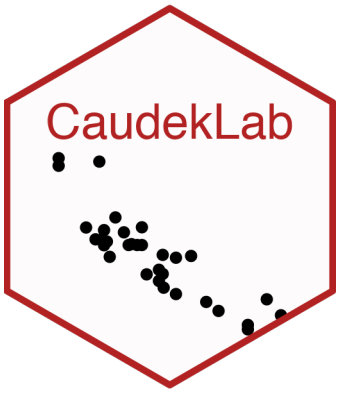
Psicometria

Corrado Caudek

2021-12-20

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

Copyright © 2022.

Data della versione presente: Dicembre 20, 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, così da ottenere una credenza "a posteriori". L'aggiornamento bayesiano richiede che le credenze siano descritte nei termini di un modello probabilistico formulato nei 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, considerando le informazioni fornite dalle nostre credenze a priori e dai dati osservati. In questo capitolo verrà descritto il significato di tutti e tre i termini a destra del segno di uguale nella formula di Bayes: la distribuzione a priori e la funzione di verosimiglianza al numeratore, la verosimiglianza marginale al denominatore.

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 il Capitolo **??**).

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). Allo stesso modo, ad esempio, la verosimiglianza del modello binomiale può essere scritta come $y \sim \text{Bin}(n, \theta)$.

[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*).

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 $Y = y$.

In questo Capitolo ci limiteremo ad introdurre le tre quantità che vengono utilizzate nella regola di Bayes: la distribuzione a priori, la verosimiglianza e la verosimiglianza marginale. Nei capitoli successivi vedremo come sia possibile, mediante queste tre distribuzione, giungere alla distribuzione a posteriori $p(\theta \mid y)$.

1.2 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[2]:

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

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

[2] In realtà, avremmo dovuto scrivere $p(\theta \mid y, \mathcal{M})$, in quanto non condizioniamo la stima di θ solo rispetto ai dati y ma anche ad un modello probabilistico $\mathcal{M}$ che viene assunto quale meccanismo generatore dei dati. Per semplicità di notazione, omettiamo il riferimento a $\mathcal{M}$.

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.

Un esempio di aggiornamento bayesiano

Per descrivere l'aggiornamento bayesiano, in questo Capitolo (così come nei successivi) considereremo i dati di Zetsche et al. (2019). Questi ricercatori si sono chiesti se gli individui depressi manifestino delle aspettative accurate circa il loro umore futuro, oppure se tali aspettative siano distorte negativamente. Esamineremo qui i 30 partecipanti dello studio di Zetsche et al. (2019) che hanno riportato la presenza di un episodio di depressione maggiore in atto. All'inizio della settimana di test, a questi pazienti è stato chiesto di valutare l'umore che si aspettavano di esperire nei giorni seguenti della settimana. Mediante una app, i partecipanti dovevano poi valutare il proprio umore in cinque momenti diversi di ciascuno dei cinque giorni successivi. Lo studio considera diverse emozioni, ma qui ci concentriamo solo sulla tristezza.

Sulla base dei dati forniti dagli autori, abbiamo calcolato la media dei giudizi relativi al livello di tristezza raccolti da ciascun partecipante tramite la app. Tale media è stata poi sottratta dall'aspettativa del livello di tristezza fornita all'inizio della settimana. La discrepanza tra aspettative e realtà è stata considerata come un evento dicotomico: valori positivi di tale differenza indicano che le aspettative circa il livello di tristezza erano maggiori del livello di tristezza effettivamente esperito — ciò significa che le aspettative future risultano negativamente distorte (evento codificato con "1"). Viceversa, si ha che le aspettative risultano positivamente distorte se la differenza descritta in precedenza assume un valore negativo (evento codificato con "0").

Nel campione dei 30 partecipanti clinici di Zetsche et al. (2019), le aspettative future di 23 partecipanti risultano distorte negativamente e quelle di 7 partecipanti risultano distorte positivamente. Chiameremo θ la probabilità dell'evento "le aspettative del partecipante sono distorte negativamente". Ci poniamo il problema di ottenere una stima a posteriori di θ avendo osservato 23 "successi" in 30 prove.[3]

1.3 Modello probabilistico

Nel caso dello studio di Zetsche et al. (2019), i dati qui considerati possono essere considerati la manifestazione di una variabile casuale Bernoulliana – 23 "successi" in 30 prove. Se i dati rappresentano una proporzione, allora possiamo adottare un modello probabilistico binomiale quale meccanismo generatore dei dati:

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

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

[3] Si noti un punto importante: dire semplicemente che la stima di θ è uguale a $23/30 = 0.77$ ci porta ad ignorare il livello di incertezza associato a tale stima. Infatti, lo stesso valore (0.77) si può ottenere come 23/30, o 230/300, o 2300/3000, o 23000/30000, ma l'incertezza di una stima pari a 0.77 è molto diversa nei quattro casi. Quando si traggono conclusioni dai dati è invece necessario quantificare il livello della nostra incertezza relativamente alla stima del parametro di interesse (nel caso presente, θ). Lo strumento ci consente di quantificare tale incertezza è la distribizione a posteriori $p(\theta \mid y)$. Ovviamente, $p(\theta \mid y)$ assume forme molto diverse nei quattro casi descritti sopra.

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.

Nel modello probabilistico che stiamo esaminando, il termine n viene trattato come una costante nota e θ come una *variabile casuale*. Dato che θ è incognito, ma abbiamo a disposione i dati y, svolgeremo l'inferenza su θ mediante la regola di Bayes per determinare la distribuzione a posteriori $p(\theta \mid y)$.

Si noti che il modello probabilistico (1.6) 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. 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. 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. 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.

1.4 Distribuzioni a priori

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 vi 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 $\frac{\alpha}{\beta} = \frac{5}{2}$.

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

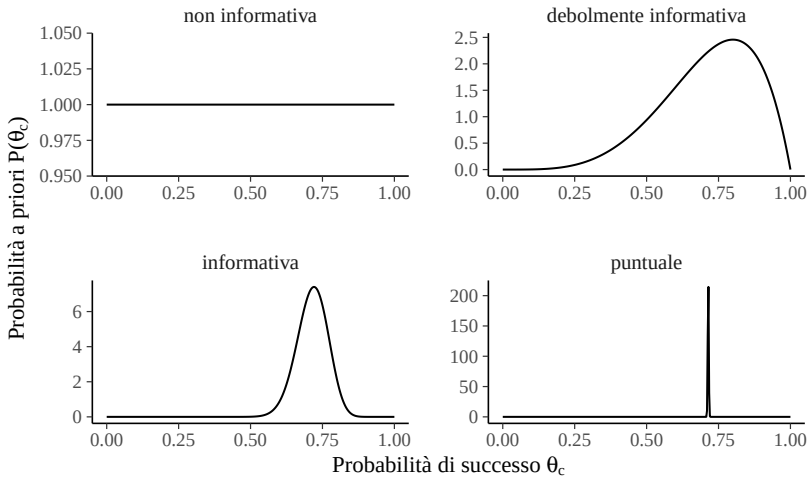


Figura 1.1: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

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 **??**.

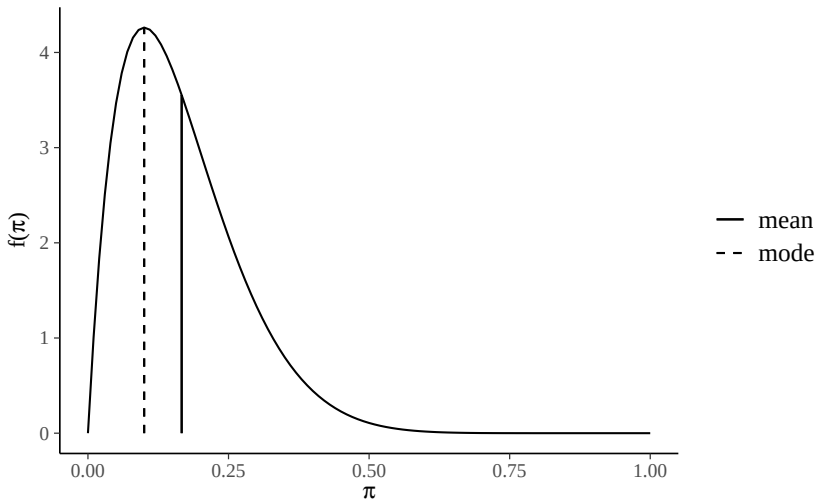
La distribuzione a priori per i dati di Zetsche et al. (2019)

In un problema concreto di analisi dei dati, la scelta della distribuzione a priori dipende dalle credenze a priori che vogliamo includere nell'analisi dei dati. Se non abbiamo alcuna informazione a priori, potremmo pensare di usare una distribuzione a priori uniforme, ovvero una Beta di parametri $\alpha = 1$ e $\beta = 1$. Questa, tuttavia, è una cattiva idea perché il risultato ottenuto non è invariante a seconda della trasformazione della scala dei dati (ad esempio, se esprimiamo l'altezza in cm piuttosto che in m). Il problema della *riparametrizzazione* verrà discusso nel Capitolo ?? **TODO**. È invece raccomandato usare una distribuzione a priori vagamente informativa, come ad esempio Beta$(2, 2)$.

Nella presente discussione, per fare un esempio, quale distribuzione a priori useremo una Beta$(2, 10)$, ovvero:

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

```
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, nel caso dell'esempio in discussione. Adotteremo una tale distribuzione a priori solo per scopi didattici, per esplorare le conseguenze di tale scelta (molto più sensato sarebbe stato usare Beta$(2, 2)$).

1.5 Verosimiglianza

Oltre alla distribuzione a priori di θ, nel numeratore della regola di Bayes troviamo la funzione di verosimigliana. Iniziamo dunque con una definizione.

Definizione 1.1. 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 è lo strumento che consente di rispondere alla seguente domanda: avendo osservato i dati y, quanto risultano (relativamente) credibili i diversi valori del parametro θ?

Spesso per indicare la verosimiglianza si scrive $\mathcal{L}(\theta)$ se è chiaro a quali valori y ci si riferisce. La verosimiglianza $\mathcal{L}$ è una curva (in generale, una superficie) nello spazio Θ del parametro (in generale, dei parametri θ) che riflette la credibilità relativa dei valori θ alla luce dei dati osservati. Notiamo un punto importante: la funzione $\mathcal{L}(\theta \mid y)$ non è una funzione di densità. Infatti, essa non racchiude un'area unitaria.

In conclusione, 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 (lasciamo come esercizio da svolgere la verifica di questa affermazione).

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:

$$\ell(\theta) = \log \mathcal{L}(\theta). \tag{1.7}$$

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:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \ell(\theta).$$

Per le proprietà del logaritmo, si ha

$$\ell(\theta) = \log \left(\prod_{i=1}^{n} f(y \mid \theta) \right) = \sum_{i=1}^{n} \log f(y \mid \theta). \tag{1.8}$$

Si noti che non è necessario lavorare con i logaritmi, anche se è fortemente consigliato, e questo perché 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.

PRATICA GUIDATA

Si trovi e si interpreti la funzione di verosimiglianza per i dati di Zetsche et al. (2019).

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 $n = 30$. Per i dati di Zetsche et al. (2019), la funzione di verosimiglianza diventa

$$\mathcal{L}(\theta \mid y) = \frac{(23 + 7)!}{23!7!}\theta^{23} + (1 - \theta)^7. \tag{1.9}$$

La definizione precedente ci dice che, per costruire la funzione di verosimiglianza, dobbiamo applicare la (1.9) 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] 9.74e-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] 3.58e-11
```

e così via. La figura 1.2 — 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)
  )
```

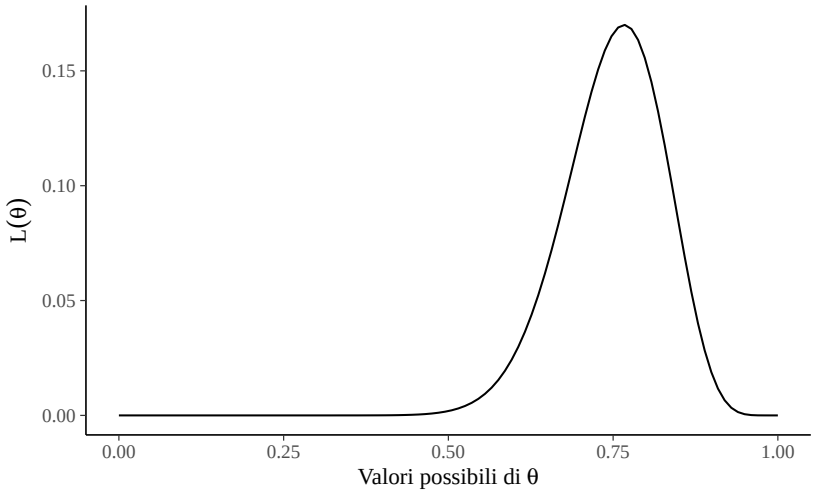


Figura 1.2: *Funzione di verosimiglianza nel caso di 23 successi in 30 prove.*

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 $23/30$ è il valore più credibile di tutti. La funzione di verosimiglianza di θ valuta la compatibilità dei dati osservati $Y = y$ con i diversi possibili valori θ. In termini più formali possiamo dire che la funzione di verosimiglianza ha la seguente interpretazione: sulla base dei dati, $\theta_1 \in \Theta$ è più credibile di $\theta_2 \in \Theta$ come indice del modello probabilistico generatore delle osservazioni se $\mathcal{L}(\theta_1) > \mathcal{L}(\theta_1)$. Alcune considerazioni sulla stima di massima verosimiglianza sono fornite nell'Appendice A.

1.6 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 il denominatore della regola di Bayes (ovvero la verosimiglianza marginale) è sempre espresso 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.

PRATICA GUIDATA

Si trovi la verosimiglianza maginale per i dati di Zetsche et al. (2019).

Supponiamo che nel numeratore bayesiano la verosimiglianza sia moltiplicata per una distribuzione uniforme, Beta(1, 1). In questo caso, il prodotto si riduce alla funzione di verosimiglianza. In riferimento ai dati di Zetsche et al. (2019), la costante di normalizzazione per si ottiene semplicemente marginalizzando la funzione di verosimiglianza

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

$$p(y = 23, n = 30) = \int_0^1 \binom{30}{23} \theta^{23}(1-\theta)^7 \, \mathrm{d}\theta. \tag{1.10}$$

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] 0.0323
```

La derivazione analitica della costante di normalizzazione qui discussa è fornita nell'Appendice B.

1.7 Distribuzione a posteriori

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

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

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); questo problema verrà discusso nel Capitolo ??

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.

Appendice A

La stima di massima verosimiglianza

La funzione di verosimiglianza rappresenta la "credibilità relativa" dei valori del parametro di interesse. Ma qual è il valore più credibile? Se utilizziamo soltanto la funzione di verosimiglianza, allora la risposta è data dalla stima di massima verosimiglinza.

Definizione A.1. Un valore di θ che massimizza $\mathcal{L}(\theta \mid y)$ sullo spazio parametrico Θ è detto *stima di massima verosimiglinza* (s.m.v.) di θ ed è indicato con $\hat{\theta}$:

$$\hat{\theta} = \text{argmax}_{\theta \in \Theta} \mathcal{L}(\theta). \tag{A.1}$$

Il paradigma frequentista utilizza la funzione di verosimiglianza quale unico strumento per giungere alla stima del valore più credibile del parametro sconosciuto θ. Tale stima corrisponde al punto di massimo della funzione di verosimiglianza. Nell'esempio presente, $\hat{\theta} = 0.6667$. Il massimo della funzione di verosimiglianza, ovvero $\hat{\theta}$, si può ottenere con metodi numerici o grafici.

In base all'approccio bayesiano, invece, il valore più credibile del parametro sconosciuto θ non corrisponde alla s.m.v.. Per l'approccio bayesiano, invece, il valore più credibile del parametro sconosciuto θ è dato dalla moda (o media, o mediana) della distribuzione a posteriori $p(\theta \mid y)$ che si ottiene combinando la verosimiglianza $p(y \mid \theta)$ con la distribuzione a priori $p(\theta)$.

A.1 Derivazione della s.m.v. per una proporzione

La s.m.v. della proporzione di successi θ in una sequenza di prove Bernoulliane è uguale data dalla proporzione di successi campionari. Questo risultato può essere dimostrato come segue.

Dimostrazione. Per n prove Bernoulliane indipendenti, le quali producono y successi e $(n-y)$ insuccessi, la funzione nucleo (ovvero, la funzione di verosimiglianza da cui sono state escluse tutte le costanti moltiplicative che non hanno alcun effetto su $\hat{\theta}$) è

$$\mathcal{L}(p \mid y) = \theta^y (1-\theta)^{n-y}.$$

La funzione nucleo di log-verosimiglianza è

$$\begin{aligned}
\ell(\theta \mid y) &= \log \mathcal{L}(\theta \mid y) \\
&= \log \left(\theta^y (1-\theta)^{n-y}\right) \\
&= \log \theta^y + \log \left((1-\theta)^{n-y}\right) \\
&= y \log \theta + (n-y) \log(1-\theta).
\end{aligned}$$

Per calcolare il massimo della funzione di log-verosimiglianza è necessario differenziare $\ell(\theta \mid y)$ rispetto a θ, porre la derivata a zero e risolvere. La derivata di $\ell(\theta \mid y)$ è:

$$\ell'(\theta \mid y) = \frac{y}{\theta} - \frac{n-y}{1-\theta}.$$

Ponendo l'equazione uguale a zero e risolvendo otteniamo la s.m.v.:

$$\hat{\theta} = \frac{y}{n}, \tag{A.2}$$

ovvero la frequenza relativa dei successi nel campione. □

Calcolo numerico

In maniera più semplice, il risultato descritto nel Paragrafo A.1 può essere ottenuto mediante una simulazione in R. Iniziamo a definire un insieme di valori possibili per il parametro incognito θ:

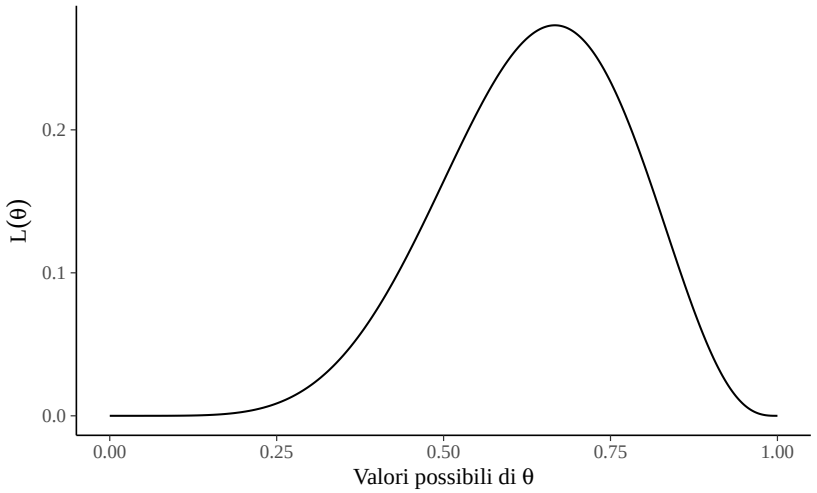
```
theta <- seq(0, 1, length.out=1e3)
```

Sappiamo che la funzione di verosimiglianza è la funzione di massa di probabilità espressa in funzione del parametro sconosciuto θ assumendo come noti i dati. Questo si può esprimere in R nel modo seguente:

```
like <- dbinom(x = 6, size = 9, prob = theta)
```

Si noti che, nell'istruzione precedente, abbiamo passato alla funzione `dbinom()` i dati, ovvero `x = 6` successi in `size = 9` prove. Inoltre, abbiamo passato alla funzione il vettore `prob = theta` che contiene 1000 valori possibili per il parametro $\theta \in [0, 1]$. Per ciascuno dei valori θ, la funzione `dbinom()` ritorna un valore che corrisopnde all'ordinata della funzione di verosimiglianza, tenendo sempre costanti i dati (ovvero, 6 successi in 9 prove). Un grafico della funzione di verosimiglianza è dato da:

```
tibble(theta, like) %>%
  ggplot(aes(x = theta, y = like)) +
  geom_line() +
  labs(
    y = expression(L(theta)),
    x = expression('Valori possibili di' ~ theta)
  )
```



Nella simulazione, il valore θ che massimizza la funzione di verosimiglianza può essere trovato nel modo seguente:

```
theta[which.max(like)]
#> [1] 0.667
```

Il valore così trovato è uguale al valore definito dalla (A.2).

La verosimiglianza del modello Normale

Ora che abbiamo capito come costruire la funzione verosimiglianza di una binomiale è relativamente semplice fare un passo ulteriore e considerare la verosimiglianza del caso di una funzione di densità, ovvero nel caso di una variabile casuale continua. Consideriamo qui il caso della Normale.

La densità di una distribuzione Normale di parametri μ e σ è

$$f(y \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\}.$$

Poniamoci ora il problema di trovare la s.m.v. dei parametri sconosciuti μ e σ nel caso in cui le n osservazioni $y = (y_1, \dots, y_n)$ sono realizzazioni indipendenti ed identicamente distribuite (di seguito, i.i.d.) della medesima variabile casuale $Y \sim \mathcal{N}(\mu, \sigma)$. Per semplicità, scriveremo $\theta = \{\mu, \sigma\}$.

Il campione osservato è un insieme di eventi, ciascuno dei quali corrisponde alla realizzazione di una variabile casuale — possiamo pensare ad uno di tali eventi come all'estrazione casuale di un valore dalla "popolazione" $\mathcal{N}(\mu, \sigma)$. Se le variabili casuali sono i.i.d., la loro densità congiunta è data da:

$$\begin{aligned} f(y \mid \theta) &= f(y_1 \mid \theta) \cdot f(y_2 \mid \theta) \cdot \ \dots \ \cdot f(y_n \mid \theta) \\ &= \prod_{i=1}^{n} f(y_i \mid \theta), \end{aligned} \tag{A.3}$$

laddove la funzione $f(\cdot)$ è la (A.1). Tenendo costanti i dati y, la funzione di verosimiglianza è:

$$\mathcal{L}(\theta \mid y) = \prod_{i=1}^{n} f(y_i \mid \theta). \tag{A.4}$$

L'obiettivo è quello di massimizzare la funzione di verosimiglianza per trovare i valori θ ottimali. Usando la notazione matematica questo si esprime dicendo che cerchiamo l'argmax della (A.4) rispetto a θ, ovvero

$$\hat{\theta} = \text{argmax}_\theta \prod_{i=1}^{n} f(y_i \mid \theta).$$

Questo problema si risolve calcolando le derivate della funzione rispetto a θ, ponendo le derivate uguali a zero e risolvendo. Saltando tutti i passaggi algebrici di questo procedimento, per μ troviamo

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{A.5}$$

e per σ abbiamo

$$\hat{\sigma} = \sqrt{\sum_{i=1}^{n} \frac{1}{n}(y_i - \mu)^2}. \tag{A.6}$$

In altri termini, la s.m.v. del parametro μ è la media del campione e la s.m.v. del parametro σ è la deviazione standard del campione.

Simulazione

Consideriamo ora un esempio che utilizza dei dati reali. I dati corrispondono ai valori BDI-II dei trenta soggetti del campione clinico di Zetsche et al. (2019):

```
d <- tibble(
  y = c(26, 35, 30, 25, 44, 30, 33, 43, 22, 43, 24,
        19, 39, 31, 25, 28, 35, 30, 26, 31, 41, 36,
        26, 35, 33, 28, 27, 34, 27, 22)
  )
```

Ci poniamo l'obiettivo di creare la funzione di verosimiglianza per questi dati, supponendo, in base ai risultati di ricerche precedenti, di sapere che i punteggi BDI-II si distribuiscono secondo una legge Normale.

Per semplificare il problema, assumeremo di conoscere σ (lo porremo uguale alla deviazione standard del campione) in modo da avere un solo parametro sconosciuto, cioè μ. Il problema è dunque quello di trovare la funzione di verosimiglianza per il parametro μ, date le 30 osservazioni del campione e dato $\sigma = s = 6.61$.

Per una singola osservazione, la funzione di verosimiglianza è la densità Normale espressa in funzione dei parametri. Per un campione di osservazioni i.i.d., ovvero $y = (y_1, y_2, \dots, y_n)$, la verosimiglianza è la funzione di densità congiunta $f(y \mid \mu, \sigma)$ espressa in funzione dei parametri, ovvero $\mathcal{L}(\mu, \sigma \mid y)$. Dato che le osservazioni sono i.i.d., la densità congiunta è data dal prodotto delle densità delle singole osservazioni.

Per semplicità, assumiamo σ noto e uguale alla deviazione standard del campione:

```
true_sigma <- sd(d$y)
true_sigma
#> [1] 6.61
```

Avendo posto $\sigma = 6.61$, per una singola osservazione y_i abbiamo

$$f(y_i \mid \mu, \sigma) = \frac{1}{6.61\sqrt{2\pi}} \exp\left\{-\frac{(y_i - \mu)^2}{2 \cdot 6.61^2}\right\},$$

dove il pedice i specifica l'osservazione y_i tra le molteplici osservazioni y, e μ è il parametro sconosciuto che deve essere determinato (nell'esempio, $\sigma = s$). La densità congiunta è dunque

$$f(y \mid \mu, \sigma) = \prod_{i=1}^{n} f(y_i \mid \mu, \sigma)$$

e, alla luce dei dati osservati, la verosimiglianza diventa

$$\begin{aligned}\mathcal{L}(\mu, \sigma \mid y) = \prod_{i=1}^{n} f(y_i \mid \mu, \sigma) = \\ \frac{1}{6.61\sqrt{2\pi}} \exp\left\{-\frac{(26 - \mu)^2}{2 \cdot 6.61^2}\right\} \times \\ \frac{1}{6.61\sqrt{2\pi}} \exp\left\{-\frac{(35 - \mu)^2}{2 \cdot 6.61^2}\right\} \times \\ \vdots \\ \frac{1}{6.61\sqrt{2\pi}} \exp\left\{-\frac{(22 - \mu)^2}{2 \cdot 6.61^2}\right\}.\end{aligned}$$

Poniamoci ora il problema di rappresentare graficamente la funzione di verosimiglianza per il parametro μ. Avendo un solo parametro sconosciuto, possiamo rappresentare la verosimiglianza con una curva. In R, definiamo la funzione di log-verosimiglianza nel modo seguente:

```
log_likelihood <- function(y, mu, sigma = true_sigma) {
  sum(dnorm(y, mu, sigma, log = TRUE))
}
```

Nella funzione `log_likelihood()`, `y` è un vettore che, nel caso presente contiene $n = 30$ valori. Per ciascuno di questi valori, la funzione `dnorm()` trova la densità Normale utilizzando il valore μ che passato a `log_likelihood()` e il valore σ uguale a 6.61 — nell'esempio, questo parametro viene assunto come noto. L'argomento `log = TRUE` specifica che deve essere preso il logaritmo. La funzione `dnorm()` è un argomento della funzione `sum()`. Ciò significa che i 30 valori così trovati, espressi su scala logaritmica, verranno sommati — sommare logaritmi è equivalente a fare il prodotto dei valori sulla scala originaria.

Se applichiamo questa funzione ad un solo valore μ otteniamo l'ordinata della funzione di log-verosimiglianza in corrispondenza del valore μ (si veda la figura (A.1)). Si noti che, per trovare un tale valore, abbiamo utilizzato le seguenti informazioni:

- i 30 dati del campione,
- il valore $\sigma = s$ fissato a 6.61,
- il singolo valore μ passato alla funzione `log_likelihood()`.

Avendo trovato un singolo punto della funzione di log-verosimiglianza, dobbiamo ripetere i calcoli precedenti per tutti i possibili valori che μ può assumere. Nel seguente ciclo `for()` viene calcolata la log-verosimiglianza di 100,000 valori possibili del parametro μ:

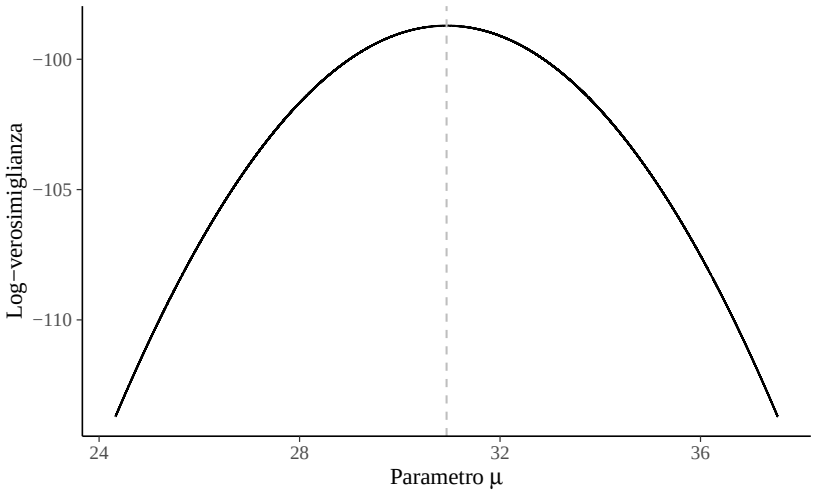
```
nrep <- 1e5
mu <- seq(
  mean(d$y) - sd(d$y),
  mean(d$y) + sd(d$y),
  length.out = nrep
)

ll <- rep(NA, nrep)
for (i in 1:nrep) {
  ll[i] <- log_likelihood(d$y, mu[i], true_sigma)
}
```

Il vettore `mu` contiene 100,000 possibili valori del parametro μ; tali valori sono stati scelti nell'intervallo $\bar{y} \pm s$. Per ciascuno di questi valori la funzione `log_likelihood()` calcola il valore di log-verosimiglianza. I 100,000 risultati vengono salvati nel vettore `ll`.

I vettori `mu` e `ll` possono dunque essere usati per disegnare il grafico della funzione di log-verosimiglianza per il parametro μ:

```
tibble(mu, ll) %>%
ggplot(aes(x = mu, y = ll)) +
  geom_line() +
  vline_at(mean(d$y), color = "gray", linetype = "dashed") +
  labs(
    y = "Log-verosimiglianza",
    x = expression("Parametro"~mu)
  )
```



Dalla figura notiamo che, per i dati osservati, il massimo della funzione di log-verosimiglianza calcolata per via numerica, ovvero 30.93, è identico alla media dei dati campionari e corrisponde al risultato teorico della (A.1).

Considerazioni conclusive

La verosimiglianza viene utilizzata sia nell'inferenza bayesiana che in quella frequentista. In entrambi i paradigmi di inferenza, il suo ruolo è quantificare la forza con la quale i dati osservati supportano i possibili valori dei parametri sconosciuti.

Nella funzione di verosimiglianza i dati (osservati) vengono trattati come fissi, mentre i valori del parametro (o dei parametri) θ vengono variati: la verosimiglianza è una funzione di θ per il dato fisso y. Pertanto, la funzione di verosimiglianza riassume i seguenti elementi: un modello statistico che genera stocasticamente i dati (in questo capitolo abbiamo esaminato due modelli statistici: quello binomiale e quello Normale), un intervallo di valori possibili per θ e i dati osservati y.

Nella statistica frequentista l'inferenza si basa solo sui dati a disposizione e qualunque informazione fornita dalle conoscenze precedenti non viene presa in considerazione. Nello specifico, nella statistica frequentista l'inferenza viene condotta massimizzando la funzione di (log) verosimiglianza, condizionatamente ai valori assunti dalle variabili casuali campionarie. Nella statistica bayesiana, invece, l'inferenza statistica viene condotta combinando la funzione di verosimiglianza con le distribuzioni a priori dei parametri incogniti θ.

La differenza fondamentale tra inferenza bayesiana e frequentista è dunque che i frequentisti non ritengono utile descrivere in termini probabilistici i parametri: i parametri dei modelli statistici vengono concepiti come fissi ma sconosciuti. Nell'inferenza bayesiana, invece, i parametri sconosciuti sono intesi come delle variabili casuali e ciò consente di quantificare in termini probabilistici il nostro grado di intertezza relativamente al loro valore.

Appendice B

Verosimiglianza marginale

Riportiamo di seguito la derivazione analitica per la costante di normalizzazione discussa nella Sezione 1.6, ovvero dell'integrale (1.9). Sia la distribuzione a priori $\theta \sim \mathrm{B}(a, b)$ e sia $y = \{y_1, \dots, y_n\} \sim \mathrm{Bin}(\theta, n)$. Scrivendo la *funzione beta* come

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

la verosimiglianza marginale diventa

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

in quanto

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

In conclusione, nel caso di una verosimiglianza binomiale $y =\sim \mathrm{Bin}(\theta, n)$ e di una distribuzione a priori $\theta \sim \mathrm{B}(a, b)$, la verosimiglianza marginale diventa uguale alla (B.1).

PRATICA GUIDATA

Si verifichi la (B.1) mediante di dati di Zetsche et al. (2019).

Per replicare mediante la (B.1) il risultato trovato per via numerica nella Sezione 1.6 assumiamo una distribuzione a priori uniforme, ovvero $\mathrm{B}(1, 1)$. I valori del problema dunque diventano i seguenti:

```
a <- 1
b <- 1
y <- 23
n <- 30
```

Definiamo

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

Il risultato cercato è

```
choose(30, 23) * B(y + a, n - y + b) / B(a, b)
```

```
## [1] 0.03225806
```

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).

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. 7).

Zetsche, U., Bürkner, P.-C. & Renneberg, B. (2019). Future expectations in clinical depression: Biased or realistic? *Journal of Abnormal Psychology*, *128*(7), 678–688 (cit. alle pp. 5, 8, 10, 11, 16, 19).

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.