

# Psicometria

Corrado Caudek

Questo documento è stato realizzato con:

- $\text{\LaTeX}$  e la classe memoir (<http://www.ctan.org/pkg/memoir>);
- R (<http://www.r-project.org/>) e RStudio (<http://www.rstudio.com/>);
- bookdown (<http://bookdown.org/>) e memoir (<https://ericmarcon.github.io/memoir/>).



Nel blog della mia pagina personale sono forniti alcuni approfondimenti degli argomenti qui trattati.

<https://ccaudek.github.io/caudeklab/>

# Indice

<b>Indice</b>	<b>iii</b>
<b>Prefazione</b>	<b>vii</b>
La psicologia e la Data Science . . . . .	vii
Come studiare . . . . .	viii
Sviluppare un metodo di studio efficace . . . . .	viii
<b>1 Il teorema di Bayes</b>	<b>1</b>
1.1 Il teorema della probabilità totale . . . . .	1
1.2 La regola di Bayes . . . . .	2
Le probabilità come grado di fiducia . . . . .	4
Aggiornamento bayesiano . . . . .	4
Considerazioni conclusive . . . . .	6
<b>Bibliografia</b>	<b>7</b>
<b>Elenco delle figure</b>	<b>9</b>



Copyright © 2022.

Data della versione presente: Dicembre 21, 2021.



# Prefazione

**Data Science per psicologi** contiene il materiale delle lezioni dell'insegnamento di *Psicometria B000286* (A.A. 2021/2022) rivolto agli studenti del primo anno del Corso di Laurea in Scienze e Tecniche Psicologiche dell'Università degli Studi di Firenze.

L'insegnamento di Psicometria si propone di fornire agli studenti un'introduzione all'analisi dei dati in psicologia. Le conoscenze/competenze che verranno sviluppate in questo insegnamento sono quelle della *Data science*, ovvero le conoscenze/competenze che si pongono all'intersezione tra statistica (ovvero, richiedono la capacità di comprendere teoremi statistici) e informatica (ovvero, richiedono la capacità di sapere utilizzare un software).

## La psicologia e la Data Science

*It's worth noting, before getting started, that this material is hard. If you find yourself confused at any point, you are normal. Any sense of confusion you feel is just your brain correctly calibrating to the subject matter. Over time, confusion is replaced by comprehension [...]*

— Richard McElreath

Sembra sensato spendere due parole su un tema che è importante per gli studenti: quello indicato dal titolo di questo Capitolo. È ovvio che agli studenti di psicologia la statistica non piace. Se piacesse, forse studierebbero Data Science e non psicologia; ma non lo fanno. Di conseguenza, gli studenti di psicologia si chiedono: “perché dobbiamo perdere tanto tempo a studiare queste cose quando in realtà quello che ci interessa è tutt'altro?” Questa è una bella domanda.

C'è una ragione molto semplice che dovrebbe farci capire perché la Data Science è così importante per la psicologia. Infatti, a ben pensarci, la psicologia è una disciplina intrinsecamente statistica, se per statistica intendiamo quella disciplina che studia la variazione delle caratteristiche degli individui nella popolazione. La psicologia studia *gli individui* ed è proprio la variabilità inter- e intra-individuale ciò che vogliamo descrivere e, in certi casi, predire. In questo senso, la psicologia è molto diversa dall'ingegneria, per esempio. Le proprietà di un determinato ponte sotto certe condizioni, ad esempio, sono molto simili a quelle di un altro ponte, sotto le medesime condizioni. Quindi, per un ingegnere la statistica è poco importante: le proprietà dei materiali sono unicamente dipendenti dalla loro composizione e restano costanti. Ma lo stesso non può dirsi degli individui: ogni individuo è unico e cambia nel tempo. E le variazioni tra gli individui, e di un individuo nel tempo, sono l'oggetto di studio proprio della psicologia: è dunque chiaro che i problemi che la psicologia si pone sono molto diversi da quelli affrontati, per esempio, dagli ingegneri. Questa è la ragione per cui abbiamo tanto bisogno della *data science* in psicologia: perché la *data science* ci consente di descrivere la variazione e il cambiamento. E queste sono appunto le caratteristiche di base dei fenomeni psicologici.

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il “lavoro sporco” (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò

che ci interessa, ovvero dei problemi psicologici slegati dai dettagli “tecnici” della *data science*? La risposta a questa domanda è che non è possibile progettare uno studio psicologico sensato senza avere almeno una comprensione rudimentale della *data science*. Le tematiche della *data science* non possono essere ignorate né dai ricercatori in psicologia né da coloro che svolgono la professione di psicologo al di fuori dell’Università. Infatti, anche i professionisti al di fuori dall’università non possono fare a meno di leggere la letteratura psicologica più recente: il continuo aggiornamento delle conoscenze è infatti richiesto dalla deontologia della professione. Ma per potere fare questo è necessario conoscere un bel po’ di *data science*! Basta aprire a caso una rivista specialistica di psicologia per rendersi conto di quanto ciò sia vero: gli articoli che riportano i risultati delle ricerche psicologiche sono zeppi di analisi statistiche e di modelli formali. E la comprensione della letteratura psicologica rappresenta un requisito minimo nel bagaglio professionale dello psicologo.

Le considerazioni precedenti cercano di chiarire il seguente punto: la *data science* non è qualcosa da studiare a malincuore, in un singolo insegnamento universitario, per poi poterla tranquillamente dimenticare. Nel bene e nel male, gli psicologi usano gli strumenti della *data science* in tantissimi ambiti della loro attività professionale: in particolare quando costruiscono, somministrano e interpretano i test psicometrici. È dunque chiaro che possedere delle solide basi di *data science* è un tassello imprescindibile del bagaglio professionale dello psicologo. In questo insegnamento verranno trattati i temi base della *data science* e verrà adottato un punto di vista bayesiano, che corrisponde all’approccio più recente e sempre più diffuso in psicologia.

## Come studiare

*I know quite certainly that I myself have no special talent. Curiosity, obsession and dogged endurance, combined with self-criticism, have brought me to my ideas.*

— Albert Einstein

Il giusto metodo di studio per prepararsi all’esame di Psicometria è quello di seguire attivamente le lezioni, assimilare i concetti via via che essi vengono presentati e verificare in autonomia le procedure presentate a lezione. Incoraggio gli studenti a farmi domande per chiarire ciò che non è stato capito appieno. Incoraggio gli studenti a utilizzare i forum attivi su Moodle e, soprattutto, a svolgere gli esercizi proposti su Moodle. I problemi forniti su Moodle rappresentano il livello di difficoltà richiesto per superare l’esame e consentono allo studente di comprendere se le competenze sviluppate fino a quel punto sono sufficienti rispetto alle richieste dell’esame.

La prima fase dello studio, che è sicuramente individuale, è quella in cui è necessario acquisire le conoscenze teoriche relative ai problemi che saranno presentati all’esame. La seconda fase di studio, che può essere facilitata da scambi con altri e da incontri di gruppo, porta ad acquisire la capacità di applicare le conoscenze: è necessario capire come usare un software (R) per applicare i concetti statistici alla specifica situazione del problema che si vuole risolvere. Le due fasi non sono però separate: il saper fare molto spesso ci aiuta a capire meglio.

## Sviluppare un metodo di studio efficace

*Memorization is not learning.*

— Richard Phillips Feynman

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l’atteggiamento mentale che descriverò qui



---

sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo. Fornisco qui sotto una breve descrizione del “metodo di studio” che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento (Burger & Starbird, 2012).

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.
- Gli errori che facciamo sono i nostri migliori maestri. Istitivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un’informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.
- C’è ovviamente un aspetto “psicologico” nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: “mi arrendo”, “non ho idea di cosa fare!”. Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose “migliori” che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c’è qualcosa che non so fare e non ho idea di come affrontare, mi dico: “oggi non ho proprio voglia di fare fatica”, non ho voglia di mettermi nello stato mentale per cui “in 10 minuti devo risolvere il problema perché dopo devo fare altre cose”. Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto “marginale” del problema, che so come affrontare, oppure considero l’aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, oppure lo stesso problema in un altro contesto. Non mi pongo l’obiettivo “risolvi il problema in 10 minuti”, ma invece quello di farmi un’idea “generale” del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se “parto per la tangente”, ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a “lavorare” al problema. Allora perché non faccio sempre così? C’è ovviamente l’aspetto dei “10 minuti” che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possiamo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.
- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: “se questo è vero, allora deve succedere quest’altra cosa”. Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c’era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.

- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.<sup>1</sup> È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per cercare di colmarle. L'atteggiamento naturale, quando non capiamo i dettagli di qualcosa, è quello di pensare: "non importa, ho capito in maniera approssimativa questo punto, non devo preoccuparmi del resto". Ma in realtà non è vero: se la nostra comprensione è superficiale, quando il problema verrà presentato in una nuova forma, non riusciremo a risolverlo. Per cui i dubbi che ci vengono quando studiamo qualcosa sono il nostro alleato più prezioso: ci dicono esattamente quali sono gli aspetti che dobbiamo approfondire per potere migliorare la nostra preparazione.
- È utile sviluppare una visione d'insieme degli argomenti trattati, capire l'obiettivo generale che si vuole raggiungere e avere chiaro il contributo che i vari pezzi di informazione forniscono al raggiungimento di tale obiettivo. Questa organizzazione mentale del materiale di studio facilita la comprensione. È estremamente utile creare degli schemi di ciò che si sta studiando. Non aspettate che sia io a fornirvi un riepilogo di ciò che dovete imparare: sviluppate da soli tali schemi e tali riassunti.
- Tutti noi dobbiamo imparare l'arte di trovare le informazioni, non solo nel caso di questo insegnamento. Quando vi trovate di fronte a qualcosa che non capite, o ottenete un oscuro messaggio di errore da un software, ricordatevi: "Google is your friend".

Corrado Caudek

Febbraio 2022

---

<sup>1</sup>Ricordatevi inoltre che gli individui tendono a sottostimare la propria capacità di apprendere (Horn & Loewenstein, 2021).

# Capitolo 1

## Il teorema di Bayes



Questo Capitolo presenterà il teorema di Bayes per calcolare la probabilità degli eventi riferiti a esperimenti casuali, ossia esperimenti di cui non si può prevedere il risultato finale ma di cui si conoscono tutti i possibili risultati. Prima di esaminare il teorema di Bayes verrà introdotta una sua componente, ovvero il teorema della probabilità totale.

### 1.1 Il teorema della probabilità totale

Il teorema della probabilità totale fa uso della legge della probabilità composta (??). Lo discuteremo qui considerando il caso di una partizione dello spazio campionario in tre sottoinsiemi, ma è facile estendere tale discussione al caso di una partizione in un qualunque numero di sottoinsiemi.

**Teorema 1.1.** *Sia  $\{F_1, F_2, F_3\}$  una partizione dello spazio campionario  $\Omega$ . Se  $E$  è un qualunque altro evento, allora:*

$$P(E) = P(E \cap F_1) + P(E \cap F_2) + P(E \cap F_3)$$

ovvero

$$P(E) = P(E | F_1)P(F_1) + P(E | F_2)P(F_2) + P(E | F_3)P(F_3). \quad (1.1)$$

Il teorema della probabilità totale afferma che, se l'evento  $E$  è costituito da tutti gli eventi elementari in  $E \cap F_1$ ,  $E \cap F_2$  e  $E \cap F_3$ , allora la probabilità  $P(E)$  è data dalla somma delle probabilità di questi tre eventi (figura 1.1).

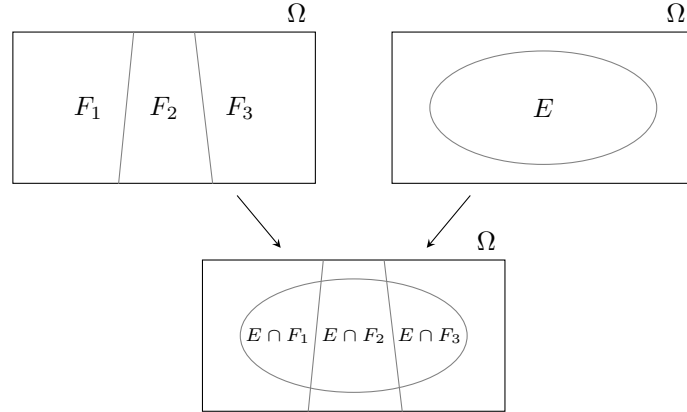
**Esempio 1.1.** Si considerino tre urne, ciascuna delle quali contiene 100 palline:

- Urna 1: 75 palline rosse e 25 palline blu,
- Urna 2: 60 palline rosse e 40 palline blu,
- Urna 3: 45 palline rosse e 55 palline blu.

Una pallina viene estratta a caso da un'urna anch'essa scelta a caso. Qual è la probabilità che la pallina estratta sia di colore rosso?

Sia  $R$  l'evento "la pallina estratta è rossa" e sia  $U_i$  l'evento che corrisponde alla scelta dell' $i$ -esima urna. Sappiamo che

$$P(R | U_1) = 0.75, \quad P(R | U_2) = 0.60, \quad P(R | U_3) = 0.45.$$



**Figura 1.1:** Partizione dello spazio campionario  $\Omega$ .

Gli eventi  $U_1$ ,  $U_2$  e  $U_3$  costituiscono una partizione dello spazio campionario in quanto  $U_1$ ,  $U_2$  e  $U_3$  sono eventi mutualmente esclusivi ed esaustivi,  $P(U_1 \cup U_2 \cup U_3) = 1.0$ . In base al teorema della probabilità totale, la probabilità di estrarre una pallina rossa è

$$\begin{aligned} P(R) &= P(R | U_1)P(U_1) + P(R | U_2)P(U_2) + P(R | U_3)P(U_3) \\ &= 0.75 \cdot \frac{1}{3} + 0.60 \cdot \frac{1}{3} + 0.45 \cdot \frac{1}{3} = 0.60. \end{aligned}$$

**Esempio 1.2.** Consideriamo un'urna che contiene 5 palline rosse e 2 palline verdi. Due palline vengono estratte, una dopo l'altra. Vogliamo sapere la probabilità dell'evento "la seconda pallina estratta è rossa".

Lo spazio campionario è  $\Omega = \{RR, RV, VR, VV\}$ . Chiamiamo  $R_1$  l'evento "la prima pallina estratta è rossa",  $V_1$  l'evento "la prima pallina estratta è verde",  $R_2$  l'evento "la seconda pallina estratta è rossa" e  $V_2$  l'evento "la seconda pallina estratta è verde". Dobbiamo trovare  $P(R_2)$  e possiamo risolvere il problema usando il teorema della probabilità totale (1.1):

$$\begin{aligned} P(R_2) &= P(R_2 | R_1)P(R_1) + P(R_2 | V_1)P(V_1) \\ &= \frac{4}{6} \cdot \frac{5}{7} + \frac{5}{6} \cdot \frac{2}{7} = \frac{30}{42} = \frac{5}{7}. \end{aligned}$$

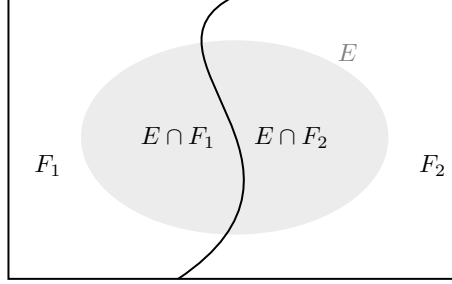
Se la prima estrazione è quella di una pallina rossa, nell'urna restano 4 palline rosse e due verdi, dunque, la probabilità che la seconda estrazione produca una pallina rossa è uguale a  $4/6$ . La probabilità di una pallina rossa nella prima estrazione è  $5/7$ . Se la prima estrazione è quella di una pallina verde, nell'urna restano 5 palline rosse e una pallina verde, dunque, la probabilità che la seconda estrazione produca una pallina rossa è uguale a  $5/6$ . La probabilità di una pallina verde nella prima estrazione è  $2/7$ .

## 1.2 La regola di Bayes

Il teorema di Bayes rappresenta uno dei fondamenti della teoria della probabilità e della statistica. Lo presentiamo qui considerando un caso specifico per poi descriverlo nella sua forma più generale.

Sia  $\{F_1, F_2\}$  una partizione dello spazio campionario  $\Omega$ . Consideriamo un terzo evento  $E \subset \Omega$  con probabilità non nulla di cui si conoscono le probabilità condizionate rispetto ad  $F_1$  e a  $F_2$ , ovvero  $P(E | F_1)$  e  $P(E | F_2)$ . È chiaro per le ipotesi fatte che se si verifica  $E$  deve anche essersi verificato almeno uno degli eventi  $F_1$  e  $F_2$ . Supponendo che si sia

verificato l'evento  $E$ , ci chiediamo: qual è la probabilità che si sia verificato  $F_1$  piuttosto che  $F_2$ ?



Per rispondere alla domanda precedente scriviamo:

$$\begin{aligned} P(F_1 | E) &= \frac{P(E \cap F_1)}{P(E)} \\ &= \frac{P(E | F_1)P(F_1)}{P(E)}. \end{aligned}$$

Sapendo che  $E = (E \cap F_1) \cup (E \cap F_2)$  e che  $F_1$  e  $F_2$  sono eventi disgiunti, ovvero  $F_1 \cap F_2 = \emptyset$ , ne segue che possiamo calcolare  $P(E)$  utilizzando il teorema della probabilità totale:

$$\begin{aligned} P(E) &= P(E \cap F_1) + P(E \cap F_2) \\ &= P(E | F_1)P(F_1) + P(E | F_2)P(F_2). \end{aligned}$$

Sostituendo il risultato precedente nella formula della probabilità condizionata  $P(F_1 | E)$  otteniamo:

$$P(F_1 | E) = \frac{P(E | F_1)P(F_1)}{P(E | F_1)P(F_1) + P(E | F_2)P(F_2)}. \quad (1.2)$$

La (1.2) si generalizza facilmente al caso di più di due eventi disgiunti, come indicato di seguito.

**Teorema 1.2.** *Sia  $E$  un evento contenuto in  $F_1 \cup \dots \cup F_k$ , dove gli eventi  $F_j, j = 1, \dots, k$  sono a due a due incompatibili e necessari. Allora per ognuno dei suddetti eventi  $F_j$  vale la seguente formula:*

$$P(F_j | E) = \frac{P(E | F_j)P(F_j)}{\sum_{j=1}^k P(F_j)P(E | F_j)}. \quad (1.3)$$

La (1.3) prende il nome di *Teorema di Bayes* e mostra che la conoscenza del verificarsi dell'evento  $E$  modifica la probabilità che abbiamo attribuito all'evento  $F_j$ . Nella (1.3) la probabilità condizionata  $P(F_j | E)$  prende il nome di probabilità *a posteriori* dell'evento  $F_j$ : il termine “a posteriori” sta a significare “dopo che è noto che si è verificato l'evento  $E$ ”. Nel capitolo ?? estenderemo questa discussione mostrando come la (1.3) possa essere formulata in un modo più generale, ovvero in modo tale che non faccia riferimento unicamente alla probabilità di eventi, ma bensì anche alle funzioni di densità di probabilità.

### Le probabilità come grado di fiducia

Il teorema di Bayes rende esplicito il motivo per cui la probabilità non può essere pensata come uno stato oggettivo, quanto piuttosto come un'inferenza soggettiva e condizionata. Il denominatore del membro di destra della (1.3) è un semplice fattore di normalizzazione. Nel numeratore compaiono invece due quantità:  $P(F_j)$  e  $P(E | F_j)$ . La probabilità  $P(F_j)$  è la probabilità *probabilità a priori* (*prior*) dell'evento  $F_j$  e rappresenta l'informazione che l'agente bayesiano possiede a proposito dell'evento  $F_j$ . Diremo che  $P(F_j)$  codifica il grado di fiducia che l'agente ripone in  $F_j$ , sul quale non possiamo porre vincoli di alcun tipo. La probabilità condizionata  $P(E | F_j)$  rappresenta invece la verosimiglianza di  $F_j$  e ci dice quant'è plausibile che si verifichi l'evento  $E$  condizionatamente al fatto che si sia verificato  $F_j$ .

Nell'interpretazione bayesiana  $P(F_j)$  rappresenta un giudizio personale dell'agente e non esistono criteri esterni che possano determinare se tale giudizio sia corretto o meno. Il teorema di Bayes descrive la regola che l'agente deve seguire per aggiornare il suo grado di fiducia in  $F_j$  alla luce di un ulteriore evento  $E$ . Per questo motivo abbiamo chiamato  $P(F_j | E)$  probabilità a posteriori: essa rappresenta infatti la nuova probabilità che l'agente assegna ad  $F_j$  affinché rimanga consistente con le nuove informazioni fornitegli da  $E$ .

La probabilità a posteriori dipende sia da  $E$ , sia dalla conoscenza a priori dell'agente  $P(F_j)$ . In questo senso è chiaro come non abbia senso parlare di una probabilità oggettiva: per il teorema di Bayes la probabilità è definita condizionatamente alla probabilità a priori, la quale a sua volta, per definizione, è un'assegnazione soggettiva. Ne segue pertanto che ogni probabilità debba essere una rappresentazione del grado di fiducia (soggettiva) dell'agente.

Se ogni assegnazione probabilistica rappresenta uno stato di conoscenza, è altresì vero che un particolare stato di conoscenza è arbitrario e dunque non deve esserci necessariamente accordo a priori tra diversi agenti. Tuttavia, alla luce di nuove informazioni, la teoria delle probabilità ci fornisce uno strumento che consente l'aggiornamento dello stato di conoscenza in un modo razionale.

### Aggiornamento bayesiano

Il teorema di Bayes consente di modificare una credenza a priori in maniera dinamica, via via che nuove evidenze vengono raccolte, in modo tale da formulare una credenza a posteriori la quale non è mai definitiva, ma può sempre essere aggiornata in base alle nuove evidenze disponibili. Questo processo si chiama *aggiornamento bayesiano*.

**Esempio 1.3.** Supponiamo che, per qualche strano errore di produzione, una fabbrica produca due tipi di monete. Il primo tipo di monete ha la caratteristica che, quando una moneta viene lanciata, la probabilità di osservare l'esito "testa" è 0.6. Per semplicità, sia  $\theta$  la probabilità di osservare l'esito "testa". Per una moneta del primo tipo, dunque,  $\theta = 0.6$ . Per una moneta del secondo tipo, invece, la probabilità di produrre l'esito "testa" è 0.4. Ovvero,  $\theta = 0.4$ .

Noi possediamo una moneta, ma non sappiamo se è del primo tipo o del secondo tipo. Sappiamo solo che il 75% delle monete sono del primo tipo e il 25% sono del secondo tipo. Sulla base di questa conoscenza *a priori* – ovvero sulla base di una conoscenza ottenuta senza avere eseguito l'esperimento che consiste nel lanciare la moneta una serie di volte per osservare gli esiti prodotti – possiamo dire che la probabilità di una prima ipotesi, secondo la quale  $\theta = 0.6$ , è 3 volte più grande della probabilità di una seconda ipotesi, secondo la quale  $\theta = 0.4$ . Senza avere eseguito alcun esperimento casuale con la moneta, questo è quello che sappiamo.

Ora immaginiamo di lanciare una moneta due volte e di ottenere il risultato seguente:  $\{T, C\}$ . Quello che ci chiediamo è: sulla base di questa evidenza, come cambiano le probabilità che associamo alle due ipotesi? In altre parole, ci chiediamo qual è la pro-

bilità di ciascuna ipotesi alla luce dei dati che sono stati osservati:  $P(H | y)$ , laddove  $y$  sono i dati osservati. Tale probabilità si chiama probabilità a posteriori. Inoltre, se confrontiamo le due ipotesi, ci chiediamo quale valore assuma il rapporto  $\frac{P(H_1|y)}{P(H_2|y)}$ . Tale rapporto ci dice quanto è più probabile  $H_1$  rispetto ad  $H_2$ , alla luce dei dati osservati. Infine, ci chiediamo come cambia il rapporto definito sopra, quando osserviamo via via nuovi risultati prodotti dal lancio della moneta.

Definiamo il problema in maniera più chiara. Conosciamo le probabilità a priori, ovvero  $P(H_1) = 0.75$  e  $P(H_2) = 0.25$ . Quello che vogliamo conoscere sono le probabilità a posteriori  $P(H_1 | y)$  e  $P(H_2 | y)$ .

Per trovare le probabilità a posteriori applichiamo il teorema di Bayes:

$$P(H_1 | y) = \frac{P(y | H_1)P(H_1)}{P(y)} = \frac{P(y | H_1)P(H_1)}{P(y | H_1)P(H_1) + P(y | H_2)P(H_2)}$$

laddove lo sviluppo del denominatore deriva da un'applicazione del teorema della probabilità totale. Inoltre,

$$P(H_2 | y) = \frac{P(y | H_2)P(H_2)}{P(y | H_1)P(H_1) + P(y | H_2)P(H_2)}.$$

Se consideriamo l'ipotesi  $H_1$  = "la probabilità di testa è 0.6", allora la verosimiglianza dei dati  $\{T, C\}$ , ovvero la probabilità di osservare questa specifica sequenza di T e C, è uguale a  $0.6 \times 0.4 = 0.24$ . Dunque,  $P(y | H_1) = 0.24$ .

Se invece consideriamo l'ipotesi  $H_2$  = "la probabilità di testa è 0.4", allora la verosimiglianza dei dati  $\{T, C\}$  è  $0.4 \times 0.6 = 0.24$ , ovvero,  $P(y | H_2) = 0.24$ . In base alle due ipotesi  $H_1$  e  $H_2$ , dunque, i dati osservati hanno la medesima plausibilità di essere osservati. Per semplicità, calcoliamo anche

$$P(y) = P(y | H_1)P(H_1) + P(y | H_2)P(H_2) = 0.24 \cdot 0.75 + 0.24 \cdot 0.25 = 0.24.$$

Le probabilità a posteriori diventano:

$$P(H_1 | y) = \frac{P(y | H_1)P(H_1)}{P(y)} = \frac{0.24 \cdot 0.75}{0.24} = 0.75,$$

$$P(H_2 | y) = \frac{P(y | H_2)P(H_2)}{P(y)} = \frac{0.24 \cdot 0.25}{0.24} = 0.25.$$

Possiamo dunque concludere dicendo che, sulla base dei dati osservati, l'ipotesi  $H_1$  ha una probabilità 3 volte maggiore di essere vera dell'ipotesi  $H_2$ .

È tuttavia possibile raccogliere più evidenze e, sulla base di esse, le probabilità a posteriori cambieranno. Supponiamo di lanciare la moneta una terza volta e di osservare croce. I nostri dati dunque sono  $\{T, C, C\}$ .

Di conseguenza,  $P(y | H_1) = 0.6 \cdot 0.4 \cdot 0.4 = 0.096$  e  $P(y | H_2) = 0.4 \cdot 0.6 \cdot 0.6 = 0.144$ . Ne segue che le probabilità a posteriori diventano:

$$P(H_1 | y) = \frac{P(y | H_1)P(H_1)}{P(y)} = \frac{0.096 \cdot 0.75}{0.096 \cdot 0.75 + 0.144 \cdot 0.25} = 0.667,$$

$$P(H_2 | y) = \frac{P(y | H_2)P(H_2)}{P(y)} = \frac{0.144 \cdot 0.25}{0.096 \cdot 0.75 + 0.144 \cdot 0.25} = 0.333.$$

In queste circostanze, le evidenze che favoriscono  $H_1$  nei confronti di  $H_2$  sono solo pari ad un fattore di 2.

Se otteniamo ancora croce in un quarto lancio della moneta, i nostri dati diventano:  $\{T, C, C, C\}$ . Ripetendo il ragionamento fatto sopra,  $P(y | H_1) = 0.6 \cdot 0.4 \cdot 0.4 \cdot 0.4 = 0.0384$  e  $P(y | H_2) = 0.4 \cdot 0.6 \cdot 0.6 \cdot 0.6 = 0.0864$ .

Dunque

$$P(H_1 | y) = \frac{0.0384 \cdot 0.75}{0.0384 \cdot 0.75 + 0.0864 \cdot 0.25} = 0.571,$$

$$P(H_2 | y) = \frac{0.0864 \cdot 0.25}{0.0384 \cdot 0.75 + 0.0864 \cdot 0.25} = 0.429.$$

e le evidenze a favore di  $H_1$  si riducono a 1.33. Se si ottenesse un altro esito croce in un sesto lancio della moneta, l'ipotesi  $H_2$  diventerebbe più probabile dell'ipotesi  $H_1$ .

In conclusione, questo esercizio ci fa capire come sia possibile aggiornare le nostre credenze sulla base delle evidenze disponibili, ovvero come sia possibile passare da un grado di conoscenza del mondo a priori a una conoscenza a posteriori. Se prima di lanciare la moneta ritenevamo che l'ipotesi  $H_1$  fosse tre volte più plausibile dell'ipotesi  $H_2$ , dopo avere osservato uno specifico campione di dati siamo giunti alla conclusione opposta. Il processo di aggiornamento bayesiano, dunque, ci fornisce un metodo per modificare il livello di fiducia in una data ipotesi, alla luce di nuove informazioni.

### Considerazioni conclusive

Il teorema di Bayes costituisce il fondamento dell'approccio più moderno della statistica, quello appunto detto bayesiano. Chi usa il teorema di Bayes non è, solo per questo motivo, "bayesiano": ci vuole ben altro. Ci vuole un modo diverso per intendere il significato della probabilità e un modo diverso per intendere gli obiettivi dell'inferenza statistica. In anni recenti, una gran parte della comunità scientifica ha riconosciuto all'approccio bayesiano il merito di consentire lo sviluppo di modelli anche molto complessi (intrattabili in base all'approccio frequentista) senza richiedere, d'altra parte, conoscenze matematiche troppo avanzate all'utente. Per questa ragione l'approccio bayesiano sta prendendo sempre più piede, anche in psicologia.



# Bibliografia

- Burger, E. B. & Starbird, M. (2012). *The 5 elements of effective thinking*. Princeton University Press. (Cit. a p. ix).
- Horn, S. & Loewenstein, G. (2021). Underestimating Learning by Doing. *Available at SSRN 3941441* (cit. a p. x).



# Elenco delle figure

1.1 Partizione dello spazio campionario $\Omega$ . . . . .	2
--	---

**Abstract** This document contains the material of the lessons of Psicometria B000286 (2021/2022) aimed at students of the first year of the Degree Course in Psychological Sciences and Techniques of the University of Florence, Italy.

**Keywords** Data science, Bayesian statistics.