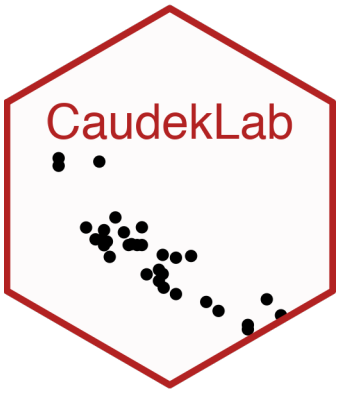
Psicometria

Corrado Caudek

2021-12-19

Questo documento è stato realizzato con:

- 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

Data della versione presente: Dicembre 19, 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 psico-

logico 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 verrano 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. 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 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.[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

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

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

Inferenza statistica bayesiana

Capitolo 1

Il problema inverso

Pensare in termini bayesiani significa aggiornare le nostre credenze combinando le credenze iniziali con le informazioni fornite dai dati. Ciò produce una credenza "a posteriori". L'aggiornamento bayesiano richiede che le credenze siano descritte nei termini di un modello probabilistico formulato in termini di uno o più parametri. La nostra incertezza riguarda il valore dei parametri. L'aggiornamento bayesiano ha lo scopo di ottenere le migliori stime possibili dei parametri del modello, alla luce delle nostre credenze a priori e dei dati osservati.

1.1 Inferenza bayesiana come un problema inverso

L'inferenza bayesiana può essere descritta come la soluziome di un problema inverso mediante la regola di Bayes, ovvero la quantificazione della plausibilità di una teoria alla luce dei dati osservati – (si veda la Sezione **??**).

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 saranno concepiti come delle variabili casuali.[1] Con x verranno 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).

[1] Nell'approccio bayesiano si fa riferimento ad un modello probabilistico $f(y \mid \theta)$ rappresentativo del fenomeno d'interesse noto a meno del valore assunto dal parametro (o dei parametri) che lo caratterizza. Si fa inoltre riferimento ad una distribuzione congiunta (di massa o di densità di probabilità) $f(y, \theta)$. Entrambi gli argomenti della funzione y e θ hanno natura di variabili casuali, laddove la nostra incertezza relativa a y è dovuta alla naturale variabilità del fenomeno indagato (*variabilità aleatoria*), mentre la nostra incertezza relativa a θ è dovuta alla mancata conoscenza del suo valore numerico (*variabilità epistemica*).

Allo stesso modo, per l'esempio presente, la verosimiglianza può essere scritta come:

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

Significato dei parametri del modello

Il parametro di un modello è un valore che influenza la credibilità dei dati. Ad esempio, il singolo parametro θ del modello binomiale determina la forma della funzione di verosimiglianza binomiale. Ricordiamo che, per il modello binomiale, la funzione di verosimiglianza è:

$$p(y \mid \theta, n) = \text{Bin}(y, n, \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}.$$

Funzioni di probabilità

Nell'aggiornamento bayesiano vengono utilizzate le seguenti distribuzioni di probabilità (o di massa di probabilità):

- la *distribuzione a priori* $p(\theta)$ — la credenza iniziale (prima di avere osservato i dati $Y = y$) riguardo a θ;
- la *funzione di verosimiglianza* $p(y \mid \theta)$ — quanto sono compatibili i dati osservati $Y = y$ con i diversi valori possibili di θ?
- la *verosimiglianza marginale* $p(y)$ — costante di normalizzazione: qual è la probabilità complessiva di osservare i dati $Y = y$? In termini formali:

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

- la *distribuzione a posteriori* $p(\theta \mid y)$ — la nuova credenza relativa alla credibilità di ciascun valore θ dopo avere osservato i dati $Y = y$.

La regola di Bayes

Assumendo un modello statistico, la formula di Bayes consente di giungere alla distribuzione a posteriori $p(\theta \mid y)$ per il parametro di interesse θ, come indicato dalla seguente catena di equazioni:

$$\begin{aligned}
p(\theta \mid y) &= \frac{p(\theta, y)}{p(y)} && \text{[definizione di probabilità condizionata]} && (1.1)\\
&= \frac{p(y \mid \theta)\, p(\theta)}{p(y)} && \text{[legge della probabilità composta]} && (1.2)\\
&= \frac{p(y \mid \theta)\, p(\theta)}{\int_\Theta p(y, \theta) \ \mathrm{d}\theta} && \text{[legge della probabilità totale]} && (1.3)\\
&= \frac{p(y \mid \theta)\, p(\theta)}{\int_\Theta p(y \mid \theta)\, p(\theta) \ \mathrm{d}\theta} && \text{[legge della probabilità composta]} && (1.4)\\
&\propto p(y \mid \theta)\, p(\theta) && && (1.5)
\end{aligned}$$

La regola di Bayes "inverte" la probabilità della distribuzione a posteriori $p(\theta \mid y)$, esprimendola nei termini della funzione di verosimiglianza $p(y \mid \theta)$ e della distribuzione a priori $p(\theta)$. L'ultimo passo è importante per la stima della distribuzione a posteriori mediante i metodi Monte Carlo a catena di Markov, in quanto per questi metodi richiedono soltanto che le funzioni di probabilità siano definite a meno di una costante di proporzionalità. In altri termini, per la maggior parte degli scopi dell'inferenza inversa, è sufficiente calcolare la densità a posteriori non normalizzata, ovvero è possibile ignorare

il denominatore bayesiano $p(y)$. La distribuzione a posteriori non normalizzata, dunque, si riduce al prodotto della varosimiglianza e della distribuzione a priori.

Possiamo dire che la regola di Bayes viene usata per aggiornare le credenze a priori su θ (ovvero, la distribuzione a priori) in modo tale da produrre le nuove credenze a posteriori $p(\theta \mid y)$ che combinano le informazioni fornite dai dati y con le credenze precedenti. La distribuzione a posteriori riflette dunque l'aggiornamento delle credenze del ricercatore alla luce dei dati. La distribuzione a posteriori $p(\theta \mid y)$ contiene tutta l'informazione riguardante il parametro θ e viene utilizzata per produrre indicatori sintetici, per la determinazione di stime puntuali o intervallari, e per la verifica d'ipotesi.

La (1.5) rende evidente che, in ottica bayesiana, la quantità di interesse θ non è fissata (come nell'impostazione frequentista), ma è una variabile casuale la cui distribuzione di probabilità è influenzata sia dalle informazioni a priori sia dai dati a disposizione. In altre parole, nell'approccio bayesiano non esiste un valore vero di θ, ma invece lo scopo è quello di fornire invece un giudizio di probabilità (o di formulare una "previsione", nel linguaggio di de Finetti). Prima delle osservazioni, sulla base delle nostre conoscenze assegnamo a θ una distribuzione a priori di probabilità. Dopo le osservazioni, correggiamo il nostro giudizio e assegniamo a θ una distribuzione a posteriori di probabilità.

1.2 Aggiornamento bayesiano per v.c. discrete

Per introdurre la procedura dell'aggiornamento bayesiano discuteremo un esempio riguardante la sfida tra Gary Kasparov e il supercomputer Deep Blue (Johnson et al., 2022). Nel 1996 Kasparov giocò sei partite contro Deep Blue vincendone tre, perdendone una, con due patte, e vincendo così il match. Nel 1997 si svolse la rivincita, sempre al meglio di sei partite. Nella rivincita, Kasparov vinse una partita, perse due partite, con tre patte, perdendo dunque il match. Quello che vogliamo fare è descrivere la nostra credenza relativa all'abilità di Kasparov di battere Deep Blue, alla luce delle credenze iniziali che possiamo avere avuto e considerati i dati relativi alla sfida del 1997.

Modello probabilistico

Se i dati rappresentano una proporzione, come in precedenza, possiamo adottare un modello probabilistico binomiale:

$$y \sim \text{Bin}(n, \theta), \tag{1.6}$$

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. Nel caso presente, il modello probabilistico è quello binomiale. Noi vogliamo

ottenere informazioni sul valore di θ conoscendo il numero y di successi osservati nel campione.

La (1.6) è un modello probabilistico. Tale modello non spiega perché, in ciascuna realizzazione, Y assuma un particolare valore. Questo modello deve piuttosto essere inteso come un costrutto matematico che ha lo scopo di riflettere alcune proprietà del processo corrispondente ad una sequenza di prove Bernoulliane. In questo senso, è simile al modello di Isaac Newton dei moti planetari che utilizza equazioni differenziali. Le equazioni non sono i pianeti, ma solo descrizioni di come si muovono i pianeti in risposta alle forze gravitazionali. Modelli come quello di Newton ci permettono di prevedere alcuni fenomeni, come il moto dei pianeti, ad esempio. Ma in generale i modelli sono solo delle approssimazioni del fenomeno che vogliono descrivere. Anche il modello di Newton, che produce previsioni estremamente accurate di ciò che possiamo osservare a occhio nudo a proposito del moto dei corpi celesti, è solo un'approssimazione dei modelli del moto e dei fenomeni gravitazionali che, in seguito, sono stati introdotti da Albert Einstein. E anche tali modelli successivi sono, a loro volta, solo un caso speciale della più generale teoria della relatività. In altre parole, modelli sempre migliori vengono proposti, laddove ogni successivo modello è migliore di quello precedente in quanto ne migliora le capacità di previsione, è più generale, o è più elegante.

Una parte del lavoro della ricerca in tutte le scienze consiste nel verificare le assunzioni dei modelli e, se necessario, nel migliorare i modelli dei fenomeni considerati. Un modello viene giudicato in relazione al suo obiettivo. Se l'obiettivo del modello molto semplice che stiamo discutendo è quello di prevedere la proporzione di casi nei quali $y_i = 1$, $i = 1, \dots, n$, allora un modello con un solo parametro come quello che abbiamo introdotto sopra può essere sufficiente. Ma l'evento $y_i = 1$ (supponiamo: superare l'esame di Psicometria, oppure risultare positivi al COVID-19) dipende da molti fattori e se vogliamo rendere conto di una tale complessità, un modello come quello che stiamo discutendo qui certamente non sarà sufficiente.

Per concludere, un modello è un costrutto matematico il cui scopo è quello di rappresentare un qualche aspetto della realtà. Il valore di un tale strumento dipende dalla sua capacità di ottenere lo scopo per cui è stato costruito.

Il problema inverso

Nel modello probabilistico che stiamo esaminando, il termine n viene trattato come una costante nota e θ come una *variabile casuale*. Il parametro θ del modello rappresenta la probabilità che ciascuna prova Bernoulliana sia un "successo". Dato che θ è incognito, ma abbiamo a disposione un campione di dati, l'inferenza su θ può essere svolta, mediante la regola di Bayes, costruendo la distribuzione a posteriori $p(\theta \mid y)$. Una volta ottenuta la distribuzione a posteriori possiamo riassumerla, ad esempio, riportando l'intervallo centrale al 95% della distribuzione di densità, ovvero

$$\Pr\left[0.025 \leq \theta \leq 0.975 \;\middle|\; Y = y\right].$$

Se vogliamo sapere, per esempio, se la probabilità di $y_i = 1$ sia maggiore di 0.5, possiamo calcolare la probabilità dell'evento

$$\Pr\left[\theta > \frac{1}{2} \;\middle|\; Y = y\right].$$

Distribuzione a priori

Per scopi didattici, Johnson et al. (2022) ipotizzano che le credenze a priori relative a π (la probabilità Kasparov che batta Deep Blue) siano le seguenti:

θ	0.2	0.5	0.8	Totale
$P(\theta)$	0.10	0.65	0.15	1.0

Anche se $\theta \in [0, 1]$ è una variabile continua, per semplificare la discussione, Johnson et al. (2022) considerano solo tre possibili valori di $\theta \in \{0.2, 0.5, 0.8\}$ e assegnano a tali valori possibili le probabilità indicate sopra. Questa tabella descrive le credenze iniziali relative alla capacità di Kasparov di battere Deep Blue; indica che le credenze iniziali pongono la massa maggiore della nostra credenza sull'evento $\theta = 0.5$ — in altre parole, più che ogni altra possibilità, a priori crediamo che Kasparov abbia solo il 50% di possibilità di battere Deep Blue.

Verosimiglianza

La sfida del 1997 ci fornisce i dati: una vittoria su sei partite; considerate le due vittorie di Deep Blue e le tre patte, Kasparov perse dunque il match. Per formulare la funzione di verosimiglianza dobbiamo utilizzare un modello statistico che descriva il "processo generatore" dei dati che abbiamo osservato. Semolifichiamo la situazione descrivendo i dati nei termini di un successo su sei prove. Ipotizziamo inoltre che le sei partite siano indipendenti le une dalle altre e che la probabilità di vittoria di Kasparov rimanga costante nelle sei partite. Descritta la situazione in questi termini, possiamo individuare nel modello binomiale il processo statistico che potrebbe avere generato i dati che abbiamo osservato. Questo modello probabilistico è formulato nei termini di un parametro: θ, ovvero la probabilità di vittoria (di Kasparov). Nel Paragrafo precedente abbiamo descritto la distribuzione di probabilità a priori del parametro θ. Poniamoci ora il problema di descrivere la verosimiglianza dei valori θ alla luce dei dati osservati (ovvero, solo una vittoria su sei partite).

Nel caso di una v.c. discreta, la verosimiglianza si ottiene utilizzando la distribuzione di massa di probabilità espressa in funzione dei parametri, quando i dati vengono tenuti costanti. Nel caso presente, la funzione di massa di probabilità è quella binomiale, ovvero

$$p(y \mid \theta) = \binom{n}{y} \cdot \theta^y (1-\theta)^{N-y}$$

e i dati sono $y = 1$ vittoria su $n = 6$ partite.

In questo esercizio, θ assume solo tre valori: $\theta_1 = 0.2$, $\theta_2 = 0.5$ e $\theta_3 = 0.8$. Svolgiamo ora i calcoli. Quando $\theta = 0.2$ otteniamo

$$p(y \mid \theta_1) = \binom{6}{1} \cdot 0.2^1 (1-0.2)^{6-1} = 0.39322,$$

con $\theta = 0.5$ otteniamo

$$p(y \mid \theta_2) = \binom{6}{1} \cdot 0.5^1 (1-0.5)^{6-1} = 0.09375,$$

con $\theta = 0.8$ otteniamo

$$p(y \mid \theta_3) = \binom{6}{1} \cdot 0.8^1 (1-0.8)^{6-1} = 0.00154.$$

Lo stesso risultato si ottiene nel modo seguente usando R:

```
theta <- c(0.2, 0.5, 0.8)
dbinom(1, 6, theta)
#> [1] 0.39322 0.09375 0.00154
```

I tre valori che abbiamo trovato costisuiscono (in questo esempio semplificato) la funzione di verosimiglianza di θ alla luce dei dati. La funzione di verosimiglianza ci dice quanto risultano compatibili i dati ($y = 1$, $n = 6$) con i possibili valori del parametro θ, considerato il modello probabilistico binomiale.

θ	0.2	0.5	0.8
$p(y = 1 \mid \theta)$	0.39322	0.09375	0.00154

Nel caso presente, i dati (ovvero, una vittoria su sei partite) risultano maggiormente compatibili con il valore $\theta = 0.2$ (Kasparov è molto meno forte di Deep Blue): la verosimiglianza di $\theta = 0.2$ è 0.39. All'altro estremo, i dati risultano poco compatibili con il valore $\theta = 0.8$ (Kasparov è molto più forte di Deep Blue): la verosimiglianza di $\theta = 0.8$ è 0.0015.

La costante di normalizzazione

Per calcolare la distribuzione a posteriori del parametro θ il prodotto tra la distribuzione a priori e la verosimiglianza va diviso per una costante di normalizzazione. La costante di normalizzazione (o verosimiglianza marginale) ha lo scopo di fare in modo che la distribuzione a posteriori di θ (quando tale v.c. è discreta) sommi ad 1.0. La *verosimiglianza marginale* $p(y = 1, n = 6)$ si ottiene marginalizzando la funzione di verosimiglianza

$$p(y = 1, n = 6 \mid \theta) = \binom{6}{1}\theta^1(1-\theta)^5.$$

sopra θ: per ogni possibile valore θ, si moltiplica il valore della verosimiglianza in corrisopndenza di θ per la sua probabilità a priori di θ; si sommano poi tutti i prodotti ottenuti in questo modo.

Nell'esempio precedente abbiamo considerato solo tre possibili valori θ: $\theta_1 = 0.2$, $\theta_2 = 0.5$ e $\theta_3 = 0.8$. A tali valori θ sono state assegnate le probabilità a priori $p(\theta_1) = 0.10$, $p(\theta_2) = 0.65$ e $p(\theta_3) = 0.15$. Date queste informazioni possiamo calcolare la verosimiglianza marginale come segue:

$$\begin{aligned} p(y = 1, n = 6) &= \binom{6}{1}\theta_1^1(1-\theta_1)^5 \cdot p(\theta_1) \\ &= \binom{6}{1}\theta_2^1(1-\theta_2)^5 \cdot p(\theta_2) \\ &= \binom{6}{1}\theta_3^1(1-\theta_3)^5 \cdot p(\theta_3), \end{aligned}$$

ovvero

$$\begin{aligned} p(y = 1, n = 6) = &\binom{6}{1}0.2^1(1-0.2)^5 \cdot 0.10 \\ &+ \binom{6}{1}0.5^1(1-0.5)^5 \cdot 0.65 \\ &+ \binom{6}{1}0.8^1(1-0.8)^5 \cdot 0.15 \\ &\approx 0.0638. \end{aligned} \tag{1.7}$$

È dunque possibile considerare la verosimiglianza marginale come una sorta di media ponderata della verosimiglianza, nella quale i "pesi" dipendono dalla credibilità dei valori del parametro.

Svolgendo i calcoli con R otteniamo

```
prior <- c(0.10, 0.25, 0.65)
sum(dbinom(1, 6, theta) * prior)
#> [1] 0.0638
```

Quello che stiamo discutendo è però un esempio artificiale perché al parametro θ sono stati attribuiti solo tre valori possibili. In realtà, θ è una variabile casuale continua. Vedremo in seguito come si affronta questo problema.

Distribuzione a posteriori

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

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

Nel caso presente abbiamo

$$p(\theta \mid y = 1) = \frac{p(y = 1 \mid \theta) p(\theta)}{p(y = 1)}, \quad \theta \in \{0.2, 0.5, 0.8\}.$$

Svolgendo i calcoli in R otteniamo

```
(dbinom(1, 6, theta) * prior) / sum(dbinom(1, 6, theta) * prior)
#> [1] 0.6167 0.3676 0.0157
```

In conclusione, dopo avere osservato i dati corrispondenti ad una sola vittoria di Kasparov sul sei partite, la *previsione* (per usare il linguaggio di de Finetti) che Kasparov sia il giocatore dominante ($\theta = 0.8$) scende da 0.65 (a priori) a 0.0157 (a posteriori). Nella distribuzione a posteriori, invece, lo scenario che riceve il supporto maggiore è quello che descrive Kasparov come il giocatore più debole ($\theta = 0.2$) – la previsione dell'evento $\theta = 0.2$ è 0.6167.

1.3 Aggiornamento bayesiano per v.c. continue

Riprendiamo ora gli stessi concetti descritti nell'esempio precedente formulandoli in un modo più generale. Descriveremo inoltre l'aggiornamento bayesiano facendo riferimento alle variabili casuali continue.

La distribuzione a priori sui parametri

Quando adottiamo un approccio bayesiano, i parametri non sono delle costanti incognite ma delle variabili casuali governate da una propria legge di distribuzione delle probabilità (probabilità a priori). La distribuzione a priori sui valori dei parametri $p(\theta)$ è parte integrante del modello statistico. Ciò implica che due modelli bayesiani possono condividere la stessa funzione di verosimiglianza, ma tuttavia devono essere considerati come modelli diversi se specificano diverse distribuzioni a priori. Ciò significa che, quando diciamo "Modello binomiale", intendiamo in realtà un'intera classe di modelli, ovvero tutti i possibili modelli che hanno la stessa verosimiglianza ma diverse distribuzioni a priori su θ.

Nell'analisi dei dati bayesiana, la distribuzione a priori $p(\theta)$ codifica le credenze del ricercatore a proposito dei valori dei parametri, prima di avere osservato i dati.

Idealmente, le credenze a priori che supportano la specificazione di una distribuzione a priori dovrebbero essere supportate da una qualche motivazione, come ad esempio i risultati di ricerche precedenti, o altre motivazioni giustificabili.

Quando una nuova osservazione (p. es., vedo un cigno bianco) corrisponde alle mie credenze precedenti (p. es., la maggior parte dei cigni sono bianchi) la nuova osservazione rafforza le mie credenze precedenti: più nuove osservazioni raccolgo (p. es., più cigni bianchi vedo), più forti diventano le mie credenze precedenti.

Tuttavia, quando una nuova osservazione (p. es., vedo un cigno nero) non corrisponde alle mie credenze precedenti, ciò contribuisce a diminuire la certezza che attribuisco alle mie credenze: tanto maggiori diventano le osservazioni non corrispondenti alle mie credenze (p. es., più cigni neri vedo), tanto più si indeboliscono le mie credenze. Fondamentalmente, tanto più forti sono le mie credenze precedenti, di tante più osservazioni incompatibili (ad esempio, cigni neri) ho bisogno per cambiare idea.

Pertanto, da una prospettiva bayesiana, l'incertezza intorno ai parametri di un modello *dopo* aver visto i dati (ovvero le distribuzioni a posteriori) deve includere anche le credenze precedenti. Se questo modo di ragionare sembra molto intuitivo, non è una coincidenza: vi sono infatti diverse teorie psicologiche che prendono l'aggiornamento bayesiano come modello di funzionamento di diversi processi cognitivi.

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) su di un singolo valore di un 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 mostra esempi di distribuzioni a priori non informative, debolmente o fortemente informative, così come una distribuzione a priori espressa nei termini di un valore puntuale per il modello Binomiale. Le distribuzione a priori illustrate di seguito sono le seguenti:

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

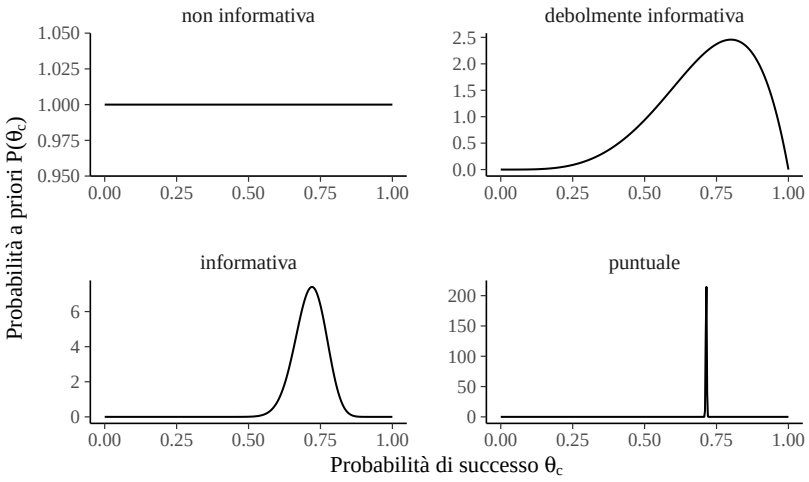


Figura 1.1: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

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 nel Capitolo **??**.

Verosimiglianza marginale

Al denominatore della regola di Bayes abbiamo la verosimiglianza marginale $p(y)$. Tale denominatore è espresso nei termini di un integrale che, tranne in pochi casi particolari, non ha una soluzione analitica. Per questa ragione, l'inferenza bayesiana procede calcolando una approssimazione della distribuzione a posteriori mediante metodi numerici.

L'esempio presentato sopra era artificiale perché al parametro θ sono stati attribuiti solo tre possibili valori. In realtà, θ può assumere tutti i possibili valori compresi nell'intervallo $[0, 1]$ e dunque la somma che dobbiamo calcolare avrà infiniti addendi. Dal punto di vista matematico, una tale somma corrisponde all'integrale:

$$p(y = 1, n = 6) = \int_0^1 \binom{6}{1} \theta^1 (1 - \theta)^5 \, \mathrm{d}\theta.$$

L'integrale precedente descrive esattamente le stesse operazioni che abbiamo discusso nell'esempio "artificiale" in cui θ poteva assumere solo tre valori, eccetto che ora dobbiamo eseguire la somma dei prodotti calcolati su tutti gli infiniti valori θ. Questo integrale

corrisponde alla marginalizzazione del parametro θ. Non è tuttavia necessario eseguire una tale operazione di maginalizzazione in forma analitica in quanto il precedente integrale può essere calcolato con R:

```
BinLik <- function(theta) {
  choose(6, 1) * theta^1 * (1 - theta)^5
}
integrate(BinLik, lower = 0, upper = 1)$value
#> [1] 0.143
```

Soluzione analitica

Qui di seguito è riportata la derivazione analitica. Sia $\theta \sim \mathrm{B}(a, b)$ e sia $y = \{y_1, \dots, y_n\} \sim \mathrm{Bin}(\theta, n)$. Ponendo

$$\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)},$$

la verosimiglianza marginale diventa

$$\begin{aligned} p(y) &= \binom{n}{y} \int p(y \mid \theta) p(\theta) \, \mathrm{d}\theta \\ &= \binom{n}{y} \int_0^1 \theta^y (1-\theta)^{n-y} \frac{1}{\mathrm{B}(a, b)} \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta \\ &= \binom{n}{y} \frac{1}{\mathrm{B}(a, b)} \int_0^1 \theta^{y+a-1} (1-\theta)^{n-y+b-1} \, \mathrm{d}\theta \\ &= \binom{n}{y} \frac{\mathrm{B}(y + a, n - y + b)}{\mathrm{B}(a, b)}, \end{aligned} \tag{1.8}$$

in quanto

$$\begin{aligned} \int_0^1 \frac{1}{\mathrm{B}(a, b)} \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta &= 1 \\ \frac{1}{\mathrm{B}(a, b)} \int_0^1 \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta &= 1 \\ \int_0^1 \theta^{a-1} (1-\theta)^{b-1} \, \mathrm{d}\theta &= \mathrm{B}(a, b). \end{aligned}$$

Continuiamo con l'esempio precedente. Per replicare il risultato trovato per via numerica con R, assumiamo una distribuzione a priori uniforme, ovvero $\mathrm{B}(1, 1)$. I valori del problema sono i seguenti:

```
a <- 1
b <- 1
y <- 1
n <- 6
```

e dunque

```
alpha <- y + a
beta <- n - y + b
```

Definiamo

```
B <- function(a, b) {
  (gamma(a) * gamma(b)) / gamma(a + b)
}
```

Il risultato cercato si ottiene con

```
choose(6, 1) * B(alpha, beta) / B(a, b)
#> [1] 0.143
```

In conclusione, nel caso di una verosimiglianza binomiale $y =\sim \text{Bin}(\theta, n)$ e di una distribuzione a priori $\theta \sim \text{B}(a, b)$, la verosimiglianza marginale diventa

$$\binom{n}{y} \frac{\text{B}(y + a, n - y + b)}{\text{B}(a, b)}. \tag{1.9}$$

La distribuzione a posteriori

Ci sono due metodi principali per calcolare la distribuzione a posteriori $p(\theta \mid y)$:

- una precisa derivazione matematica formulata nei termini della distribuzione a priori coniugata alla distribuzione a posteriori (si veda il Capitolo **??**); tale procedura però ha un'applicabilità molto limitata;
- un metodo approssimato, molto facile da utilizzare in pratica, che dipende da metodi Monte Carlo basati su Catena di Markov (MCMC).

Una volta calcolata la distribuzione a posteriori dobbiamo riassumerla in qualche modo. Questo problema verrà discusso nel Capitolo **??**.

Considerazioni conclusive

Questo Capitolo ha brevemente passato in rassegna alcuni 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. I concetti importanti che abbiamo appreso in questo Capitolo sono quelli di distribuzione a priori, verosimiglianza, verosimiglianza marginale e distribuzione a posteriori. Questi sono i concetti fondamentali della statistica bayesiana.

Bibliografia

Burger, E. B. & Starbird, M. (2012). *The 5 elements of effective thinking*. Princeton University Press. (Cit. a p. viii).

Horn, S. & Loewenstein, G. (2021). Underestimating Learning by Doing. *Available at SSRN 3941441* (cit. a p. ix).

Johnson, A. A., Ott, M. & Dogucu, M. (2022). *Bayes Rules! An Introduction to Bayesian Modeling with R*. CRC Press. (Cit. alle pp. 5, 7).

van de Schoot, R., Depaoli, S., King, R., Kramer, B., Märtens, K., Tadesse, M. G., Vannucci, M., Gelman, A., Veen, D., Willemsen, J. & Yau, C. (2021). Bayesian statistics and modelling. *Nature Reviews Methods Primer*, *1*(1), 1–26 (cit. a p. 11).

Elenco delle figure

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.