Statistica e analisi dei dati

Kevin Muka

a.a 2024-2025

Indice

Lezioni

L01 - 25/02/2025

Dispense L01-Introduzione_a_python

Introduzione a Python: panoramica sul linguaggio Python e le sue applicazioni.

Tipi di dati e operatori:

- dati semplici e operatori: numeri, stringhe e booleani; operazioni aritmetiche e logiche.
- dati strutturati: liste, tuple, dizionari e insiemi; metodi e funzioni utili per manipolare le liste (e altri tipi strutturati).

Operazioni sulle liste: creazione e manipolazione delle liste; uso di metodi e operatori specifici per le liste (*list comprehension*); esempi pratici e funzioni integrate.
Import e utilizzo di librerie per specifiche funzionalità: `numpy` per operazioni numeriche; `pandas` per la gestione di dati strutturati; `matplotlib.pyplot` per la visualizzazione grafica; `csv` e `collections` per la manipolazione dei dati.

L02 - 27/02/2025

Dispense L02-Pandas

Introduzione alla libreria `Pandas` e ai dataset: caricamento del file `heroes.csv`.

- Serie: creazione e manipolazione di serie; gestione degli indici e accesso tramite slicing, `loc` e `iloc`; calcolo delle frequenze con `value_counts`; operazioni matematiche e uso del metodo `apply`.
- Visualizzazione grafica: creazione di grafici a barre per rappresentare le frequenze; personalizzazione dell'asse delle ascisse (utilizzo di `numpy.arange` e `xticks`).
- DataFrame: creazione di DataFrame da file CSV tramite `read_csv`; accesso a righe e colonne (utilizzo di `loc`, `iloc`, `at` e `iat`); ordinamento dei dati con `sort_values` e `sort_index`; filtraggio e selezione di sottoinsiemi di dati

Librerie utilizzate: `pandas`, `matplotlib.pyplot`, `numpy`, `csv`.

L03 - 04/03/2025

Dispense L03-Dati_e_frequenze - Capitolo 2 RS

Dati e Frequenze: concetti di base; introduzione alla distinzione tra dati quantitativi e qualitativi.
Classificazione dei dati: dati qualitativi (categorie e modalità); dati quantitativi (tipologie e suddivisione).
Frequenze: calcolo e visualizzazione delle frequenze assolute e relative. Frequenze cumulate: definizione e applicazioni. Frequenze congiunte e marginali. Diagrammi:

- grafici a barre, istogrammi e diagrammi a torta.
- diagrammi di Pareto: ordinamento delle frequenze in ordine decrescente; identificazione della regola dell'80/20 per evidenziare le componenti più rilevanti.
- diagrammi stelo-foglia: rappresentazione della distribuzione dei dati quantitativi; visualizzazione della forma e concentrazione dei valori.

Suggerimenti e tecniche per la generazione dei grafici con l'uso di librerie come `matplotlib` e `pandas`.

L04 - 06/03/2025

Dispense L03-Dati_e_frequenze - Capitolo 3 RS

L05 - 11/03/2025

Dispense L03-Dati_e_frequenze: diagrammi a stelo
Dispense L04-Indici_di_dispersione - Capitolo 3 RS

L06 - 13/03/2025

Dispense L05-Indici_di_eterogeneita

Indici di dipendenza. Indici di eterogeneità: indice di Gini, indice di entropia. Alberi di decisione e "machine learning" tramite questi indici.

L07 - 18/03/2025

Dispense L05-Indici_di_eterogeneita: indici di concentrazione
Dispense L06-Trasformazione_dei_dati

L08 - 25/04/2025

Dispense L07-Analisi_della_varianza: anova
Dispense L08-Analisi_di_classificatori: classificatori

L09 - 27/04/2025

Dispense L08-Calcolo_combinatorio: calcolo combinatorio
Introduzione ai concetti basilari della probabilità.

L10 - 01/04/2025

Calcolo delle probabilità: assiomi di Kolmogorov e teoremi elementari. Spazi equiprobabili.

L11 - 03/04/2025

Probabilità condizionata, regola di fattorizzazione, teorema delle probabilità totali, teorema di Bayes.

1 Introduzione alla statistica

1.1 Definizione

Statistica La statistica è l'arte di apprendere dai dati. Si occupa della raccolta, della descrizione e dell'analisi dei dati, possibilmente permettendo di trarne delle conclusioni.

A volte un'analisi statistica comincia con un insieme di dati prestabilito, in questo caso la statistica si usa per descrivere, riassumere e analizzare i dati. In altre situazioni i dati non sono ancora disponibili e si può usare la statistica per progettare un esperimento che li generi. Se ne occupa la statistica descrittiva.

Statistica descrittiva La statistica descrittiva è la parte della statistica che descrive e riassume i dati.

Una volta che i dati sono stati raccolti e descritti, si vogliono trarre delle conclusioni. Se ne occupa la statistica inferenziale.

Statistica inferenziale La statistica inferenziale è la parte della statistica che trae conclusioni sui dati.

La statistica inferenziale si basa sul modello probabilistico che consiste nel fare un insieme di assunzioni sulle probabilità di ottenere un certo valore. La statistica inferenziale, quindi, richiede la conoscenza della teoria della probabilità. L'inferenza statistica si basa sull'assunzione che importanti aspetti del fenomeno in analisi si possano rappresentare in termini di probabilità e giunge a conclusioni usando i dati per fare inferenza su queste probabilità.

1.2 Popolazioni e campioni

Nella statistica è cruciale ottenere delle informazioni su tutto un insieme di elementi, che viene definito popolazione. Spesso la popolazione però è troppo numerosa per poter analizzare ciascuno dei suoi membri: in questo caso si sceglie e si esamina un suo sottoinsieme, che viene definito campione.

Popolazione Si definisce popolazione l'insieme di tutti gli elementi di interesse.
Campione Si definisce campione un sottoinsieme della popolazione, utile quando questa è troppo numerosa.

Affinché il campione ci dia informazioni su tutta la popolazione, esso deve essere rappresentativo di tutta la popolazione. Con rappresentativo si intende che il campione deve essere scelto in modo che tutte le parti della popolazione abbiano uguale probabilità di fare parte del campione. *Il campione deve quindi riflettere la variabilità reale della popolazione.*

Campione casuale Un campione di k membri di una popolazione si dice **campione casuale**, o talvolta campione casuale semplice, se i membri sono scelti in modo che tutte le possibili scelte dei k membri siano ugualmente probabili.

Una volta che si sceglie un campione casuale, è possibile usare l'inferenza statistica per giungere a conclusioni sull'intera popolazione studiando gli elementi del campione.

1.2.1 Campionamento casuale stratificato

Un metodo più sofisticato del campionamento casuale semplice è il campionamento casuale stratificato. Inizialmente si stratifica la popolazione in sottopopolazioni, ognuna delle quali contiene unità simili secondo determinati criteri. In seguito, da ogni strato si estrae casualmente un numero di unità proporzionale alla sua consistenza nella popolazione totale. In questo modo, le proporzioni di ciascuno strato presenti nel campione rispecchiano esattamente quelle dell'intera popolazione.

La stratificazione è particolarmente efficace per conoscere il membro *medio* della popolazione totale quando ci sono differenze tra le sottopopolazioni rispetto alla questione studiata.

Parte I

Statistica descrittiva

2 Descrivere insiemi di dati

2.1 Dati quantitativi e qualitativi

Una distinzione che si può fare sui dati osservabili riguarda il modo in cui questi sono misurati:

- dati quantitativi: l'esito della misurazione è una quantità numerica.
- dati qualitativi: l'esito della misurazione è un'etichetta appartenente a un insieme fissato di etichette. Vengono anche detti categorici o nominali.

Classificazione dati qualitativi I dati qualitativi si distinguono in binari, nominali e ordinali:

- Dati binari o booleani: l'osservazione può assumere soltanto due valori tra loro non confrontabili. Si utilizza "booleani" per evidenziare la presenza o assenza di una proprietà, mentre "binari" per indicare due etichette possibili.
- Dati nominali: non ammettono un confronto d'ordine tra i valori, ma è possibile solo stabilire una relazione di equivalenza.
- Dati ordinali: esiste una relazione d'ordine tra i valori osservabili, e quindi se due valori sono diversi è possibile stabilire quale sia il più piccolo e quale il più grande.

Classificazione dei dati quantitativi I dati quantitativi si distinguono in discreti e continui a seconda dell'insieme di valori che possono assumere:

- Dati discreti: rappresentano variabili che possono assumere un insieme numerabile di valori distinti e separati. Ad ogni valore corrisponde un significato specifico.
- Dati continui: possono teoricamente assumere qualsiasi valore all'interno di un intervallo, anche se nella pratica, per via della memorizzazione digitale, vengono approssimati a una precisione finita.

2.2 Frequenze

Frequenza assoluta La frequenza assoluta di un'osservazione x in un insieme di dati $A = \{x_1, \cdots, x_n\}$ è definita come il numero di volte in cui x compare in A.
Formalmente, se si indica con f_x la frequenza assoluta di x, si ha che $f_x = \#\{j \in \{1, \cdots, n\} \,|\, x_j = x\}$

Frequenza relativa La frequenza relativa consente di esprimere la presenza di ogni valore in termini di proporzione rispetto all'intero campione. Sia $A = \{x_1, \cdots, x_n\}$ un insieme di n dati e sia f_i la frequenza assoluta di un'osservazione x_i in A, si definisce la frequenza relativa di x_i il valore f_i/n. Si osserva che la somma di tutte le frequenze relative in un campione è sempre pari ad 1.

Frequenze cumulate

Le frequenze cumulate si ottengono quando i valori di una variabile possono essere ordinati. Il procedimento consiste nel disporre i valori in ordine crescente, calcolare le loro frequenze individuali e poi sommarle progressivamente: al primo valore si associa la sua frequenza, al secondo la somma della frequenza del primo e del secondo, al terzo la somma delle frequenze dei primi tre, e così via.

È importante notare che l'ultima frequenza cumulata rappresenta il totale dei casi osservati. Inoltre, il concetto di frequenza cumulata si applica sia alle frequenze assolute che a quelle relative: nel caso delle frequenze relative, i valori cumulati variano da 0 a 1.

Quando i dati sono numerici o comunque ordinabili, un concetto affine alle frequenze relative cumulate è quello della *funzione cumulativa empirica*, nota anche come funzione di ripartizione empirica.

Data una serie di osservazioni $x_1, \ldots, x_n$, la funzione cumulativa empirica $\hat{F} : \mathbb{R} \to [0,1]$ è definita in modo che per ogni $x \in \mathbb{R}$ essa assuma il valore pari alla frequenza relativa delle osservazioni minori o uguali a x. In altre parole:

$$\hat{F}(x) = \frac{\#\{x_i \leq x\}}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{(-\infty,x]}(x_i)$$

dove $\mathrm{I}_A : \mathbb{R} \to \{0,1\}$ è la funzione indicatrice dell'insieme A, che restituisce 1 se l'argomento appartiene ad A e 0 altrimenti: di conseguenza l'intervallo $(-\infty, x]$ include tutti i valori minori o uguali a x. Pertanto, per ogni x, $\hat{F}(x)$ rappresenta la frequenza relativa cumulata del massimo valore osservato che non supera x, e il grafico di questa funzione sarà a tratti costanti.

$$I_A(x) = \begin{cases} 1 & \text{se } x \in A \\ 0 & \text{se } x \notin A \end{cases} \quad \Rightarrow \quad I_{(-\infty,x]}(x_i) = \begin{cases} 1 & \text{se } x_i \in (-\infty, x] \\ 0 & \text{se } x_i \notin (-\infty, x] \end{cases} = \begin{cases} 1 & \text{se } x_i \leq x \\ 0 & \text{se } x_i > x \end{cases}$$

In pratica rappresenta il numero di osservazioni dei miei campioni che sono minori o uguali di una certa x, diviso per il numero totale di campioni. La divisione per n è fatta per avere dei valori relativi.

Frequenze congiunte e marginali

Quando si analizza un insieme di osservazioni, può essere particolarmente utile considerare due caratteri contemporaneamente, in modo da verificare se esiste una relazione tra i valori dei due attributi. In questo caso, il concetto di frequenza si adatta contando il numero di occorrenze in cui i due caratteri assumono contemporaneamente determinati valori. Questo conteggio porta alla definizione di *frequenza congiunta assoluta*; se invece si considera la frazione delle osservazioni, si parla di *frequenza congiunta relativa*.

Quando il numero dei possibili valori osservabili per i caratteri non è elevato, è possibile rappresentare visivamente queste frequenze tramite una *tabella delle frequenze congiunte* o *tabella di contingenza*. In tale tabella, le righe sono associate ai valori di uno dei caratteri, mentre le colonne rappresentano i valori del secondo carattere. Gli elementi all'interno della tabella indicano le frequenze congiunte (assolute o relative) per le coppie di valori.

Per facilitare ulteriori analisi, si riportano spesso nelle ultime colonne e nelle ultime righe della tabella le *frequenze marginali*, ottenute sommando rispettivamente i valori per ogni riga e per ogni colonna. Se si desiderano valori relativi, questi totali devono essere normalizzati rispetto al numero complessivo delle osservazioni.

2.3 Grafici

2.3.1 Simmetria

Simmetria Un insieme di dati si dice simmetrico attorno a un valore x_0 se, per ogni scostamento c da x_0, la frequenza dei valori $(x_0 - c)$ è uguale a quella dei valori $(x_0 + c)$. In tal caso, il valore x_0 si definisce "centro di simmetria" della distribuzione.

Quasi simmetria Se i dati non sono perfettamente simmetrici, ma la distribuzione è comunque "quasi" speculare rispetto a un punto centrale, si parla di quasi simmetria.

Un modo semplice per rendersi conto se una distribuzione è (quasi) simmetrica è rappresentarla graficamente e osservarne la forma.

2.3.2 Grafici per la frequenza

Se l'insieme di dati contiene un numero ridotto di valori distinti, lo si può rappresentrare con una *tabella delle frequenze*. Questa tabella associa a ciascun valore distinto osservato la sua frequenza assoluta. La somma di tutte le frequenze deve corrispondere al numero totale di osservazioni.
Data una variabile statistica X che può assumere vari valori, si elencano i valori distinti di X in una colonna e, a fianco di ognuno, la relativa frequenza di occorrenza nel campione.

Per costruire una tabella delle frequenze relative da un insieme di dati, bisogna innanzitutto disporre i valori dei dati in ordine crescente. Si determinano i valori distinti e quante volte ciascuno di essi compaia. Si elencano questi valori distinti affiancati dalla loro frequenza f e dalla loro frequenza relativa f/n, dove n è il numero totale di osservazioni nell'insieme di dati.

2.3.3 Grafici a bastoncini, a barre e poligonali

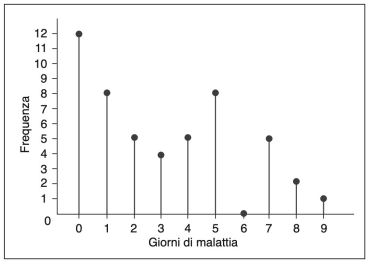


Figura 2.1 Un grafico a bastoncini.

I dati di una tabella di frequenza possono essere rappresentati graficamente in diversi modi. Uno dei più intuitivi è il *grafico a bastoncini*, in cui i valori della variabile statistica sono disposti lungo l'asse orizzontale, mentre le frequenze si riportano sull'asse verticale. Ogni valore viene quindi associato a un semplice segmento che parte dall'asse orizzontale e arriva all'altezza corrispondente alla relativa frequenza.

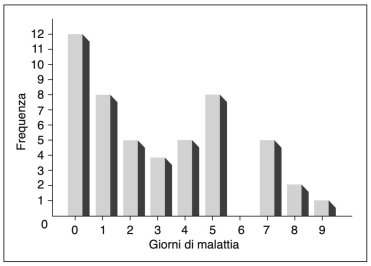


Figura 2.2 Un grafico a barre.

Un secondo tipo di rappresentazione, molto simile concettualmente, è il *grafico a barre*: anche in questo caso i valori si trovano sull'asse orizzontale e le frequenze su quello verticale, ma invece dei singoli segmenti si utilizzano barre di un certo spessore. Ciò permette di mettere in evidenza ciascuna categoria o classe di dati e risulta particolarmente efficace quando si vogliono confrontare categorie di grandezza diversa.

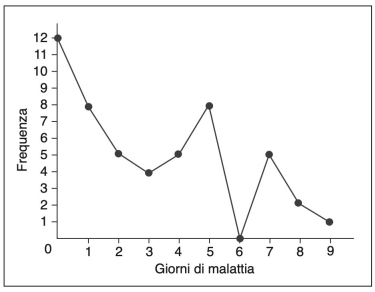


Figura 2.3 Un grafico poligonale.

Infine, esiste il *grafico poligonale*, in cui i valori (sempre disposti sull'asse orizzontale) vengono rappresentati da punti, collocati a un'altezza proporzionale alla loro frequenza, che vengono poi congiunti da segmenti. In questo modo si ottiene una linea spezzata che rende immediata la visualizzazione delle variazioni di frequenza da un valore all'altro, permettendo di apprezzare più facilmente tendenze o andamenti complessivi.

2.3.4 Diagramma ramo-foglia

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 557
26	183, 894, 982
27	671, 711, 744
28	345, 764, 913, 967

Diagramma a stelo

Un modo efficiente di rappresentare un insieme di dati di dimensioni medie consiste nell'utilizzare il *diagramma ramo-foglia* (o a stelo). Tale grafico si ottiene dividendo ciascun valore dei dati in due parti, chiamati appunto rami e foglie.

La scelta dei rami dovrebbe essere fatta in modo che il diagramma ramo-foglia che ne risulta sia informativo sui dati. Questi diagrammi sono particolarmente adatti a descrivere insiemi di dati dimensioni ridotte.

Fisicamente, questo grafico ha l'aspetto di un istogramma ruotato su un lato, con il vantaggio di contenere tutti i valori dei dati originali in ogni classe. Quando il grafico presenta troppe foglie per ogni riga, si può raddoppiare il numero di rami utilizzando due righe per ogni valore del ramo.

2.3.5 Diagramma a torta

Se i dati non sono numerici si utilizza un diagramma a torta. Si costruisce usando un cerchio suddiviso in settori, uno per ogni valore distinto dei dati. Dato un valore con frequenza relativa f/n, allora l'area del settore corrisponde all'area del cerchio moltiplicata per f/n, ovvero un arco con un angolo di $360 \cdot (f/n)$ gradi.

2.3.6 Diagrammi di Pareto

I diagrammi di Pareto sono grafici a barre ordinati in ordine decrescente di frequenza, ai quali è spesso affiancata una linea che rappresenta la frequenza cumulata. In questo modo, oltre a mostrare il numero di casi per ciascuna categoria, permettono di evidenziare quali categorie contribuiscono maggiormente al totale, facilitando l'individuazione delle cause o delle categorie più rilevanti.

2.3.7 Istogrammi

Utilizzare i grafici presentati precedentemente è un metodo efficace per descrivere un insieme di dati. Tuttavia alcuni di questi insiemi hanno troppi valori distinti per poter usare questo metodo.

Raggruppamento dei dati In questi casi si suddividono i valori in gruppi, o classi, e poi si rappresenta con un grafico il numero di valori dei dati che cadono in ciascuna classe. Il numero di classi scelte è un compromesso tra:

1. scegliere poche classi al costo di perdere molte informazioni sui valori effettivi in una classe
2. scegliere troppe classi, ottendendo frequenze troppo basse all'interno di ciascuna di esse

I valori al bordo di una classe si chiamano *estremi* della classe. Si adotta la convenzione di inclusione a sinistra, che richiede che una classe includa il suo estremo sinistro ma non quello destro.

Una volta suddivisi i dati in classi, si costruisce la tabella delle frequenze (e delle frequenze relative), e da questa si ottiene l'istogramma, un grafico a barre adiacenti che mostra la distribuzione dei dati in ciascuna classe. L'istogramma offre una visione immediata di come i valori si distribuiscono: per esempio, se sono concentrati in un certo intervallo, se ci sono "vuoti" senza osservazioni o se alcuni valori si distaccano notevolmente dagli altri. Pur non contenendo tutte le informazioni dell'insieme di dati originale, la tabella delle frequenze di classe e l'istogramma illustrano le caratteristiche fondamentali della distribuzione, come la simmetria, la dispersione e i possibili estremi isolati.

2.3.8 Diagramma di dispersione

Insieme di dati a coppie Un insieme di dati può consistere in coppie di valori che hanno una relazione di qualche tipo tra di loro. Ne viene che ogni elemento dell'insieme di dati sia costituito da un valore x e da uno y. Si indica con $(x_i, y_i),\ i = 1 \cdots n$ la i-esima coppia.

Un metodo per rappresentare un insieme di dati di questo tipo consiste nel considerare ogni elemento della coppia separatamente, producendo istogrammi (o diagrammi ramo-foglia) separati per ciascuno. Così facendo però, nonostante i due grafici ci diano molte informazioni sulle singole variabili (attributi), non si ha nessun tipo di informazione riguardo al rapporto tra queste due variabili.

Per capirne la relazione è necessario considerare i valori accoppiati di ciascun dato simultaneamente. Si possono allora rappresentare questi dati accoppiati in un diagramma rettangolare e bidimensionale, in cui l'asse x rappresenta il valore x dei dati, e l'asse y il valore y. Così facendo si ottiene un *diagramma di dispersione*.

Una delle ragioni per cui *questo* tipo di diagramma è utile consiste nella possibilità di fare previsioni sul valore y di una futura osservazione, noto il valore x. Per stimare il valore y a partire da x si cerca, in modo intuitivo, di tracciare una "retta media" che approssimi l'andamento dei punti sul diagramma, ovvero una retta che passi "il più vicino possibile" a tutti i dati.

Il diagramma di dispersione, oltre a mostrare il comportamento relativo di due variabili e ad aiutarci nelle previsioni, è utile per riconoscere i *valori anomali* (outlier) che sono i punti che non sembrano seguire il comportamento degli altri. Una volta identificati questi valori, si può decidere quali di essi siano appropriati e quali siano invece causati da errori nella raccolta dei dati.

3 Statistiche

Statistica Una statistica è una quantità numerica calcolata a partire da un insieme di dati.

3.1 Centralità

Verranno presentate le statistiche che descrivono la tendenza centrale di un insieme di dati, ossia delle statistiche che descrivono il centro di un insieme di dati. Questa proprietà che si può individuare in un insieme di dati è detta **centralità** o posizione.

Esistono tre indici di posizione: media, mediana e moda. In tutti i tre i casi si parla di campionaria, in quanto sono effettuate su dei campioni.

3.1.1 Media campionaria

Si supponga di avere un campione di n dati i cui valori sono $x_1, x_2, \cdots, x_n$. Una statistica per indicare il centro di questo insieme di dati è la media campionaria, definita come la media aritmetica dei valori dati:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Si osserva che $\bar{x}$ può non corrispondere ad uno dei dati x_i con $1 \leq i \leq n$ presi in considerazione.

Trasformazioni

Traslazione Si consideri ancora lo stesso insieme di dati. Se ciascun valore viene incrementato di una costante b, allora anche la media campionaria viene incrementata di b:

$$y_i = x_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad \bar{y} = \bar{x} + b$$

dove $\bar{y}$ e $\bar{x}$ sono le medie campionarie risettivamente degli y_i e degli x_i.

Dimostrazione: $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(x_i + b) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} x_i}_{\bar{x}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} b}_{\frac{1}{n}\cdot nb} = \bar{x} + b$$

Scalatura Se invece ciascun valore dei dati viene moltiplicato per a, lo è anche la media campionaria:

$$y_i = ax_i \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad \bar{y} = a\bar{x}$$

Dimostrazione: $$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} ax_i = a\frac{1}{n}\sum_{i=1}^{n} x_i = a\bar{x}$$

Combinazione Si faccia ora una combinazione delle due trasformazioni precedentemente illustrate:

$$y_i = ax_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad \bar{y} = a\bar{x} + b$$

Queste tre proprietà derivano dal fatto che tutte queste trasformazioni siano lineari.

Media pesata

Quando i dati sono disposti in una tabella delle frequenze, la media campionaria può essere espressa come la somma del prodotto dei valori distinti per le loro frequenze, divisi per la dimensione dell'insieme dei dati.

Per verificarlo, si supponga di disporre di una tabella delle frequenze che elenca k valori distinti $x_1, x_2, \cdots, x_k$ con le rispettive frequenze $f_1, f_2, \cdots f_k$. Ne segue che questo insieme di dati è costituito da n osservazioni, dove $n = \sum_{i=1}^{k} f_i$ e dove il valore x_i compare f_i volte per $i = 1, 2, \cdots, k$. La media campionaria per questo insieme di dati è:

$$\bar{x} = \frac{x_1 + \cdots + x_1 + x_2 + \cdots + x_2 + \cdots + x_k + \cdots + x_k}{n} = \frac{f_1 x_1 + f_2 x_2 + \cdots + f_k x_k}{n} \tag{3.1}$$

Ora, se $w_1, w_2, \cdots, w_k$ sono numeri non negativi la cui somma è 1, allora

$$w_1 x_1 + w_2 x_2 + \cdots + w_k x_k$$

prende il nome di **media pesata** dei valori $x_1, x_2, \cdots, x_k$ dove w_i è il peso di x_i.

Scrivendo l'equazione (3.1) come

$$\bar{x} = \frac{f_1}{n} x_1 + \frac{f_2}{n} x_2 + \cdots + \frac{f_k}{n} x_k$$

possiamo vedere che la media campionaria $\bar{x}$ è la media pesata dell'insieme dei valori distinti. Il peso assegnato al valore x_i è f_i/n, ossia rappresenta la frazione di volte in cui il valore x_i comare nell'insieme dei dati.

Scarti

Si supponga che l'insieme di dati sia costituito dagli n valori $x_1, \cdots, x_n$ e che $\bar{x} = \sum_{i=1}^{n} x_i/n$ sia la media campionaria. Le differenze tra ciascun valore dei dati e la media campionaria si chiamano **scarti**. Il valore dell'i-esimo scarto è $x_i - \bar{x}$

La somma di tutti gli scarti è sempre 0, ovvero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \bar{x} = n\bar{x} - n\bar{x} = 0$$

Questa uguaglianza afferma che la somma degli scarti positivi della media campionaria controbilancia esattamente la somma degli scarti negativi.

Utilizzando un linguaggio fisico, questo significa che se n pesi dotati della stessa massa vengono posti su un'asta nei punti x_i con $i = 1, \cdots, n$, allora $\bar{x}$ è il punto in cui l'asta può essere messa in equilibrio. Questo punto di equilibrio si chiama centro di gravità.

3.1.2 Mediana campionaria

La media campionaria presenta un forte punto debole come indicatore del centro di un insieme di dati: il suo valore è infatti ampliamente influenzato da eventuali valori estremi (valori fuori scala).

Si dispongano i valori dei dati in ordine crescente. Se il numero di valori è dispari, allora la mediana campionaria è il valore intermedio della lista ordinata; se è pari, allora la mediana campionaria è la media dei due valori intermedi.

Sia $x_{(i)}$ l'i-esimo dato del campione ordinato in maniera crescente, la mediana m è definita come:

$$m = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{per } n \text{ dispari} \\ \left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right)/2 & \text{per } n \text{ pari} \end{cases}$$

La media campionaria e la mediana campionaria sono due statistiche utili per descrivere la tendenza centrale di un insieme di dati. Il loro utilizzo è però molto diverso, in quanto la media campionaria (essendo una media aritmetica) prende in considerazione tutti i valori dell'insieme di dati, mentre invece la mediana campionaria, dato che considera solo uno o due valori centrali, non è influenzata dai valori estremi.

Per gli insiemi di dati che sono approssimativamente simmetrici rispetto ai valori centrali, la media campionaria e la mediana campionaria sono vicine. Entrambe le statistiche sono informative, e il loro utilizzo dipende dal contesto.

3.1.3 Percentili campionari

La mediana campionaria è un caso particolare di una statistica nota come $100p$-esimo percentile campionario, dove p indica un qualunque numero $\mathbb{R}$ nell'intervallo $[0, 1]$.

Per poter calcolare il percentile si deve poter definire un ordinamento sulle osservazioni.

$100p$-esimo percentile campionario È un valore maggiore o uguale di almeno $100p$ percento dei valori dati, e minore o uguale di almeno $100(1 - p)$ percento dei valori dati. Se due valori dei dati soddisfano questa condizione, allora il $100p$-esimo percentile campionario è la media aritmetica di essi.

La mediana campionaria è il 50-esimo percentile, ossia è il percentile campionario $100p$ quando $p = 0.5$

Supponiamo che i dati di un campione di cardinalità n siano disposti in ordine crescente. Per determinare il $100p$-esimo percentile campionario bisogna determinare quale valore sia:

- maggiore o uguale di almeno np valori dei dati
- minore o uguale di almeno $n(p - 1)$ valori dei dati

Se np non è un intero, il solo valore dei dati che soddisfa questi requisiti è quello la cui posizione è il più piccolo intero maggiore di np.
Se invece np è un intero, allora sia il valore in posizione np che il valore in posizione $np + 1$ soddisfano i due requisti, e quindi il $100p$-esimo percentile campionario è la media dei due valori.

Calcolo del $100p$-esimo percentile campionario di un insieme di dati di n elementi:

1. Si dispongono i dati in ordine crescente
2. Se np non è un intero, si determina il più piccolo intero maggiore di np. Il valore dei dati in questa posizione è il $100p$-esimo percentile campionario.
3. Se np è un intero, allora la media dei valori nelle posizioni np e $np + 1$ è il $100p$-esimo percentile campionario.

Il valore p prende il nome di *quantile di livello*, e a seconda dei valori che può assumere si ottengono statistiche diverse. In particolare si definiscono:

- **Decili**: i percentili multipli di 10, che dividono il campione in 10 parti uguali
- **Quartili**: i percentili multipli di 25, che dividono il campione in 4 parti uguali.

Il 25-esimo percentile campionario si chiama *primo quartile*. Il 50-esimo percentile campionario si chiama *mediana* o *secondo quartile*. Il 75-esimo percentile campionario si chiama *terzo quartile*.

I quartili suddividono i dati in quattro parti in modo tale che il 25% dei dati sia inferiore del primo quartile, il 25% sia compreso tra il primo e il secondo quartile, il 25% tra il secondo e il terzo quartile e il restante 25% sia maggiore del terzo quartile.

3.1.4 Moda campionaria

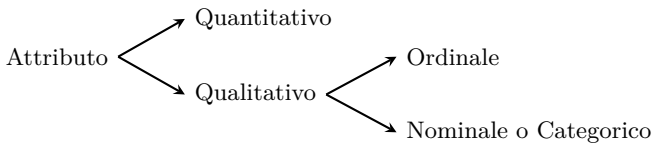
Un altro indicatore della tendenza centrale è la moda campionaria, che è il valore che si verifica con maggiore frequenza nell'insieme di dati.

Se non esiste un singolo valore che si verifica con più frequenza rispetto agli altri, allora tutti i valori con la frequenza più alta sono detti *valori modali*. In questo caso si dice che non c'è un valore unico della moda campionaria.

Questi valori si vedono facilmente in una tabella delle frequenze; sono infatti i valori con la frequenza più alta.

Riepilogo

Si considerino le varie classificazioni degli attributi:



La media si può fare solo per gli attributi quantitativi; la mediana e i percentili si possono svolgere anche sugli attributi qualitativi ordinali con cardinalità del campione dispari; la moda si può fare per qualsiasi tipo di attributo.

3.2 Dispersione

Due campioni A e B possono presentare la stessa centralità ma essere molto diversi tra loro. Si considerino:

$$A : 1, 2, 5, 6, 6 \qquad B : -40, 0, 5, 20, 35$$

Entrambi i campioni hanno la stessa media campionaria e la stessa mediana campionaria, però i valori contenuti nell'insieme B sono decisamente più sparsi di quelli nell'insieme A.

Un modo per misurare la dispersione dei dati è considerare gli scarti dei valori dei dati rispetto ad un valore centrale. Il valore centrale più usato per questo scopo è proprio la media campionaria. Se i valori dei dati sono $x_1, \cdots, n_n$ e la media campionaria è $\bar{x} = \sum_{i=1}^{n} x_i/n$, allora lo scarto del valore x_i dalla media campionaria è $x_i - \bar{x}$ con $i = 1, \cdots, n$.

Si potrebbe pensare di misurare la dispersione totale di un insieme di dati calcolando la media aritmetica degli scarti dalla media. Tuttavia, come abbiamo osservato precedentemente, $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$; questo significa che la somma degli scarti rispetto alla media campionaria è sempre uguale a 0, e di conseguenza lo è anche la media aritmetica degli scarti.

Questo avviene proprio perché gli scarti positivi e negativi si cancellano tra di loro. Si vogliono quindi considerare i singoli scarti indipendentemente dal segno. Si può ottenere questo risultato sia considerando il valore assoluto degli scarti che, come risulta più utile in pratica, il quadrato.

3.2.1 Varianza campionaria

La varianza campionaria è una misura della media degli scarti quadratici rispetto alla media campionaria. Tuttavia, per ragioni tecniche questa "media" divide la somma di n scarti quadratici per $n-1$, piuttosto che per l'usuale valore n.

La varianza campionaria si può calcolare solo per attributi quantitativi, e a differenza degli indici di centralità presenta un problema: la sua unità di misura è diversa da quella dei singoli dati del campione.

La varianza campionaria s^2 dell'insieme di dati $x_1, \cdots, x_n$ di media $\bar{x} = \left(\sum_{i=1}^{n} x_i\right)/n$ è definita come

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

L'identità algebrica che segue è utile per calcolare la varianza campionaria a mano:

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 \; - \; n\,\bar{x}^2 \tag{3.2}$$

Trasformazioni

Traslazione Si supponga di sommare una costante b a ciascuno dei valori $x_1, \cdots, x_n$ per ottenere un nuovo insieme di dati, la varianza campionaria non cambia:

$$y_i = x_i + b \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y^2 = s_x^2$$

Si ricordi che $\bar{y} = \bar{x} + b$ e quindi $y_i - \bar{y} = x_i + b - (\bar{x} + b) = x_i - \bar{x}$. Questo significa che gli scarti di y sono uguali agli scarti di x, e quindi anche le somme dei quadrati sono uguali.

La varianza campionaria quindi non cambia se sommiamo una costante a ciascun valore. *Questa proprietà può essere utilizzata insieme all'identità algebrica* (3.2) *per semplificare il calcolo della varianza campionaria.*

Scalatura Se ciascun valore dei dati viene moltiplicato per a, la varianza campionaria viene moltiplicata per il quadrato di a:

$$y_i = a x_i \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y^2 = a^2 s_x^2$$

Dimostrazione: $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[a(x_i - \bar{x})\right]^2 = a^2\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}) = a^2 s_x^2$

Combinazione Si faccia ora una combinazione delle due trasformazioni precedentemente illustrate:

$$y_i = a x_i + b \;\text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y^2 = a^2 s_x^2$$

3.2.2 Deviazione standard campionaria

La radice quadrata positiva della varianza campionaria si dice deviazione standard campionaria, e si indica con s. Questa è definita come

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

La deviazione standard campionaria, a differenza della varianza campionaria, è espressa nella stessa unità di misura dei dati originali.

Trasformazioni

Traslazione Si supponga di sommare una costante b a ciascuno dei valori $x_1, \cdots, x_n$ per ottenere un nuovo insieme di dati, la deviazione standard campionaria non cambia:

$$y_i = x_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y = s_x$$

Scalatura Se ciascun valore dei dati viene moltiplicato per a, si ottiene che $s_y^2 = a^2 s_x^2$. Calcolando la radice quadrata di entrambi i membri dell'uguaglianza si ottiene che la deviazione standard dei valori y è uguale al valore assoluto di a moltiplicato per la deviazione standard dei valori in x:

$$y_i = ax_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y = |a|\, s_x$$

Combinazione Si faccia ora una combinazione delle due trasformazioni precedentemente illustrate:

$$y_i = ax_i + b \text{ per } i = 1, \cdots, n \quad \Rightarrow \quad s_y = |a|\, s_x$$

La varianza campionaria e la deviazione standard campionaria sono due indici di dispersione che derivano dalla media campionaria.

Due altri indicatori della dispersione di un insieme di dati frequentemente utilizzati sono l'**intervallo di variazione**, ossia la differenza fra il più grande e il più piccolo valore, e lo **scarto interquartile**.

3.2.3 Scarto interquartile

Lo scarto interquartile è un indice di dispersione che deriva dalla mediana campionaria, e rappresenta la lunghezza dell'intervallo in cui si trova la metà centrale dei dati. Richiamando i quartili, possiamo dire che si tratta della lunghezza dell'intervallo compreso tra il primo quartile $Q1$ e il terzo quartile $Q3$.

Un IQR piccolo indica che la metà centrale dei dati è relativamente concentrato attorno alla mediana, mentre un IQR ampio suggerisce una maggiore dispersione nella parte centrale della distribuzione.

Come la mediana campionaria, l'IQR è un indice robusto perché non è influenzato da valori fuori scala. Questo lo rende particolarmente utile quando la distribuzione dei dati è asimmetrica o contiene anomalie.

Questo indice è fondamentale per la costruzione dei boxplot perché viene proprio utilizzato per definire quali valori siano fuori scala e quali no. Generalmente, i valori inferiori a $Q1 - 1.5 \cdot \text{IQR}$ o superiori a $Q3 + 1.5 \cdot \text{IQR}$ sono considerati outlier.

3.3 Indici di correlazione

Si consideri un insieme composto da dati accoppiati $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Per vedere la relazione relativa di queste due variabili è possibile rappresentarle in un diagramma di dispersione. Questo approccio è però qualitativo, e quindi soggetto a interpretazione.
Si vuole trovare un indice quantitativo in grado di rappresentare questa relazione oggettivamente. Questi indici sono detti di dipendenza o associazione e misurano la forza della relazione, ossia forniscono un valore numerico che indica quanto intensamente le variabili siano collegate.

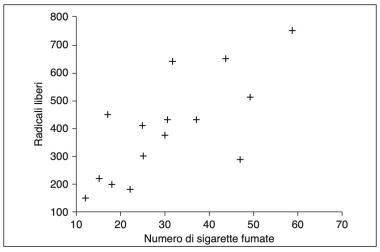
3.3.1 Covarianza campionaria

Si introduce una statistica, detta *covarianza campionaria*, che quantifica in che misura grandi valori di x corrispondano a grandi valori di y e piccoli valori di x a piccoli valori di y. Questo indice quindi misura la tendenza con cui due variabili si muovono insieme ed è definita come la media dei prodotti degli scostamenti delle variabili dalle loro medie.

Relazione tendenziale Si procede considerando una relazione di tipo tendenziale e non deterministico. Ciò significa che le affermazioni che seguiranno varranno tendenzialmente sempre: ci saranno quindi delle eccezioni, ma per lo più saranno valide.

Si supponga che un insieme sia costituito dalle coppie di valori (x_i, y_i) con $i = 1, \cdots, n$. Si calcolino le rispettive medie campionarie $\bar{x}$ e $\bar{y}$. Per la i-esima coppia di dati, si considerino $(x_i - \bar{x})$ lo scarto del valore x rispetto alla sua media campionaria e $(y - \bar{y})$ lo scarto del valore y rispetto alla sua media campionaria. Si procede ora analizzando due casi, ossia quello in cui si presenta una relazione diretta e quello con una relazione inversa.

Quando grandi valori di x tendono a essere associati con grandi valori di y, e piccoli valori di x tendono a essere associati a piccoli valori di y, allora i segni (positivi o negativi) di $(x_i - \bar{x})$ e $(y_i - \bar{y})$ tenderanno a essere gli stessi. A questo punto, se gli scarti hanno segno concorde, il loro prodotto $(x_i - \bar{x})(y_i - \bar{y})$ sarà positivo. Si ottiene che la sommatoria $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tenderà a essere un grande numero positivo.



Sigarette fumate rispetto al numero di radicali liberi.

x "grande" e y "grande"	x "piccolo" e y "piccolo"
$x \geq \bar{x}$ $\quad$ $y \geq \bar{y}$	$x < \bar{x}$ $\quad$ $y < \bar{y}$
$(x_i - \bar{x}) \geq 0$ $\quad$ $(y_i - \bar{y}) \geq 0$	$(x_i - \bar{x}) < 0$ $\quad$ $(y_i - \bar{y}) < 0$
Tendenzialmente:	Tendenzialmente:
$(x_i - \bar{x})(y_i - \bar{y}) \geq 0$	$(x_i - \bar{x})(y_i - \bar{y}) \geq 0$
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) > 0$	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) > 0$

Si individua quindi una correlazione positiva tra le due variabili poiché tendenzialmente presentano segno concorde. In questo caso si parla di relazione tra le due variabili di tipo diretta.

Per lo stesso motivo, quando grandi valori di una variabile tendono a verificarsi in corrispondenza a piccoli valori dell'altra, allora i segni di $(x_i - \bar{x})$ e $(y_i - \bar{y})$ saranno discordi e quindi la sommatoria $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tenderà ad essere un grande numero negativo.

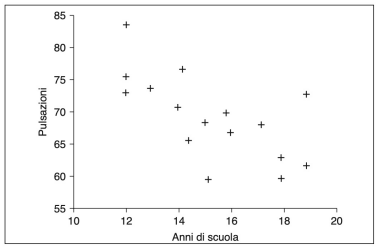


Diagramma di dispersione degli anni di scuola e delle pulsazioni.

x "grande" e y "piccola"	x "piccolo" e y "grande"
$x \geq \bar{x}$ $\quad$ $y < \bar{y}$	$x < \bar{x}$ $\quad$ $y \geq \bar{y}$
$(x_i - \bar{x}) \geq 0$ $\quad$ $(y_i - \bar{y}) < 0$	$(x_i - \bar{x}) < 0$ $\quad$ $(y_i - \bar{y}) \geq 0$
Tendenzialmente:	Tendenzialmente:
$(x_i - \bar{x})(y_i - \bar{y}) < 0$	$(x_i - \bar{x})(y_i - \bar{y}) < 0$
$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) < 0$	$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) < 0$

Si individua quindi una correlazione negativa tra le due variabili poiché tendenzialmente presentano segno discorde. In questo caso si parla di relazione tra le due variabili di tipo indiretta.

Si procede poi standardizzando la sommatoria dividendo per $n-1$, al fine di evitare che questo indice assuma valori troppo elevati. Si osserva che la formula della covarianza campionaria è riconducibile a quello della varianza campionaria, motivo per il quale si possa intuire perché si vada a dividere per $n-1$ e non direttamente per il numero totale di osservazioni.

Ricapitolando, si definisce la covarianza campionaria come:

$$s_{x,y} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \begin{cases} > 0 & \text{relazione diretta} \\ \approx 0 & \text{assenza di relazione / indipendenza} \\ < 0 & \text{relazione indiretta / inversa} \end{cases}$$

3.3.2 Coefficiente di correlazione di Pearson

La covarianza campionaria non può essere posizionata all'interno di una scala assoluta in quanto non è normalizzata e il suo valore dipende dalle osservazioni coinvolte. Si ricava perciò da questo indice il *coefficiente di correlazione lineare campionaria* (anche detto indice di correlazione di Pearson), che si indica con ρ.

Presa la covarianza campionaria, si standardizza il suo valore dividendolo per il prodotto delle due deviazioni standard campionarie delle due variabili:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

Il coefficiente di correlazione di Pearson è quindi un numero puro e, proprio come la covarianza campionaria, quando $\rho > 0$ i dati sono correlati positivamente; invece quando $\rho < 0$ sono correlati negativamente.
Non dipendendo dalle unità di misura, questo indice può essere usato per comparare dataset diversi.

Una proprietà importante, che non verrà dimostrata, è che $-1 \leq \rho \leq 1$

Relazione deterministica

Primo caso Si passi da una relazione tendenziale a una deterministica, in cui la variabile y è una trasformazione lineare della variabile x; tutti i vari indici statistici variano di conseguenza:

$$\forall i \;\; y_i = a + bx_i \quad \Rightarrow \quad \bar{y} = a + b\bar{x} \quad \Rightarrow \quad s_y^2 = b^2 s_x^2 \quad \Rightarrow \quad s_y = |\,b\,| s_x$$

Nella relazione deterministica $y = a + bx$, la costante b rappresenta la pendenza (ossia il coefficiente angolare) della retta che lega le due variabili e indica di quanto varia y all'aumentare di x. Ci si aspetta che:

- se b è positivo, all'aumento di x corrisponde un incremento di $y \Rightarrow$ relazione diretta
- se b è negativo, all'aumento di x corrisponde una diminuzione di $y \Rightarrow$ relazione inversa

Si calcoli ora il coefficiente di correlazione di Pearson:

$$\rho = \frac{b\sum_{i=1}^{n}(x_i - \bar{x})^2}{(n-1)\,|\,b\,|\,s_x^2} = \frac{b}{|\,b\,|} \cdot \frac{1}{s_x^2}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{b}{|\,b\,|} \cdot \frac{1}{\cancel{s_x^2}}\cancel{s_x^2} = \frac{b}{|\,b\,|} \;\; = \;\; \begin{cases} +1 & \text{se } b > 0 \\ -1 & \text{se } b < 0 \end{cases}$$

Questo significa che:

- l'indice ρ è uguale a $+1$ se b è una costante positiva, e se quindi le due variabili esibiscono una relazione di tipo deterministica diretta.
- l'indice ρ è uguale a -1 se b è una costante negativa, e se quindi le due variabili esibiscono una relazione di tipo deterministica indiretta.

Le conclusioni che abbiamo ottenuto con i calcoli rispecchiano le attese iniziali.

Secondo caso Si consideri ora una relazione in cui entrambe le variabili x e y sono soggette a una trasformazione lineare; i vari indici statistici variano nel seguente modo:

$$\forall i \quad \begin{array}{lllll} x_i' = a + bx_i & \Rightarrow & \bar{x}' = a + b\bar{x} & \Rightarrow & s_{x'} = |\,b\,|\,s_x \quad \Rightarrow \quad x_i' - \bar{x}' = b(x_i - \bar{x}) \\ y_i' = c + dy_i & \Rightarrow & \bar{y}' = c + d\bar{y} & \Rightarrow & s_{y'} = |\,d\,|\,s_y \quad \Rightarrow \quad y_i' - \bar{y}' = d(y_i - \bar{y}) \end{array}$$

Si procede calcolando il coefficiente di correlazione di Pearson:

$$\rho' = \frac{\sum_{i=1}^{n}(x_i' - \bar{x}')(y_i' - \bar{y}')}{(n-1)s_x' s_y'} = \frac{b\,d\,\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\,|\,b\,|\,|\,d\,|\,s_x s_y} = \frac{b\,d}{|\,b\,|\,|\,d\,|}\rho \;\; = \;\; \begin{cases} +\rho & \text{se } b \text{ concorda con } d \\ -\rho & \text{se } b \text{ discorda con } d \end{cases}$$

Ciò significa che la correlazione tra x' e y' rimane numericamente invariata rispetto a quella tra x e y e può eventualmente cambiare solo di segno:

- se i coefficienti di trasformazione b e d hanno lo stesso segno allora $\rho' = \rho$
- se i coefficienti di trasformazione b e d hanno segni opposti allora $\rho' = -\rho$

Conclusioni

Il coefficiente di correlazione di Pearson è un indicatore fondamentale per valutare la forza e la direzione di una relazione (o associazione) di tipo lineare tra due variabili, con valori che spaziano fra -1 e $+1$.

Relazione deterministica Quando due variabili presentano una relazione lineare deterministica $y = a + bx$, il coefficiente di correlazione assume valore estremo: $\rho = +1$ se $b > 0$ e $\rho = -1$ se $b < 0$. In altre parole, se tutti i punti giacciono esattamente su una retta crescente (o decrescente), la correlazione è massima (o minima).

Relazione tendenziale Nella maggior parte dei casi reali, le due variabili seguono una relazione lineare tendenziale. In questo contesto, il valore assoluto del coefficiente di correlazione, $|\rho|$, fornisce una misura di quanto le osservazioni si dispongano in prossimità di una retta:

- $|\rho| = 1$ evidenzia una perfetta relazione lineare: in altre parole, è possibile collegare tutti i valori (x_i, y_i) con $i = 1, \cdots, n$ con una retta.
- Più $|\rho|$ si avvicina a 1, e più i dati esibiscono una relazione lineare forte, anche se non perfetta: ciò significa che se anche non esiste una retta che attraversa tutti i valori dei dati, ce n'è una che passa vicino a tutti.
- Se $|\rho|$ è prossimo allo 0, non c'è evidenza di un legame lineare tra le variabili.

Il segno di ρ indica invece la direzione della relazione. Il segno è positivo quando l'approssimazione lineare è crescente (diretta), ed è invece negativo quando l'approssimazione lineare è decrescente (inversa).

È importante tenere a mente che un valore di $\rho = 0$ non implica automaticamente l'assenza di qualsiasi relazione, poiché potrebbero esistere legami non lineari che questo indice non è in grado di cogliere.

Vale inoltre la pena sottolineare che il coefficiente di correlazione di Pearson non implica in alcun modo un rapporto di causa-effetto tra le due variabili prese in considerazione. In altre parole, due variabili possono presentare un valore di correlazione elevato senza che una determini o causi l'altra. Spesso, infatti, può intervenire un terzo fattore (o più fattori) a influenzare entrambe le variabili, generando un legame che in realtà non corrisponde a un meccanismo causale diretto.

3.4 Eterogeneità

Per le variabili qualitative nominali non è possibile calcolare la varianza né gli indici che ne derivano, perché non esistono una media, una mediana o altri valori numerici di riferimento su cui misurare distanze. Risulta comunque necessario avere un indice che misuri la dispersione della distribuzione delle frequenze, detta *eterogeneità*. In particolare si dice che una variabile è distribuita in modo eterogeneo quando ogni suo valore compare con la stessa frequenza.

3.4.1 Indice di Gini

Si consideri un campione $\{x_1, \cdots, x_n\}$ in cui occorrono i valori distinti $v_1, \cdots, v_m$, e si indichi con f_j la frequenza relativa dell'elemento v_j per $j = 1, \cdots, m$. Si definisce l'*indice di eterogeneità di Gini* come:

$$I = 1 - \sum_{j=1}^{m} f_j^2$$

Una proprietà importante di questo indice è che $0 \leq I < 1$. Inoltre l'omogeneità massima dell'insieme di dati si presenta quando $I = 0$, mentre l'eterogeneità massima la si ha quando $I \to 1$. Di conseguenza, più aumenta il valore dell'indice di Gini e più aumenta il grado di eterogeneità.

Per dimostrare le limitazioni inferiori e superiori si ricordi innanzitutto che, trattandosi di frequenze relative, $0 \leq f_j \leq 1 \ \ \forall j \in \{1, \ldots, m\}$. Inoltre $\sum_{j=1}^{m} f_j = 1$. Di conseguenza si avrà:

- per almeno un j si ha $f_j > 0 \Rightarrow f_j^2 > 0 \Rightarrow \sum_{j=1}^{m} f_j^2 > 0 \quad \Rightarrow \quad I < 1$
- per ogni j, dato che $0 \leq f_j \leq 1$, si ha che $f_j^2 \leq f_j \Rightarrow \sum_{j=1}^{m} f_j^2 \leq \sum_{j=1}^{m} f_j = 1 \quad \Rightarrow \quad I \geq 0$

Si noti, come accennato precedentemente, che l'estremo inferiore si presenta quando l'insieme è massimamente omogeneo e l'estremo superiore quando è massimamente eterogeneo:

- l'eterogeneità minima la si ha quando tutti gli elementi hanno lo stesso valore, e quindi
  $\exists\, k \in [1, m] \mid f_k = 1, \ \forall j \neq k \ \ f_j = 0 \quad \Rightarrow \quad I = 1 - \sum_{j=1}^{m} f_j^2 = 1 - f_k^2 = 1 - 1 = 0$
- l'eterogeneità massima la si ha quando tutti gli elementi hanno la stessa frequenza, e quindi
  $\forall j \in [1, m] \ \ f_j = \frac{1}{m} \quad \Rightarrow \quad I = 1 - \sum_{j=1}^{m} f_j^2 = 1 - \sum_{j-1}^{m} \frac{1}{m^2} = 1 - \frac{1}{m} = \frac{m-1}{m} \ \ \to 1$ al crescere di m

Indice di Gini normalizzato Si ricordi che m è la cardinalità dell'insieme dei valori distinti. Questo indice presenta due problematiche:

1. il valore massimo che può assumere, ossia quando l'insieme di dati è massimamente eterogeneo, è $(m-1)/m$. Di conseguenza, specialmente nel caso in cui non si conosca il valore m, non si può sapere quanto questo indice debba tendere a 1 affinché si abbia la massima eterogeneità nell'insieme dei dati.
2. il suo valore dipende fortemente dal valore di m. Non è quindi possibile confrontare 2 attributi qualitativi che presentano intervalli di valori diversi, ossia m diverso.

Per ovviare a questi problemi si introduce l'*indice di Gini normalizzato*, che si ottiene dividendo l'indice di Gini per il valore massimo $(m-1)/m$ che può assumere:

$$I' = \frac{m \cdot I}{m-1}$$

Questo indice può assumere anche 1 come valore: $0 \leq I' \leq 1$. Si consideri infatti il caso in cui l'eterogeneità di un insieme di dati è massima:

$$I = \frac{m-1}{m} \quad \Rightarrow \quad I' = \frac{m \cdot I}{m-1} = \frac{m \cdot (m-1)}{(m-1) \cdot m} = 1$$

3.4.2 Indice di entropia

Si consideri un campione $\{x_1, \cdots, x_n\}$ in cui occorrono i valori distinti $v_1, \cdots, v_m$, e si indichi con f_j la frequenza relativa dell'elemento v_j per $j = 1, \cdots, m$. Si definisce l'*indice di entropia* del campione come:

$$H = \sum_{j=1}^{m} f_j \log \frac{1}{f_j} = -\sum_{j=1}^{m} f_j \log f_j$$

La funzione $I(p) = \log 1/p = -\log p$ è detta *autoinformazione* e misura la quantità di informazione ottenuta dal verificarsi di un evento con probabilità p. In altre parole, misura quanto viene ridotta l'incertezza una volta che sappiamo che l'evento si è effettivamente realizzato. Questa funzione è descrescente monotona, vale 0 quando $p = 1$ e tende a infinito per p che tende a 0.

Nel calcolo dell'entropia compare $-f_j \log f_j$. Se $f_j = 0$ questa espressione assume la forma indeterminata $0 \cdot \infty$. È però possibile estendere la definizione dell'entropia anche nei casi in cui alcune frequenze relative siano nulle. Valutando il limite $\lim_{f_j \to 0^+} -f_j \log f_j = 0$ si definisce per convenzione $-0 \log 0 = 0$.

Si effettuano le seguenti osservazioni:

- $\forall j$ vale $-f_j \log f_j \geq 0 \quad \Rightarrow \quad H \geq 0$
- $\forall j$ si ha che $-f_j \log f_j = 0 \quad \Leftrightarrow \quad f_j = 0 \vee \log f_j = 0$ e quindi $f_j = 1$. Pertanto $H = 0$ se e solo se ci si trova in condizione di massima omogeneità, e quindi tutti i dati del campione assumono lo stesso valore.
- In caso di invece massima eterogeneità si avrà $f_j = 1/m$ e quindi

$$H = \sum_{j=1}^{m} \frac{1}{m} \log m = m \left(\frac{1}{m} \log m \right) = \log m$$

  Si può dimostrare che in questo caso l'entropia assume il valore massimo.

Una proprietà importante di questo indice è quindi che $0 \leq H \leq \log m$. Più questo indice cresce, e più aumenta il grado di eterogeneità dell'insieme, viceversa più descresce e più aumenta il grado di omogeneità.

Indice di entropia normalizzato Si definisce l'*indice di entropia normalizzato* come

$$H' = \frac{H}{\log m}$$

I valori di questo indice sono compresi tra 0 e 1, infatti nel caso di massima eterogeneità si ha che:

$$H = \log m \quad \Rightarrow \quad H' = \frac{\log m}{\log m} = 1$$

Se il logaritmo è in base 2 allora l'entropia si misura in bit: ciò risulta utile quando bisogna svolgere i calcoli in un computer; è comunque possibile usare altre basi, come per sempio il logaritmo naturale e quello in base 10.

3.5 Concentrazione

Le misure di concentrazione sono strumenti statistici che consentono di comprendere come una determinata risorsa o bene – ad esempio la ricchezza – sia distribuita all'interno di una popolazione. In questo modo, è possibile valutare se tale risorsa sia distribuita in maniera equa tra tutti gli individui oppure se risulti concentrata in un numero ristretto di soggetti.

Mentre la varianza quantifica la dispersione dei singoli valori rispetto alla media, la concentrazione mette in evidenza se una piccola parte della popolazione detiene una quota sproporzionata del bene considerato.

Si consideri un campione di n osservazioni, ciascuno dei quali possiede una certa quantità di risorse. Si indichi con a_i la quantità posseduta dall'i-esimo individuo dopo aver ordinato le osservazioni in ordine crescente, ossia $a_1 \leq a_2 \leq \cdots \leq a_n$. Il valore medio della risorsa è definito come $\bar{a} = 1/n \sum_{i=1}^{n} a_i$, dove la sommatoria rappresenta la somma di tutte le dotazioni individuali. Moltiplicando il valore medio $\bar{a}$ per il numero totale degli individui n otteniamo il totale aggregato della risorsa:

$$TOT = n\,\bar{a} = \sum_{i=1}^{n} a_i$$

Qui la somma viene effettuata su tutte le osservazioni $a_1, a_2, \ldots, a_n$, cioè su tutte le dotazioni della risorsa in esame. L'ordinamento in ordine crescente serve per facilitare l'analisi della distribuzione della risorsa fra gli individui (come vedremo dopo per la curva di Lorenz).

Si possono presentare due situazioni estreme:

- caso di concentrazione minima: tutti gli elementi del campione assumono lo stesso valore: $a_1 = a_2 = \cdots = a_n = \bar{a}$
- caso di concentrazione massima: tutti gli elementi del campione assumono valore pari a 0, tranne uno: $a_1 = a_2 = \cdots = a_{n-1} = 0$ e $a_n = n\,\bar{a}$

È necessario avere un indice di concentrazione che valga 0 e 1 nei casi rispettivamente di concentrazione minima e massima, e che sia negli altri casi un valore crescente in funzione della concentrazione. Si considerino:

- la frequenza relativa cumulata degli individui fino alla i-esima osservazione:

$$F_i = \frac{i}{n} \quad \text{per } i = 1, \cdots, n \qquad \% \text{ degli inidividui}$$

- la quantità relativa cumulata fino all'i-esima osservazione:

$$Q_i = \frac{1}{TOT} \sum_{k=1}^{i} a_k \qquad \% \text{ della ricchezza}$$

Queste due quantità possiedono le seguenti proprietà:

- $0 \leq F_i, Q_i \leq 1$
- $Q_i = F_i$ nel caso di concentrazione minima
- $Q_n = F_n = 1$
- $Q_i \leq F_i$ siccome le osservazioni sono state ordinate in modo crescente.

Dimostrazione

Si vuole provare che $Q_i \leq F_i$. Pertanto si divide l'insieme ordinato in due sottogruppi, $\{a_1, \cdots, a_i\}$ e $\{a_{i+1}, \cdots, a_n\}$, e si definiscono le rispettive somme S_i e T_i:

$$S_i = \sum_{k=1}^{i} a_k \qquad T_i = \sum_{k=i+1}^{n} a_k \qquad TOT = S_i + T_i = S_n$$

Si cominci riscrivendo la disuguaglianza $Q_i \leq F_i$ in termini di S_i e T_i. In particolare, si osserva che

$$Q_i = \frac{S_i}{TOT} \leq \frac{i}{n} \iff \frac{S_i}{S_i + T_i} \leq \frac{i}{n}$$

Da quest'ultima forma, si vuole isolare da un lato della disequazione $\frac{i\,T_i}{S_i}$:

$$\frac{S_i}{S_i + T_i} \leq \frac{i}{n} \Rightarrow \frac{1}{1 + \frac{T_i}{S_i}} \leq \frac{i}{n} \Rightarrow 1 + \frac{T_i}{S_i} \geq \frac{n}{i} \Rightarrow i\left(\frac{T_i}{S_i}\right) \geq i\left(\frac{n}{i} - 1\right) \Rightarrow \frac{i\,T_i}{S_i} \geq n - i$$

Si scompone ora il termine $\frac{i\,T_i}{S_i}$ come somma sugli elementi a_k con $k > i$:

$$\frac{i\,T_i}{S_i} = \frac{i}{S_i} \sum_{k=i+1}^{n} a_k = \sum_{k=i+1}^{n} \frac{i\,a_k}{S_i}$$

Questa rielaborazione permette di sfruttare l'ordinamento $a_k \geq a_i \ \ \forall i < k$. Infatti, se $a_k \geq a_i$, allora:

$$\frac{i\,a_k}{S_i} = \frac{\overbrace{a_k + a_k + \cdots + a_k}^{\text{i volte}}}{a_1 + a_2 + \cdots + a_k} \geq 1$$

Ne consegue che

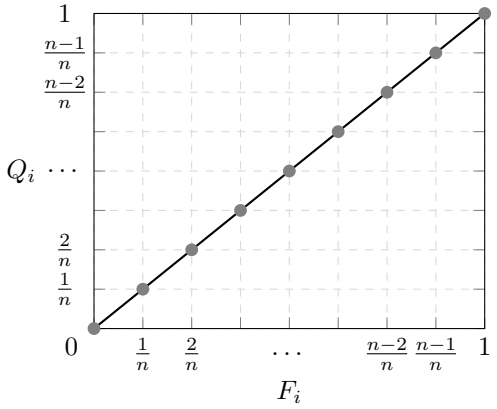
$$\frac{i\,T_i}{S_i} = \sum_{k=i+1}^{n} \frac{i\,a_k}{S_i} \geq \sum_{k=i+1}^{n} 1 = n - (i+1) + 1 = n - i$$

In tal modo si conclude che $\frac{i\,T_i}{S_i} \geq n - i$, che equivale, tramite le equivalenze iniziali, a

$$\frac{i\,T_i}{S_i} \geq n - i \Rightarrow \frac{S_i}{S_i + T_i} \leq \frac{i}{n} \Rightarrow Q_i \leq F_i$$

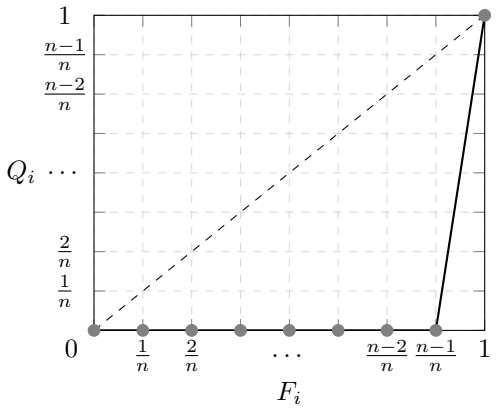
In conclusione si è dimostrato che, ordinando i dati in modo crescente, la quantità cumulata Q_i risulta sempre minore o uguale alla frequenza cumulata F_i.

Per $i = 1, \cdots, n$ le coppie (F_i, Q_i) indicano che il $100F_i\%$ della popolazione detiene il $100Q_i\%$ della quantità considerata. Si considerino ora i punti sul piano che sono identificati da queste coppie.



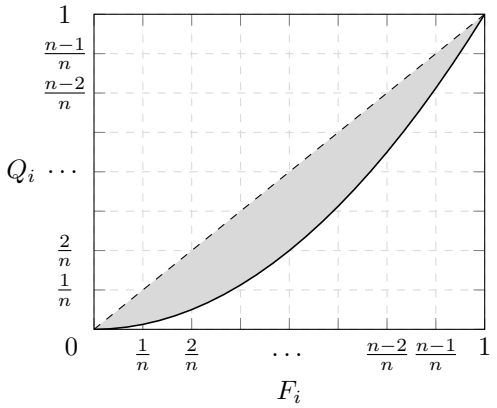
Concentrazione minima Nel caso di concentrazione minima tutti i punti (F_i, Q_i) giacciono sulla retta $F = Q$: si può dunque dire che in questo caso $F_i - Q_i = 0$ per ogni i.

$$a_i = \bar{a}\ , \cdots, \ \bar{a}\ , \bar{a} \qquad a_i = \bar{a} \quad \forall i = 1, \ldots, n$$
$$Q_i = \tfrac{\bar{a}}{TOT}, \cdots, \tfrac{(n-1)\bar{a}}{TOT}, \tfrac{n\,\bar{a}}{TOT} \qquad Q_i = \frac{i\,\bar{a}}{TOT} = \frac{i\,\bar{a}}{n\,\bar{a}} = \frac{i}{n} = F_i$$



Concentrazione massima Nel caso di concentrazione massima tutti i punti (F_i, Q_i) giacciono sulla retta $Q = 0$, tranne per l'ultimo in cui $F_n = Q_n$: dunque in questo caso $F_i - Q_i = F_i$ per $i = 1, \cdots, n-1$ e $F_n - Q_n = 0$.

$$a_i = 0\,, 0\,, \quad \cdots, \quad 0 \quad , TOT$$
$$Q_i = 0\,, 0\,, \quad \cdots, \quad 0 \quad , 1$$
$$F_i = \frac{1}{n}, \frac{2}{n}, \quad \cdots, \quad \frac{n-1}{n}, 1$$



Concentrazione intermedia Nei casi intermedi si avrà che i punti staranno su una curva sotto la bisettrice del I° e III° quadrante $F = Q$, dato che $Q_i \leq F_i$ per qualsiasi $i = 1, \cdots, n$. Più la curva si avvicina alla bisettrice, e più la concentrazione è bassa, mentre più si allontana e più la concentrazione è alta.

La curva dei punti (F_i, Q_i) è detta *curva di Lorenz*. L'area compresa tra la curva di Lorenz e la retta di equidistribuzione (la bisettrice) è detta area di concentrazione e può essere utilizzata come base per la definizione di appositi indici di concentrazione: maggiore infatti è la concentrazione osservata e maggiore sarà quell'area.

3.5.1 Indice di concentrazione di Gini

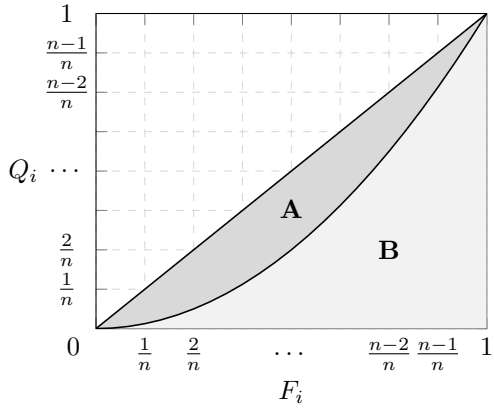
I diagrammi illustrati precedentemente sono però degli strumenti qualitativi. Si vuole perciò introdurre un indice numerico calcolato a partire dai dati, il cui valore possa facilemnte essere confrontato con due estremi minimo e massimo. Si definisce quindi l'*indice di concentrazione di Gini*:

$$G = \sum_{i=1}^{n-1}(F_i - Q_i)$$

Si osserva che non viene considerato il caso in cui $i = n$ in quanto $F_n - Q_n = 0$ sempre.

Per interpretare questo indice come l'area nel diagramma (F, Q), occorre introdurre una somma di Riemann. Ogni differenza $(F_i - Q_i)$ va infatti molitplicata per l'ampiezza in ascissa ΔF_i: trovandoci in un contesto discreto con dati equispaziati, ciò si traduce in $\Delta F_i = 1/n$. Di conseguenza, la somma $\sum(F_i - Q_i)$ diventa n volte l'area effettiva. Dividendo quindi questa sommatoria per n, si ottiene un indice G_{area} che riflette la superficie compresa tra la retta di equidistribuzione e la curva di Lorenz:

$$G_{area} = \sum_{i=1}^{n-1}[(F_i - Q_i)\Delta F_i] = \frac{1}{n}\sum_{i=1}^{n-1}(F_i - Q_i)$$



Questo indice misura l'area compresa tra la bisettrice $F = Q$ e la curva di Lorenz. Nel grafico è rappresentata da A.

Nel caso di concentrazione minima la curva di Lorenz si appiattisce e coincide con la bisettrice: di conseguenza l'indice avrà valore minimo 0.

Nel caso di concentrazione massima la curva di Lorenz coincide con l'asse orizzontale, e quindi l'area è rappresentata da $A+B$. L'indice avrà valore massimo $1/2$, ossia l'area della porzione di piano compreso tra la bisettrice e l'asse orizzontale.

Calcolando algebricamente il valore massimo di G_{area} si trova che in realtà questo non assume mai valore pari a $1/2$. Ricordando che nel caso di concentrazione massima $F_i - Q_i = F_i$ per $i, \cdots, n-1$:

$$G_{area}(\max) = \frac{1}{n}\sum_{i=1}^{n-1}F_i = \frac{1}{n}\sum_{i=1}^{n-1}\frac{i}{n} = \frac{1}{n^2}\sum_{i=1}^{n-1}i = \frac{1}{n^2}\frac{(n-1)(n-\cancel{1}+\cancel{1})}{2} = \frac{n-1}{2n}$$

Di conseguenza si trova che per n grande, nel caso di concentrazione massima, il valore di G_{area} tende a $1/2$. Si è dimostrato che $0 \leq G_{area} < 1/2$.

Indice di concentrazione di Gini normalizzato Si consideri G_{area}, lo si normalizza dividendolo per il suo valore massimo $\sum F_i$ che si presenta nel caso di concentrazione massima:

$$G' = \frac{2n}{n-1}\sum_{i=1}^{n-1}[(F_i - Q_i)\Delta F_i] = \frac{2n}{n-1}\frac{1}{n}\sum_{i=1}^{n-1}(F_i - Q_i) = \frac{2}{n-1}\sum_{i=1}^{n-1}(F_i - Q_i)$$

Si osserva che $0 \leq G' \leq 1$. Si noti che si è arrivati a tale conclusione partendo da G_{area}, ma è possibile arrivare al medesimo risultato dividendo G per il suo valore massimo $(n-1)/2$.

Considerando il grafico, questo indice indica il rapporto $A/(A+B)$, dove $(A+B)$ rappresenta proprio l'area dell'indice non normalizzato nel caso di concentrazione massima.

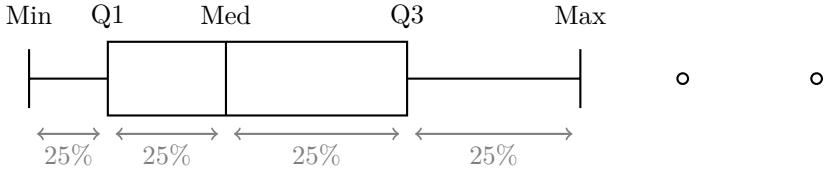
4 Altro

4.1 Altri grafici

4.1.1 Box Plot

Per visualizzare alcune statistiche riassuntive di un insieme di dati si usa un *box plot* (diagramma a scatola). Per realizzarlo tracciamo un segmento orizzontale dal minore al maggiore dei dati. Al segmento sovrapponiamo un rettangolo che si estende dal primo al terzo quartile. Il rettangolo è diviso in due parti da un segmento verticale in corrispondenza della mediana campionaria.

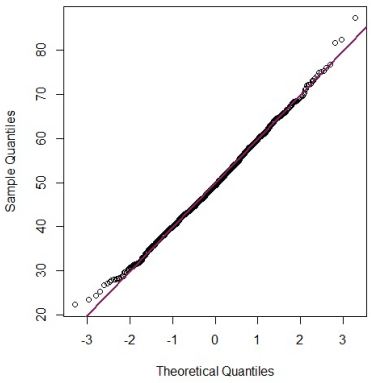
La lunghezza della base del rettangolo corrisponde allo scarto interquartile.



In un box plot ciascuno dei quattro segmenti contiene il 25% delle osservazioni, ossia un quarto dei dati, però la lunghezza di ciascun segmento sulla scala orizzontale dipende dalla distribuzione dei valori. Se i dati sono più concentrati in un certo intervallo, quel tratto sarà più corto. I quartili quindi dividono le osservazioni in parti uguali sul numero dei dati, non sulla distanza numerica.

I pallini a destra del box plot rappresentano dei valori fuori scala, determinati tramite l'utilizzo dell'IQR.

4.1.2 Q-Q Plot



Un diagramma Q-Q (o diagramma quantile-quantile) è una rappresentazione grafica qualitativa che permette di verificare le similarità tra le distribuzioni di due campioni diversi (utile per vedere quindi se seguono una stessa distribuzione).

Questi diagrammi si basano sul fatto che i quantili campionari rappresentino l'approssimazione di quantili teorici che, se considerati tutti insiemi, individuano la distribuzione dei dati.

Ogni asse cartesiano di questo diagramma contiene i quantili dei due campioni presi in considerazione. Poiché i quantili sono ordinati in modo crescente, anche il grafico risultante sarà crescente, o perlomeno non decrescente.

Se due campioni hanno una distribuzione uguale, allora estraendo da entrambi il quantile di un livello fissato si dovranno ottenere due numeri vicini. In questo caso i punti del diagramma Q-Q tenderanno ad allinearsi alla bisettrice del I° e III° quadrante.

4.2 Distribuzioni normali

Dati normali Un insieme di dati si dice *normale* se il rispettivo istogramma ha le proprietà seguenti:

- L'istogramma è simmetrico rispetto all'intervallo centrale
- Ha il punto massimo in corrispondenza dell'intervallo centrale
- Spostandoci dal centro verso destra o verso sinistra, l'altezza diminuisce in modo tale che l'intero istogramma è a forma di campana.

Se l'istogramma di un insieme di dati è vicino a essere un istogramma normale, allora diciamo che l'insieme di dati è approssimativamente normale. Inoltre l'insieme di dati si dice assimmetrico a destra o a sinistra a seconda di quale sia la coda più lunga.

A causa della simmetria dell'istogramma normale, la media e la mediana di un insieme di dati approssimativamente normale sono uguali o molto prossime.

Regola empirica Siano $\bar{x}$ e s rispettivamente la media e la deviazione standard campionarie di un insieme di dati approssimativamente normale. La *regola empirica* specifica le proporzioni approssimate delle osservazioni che si trovano a una distanza di s, $2s$ e $3s$ da $\bar{x}$:

- circa il 68% delle osservazioni rientrano nell'intervallo $\bar{x} \pm s$
- circa il 95% delle osservazioni ritentrano nell'intervallo $\bar{x} \pm 2s$
- circa il 99.7% delle osservazioni ritentrano nell'intervallo $\bar{x} \pm 3s$

Insiemi di dati bimodali Un insieme di dati ottenuto campionando una popolazione costituita da sottogruppi eterogenei non è di solito normale. Piuttosto, l'istogramma di un insieme di dati di questo tipo spesso rassomiglia a una sovrapposizione di istogrammi normali e quindi spesso ha due o più massimi locali. Questi massimi locali si comportano come mode. Un insieme di dati il cui istogramma ha due massimi locali si dice quindi *bimodale*.

In questi casi, quando nei dati si hanno due popolazioni ben distinte per quanto riguarda un certo attributo, ha senso dividere i dati in base a queste popolazioni e ottenere un insieme normale.

4.3 Trasformazione dei dati

Può risultare utile una rielaborazione dei dati iniziali per diversi motivi: per poterli confrontare con altri dati riportandoli ad un intervallo predefinito, per poter confrontare la loro distribuzione di frequenza con quella di altri dati oppure per renderli più facilmente leggibili.

Si consideri un insieme di valori distinti $V = \{v_1, v_2, \cdots, v_m\}$, ognuno con la propria frequenza relativa $f_1, f_2, \cdots, f_m$. Si consideri anche una funzione f che trasformi i valori di V in valori appartenenti all'insieme $V' = \{v'_1, v'_2, \cdots, v'_m\}$: si ha perciò che $\forall j \in \{1, \ldots, m\}$ $f(v_j) = v'_j$.

Si prenderanno in esame solo funzioni *iniettive*: per questo tipo di trasformazioni i valori delle frequenze relative per l'insieme V' rimangono i medesimi di quelli per l'insieme V, ossia $f_1, f_2, \cdots, f_m$.

Si analizzera come variano gli indici statistici, e di conseguenza il grafico della distribuzione, a seconda delle trasformazioni che verranno effettuate. Verranno analizzate solo le trasformazioni che prevedono di applicare ai dati una funzione *lineari*. Fissate perciò due costanti $a, b \in \mathbb{R}$ si avrà che $v' = f(v) = a\,v + b$.

4.3.1 Traslazione

Se si vogliono traslare i dati di una quantità costante $k \in \mathbb{R}$, si applica la trasformazione $f(x) = x + k$. Per $k > 0$ si trasla verso destra e per $k < 0$ si trasla verso sinistra.

Questa trasformazione è utile quando i valori osservati sono molto grandi e sono poco dispersi attorno ad un valore centrale. Si osserva che:

- gli indici di centralità, quali media, quantili e mediana, vengono traslati della stessa quantità k.
- gli indici di dispersione, quali range (dei dati), IQR, varianza e devizazione standard, dell'insieme traslato V' rimangono invece gli stessi dell'insieme di partenza V.

La traslazione è una trasformazione iniettiva, e quindi i dati osservati varieranno nei loro valori ma le relative frequenze rimarranno uguali. Qualora si rappresentasse graficamente la distribuzione dei dati originali e quella dei dati trasformati, si osserverebbe che la forma dei due grafici non subirebbe alterazioni: ciò che cambia sono solo i valori dei dati l'ungo l'asse delle ascisse.

Si conclude che la traslazione comporta uno spostamento dell'*origine* del sistema di riferimento, ovvero il punto in cui si trova lo zero sull'asse delle ascisse. Le relazioni interne, e di conseguenza le proprietà della distribuzione come la forma, la simmetria e gli indici di dispersione, rimangono invariate.

4.3.2 Scalatura

Se si vogliono dilatare o contrarre i dati di un fattore costante $h \in \mathbb{R}^+$ si applica la trasformazione $f(x) = hx$. Se $h > 1$ il range dei valori risulta aumentato, ed è stata quindi applicata una dilatazione, mentre per $0 < h < 1$ si applica una contrazione del range dei valori.

Si noti che non viene considerato il caso in cui $h < 0$ in quanto, oltre alla dilatazione o contrazione, i dati vengono specchiati rispetto all'asse delle ordinate.

Si consideri $h = 1/k$. Se k è minore del valore minimo nel campione, allora tutti i valori trasformati saranno maggiori di 1, mentre se k è maggiore del valore massimo, allora tutti i valori trasformati saranno minori di 1[1].

Si osserva che:

- gli indici di centralità vengono scalati della stessa quantità h.
- il range di variazione e l'IQR vengono scalati della stessa quantità h.
- la varianza viene scalata di una quantità h^2 mentre la deviazione standard viene scalata di $|h|$.

Anche in questo caso si verifica che i dati osservati varieranno nei loro valori ma le relative frequenze rimarranno uguali. I grafici che rappresentano la distribuzione dei dati originali e quella dei dati trasformati hanno quindi la medesima forma ma valori diversi sull'asse delle ascisse.

4.3.3 Cambiamento di origine e scala

Si abbia un insieme di valori nell'intervallo (a, b) e vogliamo adattarli in modo che appartengano all'intervallo (c, d), la trasformazione da applicare sarà:

$$f(x) = c + \frac{d-c}{b-a}(x-a)$$

Dimostrazione

Metodo delle rette La funzione che trasforma i valori nell'intervallo (a, b) all'interno dell'intervallo (c, d) è una retta, la cui equazione si ricava tramite la formula della retta passante per due punti. Sia $(a, c) = (x_0, y_0)$ e $(b, d) = (x_1, y_1)$:

$$\frac{y-y_0}{y_1-y_0} = \frac{x-x_0}{x_1-x_0} \;\Rightarrow\; \frac{x'-x'_0}{x'_1-x'_0} = \frac{x-x_0}{x_1-x_0} \;\Rightarrow\; \frac{x'-c}{d-c} = \frac{x-a}{b-a} \;\Rightarrow\; f(x) = x' = c + \frac{d-c}{b-a}(x-a)$$

Metodo delle condizioni Si consideri la funzione $f : [a, b] \to [c, d]$. In generale, una trasformazione lineare ha la forma $f(x) = mx + q$. Per determinare m e q si impongono le condizioni:

1. $f(a) = c$, ossia $ma + q = c$
2. $f(b) = d$, ossia $mb + q = d$

Sottraendo la prima dalla seconda si ottiene:

$$m(b-a) = d-c \;\Rightarrow\; m = \frac{d-c}{b-a}$$

Sostituendo m nella prima equazione:

$$\frac{d-c}{b-a}a + q = c \;\Rightarrow\; q = c - \frac{d-c}{b-a}a$$

Di conseguenza la funzione diventa:

$$f(x) = \frac{d-c}{b-a}x + \left(c - \frac{d-c}{b-a}a\right) = c + \frac{d-c}{b-a}(x-a)$$

[1] Questa proprietà vale solo quando i dati del campione $\in \mathbb{R}^+$. Non vale se sono presenti valori negativi.

Scegliendo un apposito intervallo (c, d), è possibile trasformare un campione di dati affinché le osservazioni siano compresi tra due valori significativi. Utilizzando l'equazione ottenuta:

- nel caso in cui si vogliano mappare i valori in $(0, 1)$ si applica una funzione $f(x) = \dfrac{x-a}{b-a}$
- nel caso in cui si vogliano mappare i valori in $(-1, 1)$ si applica una funzione $f(x) = 2\dfrac{x-a}{b-a} - 1$

Si osserva che questa mappatura lineare conserva l'ordine dei dati.

Standardizzazione

La standardizzazione è un caso particolare di cambiamento di origine e scala, e consiste nello traslare verso sinistra rispetto alla media dei valori delle osservazioni, per poi applicare una scala il cui fattore è il reciproco della deviazione standard dei valori. Indicando con $\bar{x}$ e con s_x rispettivamente la media campionaria e la deviazione standard campionaria dei valori, la trasformazione che si ottiene sarà:

$$f(x) = \frac{x - \bar{x}}{s_x}$$

La standardizzazione è quindi un'operazione di trasformazione lineare che prevede una centratura, ossia la sottrazione della media, e una uniformazione, ossia la divisione per la deviazione standard. Tramite la centratura si ottiene un nuovo campione la cui media sia zero, e tramite l'uniformazione si rimuove l'unità di misura alle osservazioni, rendendoli numeri adimensionali, e si ha come deviazione standard per il campione 1. Infatti:

$$\bar{x}' = \frac{\bar{x} - \bar{x}}{s_x} = 0 \qquad s^2_{x'} = \frac{1}{s^2_x} = 1 \quad \Rightarrow \quad s_{x'} = 1$$

4.3.4 Trasformazioni logaritmiche

A volte i valori di una variabile osservata sono molto grandi oppure molto distanziati. In questi casi può essere utile considerare non tanto il valore originale ma, pensando a tale valore come potenza di una data base, ragionare in termini del relativo esponente. Ciò corrisponde ad applicare una trasformazione logaritmica del seguente tipo:

$$f(x) = \log x$$

La scelta della base del logaritmo in questa funzione dipende dal contesto in cui bisogna applicarlo: basi comuni sono il 10, la costante e oppure il 2.

Nel caso i valori siano molto distanziati tra loro e caratterizzati da una distribuzione di frequenza unimodale fortemente asimmetrica, la trasformazione logaritmica permette di ottenere una distribuzione di frequenza più simmetrica. Questo tipo di trasformazione ha molti altri vantaggi, dovuti al fatto che l'operazione di prodotto (o quoziente) tra due valori viene trasformata nella somma (o nella differenza) dei rispettivi logaritmi.

4.4 Alberi di decisione

Gli indici di eterogeneità non servono solo per misurare la dispersione delle frequenze nelle variabili qualitative, ma trovano anche un'applicazione fondamentale nella costruzione degli alberi di decisione. In un albero di decisione, ogni oggetto da classificare è descritto da un vettore di attributi, e la classificazione avviene valutando, a partire dalla radice dell'albero, condizioni sui valori di tali attributi.

In pratica, ad ogni nodo dell'albero viene associata una condizione (o test) che suddivide il campione in sottoinsiemi: si percorre la freccia corrispondente in base al risultato del test, fino a raggiungere una foglia, la quale indica la classe assegnata. La scelta del test in ciascun nodo è guidata proprio dagli indici di eterogeneità: l'obiettivo è quello di ridurre l'eterogeneità dei dati nei nodi figli rispetto a quella del nodo padre.

Si cerca quindi di porre nei vari nodi domande che permettono di ottenere sottogruppi il più omogenei possibile. Così facendo si riduce il numero di domande da fare.

Ad esempio, utilizzando l'indice di Gini si seleziona la condizione che minimizza l'indice nei gruppi risultanti, cioè quella che porta a sottoinsiemi in cui la distribuzione delle classi è il più possibile concentrata in una sola categoria. Analogamente, se si impiega l'indice di entropia, si cerca la divisione che riduce al minimo l'incertezza (ovvero l'entropia) nei nodi successivi. In entrambi i casi, il criterio adottato assicura che, procedendo lungo l'albero, si raggiungano foglie contenenti gruppi di oggetti omogenei rispetto alla classe di appartenenza.

Così, l'impiego degli indici di eterogeneità consente di valutare quantitativamente la bontà delle suddivisioni, contribuendo a costruire alberi di decisione efficaci per il compito di classificazione.

4.5 Analisi di classificatori

Un classificatore è un meccanismo che, dati degli oggetti su cui si desidera effettuare una distinzione, associa a ciascun oggetto una classe tra quelle disponibili. Matematicamente, ciò equivale ad avere un insieme Ω (che raccoglie i campioni da classificare), e un insieme di etichettte $\mathcal{C}$. Il classificatore è quindi una funzione $c : \Omega \to \mathcal{C}$ che. a ogni elemento $x \in \Omega$, associa una delle etichette in $\mathcal{C}$.

Classificatore binario Sia Ω l'insieme degli oggetti su cui vogliamo effettuare la classificazione, e supponiamo che a ciascun oggetto $x \in \Omega$ sia associata, in modo noto, un'etichetta $\gamma(x) \in \mathcal{C} = \{\text{positivo}, \text{negativo}\}$.

Un classificatore binario è una funzione $c : \Omega \to \{\text{positivo}, \text{negativo}\}$ che, a partire da un oggetto x, ne prevede la classe, positiva o negativa che sia.

Una volta definito il classificatore c, la sua prestazione rispetto a un insieme di oggetti di test $\mathcal{D} \subseteq \Omega$, di cui sono note le classi vere $\gamma(x)$, si può valutare confrontando l'uscita $c(x)$ con la classe effettiva $\gamma(x)$. Da questo confronto si possono verificare quattro possibili casi: true positive, false negative, true negative e false positive.

I valori ottenuti possono essere organizzati in delle matrici, dette *matrici di confusione*.

		Valore Effettivo P	N
Valore Predetto	P	TP	FP
	N	FN	TN

TP: oggetti x con $\gamma(x) =$ positivo e $c(x) =$ positivo
FN: oggetti x con $\gamma(x) =$ positivo e $c(x) =$ negativo
TN: oggetti x con $\gamma(x) =$ negativo e $c(x) =$ negativo
FP: oggetti x con $\gamma(x) =$ negativo e $c(x) =$ positivo

Si osserva che sommando i valori sulla prima colonna si ha la cardinalità degli elementi realmente positivi, mentre sommando quelli sulla seconda si ottiene la cardinalità degli elementi realmente negativi. Sommando queste due cardinalità si ottiene la cardinalità del campione.

[To be done]

4.5.1 Classificatori costanti

4.5.2 Classificatori ideali

4.5.3 Classificatori casuali

4.5.4 Classificatori a soglia

Parte II

Teoria della probabilità

5 Calcolo combinatorio

Principio fondamentale del calcolo combinatorio Se ci sono s_1 modi per operare una scelta e, per ciascuno di essi, ci sono s_2 modi per operare una seconda scelta e, per ciascuno di essi, ci sono s_3 modi per operare una terza scelta e così via fino a s_t modi per operare la t-esima scelta, allora il numero delle sequenze di possibili scelte è

$$s_1 \cdot s_2 \cdots s_t = \prod_{i=1}^{t} s_i$$

Si osserva che questo risultato corrisponde a calcolare il numero delle foglie di un albero di profondità t il cui primo livello ha s_1 nodi, ciascuno dei quali ha s_2 figli, ciascuno dei quali ha s_3 figli e così via.

5.1 Permutazioni

Si consideri un insieme di n oggetti $A = \{a_1, \cdots, a_n\}$. Una permutazione di questi n oggetti è una sequenza ordinata in cui compaiono tutti e soli gli elementi dell'insieme A.

5.1.1 Permutazioni semplici

Se gli n oggetti di A sono tutti distinguibili, allora si parla di permutazione semplice.

Il numero totale di permutazioni semplici si ottiene applicando il principio fondamentale del calcolo combinatorio: il primo elemento della configurazione può essere scelto in n modi, il secondo in $(n-1)$, il terzo in $(n-2)$ e così via, fino all'ultimo che può essere scelto in un solo modo, essendo rimasto un solo oggetto disponibile.

Indicando con p_n il numero delle possibili permutazioni di un insieme di n elementi distinguibili, si ottiene:

$$p_n = n \cdot (n-1) \cdot (n-2) \cdots 1 = n!$$

Come casi particolari, si ottiene $p_1 = 1! = 1$, mentre, per convenzione, si pone $p_0 = 0! = 1$.

5.1.2 Permutazioni con ripetizioni

Se gli n oggetti di A non sono tutti distinguibili, ma sono suddivisi in r gruppi di oggetti indistinguibili tra loro, con numerosità rispettive $k_1, k_2, \cdots, k_r$, allora una sequenza ordinata che includa tutti gli oggetti è detta *permutazione con ripetizioni*. Poiché ogni oggetto appartiene a un solo gruppo, si ha $\sum_{i=1}^{r} k_i = n$, da cui segue $r \leq n$.

Indicando con P_n il numero delle possibili permutazioni di un insieme di n elementi non tutti distinguibili, si ottiene:

$$P_n^{k_1,k_2,\cdots,k_r} = \frac{n!}{k_1!\, k_2! \cdots k_r!} = \binom{n}{k_1, k_2, \cdots, k_r}$$

Questa formula si ottiene dividendo il numero totale di permutazioni di n oggetti distinti per il numero delle permutazioni indistinguibili, ovvero quelle interne ai singoli gruppi di oggetti uguali, che non alterano la configurazione complessiva. La quantità ottenuta è detta *coefficiente multinomiale*.

Si osserva che $p_n = P_n^{k_1,\cdots,k_r} \cdot k_1! \cdots k_r!$ e che, nel caso in cui tutti i gruppi abbiano numerosità unitaria, ossia $k_i = 1$ per ogni i, si ottiene la formula delle permutazioni semplici.

5.2 Disposizioni

Si consideri un insieme di n oggetti $A = \{a_1, \ldots, a_n\}$. Una disposizione di k oggetti tratti dall'insieme A è una sequenza ordinata di k elementi in cui l'ordine è rilevante e gli oggetti possono essere scelti con o senza ripetizione, a seconda del contesto.

5.2.1 Disposizioni semplici

Se gli n oggetti di A sono tutti distinguibili e non sono ammesse ripetizioni, allora si parla di disposizione semplice. Ne segue che $k \leq n$.

Il numero totale di disposizioni semplici si ottiene applicando il principio fondamentale del calcolo combinatorio: il primo elemento della configurazione può essere scelto in n modi, il secondo in $(n-1)$, il terzo in $(n-2)$ e così via, fino al k-esimo che può essere scelto in $(n-k+1)$ modi diversi.

Indicando con $d_{n,k}$ il numero delle possibili disposizioni semplici di k elementi tra n oggetti distinti, si ottiene:

$$d_{n,k} = n\,(n-1)\ldots(n-k+1) = n\,(n-1)\ldots(n-k+1)\,\frac{(n-k)!}{(n-k)!} = \frac{n!}{(n-k)!}$$

Le permutazioni semplici possono essere interpretate come un caso particolare delle disposizioni semplici, in cui il numero di elementi scelti coincide con il numero totale disponibile, ossia quando $k = n$.

5.2.2 Disposizioni con ripetizioni

Se gli n oggetti dell'insieme A sono tutti distinguibili e ogni elemento può comparire più volte nella sequenza, si parla di disposizione con ripetizione.

In questo caso, ogni posizione della sequenza può essere occupata da uno qualunque degli n elementi, senza alcuna restrizione, e quindi ognuna delle k posizioni può essere riempita in n modi indipendenti dagli altri.

Indicando con $D_{n,k}$ il numero delle disposizioni con ripetizione di k elementi tratti da un insieme di n oggetti distinti, si ottiene:

$$D_{n,k} = \underbrace{n \cdot n \ldots n}_{k \text{ volte}} = n^k$$

Tale formula vale per ogni intero $k \geq 0$, indipendentemente dalla cardinalità dell'insieme di partenza.

Quando $k = 1$, si ottiene $D_{n,1} = n$; quando $k = 0$, si pone per convenzione $D_{n,0} = 1$, in quanto esiste un'unica sequenza vuota di lunghezza zero.

5.3 Combinazioni

Si consideri un insieme di n oggetti $A = \{a_1, \ldots, a_n\}$. Una combinazione di k oggetti tratti dall'insieme A è un insieme di k elementi in cui l'ordine non è rilevante e gli oggetti possono essere scelti con o senza ripetizione, a seconda del contesto.

5.3.1 Combinazioni semplici

Una combinazione semplice di k oggetti tratti da un insieme A di n oggetti distinguibili è definita come un sottoinsieme di k elementi di A, in cui l'ordine non è rilevante e non è consentito ripetere lo stesso oggetto più volte. Ne consegue che k debba soddisfare la condizione $0 \leq k \leq n$.

Per determinarne il numero, si consideri il numero di disposizioni semplici di k elementi su n, vale a dire tutte le possibili sequenze ordinate di k oggetti distinti scelti da A. Ogni sottoinsieme di k elementi può essere riordinato in $k!$ modi diversi, ossia in un numero pari a quello delle permutazioni dei suoi k oggetti. Per convertire quindi il conteggio delle sequenze ordinate in quello dei sottoinsiemi, in cui l'ordine è irrilevante, è necessario dividere $d_{n,k}$ per $k!$.

Indicando con $c_{n,k}$ il numero delle combinazioni semplici di k elementi tratti da un insieme di n oggetti distinti, si ottiene:

$$c_{n,k} = \frac{d_{n,k}}{p_k} = \frac{n!}{(n-k)!}\frac{1}{k!} = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}$$

La quantità ottenuta è detta *coefficiente binomiale* n su k ed esprime il numero di tutti i possibili sottoinsiemi di cardinalità k che si possono formare a partire da n oggetti distinti. Si osservi come per definizione $c_{n,k} < d_{n,k}$.

5.3.2 Combinazioni con ripetizione

Se ogni elemento di A può comparire più volte nella combinazione, ignorando comunque l'ordine, si parla di combinazione con ripetizione. In tal caso si possono scegliere k elementi (con possibile duplicazione) tra gli n oggetti di A, senza considerare l'ordine in cui vengono selezionati.

Indicando con $C_{n,k}$ il numero delle combinazioni con ripetizione di k elementi tratti da un insieme di n oggetti distinti, si ottiene:

$$C_{n,k} = c_{n+k-1,k} = \binom{n+k-1}{k}$$

Dimostrazione

Sia $A = \{a_1, a_2, \ldots, a_n\}$ un insieme di n oggetti distinti. Vogliamo contare il numero di combinazioni con ripetizione di k elementi da A. Poiché in questo contesto l'ordine non è rilevante e le ripetizioni sono permesse, possiamo associare ogni combinazione (ossia, ogni multinsieme) a una sequenza non decrescente di indici. In particolare, consideriamo una sequenza

$$m_1, m_2, \ldots, m_k \quad \text{con} \quad 1 \le m_1 \le m_2 \le \cdots \le m_k \le n$$

Qui, i numeri m_i non rappresentano direttamente gli oggetti di A, ma sono gli indici che li identificano: l'indice m_i corrisponde all'oggetto a_{m_i} in A. In questo modo, ogni scelta di k elementi da A, con ripetizione, è associata a una sequenza non decrescente di indici.

Per facilitare il conteggio, trasformiamo questa sequenza non decrescente in una sequenza strettamente crescente mediante la trasformazione

$$n_i = m_i + (i-1) \quad \text{per } i = 1, 2, \ldots, k$$

Dal momento che $m_i \le m_{i+1}$, si ha

$$n_i = m_i + (i-1) < m_{i+1} + i = n_{i+1}$$

che garantisce che la nuova sequenza $n_1, n_2, \ldots, n_k$ sia strettamente crescente.

Osserviamo inoltre che il primo elemento soddisfa $n_1 = m_1 \ge 1$, mentre l'ultimo elemento è

$$n_k = m_k + (k-1) \le n + (k-1)$$

quindi ogni n_i appartiene all'insieme $\{1, 2, \ldots, n+k-1\}$.

La trasformazione appena definita stabilisce una corrispondenza biunivoca tra le sequenze non decrescenti di indici (che rappresentano le combinazioni con ripetizione di k elementi da A) e le sequenze strettamente crescenti di k numeri presi da $\{1, 2, \ldots, n+k-1\}$. Queste ultime sono esattamente le combinazioni semplici di k elementi da un insieme di $n+k-1$ elementi, il cui numero è dato da

$$\binom{n+k-1}{k}$$

Pertanto, il numero di combinazioni con ripetizione $C_{n,k}$ è proprio quel valore.

Questa dimostrazione evidenzia come il problema delle combinazioni con ripetizione possa essere ridotto a quello delle combinazioni semplici, tramite una trasformazione che converte una sequenza non decrescente di indici in una sequenza strettamente crescente.

Coefficienti combinatori

I coefficienti combinatori misurano il numero di modi in cui si possono selezionare o distribuire gli elementi.

Coefficiente binomiale Rappresenta il numero di modi in cui si possono scegliere k elementi da un insieme di n elementi, senza considerare l'ordine. Il suo valore è dato da:

$$\binom{n}{k} = \binom{n}{k,\, n-k} = \frac{n!}{k!\,(n-k)!}$$

Coefficiente multinomiale Generalizza il concetto del coefficiente binomiale e indica il numero di modi per suddividere n elementi in r gruppi distinti di dimensioni $k_1, \ldots, k_r$ dove $k_1 + \cdots + k_r = n$. Il suo valore è dato da:

$$\binom{n}{k_1, k_2, \cdots, k_r} = \frac{n!}{k_1!\,k_2!\,\cdots\,k_r!}$$

Si osserva che il coefficiente binomiale è il caso particolare del coefficiente multinomiale quando si divide l'insieme in due gruppi: uno di grandezza k e l'altro di grandezza $n-k$. La somma delle cardinalità dei due gruppi distinti continua a essere n, infatti $k + (n-k) = n$.

6 Introduzione alla Probabilità

Il concetto di probabilità di un evento, quando si effettua un esperimento, è passabile di diverse interpretazioni filosofiche:

1. Interpretazione frequentista: la probabilità di un evento viene intesa come il limite del rapporto tra il numero di volte in cui l'evento si verifica e il numero totale di prove, quando queste sono ripetute indefinitamente.
2. Interpretazione soggettivistica: la probabilità non è vista come una proprietà oggettiva dell'esito, ma come una misura del livello di fiducia che lo studioso ripone nel verificarsi dell'evento.

Indipendentemente dall'approccio che si favorisce, utilizzando un approccio matematico ed i suoi strumenti, come per esempio la notazione insiemistica, è possibile formalizzare le regole e gli assiomi della teoria della probabilità.

6.1 Definizioni

Prima di enunciare gli assiomi della teoria della probabilità, occorre introdurre alcuni concetti fondamentali relativi agli esperimenti e ai loro esiti.

6.1.1 Spazio degli esiti

Esperimento Un esperimento è un procedimento o una prova condotta in condizioni controllate, il cui risultato è incerto.

Esito L'esito è un possibile risultato ottenuto da un esperimento. Si indica con ω e appartiene allo spazio degli esiti: $\omega \in \Omega$.

Spazio degli esiti Lo spazio degli esiti (o insieme universo o spazio campionario) è l'insieme dei possibili esiti dell'esperimento, e si indica con Ω. L'universo può essere:

- finito o infinito, a seconda del numero di esiti possibili
- discreto se gli esiti sono isolati e contabili, o continuo se gli esiti formano un continuum. In questo contesto, la distinzione tra spazi discreti e continui riguarda la struttura complessiva di Ω, e non le proprietà intrinseche degli elementi stessi

6.1.2 Evento

Un evento E è un sottoinsieme dello spazio degli esiti, perciò $E \subseteq \Omega$. Un evento formato da un solo esito $\{\omega\}$ è detto evento elementare. Per definizione, Ω rappresenta l'evento certo mentre $\varnothing$ è l'evento impossibile.

Operazioni

Dati due eventi $E, F \subseteq \Omega$, è possibile applicare le operazioni fondamentali degli insiemi:

- Unione $E \cup F$: è l'evento che si verifica quando si verifica almeno uno tra E e F.
  Per un esito x si ottiene che $x \in E \cup F \Leftrightarrow x \in E \vee x \in F$
- Intersezione $E \cap F$: è l'evento che si verifca quando si verificano entrambi gli eventi E e F.
  Per un esito x si ottiene che $x \in E \cap F \Leftrightarrow x \in E \wedge x \in F$
  Se $E \cap F = \varnothing$ allora E e F si dicono *mutualmente esclusivi* o *eventi disgiunti*.
- Complementare E^C di E: è l'evento che si verifica quando non si verifica E . Si indica anche con $\overline{E}$.
  Per un esito x si ottiene che $x \in E^C \Leftrightarrow x \notin E$
  Si osserva che $E^C = \Omega - E$. Vale anche la relazione $\Omega^C = \varnothing$.
- Differenza $E \setminus F$: è l'evento che si verifica quando E si verifica ma non F. Per un esito x si ottiene che $x \in E \setminus F \Leftrightarrow x \in E \wedge x \notin F$
  Si osserva che questa operazione non è simmetrica, infatti $E \setminus F \neq F \setminus E$.

È possibile definire l'unione o l'intersezione di più eventi. Si considerino gli eventi $E_1, E_2, \ldots, E_n$:

- la loro unione $\bigcup_{i=1}^{n} E_i = E_1 \cup \cdots \cup E_n$ è l'evento formato da tutti gli esiti che appartengono ad almeno uno degli eventi E_i
- la loro intersezione $\bigcap_{i=1}^{n} E_i = E_1 \cap \cdots \cap E_n$ è l'evento formato da tutti gli esiti che appartengono a tutti gli eventi E_i

In altre parole, l'unione degli E_i si verifica se almeno uno degli eventi E_i si verifica, mentre l'intersezione degli E_i si verifica solo se tutti gli E_i si verificano.

Inclusione È inoltre possibile definire delle relazioni di inclusione e uguaglianza tra eventi. Siano $E, F \subseteq \Omega$ due eventi, si dice che l'evento E è contenuto in F, e si scrive $E \subseteq F$, se tutti gli esiti di E appartengono anche a F. Formalmente, si può indicare questa relazione come $E \subseteq F \Leftrightarrow \forall \omega \in E : \omega \in F$. Questo significa che se si verifica E, allora si verifica anche F. Si osserva che se $E \subseteq F \wedge F \subseteq E \Rightarrow E = F$.

Proprietà

Per l'unione e l'intersezione valgono le seguenti proprietà (verranno presentate solo sull'unione):

- commutatività: $E \cup F = F \cup E$
- associatività: $E \cup F \cup G = (E \cup F) \cup G = E \cup (F \cup G)$
- distributività: $E \cup (F \cap G) = (E \cup F) \cap (E \cup G)$
- leggi di assorbimento: $E \cup (E \cap F) = E$ e $E \cap (E \cup F) = E$
- leggi di De Morgan: $\overline{E \cup F} = \overline{E} \cap \overline{F}$ e $\overline{E \cap F} = \overline{E} \cup \overline{F}$

  Dimostrazione

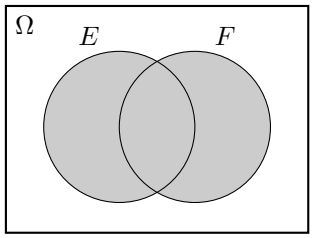
  1. $x \in \overline{E \cup F} \Rightarrow x \notin E \cup F \Rightarrow x \notin E \wedge x \notin F \Rightarrow x \in \overline{E} \wedge x \in \overline{F} \Rightarrow x \in \overline{E} \cap \overline{F} \Rightarrow \overline{E \cup F} \subseteq \overline{E} \cap \overline{F}$
  2. $x \in \overline{E} \cap \overline{F} \Rightarrow x \in \overline{E} \wedge x \in \overline{F} \Rightarrow x \notin E \wedge x \notin F \Rightarrow x \notin E \cup F \Rightarrow x \in \overline{E \cup F} \Rightarrow \overline{E} \cap \overline{F} \subseteq \overline{E \cup F}$

  Da entrambe le inclusioni si ottiene che $\overline{E \cup F} = \overline{E} \cap \overline{F}$.

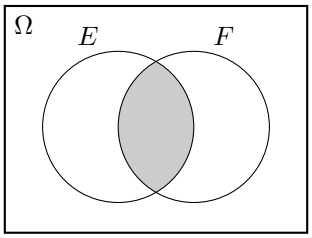
Un modo rigoroso per dimostrare queste proprietà consiste nel verificare che ogni esito appartenente all'evento al primo membro è anche contenuto nell'evento al secondo membro, e viceversa, proprio come si è fatto pocanzi tramite la dimostrazione della legge di De Morgan.

Diagrammi di Venn Un tipo di rappresentazione grafica degli eventi, utile per illustrare le relazioni logiche che li legano, sono i *diagrammi di Venn*. Lo spazio degli esiti Ω è rappresentano da un rettangolo che contiene il resto della figura. Gli eventi da prendere in considerazione sono invece rappresentati da cerchi disegnati all'interno del rettangolo. A questo punto si evidenziano gli eventi complessi rilevanti.

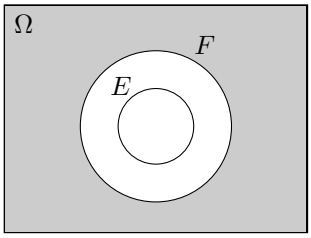
Si illustrano le operazioni di unione, intersezione, complemento e inclusione tramite i diagrammi di Venn:



$E \cup F$



$E \cap F$



$\overline{F}$, $E \subseteq F$

6.1.3 Algebra degli eventi

Un'algebra degli eventi $\mathcal{A}$ su Ω è una collezione di sottoinsiemi di Ω, ossia $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, tale che:

1. $\Omega \in \mathcal{A}$: l'evento certo fa parte dell'algebra
2. $\forall E \in \mathcal{A} \Rightarrow \overline{E} \in \mathcal{A}$: chiusura rispetto al complementare
3. $\forall E, F \in \mathcal{A} \Rightarrow E \cup F \in \mathcal{A}$: chiusura rispetto all'unione di due eventi

Da queste proprietà discendono varie conseguenze:

- $\varnothing \in \mathcal{A}$ perché $\varnothing = \overline{\Omega}$
- Per induzione, $\mathcal{A}$ è chiusa per l'unione finita, infatti $\forall E_1, E_2, \ldots, E_n \in \mathcal{A} \quad \bigcup_{i=1}^{n} E_i \in \mathcal{A}$.
- $\mathcal{A}$ è chiusa rispetto all'intersezione di due eventi: $A \cap B = \overline{\overline{A} \cup \overline{B}}$, e per induzione anche rispetto all'intersezione finita: $\forall E_1, E_2, \ldots, E_n \in \mathcal{A} \quad \bigcap_{i=1}^{n} E_i \in \mathcal{A}$
- $\mathcal{A}$ è chiusa rispetto alla differenza finita: $E \setminus F = E \cap \overline{F}$

Se Ω è finito, l'algebra più grande che si possa considerare è $\mathcal{P}(\Omega)$, l'insieme di tutte le parti di Ω.

σ-algebra

Sia $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ un'algebra su Ω. Se per ogni famiglia numerabile di insiemi $\{E_i\}_{i \in \mathbb{N}} \subseteq \mathcal{A}$ vale

$$\bigcup_{i=1}^{\infty} E_i \in \mathcal{A}$$

allora $\mathcal{A}$ si dice σ-algebra e si indica con $\mathcal{F}$. Da ciò discende, per De Morgan, anche la chiusura rispetto alle intersezioni numerabili. Gli elementi di una σ-algebra si dicono insiemi *misurabili*.

Insieme numerabile Un insieme è detto numerabile se i suoi elementi sono in numero finito oppure se possono essere messi in corrispondenza biunivoca con $\mathbb{N}$. Se un insieme numerabile ha un numero infinito di elementi, viene detto infinito numerabile, e dato che può essere messo in corrispondenza biunivoca con i numeri naturali, si può dire che un insieme è infinito numerabile se ha la cardinalità di $\mathbb{N}$.

Se Ω è finito, l'insieme di tutte le parti ha cardinalità $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$ ed è quindi finito anch'esso. In questo caso, ogni algebra $\mathcal{A} \subseteq \mathcal{P}(\Omega)$, che è chiusa per unioni finite, è automaticamente una σ-algebra, poiché ogni unione numerabile coincide con una unione finita. $\mathcal{P}(\Omega)$ rappresenta la σ-algebra più grande possibile.
Nel caso in cui Ω sia infinito, $\mathcal{P}(\Omega)$ è certamente una σ-algebra, ma in genere non è quella che si usa in contesti di misura, perché spesso si considerano σ-algebre proprie (strettamente più piccole).

Spazio misurabile Una volta fissata una σ-algebra $\mathcal{F}$ su Ω, la coppia $(\Omega, \mathcal{F})$ si chiama *spazio misurabile*. Qui $\mathcal{F}$ individua i sottoinsiemi di Ω considerati misurabili, ossia quelli ai quali sarà in seguito possibile associare una misura in modo coerente. Lo spazio misurabile è dunque la struttura formata dallo spazio degli esiti Ω e dalla famiglia $\mathcal{F}$ di sottoinsiemi ammessi.

Isomorfismo tra algebre

Due spazi misurabili $(\Omega_1, \mathcal{F}_1)$ e $(\Omega_2, \mathcal{F}_2)$ si dicono isomorfi se esiste una funzione biunivoca $\phi : \mathcal{F}_1 \to \mathcal{F}_2$ che preserva le operazioni fondamentali, cioè:

- Per ogni $E \in \mathcal{F}_1$, vale $\phi(\overline{E}) = \overline{\phi(E)}$
- Per ogni coppia $E, F \in \mathcal{F}_1$, vale $\phi(E \cup F) = \phi(E) \cup \phi(F)$

La mappa ϕ è un isomorfismo di algebre booleane, il che implica che le due strutture hanno la stessa struttura misurabile, pur essendo definite su spazi degli esiti differenti. Questo significa che, per ogni proprietà, operazione o misura che possiamo definire su una delle algebre, c'è una corrispondenza diretta nell'altra.

Esempio Si consideri il lancio di una moneta, per il quale la σ-algebra è $\mathcal{F}_M = \{\varnothing, \{T\}, \{C\}, \Omega_M\}$, dove T sta per testa, C per croce e $\Omega_M = \{T, C\}$.

Per il lancio di un dado, supponiamo di considerare solo due eventi, ottenuti partizionando lo spazio degli esiti $\Omega_D = \{1, 2, 3, 4, 5, 6\}$ in $\mathcal{F}_D = \{\varnothing, \{1, 2\}, \{3, 4, 5, 6\}, \Omega_D\}$. $\mathcal{F}_D$ è un'algebra ammissibile diversa dall'insieme delle parti $\mathcal{P}(\Omega_D)$.

Si definisce la mappa $\phi : \mathcal{F}_D \to \mathcal{F}_M$ mediante:

- $\phi(\varnothing) = \varnothing$
- $\phi(\{1, 2\}) = \{T\}$
- $\phi(\{3, 4, 5, 6\}) = \{C\}$
- $\phi(\Omega_D) = \Omega_M$

È facile verificare che ϕ preserva il complementare e le unioni, quindi $\mathcal{F}_D$ e $\mathcal{F}_M$ sono isomorfe. In questo modo, funzionalmente, il lancio del dado (con questa specifica scelta di σ-algebra) si comporta come il lancio della moneta, pur essendo l'esperimento originariamente a sei esiti.

6.1.4 Assiomi di Kolmogorov

Sperimentando un esperimento ripetutamente, mantenendo costanti le condizioni, si osserva empiricamente che la frazione di casi in cui si realizza un evento E tende, al crescere del numero dei tentativi, a stabilizzarsi in un valore costante, che dipende unicamente da E. Questo valore costante è quello che intendiamo come probabilità dell'evento E.

Si consideri lo spazio misurabile $(\Omega, \mathcal{F})$. Definiamo la *misura di probabilità* come una funzione σ-additiva $\mathbb{P} : \mathcal{F} \to [0, 1]$ che assegna a ciascun evento $E \in \mathcal{F}$ il numero $\mathbb{P}(E)$, ossia la probabilità che E si verifichi.

La funzione $\mathbb{P}$ deve soddisfare i seguenti assiomi (assiomi di Kolmogorov):

1. Non negatività:
   $\forall E \in \mathcal{F} \quad \mathbb{P}(E) \geq 0$
2. Normalizzazione:
   $\mathbb{P}(\Omega) = 1$ l'evento certo ha probabilità 1
3. Additività numerabile (o σ-additività):
   Se $\{E_i\}_{i \in \mathbb{N}} \subseteq \mathcal{F}$ è una famiglia di eventi disgiunti (cioè, $E_i \cap E_j = \varnothing$ per ogni $i \neq j$), allora
   $$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} E_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(E_i).$$

Questi assiomi formalizzano l'osservazione empirica: la probabilità, definita come la frequenza relativa limite, viene interpretata come una misura che assegna ad ogni evento misurabile un valore compreso tra 0 e 1, rispettando le proprietà di coerenza e additività.

Si osserva che per garantire una collezione di n eventi $\{E_1, E_2, \dots, E_n\}$ sia considerata disgiunta, è necessario e sufficiente provare che ogni coppia di eventi distinti E_i e E_j, con $i \neq j$, non abbia nessun elemento in comune, ossia che $E_i \cap E_j = \varnothing \quad \forall i \neq j$.

σ-additività Sia $\mathcal{A}$ un'algebra di insiemi. Una funzione $\mu : \mathcal{A} \to (-\infty, +\infty)$ è detta (finitamente) additiva se $\forall A, B \in \mathcal{A}$ disgiunti si ha $\mu(A \cup B) = \mu(A) + \mu(B)$. La funzione è detta numerabilmente additiva o σ-additiva se per ogni successione $A_1, \dots, A_n, \dots \in \mathcal{A}$ tra loro disgiunti e tali che la loro unione numerabile stia ancora in $\mathcal{A}$ si ha:

$$\mu\Big(\bigcup_{n=1}^{\infty} A_n\Big) = \sum_{n=1}^{\infty} \mu(A_n)$$

Ogni funzione σ-additiva è una funzione (finitamente) additiva, ma non vale il contrario

Proprietà

Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Sono allora vere le seguenti proprietà che verranno dimostrate.

Teorema 1 $\forall E \in \mathcal{F} \quad \mathbb{P}(\overline{E}) = 1 - \mathbb{P}(E)$

Dimostrazione:

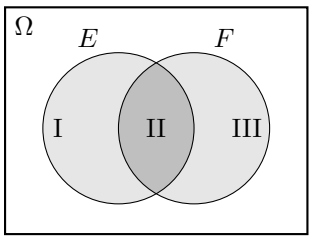
- $E \cup \overline{E} = \Omega \wedge E \cap \overline{E} = \varnothing \;\Rightarrow$ I due insiemi sono disgiunti
- Dato che i due insiemi sono disgiunti, è possibile applicare il terzo assioma:

$$1 \overset{K2}{=} \mathbb{P}(\Omega) = \mathbb{P}(E \cup \overline{E}) \overset{K3}{=} \mathbb{P}(E) + \mathbb{P}(\overline{E})$$

$$\mathbb{P}(E) + \mathbb{P}(\overline{E}) = 1 \;\Rightarrow\; \mathbb{P}(\overline{E}) = 1 - \mathbb{P}(E)$$

Teorema 2 $\forall E, F \in \mathcal{F} \quad \mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

Dimostrazione:



Suddivisione di $E \cup F$

- Si suddivide l'evento $E \cup F$ in tre eventi distinti:
  1. I $= E \cap \overline{F}$
  2. II $= E \cap F$
  3. III $= \overline{E} \cap F$
- Dal diagramma di Venn si osserva che questi tre eventi sono disgiunti a due a due. Si dimostra ora algebricamente che I e II sono disgiunti.

- I e II sono disgiunti $\Leftrightarrow$ $(\mathrm{I} \cup \mathrm{II} = E) \wedge (\mathrm{I} \cap \mathrm{II} = \varnothing)$

$$(E \cap \overline{F}) \cup (E \cap F) = E \cup (\overline{F} \cap F) = E \cup \varnothing = E$$

$$(E \cap \overline{F}) \cap (E \cap F) = (E \cap E) \cap (F \cap \overline{F}) = E \cap \varnothing = \varnothing$$

- Si è quindi dimostrato che I e II sono due eventi disgiunti. Dimostrandolo analogamente per le altre coppie, è possibile poi applicare il terzo assioma su questi tre eventi disgiunti:

$$\mathbb{P}(E \cup F) = \mathbb{P}(\mathrm{I} \cup \mathrm{II} \cup \mathrm{III}) \overset{K3}{=} \underbrace{\mathbb{P}(\mathrm{I}) + \mathbb{P}(\mathrm{II})}_{\mathbb{P}(E)} + \underbrace{\mathbb{P}(\mathrm{III}) + \mathbb{P}(\mathrm{II})}_{\mathbb{P}(F)} - \mathbb{P}(\mathrm{II}) = \mathbb{P}(E) + \mathbb{P}(F) + \mathbb{P}(E \cap F)$$

Teorema 3 $\mathbb{P}(\varnothing) = 0$

Dimostrazione:

- $\overline{\Omega} = \varnothing$
- $\mathbb{P}(\overline{\Omega}) \overset{T1}{=} 1 - \mathbb{P}(\Omega) \;\Rightarrow\; \mathbb{P}(\varnothing) = 1 - \mathbb{P}(\Omega) \overset{K2}{\Rightarrow} \mathbb{P}(\varnothing) = 1 - 1 = 0$

Teorema 4 $\forall E \in \mathcal{F} \quad \mathbb{P}(E) \leq 1$

Dimostrazione:

- $\mathbb{P}(\overline{E}) \overset{T1}{=} 1 - \mathbb{P}(E) \overset{K1}{\Rightarrow} \mathbb{P}(\overline{E}) = 1 - \mathbb{P}(E) \geq 0 \;\Rightarrow\; \mathbb{P}(E) \leq 1$

Teorema 5 $\forall E, F \in \mathcal{F} \mid E \subseteq F \quad \mathbb{P}(E) \leq \mathbb{P}(F)$

Dimostrazione:

- Dato che $E \subseteq F$, allora si può scrivere $F = E \cup (F \setminus E)$ con $E \cap (F \setminus E) = \varnothing$
- E e $F \setminus E$ sono quindi due eventi disgiunti ed è quindi possibile applicare il terzo assioma:

$$\mathbb{P}(F) \overset{K3}{=} \mathbb{P}(E) + \mathbb{P}(F \setminus E) \overset{K1}{\Rightarrow} \mathbb{P}(F \setminus E) \geq 0 \;\Rightarrow\; \mathbb{P}(F) \geq \mathbb{P}(E)$$

6.1.5 Spazio di probabilità

Se $\mathcal{F}$ è una σ-algebra definita sullo spazio degli esiti Ω e $\mathbb{P}$ è una misura di probabilità definita su $\mathcal{F}$ che soddisfa gli assiomi di Kolmogorov, allora la terna $(\Omega, \mathcal{F}, \mathbb{P})$ è detta spazio di probabilità.

Spazio equiprobabile

Se nello spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ lo spazio degli esiti Ω è finito e $\forall \omega \in \Omega$ si ha che $\mathbb{P}(\{\omega\})$ è costante, allora lo spazio di dice equiprobabile.

Essendo Ω finito, lo si può considerare come $\{\omega_1, \omega_2, \ldots, \omega_N\}$, e di conseguenza la sua cardinalità è $|\Omega| = N$. L'equiprobabilità degli esiti si scrive come $\mathbb{P}(\{\omega_1\}) = \mathbb{P}(\{\omega_2\}) = \cdots = \mathbb{P}(\{\omega_N\}) = p$.

Dagli assiomi 1 e 3 segue che

$$1 \overset{K1}{=} \mathbb{P}(\Omega) = \mathbb{P}(\{\omega_1\} \cup \cdots \cup \{\omega_N\}) \overset{K3}{=} \mathbb{P}(\{\omega_1\}) + \cdots + \mathbb{P}(\{\omega_N\}) = Np$$

da cui si deduce che

$$\mathbb{P}(\{\omega_i\}) = \frac{|\{\omega_i\}|}{|\Omega|} = p = \frac{1}{N} \quad \forall i \in \{1, \ldots, N\}$$

Si consideri un evento $E \subseteq \Omega$, ed essendo E finito sia la sua cardinalità $|E| = k$. Inidichiamo genericamente gli elementi di E con $\{e'_1, \ldots, e'_k\}$ per sottolineare che sono k elementi arbitrari di Ω, e non necessariamente i primi k. Riapplicando gli assiomi si ha:

$$\mathbb{P}(E) = \mathbb{P}(\{e'_1, \ldots, e'_k\}) = \mathbb{P}(\{e'_1\} \cup \cdots \cup \{e'_k\}) \overset{K3}{=} \sum_{i=1}^{k} \mathbb{P}(\{e'_i\}) = \sum_{i=1}^{k} p = pk = \frac{k}{N} = \frac{|E|}{|\Omega|}$$

Si definisce $\mathbb{P}(E) = \dfrac{|E|}{|\Omega|} = \dfrac{\#\text{ casi favorevoli}}{\#\text{ casi possibili}}$ come la regola classica per il calcolo delle probabilità.

Se Ω è infinito non è possibile definire una probabilità equiprobabile nel senso in cui ogni esito riceve la stessa probabilità positiva p. Infatti se $|\Omega| = \infty$ allora $p \to 0$, ma se $\forall \omega \in \Omega$ si ha che $\mathbb{P}(\{\omega\}) = 0$, allora gli assiomi di Kolmogorov non sono più soddisfatti e si giunge ad un assurdo. Questo vale per spazi discreti.

6.2 Probabilità condizionata

Si definisce *probabilità condizionata* la probabilità che si verifichi un evento E sapendo che si è già verificato un altro evento F. La probabilità condizionata di E dato F si indica con $\mathbb{P}(E|F)$, oppure $\mathbb{P}_F(E)$, e si può definire a patto che la probabilità di F sia diversa da zero.

La probabilità condizionata subentra tutte le volte che si vuole calcolare la probabilità di un evento E, detto *evento condizionato*, assumendo che si sia già verificato un altro evento F, detto *evento condizionante*. L'incertezza dell'evento E è quindi solo parziale ed è limitata al sottoinsieme degli esiti in cui F si verifica.

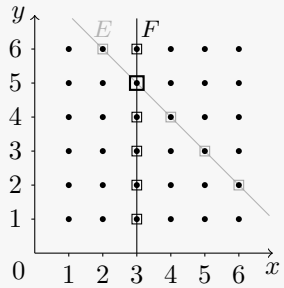
Utilizzando la definizione classica di probabilità, vale la seguente formula sulla probabilità condizionata:

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} \quad \text{con } \mathbb{P}(F) \neq 0$$

Infatti, se si è verificato l'evento F, affinché si verifichi anche E il caso deve aver favorito un elemento che sta sia in E che in F, ovvero che appartiene all'intersezione $E \cap F$. In secondo luogo, il verificarsi di F restringe lo spazio degli esiti ai soli elementi di F, escludendo quelli che non vi appartengono. L'evento condizionante diventa quindi il nuovo spazio degli esiti, sostituendo Ω.

Nel caso in cui $F = \varnothing$, e quindi $\mathbb{P}(F) = 0$, non è possibile calcolare $P(E|F)$ che è perciò detta indefinita.

Esempio Si immagini di tirare due dadi. Lo spazio degli esiti di questo esperimento è descritto da $\Omega = \{(x, y) \mid x, y \in \{1, \dots, 6\}\}$ dove si intende che si ottiene l'esito (x, y) se il risultato del primo dado è x e quello del secondo y. Si supponga che entrambi i dadi non siano truccati, e di trovarci quindi in uno spazio equiprobabile dove $\mathbb{P}((x, y)) = 1/|\Omega| = 1/36$.



Sia $E = \{(x, y) \in \Omega \mid x + y = 8\}$ l'evento che si verifica quando la somma dei due dadi lanciati vale 8. Graficamente, queste coppie (x, y) stanno sulla retta $x + y = 8$ nel diagramma: si hanno quindi 5 possibili coppie valide.

Se si calcola la probabilità di questo evento, si ottiene:

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{5}{36}$$

Si supponga ora che il primo dado sia risultato in un 3, si vuole ancora calcolare la probabilità che E si verifichi. Possedendo questa informazione, si definisce $F = \{(x, y) \in \Omega \mid x = 3\}$ come l'evento condizionante. Graficamente, si osserva che l'evento F contiene esattamente 6 esiti; di conseguenza $\mathbb{P}(F) = |F|/|\Omega| = 1/6$.

Calcolando $\mathbb{P}(E|F)$ si ottiene la probabilità di E sapendo che F si è verificato. Per definizione:

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \frac{1/36}{1/6} = \frac{1}{6}$$

$E \cap F$ è infatti l'insieme degli esiti che soddisfano sia $x + y = 8$ che $x = 3$. Graficamente, si osserva che le due rette si incontrano in un unico punto, e di conseguenza $|E \cap F| = 1$.

Nel diagramma, l'evento F è rappresentato dalla retta verticale $x = 3$ mentre E è rappresentato dalla retta obliqua $x + y = 8$. Una volta saputo che il primo dado risulta in un 3, rimangono solo 6 possibili esiti, ossia quelli della retta verticale: lo spazio degli esiti è quindi ridotto da Ω a F. Tra questi punti, solo uno realizza la somma 8, ossia quella all'incrocio delle due rette.

Si noti come $\mathbb{P}(E)$ differisca da $\mathbb{P}(E|F)$: l'informazione sul primo dado modifica la probabilità che la somma dei due dadi sia 8, mostrando il ruolo decisivo della probabilità condizionata.

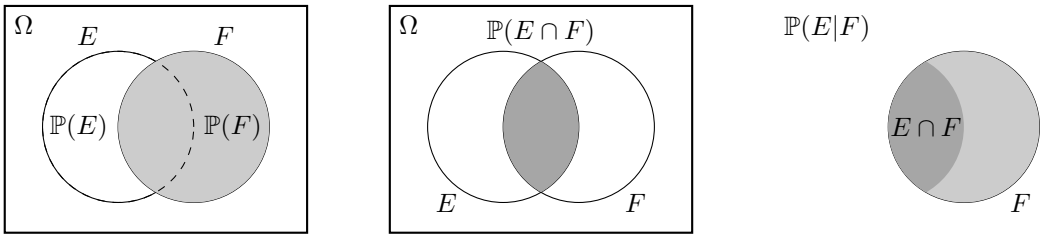
Si osserva che la definizione di probabilità condizionata è compatibile con l'interpretazione frequentista della probabilità degli eventi. Quest'ultima considera la probabilità come il limite del rapporto tra il numero di volte in cui si verifica un evento e il totale delle prove, al crescere indefinito di queste ultime. Pertanto, la probabilità condizionata rappresenta la frequenza relativa con cui E si verifica tra le prove in cui F è accaduto, rendendo la definizione coerente con l'approccio empirico.

Definizione rigorosa Dato uno spazio misurabile $(\Omega, \mathcal{F})$ di misura $\mathbb{P}$, ogni evento F eredita una struttura di spazio misurato $(F, \mathcal{A}_F, \mathbb{P})$, restringendo gli insiemi misurabili a quelli contenuti in F, ed induce una nuova misura $\mathbb{P}'_F(E) = \mathbb{P}(E \cap F)$ su $(\Omega, \mathcal{F})$, con $\mathbb{P}'_F(\Omega) = P(F)$.
Se $(\Omega, \mathcal{F}, \mathbb{P})$ è uno spazio di probabilità (valgono quindi gli assiomi di Kolmogorov, tra cui $\mathbb{P}(\Omega) = 1$) e F non è trascurabile (ossia $\mathbb{P}(F) \neq 0$), allora riscalando $\mathbb{P}'_F$ a $\mathbb{P}_F = \frac{1}{\mathbb{P}(F)} \mathbb{P}'_F$ si ottiene lo spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P}_F)$ condizionato dall'evento F.

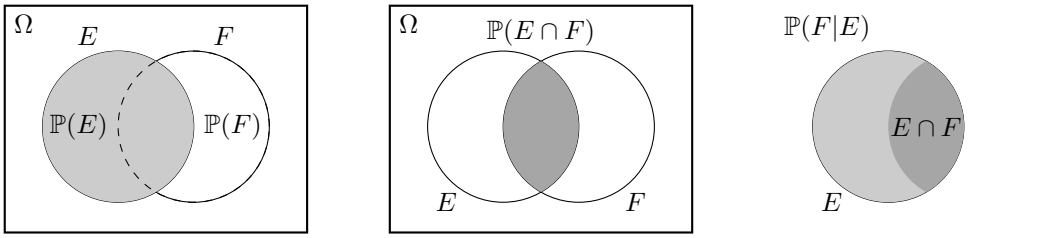
6.2.1 Teorema delle probabilità totali

Regola di fattorizzazione

Siano E e F due eventi in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. Se $\mathbb{P}(F) \neq 0$, moltiplicando entrambi i membri della formula della probabilità condizionata di E dato F per $\mathbb{P}(F)$ si ottiene $\mathbb{P}(E \cap F) = \mathbb{P}(E|F)\,\mathbb{P}(F)$



Allo stresso modo, se $\mathbb{P}(E) \neq 0$, moltiplicando entrambi i membri della formula della probabilità condizionata di F dato E per $\mathbb{P}(E)$ si ottiene $\mathbb{P}(F \cap E) = \mathbb{P}(F|E)\,\mathbb{P}(E)$



Essendo però $\mathbb{P}(F \cap E) = \mathbb{P}(E \cap F)$, si può concludere che:

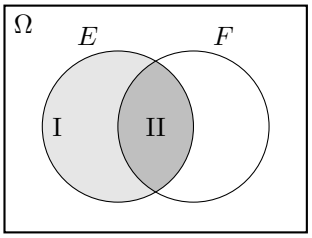
$$\mathbb{P}(E \cap F) = \mathbb{P}(E|F)\,\mathbb{P}(F) = \mathbb{P}(F|E)\,\mathbb{P}(E) \tag{6.2.1}$$

Questa formula è detta *regola di fattorizzazione* e discende in maniera diretta dalla definizione di probabilità condizionata. Essa afferma che la probabilità dell'evento $E \cap F$ può essere vista sotto due prospettive equivalenti, a seconda dell'evento che decidiamo di considerare come condizionante.

Questa reciprocità nasce dal fatto che gli eventi $E \cap F$ e $F \cap E$ sono identici in quanto la loro intersezione è commutativa. Quindi $\mathbb{P}(E|F)$ e $\mathbb{P}(F|E)$ sono due modi diversi di esplorare lo stesso evento $E \cap F$, solo che prendono come spazio degli esiti di riferimento due eventi diversi, rispettivamente F ed E.

La regola di fattorizzazione ci permette di spezzare la probabilità di un evento E in parti più semplici, legate a condizioni note. Partendo dalla considerazione che qualsiasi evento può essere suddiviso rispetto a un altro o più eventi che lo partizionano, si ottiene la *formula delle probabilità totali*.

Formula binaria delle probabilità totali Siano E e F due eventi in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. Poiché F e il complementare $\overline{F}$ costituiscono una partizione di Ω, si può suddividere E in due parti disgiunte:



Suddivisione di E

$$E = \mathrm{I} \cup \mathrm{II} = (E \cap \overline{F}) \cup (E \cap F)$$

Poiché gli insiemi $E \cap F$ e $E \cap \overline{F}$ sono disgiunti, è possibile applicare il terzo assioma di Kolmogorov:

$$\mathbb{P}(E) = \mathbb{P}(E \cap \overline{F}) + \mathbb{P}(E \cap F)$$

Applicando la regola di fattorizzazione si ottiene:

$$\mathbb{P}(E) = \mathbb{P}(E|F)\,\mathbb{P}(F) + \mathbb{P}(E|\overline{F})\,\mathbb{P}(\overline{F})$$

Si ottiene quindi una versiona binaria del teorema delle probabilità totali, limitata alla partizione $\{F, \overline{F}\}$:

$$\mathbb{P}(E) = \mathbb{P}(E|F)\,\mathbb{P}(F) + \mathbb{P}(E|\overline{F})\,\mathbb{P}(\overline{F}) \tag{6.2.2}$$

Si osserva che $\mathbb{P}(\overline{F}) = 1 - \mathbb{P}(F)$, di conseguenza andando a sostituire sopra si ha:

$$\mathbb{P}(E) = \mathbb{P}(E|F)\,\mathbb{P}(F) + \mathbb{P}(E|\overline{F})\,(1 - \mathbb{P}(F))$$

Formula estesa delle probabilità totali Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Si consideri una *partizione* $\{F_1, F_2, \ldots, F_n\}$ di Ω, ovvero un insieme di eventi tali che:

- $F_i \neq \varnothing \quad \forall i \in \{1, \ldots, n\}$
- $F_i \cap F_j = \varnothing \quad \forall i \neq j$
- $\bigcup_{i=1}^{n} F_i = \Omega$

Per un evento $E \subseteq \Omega$, possiamo scrivere E come unione disgiunta di più parti:

$$E \;=\; (E \cap F_1) \;\cup\; (E \cap F_2) \;\cup\; \cdots \;\cup\; (E \cap F_n) = \bigcup_{i=1}^{n} (E \cap F_i)$$

dove $(E \cap F_i) \cap (E \cap F_j) = \varnothing$ per $\forall i, j \mid i \neq j$.

Essendo E l'unione di eventi disgiunti, è possibile applicare il terzo assioma di Kolmogorov:

$$\mathbb{P}(E) \;=\; \mathbb{P}\Big(\bigcup_{i=1}^{n} (E \cap F_i)\Big) \;\overset{K3}{=}\; \sum_{i=1}^{n} \mathbb{P}(E \cap F_i)$$

Tramite la regola di fattorizzazione, per ogni F_i con $\mathbb{P}(F_i) \neq 0$ si ottiene $\mathbb{P}(E \cap F_i) = \mathbb{P}(E|F_i)\,\mathbb{P}(F_i)$

Sommando su tutti gli indici i si ottiene dunque la *formula delle probabilità totali in forma estesa*:

$$\mathbb{P}(E) = \sum_{i=1}^{n} \mathbb{P}(E|F_i)\,\mathbb{P}(F_i) \tag{6.2.3}$$

Questa relazione generalizza il caso binario $\{F, \overline{F}\}$ e permette di calcolare $\mathbb{P}(E)$ suddividendo lo spazio degli esiti in una partizione $\{F_1, F_2, \ldots, F_n\}$. In tal modo, ciascun insieme F_i ha probabilità $\mathbb{P}(F_i)$ e, all'interno di ciascuno, si considera la probabilità condizionata $\mathbb{P}(E|F_i)$. Sommando tutti i contributi $\mathbb{P}(E|F_i)\,\mathbb{P}(F_i)$ si ottiene $\mathbb{P}(E)$.

6.3 Teorema di Bayes

Una volta chiarite la regola di fattorizzazione e la formula (o teorema) delle probabilità totali, è naturale introdurre il teorema di Bayes, che fornisce un modo per capovolgere il condizionamento di un evento E rispetto a un altro evento F.

Siano E e F due eventi di uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ con $\mathbb{P}(E) \neq 0$. Allora vale la seguente formula:

$$\mathbb{P}(F|E) = \frac{\mathbb{P}(E|F)\,\mathbb{P}(F)}{\mathbb{P}(E|F)\,\mathbb{P}(F) + \mathbb{P}(E|\overline{F})\,\mathbb{P}(\overline{F})} = \frac{\mathbb{P}(E|F)\,\mathbb{P}(F)}{\mathbb{P}(E)} \tag{6.3.1}$$

Dimostrazione Tramite la formula (6.2.1) si è visto che $\mathbb{P}(E \cap F)$ può essere scritta in 2 modi equivalenti:

$$\mathbb{P}(E \cap F) = \mathbb{P}(E|F)\,\mathbb{P}(F) = \mathbb{P}(F|E)\,\mathbb{P}(E)$$

Dato che $\mathbb{P}(E) \neq 0$, isolando $\mathbb{P}(F|E)$ si ottiene proprio la formula di Bayes. Si ricordi che tramite la formula binaria delle probabilità totali (6.2.2) si ha che $\mathbb{P}(E|F)\,\mathbb{P}(F) + \mathbb{P}(E|\overline{F})\,\mathbb{P}(\overline{F}) = \mathbb{P}(E)$.

Mentre $\mathbb{P}(E|F)$ descrive la probabilità che E accada dopo che F è avvenuto, $\mathbb{P}(F|E)$ sposta l'attenzione su F, supponendo di aver già osservato E.
Inoltre il denominatore $\mathbb{P}(E)$ funge da normalizzatore: rappresenta la probabilità totale di E e assicura che la probabilità condizionata $\mathbb{P}(F|E)$ sia un numero tra 0 e 1.

Forma estesa Estendendo il ragionamento a una partizione generale di Ω, si ottiene la forma estesa del teorema di Bayes. Sia $\{F_1, \ldots, F_n\}$ una partizione di Ω e $\mathbb{P}(F_i) \neq 0$ per ogni i, e sia E sia un evento tale per cui $\mathbb{P}(E) \neq 0$, allora:

$$\mathbb{P}(F_i|E) = \frac{\mathbb{P}(E|F_i)\,\mathbb{P}(F_i)}{\displaystyle\sum_{k=1}^{n} \mathbb{P}(E|F_k)\,\mathbb{P}(F_k)} = \frac{\mathbb{P}(E|F_i)\,\mathbb{P}(F_i)}{\mathbb{P}(E)} \tag{6.3.2}$$

dove il denominatore è $\mathbb{P}(E)$ per via della formula estesa delle probabilità totali.

6.3.1 Classificatori naive Bayes

Un classificatore è un meccanismo che, dati degli oggetti (individui) su cui si vuole effettuare una distinzione, associa a ciascun oggetto una classe tra quelle disponibili. Per esempio, potremmo suddividere gli individui in "positivi" o "negativi" rispetto a una determinata condizione.

Nel contesto di un classificatore bayesiano, si sfrutta il teorema di Bayes per valutare la probabilità che un individuo appartenga a una certa classe, sulla base delle proprietà che abbiamo osservato per quell'individuo. In generale, se consideriamo:

- n proprietà (o variabili aleatorie) $X_1, \ldots, X_n$, con valori $\{x_1, \ldots, x_n\}$
- m classi $\{y_1, \ldots, y_m\}$ (ognuna corrisponde a un evento $\{Y = y_k\}$)

Per un individuo di cui abbiamo misurato $(x_1, \ldots, x_n)$ come realizzazioni di $X_1, \ldots, X_n$, vorremmo attribuirgli la classe $\{Y = y_k\}$ che risulta più "probabile" alla luce di tali proprietà. Il teorema di Bayes ci dice che:

$$\mathbb{P}\big(Y = y_k \mid X_1 = x_1,\ \ldots,\ X_n = x_n\big) \;=\; \frac{\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n \mid Y = y_k\big)\,\mathbb{P}(Y = y_k)}{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}$$

Per classificare l'individuo, dobbiamo scegliere la classe $\{Y = y_k\}$ che massimizza la probabilità a posteriori $\mathbb{P}(Y = y_k \mid X_1 = x_1, \ldots, X_n = x_n)$. Tuttavia, la stima diretta della probabilità congiunta $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid Y = y_k)$ può risultare molto onerosa, poiché richiede di considerare tutte le combinazioni dei valori $(x_1, \ldots, x_n)$.

Il classificatore naive Bayes semplifica tale stima assumendo che, condizionatamente alla classe $Y = y_k$, le variabili $X_1, \ldots, X_n$ siano approssimativamente indipendenti. In formule:

$$\mathbb{P}\big(X_1 = x_1, \ldots, X_n = x_n \mid Y = y_k\big) \;\approx\; \prod_{i=1}^{n} \mathbb{P}\big(X_i = x_i \mid Y = y_k\big)$$

Sostituendo questa ipotesi nella versione bayesiana precedente, si ottiene:

$$\mathbb{P}\big(Y = y_k \mid X_1 = x_1,\, \ldots,\, X_n = x_n\big) \;\approx\; \frac{\mathbb{P}(Y = y_k) \prod_{i=1}^{n} \mathbb{P}\big(X_i = x_i \mid Y = y_k\big)}{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)}$$

Il denominatore $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$ non dipende dalla classe y_k ma solo dai valori osservati $(x_1, \ldots, x_n)$. Per la decisione di classificazione, cioè per confrontare le probabilità di classi diverse, esso funge da costante di normalizzazione, la stessa per ogni classe candidata. Di conseguenza, è sufficiente determinare la classe $\{Y = y_{k^*}\}$ che massimizza il prodotto:

$$\mathbb{P}(Y = y_k) \prod_{i\,=\,1}^{n} \mathbb{P}(X_i = x_i \mid Y = y_k)$$

In pratica, per classificare un individuo con proprietà $(x_1, \ldots, x_n)$, si calcola per ogni classe y_k il prodotto $\mathbb{P}(Y = y_k) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid Y = y_k)$ e si sceglie la classe che ne produce il valore più alto. In notazione compatta:

$$k^* = \arg \max_{k \in \{1,\ldots,m\}} \Big[\mathbb{P}(Y = y_k) \prod_{i=1}^{n} \mathbb{P}(X_i = x_i \mid Y = y_k) \Big]$$

L'ipotesi di indipendenza condizionale riduce drasticamente il numero di stime necessarie per calcolare le probabilità, passando da una modellazione congiunta (potenzialmente esponenziale) a una sommatoria di stime "marginali" ($\sum_i |\mathcal{X}i|$ invece di $\prod i|\mathcal{X}_i|$).

Sebbene nella pratica le variabili X_i possano non essere completamente indipendenti all'interno di una stessa classe (da cui l'aggettivo naive), l'approssimazione risulta spesso efficace in molti scenari, a fronte di una grande semplicità computazionale.

6.4 Eventi indipendenti

In generale, la probabilità condizionata $\mathbb{P}(E|F)$ differisce da $\mathbb{P}(E)$, poiché il verificarsi di F fornisce informazioni che possono modificare la probabilità che si verifichi E. Tuttavia, se si ha $\mathbb{P}(E|F) = \mathbb{P}(E)$, allora si dice che gli eventi E e F sono *indipendenti*. Questo significa che la conoscenza del verificarsi di F non influisce sulla probabilità che E si realizzi.

Partendo dalla definizione di probabilità condizionata, l'uguaglianza $\mathbb{P}(E|F) = \mathbb{P}(E)$, per $\mathbb{P}(F) \neq 0$, implica

$$\mathbb{P}(E|F) = \frac{\mathbb{P}(E \cap F)}{\mathbb{P}(F)} = \mathbb{P}(E)$$

Moltiplicando entrambi i membri per $\mathbb{P}(F)$ si ottiene una definizione simmetrica di indipendenza:

$$E,\, F \text{ indipendenti} \quad \Longleftrightarrow \quad \mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F) \tag{6.4.1}$$

Analogamente, ponendo $\mathbb{P}(F|E) = \mathbb{P}(F)$ per $\mathbb{P}(E) \neq 0$ si giunge alla medesima conclusione.

Questa relazione evidenzia che, se E è indipendente da F, anche F risulta indipendente da E, poiché entrambi gli enunciati implicano l'uguaglianza della probabilità dell'intersezione al prodotto delle probabilità marginali.

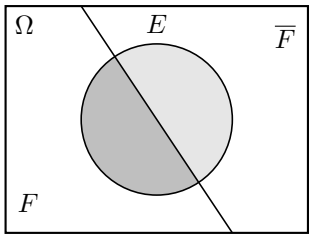
La nozione di indipendenza si conserva rispetto ad alcune operazioni insiemistiche elementari tra eventi. In particolare, se due eventi sono indipendenti, anche semplici combinazioni di essi, come intersezioni, unioni o complementi, possono preservare la proprietà di indipendenza.

Nel seguito, si dimostra questo fatto per quanto riguarda l'operazione di complemento.

Teorema Se E e F sono indipendenti, allora lo sono anche E e $\overline{F}$.

Dimostrazione:

- Affinché E e $\overline{F}$ siano indipendenti, bisogna dimostrare che $\mathbb{P}(E \cap \overline{F}) = \mathbb{P}(E)\,\mathbb{P}(\overline{F})$



$E = (E \cap F) \cup (E \cap \overline{F})$

- Osservando il diagramma, è possibile suddividere E in una partizione $E = \{E \cap F, E \cap \overline{F}\}$:
  1. $(E \cap F) \cup (E \cap \overline{F}) = E$
  2. $(E \cap F) \cap (E \cap \overline{F}) = \varnothing$

  Diventa quindi possibile applicare il terzo assioma di Kolmogorov

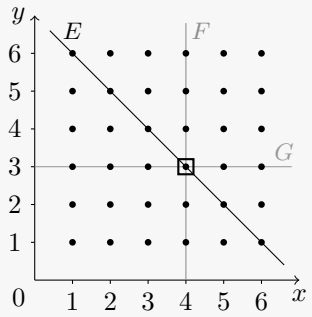
- $\mathbb{P}(E) \overset{K3}{=} \mathbb{P}(E \cap F) + \mathbb{P}(E \cap \overline{F}) \;\Rightarrow\; \mathbb{P}(E \cap \overline{F}) = \mathbb{P}(E) - \mathbb{P}(E \cap F)$
- Dato che E e F sono indipendenti, allora vale $\mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F)$. Sostituendo, si ottiene:

$$\mathbb{P}(E \cap \overline{F}) = \mathbb{P}(E) - \mathbb{P}(E)\,\mathbb{P}(F)$$
$$\mathbb{P}(E \cap \overline{F}) = \mathbb{P}(E)\,(1 - \mathbb{P}(F)) = \mathbb{P}(E)\,\mathbb{P}(\overline{F})$$

Estensione dell'indipendenza

Si osserva che non è possibile estendere l'indipendenza a più eventi richiedendo solo l'indipendenza a coppie, similmente a quanto invece si fa per provare la disgiunzione tra più eventi.

Esempio Si immagini di tirare due dadi. Lo spazio degli esiti di questo esperimento è descritto da $\Omega = \{(x, y) \mid x, y \in \{1, \ldots, 6\}\}$ dove si intende che si ottiene l'esito (x, y) se il risultato del primo dado è x e quello del secondo y. Si supponga che entrambi i dadi non siano truccati, e di trovarci quindi in uno spazio equiprobabile dove $\mathbb{P}((x, y)) = 1/|\Omega| = 1/36$.



Si considerano i seguenti eventi:

$E = \{(x, y) \in \Omega \mid x + y = 7\} = \{\text{somma dei dadi è } 7\}$
$F = \{(x, y) \in \Omega \mid x = 4\} = \{4 \text{ sul primo dado}\}$
$G = \{(x, y) \in \Omega \mid y = 3\} = \{3 \text{ sul secondo dado}\}$

Calcolando le probabilità di ciascun evento, si trova che

$$\mathbb{P}(E) = \mathbb{P}(F) = \mathbb{P}(F) = 1/6$$

Osservando il grafico a lato si osserva, infatti, che ogni evento, rappresentato dalla propria retta, contiene 6 esiti. Dividendo questa quantità per $|\Omega| = 36$ si ottiene proprio 1/6.

Gli eventi sono indipendenti a coppie, infatti:

$\mathbb{P}(E \cap F) = 1/36 = \mathbb{P}(E)\,\mathbb{P}(F)$

$\mathbb{P}(E \cap G) = 1/36 = \mathbb{P}(E)\,\mathbb{P}(G)$

$\mathbb{P}(F \cap G) = 1/36 = \mathbb{P}(F)\,\mathbb{P}(G)$

Se si calcola $\mathbb{P}(E|F \cap G) = 1$, si osserva che la probabilità di E dato $F \cap G$ risulta diversa dalla probabilità marginale $\mathbb{P}(E)$. Questo implica che E è dipendente dal verificarsi di $F \cap G$, e di conseguenza i tre eventi E, F e G non sono indipendenti nel senso globale.
Infatti, affinché valga l'indipendenza complessiva, dovrebbe risultare $\mathbb{P}(E|F \cap G) = \mathbb{P}(E)$, condizione che in questo caso non è soddisfatta.

Dati tre eventi E, F e G, questi sono indipendenti se e solo se:

- $\mathbb{P}(E \cap F) = \mathbb{P}(E)\,\mathbb{P}(F)$
- $\mathbb{P}(E \cap G) = \mathbb{P}(E)\,\mathbb{P}(G)$
- $\mathbb{P}(F \cap G) = \mathbb{P}(F)\,\mathbb{P}(G)$
- $\mathbb{P}(E \cap F \cap G) = \mathbb{P}(E)\,\mathbb{P}(F)\,\mathbb{P}(G)$

Si può osservare come anche in questo contesto valga quanto discusso in precedenza: se gli eventi E, F e G sono indipendenti nel senso globale, allora anche eventi ottenuti tramite semplici operazioni insiemistiche (come intersezione, unione o complementare) risultano indipendenti senza necessità di ulteriori verifiche. Questa proprietà conferma che l'indipendenza si estende naturalmente agli eventi costruiti a partire da eventi già indipendenti.

Teorema Se E, F e G sono indipendenti, allora anche E e $F \cup G$ sono indipendenti.

Dimostrazione:

- Affinché E e $F \cup G$ siano indipendenti, bisogna dimostrare che $\mathbb{P}(E \cap (F \cup G)) = \mathbb{P}(E)\,\mathbb{P}(F \cup G)$
- Si applica la proprietà distributiva su $E \cap (F \cup G)$:

$$\mathbb{P}(E \cap (F \cup G)) = \mathbb{P}((E \cap F) \cup (E \cap G)) = \mathbb{P}(E \cap F) + \mathbb{P}(E \cap G) - \mathbb{P}(\underbrace{(E \cap F) \cap (E \cap G)}_{E \cap F \cap G}) =$$

$$= \mathbb{P}(E)\,\mathbb{P}(F) + \mathbb{P}(E)\,\mathbb{P}(G) - \mathbb{P}(E)\,\mathbb{P}(F)\,\mathbb{P}(G) = \mathbb{P}(E)[\mathbb{P}(F) + \mathbb{P}(G) - \underbrace{\mathbb{P}(F)\,\mathbb{P}(G)}_{\mathbb{P}(F \cap G)}]$$

- Si osserva che $\mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G)$ corrisponde a $\mathbb{P}(F \cup G)$ dagli assiomi di Kolmogorov, di conseguenza si è dimostrato il teorema:

$$\mathbb{P}(E)\,[\mathbb{P}(F) + \mathbb{P}(G) - \mathbb{P}(F \cap G)] = \mathbb{P}(E)\,\mathbb{P}(F \cup G)$$

Generalizzazione dell'indipendenza Si abbiano n eventi $E_1, \ldots, E_n \subseteq \Omega$, questi sono indipendenti se e solo se $\forall r \leq n \quad \forall\, 1 \leq \alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_r \leq n$ con $\alpha_i \in \mathbb{N}$ si ha che

$$\mathbb{P}\Big(\bigcap_{j=1}^{r} E_{\alpha_j}\Big) = \prod_{j=1}^{r} \mathbb{P}(E_{\alpha_j})$$

Questo significa che, dati più eventi, l'indipendenza globale richiede che ogni intersezione di un numero qualsiasi di essi abbia probabilità uguale al prodotto delle probabilità dei singoli eventi coinvolti.

7 Variabili aleatorie

Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e una σ-algebra $\mathcal{E}$ su $\mathbb{R}$, una variabile aleatoria $X : \Omega \to \mathbb{R}$ è definita come una funzione misurabile, ossia tale che per ogni $A \in \mathcal{E}$ l'evento

$$\{X \in A\} \equiv \{\omega \in \Omega \mid X(\omega) \in A\}$$

appartenga alla σ-algebra $\mathcal{F}$ di Ω. In altre parole, questa condizione garantisce che, per ogni insieme misurabile A in $\mathbb{R}$, il corrispondente sottoinsieme di Ω (cioè l'insieme degli esiti per cui X assume valori in A) sia un evento a cui è possibile assegnare una probabilità.

In particolare, se si considera $A = \{\alpha\}$ per un qualsiasi $\alpha \in \mathbb{R}$, allora si ottiene l'evento

$$\{X = \alpha\} \equiv \{\omega \in \Omega \mid X(\omega) = \alpha\}$$

Questa forma è particolarmente usata quando si trattano variabili aleatorie discrete.

Si osserva che per ogni insieme misurabile $A \in \mathcal{E}$, la probabilità che la variabile aleatoria X assuma valori in A, ossia la probabilità dell'evento $\{X \in A\}$, è definita come

$$\mathbb{P}(X \in A) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) \in A\}) = \mathbb{P}(X^{-1}(A))$$

Definizione rigorosa Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e una σ-algebra $\mathcal{E}$ su $\mathbb{R}$, una variabile aleatoria $X : \Omega \to \mathbb{R}$ è definita come una funzione misurabile, ossia una funzione che trasforma gli esiti dello spazio campionario Ω in valori reali in modo da preservare la struttura misurabile. Questo significa che per ogni insieme $B \in \mathcal{E}$ (ossia per ogni insieme misurabile in $\mathbb{R}$) l'insieme preimmagine

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$$

appartiene alla σ-algebra $\mathcal{F}$ su Ω.

In altre parole, la funzione X rispetta la struttura misurabile: essa trasforma gli esiti possibili in Ω in valori misurabili in $\mathbb{R}$, permettendo così l'applicazione degli strumenti della teoria della probabilità alla variabile X.

I singoli valori che la variabile aleatoria X può assumere sono detti *specificazioni* o *variabili osservabili*. L'insieme di tutte le specificazioni di X costituisce il *dominio di supporto* della variabile, e si indica con D_X.

Una variabile aleatoria si dice *discreta* se può assumere solo un numero finito o infinito numerabile di valori, mentre si dice continua se può assumere valori in un intervallo (o insieme) non numerabile di $\mathbb{R}$.

7.0.1 Funzione di ripartizione

La funzione di ripartizione, o funzione di distribuzione cumulativa, di una variabile aleatoria X a valori reali è la funzione che associa a ciascun valore x la probabilità che la variabile X assuma valori minori o uguali a x.

In altre parole, è la funzione $F_X : \mathbb{R} \to [0, 1]$ definita da

$$F_X(x) = \mathbb{P}(X \leq x) \quad \forall x \in \mathbb{R}$$

La definizione della variabile aleatoria X come funzione misurabile implica che $\forall x \in \mathbb{R}$ l'insieme $\{X \leq x\}$, che è l'immagine inversa dell'intervallo $(-\infty, x]$, sia un evento, ossia appartenga alla σ-algebra $\mathcal{F}$ dello spazio di probabilità su cui X è definita. Senza questa condizione di misurabilità, non avrebbe senso assegnare una probabilità a tali eventi e, di conseguenza, la funzione di ripartizione non sarebbe ben definita.

Proprietà La funzione di ripartizione F_X possiede le seguenti proprietà:

- $0 \leq F_X(x) \leq 1 \ \ \forall x \in \mathbb{R}$
- Monotonicità: F_X è non decrescente, ossia se $x_1 < x_2$, allora $F_X(x_1) \leq F_X(x_2)$. Questo riflette il fatto che $\{X \leq x_1\} \subseteq \{X \leq x_2\}$.
- Continuità a destra: per ogni $x \in \mathbb{R} \ \lim_{y \to x^+} F_X(y) = F_X(x)$
- Limiti estremi: $\lim_{x \to -\infty} F_X(x) = 0$ e $\lim_{x \to +\infty} F_X(x) = 1$

La funzione di ripartizione non è necessariamente continua a sinistra.

Dalla definizione segue che la probabilità che X risieda in un intervallo semichiuso $(a, b]$, dove $a < b$, è

$$\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$$

Dimostrazione Si consideri l'evento $\{a < X \leq b\}$ con $a < b$. È possibile affermare che

$$\{X \leq b\} = \{X \leq a\} \cup \{a < X \leq b\}$$

dove $\{X \leq a\}$ e $\{a < X \leq b\}$ sono due eventi disgiunti. Utilizzando il terzo assioma di Kolmogorov:

$$\mathbb{P}(X \leq b) = \mathbb{P}(X \leq a) + \mathbb{P}(a < X \leq b)$$

Riscrivendo il tutto utilizzando la funzione di ripartizione, si ha

$$F_X(b) = F_X(a) + \mathbb{P}(a < X \leq b)$$

Da cui si ottiene che $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$. La tesi è quindi dimostrata.

Si può quindi esprimere la probabilità di un qualsiasi evento relativo a X in termini della funzione di ripartizione, sfruttando la rappresentazione degli eventi come unioni di intervalli semiaperti.

Sia $A \subseteq \mathbb{R}$ un evento della variabile X, allora si intende in realtà l'evento $\{X \in A\} \equiv \{\omega \in \Omega \mid X(\omega) \in A\}$, ossia il sottoinsieme di Ω formato da tutti gli esiti ω per cui la trasformazione X produce un valore appartentente all'insieme $A \subseteq \mathbb{R}$.

A può essere scritto eventualmente come un'unione disgiunta di intervalli della forma $(a, b]$. Di conseguenza, grazie alla proprietà già dimostrata per gli intervalli semiaperti, ossia $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$, si ottiene che

$$\mathbb{P}(X \in A) = \sum_k \Big[F_X(b_k) - F_X(a_k)\Big]$$

dove la somma è presa sulla rappresentazione disgiunta di A in intervalli semiaperti $(a_k, b_k]$.

7.1 Variabili aleatorie discrete

Le variabili aleatorie discrete sono variabili il cui dominio di supporto è un insieme numerabile:

$$D_X = \{x \in \mathbb{R} \mid \mathbb{P}(X = x) > 0\}$$

Poiché il supporto D_X di una variabile aleatoria discreta è numerabile e, in genere, consiste di punti isolati, esso può essere ordinato in una sequenza $(x_1, x_2, \dots)$. In questo contesto, ogni specificazione, salvo l'eventuale massimo, ha un successivo nell'ordine naturale dei numeri reali.

Questo insieme si determina esaminando i valori per cui la funzione di massa di probabilità non è zero.

Le probabilità di un evento $\{X \in A\}$, per ogni $A \subseteq D_X$, si ottiene sommando le probabilità dei valori di X che ricadono in A:

$$\mathbb{P}(X \in A) = \sum_{x \in A \cap D_X} \mathbb{P}(X = x)$$

Proprietà La somma delle probabilità di tutte le specificazioni deve essere 1:

$$\sum_{x \in D_X} \mathbb{P}(X = x) = 1$$

Dimostrazione Se D_X è il supporto di X, allora Ω si può scrivere come unione disgiunta di tutti gli eventi $\{X = x\}$ con $x \in D_X$, ossia:

$$\Omega = \bigcup_{x \in D_X} \{X = x\}$$

perché ogni esito ω in Ω genera un valore $X(\omega)$ che appartiene al supporto. Essendo gli eventi $\{X = x\}$ a due a due disgiunti, è possibile applicare il terzo assioma di Kolmogorov:

$$\mathbb{P}\Big(\bigcup_{x \in D_X} \{X = x\} \Big) = \sum_{x \in D_X} \mathbb{P}(X = x) = 1$$

Infine, poiché l'unione copre interamente Ω, per il primo assioma si ottiene:

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

7.1.1 Funzione di probabilità

Data una variabile aleatoria discreta X, la sua *funzione di (massa di) probabilità* è una funzione di variabile reale che assegna ad ogni valore di X la probabilità dell'evento elementare $\{X = x\}$.

- Nel caso in cui X sia continua, ogni singolo punto ha probabilità zero, e dunque la funzione di massa perde di significato; in tal caso si usa la funzione di densità.

Formalmente, data una variabile aleatoria $X : \Omega \to \mathbb{R}$, la funzione di probabilità è la funzione $p_X : \mathbb{R} \to [0, 1]$ definita da

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\{\omega \in \Omega \mid X(\omega) = x\}) \quad \forall x \in \mathbb{R}$$

che associa ad ogni valore x assunto da X la probabilità che X assuma esattamente quella specificazione.

Proprietà La funzione di probabilità rispetta le seguenti proprietà:

- $\forall x \in \mathbb{R} \;\; p_X(x) \geq 0$: trattandosi del calcolo di una probabilità, questa funzione non può assumere valori negativi
- $p_X(x) \neq 0$ solo per $x \in \mathbb{R} \wedge x \in D_X$
- per $x \notin D_X$ si assume che $p_X(x) = 0$ [2]

Come già dimostrato in precedenza, la somma delle probabilità su tutto il supporto di X deve essere pari a 1. Di conseguenza, se D_X è numerabile, scrivendo $p_X(x) = \mathbb{P}(X = x)$ si ottiene

$$\sum_{x \in D_X} p_X(x) = 1$$

[2] L'uso della funzione indicatrice $I_{D_X}(x)$ garantisce formalmente che la funzione di probabilità sia definita unicamente sui valori appartenenti al supporto D_X. Infatti, per ogni $x \notin D_X$ abbiamo $I_{D_X}(x) = 0$, e quindi $p_X(x) = \mathbb{P}(X = x) \cdot I_{D_X}(x) = 0$. Questo esplicita il fatto che ogni x non preso in considerazione (ossia, per cui $\mathbb{P}(X = x) = 0$) viene escluso dalla definizione di p_X.

Funzione indicatrice Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria discreta a valori reali con dominio di supporto $D_X = \{x \in \mathbb{R} \mid \mathbb{P}(X = x) > 0\}$. La funzione di massa di probabilità p_X può essere espressa in forma compatta mediante la funzione indicatrice. Formalmente, si scrive

$$p_X(x) = \mathbb{P}(X = x)\, I_{D_X}(x)$$

dove $I_{D_X} : \mathbb{R} \to \{0, 1\}$ è la funzione indicatrice dell'insieme D_X, definita come

$$I_{D_X}(x) = \begin{cases} 1 & \text{se } x \in D_X \\ 0 & \text{altrimenti} \end{cases}$$

Nel caso in cui X sia uniformemente distribuita su D_X, cioè se $\mathbb{P}(X = x) = c$ per ogni $x \in D_X$ (con c costante tale che $\sum_{x \in D_X} c = 1$), la funzione di probabilità si semplifica a $p_X(x) = c \cdot I_{D_X}(x)$.

Relazione tra funzione di ripartizione e di probabilità

Sia F_X la funzione di ripartizione di X e sia p_X la sua funzione di probabilità. La prima rappresenta la somma cumulativa dei valori della seconda.

Per definizione si ha che $F_X(x) = \mathbb{P}(X \leq x)$ e, per una variabile discreta con dominio di supporto numerabile, possiamo esprimere $F_X(x)$ come la somma delle probabilità associate a tutti i valori $y \in D_X$ tali che $y \leq x$. In altre parole

$$F_X(x) = \sum_{y \leq x} p_X(y)$$

Questo significa che, per ogni valore x, la funzione $F_X(x)$ è data dal contributo cumulativo dei salti indotti da ciascun valore specifico di X minore o uguale a x. In particolare, se x_0 è un punto in cui X può assumere un valore, ossia $x_0 \in D_X$, allora il salto di F_X in x_0 è proprio

$$F_X(x_0) - \lim_{x \to x_0^-} F_X(x) = p_X(x_0)$$

Indicando con $F_X(x^-)$ il limite sinistro della F_X in x, si può riscrivere l'equazione precedente in

$$p_X(x) = F_X(x) - F_X(x^-)$$

Da ciò si deduce che se X è una variabile aleatoria continua, tale valore è nullo in ogni punto poiché la sua funzione di ripartizione è continua, e di conseguenza $F_X(x) = F_X(x^-)$

Proposizione Quando si conosce la funzione di probabilità p_X, oppure la funzione di ripartizione F_X, di una variabile aleatoria X qualsiasi, si hanno abbastanza informazioni per poter calcolare la probabilità di ogni evento che dipenda solo da tale variabile aleatoria. Si dice in questo caso che si conosce la *distribuzione* o *legge* della variabile aleatoria considerata.
Affermare quindi che X e Y hanno la stessa distribuzione significa che le rispettive funzioni di ripartizioni sono identiche, $X \sim F_X \equiv F_Y \sim Y$, e quindi anche che $\mathbb{P}(X \in A) = \mathbb{P}(Y \in A)$ per ogni insieme di valori $A \subseteq \mathbb{R}$.

7.2 Variabili multivariate

Risulta necessario talvolta ridurre un esperimento casuale a più variabili aleatorie, in quanto l'oggetto di interesse sono proprio le relazioni presenti tra due o più grandezze numeriche. La coppia (X, Y), con X e Y variabili aleatorie, è detta *variabile aleatoria bivariata*; generalizzando, una *variabile aleatoria multivariata* è un vettore aleatorio $X = (X_1, X_2, \ldots, X_n)$, dove X_i è una variabile aleatoria a valori reali.

Verranno ora presentate le definizioni e le proprietà delle variabili aleatorie bivariate, che possono essere però estese sul caso di variabili multivariate.

Funzione di ripartizione congiunta

Considerata una variabile aleatoria bivariata (X, Y), la funzione di ripartizione congiunta è definita come

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y)$$

dove la virgola denota l'intersezione tra i due eventi $\{X \leq x\}$ e $\{Y \leq y\}$.

La conoscenza di questa funzione permette di calcolare la probabilità di tutti gli eventi che dipendono, singolarmente o congiuntamente, da X e Y. La funzione di ripartizione di X può essere ottenuta dalla funzione di ripartizione congiunta come

$$F_X(x) = \lim_{y \to +\infty} F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq +\infty) = F_{X,Y}(x, +\infty)$$

Analogamente, la funzione di ripartizione di Y è

$$F_Y(y) = \lim_{x \to +\infty} F_{X,Y}(x, y) = \mathbb{P}(X \leq +\infty, Y \leq y) = F_{X,Y}(+\infty, y)$$

Le funzioni di ripartizioni F_X e F_Y sono dette *marginali*.

Funzione di probabilità congiunta

Nel caso in cui X e Y siano variabili aleatorie discrete, la funzione di probabilità congiunta è definita come

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$$

Le funzioni di probabilità p_X e p_Y si possono ricavare da quella congiunta notando che, siccome Y deve assumere uno dei valori y, l'evento $\{X = x\}$ può essere visto come l'unione al variare degli y degli eventi $\{X = x, Y = y\}$, che sono mutualmente esclusivi. In formule:

$$\{X = x\} = \bigcup_{y \in D_Y} \{X = x, Y = y\}$$

Tramite il terzo assioma di Kolmogorov, si ha quindi

$$\mathbb{P}(X = x) = \sum_{y \in D_Y} \mathbb{P}(X = x, Y = y) = \sum_{y \in D_Y} p_{X,Y}(x, y) = p_X(x)$$

Analogamente, si può dimostrare che

$$p_Y(y) = \sum_{x \in D_X} p_{X,Y}(x, y)$$

Queste due funzioni di probabilità sono dette *marginali*. È importante notare che sebbene le funzioni di massa di probabilità marginali si possono sempre ricavare da quella congiunta, il viceversa è falso.

7.2.1 Variabili indipendenti

Due variabili aleatorie X e Y sono dette *indipendenti* se tutti gli eventi relativi alla prima sono indipendenti da tutti gli eventi relativi alla seconda.

Formalmente, due variabili aleatorie che riguardano lo stesso esperimento casuale sono indipendenti se e solo se, per ogni insieme $A, B \subseteq \mathbb{R}$, si ha

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

ovvero, se per ogni scelta di A e B, gli eventi $\{X \in A\}$ e $\{Y \in B\}$ sono indipendenti. In caso contrario X e Y sono dette *dipendenti*.

Proprietà

Siano X e Y due variabili aleatorie indipendenti. Allora valgono le seguenti proprietà:

- $F_{X,Y}(x,y) = F_X(x)\,F_Y(y)$
- $p_{X,Y}(x,y) = p_X(x)\,p_Y(y)$

  Dimostrazione

  1. X, Y indipendenti $\Rightarrow$ $p_{X,Y}(x,y) = p_X(x)\,p_Y(y) \quad \forall x, y$

     Si fissino x, y : $\quad p_{X,Y}(x,y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X \in \underbrace{\{x\}}_{:=A}, Y \in \underbrace{\{y\}}_{:=B})$

     $$= \mathbb{P}(X \in A, Y \in B) \overset{(1)}{=} \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$
     $$= \mathbb{P}(X = x)\,\mathbb{P}(Y = y) = p_X(x)\,p_Y(y)$$

     (1): per ipotesi di indipendenza

  2. $p_{X,Y}(x,y) = p_X(x)\,p_Y(y) \quad \forall x, y \;\Rightarrow\; X, Y$ indipendenti

     Si fissino $A, B \subseteq \mathbb{R}$: $\quad \mathbb{P}(X \in A, Y \in B) = \displaystyle\sum_{x \in A, y \in B} \mathbb{P}(X = x, Y = y) = \sum_{x \in A} \sum_{y \in B} p_{X,Y}(x,y)$

     $$= \sum_{x \in A} \sum_{y \in B} p_X(x)\,p_Y(y) = \sum_{x \in A} p_X(x) \sum_{y \in B} p_Y(y)$$
     $$= \mathbb{P}(X \in A)\,\mathbb{P}(Y \in B)$$

     Di conseguenza X e Y sono indipendenti.

  Si è dimostrata perciò la tesi in entrambi i versi.

È possibile estendere l'indipendenza a più variabili aleatorie. In questo caso, si dice che $X_1, X_2, \ldots, X_n$ sono indipendenti se

$$\forall A_1, \ldots, A_n \subseteq \mathbb{R} \quad \mathbb{P}\Big(\bigcap_{i=1}^{n} X_i \in A_i\Big) = \prod_{i=1}^{n} \mathbb{P}(X_i \in A_i)$$

7.3 Valore atteso

7.3.1 Valore atteso di una variabile discreta

Sia X una variabile aleatoria discreta che può assumere i valori $D_X = \{x_1, \dots, x_n, \dots\}$, il valore atteso di X, che si indica con $\mathbb{E}[X]$, è il numero [3]

$$\mathbb{E}[X] = \sum_{x \in D_X} x\, \mathbb{P}(X = x) = \sum_{x \in D_X} x\, p_X(x) \tag{7.1.2}$$

Pertanto, il valore atteso rappresenta la media pesata delle specificazioni di X, usando come pesi le probabilità che tali valori vengano assunti da X. Perciò $\mathbb{E}[X]$ è un indice di centralità della distribuzione di X.

Analogamente alla media campionaria, il valore atteso può non corrispondere a una specificazione della variabile aleatoria X. Inoltre $\mathbb{E}[X]$ presenta la stessa unità di misura delle specificazioni.

Funzione indicatrice di un evento La funzione indicatrice di un evento $A \subseteq \Omega$ è definita come una funzione $I_A : \Omega \to \{0, 1\}$ tale che

$$I_A(\omega) = \begin{cases} 1 & \text{sse } A \text{ si verifica, cioè se } \omega \in A \\ 0 & \text{altrimenti} \end{cases}$$

Utilizzando questa definizione, si può dimostrare che il valore atteso della funzione indicatrice coincide con la probabilità dell'evento, infatti:

$$\mathbb{E}[I_A] = \sum_{\omega \in \Omega} I_A(\omega)\, \mathbb{P}(\{\omega\})$$

Notando che $I_A(\omega) = 1$ solo se $\omega \in A$ e 0 altrimenti, la somma si riduce a

$$\mathbb{E}[I_A] = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) = \mathbb{P}(A)$$

Proprietà

Proposizione Se X è una variabile aleatoria discreta con funzione di probabilità p_X, allora, per ogni funzione reale g vale [4]

$$\mathbb{E}[g(X)] = \sum_{x \in D_X} g(x)\, \mathbb{P}(X = x) = \sum_{x \in D_X} g(x)\, p_X(x) \tag{7.1.3}$$

Si consideri infatti una variabile aleatoria X di cui si conosce la distribuzione. Anziché calcolare il valore atteso di X, può essere conveniente calcolare il valore atteso di una funzione $g(X)$, dove g è una funzione $g : \mathbb{R} \to \mathbb{R}$. Si nota che $g(X)$ è anch'essa una variabile aleatoria, e quindi è possibile calcolarne la distribuzione in un qualche modo; dopo averla ottenuta si può calcolare $\mathbb{E}[g(X)]$ con la sua definizione usuale.

[3] Si osserva che $\mathbb{E}[X]$ è definito solo se la serie (7.1.2) converge in valore assoluto, ovvero deve valere

$$\sum_{x \in D_X} |x|\, p_X(x) < \infty$$

In caso contrario si dice che X non ha valore atteso.

[4] Anche in questo caso si richiede che la serie converga in valore assoluto affinché $\mathbb{E}[g(X)]$ sia definito:

$$\sum_{x \in D_X} |g(x)|\, p_X(x) < \infty$$

Proposizione Per ogni coppia di costanti reali a e b, si ha $\mathbb{E}[aX+b] = a\,\mathbb{E}[X] + b$.

Dimostrazione:

$$\mathbb{E}[aX+b] = \sum_{x\in D_X} (ax+b)\,p_X(x) = a\sum_{x\in D_X} x\,p_X(x) + b\sum_{x\in D_X} p_X(x) = a\,\mathbb{E}[X] + b$$

Il valore atteso è quindi un operatore lineare, proprio come la media campionaria.

Si presentano due casi:

- se $a = 0$, si ha $\mathbb{E}[b] = b$ e quindi il valore atteso di una costante è la costante stessa. Una costante è infatti una variabile aleatoria degenere che assume un unico valore con probabilità 1.
- se $b = 0$ si ottiene che $\mathbb{E}[aX] = a\,\mathbb{E}[X]$. Di conseguenza il valore atteso scala rispetto alle costanti moltiplicative.

Proposizione Il valore atteso è lineare rispetto alla somma di variabili aleatorie: $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

Dimostrazione:

È possibile estendere la formula (7.1.3) in una variante in due dimensioni: se X e Y sono variabili aleatorie e g è una qualunque funzione di due variabili, allora, se $\mathbb{E}[g(X,Y)]$ è definito, vale

$$\mathbb{E}[g(X,Y)] = \begin{cases} \displaystyle\sum_{x\in D_X}\sum_{y\in D_Y} g(x,y)\,p_{X,Y}(x,y) & \text{nel discreto} \\ \displaystyle\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\,f_{X,Y}(x,y)\,dx\,dy & \text{nel continuo} \end{cases}$$

Si consideri la funzione $g(x,y) = x+y$ e si applichi la formula precedente. Verrà presentata la dimostrazione per il caso discreto, ma il ragionamento è analogo nel caso continuo:

$$\begin{aligned} \mathbb{E}[X+Y] &= \sum_{x\in D_X}\sum_{y\in D_Y} (x+y)\,p_{X,Y}(x,y) = \sum_{x\in D_X}\sum_{y\in D_Y} x\,p_{X,Y}(x,y) + \sum_{x\in D_X}\sum_{y\in D_Y} y\,p_{X,Y}(x,y) \\ &= \sum_{x\in D_X} x\left(\sum_{y\in D_Y} p_{X,Y}(x,y)\right) + \sum_{y\in D_Y} y\left(\sum_{x\in D_X} p_{X,Y}(x,y)\right) \\ &\overset{(1)}{=} \sum_{x\in D_X} x\,p_X(x) + \sum_{y\in D_Y} y\,p_Y(y) = \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

(1): le sommatorie all'interno delle parentesi tonde rappresentano rispettivamente le funzioni di massa di probabilità marginali di X e Y.

Applicando ricorsivamente questa equazione si può estendere la linearità del valore atteso a un numero finito di variabili aleatorie:

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$$

Formalmente, se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie discrete, si ha

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n}\mathbb{E}[X_i] = \sum_{i=1}^{n}\sum_{x\in D_{X_i}} x\,p_{X_i}(x) = \sum_{x\in D_X} x\left(\sum_{i=1}^{n} p_{X_i}(x)\right)$$

Proposizione Se X e Y sono variabili aleatorie indipendenti, allora $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.

Dimostrazione:

Si consideri la funzione $g(x, y) = xy$ e si applichi la formula (7.1.3):

$$\mathbb{E}[XY] = \sum_{x \in D_X} \sum_{y \in D_Y} xy\, p_{X,Y}(x, y) \overset{(1)}{=} \sum_{x \in D_X} \sum_{y \in D_Y} xy\, p_X(x)\, p_Y(y)$$

$$= \sum_{x \in D_X} x\, p_X(x) \sum_{y \in D_Y} y\, p_Y(y) = \mathbb{E}[X]\,\mathbb{E}[Y]$$

(1): per ipotesi di indipendenza

Teorema Sia $X \geq 0$ una variabile aleatoria discreta a specificazioni non negative, allora

$$\mathbb{E}[X] = \int_0^{+\infty} [1 - F_X(x)]\, dx$$

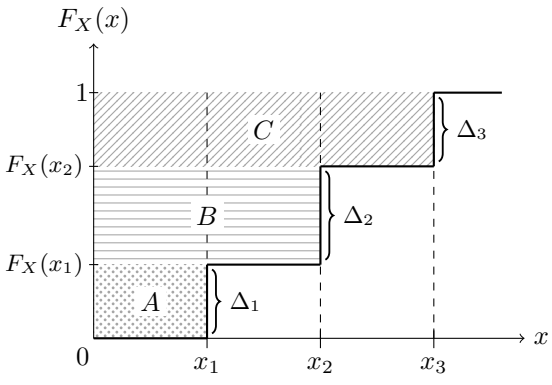
Dimostrazione:

Siano $x_1, x_2, \ldots, x_n$ le specificazioni non nulle assunte da X e siano $F_X(x_1), F_X(x_2), \ldots, F_X(x_n)$ i valori assunti dalla funzione di ripartizione in corrispondenza di tali specificazioni. La funzione di ripartizione è definita come la somma cumulativa delle probabilità associate a ciascun valore di X, e quindi il salto della funzione di ripartizione in corrispondenza di x_i è dato da

$$\Delta_i = F_X(x_i) - F_X(x_{i-1}) = \mathbb{P}(X \leq x_i) - \mathbb{P}(X \leq x_{i-1}) = p_X(x_i) \quad \forall i = 1, \ldots, n$$

Si osserva che ciascun salto Δ_i, se moltiplicato per la corrispondente specificazione x_i, contribuisce al valore atteso $\mathbb{E}[X]$; sommando i contributi di tutti i salti, si ottiene pertanto proprio il valore atteso:

$$\mathbb{E}[X] = \sum_{i=1}^{n} x_i\, p_X(x_i) = \sum_{i=1}^{n} x_i\, \Delta_i$$



Si consideri ora il grafico della funzione di ripartizione $F_X(x)$ presentato a sinistra: in questo caso si assumono per semplicità solo tre specificazioni. Per quanto detto precedentemente, il valore atteso di X è dato dalla somma dei contributi dei salti della funzione di ripartizione, che sono rappresentati dalle aree A, B e C. Ma la somma di queste aree corrisponde proprio all'area sopra la curva della funzione di ripartizione, che è pari a

$$\int_0^{+\infty} [1 - F_X(x)]\, dx$$

Si è quindi dimostrato che il valore atteso di una variabile aleatoria discreta non negativa corrisponde all'integrale presentato, provando quindi la tesi.

Osservazione $\mathbb{E}[(X-c)^2] \geq \mathbb{E}[(X-\mu)^2]$

Vi è una interessante proprietà della media che emerge quando si vuole predire con il minore errore possibile il valore che verrà assunto da una variabile aleatoria. Si supponga di voler predire il valore di X: se si sceglie un numero reale c e si dice che X sarà uguale a c, il quadrato dell'errore che si commetterà sarà $(X-c)^2$. Si dimostra di seguito che la media dell'errore al quadrato (detto errore quadratico medio) è minima quando c coincide con il valore atteso di X. Infatti, detta $\mu = \mathbb{E}[X]$, si ha

$$\begin{aligned}\mathbb{E}[(X-c)^2] &= \mathbb{E}[(X-\mu+\mu-c)^2] = \mathbb{E}[(X-\mu)^2 + 2(X-\mu)(\mu-c) + (\mu-c)^2] \\ &= \mathbb{E}[(X-\mu)^2] + 2(\mu-c)\,\mathbb{E}[X-\mu] + (\mu-c)^2 \\ &\overset{(1)}{=} \mathbb{E}[(X-\mu)^2] + (\mu-c)^2 \\ &\geq \mathbb{E}[(X-\mu)^2]\end{aligned}$$

(1): infatti $\mathbb{E}[X-\mu] = \mathbb{E}[X] - \mu = 0$.

Perciò la migliore previsione di X, in termini di minimizzazione dell'errore quacratico medio, è proprio il suo valore atteso.

7.4 Varianza

La varianza misura quanto i valori di una variabile aleatoria si dispergano intorno alla media. Siccome i valori di X sono distribuiti attorno al suo valore atteso, un approccio per misurare la loro variabilità potrebbe essere quantificare la loro distanza da $\mathbb{E}[X]$, ad esempio calcolando quanto valga $\mathbb{E}[|X-\mu|]$. Tuttavia, il valore assoluto comporta alcuni problemi di calcolo, e si predilige pertanto l'elevamento al quadrato.

Sia X una variabile aleatoria e sia il suo valore atteso $\mu = \mathbb{E}[X]$. La varianza di X, che si denota con $\mathrm{Var}(X)$ oppure σ^2, è (se esiste) la quantità:

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

Teorema Sia X una variabile aleatoria, allora $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Dimostrazione:

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}[X^2 - 2\mu X + \mu^2] = \mathbb{E}[X^2] - 2\mu\,\mathbb{E}[X] + \mu^2 = \mathbb{E}[X^2] - 2\mu^2 + \mu^2 = \mathbb{E}[X^2] - \mu^2$$

7.4.1 Proprietà

Proposizione Per ogni coppia di costanti reali a e b, si ha $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$.

Dimostrazione:

$$\mathrm{Var}(aX+b) = \mathbb{E}[(aX+b-(a\mu+b))^2] = \mathbb{E}[(aX-a\mu)^2] = \mathbb{E}[(a(X-\mu))^2] = a^2\,\mathbb{E}[(X-\mu)^2] = a^2\,\mathrm{Var}(X)$$

La varianza non è quindi un operatore lineare, proprio come la varianza campionaria. Si osserva che ha il quadrato dell'unità di misura della variabile aleatoria X.

Si presentano due casi:

- se $a = 0$, si ha $\mathrm{Var}(b) = 0$ e quindi la varianza di una costante è zero. Infatti, una costante è una variabile aleatoria degenere che assume un unico valore con probabilità 1.
- se $b = 0$, si ottiene che $\mathrm{Var}(aX) = a^2\,\mathrm{Var}(X)$. Di conseguenza la varianza scala al quadrato rispetto alle costanti moltiplicative.

Sia I_A la funzione indicatrice di un evento $A \subseteq \Omega$. Notando che $I_A^2 = I_A$ per idempotenza (infatti i valori possibili di I_A sono solo 0 e 1, che elevati al quadrato rimangono invariati), si ha:

$$\mathrm{Var}(I_A) = \mathbb{E}[I_A^2] - \mathbb{E}[I_A]^2 = \mathbb{E}[I_A] - \mathbb{E}[I_A]^2 = \mathbb{P}(A) - \mathbb{P}(A)^2 = \mathbb{P}(A)(1 - \mathbb{P}(A)) = \mathbb{P}(A)\,\mathbb{P}(\overline{A})$$

Deviazione standard A partire dalla varianza, è possibile deifinire la deviazione standard di una variabile aleatoria X come

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}[(X-\mu)^2]} = \sqrt{\mathbb{E}[X^2] - \mathbb{E}[X]^2}$$

La deviazione standard possiede la stessa unità di misura della variabile aleatoria presa in considerazione.

Linearità

Se il valore atteso è lineare rispetto alla somma di variabili aleatorie, in generale non si può dire lo stesso per la varianza. Infatti, ad esempio

$$\mathrm{Var}(X + X) = \mathrm{Var}(2X) = 4\,\mathrm{Var}(X) \neq \mathrm{Var}(X) + \mathrm{Var}(X)$$

Proposizione Si considerino due variabili aleatorie qualsiasi X e Y. La varianza della loro somma è data da

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$$

Dimostrazione:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= \mathbb{E}[(X+Y)^2] - \mathbb{E}[X+Y]^2 = \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X] + \mathbb{E}[Y])^2 \\
&= \mathbb{E}[X^2] + 2\mathbb{E}[XY] + \mathbb{E}[Y^2] - (\mathbb{E}[X]^2 + 2\,\mathbb{E}[X]\,\mathbb{E}[Y] + \mathbb{E}[Y]^2) \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 + 2\,\mathbb{E}[XY] - 2\,\mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)
\end{aligned}$$

Utilizzando questa formula sul caso precedente, si ottiene infatti

$$\mathrm{Var}(X + X) = \mathrm{Var}(X) + \mathrm{Var}(X) + 2\mathrm{Cov}(X, X) = 2\,\mathrm{Var}(X) + 2\,\mathrm{Var}(X) = 4\,\mathrm{Var}(X)$$

Questa formula può essere estesa a un numero finito di variabili aleatorie, ottenendo che, se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie, allora

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} \mathrm{Cov}(X_i, X_j)$$

Proposizione Se X e Y sono variabili aleatorie indipendenti, allora $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Dimostrazione:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y) \overset{(1)}{=} \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot 0 = \mathrm{Var}(X) + \mathrm{Var}(Y)$$

(1): per ipotesi di indipendenza

7.5 Covarianza

Si considerino due variabili aleatorie X e Y di valore atteso μ_X e μ_Y rispettivamente. La loro *covarianza*, che si indica con $\mathrm{Cov}(X,Y)$, è definita come, se esiste, la quantità

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

Teorema Siano X e Y due variabili aleatorie, allora $\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

Dimostrazione:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] = \mathbb{E}[XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y] \\ &= \mathbb{E}[XY] - \mu_X\mathbb{E}[Y] - \mu_Y\mathbb{E}[X] + \mu_X\mu_Y = \mathbb{E}[XY] - \mu_X\mu_Y - \cancel{\mu_Y\mu_X} + \cancel{\mu_X\mu_Y} \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}$$

7.5.1 Proprietà

Dalla definizione di covarianza si deducono le seguenti proprietà:

- $\mathrm{Cov}(X,Y) = \mathrm{Cov}(Y,X)$ simmetria
- $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$
- $\mathrm{Cov}(X+b,Y) = \mathrm{Cov}(X,Y) = \mathrm{Cov}(X,Y+b)$
- $\mathrm{Cov}(aX,Y) = a\,\mathrm{Cov}(X,Y) = \mathrm{Cov}(X,aY)$

  Dimostrazione:

  $$\mathrm{Cov}(aX,Y) = \mathbb{E}[(aX-a\mu_X)(Y-\mu_Y)] = \mathbb{E}[a(X-\mu_X)(Y-\mu_Y)] = a\,\mathrm{Cov}(X,Y)$$

Lemma Siano X,Y e Z tre variabili aleatorie, allora $\mathrm{Cov}(X+Y,Z) = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)$.

Dimostrazione:

$$\begin{aligned}\mathrm{Cov}(X+Y,Z) &= \mathbb{E}[(X+Y)Z] - \mathbb{E}[X+Y]\,\mathbb{E}[Z] = \mathbb{E}[XZ+YZ] - (\mathbb{E}[X]+\mathbb{E}[Y])\,\mathbb{E}[Z] \\ &= \mathbb{E}[XZ] + \mathbb{E}[YZ] - \mathbb{E}[X]\,\mathbb{E}[Z] - \mathbb{E}[Y]\,\mathbb{E}[Z] = \mathrm{Cov}(X,Z) + \mathrm{Cov}(Y,Z)\end{aligned}$$

Questo lemma può essere generalizzato a più di due variabili aleatorie, ottenendo che, se $X_1,\dots,X_n$ sono variabili aleatorie, allora

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, Y\right) = \sum_{i=1}^{n}\mathrm{Cov}(X_i,Y)$$

Proposizione Siano $X_1,\dots,X_n$ e $Y_1,\dots,Y_n$ variabili aleatorie qualsiasi, allora

$$\mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i,Y_j)$$

7.5.2 Indipendenza

Teorema Siano X e Y due variabili aleatorie indipendenti, allora $\text{Cov}(X,Y) = 0$.

Dimostrazione:

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \overset{(1)}{=} \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

(1): per ipotesi di indipendenza

Se due variabili aleatorie non sono indipendenti, la loro covarianza è un importante indicatore della relazione che sussiste tra loro. Si consideri la situazione in cui X e Y sono le funzioni indicatrici di due eventi $A, B \subseteq \mathcal{F}$:

$$X = I_A = \begin{cases} 1 & \text{sse } A \text{ si verifica} \\ 0 & \text{altrimenti} \end{cases} \qquad Y = I_B = \begin{cases} 1 & \text{sse } B \text{ si verifica} \\ 0 & \text{altrimenti} \end{cases}$$

Si ricorda che $\mathbb{E}[X] = \mathbb{P}(A)$ e $\mathbb{E}[Y] = \mathbb{P}(B)$. Si osserva che anche XY è una funzione indicatrice:

$$XY = I_{A\cap B} = \begin{cases} 1 & \text{sse } X = 1 \text{ e } Y = 1 \\ 0 & \text{altrimenti} \end{cases}$$

Sapendo che $\mathbb{E}[XY] = \mathbb{P}(X = 1, Y = 1) = \mathbb{P}(A \cap B)$, si ottiene che

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{P}(X = 1, Y = 1) - \mathbb{P}(X = 1)\,\mathbb{P}(Y = 1)$$

da cui si deduce che

$$\begin{aligned} \text{Cov}(X,Y) > 0 &\iff \mathbb{P}(X = 1, Y = 1) > \mathbb{P}(X = 1)\,\mathbb{P}(Y = 1) \\ &\iff \frac{\mathbb{P}(X = 1, Y = 1)}{\mathbb{P}(Y = 1)} > \mathbb{P}(X = 1) \\ &\iff \mathbb{P}(X = 1 | Y = 1) > \mathbb{P}(X = 1) \end{aligned}$$

Perciò la covarianza di X e Y è positiva se condizionando a $\{Y = 1\}$ aumenta la probabilità di $X = 1$.

Indice di correlazione lineare In generale si può dimostrare che un valore positivo di $\text{Cov}(X,Y)$ indica che X e Y tendenzialmente assumono valori grandi o piccoli contemporaneamente. La forza della relazione tra X e Y è misurata propriamente dal *coefficiente di correlazione lineare*, un numero puro che tiene conto anche delle deviazioni standard di X e Y. Viene indicato con $\text{Corr}(X,Y)$ e definito come:

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Si può dimostrare che questa quantità è sempre compresa tra -1 e +1. Valgono le seguenti affermazioni:

- $\text{Corr}(X,Y) = 1$ se X e Y sono perfettamente correlati positivamente, cioè se esiste una relazione lineare crescente tra X e Y.
- $\text{Corr}(X,Y) = -1$ se X e Y sono perfettamente correlati negativamente, cioè se esiste una relazione lineare decrescente tra X e Y.
- $\text{Corr}(X,Y) = 0$ se X e Y sono incorrelati, cioè se non esiste alcuna relazione lineare tra X e Y. Ciò non implica che X e Y siano indipendenti, in quanto potrebbero esistere relazioni non lineari che questo coefficiente non è in grado di cogliere.

7.6 Disuguaglianze

L'importanza delle disuguaglianze di Markov e Chebyshev, che verranno presentate di seguito, sta nel fatto che permettono di limitare le probabilità di eventi rari che riguardano variabili aleatorie di cui si conosce solo il valore atteso oppure il valore atteso e la varianza. Naturalmente, quando la distribuzione è nota, è possibile calcolare queste probabilità esattamente e non vi è quindi la necesittà di ridursi alle disuguaglianze.

7.6.1 Disuguaglianza di Markov

Sia $X \geq 0$ una variabile aleatoria qualsiasi a specificazioni non negative, allora

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \quad \forall a > 0$$

Dimostrazione:

Si mostra la dimostrazione per il caso discreto, ma il ragionamento è analogo nel caso continuo.

$$\mathbb{E}[X] = \sum_{x \in D_X} x\,\mathbb{P}(X = x) = \underbrace{\sum_{x<a} x\,\mathbb{P}(X = x)}_{\geq 0 \text{ per ipotesi}} + \sum_{x \geq a} x\,\mathbb{P}(X = x)$$

$$\overset{(1)}{\geq} \sum_{x \geq a} x\,\mathbb{P}(X = x) \overset{(2)}{\geq} \sum_{x \geq a} a\,\mathbb{P}(X = x) = a \underbrace{\sum_{x \geq a} \mathbb{P}(X = x)}_{\text{eventi disgiunti}}$$

$$\overset{K3}{=} a\,\mathbb{P}(X \geq a)$$

(1): perché il primo addendo è positivo
(2): perché $x \geq a$ nella sommatoria

Si è quindi dimostrato che $\mathbb{E}[X] \geq a\,\mathbb{P}(X \geq a) \Rightarrow \mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$, provando di fatto la tesi.

Si osserva che è possibile utilizzare questa disuguaglianza anche nel verso opposto, ossia

$$\mathbb{P}(X < a) = 1 - \mathbb{P}(X \geq a) \geq 1 - \frac{\mathbb{E}[X]}{a} \quad \forall a > 0$$

7.6.2 Disuguaglianza di Chebyshev

Sia X una variabile aleatoria qualsiasi con media μ e varianza σ^2, allora

$$\mathbb{P}(|X - \mu| \geq r) \leq \frac{\sigma^2}{r^2} \quad \forall r > 0$$

Dimostrazione:

Si osservi che gli eventi $\{|X - \mu| \geq r\}$ e $\{(|X - \mu|)^2 \geq r^2\}$ si implicano a vicenda e sono quindi equiprobabili:

$$\mathbb{P}(|X - \mu| \geq r) = \mathbb{P}((X - \mu)^2 \geq r^2)$$

Si consideri perciò la variabile aleatoria $Y = (X - \mu)^2$. Essendo le sue specificazioni non negative, è possibile applicarle la disuguaglianza di Markov con $a = r^2$:

$$\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a} \quad \forall a > 0$$

$$\Rightarrow \mathbb{P}((X - \mu)^2 \geq r^2) \leq \frac{\mathbb{E}[(X - \mu)^2]}{r^2} = \frac{\sigma^2}{r^2}$$

La disuguaglianza finale implica a sua volta che $\mathbb{P}(|X-\mu| \geq r) \leq \dfrac{\sigma^2}{r^2}$, provando di fatto la tesi.

Si osserva che è possibile utilizzare questa disuguaglianza anche nel verso opposto, ossia

$$\mathbb{P}(|X-\mu| < r) = 1 - \mathbb{P}(|X-\mu| \geq r) \geq 1 - \frac{\sigma^2}{r^2} \quad \forall r > 0$$

Inoltre, se nella disugualianza di Chebyshev si pone $r = k\sigma$, essa assume la seguente forma:

$$\mathbb{P}(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \forall k > 0$$

In altri termini, la probabilità che una variabile aleatoria differisca dalla sua media per più di k volte la deviazione standard è al più $\dfrac{1}{k^2}$.

8 Modelli di distribuzione

In ambito statistico, un modello di distribuzione è una rappresentazione teorica che specifica, attraverso una funzione di probabilità (discreta) o di densità (continua), la legge con cui una variabile aleatoria assume i propri valori nello spazio campionario. L'adozione di tale modello consente di trascendere l'osservazione empirica di una singola realizzazione o di un campione finito, attribuendo invece alla variabile un comportamento stocastico stabilito a priori mediante assunzioni strutturate.

In sintesi, il ragionamento si fonda su tre passaggi essenziali. Anzitutto si introduce la variabile aleatoria X, intesa come funzione misurabile che collega ogni evento elementare ω dello spazio campionario Ω a un numero reale. A questa funzione si associa poi la sua distribuzione: una misura di probabilità, indicata con P_X, che a ogni insieme boreliano B assegna la probabilità che X vi cada, ossia $P(X \in B)$; tale legge può essere descritta tramite funzione di ripartizione, massa di probabilità o densità. Infine, adottare un modello di distribuzione significa disporre di un linguaggio per calcolare momenti, quantili e altre probabilità rilevanti, così da fornire la base matematica per inferenze, verifiche di ipotesi e previsioni sul fenomeno in esame.

8.1 Modelli discreti

Quando il fenomeno osservato può assumere solo un insieme numerabile di valori, parliamo di distribuzioni discrete. Ciascun modello discreto è descritto da una funzione di massa di probabilità che assegna a ogni possibile valore k la probabilità $P(X = k)$.

8.1.1 Modello di Bernoulli

Si supponga che venga realizzato un esperimento di Bernoulli, ossia un esperimento che può avere solo due esiti possibili, positivo e negativo. Si definisce una variabile aleatoria X in modo tale che $X = 1$ nel primo caso e $X = 0$ nel secondo: il supporto è quindi $D_X = \{0, 1\}$.

Per identificare univocamente una distribuzione basta conoscerne la funzione di massa di probabilità, che in questo caso è definita come

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

dove con p si indica la probabilità che l'esperimento registri un successo. Deve essere ovviamente $0 \leq p \leq 1$.

Una variabile aleatoria X si dice di Bernoulli con parametro $p \in [0, 1]$ e si indica con $X \sim B(p)$ se la sua funzione di massa di probabilità è definita come sopra. In altri termini, X si dice bernoulliana se può assumere solo i valori 0 e 1.

È possibile definire più formalmente la funzione di massa di probabilità come

$$p_X(x) = p^x (1 - p)^{1-x} \, I_{\{0,1\}}(x)$$

Per essere una funzione di massa di probabilità, $p_X(x)$ deve soddisfare le seguenti condizioni:

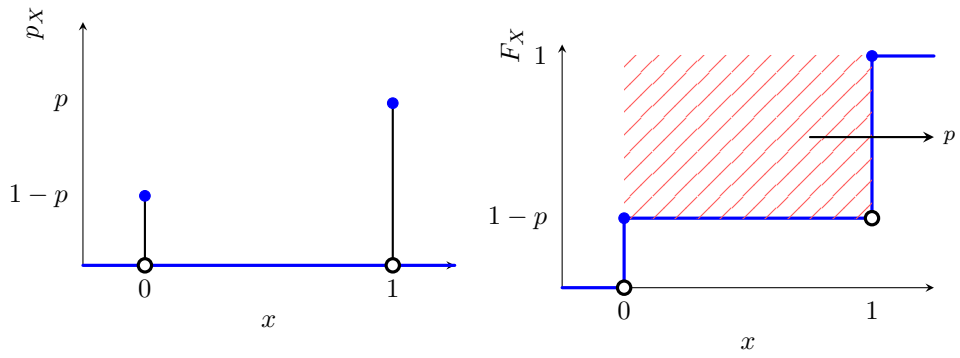
- $p_X(x) \geq 0$ per ogni $x \in \mathbb{R}$;
- $\sum_{x \in D_X} p_X(x) = 1$.

La prima condizione è soddisfatta per ogni $x \in \mathbb{R}$, mentre la seconda condizione è verificata come segue:

$$\sum_{x \in D_X} p_X(x) = p^1 (1 - p)^{1-1} + p^0 (1 - p)^{1-0} = p + (1 - p) = 1$$

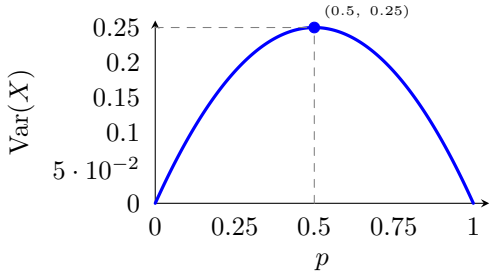
La funzione di ripartizione di una variabile aleatoria bernoulliana è definita come

$$F_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ 1-p & \text{se } 0 \le x < 1 \\ 1 & \text{se } x \ge 1 \end{cases}$$



Il suo valore atteso è dato da $\mathbb{E}[X] = 1 \cdot p + 0 \cdot (1-p) = p$ ed è quindi pari alla probabilità che la variabile aleatoria assuma il valore 1.

La varianza è $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \overset{(1)}{=} \mathbb{E}[X] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$ (1): per idempotenza



Si osserva che nel caso si scelga $p = 0$ oppure $p = 1$, la variabile aleatoria assume il valore 0 o 1 con probabilità 1, rispettivamente. In questo caso si ha una variabile aleatoria degenere e la varianza è nulla.

8.1.2 Modello binomiale

Si supponga di realizzare $n \in \mathbb{N}$ ripetizioni indipendenti di un esperimento bernoulliano di parametro p. Se X denota il numero totale di successi, allora X si dice variabile aleatoria binomiale con parametri n e p e si indica con $X \sim B(n,p)$.

Formalmente, dati $n \in \mathbb{N}$ e $X_1, \dots, X_n \sim B(p)$ variabili aleatorie indipendenti, si definisce una variabile aleatoria X binomiale come

$$X = \sum_{i=1}^{n} X_i \sim B(n,p)$$

dove X_i è la funzione indicatrice del successo dell'i-esimo esperimento:

$$X_i = \begin{cases} 1 & \text{se la prova } i\text{-esima ha successo} \\ 0 & \text{altrimenti} \end{cases}$$

Il dominio di supporto è quindi $D_X = \{0, 1, \dots, n\}$. La funzione di massa di probabilità per una variabile aleatoria binomiale di parametri (n,p) è data da

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \, I_{\{0,\dots,n\}}(x)$$

dove il coefficiente binomiale

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

rappresenta il numero di combinazioni differenti che si possono ottenere scegliendo x successi tra n prove.

La correttezza della funzione di massa di probabilità è garantita dalle seguenti condizioni:

- $p_X(x) \geq 0$ per ogni $x \in \mathbb{R}$;
- $\sum_{x \in D_X} p_X(x) = 1$.

La prima condizione è soddisfatta per ogni $x \in \mathbb{R}$, mentre la seconda condizione è verificata come segue:

$$\sum_{x \in D_X} p_X(x) = \sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} \overset{(1)}{=} (p + (1-p))^n = 1$$

(1): per la formula delle potenze del binomio $(x+y)^n = \sum_{i=0}^{n} \binom{n}{i} x^i \, y^{n-i}$

La funzione di ripartizione di una variabile aleatoria binomiale è definita come

$$F_X(x) = \mathbb{P}(X \leq \lfloor x \rfloor) = \sum_{i=0}^{\lfloor x \rfloor} p_X(i)$$

Grafici di aggiungere

Si osserva che indipendetemente dal valore di p, il grafico di p_X cresce e poi decresce in maniera simmetrica. Ciò che p stabilisce è se questo grafico sia spostato verso 0 o verso n, il che è ragionevole in quanto, come si vedrà ora, il valore atteso di X è np.

Per definizione di variabile aleatoria binomiale, essa è la somma di n variabili aleatorie bernoulliane, ognuna delle quali ha valore atteso p. Sfruttando la linearità del valore atteso, si ottiene

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p = n\,p$$

Ragionamento analogo si può fare per la varianza, notando che è possibile sfruttare l'indipendenza tra le variabili bernoulliane X_i per evitare il termine di covarianza:

$$\mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) = \sum_{i=1}^{n} p(1-p) = n\,p(1-p)$$

Come nel modello di Bernoulli, anche in questo caso la varianza assume il grafico di una parabola. La varianza aumenta sia quando p tende a $1/2$, ma anche all'aumentare di n.

Riproducibilità

Si considerino due variabili aleatorie $X \sim B(n,p)$ e $Y \sim B(m,p)$ indipendenti tra loro. Queste due variabili non seguono la stessa distribuzione, ma sono correlate in quanto seguono lo stesso modello di distribuzione e, in aggiunta, condividono lo stesso parametro p.

La somma di due variabili aleatorie binomiali con lo stesso parametro p è una variabile aleatoria binomiale con parametri $n+m$ e p:

$$X + Y = \sum_{i=1}^{n} X_i + \sum_{i=1}^{m} Y_i = \sum_{i=1}^{n+m} Z_i \ \sim B(n+m,p)$$

Si osserva che $\forall i = 1, \dots, n \quad Z_i = X_i$ e $\forall i = 1, \dots, m \quad Z_{n+i} = Y_i$.

Si dice perciò che questo modello gode della proprietà di riproducibilità tra variabili binomiali con lo stesso parametro p.

8.1.3 Modello uniforme discreto

Questo modello si presenta quando l'esperimento casuale può restituire n esiti distinti, ciascuno dei quali ha la stessa probabilità di verificarsi. Una variabile aleatoria X che codifica tale esperimento si dice uniforme discreta con parametro n e si indica con $X \sim U(n)$.

Il suo dominio di supporto è quindi $D_X = \{1, \dots, n\}$ e $\forall i \in D_X \;\; \mathbb{P}(X = i) = 1/n$.

La funzione di massa di probabilità è definita come

$$p_X(x) = \frac{1}{n} I_{\{1,\dots,n\}}(x)$$

La correttezza della funzione di massa di probabilità è banale, in quanto

$$\sum_{x \in D_X} p_X(x) = \sum_{i=1}^{n} \frac{1}{n} = 1$$

La funzione di ripartizione di una variabile aleatoria uniforme discreta è definita come

$$F_X(x) = \mathbb{P}(X \leq x) = \sum_{i \leq x} p_X(i) = \sum_{i=1}^{\lfloor x \rfloor} p_X(i) = \sum_{i=1}^{\lfloor x \rfloor} \frac{1}{n} = \frac{\lfloor x \rfloor}{n}$$

Il valore atteso è dato da

$$\mathbb{E}[X] = \sum_{x=1}^{n} x\, p_X(x) = \sum_{x=1}^{n} \frac{x}{n} = \frac{1}{n} \sum_{x=1}^{n} x = \frac{1}{n} \frac{n(n+1)}{2} = \frac{n+1}{2}$$

La varianza invece si calcola come

$$\mathbb{E}[X^2] = \sum_{x=1}^{n} x^2\, p_X(x) = \sum_{x=1}^{n} \frac{x^2}{n} = \frac{1}{n} \sum_{x=1}^{n} x^2 = \frac{1}{n} \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}$$

$$\begin{aligned} \mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= \frac{(n+1)(2n+1)}{6} - \frac{(n+1)^2}{4} = (n+1)\left(\frac{2n+1}{6} - \frac{n+1}{4}\right) \\ &= (n+1)\left(\frac{4n+2-3n-3}{12}\right) = \frac{(n+1)(n-1)}{12} = \frac{(n^2-1)}{12} \end{aligned}$$

8.1.4 Modello geometrico

Si supponga di realizzare un esperimento di Bernoulli con parametro p e di contare il numero di prove necessarie affinché si verifichi il primo successo. La variabile aleatoria X che codifica questo esperimento si dice geometrica con parametro p e si indica con $X \sim G(p)$.
X quindi conta il numero di fallimenti che precedono il primo successo.

Formalmente, dati $X_1, X_2, \dots$ variabili aleatorie indipendenti e identicamente distribuite (i.i.d.) di Bernoulli con parametro p, si definisce una variabile aleatoria X geometrica come

$$X = \sum_{i=1}^{+\infty} X_i \cdot I_{\{X_i=0\}} \sim G(p)$$

Si osserva che se $p = 1$ allora $X = 0$ con probabilità 1, mentre se $p = 0$ allora $X \to +\infty$ con probabilità 1. In entrambi i casi la variabile aleatoria è degenere. Si considera per tale motivo $p \in (0, 1]$.

Funzione di massa di probabilità

Il dominio di supporto è $D_X = \mathbb{N} \cup \{0\}$. La funzione di massa di probabilità è definita come

$$p_X(x) = p(1-p)^x \, I_{\mathbb{N}\cup\{0\}}(x)$$

Infatti calcolare $p_X(x) = \mathbb{P}(X = x)$ equivale a calcolare la probabilità che i primi x esperimenti siano fattili e che il $(x+1)$-esimo sia un successo, ossia:

$$\mathbb{P}(X = x) = \mathbb{P}\left(\bigcap_{i=0}^{x} X_i = 0 \cap X_{x+1} = 1\right) \overset{(1)}{=} \prod_{i=0}^{x} \mathbb{P}(X_i = 0) \cdot \mathbb{P}(X_{x+1} = 1) = \prod_{i=0}^{x} (1-p) \cdot p = p\,(1-p)^x$$

(1): per indipendenza

La correttezza della funzione di massa di probabilità è così dimostrata:

$$\sum_{x=0}^{+\infty} = \sum_{x=0}^{+\infty} p(1-p)^x = p \sum_{x=0}^{+\infty} (1-p)^x \overset{(1)}{=} p\,\frac{1}{1-(1-p)} = p\,\frac{1}{p} = 1$$

(1): per la formula della somma geometrica: $\displaystyle\sum_{i=0}^{+\infty} \alpha^i$ converge a $\dfrac{1}{1-\alpha}$ per $-1 < \alpha < 1$. Nel nostro caso $\alpha = p$, che è una probabilità ed è quindi compresa tra 0 e 1.

Il grafico della funzione di massa di probabilità presenta un decadimento esponenziale, ed è tanto più ripido quanto più è grande il valore di p.

Funzione di ripartizione

Prima di calcolare la funzione di ripartizione, si osservi che

$$\mathbb{P}(X > n) = \sum_{x=n+1}^{+\infty} p_X(x) = \sum_{x=n+1}^{+\infty} p\,(1-p)^x = p\,(1-p)^{n+1} \sum_{x=n+1}^{+\infty} (1-p)^{x-(n+1)}$$

$$\overset{(1)}{=} p\,(1-p)^{n+1} \sum_{y=0}^{+\infty} (1-p)^y \overset{(2)}{=} p\,(1-p)^{n+1} \cdot \frac{1}{1-(1-p)} = p\,(1-p)^{n+1} \cdot \frac{1}{p} = (1-p)^{n+1}$$

(1): ponendo $y = x - (n+1)$

(2): per la formula della serie geometrica

Ora è possibile calcolare la funzione di ripartizione:

$$F_X(n) = \mathbb{P}(X \le n) = 1 - \mathbb{P}(X > n) = 1 - (1-p)^{n+1} \qquad \forall n \in \mathbb{N} \cup \{0\}$$

$$F_X(x) = \mathbb{P}(X \le x) = \mathbb{P}(X \le \lfloor x \rfloor) = 1 - (1-p)^{\lfloor x \rfloor + 1} \qquad \forall x \in \mathbb{R}$$

Assenza di memoria L'assenza di memoria è una proprietà di cui gode il modello geometrico, ed è l'unico a possederla tra i vari modelli discreti. Si consideri $X \sim G(p)$ e si provi a calcolare $\mathbb{P}(X > i+j \mid X \geq i)$, ovvero la probabilità condizionata di ottenere un successo dopo $i+j$ prove, sapendo che ne sono già state effettuate i senza successo. Si ha:

$$\mathbb{P}(X \geq n) = \mathbb{P}(X > n-1) = (1-p)^n$$

$$\mathbb{P}(X \geq i+j \mid X \geq i) = \frac{\mathbb{P}(X \geq i+j, X \geq i)}{\mathbb{P}(X \geq i)} = \frac{\mathbb{P}(X \geq i+j)}{\mathbb{P}(X \geq i)} = \frac{(1-p)^{i+j}}{(1-p)^i} = (1-p)^j = \mathbb{P}(X \geq j)$$

Si ottiene quindi che $\mathbb{P}(X > i+j \mid X \geq i) = \mathbb{P}(X > j)$, ossia la probabilità di ottenere un successo dopo $i+j$ prove, sapendo che ne sono già state effettuate i senza successo, è uguale alla probabilità di ottenere un successo dopo j prove. In altre parole, il numero di prove necessarie per ottenere il primo successo non dipende da quante prove siano già state effettuate.

Valore atteso e varianza

Prima di calcolare il valore atteso, si osservi che

$$\sum_{i=0}^{+\infty} i\alpha^i = \alpha \sum_{i=0}^{+\infty} i\alpha^{i-1} \overset{(1)}{=} \alpha \sum_{i=1}^{+\infty} \frac{d}{d\alpha}\alpha^i = \alpha \frac{d}{d\alpha}\left[\sum_{i=0}^{+\infty} \alpha^i\right] \overset{(2)}{=} \alpha \frac{d}{d\alpha}\left(\frac{1}{1-\alpha}\right) \overset{(3)}{=} \alpha \frac{1}{(1-\alpha)^2} = \frac{\alpha}{(1-\alpha)^2}$$

(1): $\dfrac{d}{dx}x^i = i\,x^{i-1}$

(2): per la formula della serie geometrica, considerando $|\alpha| < 1$

(3): $\dfrac{d}{dx}\dfrac{1}{x} = -\dfrac{1}{x^2}$

Ora è possibile calcolare il valore atteso:

$$\mathbb{E}[X] = \sum_{x=0}^{+\infty} x\, p_X(x) = \sum_{x=0}^{+\infty} x\, p(1-p)^x = p\sum_{x=0}^{+\infty} x(1-p)^x \overset{(1)}{=} p\frac{1-p}{[1-(1-p)]^2} = p\frac{1-p}{p^2} = \frac{1-p}{p}$$

(1): per l'osservazione appena fatta, che si può utilizzare in quanto $0 < 1-p < 1$ e quindi la serie converge

La varianza è data da:

$$\mathbb{E}[X^2] = \sum_{x=0}^{+\infty} x^2\, p(1-p)^x = p(1-p)\sum_{x=0}^{+\infty} x^2(1-p)^{x-1} \overset{(1)}{=} p(1-p)\sum_{x=0}^{+\infty} \frac{d}{dp}\left[-x(1-p)^x\right]$$

$$= p(1-p)\frac{d}{dp}\left[\sum_{x=0}^{+\infty} -x(1-p)^x\right] = -p(1-p)\frac{d}{dp}\left[\sum_{x=0}^{+\infty} x(1-p)^x\right] \overset{(2)}{=} -p(1-p)\frac{d}{dp}\left(\frac{1-p}{p^2}\right)$$

$$= -\not{p}(1-p)\frac{-p^2 - 2p(1-p)}{p^{\not 3}} = -(1-p)\frac{-p-2(1-p)}{p^2} = (1-p)\frac{p+2-2p}{p^2} = \frac{(1-p)(2-p)}{p^2}$$

(1): $\dfrac{d}{dx}x(1-p)^x = -x^2(1-p)^{x-1} \quad \Rightarrow \quad \dfrac{d}{dx}[-x(1-p)^x] = x^2(1-p)^{x-1}$

(2): per l'osservazione fatta in precedenza

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(1-p)(2-p)}{p^2} - \left(\frac{1-p}{p}\right)^2 = \frac{(1-p)(2-p)-(1-p)^2}{p^2}$$

$$= \frac{(1-p)(2-\not{p}-1+\not{p})}{p^2} = \frac{1-p}{p^2}$$

Parte III

Statistica inferenziale

9 Analisi della varianza

To do (lezione 08)