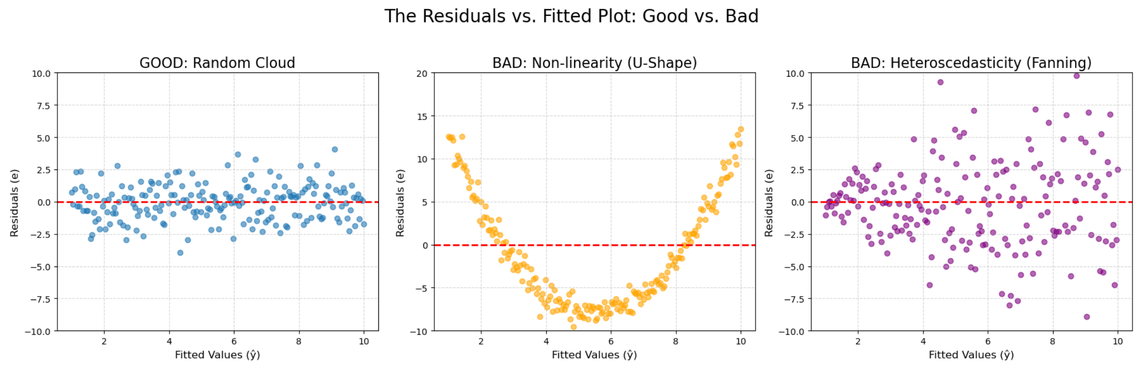

Figura 8.5: Confronto tra diversi pattern nei grafici Residuals vs. Fitted. A sinistra un caso "buono", in cui i residui formano una nube casuale attorno allo zero, indicando che il modello lineare è appropriato. Al centro un caso "cattivo", dove i residui mostrano una chiara struttura a U, segnalando non-linearità. A destra un altro caso "cattivo", in cui la dispersione dei residui aumenta con i valori stimati, evidenziando eteroschedasticità (effetto "fanning").

Grafici Q-Q (Quantile-Quantile). Un altro grafico utile è il grafico Q-Q (Quantile-Quantile), che confronta la distribuzione dei residui con una distribuzione normale teorica. In questo grafico, i quantili dei residui sono tracciati contro i quantili di una distribuzione normale. Se i residui seguono una distribuzione normale, i punti nel grafico Q-Q dovrebbero

[7] L'eteroschedasticità è una forma di non costanza della varianza degli errori.

allinearsi lungo una linea retta diagonale. Eventuali deviazioni significative da questa linea indicano che i residui non seguono una distribuzione normale, suggerendo che il modello di regressione potrebbe non essere appropriato per i dati.

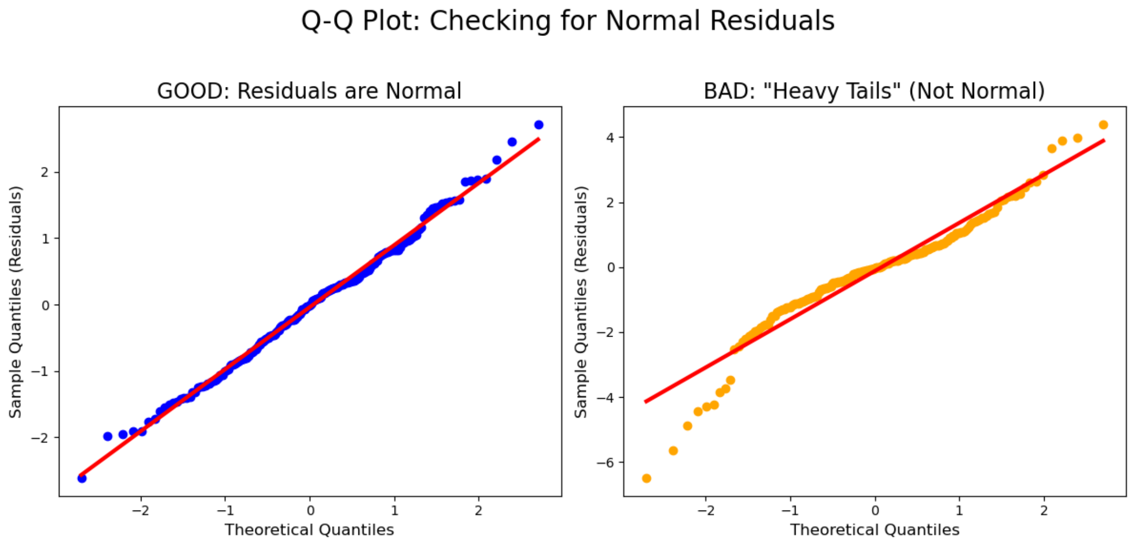

Figura 8.6: Confronto tra due Q–Q plot dei residui. A sinistra un caso "buono", in cui i punti seguono approssimativamente la linea rossa, indicando che i residui possono essere considerati normalmente distribuiti. A destra un caso "cattivo", dove le code pesanti (heavy tails) producono deviazioni marcate dalla linea teorica, mostrando che i residui non seguono una distribuzione normale.

8.4 Regressione lineare multivariata

La regressione lineare multivariata estende il concetto di regressione lineare semplice includendo più variabili indipendenti. In questo caso, il modello di regressione può essere espresso come:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Dove $X_1, X_2, \ldots, X_p$ sono le variabili indipendenti e p è il numero di variabili indipendenti nel modello. I coefficienti $\beta_1, \beta_2, \ldots, \beta_p$ rappresentano l'effetto di ciascuna variabile indipendente sulla variabile dipendente Y, tenendo conto dell'effetto delle altre variabili indipendenti nel modello.

Questa generalizzazione viene utilizzata quando si vuole studiare l'effetto di più variabili esplicative su una variabile di interesse, permettendo di controllare per l'influenza di altre variabili e di ottenere stime più accurate degli effetti delle singole variabili indipendenti.

Nell'esempio della tabella 8.1, si potrebbe costruire un modello:

$$mpg \approx= \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{weight} + \beta_3 \cdot \text{model\_year} + \epsilon$$

e ottenere un valore di $R^2 = 0.808$ rispetto a $R^2 = 0.682$ del modello semplice con solo *horsepower* come variabile indipendente, suggerendo che l'inclusione di più variabili indipendenti migliora la capacità del modello di spiegare la variabilità nella variabile dipendente *mpg*.

8.4.1 Interpretazione geometrica

In un contesto multivariato, la regressione lineare può essere interpretata geometricamente come la ricerca di un iperpiano che meglio si adatta ai dati in uno spazio a più dimensioni. Ogni variabile indipendente X_j rappresenta una dimensione nello spazio, e la variabile dipendente Y rappresenta l'altezza dell'iperpiano in quella posizione.

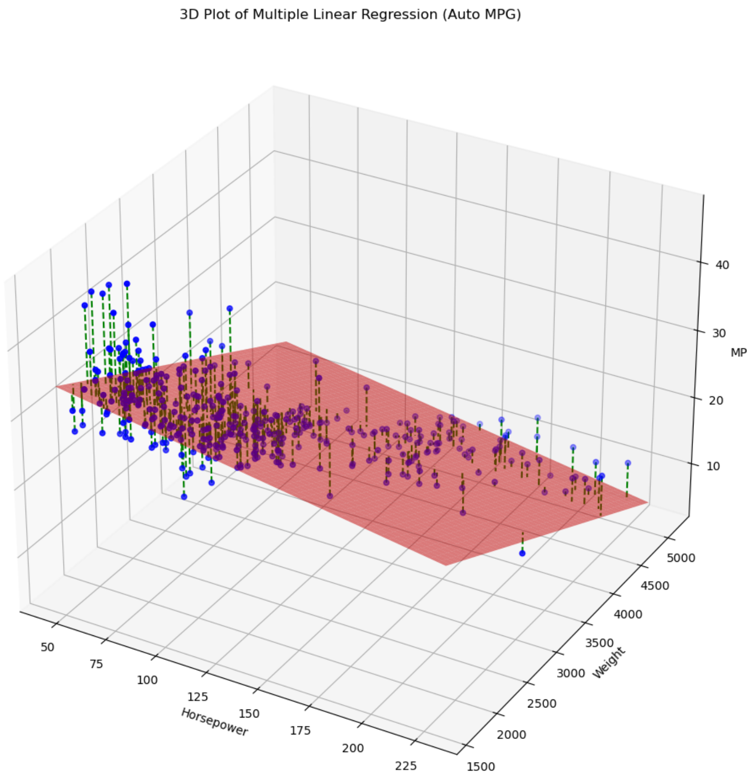

Figura 8.7: Rappresentazione tridimensionale della regressione lineare multipla sul dataset Auto MPG. I punti indicano i valori osservati di *mpg* in funzione di *horsepower* e *weight*, mentre il piano rosso rappresenta il piano di regressione stimato. Le linee verticali mostrano la distanza tra i valori osservati e quelli predetti dal modello, evidenziando l'errore di regressione.

8.4.2 Interpretazione statistica

Generalmente interpretare statisticamente un modello di regressione lineare multivariata è simile all'interpretazione della regressione lineare semplice. Dato un modello:

$$Y = \beta_0 + \beta \cdot x$$

dove x è un vettore di variabili indipendenti e β è un vettore di coefficienti associati a ciascuna variabile indipendente. Si può interpretare:

- Il valore di β_0 come l'intercetta del modello, ovvero il valore atteso di Y quando tutte le variabili indipendenti sono uguali a zero.
- Il valore di un certo β_i, con $i \in [1, p]$, come il cambiamento atteso in Y associato a un aumento unitario della variabile indipendente X_i, mantenendo costanti tutte le altre variabili indipendenti nel modello. Questo permette di isolare l'effetto di ciascuna variabile indipendente sulla variabile dipendente, tenendo conto dell'influenza delle altre variabili nel modello.

Come variano i coefficienti al variare delle variabili indipendenti. Da questa interpretazione statistica, possiamo dire che: aggiungere nuove variabili al regressore cambia il modo in cui la varianza di Y viene "spiegata" dai coefficienti. Nel modello semplice

$$mpg = \beta_0 + \beta_1 \cdot \text{horsepower}$$

otteniamo ad esempio $\hat{\beta}_0 \approx 39.94$ e $\hat{\beta}_1 \approx -0.16$. Quando introduciamo anche il *weight*,

$$mpg = \beta_0 + \beta_1 \cdot \text{horsepower} + \beta_2 \cdot \text{weight},$$

le stime diventano circa $\hat{\beta}_0 \approx 45.64$ e $\hat{\beta}_1 \approx -0.05$. Questo non significa che uno dei due modelli sia "sbagliato", ma che nel primo caso l'effetto di *horsepower* ingloba anche parte della variabilità dovuta al *weight*[8] (e ad altre variabili non osservate), mentre nel secondo caso ogni coefficiente descrive l'effetto della propria variabile a parità delle altre.

8.4.3 Stima dei coeffficienti di regressione

Dato il modello generale di regressione lineare multivariata:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

Possiamo definire la **funzione di costo** come la somma dei quadrati dei residui (RSS):

$$RSS(\beta_0, \beta_1, \ldots, \beta_p) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}))^2$$

dove n è il numero di osservazioni, y_i è il valore osservato della variabile dipendente per l'osservazione i, e x_{ij} è il valore della variabile indipendente X_j per l'osservazione i.

I valori $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ che minimizzano la funzione di costo RSS possono essere trovati

[8] Ecco perché il test R^2 non spiega completamente la bontà del modello.

risolvendo il seguente sistema di equazioni:

$$\begin{cases} y^{(1)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(1)} + \hat{\beta}_2 x_2^{(1)} + \ldots + \hat{\beta}_p x_p^{(1)} \\ y^{(2)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(2)} + \hat{\beta}_2 x_2^{(2)} + \ldots + \hat{\beta}_p x_p^{(2)} \\ \vdots \\ y^{(n)} = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(n)} + \hat{\beta}_2 x_2^{(n)} + \ldots + \hat{\beta}_p x_p^{(n)} \end{cases}$$

trasformato in forma matriciale come:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Dove:

- $\mathbf{Y}$ è il vettore colonna dei valori osservati della variabile dipendente.
- $\mathbf{X}$ è la matrice delle variabili indipendenti, con una colonna di 1s per l'intercetta, chiamata anche **design matrix**.
- $\boldsymbol{\beta}$ è il vettore colonna dei coefficienti di regressione da stimare.
- $\boldsymbol{\epsilon}$ è il vettore colonna degli errori.

Dalla notazione sopra, possiamo derivare la soluzione per i coefficienti stimati ricordando che la somma dei quadrati dei residui può essere espressa come:

$$RSS(\beta) = \sum_{i=1}^{n} (e^{(i)})^2 = e^T e$$

Minimizzando la funzione di costo, utilizzando sempre il metodo dei *minimi quadrati ordinari* (OLS), otteniamo la seguente formula per i coefficienti stimati:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

8.4.4 F-Test

Per valutare la significatività complessiva di un modello di regressione lineare multivariata, si può utilizzare il test F. Questo test confronta un modello completo (con tutte le variabili indipendenti) con un modello ridotto (senza alcune o tutte le variabili indipendenti) per determinare se l'inclusione delle variabili indipendenti migliora significativamente l'adattamento del modello ai dati. Nella realtà, non è altro che definire un test d'ipotesi:

$$H_0 : \text{Tutte le variabili indipendenti non hanno effetto su } Y \Leftrightarrow \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

$$H_a : \text{Almeno una variabile indipendente ha un effetto su } Y \Leftrightarrow \exists i : \beta_i \neq 0$$