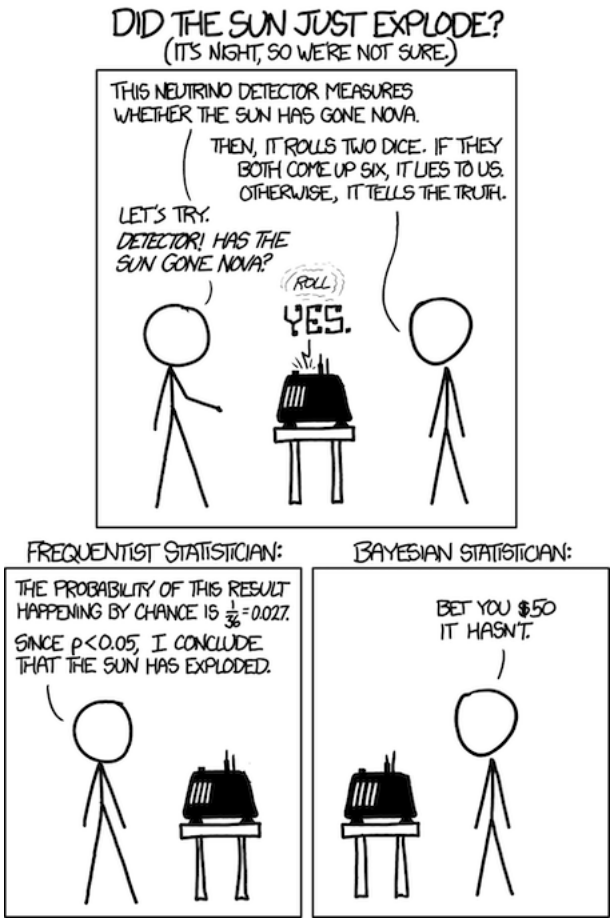
Corrado Caudek

Data Science per psicologi

Psicometria – AA 2021/2022



Indice

Elenco delle figure

Elenco delle tabelle

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. *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 un insieme di 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

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 pon-

te, 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

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

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.

- 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. Istintivamente 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 affontare, 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, opppure 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 possia-

mo 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.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare 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

[1] Ricordatevi inoltre che gli individui tendono a sottostimare la propria capacità di apprendere (Horn and Loewenstein, 2021).

un software, ricordatevi: “Google is your friend”!

Corrado Caudek
Marzo 2022

Parte I

Il calcolo delle probabilità

1

Il ragionamento scientifico: deduzione e induzione

In questa parte della dispensa verrà introdotta la teoria delle probabilità. Prima di entrare nei dettagli, cerchiamo di capire perché la probabilità è cruciale per la ricerca scientifica. Ingenuamente, potremmo pensare che il modo migliore di procedere nella ricerca scientifica sia quello di usare la logica deduttiva (aristotelica) – con un tale metodo, infatti, siamo sicuri di non commettere errori. Un esempio è il sillogismo, come

- tutti gli uomini sono mortali,
- Socrate è un uomo,
- quindi, Socrate è mortale.

La logica deduttiva, però, non può essere utilizzata in psicologia, né in alcun'altra scienza empirica. Nel sillogismo, la correttezza del ragionamento dipende dalla sua struttura e non dal significato delle parole (come uomo, mortale, ecc.). Nelle scienze empiriche, però, il "significato delle parole" è cruciale. Le "parole" usate nel sillogismo corrispondono ai "concetti teorici" (detti, in psicologia, costrutti) delle teorie scientifiche. Il problema è che la *corrispondenza* tra relazioni tra costrutti teorici, da una parte, e relazioni tra i fenomeni empirici, dall'altra, dipende dalla validità delle teorie. In fisica, ad esempio, i concetti teorici di massa (m), peso (P) e forza di gravità (g) consentono di descrivere accuratamente ciò che si osserva nel mondo empirico: $P = m \cdot g$. Non è così, invece, nelle scienze sociali, dove le relazioni tra costrutti sono in grado di descrivere *solo in parte* le relazioni tra i corrispondenti fenomeni empirici. Nelle scienze sociali, dunque, la ricerca procede mediante l'inferenza induttiva. Non siamo mai completamente sicuri della verità di una proposizione: al valore di verità di una proposizione possiamo solo assegnare un giudizio probabilistico. L'approccio bayesiano è una scuola di pensiero che usa la probabilità per quantificare il grado di credenza che viene attribuito al valore di verità di una proposizione. L'inferenza statistica bayesiana è un tipo di inferenza induttiva che ha lo scopo di quantificare quanto sia

plausibile la proposizione A dopo aver osservato l'evento B. Per quantificare la plausibilità di una proposizione l'inferenza statistica bayesiana utilizza la teoria delle probabilità. Una discussione dell'inferenza statistica bayesiana richiede dunque, preliminarmente, la conoscenze della teoria delle probabilità.

1.1 Proposizioni e modelli statistici

Nell'inferenza bayesiana, le proposizioni di una teoria scientifica sono espresse nella forma di un modello statistico, ovvero mediante una legge generale che descrive il modo in cui un fenomeno si manifesta. Tale legge generale viene anche detta *processo generativo dei dati*. Ma che cos'è un modello generativo dei dati? Per fare un esempio, consideriamo il quoziente d'intelligenza. Sappiamo che il punteggio totale della *Wechsler Adult Intelligence Scale* ha, nella popolazione, media 100 e deviazione standard 15 (dato che il test WAIS è stato costruito in modo da avere una tale proprietà). Quindi, se prendiamo un campione abbastanza grande di persone, i valori del QI di tali persone avranno, circa, media uguale a 100 e deviazione standard uguale a 15. Se con tali dati costruiamo un istogramma, sappiamo anche che il profilo di tale istogramma sarà ben descritto da una funzione matematica che va sotto il nome di *legge gaussiana*. La rappresentazione grafica della funzione gaussiana è la classica curva a campana che avrete visto tante volte in passato. La funzione gaussiana dipende da due parametri: la media (solitamente indicata con μ) e la deviazione standard (σ). Se cambiamo questi parametri, ma usiamo sempre la stessa formula, otteniamo una curva diversa. Per esempio, se consideriamo solo la sotto-popolazione dei bambini plus-dotati, la distribuzione dei punteggi QI sarà una gaussiana centrata su 130, con una qualche deviazione standard. In questo esempio, la gaussiana è il modello generatore dei dati e i parametri sono μ e σ. Per altri fenomeni, come ad esempio i tempi di reazione nel compito Stroop, o la gravità della sintomatologia ansiosa negli adulti misurata attraverso il test *Beck Anxiety Inventory*, il modello gaussiano non è più appropriato ed è necessario *ipotizzare* un diverso processo generativo dei dati.

In generale, l'inferenza induttiva bayesiana procede *ipotizzando* un modello generativo dei dati per poi, sulla base dei dati osservati in un campione e sulla base delle nostre credenze a priori, *inferire* i valori plausi-

bili dei parametri del modello. In questo processo inferenziale, possiamo individuare cinque fonti di incertezza:

1. incertezza sui parametri dei modelli;
2. incertezza su quale sia il modello migliore;
3. incertezza su cosa fare con l'output dei (migliori) modelli;
4. incertezza sul funzionamento del software che produce i risultati;
5. incertezza sul fatto che il/i modello/i (migliore/i) siano coerenti con altri campioni di dati.

- L'approccio bayesiano usa la teoria delle probabilità per descrivere l'incertezza relativa ai punti (1) e (2);
- l'approccio bayesiano si collega alla teoria delle decisioni, che prescrive come affrontare il problema descritto nel punto (3);
- il software utilizzato (nel nostro caso, Stan) fa tutto ciò che è possibile per mitigare la preoccupazione (4);
- l'approccio bayesiano consente di quantificare l'incertezza descritta al punto (5).

1.2 Oggettività e soggettività

Facendo delle assunzioni non controverse, l'inferenza bayesiana consente di aggiornare le credenze a priori sui valori (sconosciuti) dei parametri θ di un modello statistico alla luce di nuovi dati $y_1, y_2, \dots, y_N$ che vengono osservati. L'approccio bayesiano è etichettato come "soggettivo" perché non ci dice quale valore dovrebbe essere assegnato ai parametri prima di avere osservato i dati $y_1, y_2, \dots, y_N$. In realtà, l'aggiornamento bayesiano è il modo più razionale di procedere: se, prima di avere osservato i dati, il ricercatore ha una credenza assurda relativamente al valore di θ, dopo avere osservato $y_1, y_2, \dots, y_N$ le sue credenze *a posteriori* su θ, aggiornate secondo i principi bayesiani, saranno meno assurde. Il problema di questo modo di procedere non è che, a priori, i ricercatori (o chiunque altro) possono avere delle credenze sbagliate, ma bensì il fatto che, avendo osservato $y_1, y_2, \dots, y_N$, le credenze su θ non vengono aggiornate secondo i principi bayesiani. Infatti, in alcune situazioni, l'osservazione di dati che contraddicono le credenze pregresse non fa altro che rafforzare tali convinzioni errate – il problema, dunque, non nasce dalla "soggettività"

dell'approccio bayesiano, ma quanto dal fatto di non seguire un tale modo di procedere!

Lo scopo delle prossime sezioni della dispensa è quello di introdurre quei concetti base della teoria delle probabilità che risultano necessari per una presentazione delle procedure dell'inferenza induttiva bayesiana.

1.3 Che cos'è la probabilità?

La definizione della probabilità è un problema estremamente dibattuto ed aperto. Sono state fornite due possibili soluzioni al problema di definire il concetto di probabilità. (a) La natura della probabilità è "ontologica" (ovvero, basata sulla metafisica): la probabilità è una proprietà della della realtà, del mondo, di come sono le cose, indipendentemente dalla nostra esperienza. È una visione che qualcuno chiama "oggettiva".

(b) La natura della probabilità è "epistemica" (ovvero, basata sulla conoscenza): la probabilità si riferisce alla conoscenza che abbiamo del mondo, non al mondo in sé. Di conseguenza è detta, in contrapposizione alla precedente definizione, "soggettiva".

In termini epistemici, la probabilità fornisce una misura della nostra incertezza sul verificarsi di un fenomeno, alla luce delle informazioni disponibili. Potremmo dire che c'è una "scala" naturale che ha per estremi il vero (1: evento certo) da una parte ed il falso (0: evento impossibile) dall'altra. La probabilità è la quantificazione di questa scala: quantifica lo stato della nostra incertezza rispetto al contenuto di verità di una proposizione (ovvero, quantifica la plausibilità di una proposizione).

- Nell'interpretazione frequentista della probabilità, la probabilità $P(A)$ rappresenta la frequenza relativa a lungo termine nel caso di un grande numero di ripetizioni di un esperimento casuale sotto le medesime condizioni. L'evento A deve essere una proposizione relativa alle variabili casuali[1].

[1] Viene stressata qui l'idea che ciò di cui parliamo è qualcosa che emerge nel momento in cui è possibile ripetere l'esperimento casuale tante volte sotto le medesime condizioni. Le variabili casuali, infatti, forniscono una quantificazione dei risultati che si ottengono ripetendo tante volte l'esperimento casuale sotto le medesime condizioni.

- Nell'interpretazione bayesiana della probabilità $P(A)$ rappresenta il grado di credenza, o plausibilità, a proposito di A, dove A può essere qualsiasi proposizione logica.

In questo insegnamento utilizzeremo l'interpretazione bayesiana della probabilità. Possiamo citare De Finetti, ad esempio, il quale ha formulato la seguente definizione "soggettiva" di probabilità la quale risulta applicabile anche ad esperimenti casuali i cui eventi elementari non siano ritenuti ugualmente possibili e che non siano necessariamente ripetibili più volte sotto le stesse condizioni:

Definizione 1.1. La probabilità di un evento E è la quota $p(E)$ che un individuo reputa di dover pagare ad un banco per ricevere "1" ovvero "0" verificandosi o non verificandosi E. Le valutazioni di probabilità degli eventi devono rispondere ai pricipi di equità e coerenza.

I principi di equità e coerenza sono definiti come segue.

Definizione 1.2. Una scommessa risponde ai pricipi di *equità* se il ruolo di banco e giocatore sono scambiabili in ogni momento del gioco e sempre alle stesse condizioni; *coerenza* se non vi sono combinazioni di scommesse che consentano (sia al banco che al giocatore) di realizzare perdite o vincite certe.

Secondo de Finetti (1931), *"nessuna scienza ci permetterà di dire: il tale fatto accadrà, andrà così e così, perché ciò è conseguenza di tale legge, e tale legge è una verità assoluta, ma tanto meno ci condurrà a concludere scetticamente: la verità assoluta non esiste, e quindi tale fatto può accadere e può non accadere, può andare così e può andare in tutt'altro modo, nulla io ne so. Quel che si potrà dire è questo: io prevedo che il tale fatto avverrà, e avverrà nel tal modo, perché l'esperienza del passato e l'elaborazione scientifica cui il pensiero dell'uomo l'ha sottoposta mi fanno sembrare ragionevole questa previsione."*

In altri termini, de Finetti ritiene che la probabilità debba essere concepita non come una proprietà "oggettiva" dei fenomeni ("la probabilità di un fenomeno ha un valore determinato che dobbiamo solo scoprire"), ma bensì come il "grado di fiducia – in inglese *degree of belief* – di un dato soggetto, in un dato istante e con un dato insieme d'informazioni, riguardo al verificarsi di un evento". Per denotare sia la probabilità

(soggettiva) di un evento sia il concetto di *valore atteso* (che descriveremo in seguito), de Finetti (1970) utilizza il termine "previsione" (e lo stesso simbolo P): *"la previsione [...] consiste nel considerare ponderatamente tutte le alternative possibili per ripartire fra di esse nel modo che parrà più appropriato le proprie aspettative, le proprie sensazioni di probabilità."*

1.4 Variabili casuali e probabilità di un evento

Esaminiamo qui di seguito alcuni concetti di base della teoria delle probabilità.

1.4.1 Eventi e probabilità

Nella teoria delle probabilità il risultato "testa" nel lancio di una moneta è chiamato *evento*.[2] Ad esempio, $Y = 1$ denota l'evento in cui il lancio di una moneta produce come risultato testa. Il funzionale $Pr[\cdot]$ definisce la probabilità di un evento. Ad esempio, per il lancio di una moneta equilibrata, la probabilità dell'evento "il risultato del lancio della moneta è testa" è scritta come

$$Pr[Y = 1] = 0.5.$$

Se la moneta è equilibrata dobbiamo anche avere $Pr[Y = 0] = 0.5$. I due eventi $Y = 1$ e $Y = 0$ sono *mutuamente esclusivi* nel senso che non possono entrambi verificarsi contemporaneamente. Nella notazione probabilistica,

$$Pr[Y = 1 \; e \; Y = 0] = 0.$$

Gli eventi $Y = 1$ e $Y = 0$ di dicono *esaustivi*, nel senso che almeno uno di essi deve verificarsi e nessun altro tipo di evento è possibile. Nella notazione probabilistica,

$$Pr[Y = 1 \; o \; Y = 0] = 1.$$

[2] Per un ripasso delle nozioni di base della teoria degli insiemi, si veda l'Appendice **??**.

Il connettivo logico "o" specifica eventi *disgiunti*, ovvero eventi che non possono verificarsi contemporaneamente (eventi *incompatibili*) e per i quali, perciò, la probabilità della loro congiunzione è $P(A\ e\ B) = 0$. Il connettivo logico "e", invece, specifica eventi *congiunti*, ovvero eventi che possono verificarsi contemporaneamente (eventi *compatibili*) e per i quali, perciò, la probabilità della loro congiunzione è $Pr(A\ e\ B) > 0$.

1.4.2 Spazio campionario e risultati possibili

Anche se il lancio di una moneta produce sempre uno specifico risultato nel mondo reale, noi possiamo anche immaginare i possibili risultati alternativi che si sarebbero potuti osservare. Quindi, anche se in uno specifico lancio la moneta dà testa ($Y = 1$), possiamo immaginare la possibilità che il lancio possa avere prodotto croce ($Y = 0$). Tale ragionamento controfattuale è la chiave per comprendere la teoria delle probabilità e l'inferenza statistica.

I risultati possibili che si possono osservare come conseguenza del lancio di una moneta determinano i valori possibili che la variabile casuale può assumere. L'insieme di tutti i risultati possibili è chiamato *spazio campionario*. Lo spazio campionario può essere concettualizzato come un'urna contenente una pallina per ogni possibile risultato del lancio della moneta. Su ogni pallina è scritto il valore della variabile casuale. Uno specifico lancio di una moneta – ovvero, l'osservazione di uno specifico valore di una variabile casuale – è chiamato *esperimento casuale*.

Il lancio di un dado ci fornisce l'esempio di un altro esperimento casuale. Supponiamo di essere interessati all'evento "il lancio del dado produce un numero dispari". Un *evento* seleziona un sottoinsieme dello spazio campionario: in questo caso, l'insieme dei risultati $\{1, 3, 5\}$. Se esce 3, per esempio, diciamo che si è verificato l'evento "dispari" (ma l'evento "dispari" si sarebbe anche verificato anche se fosse uscito 1 o 5).

1.4.3 Variabili casuali

Sia Y il risultato del lancio di moneta equilibrata, non di un generico lancio di una moneta, ma un'istanza specifica del lancio di una specifica moneta in un dato momento. Definita in questo modo, Y è una *variabile casuale*, ovvero una variabile i cui valori non possono essere previsti con esattezza. Se la moneta è equilibrata, c'è una probabilità del 50% che il lancio della moneta dia come risultato "testa" e una probabilità del 50%

che dia come risultato "croce". Per facilitare la trattazione, le variabili casuali assumono solo valori numerici. Per lo specifico lancio della moneta in questione, diciamo, ad esempio, che la variabile casuale Y assume il valore 1 se esce testa e il valore 0 se esce croce.

Una variabile casuale può essere *discreta* o *continua*. Una variabile casuale discreta può assumere un numero finito di valori $x_1, \dots, x_n$, in corrispondenza degli eventi $E_i, \dots, E_n$ che si verificano con le rispettive probabilità $p_1, \dots, p_n$. Un esempio è il punteggio totale di un test psicometrico costituito da item su scala Likert. Invece un esempio di una variabile casuale continua è la distanza tra due punti, che può assumere infiniti valori all'interno di un certo intervallo. L'intervallo dei valori che può assumere la variabile casuale è detto *supporto* della sua distribuzione di probabilità, che può essere finito (come nel caso di una variabile casuale uniforme di supporto $[a, b]$) o infinito (nel caso di una variabile causale gaussiana il cui supporto coincide con la retta reale).

1.5 Usare la simulazione per stimare le probabilità

I metodi basati sulla simulazione consentono di stimare le probabilità degli eventi in un modo diretto, se siamo in grado di generare molteplici e casuali realizzazioni delle variabili casuali coinvolte nelle definizioni degli eventi. Per simulare il lancio di una moneta equilibrata in R iniziamo a definire un vettore che contiene i possibili risultati del lancio della moneta (ovvero i possibili valori della variabile casuale Y):

```
coin <- c(0, 1)
```

L'estrazione casuale di uno di questi due possibili valori (ovvero, la simulazione di uno specifico lancio di una moneta) si realizza con la funzione `sample()`:

```
sample(coin, size = 1)
#> [1] 0
```

In maniera equivalente, la stessa operazione si può realizzare mediante l'istruzione

```
rbinom(1, 1, 0.5)
#> [1] 1
```

Supponiamo di ripetere questo esperimento casuale 100 volte e di registrare i risultati così ottenuti. La stima della probabilità dell'evento $Pr[Y = 1]$ è data dalla frequenza relativa del numero di volte in cui abbiamo osservato l'evento di interesse $(Y = 1)$:

```
M <- 100
y <- rep(NA, M)
for (m in 1:M) {
  y[m] <- rbinom(1, 1, 0.5)
}
estimate <- sum(y) / M

cat("estimated Pr[Y = 1] =", estimate)
#> estimated Pr[Y = 1] = 0.53
```

Ripetiamo questa procedura 10 volte.

```
flip_coin <- function(M) {
  y <- rep(NA, M)
  for (m in 1:M) {
    y[m] <- rbinom(1, 1, 0.5)
  }
  estimate <- sum(y) / M
  cat("estimated Pr[Y = 1] =", estimate, "\n")
}
```

```
for (i in 1:10) {
  flip_coin(100)
}
#> estimated Pr[Y = 1] = 0.44
#> estimated Pr[Y = 1] = 0.52
#> estimated Pr[Y = 1] = 0.46
#> estimated Pr[Y = 1] = 0.57
#> estimated Pr[Y = 1] = 0.47
```

```
#> estimated Pr[Y = 1] = 0.46
#> estimated Pr[Y = 1] = 0.48
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.47
#> estimated Pr[Y = 1] = 0.62
```

Dato che la moneta è equilibrata, la stima delle probabilità dell'evento $Pr[Y = 1]$ è simile a al valore che ci aspettiamo ($Pr[Y = 1] = 0.5$), ma il risultato ottenuto nelle varie simulazioni non è sempre esatto. Proviamo ad aumentare il numero di lanci in ciascuna simulazione:

```
for (i in 1:10) {
  flip_coin(1000)
}
#> estimated Pr[Y = 1] = 0.497
#> estimated Pr[Y = 1] = 0.529
#> estimated Pr[Y = 1] = 0.493
#> estimated Pr[Y = 1] = 0.511
#> estimated Pr[Y = 1] = 0.506
#> estimated Pr[Y = 1] = 0.52
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.495
#> estimated Pr[Y = 1] = 0.489
#> estimated Pr[Y = 1] = 0.496
```

In questo secondo caso, gli errori tendono ad essere più piccoli della simulazione precedente. Cosa succede se in ciascuna simulazione esaminiamo i risultati di 10,000 lanci della moneta?

```
for (i in 1:10) {
  flip_coin(1e4)
}
#> estimated Pr[Y = 1] = 0.4885
#> estimated Pr[Y = 1] = 0.4957
#> estimated Pr[Y = 1] = 0.4902
#> estimated Pr[Y = 1] = 0.5032
#> estimated Pr[Y = 1] = 0.5048
#> estimated Pr[Y = 1] = 0.4931
```

```
#> estimated Pr[Y = 1] = 0.4965
#> estimated Pr[Y = 1] = 0.499
#> estimated Pr[Y = 1] = 0.4979
#> estimated Pr[Y = 1] = 0.4973
```

Ora le stime ottenute sono molto vicine alla vera probabilità che vogliamo stimare (cioè 0.5, perché la moneta è equilibrata). I risultati delle simulazioni precedenti pongono dunque il problema di determinare quale sia il numero di lanci di cui abbiamo bisogno per assicurarci che le stime siano accurate (ovvero, vicine al valore corretto della probabilità)

1.6 La legge dei grandi numeri

La visualizzazione mediante grafici contribuisce alla comprensione dei concetti della statistica e della teoria delle probabilità. Un modo per descrivere qjello che accade all'aumentare del numero M di ripetizioni del lancio della moneta consiste nel registrare la stima della probabilità dell'evento $Pr[Y = 1]$ in funzione del numero di ripetizioni dell'esperimento casuale per ogni $m \in 1 : M$. Un grafico dell'andamento della stima di $Pr[Y = 1]$ in funzione di m si ottiene nel modo seguente.

```
nrep <- 1e4
estimate <- rep(NA, nrep)
flip_coin <- function(m) {
  y <- rbinom(m, 1, 0.5)
  phat <- sum(y) / m
  phat
}
for (i in 1:nrep) {
  estimate[i] <- flip_coin(i)
}
d <- data.frame(
  n = 1:nrep,
  estimate
)
d %>%
```

```
  ggplot(
    aes(x = n, y = estimate)
  ) +
  geom_line() +
  theme(legend.title = element_blank()) +
  labs(
    x = "Numero di lanci della moneta",
    y = "Stima Pr[Y = 1]"
  )
```

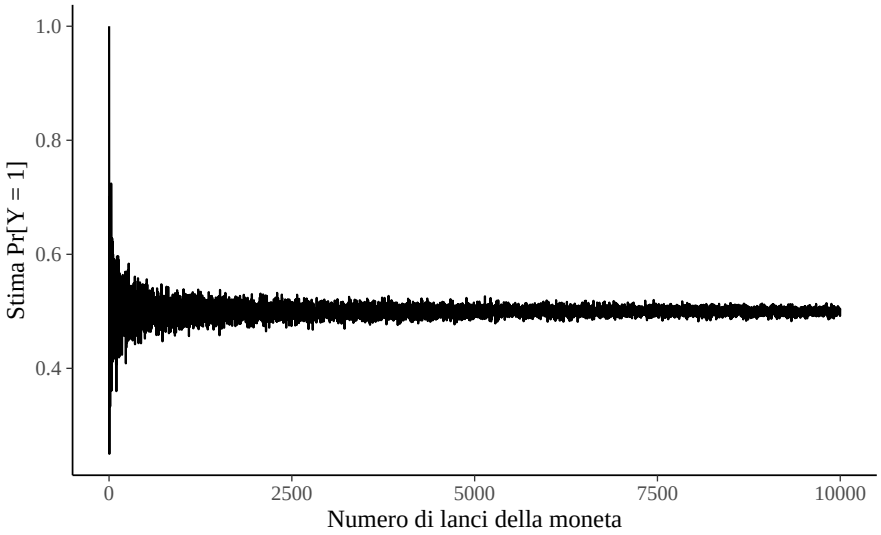


Figura 1.1: *Stima della probabilità di successo in funzione del numero di lanci di una moneta.*

Dato che il grafico 1.1 espresso su una scala lineare non rivela chiaramente l'andamento della simulazione, possiamo usare un grafico che impone una scala logaritmica sull'asse delle ascisse (x). Su scala logaritmica, i valori tra 1 e 10 vengono tracciati all'incirca con la stessa ampiezza che si osseerva per valori tra 50 e 700, eccetera.

```
d %>%
  ggplot(
    aes(x = n, y = estimate)
```

```
) +
geom_line() +
scale_x_log10(
  breaks = c(
    1, 3, 10, 50, 200,
    700, 2500, 10000
  )
) +
theme(legend.title = element_blank()) +
labs(
  x = "Numero di lanci della moneta",
  y = "Stima Pr[Y = 1]"
)
```

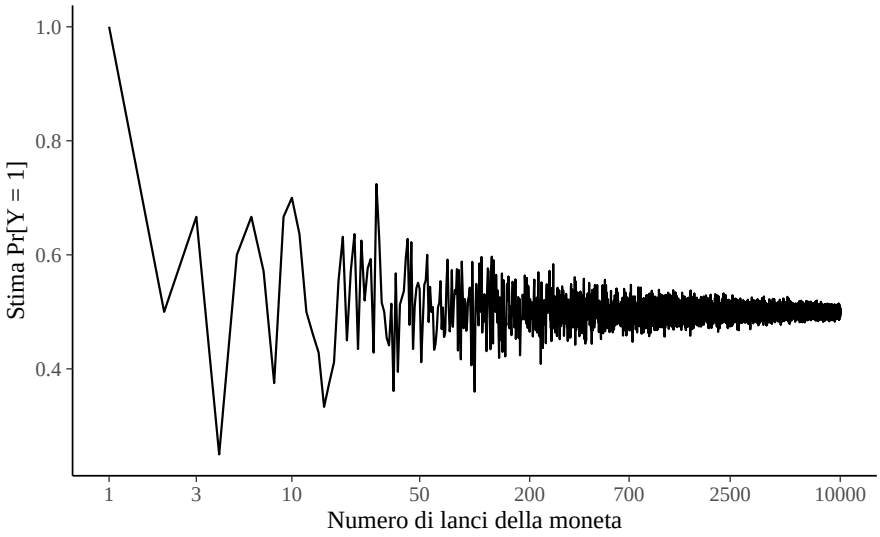


Figura 1.2: *Stima della probabilità di successo in funzione del numero di lanci di una moneta – scala logaritmica.*

La *legge dei grandi numeri* ci dice che, all'aumentare del numero di ripetizioni dell'esperimento casuale, la media dei risultati ottenuti tenderà ad avvicinarsi al valore atteso, man mano che verranno eseguite più prove. Nel caso presente, la figura 1.2 mostra appunto che, all'aumentare del numero M di lanci della moneta, la stima di $Pr[Y = 1]$ tende a convergere al vero valore di 0.5.

1.7 Variabili casuali multiple

Le variabili casuali non esistono isolatamente. Abbiamo iniziato con una singola variabile casuale Y che rappresenta il risultato di un singolo, specifico lancio di una moneta equlibrata. Ma supponiamo ora di lanciare la moneta tre volte. Ciò suggerisce che i risultati di ciascuno dei tre lanci possono essere rappresentati da una diversa variabile casuale, ad esepio, Y_1, Y_2, Y_3. Possiamo assumere che ogni lancio sia indipendente, ovvero che non dipenda dal risultato degli altri lanci. Ognuna di queste variabili Y_n per $n \in 1:3$ ha $Pr[Y_n = 1] = 0.5$ e $Pr[Y_n = 0] = 0.5$.

È possibile combinare più variabili casuali usando le operazioni aritmetiche. Se Y_1, Y_2, Y_3 sono variabili casuali che rappresentano tre lanci di una moneta equilibrata (o un lancio di tre monete equilibrate), possiamo definire la somma di tali variabili casuali come

$$Z = Y_1 + Y_2 + Y_3.$$

Possiamo simulare i valori assunti dalla variabile casuale Z simulando i valori di Y_1, Y_2, Y_3 per poi sommarli.

```
y1 <- rbinom(1, 1, 0.5)
y2 <- rbinom(1, 1, 0.5)
y3 <- rbinom(1, 1, 0.5)
c(y1, y2, y3)
#> [1] 1 0 1
z <- sum(c(y1, y2, y3))
cat("z =", z, "\n")
#> z = 2
```

ovvero,

```
y <- rep(NA, 3)
for (i in 1:3) {
  y[i] <- rbinom(1, 1, 0.5)
}
y
```

```
#> [1] 0 1 1
z <- sum(y)
cat("z =", z, "\n")
#> z = 2
```

oppure, ancora più semplicemente:

```
y <- rbinom(3, 1, 0.5)
y
#> [1] 1 0 1
z <- sum(y)
cat("z =", z, "\n")
#> z = 2
```

Possiamo ripetere questa simulazione $M = 1e5$ volte:

```
M <- 1e5
z <- rep(NA, M)
for (i in 1:M) {
  y <- rbinom(3, 1, 0.5)
  z[i] <- sum(y)
}
```

e calcolare una stima della probabilità che la variabile casuale Z assuma i valori 0, 1, 2, 3:

```
table(z) / M
#> z
#>      0      1      2      3
#> 0.1258 0.3750 0.3748 0.1244
```

Nel caso di 4 monete equilibrate, avremo:

```
M <- 1e5
z <- rep(NA, M)
for (i in 1:M) {
  y <- rbinom(4, 1, 0.5)
```

```
  z[i] <- sum(y)
}
table(z) / M
#> z
#>       0       1       2       3       4
#> 0.06340 0.24917 0.37360 0.25022 0.06361
```

Una variabile casuale le cui modalità possono essere costituite solo da numeri interi è detta *variabile casuale discreta*:

$$\mathbb{Z} = \dots, -2, -1, 0, 1, 2, \dots$$

1.8 Funzione di massa di probabilità

È conveniente avere una funzione che associa ogni possibile valore di una variabile casuale alla sua probabilità. In generale, ciò è possibile se e solo se la variabile casuale è discreta, così com'è stata definita nel Paragrafo precedente.

Ad esempio, se consideriamo $Z = Y_1 + \cdots + Y_4$ come il numero di risultati "testa" in 4 lanci della moneta, allora possiamo definire la seguente funzione:

$$\begin{array}{lcll} p_Z(0) & = & 1/16 & \text{TTTT} \\ p_Z(1) & = & 4/16 & \text{HTTT, THTT, TTHT, TTTH} \\ p_Z(2) & = & 6/16 & \text{HHTT, HTHT, HTTH, THHT, THTH, TTTH} \\ p_Z(3) & = & 4/16 & \text{HHHT, HHTH, HTHH, THHH} \\ p_Z(4) & = & 1/16 & \text{HHHH} \end{array}$$

Il lancio di quattro monete può produrre sedici possibili risultati. Dato che i lanci sono indipendenti e le monete sono equilibrate, ogni possibile risultato è ugualmente probabile. Nella tabella in alto, le sequenze dei risultati possibili del lancio delle 4 monete sono riportate nella colonna più a destra. Le probabilità si ottengono dividendo il numero di sequenze che producono lo stesso numero di eventi testa per il numero dei risultati possibili.

La funzione p_Z è stata costruita per mappare un valore u per Z alla probabilità dell'evento $Z = u$. Convenzionalmente, queste probabilità sono scritte come

$$p_Z(z) = \Pr[Z = z].$$

La parte a destra dell'uguale si può leggere come: "la probabilità che la variabile casuale Z assuma il valore z".

Una funzione definita come sopra è detta *funzione di massa di probabilità* della variabile casuale Z. Ad ogni variabile casuale discreta è associata un'unica funzione di massa di probabilità.

Una rappresentazione grafica della stima della funzione di massa di probabilità per l'esperimento casuale del lancio di quattro monete equilibrate è fornita nella figura 1.3.

```
set.seed(1234)
M <- 1e5
nflips <- 4
u <- rbinom(M, nflips, 0.5)
x <- 0:nflips
y <- rep(NA, nflips + 1)
for (n in 0:nflips) {
  y[n + 1] <- sum(u == n) / M
}
bar_plot <-
  data.frame(Z = x, count = y) %>%
  ggplot(
    aes(x = Z, y = count)
  ) +
  geom_bar(stat = "identity") +
  scale_x_continuous(
    breaks = 0:4,
    labels = c(0, 1, 2, 3, 4)
  ) +
  labs(
    y = "Probabilità stimata Pr[Z = z]"
  )
bar_plot
```

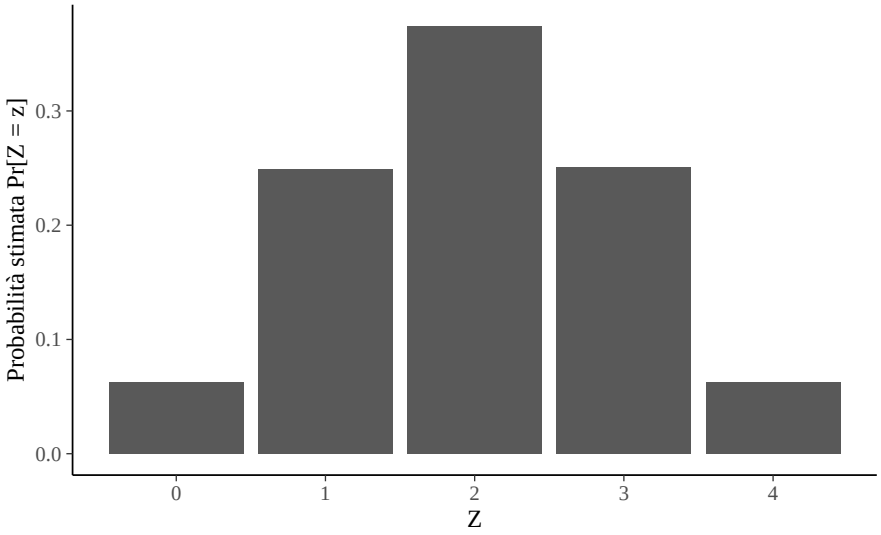


Figura 1.3: *Grafico di $M = 100\,000$ simulazioni della funzione di massa di probabilità di una variabile casuale definita come il numero di teste in quattro lanci di una moneta equilibrata.*

Se A è un sottoinsieme della variabile casuale Z, allora denotiamo con $P_z(A)$ la probabilità assegnata ad A dalla distribuzione P_z. Mediante una distribuzione di probabilità P_z è dunque possibile determinare la probabilità di ciascun sottoinsieme $A \subset Z$ come

$$P_z(A) = \sum_{z \in A} P_z(Z).$$

Esempio 1.1. Nel caso dell'esempio discusso nella Sezione 1.8, la probabilità che la variabile casuale Z sia un numero dispari è

$$Pr(\text{Z è un numero dispari}) = P_z(Z = 1) + P_z(Z = 3) = \frac{4}{16} + \frac{4}{16} = \frac{1}{2}.$$

Commenti e considerazioni finali

In questo capitolo abbiamo visto come si costruisce lo spazio campionario di un esperimento casuale, quali sono le proprietà di base della

probabilità e come si assegnano le probabilità agli eventi definiti sopra uno spazio campionario discreto. Abbiamo anche introdotto le nozioni di "variabile casuale", ovvero di una variabile che prende i suoi valori casualmente. E abbiamo descritto il modo di specificare la probabilità con cui sono presi i differenti valori, ovvero la funzione di distribuzione probabilistica $F(X) = Pr(X < x)$, e la funzione di massa di probabilità. Le procedure di analisi dei dati psicologici che discuteremo in seguito faranno un grande uso di questi concetti e della notazione qui introdotta.

2

Probabilità condizionata

Il fondamento della statistica bayesiana è il teorema di Bayes e il fondamento del teorema di Bayes è la probabilità condizionata. In questo Capitolo, inizieremo a presentare la probabilità condizionata. Nel Capitolo successivo, partendo dalla definizione di probabilità condizionata, deriveremo il teorema di Bayes.

2.1 Probabilità condizionata su altri eventi

L'attribuzione di una probabilità ad un evento è sempre condizionata dalle conoscenze che abbiamo a disposizione. Per un determinato stato di conoscenze, attribuiamo ad un dato evento una certa probabilità di verificarsi; ma se il nostro stato di conoscenze cambia, allora cambierà anche la probabilità che attribuiremo all'evento in questione. Infatti, si può pensare che tutte le probabilità siano probabilità condizionate, anche se l'evento condizionante non è sempre esplicitamente menzionato. Consideriamo il seguente problema.

Esercizio 2.1. Supponiamo che lo screening per la diagnosi precoce del tumore mammario si avvalga di test che sono accurati al 90%, nel senso che il 90% delle donne con cancro e il 90% delle donne senza cancro saranno classificate correttamente. Supponiamo che l'1% delle donne sottoposte allo screening abbia effettivamente il cancro al seno. Ci chiediamo: qual è la probabilità che una donna scelta casualmente abbia una mammografia positiva e, se ce l'ha, qual è la probabilità che abbia davvero il cancro?

Per risolvere questo problema, supponiamo che il test in questione venga somministrato ad un grande campione di donne, diciamo a 1000 donne. Di queste 1000 donne, 10 (ovvero, l'1%) hanno il cancro al seno. Per queste 10 donne, il test darà un risultato positivo in 9 casi (ovvero, nel

90% dei casi). Per le rimanenti 990 donne che non hanno il cancro al seno, il test darà un risultato positivo in 99 casi (se la probabilità di un vero positivo è del 90%, la probabilità di un falso positivo è del 10%). Questa situazione è rappresentata nella figura 2.1.

Combinando i due risultati precedenti, vediamo che il test dà un risultato positivo per 9 donne che hanno effettivamente il cancro al seno e per 99 donne che non ce l'hanno, per un totale di 108 risultati positivi. Dunque, la probabilità di ottenere un risultato positivo al test è $\frac{108}{1000} = 11\%$. Ma delle 108 donne che hanno ottenuto un risultato positivo al test, solo 9 hanno il cancro al seno. Dunque, la probabilità di avere il cancro, dato un risultato positivo al test, è pari a $\frac{9}{108} = 8\%$.

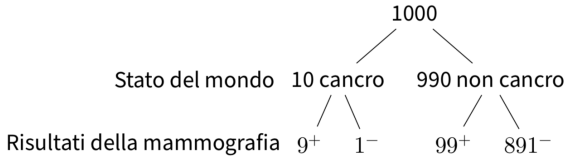


Figura 2.1: *Rappresentazione ad albero che riporta le frequenze attese dei risultati di una mammografia in un campione di 1,000 donne.*

Nell'esercizio precedente, la probabilità dell'evento "ottenere un risultato positivo al test" è una probabilità non condizionata, mentre la probabilità dell'evento "avere il cancro al seno, dato che il test ha prodotto un risultato positivo" è una probabilità condizionata. In termini generali, la probabilità condizionata $P(A \mid B)$ rappresenta la probabilità che si verifichi l'evento A sapendo che si è verificato l'evento B (oppure: la probabilità di A in una prova valida solo se si verifica anche B). Ciò ci conduce alla seguente definizione.

Definizione 2.1. Dato un qualsiasi evento A, si chiama *probabilità condizionata* di A dato B il numero

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{con } P(B) > 0, \tag{2.1}$$

dove $P(A \cap B)$ è la probabilità congiunta dei due eventi, ovvero la probabilità che si verifichino entrambi.

Nella (2.1) possiamo distinguere tra la probabilità congiunta $P(A \cap B)$, la probabilità condizionata $P(A \mid B)$ e la probabilità marginale $P(B)$.

Esercizio 2.2. Da un mazzo di 52 carte (13 carte per ciascuno dei 4 semi) ne viene estratta una in modo casuale. Qual è la probabilità che esca una figura di cuori? Sapendo che la carta estratta ha il seme di cuori, qual è la probabilità che il valore numerico della carta sia 7, 8 o 9?

Ci sono 13 carte di cuori, dunque la risposta alla prima domanda è 1/4. Per rispondere alla seconda domanda consideriamo solo le 13 carte di cuori; la probabilità cercata è dunque 3/13.

2.1.1 La fallacia del condizionale trasposto

Un errore comune che si commette è quello di credere che $P(A \mid B)$ sia uguale a $P(B \mid A)$. Tale fallacia ha particolare risalto in ambito forense tanto che è conosciuta con il nome di "fallacia del procuratore". In essa, una piccola probabilità dell'evidenza, data l'innocenza, viene erroneamente interpretata come la probabilità dell'innocenza, data l'evidenza.

Consideriamo il caso di un esame del DNA. Un esperto forense potrebbe affermare, ad esempio, che "se l'imputato è innocente, c'è solo una possibilità su un miliardo che vi sia una corrispondenza tra il suo DNA e il DNA trovato sulla scena del crimine". Ma talvolta questa probabilità è erroneamente interpretata come avesse il seguente significato: "date le prove del DNA, c'è solo una possibilità su un miliardo che l'imputato sia innocente".

Le considerazioni precedenti risultano più chiare se facciamo nuovamente riferimento all'esercizio sul tumore mammario descritto sopra. In tale esercizio abbiamo visto come la probabilità di cancro dato un risultato positivo al test sia uguale a 0.08. Tale probabilità è molto diversa dalla probabilità di un risultato positivo al test data la presenza del cancro. Infatti, questa seconda probabilità è uguale a 0.90 ed è descritta nel problema come una delle caratteristiche del test in questione.

2.2 Legge delle probabilità composte

Dalla definizione di probabilità condizionata è possibile esprimere la probabilità congiunta come prodotto di due probabilità, una condizionata

e una marginale. La legge delle probabilità composte afferma che la probabilità che si verifichino due eventi A e B è pari alla probabilità di uno dei due eventi moltiplicato con la probabilità dell'altro evento condizionato al verificarsi del primo. Per esempio se conosciamo la probabilità marginale $P(B)$ e la probabilità condizionata $P(A \mid B)$ otteniamo

$$P(A \cap B) = P(B)P(A \mid B), \tag{2.2}$$

mentre se conosciamo la probabilità marginale $P(A)$ e la probabilità condizionata $P(B \mid A)$ otteniamo

$$P(A \cap B) = P(A)P(B \mid A).$$

L'equazione (2.2) si estende al caso di n eventi $A_1, \ldots, A_n$ nella forma seguente:

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j\right) \tag{2.3}$$

Per esempio, nel caso di quattro eventi abbiamo

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot \\ P(A_4 \mid A_1 \cap A_2 \cap A_3). \end{aligned}$$

Esercizio 2.3. Da un'urna contenente 6 palline bianche e 4 nere si estrae una pallina per volta, senza reintrodurla nell'urna. Indichiamo con B_i l'evento: "esce una pallina bianca alla i-esima estrazione" e con N_i l'estrazione di una pallina nera. L'evento: "escono due palline bianche nelle prime due estrazioni" è rappresentato dalla intersezione $\{B_1 \cap B_2\}$ e la sua probabilità vale, per la (2.2)

$$P(B_1 \cap B_2) = P(B_1)P(B_2 \mid B_1).$$

$P(B_1)$ vale 6/10, perché nella prima estrazione Ω è costituito da 10 elementi: 6 palline bianche e 4 nere. La probabilità condizionata $P(B_2 \mid B_1)$ vale 5/9, perché nella seconda estrazione, se è verificato l'evento B_1, lo spazio campionario consiste di 5 palline bianche e 4 nere. Si ricava pertanto:

$$P(B_1 \cap B_2) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3}.$$

In modo analogo si ha che

$$P(N_1 \cap N_2) = P(N_1)P(N_2 \mid N_1) = \frac{4}{10} \cdot \frac{3}{9} = \frac{4}{30}.$$

Se l'esperimento consiste nell'estrazione successiva di 3 palline, la probabilità che queste siano tutte bianche vale, per la (2.3):

$$P(B_1 \cap B_2 \cap B_3) = P(B_1)P(B_2 \mid B_1)P(B_3 \mid B_1 \cap B_2),$$

dove la probabilità $P(B_3 \mid B_1 \cap B_2)$ si calcola supponendo che si sia verificato l'evento condizionante $\{B_1 \cap B_2\}$. Lo spazio campionario per questa probabilità condizionata è costituito da 4 palline bianche e 4 nere, per cui $P(B_3 \mid B_1 \cap B_2) = 1/2$ e quindi:

$$P(B_1 \cap B_2 \cap B_3) = \frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} = \frac{1}{6}.$$

La probabilità dell'estrazione di tre palline nere è invece:

$$\begin{aligned} P(N_1 \cap N_2 \cap N_3) &= P(N_1)P(N_2 \mid N_1)P(N_3 \mid N_1 \cap N_2) \\ &= \frac{4}{10} \cdot \frac{3}{9} \cdot \frac{2}{8} = \frac{1}{30}. \end{aligned}$$

2.3 L'indipendenza stocastica

Un concetto molto importante per le applicazioni statistiche della probabilità è quello dell'indipendenza stocastica. La definizione (2.1) esprime il concetto intuitivo di indipendenza di un evento da un altro, nel senso che il verificarsi di A non influisce sulla probabilità del verificarsi di B, ovvero non la condiziona. Infatti, per la definizione (2.1) di probabilità condizionata, si ha che, se A e B sono due eventi indipendenti, risulta:

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A).$$

Possiamo dunque dire che due eventi A e B sono indipendenti se

$$\begin{aligned} P(A \mid B) &= P(A), \\ P(B \mid A) &= P(B). \end{aligned} \tag{2.4}$$

Si noti inoltre che, se due eventi con probabilità non nulla sono statisticamente indipendenti, la legge delle probabilità totali espressa dalla (2.5)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{2.5}$$

si modifica nella relazione seguente:

$$P(A \cup B) = P(A) + P(B) - P(A)P(B). \tag{2.6}$$

Esercizio 2.4. Nel lancio di due dadi non truccati, si considerino gli eventi: $A = \{$esce un 1 o un 2 nel primo lancio$\}$ e $B = \{$il punteggio totale è 8$\}$. Gli eventi A e B sono indipendenti?

Rappresentiamo qui sotto lo spazio campionario dell'esperimento casuale.

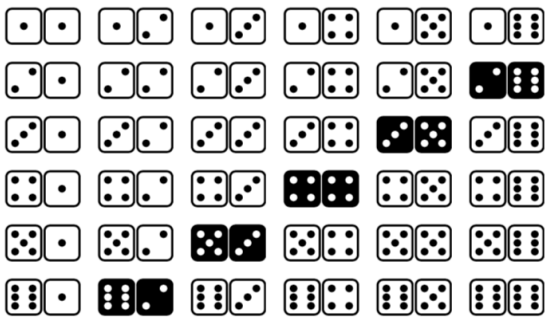


Figura 2.2: *Rappresentazione dello spazio campionario dei risultati dell'esperimento casuale corrispondente al lancio di due dadi bilanciati. Sono evidenziati gli eventi elementari che costituiscono l'evento A: esce un 1 o un 2 nel primo lancio.*

Gli eventi A e B non sono statisticamente indipendenti. Infatti, le loro probabilità valgono $P(\mathrm{A}) = 12/36$ e $P(\mathrm{B}) = 5/36$ e la probabilità della loro intersezione è

$$P(A \cap B) = 1/36 = 3/108 \neq P(A)P(B) = 5/108.$$

Osservazione. Il concetto di indipendenza è del tutto differente da quello di incompatibilità. Si noti infatti che due eventi A e B incompatibili (per i quali si ha $A \cap B = \emptyset$) sono statisticamente dipendenti, poiché il verificarsi dell'uno esclude il verificarsi dell'altro: $P(A \cap B) = 0 \neq P(A)P(B)$.

2.4 Il teorema della probabilità assoluta

Il teorema della probabilità assoluta è presente al denominatore del teorema di Bayes (che esamineremo nel prossimo capitolo) e verrà qui presentato considerando 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.

$$\mathrm{P}(E) = \mathrm{P}(E \cap H_1) + \mathrm{P}(E \cap H_2) + \mathrm{P}(E \cap H_3)$$

ovvero

$$\mathrm{P}(E) = \mathrm{P}(E \mid F_1)\mathrm{P}(F_1) + \mathrm{P}(E \mid F_2)\mathrm{P}(F_2) + \mathrm{P}(E \mid F_3)\mathrm{P}(F_3). \quad (2.7)$$

Il teorema della probabilità assoluta 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 $\mathrm{P}(E)$ è data dalla somma delle probabilità di questi tre eventi (figura 2.3). La (2.7) costituisce il denominatore del teorema di Bayes.

Esercizio 2.5. 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

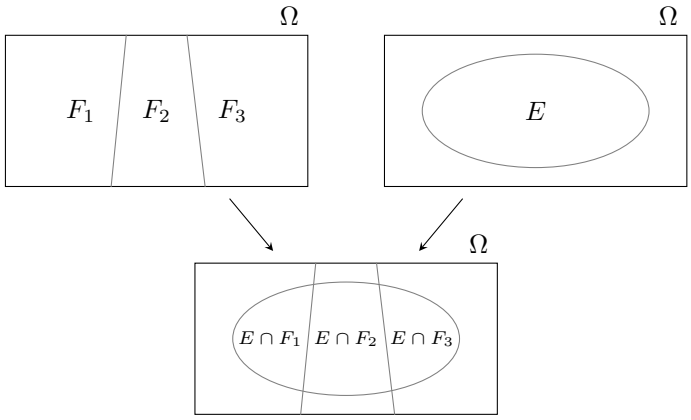


Figura 2.3: *Partizione dello spazio campionario* Ω.

$$\mathrm{P}(R \mid U_1) = 0.75, \quad \mathrm{P}(R \mid U_2) = 0.60, \quad \mathrm{P}(R \mid U_3) = 0.45.$$

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, $\mathrm{P}(U_1 \cup U_2 \cup U_3) = 1.0$. In base al teorema della probabilità assoluta, la probabilità di estrarre una pallina rossa è dunque

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

Commenti e considerazioni finali

La probabilità condizionata è importante perché ci fornisce uno strumento per precisare il concetto di indipendenza statistica. Una delle domande più importanti delle analisi statistiche è infatti quella che si chiede se due variabili sono associate tra loro oppure no. In questo Capitolo abbiamo discusso il concetto di indipendenza (come contrapposto al concetto di associazione – si veda il Capitolo **??**). In seguito vedremo come sia possibile fare inferenza sull'associazione tra variabili.

3

Il teorema di Bayes

Il teorema di Bayes assume un ruolo fondamentale nell'interpretazione soggettivista della probabilità perché descrive l'aggiornamento della fiducia che si aveva nel verificarsi di una determinata ipotesi H (identificata con la probabilità assegnata all'ipotesi stessa) in conseguenza del verificarsi dell'evidenza E.

3.1 Il teorema di Bayes

Teorema 3.1. *Sia $(H_i)_{i\geq 1}$ una partizione dell'evento certo Ω tale che*

1. $\bigcup_{i=1}^{\infty} H_i = \Omega$;
2. $H_j \cap H_j = \emptyset, i \neq j$;
3. $P(H_i) > 0, i = 1, \dots, \infty$,

Sia $E \subseteq \Omega$ un evento tale che $p(E) > 0$, allora, per $i = 1, \dots, \infty$:

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{\sum_{j=1}^{\infty} P(H_j)P(E \mid H_j)}. \tag{3.1}$$

La formula di Bayes contiene tre concetti fondamentali. I primi due distinguono il grado di fiducia precedente al verificarsi dell'evidenza E da quello successivo al verificarsi dell'evidenza E. Pertanto, dati gli eventi $H, E \subseteq \Omega$

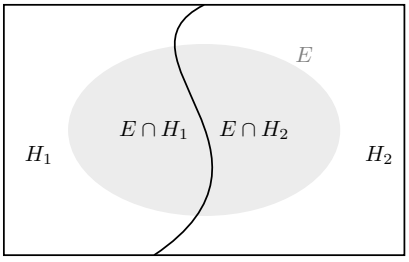
- si definisce *probabilità a priori* la probabilità che viene attribuita al verificarsi di H prima di sapere che si è verificato l'evento E, seguendo l'approccio bayesiano, ovvero tenendo conto delle caratteristiche cognitive del decisore (esperienza, modo di pensare, ecc.);

- si definisce *probabilità a posteriori* la probabilità assegnata ad H, una volta che sia noto E, ovvero l'aggiornamento della probabilità a priori alla luce della nuova evidenza E.

Il terzo concetto definisce la probabilità che ha l'evento E di verificarsi quando è vera l'ipotesi H, ovvero la probabilità dell'evidenza in base all'ipotesi. Pertanto, dati gli eventi $H, E \subseteq \Omega$

- si definisce *verosimiglianza* di H dato E, la probabilità condizionata che si verifichi E, se è vera H: $P(E \mid H)$.

Per fare un esempio concreto, considerando una partizione dell'evento certo Ω in due soli eventi che chiamiamo ipotesi H_1 e H_2. Supponiamo conosciute le probabilità a priori $P(H_1)$ e $P(H_2)$. Consideriamo un terzo evento $E \subseteq \Omega$ con probabilità non nulla di cui si conosce la verosimiglianza, ovvero si conoscono le probabilità condizionate $\mathrm{P}(E \mid H_1)$ e $P(E \mid H_2)$. Supponendo che si sia verificato l'evento E, vogliamo conoscere le probabilità a posteriori delle ipotesi, ovvero $P(H_1 \mid E)$ e $P(H_2 \mid E)$.



Per trovare le probabilità cercate scriviamo:

$$\begin{aligned}\mathrm{P}(H_1 \mid E) &= \frac{\mathrm{P}(E \cap H_1)}{\mathrm{P}(E)} \\ &= \frac{\mathrm{P}(E \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(E)}.\end{aligned}$$

Sapendo che $E = (E \cap H_1) \cup (E \cap H_2)$ e che H_1 e H_2 sono eventi disgiunti, ovvero $H_1 \cap H_2 = \emptyset$, ne segue che possiamo calcolare $\mathrm{P}(E)$ utilizzando il teorema della probabilità assoluta:

$$\begin{aligned}\mathrm{P}(E) &= \mathrm{P}(E \cap H_1) + \mathrm{P}(E \cap H_2) \\ &= \mathrm{P}(E \mid H_1)P(H_1) + \mathrm{P}(E \mid H_2)\mathrm{P}(H_2).\end{aligned}$$

Sostituendo tale risultato nella formula precedente otteniamo:

$$\mathrm{P}(H_1 \mid E) = \frac{\mathrm{P}(E \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(E \mid H_1)\mathrm{P}(H_1) + \mathrm{P}(E \mid H_2)P(H_2)}. \tag{3.2}$$

Un lettore attento si sarà reso conto che, in precedenza, abbiamo già applicato il teorema di Bayes quando abbiamo risolto l'esercizio riportato nella Sezione 2.1. In quel caso, le due ipotesi erano "malattia presente", che possiamo denotare con M, e "malattia assente", M^{C}. L'evidenza E è costituita dal risultato positivo al test, ovvero $+$. Con questa nuova notazione la (3.2) diventa:

$$\mathrm{P}(M \mid +) = \frac{\mathrm{P}(+ \mid M)\mathrm{P}(M)}{\mathrm{P}(+ \mid M)\mathrm{P}(M) + \mathrm{P}(+ \mid M^{\mathrm{C}})\mathrm{P}(M^{\mathrm{C}})}$$

Inserendo i dati nella formula, otteniamo

$$\begin{aligned}\mathrm{P}(M \mid +) &= \frac{0.9 \cdot 10/1000}{0.9 \cdot 10/1000 + 99/990 \cdot 990/1000} \\ &= \frac{9}{108}.\end{aligned}$$

Commenti e considerazioni finali

Il teorema di Bayes rende esplicito il motivo per cui la probabilità non possa essere pensata come uno stato oggettivo, quanto piuttosto come un'inferenza soggettiva e condizionata. Il denominatore del membro di destra della (3.1) è un semplice fattore di normalizzazione. Nel numeratore compaiono invece due quantità: $\mathrm{P}(H_i)$ e $\mathrm{P}(E \mid H_i)$. La probabilità $\mathrm{P}(H_i)$ è la probabilità *probabilità a priori* (*prior*) dell'ipotesi H_i e rappresenta l'informazione che l'agente bayesiano possiede a proposito dell'ipotesi H_i. Diremo che $\mathrm{P}(H_i)$ codifica il grado di fiducia che l'agente ripone in H_i precedentemente al verificarsi dell'evidenza E. Nell'interpretazione bayesiana, $\mathrm{P}(H_i)$ rappresenta un giudizio personale dell'agente e non esistono criteri esterni che possano determinare se tale giudizio sia coretto o meno. La probabilità condizionata $\mathrm{P}(E \mid H_i)$ rappresenta invece la verosimiglianza di H_i dato E e descrive la plausibilità che si verifichi l'evento E se è vera l'ipotesi H_i. Il teorema di Bayes descrive la regola che l'agente deve seguire per aggiornare il suo grado di fiducia

nell'ipotesi H_i alla luce del verificarsi dell'evento E. La $\mathrm{P}(H_i \mid E)$ è chiamata probabilità a posteriori dato che rappresenta la nuova probabilità che l'agente assegna all'ipotesi H_i affinché rimanga consistente con le nuove informazioni fornitegli da E.

La probabilità a posteriori dipende sia dall'evidenza E, sia dalla conoscenza a priori dell'agente $\mathrm{P}(H_i)$. È dunque 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 deve essere considerata come una rappresentazione del grado di fiducia soggettiva dell'agente. Dato che ogni assegnazione probabilistica rappresenta uno stato di conoscenza e che ciascun particolare stato di conoscenza è arbitrario, un accordo tra agenti diversi non è richiesto. Tuttavia, la teoria delle probabilità ci fornisce uno strumento che, alla luce di nuove evidenze, consente di aggiornare in un modo razionale il grado di fiducia che attribuiamo ad un'ipotesi, via via che nuove evidenze vengono raccolte, in modo tale da formulare un'ipotesi 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*. Vedremo nel Capitolo 4 come estendere la (3.1) al caso continuo.

Parte II

Inferenza bayesiana

4

Flusso di lavoro bayesiano

La moderna statistica bayesiana viene per lo più eseguita utilizzando un linguaggio di programmazione probabilistico implementato su computer. Ciò ha cambiato radicalmente il modo in cui venivano eseguite le statistiche bayesiane anche fin pochi decenni fa. La complessità dei modelli che possiamo costruire è aumentata e la barriera delle competenze matematiche e computazionali che sono richieste è diminuita. Inoltre, il processo di modellazione iterativa è diventato, sotto molti aspetti, molto più facile da eseguire. Anche se formulare modelli statistici complessi è diventato più facile che mai, la statistica è un campo pieno di sottigliezze che non scompaiono magicamente utilizzando potenti metodi computazionali. Pertanto, avere una buona preparazione sugli aspetti teorici, specialmente quelli rilevanti per la pratica, è estremamente utile per applicare efficacemente i metodi statistici.

4.1 Modellizzazione bayesiana

Riprendiamo ora la (3.1) per formulare il teorema di Bayes nel caso delle variabili casuali continue. Si può descrivere l'aggiornamento bayesiano facendo riferimento ad una variabile casuale Y di cui si conosce la distribuzione a meno di un parametro θ. Secondo l'approccio bayesiano, è possibile modellare l'incertezza su tale parametro mediante una variabile casuale continua Θ avente come supporto l'insieme dei valori ammissibili per il parametro cercato. La funzione di densità $p(\theta)$ di tale variabile casuale prende il nome di *distribuzione a priori* e rappresenta la sintesi delle opinioni e delle informazioni che si hanno sul parametro prima dell'osservazione dei dati. L'aggiornamento dell'incertezza su θ è determinata dal verificarsi dell'evidenza y, ovvero dall'osservazione dei risultati di un esperimento casuale. Le informazioni provenienti dal campione osservato $y = (y_1, \dots, y_n)$ sono contenute nella funzione $p(y \mid \theta)$,

che, osservata come funzione di θ per y, prende il nome di *funzione di verosimiglianza*. L'aggiornamento delle conoscenze a priori incorporate nella distribuzione iniziale $p(\theta)$ in base all'evidenza y avviene attraverso il teorema di Bayes

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_\Theta p(y \mid \theta)p(\theta)\ \mathrm{d}\theta} \quad \theta \in \Theta \tag{4.1}$$

in cui $p(\theta \mid y)$ risulta proporzionale al prodotto della probabilità a priori e della verosimiglianza e prende il nome di *distribuzione a posteriori*. Si noti che l'integrale al denominatore della (4.1) è spesso di difficile risoluzione analitica per cui l'inferenza bayesiana solitamente procede attraverso metodi di ricampionamento e metodi iterativi, quali le Catene di Markov Monte Carlo (MCMC).

Martin et al. (2022) descrive la modellazione bayesiana distinguendo tre passaggi.

1. Dati alcuni dati e alcune ipotesi su come questi dati potrebbero essere stati generati, si progetta un modello statistico combinando e trasformando variabili casuali.
2. Si usa il teorema di Bayes per condizionare il modello ai dati. Questo processo viene chiamato "inferenza" e come risultato si ottiene una distribuzione a posteriori.
3. Si critica il modello utilizzando criteri diversi, inclusi i dati e la nostra conoscenza del dominio, per verificare se abbia senso. Poiché in generale siamo incerti sul modello, a volte si confrontano modelli diversi.

Questi tre passaggi vengono eseguiti in modo iterativo e danno luogo a quello che è chiamato "flusso di lavoro bayesiano" (*bayesian workflow*).

4.1.1 Notazione

Per fissare la notazione, nel seguito y rappresenterà i dati e θ rappresenterà i parametri incogniti di un modello statistico. Sia y che θ vengono concepiti come variabili casuali. Con x vengono invece denotate le quantità note, come ad esempio i predittori del modello lineare. Per rappresentare in un modo conciso i modelli probabilistici viene usata una notazione particolare. Ad esempio, invece di scrivere $p(\theta) = \text{Beta}(1, 1)$ scriviamo $\theta \sim \text{Beta}(1, 1)$. Il simbolo "$\sim$" viene spesso letto "è distribuito

come". Possiamo anche pensare che significhi che θ costituisce un campione casuale estratto dalla distribuzione Beta(1, 1). Allo stesso modo, ad esempio, la verosimiglianza del modello binomiale può essere scritta come $y \sim \text{Bin}(n, \theta)$.

4.2 Distribuzioni a priori

Quando adottiamo un approccio bayesiano, i parametri della distribuzione di riferimento non venono considerati come delle costanti incognite ma bensì vengono trattati come variabili casuali; di conseguenza, i parametri assumono una particolare distribuzione che nelle statistica bayesiana viene definita "a priori". I parametri (o il parametro), che possiamo indicare con θ, possono assumere delle distribuzioni a priori differenti: a seconda delle informazioni disponibili bisogna selezionare una distribuzione di θ in modo tale che venga assegnata una probabilità maggiore a quei valori che si ritengono più plausibili per θ. Idealmente, le credenze a priori che portano alla specificazione di una distribuzione a priori dovrebbero essere supportate da una qualche motivazione, come ad esempio i risultati di ricerche precedenti.

4.2.1 Tipologie di distribuzioni a priori

Possiamo distinguere tra diverse distribuzioni a priori in base a quanto fortemente impegnano il ricercatore a ritenere come plausibile un particolare intervallo di valori dei parametri. Il caso più estremo è quello che rivela una totale assenza di conoscenze a priori, il che conduce alle *distribuzioni a priori non informative*, ovvero quelle che assegnano lo stesso livello di credibilità a tutti i valori dei parametri. Le distribuzioni a priori informative, d'altra parte, possono essere *debolmente informative* o *fortemente informative*, a seconda della forza della credenza che esprimono. Il caso più estremo di credenza a priori è quello che riassume il punto di vista del ricercatore nei termini di un *unico valore* del parametro, il che assegna tutta la probabilità (massa o densità) ad di un singolo valore del parametro. Poiché questa non è più una distribuzione di probabilità, sebbene ne soddisfi la definizione, in questo caso si parla di una *distribuzione a priori degenerata*. La figura seguente motra alcuni esempi di distribuzioni a priori per il modello Binomiale:

- distribuzione *non informativa*: $\theta_c \sim \text{Beta}(1, 1)$;
- distribuzione *debolmente informativa*: $\theta_c \sim \text{Beta}(5, 2)$;
- distribuzione *fortemente informativa*: $\theta_c \sim \text{Beta}(50, 20)$;
- *valore puntuale*: $\theta_c \sim \text{Beta}(\alpha, \beta)$ con $\alpha, \beta \to \infty$ e $\frac{\alpha}{\beta} = \frac{5}{2}$.

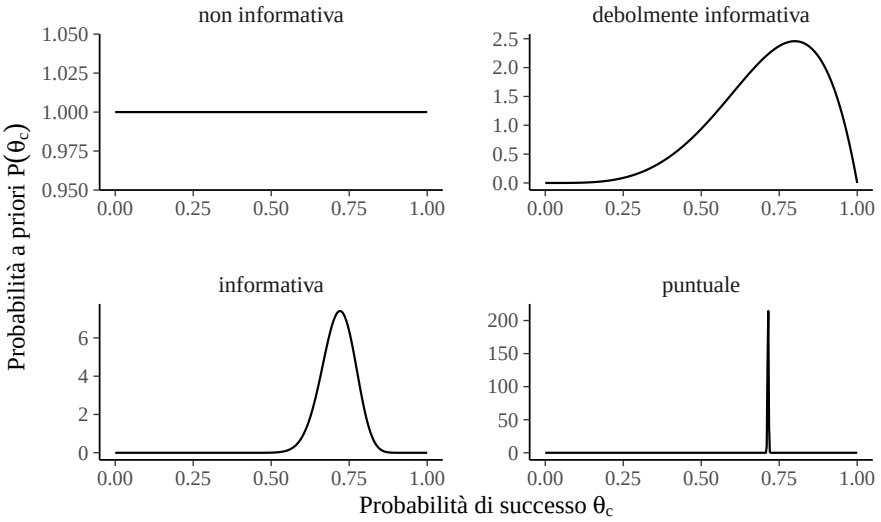


Figura 4.1: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

4.2.2 Selezione della distribuzione a priori

La selezione delle distribuzioni a priori è stata spesso vista come una delle scelte più importanti che un ricercatore fa quando implementa un modello bayesiano in quanto può avere un impatto sostanziale sui risultati finali. La soggettività delle distribuzioni a priori è evidenziata dai critici come un potenziale svantaggio dei metodi bayesiani. A questa critica, van de Schoot et al. (2021) rispondono dicendo che, al di là della scelta delle distribuzioni a priori, ci sono molti elementi del processo di inferenza statistica che sono soggettivi, ovvero la scelta del modello statistico e le ipotesi sulla distribuzione degli errori. In secondo luogo, van de Schoot et al. (2021) notano come le distribuzioni a priori svolgono due importanti ruoli statistici: quello della "regolarizzazione della stima", ovvero, il processo che porta ad indebolire l'influenza indebita di osservazioni estreme, e quello del miglioramento dell'efficienza della stima, ovvero, la facilitazione dei processi di calcolo numerico di stima

della distribuzione a posteriori. L'effetto della distribuzione a priori sulla distribuzione a posteriori verrà discusso in dettaglio nel Capitolo **??**.

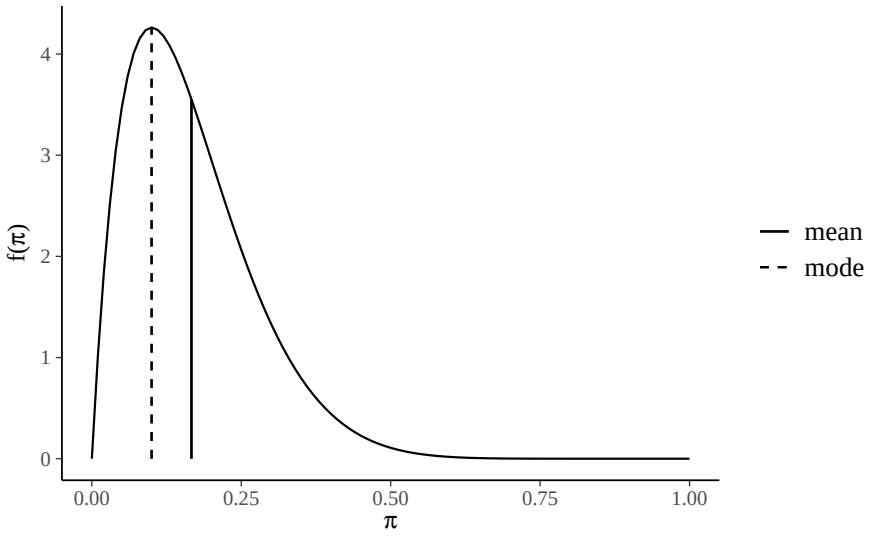
4.2.3 Un'applicazione empirica

Per introdurre la modelizzazione bayesiana useremo qui i dati riportati da Zetsche et al. (2019) (si veda l'appendice **??**). Tali dati corrispondono a 23 "successi" in 30 prove e possono dunque essere considerati la manifestazione di una variabile casuale Bernoulliana.

Se non abbiamo alcuna informazione a priori su θ (ovvero, la probabilità che l'aspettativa dell'umore futuro del partecipante sia distorta negativamente), potremmo pensare di usare una distribuzione a priori uniforme, ovvero una Beta di parametri $\alpha = 1$ e $\beta = 1$. Una tale scelta, tuttavia, è sconsigliata in quanto è più vantaggioso usare una distribuzione debolmente informativa, come ad esempio Beta$(2, 2)$, che ha come scopo la regolarizzazione, cioè quello di mantenere le inferenze in un intervallo ragionevole. Qui useremo una Beta$(2, 10)$.

$$p(\theta) = \frac{\Gamma(12)}{\Gamma(2)\Gamma(10)}\theta^{2-1}(1-\theta)^{10-1}.$$

```
bayesrules::plot_beta(alpha = 2, beta = 10, mean = TRUE, mode = TRUE)
```



La Beta$(2, 10)$ esprime la credenza che θ assume valori < 0.5, con il valore più plausibile pari a circa 0.1. Questo è assolutamente implausibile per il caso dell'esempio in discussione: la Beta$(2, 10)$ verrà usata solo per scopi didattici, ovvero, per esplorare le conseguenze di tale scelta sulla distribuzione a posteriori.

4.3 La funzione di verosimiglianza

Iniziamo con una definizione.

Definizione 4.1. La *funzione di verosimiglianza* $\mathcal{L}(\theta \mid y) = f(y \mid \theta), \theta \in \Theta$, è la funzione di massa o di densità di probabilità dei dati y vista come una funzione del parametro sconosciuto (o dei parametri sconosciuti) θ.

Detto in altre parole, le funzioni di verosimiglianza e di (massa o densità di) probabilità sono formalmente identiche, ma è completamente diversa la loro interpretazione. Nel caso della funzione di massa o di densità di probabilità la distribuzione del vettore casuale delle osservazioni campionarie y dipende dai valori assunti dal parametro (o dai parametri) θ; nel caso della la funzione di verosimiglianza la credibilità assegnata a ciascun possibile valore θ viene determinata avendo acquisita l'informazione campionaria y che rappresenta l'elemento condizionante. In altri termini, la funzione di verosimiglianza descrive in termini relativi il sostegno empirico che $\theta \in \Theta$ riceve da y. Infatti, la funzione di verosimiglianza assume forme diverse al variare di y. Possiamo dunque pensare alla funzione di verosimiglianza come alla risposta alla seguente domanda: avendo osservato i dati y, quanto risultano (relativamente) credibili i diversi valori del parametro θ? In termini più formali possiamo dire: sulla base dei dati, $\theta_1 \in \Theta$ risulta più credibile di $\theta_2 \in \Theta$ quale indice del modello probabilistico generatore dei dati se $\mathcal{L}(\theta_1) > \mathcal{L}(\theta_1)$.

Notiamo un punto importante: la funzione $\mathcal{L}(\theta \mid y)$ non è una funzione di densità. Infatti, essa non racchiude un'area unitaria.

4.3.1 Notazione

Seguendo una pratica comune, in questa dispensa spesso useremo la notazione $p(\cdot)$ per rappresentare due quantità differenti, ovvero la funzione

di verosimiglianza e la distribuzione a priori. Questo piccolo abuso di notazione riflette il seguente punto di vista: anche se la verosimiglianza non è una funzione di densità di probabilità, noi non vogliamo stressare questo aspetto, ma vogliamo piuttosto pensare alla verosimiglianza e alla distribuzione a priori come a due elementi che sono egualmente necessari per calcolare la distribuzione a posteriori. In altri termini, per così dire, questa notazione assegna lo stesso status epistemologico alle due diverse quantità che si trovano al numeratore della regola di Bayes.

4.3.2 La log-verosimiglianza

Dal punto di vista pratico risulta più conveniente utilizzare, al posto della funzione di verosimiglianza, il suo logaritmo naturale, ovvero la funzione di log-verosimiglianza:

$$\ell(\theta) = \log \mathcal{L}(\theta). \tag{4.2}$$

Poiché il logaritmo è una funzione strettamente crescente (usualmente si considera il logaritmo naturale), allora $\mathcal{L}(\theta)$ e $\ell(\theta)$ assumono il massimo (o i punti di massimo) in corrispondenza degli stessi valori di θ (per un approfondimento, si veda l'Appendice **??**):

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \ell(\theta) = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta).$$

Per le proprietà del logaritmo, si ha

$$\ell(\theta) = \log \left(\prod_{i=1}^{n} f(y \mid \theta) \right) = \sum_{i=1}^{n} \log f(y \mid \theta). \tag{4.3}$$

Si noti che non è necessario lavorare con i logaritmi, ma è fortemente consigliato. Il motivo è che i valori della verosimiglianza, in cui si moltiplicano valori di probabilità molto piccoli, possono diventare estremamente piccoli – qualcosa come 10^{-34}. In tali circostanze, non è sorprendente che i programmi dei computer mostrino problemi di arrotondamento numerico. Le trasformazioni logaritmiche risolvono questo problema.

4.3.3 Un'applicazione empirica

Se i dati di Zetsche et al. (2019) possono essere riassunti da una proporzione allora è sensato adottare un modello probabilistico binomiale quale meccanismo generatore dei dati:

$$y \sim \text{Bin}(n, \theta), \tag{4.4}$$

laddove θ è la probabiltà che una prova Bernoulliana assuma il valore 1 e n corrisponde al numero di prove Bernoulliane. Questo modello assume che le prove Bernoulliane y_i che costituiscono il campione y siano tra loro indipendenti e che ciascuna abbia la stessa probabilità $\theta \in [0, 1]$ di essere un "successo" (valore 1). In altre parole, il modello generatore dei dati avrà una funzione di massa di probabilità

$$p(y \mid \theta) \; = \; \text{Bin}(y \mid n, \theta).$$

Nei capitoli precedenti è stato mostrato come, sulla base del modello binomiale, sia possibile assegnare una probabilità a ciascun possibile valore $y \in \{0, 1, \dots, n\}$ *assumendo noto il valore del parametro* θ. Ma ora abbiamo il problema inverso, ovvero quello di fare inferenza su θ alla luce dei dati campionari y. In altre parole, riteniamo di conoscere il modello probabilistico che ha generato i dati, ma di tale modello non conosciamo i parametri: vogliamo dunque ottenere informazioni su θ avendo osservato i dati y.

Per i dati di Zetsche et al. (2019) la funzione di verosimiglianza corrisponde alla funzione binomiale di parametro $\theta \in [0, 1]$ sconosciuto. Abbiamo osservato un "successo" 23 volte in 30 "prove", dunque, $y = 23$ e $n = 30$. La funzione di verosimiglianza diventa

$$\mathcal{L}(\theta \mid y) = \frac{(23 + 7)!}{23!7!}\theta^{23} + (1 - \theta)^7. \tag{4.5}$$

Per costruire la funzione di verosimiglianza dobbiamo applicare la (4.5) tante volte, cambiando ogni volta il valore θ ma *tenendo sempre costante il valore dei dati*. Per esempio, se poniamo $\theta = 0.1$

$$\mathcal{L}(\theta \mid y) = \frac{(23 + 7)!}{23!7!}0.1^{23} + (1 - 0.1)^7$$

otteniamo

```
dbinom(23, 30, 0.1)
#> [1] 9.737e-18
```

Se poniamo $\theta = 0.2$

$$\mathcal{L}(\theta \mid y) = \frac{(23+7)!}{23!7!} 0.2^{23} + (1-0.2)^7$$

otteniamo

```
dbinom(23, 30, 0.2)
#> [1] 3.581e-11
```

e così via. La figura 4.2 — costruita utilizzando 100 valori equispaziati $\theta \in [0, 1]$ — fornisce una rappresentazione grafica della funzione di verosimiglianza.

```
n <- 30
y <- 23
theta <- seq(0, 1, length.out = 100)
like <- choose(n, y) * theta^y * (1 - theta)^(n - y)
tibble(theta, like) %>%
  ggplot(aes(x = theta, y = like)) +
  geom_line() +
  labs(
    y = expression(L(theta)),
    x = expression("Valori possibili di" ~ theta)
  )
```

Come possiamo interpretare la curva che abbiamo ottenuto? Per alcuni valori θ la funzione di verosimiglianza assume valori piccoli; per altri valori θ la funzione di verosimiglianza assume valori più grandi. Questi ultimi sono i valori di θ più credibili e il valore 23/30 (la moda della funzione di verosimiglianza) è il valore più credibile di tutti.

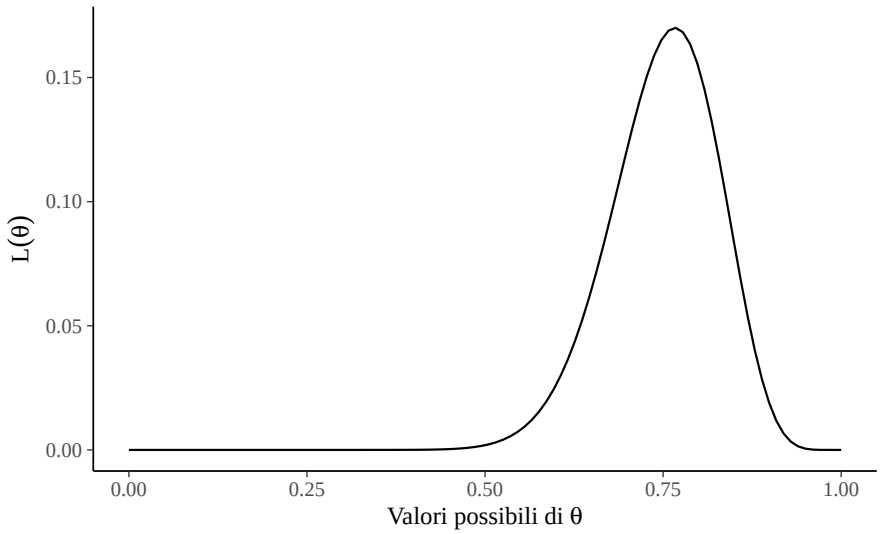


Figura 4.2: *Funzione di verosimiglianza nel caso di 23 successi in 30 prove.*

4.4 La verosimiglianza marginale

Per il calcolo di $p(\theta \mid y)$ è necessario dividere il prodotto tra la distribuzione a priori e la verosimiglianza per una costante di normalizzazione. Tale costante di normalizzazione, detta *verosimiglianza marginale*, ha lo scopo di fare in modo che $p(\theta \mid y)$ abbia area unitaria.

Si noti che, nel caso di variabili continue, la verosimiglianza marginale è espressa nei termini di un integrale. Tranne in pochi casi particolari, tale integrale non ha una soluzione analitica. Per questa ragione, l'inferenza bayesiana procede calcolando una approssimazione della distribuzione a posteriori mediante metodi numerici.

4.4.1 Un'applicazione empirica

Consideriamo nuovamente i dati di Zetsche et al. (2019). Supponiamo che nel numeratore bayesiano la verosimiglianza sia moltiplicata per una distribuzione uniforme, ovvero Beta$(1, 1)$. In tali circostanze, il prodotto si riduce alla funzione di verosimiglianza. Per i dati di Zetsche et al. (2019), dunque, la costante di normalizzazione si ottiene marginalizzando

la funzione di verosimiglianza $p(y = 23, n = 30 \mid \theta)$ sopra θ, ovvero risolvendo l'integrale:

$$p(y = 23, n = 30) = \int_0^1 \binom{30}{23} \theta^{23}(1-\theta)^7 \, \mathrm{d}\theta. \tag{4.6}$$

Una soluzione numerica si trova facilmente usando R:

```
like_bin <- function(theta) {
  choose(30, 23) * theta^23 * (1 - theta)^7
}
integrate(like_bin, lower = 0, upper = 1)$value
#> [1] 0.03226
```

La derivazione analitica è fornita nell'Appendice **??**.

4.5 Distribuzione a posteriori

La distribuzione a postreriori si trova applicando il teorema di Bayes:

$$\text{probabilità a posteriori} = \frac{\text{probabilità a priori} \cdot \text{verosimiglianza}}{\text{costante di normalizzazione}}$$

Una volta trovata la distribuzione a posteriori, possiamo usarla per derivare altre quantità di interesse. Questo viene generalmente ottenuto calcolando il seguente valore atteso:

$$J = \int f(\theta) p(\theta \mid y) \, \mathrm{d}y$$

Se $f(\cdot)$ è la funzione identità, ad esempio, J risulta essere la media di θ:

$$\bar{\theta} = \int_\Theta \theta p(\theta \mid y) \, \mathrm{d}\theta.$$

Ripeto qui quanto detto sopra: le quantità di interesse della statistica bayesiana(costante di normalizzazione, valore atteso della distribuzione a posteriori, ecc.) contengono integrali che risultano, nella maggior parte

dei casi, impossibili da risolvere analiticamente. Per questo motivo, si ricorre a metodi di stima numerici, in particolare a quei metodi Monte Carlo basati sulle proprietà delle catene di Markov (MCMC). Questo argomento verrà discusso nel Capitolo **??**.

4.6 Distribuzione predittiva a priori

La distribuzione a posteriori è l'oggetto centrale nella statistica bayesiana, ma non è l'unico. Oltre a fare inferenze sui valori dei parametri, potremmo voler fare inferenze sui dati. Questo può essere fatto calcolando la *distribuzione predittiva a priori*:

$$p(y^*) = \int_\Theta p(y^* \mid \theta) p(\theta) \, \mathrm{d}\theta. \tag{4.7}$$

La (4.7) descrive la distribuzione prevista dei dati in base al modello (che include la distribuzione a priori e la verosimiglianza), ovvero descrive i dati y^* che ci aspettiamo di osservare, dato il modello, prima di avere osservato i dati del campione.

È possibile utilizzare campioni dalla distribuzione predittiva a priori per valutare e calibrare i modelli utilizzando le nostre conoscenze dominio-specifiche. Ad esempio, ci possiamo chiedere: "È sensato che un modello dell'altezza umana preveda che un essere umano sia alto -1.5 metri?". Già prima di misurare una singola persona, possiamo renderci conto dell'assurdità di questa domanda. Se la distribuzione prevista dei dati consente domande di questo tipo (ovvero, prevede di osservare dati che risultano insensati alla luce delle nostre conoscenze dominio-specifiche), è chiaro che il modello deve essere riformulato.

4.7 Distribuzione predittiva a posteriori

Un'altra quantità utile da calcolare è la distribuzione predittiva a posteriori:

$$p(\tilde{y} \mid y) = \int_{\Theta} p(\tilde{y} \mid \theta) p(\theta \mid y) \, \mathrm{d}\theta. \tag{4.8}$$

Questa è la distribuzione dei dati attesi futuri $\tilde{y}$ alla luce della distribuzione a posteriori $p(\theta \mid y)$, che a sua volta è una conseguenza del modello adottato (distribuzione a priori e verosimiglianza) e dei dati osservati. In altre parole, questi sono i dati che il modello si aspetta dopo aver osservato i dati de campione. Dalla (4.8) possiamo vedere che le previsioni sui dati attesi futuri sono calcolate integrando (o marginalizzando) sulla distribuzione a posteriori dei parametri. Di conseguenza, le previsioni calcolate in questo modo incorporano l'incertezza relativa alla stima dei parametri del modello.

Commenti e considerazioni finali

Questo Capitolo ha brevemente passato in rassegna i concetti di base dell'inferenza statistica bayesiana. In base all'approccio bayesiano, invece di dire che il parametro di interesse di un modello statistico ha un valore vero ma sconosciuto, diciamo che, prima di eseguire l'esperimento, è possibile assegnare una distribuzione di probabilità, che chiamano stato di credenza, a quello che è il vero valore del parametro. Questa distribuzione a priori può essere nota (per esempio, sappiamo che la distribuzione dei punteggi del QI è normale con media 100 e deviazione standard 15) o può essere del tutto arbitraria. L'inferenza bayesiana procede poi nel modo seguente: si raccolgono alcuni dati e si calcola la probabilità dei possibili valori del parametro alla luce dei dati osservati e delle credenze a priori. Questa nuova distribuzione di probabilità è chiamata "distribuzione a posteriori" e riassume l'incertezza dell'inferenza.

Bibliografia

de Finetti, B. (1931). Probabilismo. *Logos*, pages 163–219.

de Finetti, B. (1970). *Teoria delle probabilità: sintesi introduttiva con appendice critica*. Einaudi.

Horn, S. and Loewenstein, G. (2021). Underestimating learning by doing. *Available at SSRN 3941441*.

Martin, O. A., Kumar, R., and Lao, J. (2022). *Bayesian Modeling and Computation in Python*. CRC Press.

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J., and Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primer*, 1(1):1–26.

Zetsche, U., Bürkner, P.-C., and Renneberg, B. (2019). Future expectations in clinical depression: Biased or realistic? *Journal of Abnormal Psychology*, 128(7):678–688.