Riassunto Statistica

Dimix

2025-03-14

Indice

Capitolo 1 – Introduzione alla Statistica

Definizione

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

La **statistica descrittiva** è quella parte della statistica che si occupa di descrivere e riassumere i dati

La **statistica inferenziale** è quella parte della statistica che si occupa di trarre conclusione dai dati

La statistica inferenziale si basa sul modello probabilistico che consiste nel fare un insieme di assunzioni sulle probabilità di ottenere un certo valore, per cui essa 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.

Popolazione e campioni

Nella statistica è cruciale ottenere delle informazioni su tutto un insieme di elementi, che vengono definiti **popolazione**. Spesso la popolazione pero 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**

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

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

Una volta scelto il campione casuale, si può utilizzare l'inferenza statistica per giungere a conclusioni sull'intera popolazione studiando gli elementi del campione.

Campione casuale stratificato

Un metodo più sofisticato del campionamento casuale semplice è il **campionamento casuale stratificato**. inizialmente si stratifica la popolazione in *sottopopolazioni*, ognuna dalle 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 ciascun strato presenti nel campione rispecchiano esattamente quelle dell'intera popolazione.

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

Statistica descrittiva

Capitolo 2 - Descrivere insiemi di dati

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 dei dati qualitativi

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

- Dati binari o booleani: possono assumere soltanto due valori tra loro non confrontabili, si utilizza *booleani* per evidenziare la presenza/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 stabilire una relazione di equivalenza.
- Dati ordinali: se abbiamo due valori diversi riusciamo a stabilire quale sia il più piccolo e quale il più grande, quindi esiste una relazione d'ordine tra i valori.

Classificazione dei dati quantitativi

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

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

Frequenze

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.

In modo formale possiamo indicarla con f_x la frequenza assoluta di x, si ha che $f_x = \#\{j \in \{1, \cdots, n\} \mid x_j = x\}$

La **frequenza relativa** consente di esprimere la presenza di ogni valori 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, possiamo definire frequenza relativa di x_i il valore f_i/n.

Teniamo presente che la somma di tutte le frequenze relative in un campione è sempre uguale ad 1.

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 termo la somma delle frequenze dei primi tre e cosi via.

l'ultima frequenza cumulata rappresenta il totale dei casi osservati. inoltre possiamo applicare il concetto di frequenza cumulata sia alle frequenze assolute che a quelle relative, in caso di frequenze relative i valori cumulati variano da 0 a 1.

Quando i dati sono numerici o comunque ordinabili, un concetto affine alle frequenze cumulate è quello della **funzione cumulativa empirica**, nota anche come funzione di ripartizione empirica. Data una serie di osservazioni $x_1, \cdots, x_n$, la funzione empirica $\widehat{F} : \mathbb{R} \to [0, 1]$ è definita in modo che per ogni $x \in \mathbb{R}$ essa assume 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 o 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 grafo di questa funzione sarà a tratti costanti.

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

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

Le **frequenze congiunte e marginali**, quando si analizza un insieme di osservazioni, può essere 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 occorenze in cui i due caratteri assumono contemporaneamente determinati valori. Questo conteggio porta alla definizione di *frequenza congiunta assoluta*, se invece si considera una frazione delle osservazioni, si parla di *frequenza congiunta relativa*.

Se il numero dei possibili valori osservabili per i caratteri non è elevato, possiamo rappresentare visivamente queste frequenze tramite una *tabella delle frequenze congiunte* o *tabella di contigenza*. in questa 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 indicazione le frequenze congiunte 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.

Grafici

La **simmetria**, un insieme di dati è detto 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 viene detto **centro di simmetria** della distribuzione.

La **quasi simmetria** si presenta quando i dati non sono permettamente simmetrici, ma la distribuzione rimane quasi specuplare rispetto a un punto centrale.

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

Grafici per la frequenza

Se l'insieme di dati contiene un numero ridotto di valori distinti allora è rappresentabile 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 statitisca $\boldsymbol{X}$ che può assumere vari valori, si elencano i valori distinti di $\boldsymbol{X}$ in una colonna e, a fianco di ognuno, la relativa frequenza di occorenza nel campione.

Per costruire la 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 totalte di osservazioni nell'insieme di dati.

Grafici a bastoncini, a barre e poligonali

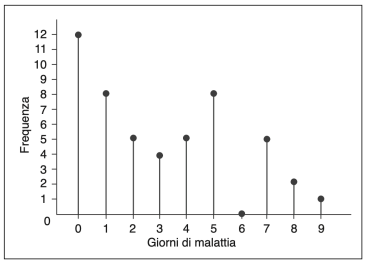


Figura 2.1 Un grafico a bastoncini.

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.

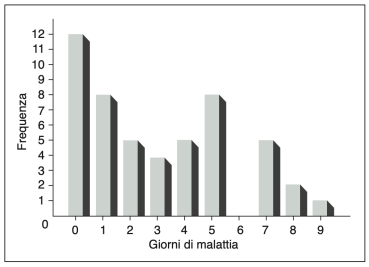


Figura 2.2 Un grafico a barre.

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.

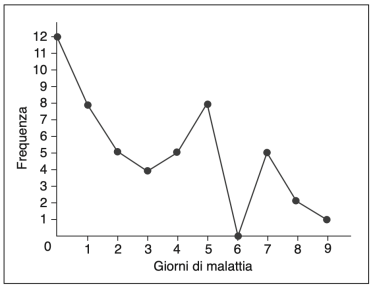


Figura 2.3 Un grafico poligonale.

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.

Diagramma a torta

In caso di dati non sono numerici si utilizza un diagramma a torta, esso consiste in 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.

Diagramma ramo-foglia

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575
26	183, 894, 982
27	671, 711, 744
28	345, 764, 913, 967

Tabella 1: Diagramma a stelo

Un modo efficiente di rappresentare un insieme di dati di dimensioni medie consiste nell'utizzare 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.

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.

Diagramma di Pareto

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

Istogrammi e raggruppamento dei dati

Utilizzare i grafici precedenti è un metodo efficace per descrivere un insieme di dati, ma alcuni di questi insiemi hanno troppi valori distinti per poter usare questo metodo. Perciò è necessario suddividere i valori in gruppi, o classi, e poi rappresentare con un grafico il numero di valori dei dati che cadono in ciascuna classe. Il numero di classi scelte è un compromesso tra:

- Scegliere poche classi al costo di perdere molte informazioni sui valori effettivi in una classe.
- Scegliere troppe classi, ottenendo 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, cioè una classe include 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.

Diagramma di dispersione e 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. Cosi facendo però, nonostante i due grafici ci diano molter informazioni sulle singole variabili (attributi), non si ha nessun tipo di informazione riguardo al rapporo tra queste due variabili.

Per capirne la relazione è necessario considerare i valori accoppiati di ciascun dato simultaneamente. Si possono allora rappresentare questi dati accopiati in un diagramma rettangolare e bidimensionale, in cui l'asse x rappresenta il valore x dei dati, e l'asse y il valore y. Cosi facendo si ottiene un **digramma 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 trammare una **retta media** che approssimi l'andamento dei punti sul diagramma, ovvero una retta che passi *il più vicino possibile* a tutti i dati.

- in pratica, si ricorre a metodi di regressione lineare, come il *metodo dei minimi quadrati*, che permette di trovare l'equazione della retta (del tipo $y = a + bx$) minimizzando la somma delle distanze (al quadrato) tra i valori osservati (x_i, y_i) e i valori $\widehat{y}_i$ previsti dalla retta. Una volta trovata questa retta di miglior adattamento, per un qualunque valore x che possa presentarsi in futuro, si ottiene la stima di y semplicemente sostituiendo x nella equazione $y = a + bx$.

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 invece siano causati da errori nella raccolta dati.

Capitolo 3 - Statistiche

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

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 e tre i casi si parla di campionaria, in quanto vengono effettuare su dei campioni.

Media campionaria

Dato 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, cioè la media aritmetica dei valori dati:

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

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

Trasformazioni

La **traslazione** dello stesso insieme di dati, si verifica quando 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 degli y_i e degli x_i.

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

La **scalatura** è quando 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: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} ax_i = a\frac{1}{n}\sum_{i=1}^{n} x_i = a\bar{x}$

La **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 sono lineari.

Media pesata

Quindi se abbiamo un insieme di dati organizzato in una tabella delle frequenze, la media campionara può essere calcolata moltiplicando ciascun valore distino per la sua frequenza, sommando tutti questi prodotti e poi dividendo il risultato per il numero totale di osservazioni.

In modo formale, supponiamo di avere k valori distinti $x_1, x_2, \cdots, x_k$ con frequenze corrispondenti $f_1, f_2, \cdots, f_k$. Il numero totale di osservazioni è dato da: $n = \sum_{i=1}^{k} f_i$.

La media campionaria $\bar{x}$ è quindi calcolata come:

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

Ora, consideriamo $w_1, w_2, \cdots, w_k$ numeri non negativi la cui somma è pari ad 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. Quindi scriviamo 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 perso assegnato al valore x_i è f_i/n, ossia rappresenta la frazione di volte in cui il valore x_i compare nell'insieme dei dati.

Scarti

Si supponga che l'insieme di dati sia costituito dagli n valori $x_1, x_2, \cdots, x_n$ 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, dato che:

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

Questa uguaglianza afferma che la somma degli scarti positivi e la somma degli scarti negativi della media campionaria si controbilanciano.

Detto tramite un linguaggio fisico, questo significa che se n pesi dotati delal 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 equilibri. Questo punto di equilibrio viene detto centro di gravità

Mediana campionaria

Le media campionaria presenta un notevole punto debole come indicatore del centro di un insieme di dati, il suo valore infatti è ampliamente influenzato da eventuali valori fuori scala.

Per calcolare la **mediana campionaria** andiamo a disporre il valore dei dati in ordine crescente. Se il numero di valori è dispari, allora il valore intermedio della lista ordinata è la mediana campionaria, mentre se è pari, la media dei due valori centrali è la mediana campionaria.

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 prende in considerazione tutti i valori dell'insieme di dati, mentre la mediana campionaria considera solo uno o due valori centrali e quindi non è influenzata dai valori fuori scala.

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

Percentili campionari

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

Per calcolare il percentile è necessario definire un ordinamento sulle osservazioni.

il **100p-esimo percentile campionario** è un valore maggiore o uguale di almeno 100p percento dei valori dati, e minore o guale 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 *np+1* soddisfano i due requisiti, 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*, e 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% compreso tra il primo e il secondo, il 25% tra il secondo e il terzo e il restante 25% sia maggiore del terzo quartile.

Moda campionaria

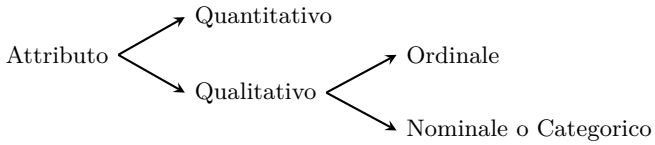
Un ulteriore indicatore della tendenza centrale è la moda campionaria, che è il valore che si verifica con maggiore frequenza nell'insieme di dati.

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

Tramite una tabella delle frequenze riusciamo a dedurre facilmente questi valori, sono quelli con la maggiore frequenza.

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 in caso di cardinalità del campione dispari, se la cardinalità è pari sarà necessaria la media e quindi possibili solo su attributi quantitativi, la moda si può fare per qualsiasi tipo di attributo.

Dispersione

Due campioni A e B possono presentare la stessa centralità ma essere molto diversi tra loro, per esempio:

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

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

Un modo per misurare la dispersione dei dati è quello di considerare gli scarti dei valori dei dati rispetto ad un valore centrale. Il valore centrale più usato per questo scopo è la media campionaria. Se i valori dei dati sono $x_1, \cdots, x_n$ 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 sappiamo che la $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, questo significa che la somma degli scarti rispetto alla media campionaria è sempre uguale a 0 e quindi anche la media aritmetica degli scarti lo sarà.

Per fare in modo che non valga 0, andiamo a considerare i singoli scarti indipendentemente dal segno, questo possiamo farlo considerando il valore assoluto degli scari oppure il quadrato.

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

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

La formula 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

La **traslazione** cioè quando a ciascun valore dei dati viene sommata una costante b 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 alla formula (3.2) per semplificare il calcolo della varianza campionaria.*

La **scalatura** cioè quando ciascun valore dei dati viene moltiplicato per a, la varianza campionaria viene moltiplicata per il quadrato di a:

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

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

La **combinazione** consiste nella combinazione delle due trasformazioni precedenti:

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

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

La **traslazione** cioè la somma di 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:

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

La **scalatura** cioè quando 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:

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

La **combinazione** consiste nella combinazione delle due trasformazioni precedenti:

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

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**.

Scarto interquantile

Lo scarto interquantile 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 concentrata attorno alla mediana, mentre un IQR ampio suggerisce una maggiore dispersione nella parte centrale della distribuzione.

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

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

Indici di dipendenza

Consideriamo un insieme composto da dati accoppiati $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Per vedere la relazione relativa di queste due variabili è possibile rappresentarle in un diagramma di dispersione. Questo approccio è pero qualitativo e quindi soggetto a interpretazione. Vogliamo 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.

Covarianza campionaria

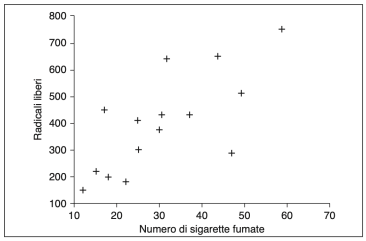
La **covarianza campionaria** è una statistica che quantifica in che misura grandi valori di x corrispondano a grandi valori di y e piccoli valori di x a piccoli valori di y. Questo indice quindi misura la tendenza con cui due variabili si muovono insieme ed è definita come la media dei prodotti degli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

Relazione tendenziale

Procediamo considerando una **relazione** di tipo **tendenziale** e non deterministico. Ciò significa che le affermazioni che seguiranno varrano tendenzialmente sempre: ci saranno quindi delle eccezioni ma per lo più saranno valide.

Si supponga che un insieme sia composto dalle coppie di valori (x_i, y_i) con $i = 1, \cdots, n$ e calcolando le rispettive medie campionarie $\bar{x}$ e $\bar{y}$. Per la i-esima coppia di dati, consideriamo $(x_i - \bar{x})$ e $(y_i - \bar{y})$ cioè gli scarti dei valori x, y rispetto alla loro media campionaria.

Possiamo notare che quando grandi valori di x tendono ad essere associati a grandi valori di y, e piccoli valori di x tendono ad 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. Quindi se gli scarti hanno segno concorde, il loro prodotto $(x_i - \bar{x})(y_i - \bar{y})$ sarà positivo. Otteniamo che la sommatoria $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$ tenderà ad 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.

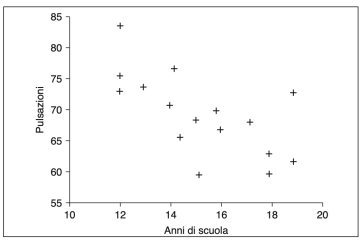


Diagramma di dispersione degli anni di scuola e delle pulsazioni.

x "grande" e y "piccola"	x "piccolo" e y "grande"
$x \geq \bar{x}$	$x < \bar{x}$
$(x_i - \bar{x}) \geq 0$	$(x_i - \bar{x}) < 0$
$y < \bar{y}$	$y \geq \bar{y}$
$(y_i - \bar{y}) < 0$	$(y_i - \bar{y}) \geq 0$
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.

Andiamo ad standardizzare la sommatoria dividendo per n-1, al fine di evitare che questo indice assuma valori troppo elevati. Possiamo osservare che la formula della covarianza campionaria e riconducibile a quella della varianza campionaria, motivo per il quale si possa intuire perchè si vada a dividere per n-1 e non direttamente per il numero totale di osservazioni (*correzzione sottostima*).

Infine possiamo definire la covarianza campionaria come:

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

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.

Possiamo standardizzare il valore della covarianza campionaria 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, mentre quando $\rho < 0$ sono correlati negativamente. Non dipendendo dalle unità di misura, questo indice può essere usato per comparare dataset diversi.

Una proprietà importante è che vale $-1 \leq \rho \leq 1$.

Relazione deterministica

Come **primo caso** vogliamo passare da una relazione tendenziale a una deterministica, nella quale 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 della retta che lega le due variabili e indica di quanto varia y all'aumentare di x. possiamo aspettarci:

- 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 un decremento di $y \Rightarrow$ relazione inversa

Si calcoli ora il coefficiente di correlazione di Pearson:

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

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 ottenute con i calcoli rispecchiano le nostre attese iniziali.

Come **secondo caso** consideriamo una relazione in cui entrambi le variabili x e y sono soggette ad una trasformazione lineare, i vari indici statistici variano nel seguente modo:

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

Procediamo a calcolare il coefficiente di correlazione di Pearson:

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

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 diversi 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 variabli 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 che 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 coefficente di correlazione $|\rho|$, fornisce una misura di quanto le osservazioni si dispongano in prossimità di una retta:

- $|\rho| = 1$ evidenzia una perfetta relazione lineare, cioè 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, se è positivo allora l'approssimazione lineare è crescente (diretta) mentre è 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, poichè 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 causa-effetto tra le due variabili prese in considerazione, cioè due variabili possono presentare un valore di correlazione elevato senza che una determini o causi l'altra. Spesso infatti può intervenire un terzo fatto, o più fattori, a influenzare entrambe le variabili, generando un legame che in realtà non corrisponde a un meccanismo causale diretto.

Indici di eterogeneità

Sappiamo che per le variabili qualitative nominali non è possibile calcolare la varianza nè gli indici che ne derivano, perchè non esistono una media, una mediana o altri valori numerici di riferimento su cui misurare le distanze. Ma è 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.

Indice di Gini

Consideriamo un campione $x_1, \cdots, x_n$ in cui occorrono i valori distinti $v_1, \cdots, v_m$ e indichiamo con f_j la frequenza relativa dell'elemento v_j per $j = 1, \cdots, m$. Definiamo l'**indice di eterogenità di Gini** come:

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

Una proprietà importante di questo indice è che vale $0 \leq I < 1$ ed inoltre l'omogeneità massima dell'insieme di dati si presenta quando $I = 0$, mentre l'eteregeneità massima quando $I = 1$. Di conseguenza, all'aumentare dell'indice di Gini aumenta l'eterogeneità.

Per dimostrare le limitazioni inferiori e superiori, ricordiamo che trattandosi di frequenze relative, vale $0 \leq f_j \leq 1 \ \ \forall j \in [1, m]$ ed inoltre vale $\sum_{j=1}^{m} f_j = 1$. 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 $f_j^2 \leq f_j \Rightarrow \sum_{j=1}^{m} f_j^2 \leq \sum_{j=1}^{m} f_j = 1 \quad \Rightarrow \quad I \geq 0$

Notiamo che l'estremo inferiore si presenta quando l'insieme è massimamente omogeneo e l'estremo superiore quando è massimamente eterogeno, quindi:

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

Indice di Gini normalizzato

Ricordiamo che m è la cardinalità dell'insieme dei valori distinti. Questo indice presenta due problematiche:

- il valore massimo che può assumere, ossia quando l'insieme 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 affinchè si abbia la massima eterogeneità nell'insieme dei dati.
- il suo vaore dipende fortemente dal valore di m. Non è quindi possibile confrontare due attributi qualitativi che presentano intervalli di valori diversi, ovvero m diverso.

Per eliminare questi problemi introduciamo 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: $0 \leq I' \leq 1$. Consideriamo infattti il caso in cui l'eterogeneità dell'insieme è massima:

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

Indice di entropia

Consideriamo un campione $x_1, \cdots, x_n$ in cui occorrono i valori distinti $v_1, \cdots, v_m$ e indichiamo con f_j la frequenza relativa dell'elemento v_j per $j = 1, \cdots, m$. Possiamo definire 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 vento 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 monotona decrescente, vale 0 quando $p = 1$ e tende a infinito per p tendente a 0.

Nel calcolo dell'entropia compare $-f_j \log f_j$. Se $f_j = 0$ questa espressione assume la forma indeterminata $0 \cdot \infty$. Però possiamo estendere la definizione di 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 $-0 \log 0 = 0$.

Effettuiamo le seguenti osservazioni:

- $\forall j$ vale $-f_j \log f_j \geq 0 \ \Rightarrow \ H \geq 0$
- $\forall j$ si ha che $-f_j \log f_j = 0 \ \Leftrightarrow \ f_j = 0 \vee \log f_j = 0$ 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 che vale $0 \leq H \leq \log m$. Più questo indice cresce più aumenta il grado di eterogeneità dell'insieme e vale il viceversa più decresce più aumenta il grado di omogeneità.

Indice di entropia normalizzato

Definiamo 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:

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

Nel nostro caso è utile avere il logaritmo in base 2 in modo da poter misurare l'entropia di bit, che risulta utile quando bisogna svolgere i calcoli in un computer, ma volendo si può utilizzare una qualsiasi base, come il logaritmo naturale o in base 10.

Concentrazione

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

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

Consideriamo un campione di n osservazioni, ciascuno dei quali possiene una certa quantità di risorse. inchiamo con a_i la quantità posseduta dall'i-esimo individuo dopo aver ordinato le osservazioni in ordine crescente, cioè $a_1 \leq a_2 \leq \cdots \leq a_n$. 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. Andando a moltiplicare il valore medio $\bar{a}$ per il numero totale di individui n ottieniamo il totale aggregato della risorsa:

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

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

Possono presentarsi due situzioni estreme:

- Caso di concentrazione minima: tutti gli elementi del campione assumono lo stesso valore, cioè $a_1 = a_2 = \cdots = a_n = \bar{a}$
- Caso di concentazione 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 rispettivi di concentrazione minima e massima, e che negli altri casi sia un valore crescente in funzione della concentrazione. Si considerino:

- La frequenza relativa cumulata degli individui fino all'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:

- $0 \leq F_i, Q_i \leq 1$
- $Q_i = F_i$ in caso di concentrazione minima
- $Q_n = F_n = 1$
- $Q_i \leq F_i$ siccome le osservazioni sono in ordine crescente

Dimostrazione

Vogliamo dimostrare che $Q_i \leq F_i$. Pertanto andiamo a dividere l'insieme ordinato in due sottogruppi, $\{a_1, \cdots, a_i\}$ e $\{a_{i+1}, \cdots, a_n\}$ e definiamo le rispettive somme S_i e T_i:

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

Riscriviamo la disuguaglianza $Q_i \leq F_i$ in termini di S_i e T_i, in particolare, osserviamo 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, vogliamo isolare da un lato della disequazione $\frac{i\,T_i}{S_i}$:

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

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

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

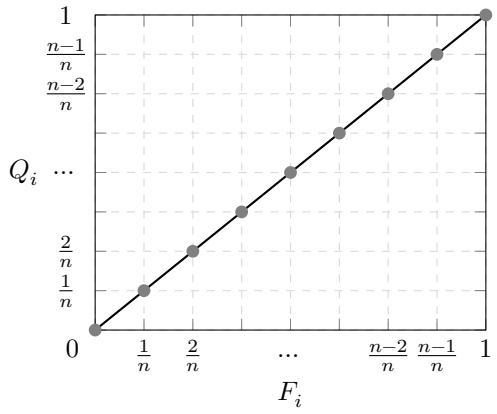
Ne consegue che

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

In tal modo si conclude che $\frac{i\,T_i}{S_i} \geq n - i$. Poiché $n - i \geq n - 1$ quando $i \geq 1$, abbiamo dunque dimostrato

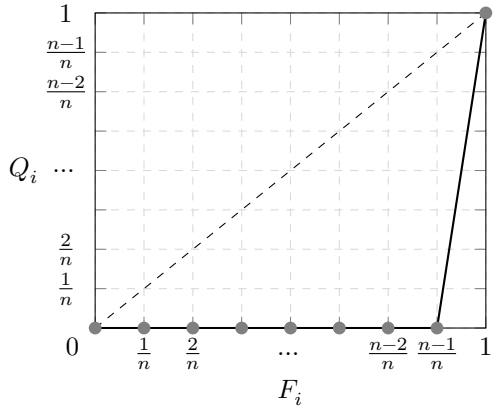
$$\frac{i\,T_i}{S_i} \geq n - 1 \ \ \Rightarrow \ \ Q_i \leq F_i$$

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. Consideriamo ora i punti sul piano che sono indicati 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.

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



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 $F_n - Q_n = 0$.

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

Nei casi intermedi si avrà dunque che i punti staranno su una curva sotto la bisettrice del I° e III° quadrante $F = Q$, cado 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 altra.

Indice di concentrazione di Gini TO DO

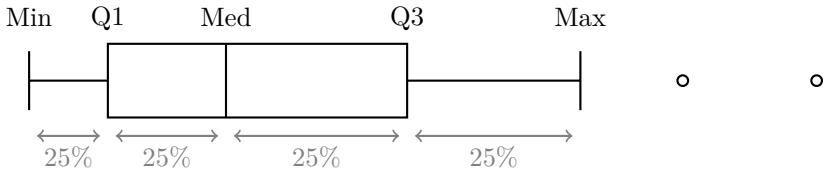
Capitolo 4 - Altro

Altri grafici

Box Plot

Se vogliamo visualizzare alcune statistiche riassuntive di un insieme di dati usiamo un **box plot** (diagramma a scatola). Per realizzarlo tracciamo un segmento orizzontale dal minore al maggiore dei dati. A questo 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.

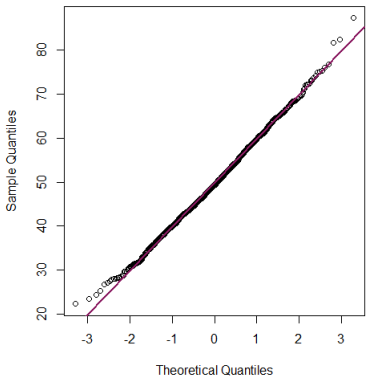
La lunghezza della base del rettangolo corrisponde allo scarto interquantile.



In un box plot ciascuno dei quattro segmenti contiene il 25% delle osservazioni, 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 dividono le osservazioni in parti uguali sul numero dei dati ma non sulla distanza numerica.

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

Q-Q Plot



Un diagramma Q-Q, o diagramma qauntile-quantile, è una rappresentazione grafica qualitativa che permette di verificare le similarità tra le distribuzioni di due campioni diversi. Può tornare 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, se considerati tutti insieme, individuano la distribuzione dei dati.

Ogni asse cartesiano di questo diagramma contiene i quantili dei due campioni presi in considerazione. Poichè i quantili sono ordinati in modo creascente anche il grafico lo sarà, 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.

Distribuzioni normali

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

- L'istogramma è simmetrico rispetto all'intervallo centrale.
- Ha il punto massimo in corrispondenza dell'intervallo centrale.
- Spostandoci dal centro verso destra o 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 ad essere un istrogramma normale, allora diciamo che l'insieme di dati è approssimativamente normale. Inoltre l'insieme di dati si dice asimmetrico 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.

Siano $\bar{x}$ e s rispettivamente la media e la deviazione standard campionarie di un insieme di dati approssivativamente 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 rientrano nell'intervallo $\bar{x} \pm 2s$
- circa il 99,7% delle osservazioni rientrano nell'intervallo $\bar{x} \pm 3s$

Un insieme di dati ottenuto campionando una popolazione costituita da sottogruppi eterogenei non è di solito normale. L'istogramma di un insieme di dati di questo tipo spesso assomiglia ad 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 **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.

Traslazione dei dati TO DO

Alberi di decisione

Gli indici di eterogeneità oltre ad misurare la dispersione delle frequenze nelle variabili qualitative, sono fondamentali anche 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, partendo dalla radice, condizioni sui valori di tali attributi.

In pratica, ad ogni nodo viene associata una condizione/test che suddivide il campione in sottoinsiemi: si segue il percorso corrispondente in base al risultato del test, fino al raggiungimento di una foglia, la quale indica la classe assegnata. La scelta del test in ciascun nodo è guidata proprio dagli indici di eterogeneità; l'obbiettivo è quello di ridurre l'eterogeneità dei dati nei nodi figli rispetto a quella del nodo padre. Si cerca quindi di porre condizioni ai vari nodi per ottenere dei sottogruppi il più omogenei possibili e con il minor numero di nodi possibili.

Ad esempio, tramite 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 usa l'indice di Entropia si cerca la divisione che riduce al minimo l'incertezza nei nodi successivi. In entrambi i casi procedento lungo l'albero si raggiungono foglie contenenti gruppi di oggetti più omogenei rispetto alla classe di appartenenza.

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

Analisi di classificatori TO DO

Classificatori costanti TO DO

Classificatori ideali TO DO

Classificatori casuali TO DO

Classificatori a soglia TO DO

Teoria delle probabilità

Capitolo 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 cosi via fino a s_t modi per operare la t-esima scelta, allora il numero delle sequenze di possibili scelte è

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

Osserviamo che il 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, e cosi via per ogni nodo.

Permutazioni

Consideriamo 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.

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 cosi via, fino all'ultimo elemento che può essere scelto in un solo modo.

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

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

Casi particolari: $p_1 = 1! = 1$ mentre, $p_0 = 0! = 1$ per convenzione.

Permutazioni con ripetizioni

Se gli n oggetti di A non sono tutti distinguibili, ma 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. Poichè 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 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*.

Osserviamo 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.

Disposizioni

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

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 dlela configurazione può essere scelto in n nodi, il secondo in $(n-1)$, il terzo in $(n-2)$ e cosi 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:

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

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 con k = n.

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:

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

Tale formula vale per ogni intero $k \geq 0$, indipendentemente dalla cardinalità 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.

Combinazioni

Consideriamo un insieme di n oggetti $A = a_1, \cdots, a_n$. 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.

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 determinare il numero, si considere 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 un numero pari a quello delle permutazioni dei suoi k oggetti. Per convertite quindi il conteggio delle sequenze ordinati in quello dei sottoinsiemi, in cui l'ordine è irrilevante, è ncessario 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 $c_{n,k} < d_{n,k}$.

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

TO DO

Coefficienti combinatori

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

Il **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 data da:

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

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

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

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 $k + (n-k) = n$.

Capitolo 6 - Probabilità

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

- 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.
- 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 nozione insiemistica, è possibile formalizzare regole e gli assiomi della teoria della probabilità.

Definizioni

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

Spazio degli esiti

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

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

Lo **spazio degli esiti**, anche detto 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 intriseche degli elementi stessi.

Evento

Un evento E è un sottoinsieme dello spazio degli esiti, percio $E \subseteq \Omega$. Un evento formato solo da un solo esisto ω è detto evento elementare, Ω rappresenta l'evento certo mentre $\emptyset$ 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 almeno uno tra E e F si verifica. 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 verifica 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 $E \cap F = \emptyset$ allora E e F si dicono *mutuamente 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 $\Omega^C = \emptyset$.
- 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 simmetria, infatti $E \setminus F \neq F \setminus E$.

È possibile definire l'unione o l'intersezione di più eventi. Consideriamo gli eventi $E_1, \cdots, E_n$:

- 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.

È inoltre possibile definire delle relazioni di **inclusione** e ugualianza 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 esisti 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. Osserviamo 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 $\overline{E \cap F} = \overline{E} \cup \overline{F}$

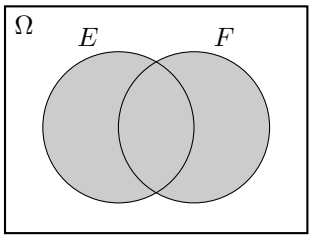
Dimostrazione

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

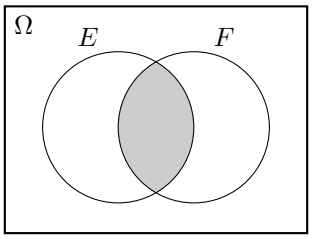
Da entrambe le inclusioni si ottiene che $\overline{E \cup F} = \overline{E} \cap \overline{F}$.

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

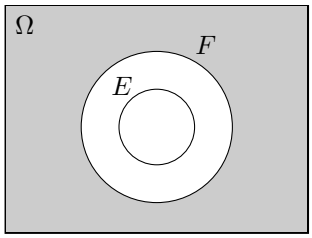
Un tipo di rappresentazione grafica degli eventi, utile per illustrare le relazioni logiche che li legano, sono i **diagrammi di Venn**. Lo spazio degli esiti Ω è rappresentato da un rettangolo che contiene il resto della figura. Gli eventi da prendere in considerazione sono invece rappresentati da cerchi 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:



$E \cup F$



$E \cap F$



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

Algebra degli eventi

Un'algebra degli eventi $\mathcal{A}$ su Ω è una collezione di sottoinsiemi di Ω, ossia 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:

- $\emptyset \in \mathcal{A}$ perché $\emptyset = \overline{\Omega}$
- 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.

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, poichè 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, perchè spesso si considerano σ-algebre proprie, ovvero strettamente più piccole.

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 $\mathcal{F}_M = \{\emptyset, \{T\}, \{C\}, \Omega_M\}$, 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 $\mathcal{F}_D = \{\emptyset, \{1, 2\}, \{3, 4, 5, 6\}, \Omega_D\}$. $\mathcal{F}_D$ è un'algebra ammissibile diversa dall'insieme delle parti $\mathcal{P}(\Omega)$.
Si definisce la mappa $\phi : \mathcal{F}_D \to \mathcal{F}_M$ mediante:

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

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.

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 cresce del numero dei tentativi, a stabilizzarsi in un valore costante, che dipende univocamente da E. Questo valore costante è quello che intendiamo come probabilità dell'evento E.

Si consideri lo spazio misurabile $(\Omega, \mathcal{F})$. Definiziamo 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 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, $E_i \cap E_j = \emptyset$ 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.

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 numerabile additiva o **σ-additiva** se per ogni successione $A_1, \cdots, A_n, \cdots \in \mathcal{A}$ 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, e verranno dimostrate.

Teorema 1

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

Dimostrazione:

$E \cup \overline{E} = \Omega \wedge E \cap \overline{E} = \emptyset \;\Rightarrow$ I due insiemi sono disgiunti

Dato che i due insiemi sono disgiunti, è possibile applicare il terzo assioma:

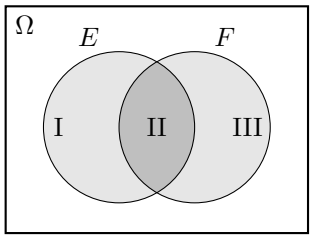
$1 \overset{K2}{=} \mathbb{P}(\Omega) = \mathbb{P}(E \cup \overline{E}) \overset{K3}{=} \mathbb{P}(E) + \mathbb{P}(\overline{E})$

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

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. $\mathrm{I} = E \cap \overline{F}$
2. $\mathrm{II} = E \cap F$
3. $\mathrm{III} = \overline{E} \cap F$

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. Analogamente sarà possibile dimostrarlo per le altre coppie.

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

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

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

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 $\mathbb{P}(\emptyset) = 0$

Dimostrazione:

$\overline{\Omega} = \emptyset$

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

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 $E \cap (F \setminus E) = \emptyset$

E e $F \setminus E$ sono quindi due eventi disgiunti ed è quindi possibile applicare il terzo assioma:

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

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 di probabilità 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 $\{\omega_1, \omega_2, \cdots, \omega_N\}$, 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 $\mathbb{P}(\{\omega_i\}) = p = 1/N$ per $i \in [1, N]$. Generalizzando si ha:

$$p = \frac{|\{\omega_i\}|}{|\Omega|}$$

Si consideri un evento $E \subseteq \Omega$, ed essendo E finito sia la sua cardinalità $|E| = k$. Riapplicando gli assiomi:

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

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.

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 condizionate di E dato F si indica con $\mathbb{P}(E|F)$, oppure $\mathbb{P}_F(E)$, e si può definire a partto che la probabilità di F sia diverso da zero.

La probabilità condizionata subentra tutte le vole 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.

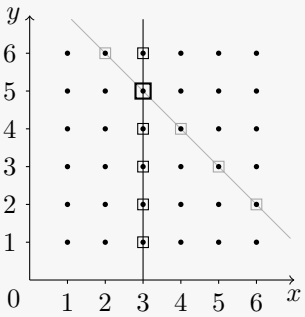
Utilizzando la definizione classica di probabilità, vale la seguente formula sulla probabilità condizionata:

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

Infatti, se si è verificato l'evento F, affinchè 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 esisti, sostituendo Ω.

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

Esempio Si immagini di tirare due dadi. Lo spazio degli esiti di questo esperimento è descritto da $\Omega = \{(x, y) \mid x, y \in [1, 6]\}$ 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 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.

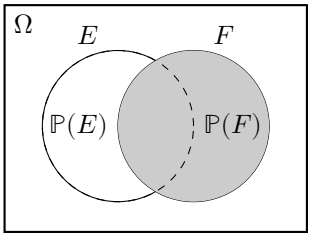
Definizione rigorosa: dato uno spazio misurabile $(\Omega, \mathcal{F})$ di misura $\mathbb{P}$, ogni evento F eredita una struttura di spazio misurato $(F, \mathcal{A}_F, \mathbb{P})$, restrigendo gli insiemi misurabili a quelli contenuti in F, ed induce una nuova misura $\mathbb{P}'_F(E) = \mathbb{P}(E \cap F)$ su $(\Omega, \mathcal{F})$, con $\mathbb{P}'_F(\Omega) = P(F)$.

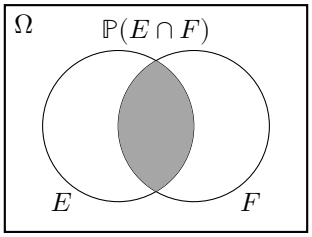
Se $(\Omega, \mathcal{F}, \mathbb{P})$ è uno spazio di probabilità (valgono quindi gli assiomi di Kolmogorov, tra cui $\mathbb{P}(\Omega) = 1$) e F non è trascurabile (ossia $\mathbb{P}(F) \neq 0$), allora riscalando $\mathbb{P}'_F$ a $\mathbb{P}_F = \frac{1}{\mathbb{P}(F)} \mathbb{P}'_F$ si ottiene lo spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P}_F)$ condizionato dall'evento F.

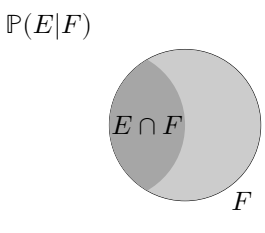
Teorema delle probabilità totali

Regola di fattorizzazione

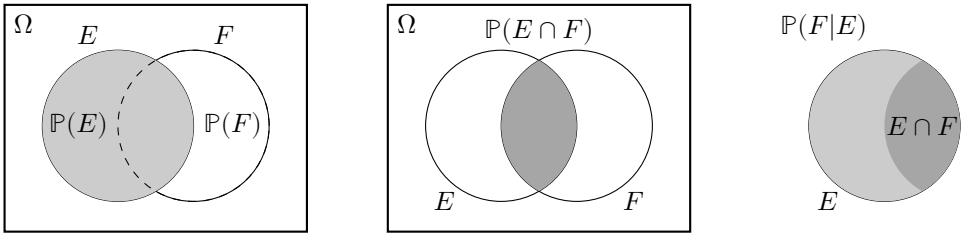
Siano E e F due eventi in uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$. Se $\mathbb{P}(F) \neq 0$, moltiplicando entrambi i membri della formula della probabilità condizionata di E dato F per $\mathbb{P}(F)$ si ottiene $\mathbb{P}(E \cap F) = \mathbb{P}(E|F), \mathbb{P}(F)$







Allo stesso modo, se $\mathbb{P}(E) \neq 0$, 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:

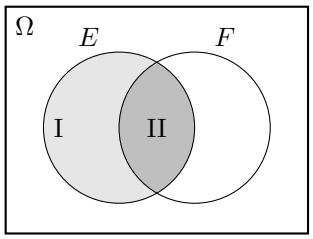
$$\mathbb{P}(E \cap F) = \mathbb{P}(E|F), \mathbb{P}(F) = \mathbb{P}(F|E), \mathbb{P}(E)$$

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 e 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 $(\Omega, \mathcal{F}, \mathbb{P})$. Poiche 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 versione binaria del teorema delle probabilità totali, limitata alla partizione $\{F, \overline{F}\}$:

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

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:

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

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 $(E \cap F_i) \cap (E \cap F_j) = \emptyset$ 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 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 tutti gli indici i si ottiene dunque *la formula delle probabilità totali in forma estesa*:

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

Questa relazione generalizza il caso binario $\{F, \overline{F}\}$ e permette di calcolare $\mathbb{P}(E)$ suddividendo lo spazio deli esiti in una partizione $\{F_1, F_2, \dots, F_n\}$. In tal modo, ciascun insieme F_i ha la 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 $\mathbb{P}(E)$.

Teorema di Bayes

Una volta chiarite la regola di fattorizzazione e la formula delle probabilità totali, possiamo introdurre il teorema di Bayes, questo 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:

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

Dimostrazione Tramite la formula della regola di fattorizzazione 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 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, $\mathbb{P}(F|E)$ equivale al contrario. 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, otteniamo la forma estesa del teorema di Bayes, sia $F_1, \cdots, F_n$ una partizione di Ω e $\mathbb{P}(F_i) \neq 0$ per ogni i, e sia E un evento tale per cui $\mathbb{P}(E) \neq 0$ allora:

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

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

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 quelli disponibili. Per esempio, potremmo suddividere gli individui in "positivi" o "negativi" rispetto a una determinata condizione.

Nel contesto di un classificatore bayesano, 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 cosideriamo:

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

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

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

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

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

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

Sostituendo questa ipotesi nella versione bayesiana precedente, si ottiene:

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

Il denominatore $\mathbb{P}(X_1 = x_1, \cdots, X_n = x_n)$ non dipende dalla classe y_k ma solo dai valori osservati $(x_1, \cdots, x_n)$. 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 $(x_1, \cdots, x_n)$, 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:

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

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.

Eventi indipendenti

In generale, la probabilità condizionata $\mathbb{P}(E|F)$ differisce da $\mathbb{P}(E)$, poichè il verificarsi di F fornisce informazioni che possono modificare la probabilità che si verifiche 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:

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

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

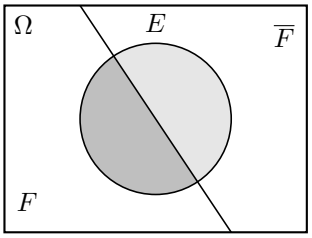
Questa relazione evidenzia che, se E è indipendente da F, allora anche F è indipendente da E, poichè entrambi gli enunciati implicano l'uguaglianza della probabilità dell'intersezione al prodotto delle probabilità marginali. La nozione di indipendenza si conserva rispetto ad alcune operazioni insiemistiche elemetari 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.

Dimostriamo questo fatto per quanto riguarda l'operazione di complemento.

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

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. $(E \cap F) \cap (E \cap \overline{F}) = \emptyset$

  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:

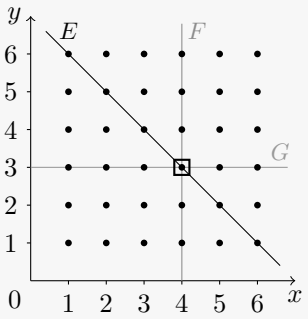
$$\mathbb{P}(E \cap \overline{F}) = \mathbb{P}(E) - \mathbb{P}(E)\,\mathbb{P}(F)$$

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

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 $\Omega = \{(x, y) \mid x, y \in [1, 6]\}$ 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 risultano indipendenti senza necessità di ulteriori verifiche. Questa proprietà conferma che l'indipendenza si estende naturalmente agli eventi costruiti a partire da eventi 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 $E_1, \cdots, E_n \subseteq \Omega$, questi sono indipendenti se e solo se $\forall r \leq n \quad \forall\, 1 \leq \alpha_1 \leq \alpha_2 \leq ... \leq \alpha_r \leq n$ 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.

Statistica inferenziale

Capitolo 7 - Analisi della varianza

Formulario

Frequenza cumulata:

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