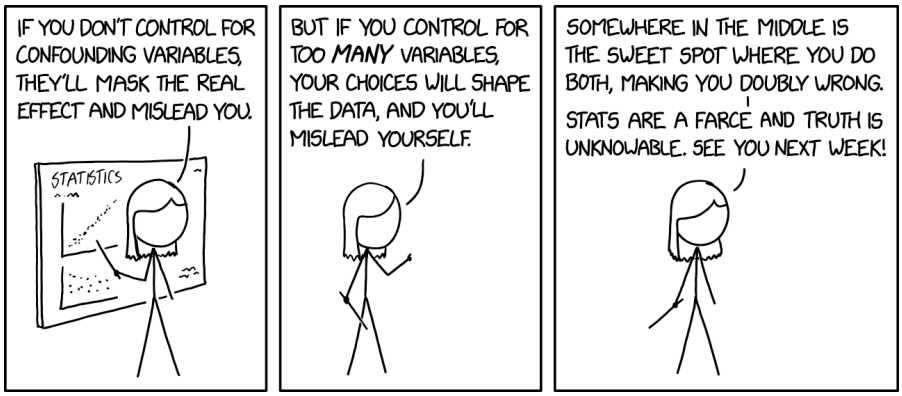
Corrado Caudek

Data Science per psicologi

Psicometria – AA 2021/2022



Indice

Elenco delle figure

Elenco delle tabelle

Prefazione

Data Science per psicologi contiene il materiale delle lezioni dell'insegnamento di *Psicometria B000286* (A.A. 2021/2022) rivolto agli studenti del primo anno del Corso di Laurea in Scienze e Tecniche Psicologiche dell'Università degli Studi di Firenze. *Psicometria* si propone di fornire agli studenti un'introduzione all'analisi dei dati in psicologia. Le conoscenze/competenze che verranno sviluppate in questo insegnamento sono quelle della Data science, ovvero un insieme di conoscenze/competenze che si pongono all'intersezione tra statistica (ovvero, richiedono la capacità di comprendere teoremi statistici) e informatica (ovvero, richiedono la capacità di sapere utilizzare un software).

La psicologia e la Data science

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

C'è una ragione molto semplice che dovrebbe farci capire perché la Data science è così importante per la psicologia. Infatti, a ben pensarci, la psicologia è una disciplina intrinsecamente statistica, se per statistica intendiamo quella disciplina che studia la variazione delle caratteristiche degli individui nella popolazione. La psicologia studia *gli individui* ed è proprio la variabilità inter- e intra-individuale ciò che vogliamo descrivere e, in certi casi, predire. In questo senso, la psicologia è molto diversa dall'ingegneria, per esempio. Le proprietà di un determinato ponte sotto certe condizioni, ad esempio, sono molto simili a quelle di un altro pon-

te, sotto le medesime condizioni. Quindi, per un ingegnere la statistica è poco importante: le proprietà dei materiali sono unicamente dipendenti dalla loro composizione e restano costanti. Ma lo stesso non può dirsi degli individui: ogni individuo è unico e cambia nel tempo. E le variazioni tra gli individui, e di un individuo nel tempo, sono l'oggetto di studio proprio della psicologia: è dunque chiaro che i problemi che la psicologia si pone sono molto diversi da quelli affrontati, per esempio, dagli ingegneri. Questa è la ragione per cui abbiamo tanto bisogno della Data science in psicologia: perché la Data science ci consente di descrivere la variazione e il cambiamento. E queste sono appunto le caratteristiche di base dei fenomeni psicologici.

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il "lavoro sporco" (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò che ci interessa, ovvero dei problemi psicologici slegati dai dettagli "tecnici" della Data science? La risposta a questa domanda è che non è possibile progettare uno studio psicologico sensato senza avere almeno una comprensione rudimentale della Data science. Le tematiche della Data science non possono essere ignorate né dai ricercatori in psicologia né da coloro che svolgono la professione di psicologo al di fuori dell'Università. Infatti, anche i professionisti al di fuori dall'università non possono fare a meno di leggere la letteratura psicologica più recente: il continuo aggiornamento delle conoscenze è infatti richiesto dalla deontologia della professione. Ma per potere fare questo è necessario conoscere un bel po' di Data science! Basta aprire a caso una rivista specialistica di psicologia per rendersi conto di quanto ciò sia vero: gli articoli che riportano i risultati delle ricerche psicologiche sono zeppi di analisi statistiche e di modelli formali. E la comprensione della letteratura psicologica rappresenta un requisito minimo nel bagaglio professionale dello psicologo.

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

adottato un punto di vista bayesiano, che corrisponde all'approccio più recente e sempre più diffuso in psicologia.

Come studiare

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

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

Sviluppare un metodo di studio efficace

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l'atteggiamento mentale che descriverò qui sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo. Fornisco qui sotto una breve descrizione del "metodo di studio" che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento.

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.
- Gli errori che facciamo sono i nostri migliori maestri. Istintivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un'informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.
- C'è ovviamente un aspetto "psicologico" nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: "mi arrendo", "non ho idea di cosa fare!". Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose "migliori" che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c'è qualcosa che non so fare e non ho idea di come affontare, mi dico: "oggi non ho proprio voglia di fare fatica", non ho voglia di mettermi nello stato mentale per cui "in 10 minuti devo risolvere il problema perché dopo devo fare altre cose". Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto "marginale" del problema, che so come affrontare, oppure considero l'aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, opppure lo stesso problema in un altro contesto. Non mi pongo l'obiettivo "risolvi il problema in 10 minuti", ma invece quello di farmi un'idea "generale" del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se "parto per la tangente", ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a "lavorare" al problema. Allora perché non faccio sempre così? C'è ovviamente l'aspetto dei "10 minuti" che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possia-

mo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.

- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: "se questo è vero, allora deve succedere quest'altra cosa". Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c'era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.
- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare di colmarle. L'atteggiamento naturale, quando non capiamo i dettagli di qualcosa, è quello di pensare: "non importa, ho capito in maniera approssimativa questo punto, non devo preoccuparmi del resto". Ma in realtà non è vero: se la nostra comprensione è superficiale, quando il problema verrà presentato in una nuova forma, non riusciremo a risolverlo. Per cui i dubbi che ci vengono quando studiamo qualcosa sono il nostro alleato più prezioso: ci dicono esattamente quali sono gli aspetti che dobbiamo approfondire per potere migliorare la nostra preparazione.
- È utile sviluppare una visione d'insieme degli argomenti trattati, capire l'obiettivo generale che si vuole raggiungere e avere chiaro il contributo che i vari pezzi di informazione forniscono al raggiungimento di tale obiettivo. Questa organizzazione mentale del materiale di studio facilita la comprensione. È estremamente utile creare degli schemi di ciò che si sta studiando. Non aspettate che sia io a fornirvi un riepilogo di ciò che dovete imparare: sviluppate da soli tali schemi e tali riassunti.
- Tutti noi dobbiamo imparare l'arte di trovare le informazioni, non solo nel caso di questo insegnamento. Quando vi trovate di fronte a qualcosa che non capite, o ottenete un oscuro messaggio di errore da

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

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

Corrado Caudek
Marzo 2022

Parte I

Inferenza statistica bayesiana

1

Inferenza bayesiana

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

1.1 Modellizzazione bayesiana

L'analisi bayesiana corrisponde alla costruzione di un modello statistico che si può rappresentare con una quaterna

$$(\Omega, p(y \mid \theta), p(\theta), \theta \in \Theta),$$

dove Ω è l'insieme di tutti i possibili risultati ottenuti dall'esperimento casuale e $p(y \mid \theta)$ è una famiglia di leggi di probabilità, indicizzata dal parametro $\theta \in \Theta$, che descrive l'incertezza sull'esito dell'esperimento. Secondo l'approccio bayesiano, il parametro incognito θ è considerato una variabile casuale che segue la legge di probabilità $p(\theta)$. L'incertezza su θ è la sintesi delle opinioni e delle informazioni che si hanno sul parametro prima di avere osservato il risultato dell'esperimento e prende il nome di *distribuzione a priori*. La costruzione del modello statistico

passa attraverso la scelta di una densità $p(y \mid \theta)$ che rappresenta, in senso probabilistico, il fenomeno d'interesse, e attraverso la scelta di una distribuzione a priori $p(\theta)$.

In ambito bayesiano le informazioni che si hanno a priori sul parametro di interesse θ, contenute in $p(\theta)$, vengono aggiornate attraverso quelle provenienti dal campione osservato $y = (y_1, \dots, y_n)$ contenute nella funzione $p(y \mid \theta)$, che, osservata come funzione di θ per y, prende il nome di *funzione di verosimiglianza*. L'aggiornamento delle informazioni avviene attraverso la formula di Bayes

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_\Theta p(y \mid \theta)p(\theta)\, \mathrm{d}\theta} \quad \theta \in \Theta, \tag{1.1}$$

in cui $p(\theta \mid y)$ prende il nome di *distribuzione a posteriori*.

Il denominatore del Teorema di Bayes (1.1), che costituisce la costante di normalizzazione, è la densità marginale dei dati descritta nel Paragrafo XX. In ambito bayesiano la distribuzione a posteriori viene utilizzata per calcolare le principali quantità di interesse dell'inferenza, ad esempio le medie a posteriori di funzioni di θ.

Le quantità di interesse della statistica bayesiana sono espresse nei termini di integrali che risultano, nella maggior parte dei casi, impossibili da risolvere analiticamente e, per questo motivo, si ricorre a metodi di numerici, in particolare a quei metodi Monte Carlo basati sulle proprietà delle catene di Markov (MCMC).

Seguendo (Martin et al., 2022), possiamo descrivere la modellazione bayesiana distinguendo tre passaggi.

1. Dati alcuni dati e alcune ipotesi su come questi dati potrebbero essere stati generati, progettiamo un modello combinando e trasformando variabili casuali.
2. Usiamo il teorema di Bayes per condizionare i nostri modelli ai dati disponibili. Chiamiamo questo processo "inferenza" e come risultato otteniamo una distribuzione a posteriori.
3. Critichiamo il modello verificando se il modello abbia senso utilizzando criteri diversi, inclusi i dati e la nostra conoscenza del dominio. Poiché generalmente siamo incerti sui modelli stessi, a volte confrontiamo modelli diversi.

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

1.1.1 Notazione

Per fissare la notazione, nel seguito y rappresenterà i dati e θ rappresenterà i parametri incogniti di un modello statistico. Sia y che θ vengono concepiti come variabili casuali. Con x vengono invece denotate le quantità note, come ad esempio i predittori del modello lineare. Per rappresentare in un modo conciso i modelli probabilistici viene usata una notazione particolare. Ad esempio, invece di scrivere $p(\theta) = \text{Beta}(1, 1)$ scriviamo $\theta \sim \text{Beta}(1, 1)$. Il simbolo "$\sim$" viene spesso letto "è distribuito come". Possiamo anche pensare che significhi che θ costituisce un campione casuale estratto dalla distribuzione Beta(1, 1). Allo stesso modo, ad esempio, la verosimiglianza del modello binomiale può essere scritta come $y \sim \text{Bin}(n, \theta)$.

1.2 Distribuzioni a priori

Quando adottiamo un approccio bayesiano, i parametri della distribuzione di riferimento non venono considerati come delle costanti incognite ma bensì vengono trattati come variabili casuali e, di conseguenza, i parametri assumono una particolare distribuzione che nelle statistica bayesiana viene definita come "a priori". I parametri (o il parametro), che possiamo indicare con θ, possono assumere delle distribuzioni a priori differenti; a seconda delle informazioni disponibili bisogna cercare di assegnare una distribuzione di θ in modo tale che venga assegnata una probabilità maggiore a quei valori che si ritengono più plausibili per θ.

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.

1.2.1 Tipologie di distribuzioni a priori

Possiamo distinguere tra diverse distribuzioni a priori in base a quanto fortemente impegnano il ricercatore a ritenere come plausibile un particolare intervallo di valori dei parametri. Il caso più estremo è quello che rivela una totale assenza di conoscenze a priori, il che conduce alle *distribuzioni a priori non informative*, ovvero quelle che assegnano lo stesso livello di credibilità a tutti i valori dei parametri. Le distribuzioni a priori informative, d'altra parte, possono essere *debolmente informative* o *fortemente informative*, a seconda della forza della credenza che esprimono. Il caso più estremo di credenza a priori è quello che riassume il punto di vista del ricercatore nei termini di un *unico valore* del parametro, il che assegna tutta la probabilità (massa o densità) 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 $\frac{\alpha}{\beta} = \frac{5}{2}$.

1.2.2 Selezione della distribuzione a priori

La selezione delle distribuzioni a priori è stata spesso vista come una delle scelte più importanti che un ricercatore fa quando implementa un modello bayesiano in quanto può avere un impatto sostanziale sui risultati finali. La soggettività delle distribuzioni a priori è evidenziata dai critici come un potenziale svantaggio dei metodi bayesiani. A questa critica, van de Schoot et al. (2021) rispondono dicendo che, al di là della

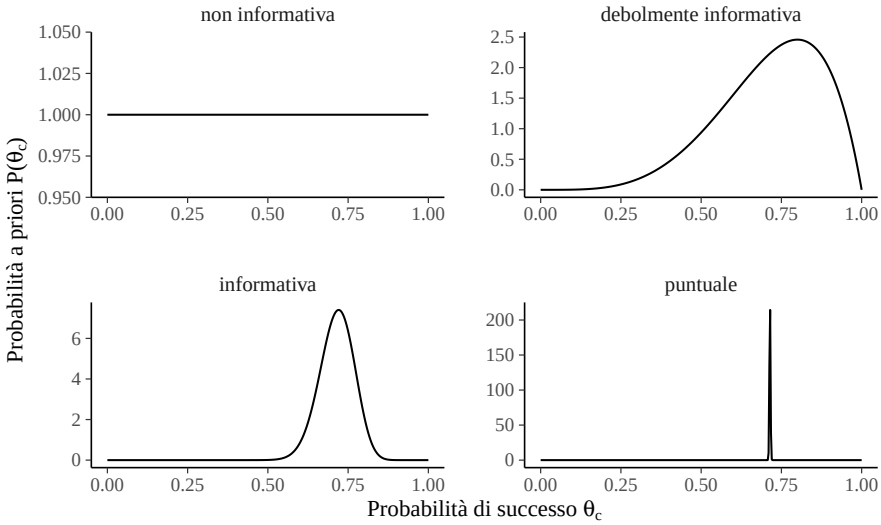


Figura 1.1: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

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 **??**.

1.2.3 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 delle trasformazioni di scala dei dati. È invece raccomandato usare una distribuzione a priori

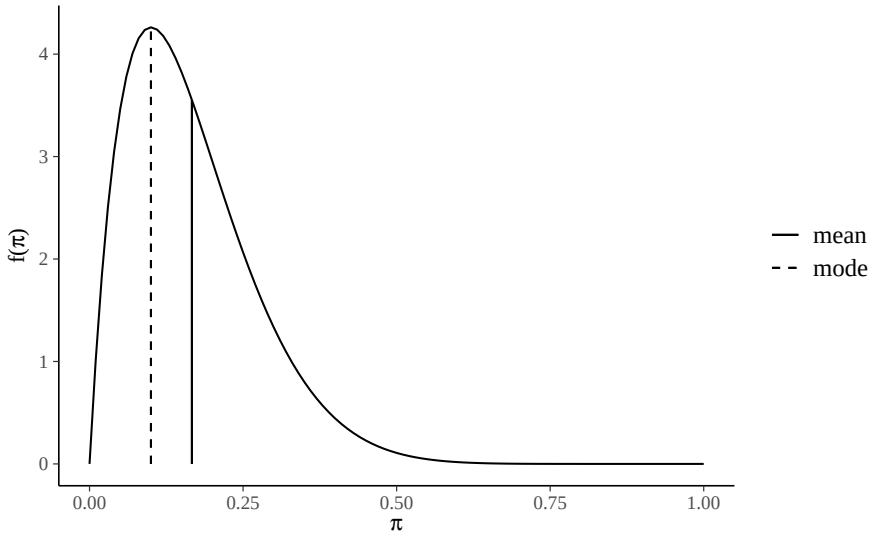
debolmente informativa, come ad esempio Beta$(2, 2)$.

Nella presente discussione, per fare un esempio concreto, useremo i dati riportati da Zetsche et al. (2019) (si veda l'appendice **??**). Tali dati possono essere considerati la manifestazione di una variabile casuale Bernoulliana e corrispondono a 23 "successi" in 30 prove.

In tale contesto, ha senso usare una distribuzione a priori debolmente informativa, per esempio, Beta$(2, 2)$. Nel presente Capitolo, quale distribuzione a priori, useremo una Beta$(2, 10)$:

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

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



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

1.3 La funzione di verosimiglianza

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

$$y \sim \text{Bin}(n, \theta), \tag{1.2}$$

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

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

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

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) 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. Per un approfondimento della stima di massima verosimiglianza si veda l'Appendice **??**.

1.3.1 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:

$$\ell(\theta) = \log \mathcal{L}(\theta). \tag{1.3}$$

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:

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

Per le proprietà del logaritmo, si ha

$$\ell(\theta) = \log \left(\prod_{i=1}^{n} f(y \mid \theta) \right) = \sum_{i=1}^{n} \log f(y \mid \theta). \tag{1.4}$$

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

Osservazione. Seguendo una pratica comune, in questa dispensa spesso useremo la notazione $p(\cdot)$ per rappresentare due quantità differenti, ovvero la funzione di verosimiglianza e la distribuzione a priori. Questo piccolo abuso di notazione riflette il seguente punto di vista: anche se la verosimiglianza non è una funzione di densità di probabilità, noi non vogliamo stressare questo aspetto, ma vogliamo piuttosto pensare alla verosimiglianza e alla distribuzione a priori come a due elementi che sono egualmente necessari per calcolare la distribuzione a posteriori. In altri termini, per così dire, questa notazione assegna lo stesso status epistemologico alle due diverse quantità che si trovano al numeratore della regola di Bayes.

Esercizio 1.1. Per i dati di Zetsche et al. (2019), ovvero 23 "successi" in 30 prove, si trovi e si interpreti la funzione di verosimiglianza.

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

$$\mathcal{L}(\theta \mid y) = \frac{(23 + 7)!}{23!7!}\theta^{23} + (1 - \theta)^7. \tag{1.5}$$

Per costruire la funzione di verosimiglianza dobbiamo applicare la (1.5) tante volte, cambiando ogni volta il valore θ ma *tenendo sempre costante il valore dei dati*. Per esempio, se poniamo $\theta = 0.1$

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

otteniamo

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

Se poniamo $\theta = 0.2$

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

otteniamo

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

e così via. La figura 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)
  )
```

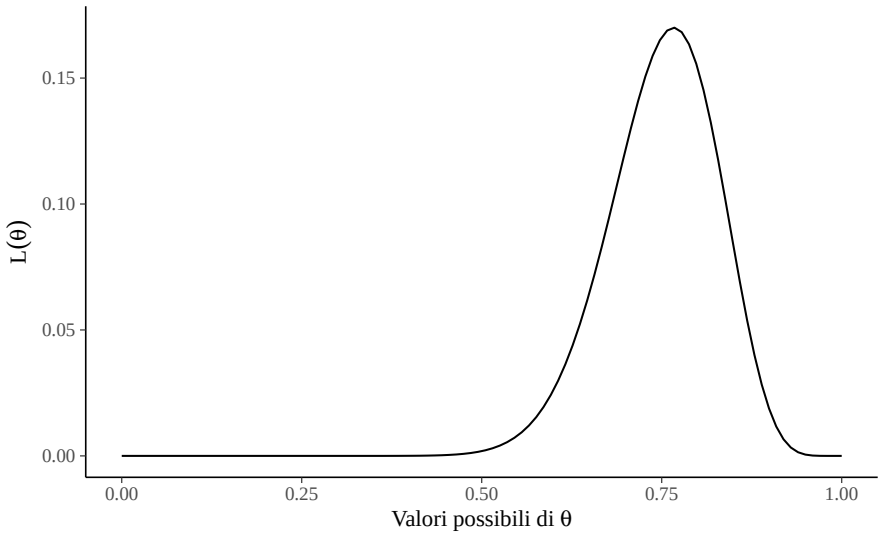


Figura 1.2: *Funzione di verosimiglianza nel caso di 23 successi in 30 prove.*

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 $\mathcal{L}(\theta_1) > \mathcal{L}(\theta_1)$.

1.4 La verosimiglianza marginale

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

Si noti che 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.

Esercizio 1.2. 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:

$$p(y = 23, n = 30) = \int_0^1 \binom{30}{23} \theta^{23} (1 - \theta)^7 \, \mathrm{d}\theta. \tag{1.6}$$

Una soluzione numerica si trova facilmente usando R:

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

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

1.5 Distribuzione a posteriori

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

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

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

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

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

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

1.6 Distribuzione predittiva a priori

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

$$p(y^*) = \int_\Theta p(y^* \mid \theta) p(\theta) \, \mathrm{d}\theta. \tag{1.7}$$

La (1.7) descrive la distribuzione prevista dei dati in base al modello (che include la distribuzione a priori e la verosimiglianza). Questi sono i dati y^* che ci aspettiamo, dato il modello, prima di avere osservato i dati del campione.

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

Osservazione. Si dice comunemente che l'adozione di una prospettiva probabilistica per la modellazione conduce all'idea che i modelli generano dati. Se i modelli generano dati, possiamo creare modelli adatti per i nostri dati solo pensando a come i dati potrebbero essere stati generati. Inoltre, questa idea non è solo un concetto astratto. Assume una concreta nella forma della distribuzione predittiva a priori. Se la distribuzione predittiva a priori non ha senso, come abbiamo detto sopra, diventa necessario riformulare il modello.

1.7 Distribuzione predittiva a posteriori

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

$$p(\tilde{y} \mid y) = \int_{\Theta} p(\tilde{y} \mid \theta) p(\theta \mid y) \, \mathrm{d}\theta. \tag{1.8}$$

Questa è la distribuzione dei dati attesi futuri $\tilde{y}$ alla luce della distribuzione a posteriori $p(\theta \mid y)$, che a sua volta è una conseguenza del modello (distribuzione a priori e verosimiglianza) e dei dati osservati. In altre parole, questi sono i dati che il modello si aspetta dopo aver osservato i dati y. Dalla (1.8) possiamo vedere che le previsioni sui dati attesi futuri sono calcolate integrando (o marginalizzando) sulla distribuzione a posteriori dei parametri. Di conseguenza, le previsioni calcolate in questo modo incorporano l'incertezza relativa alla stima dei parametri del modello.

Commenti e considerazioni finali

Questo Capitolo ha brevemente passato in rassegna 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.

2

Distribuzioni a priori coniugate

Obiettivo di questo Capitolo è fornire un esempio di derivazione della distribuzione a posteriori scegliendo quale distribuzione a priori una distribuzione coniugata. Esamineremo qui il modello Beta-Binomiale.

2.1 Pensare a una proporzione "in termini soggettivi"

Nei problemi tradizionali di teoria delle probabilità ci sono molti esempi che riguardano l'estrazizone di palline colorate da un'urna. In questi esempi, ci viene fornito il numero di palline di vari colori nell'urna e ci viene chiesto di calcolare le probabilità di vari eventi. Ad esempio, in una scatola ci sono 40 palline bianche e 20 rosse. Se estrai due palline a caso, qual è la probabilità che entrambe siano bianche?

L'approccio bayesiano considera uno scenario diverso: quello in cui non conosciamo le proporzioni delle palline colorate nell'urna. Cioè, nell'esempio precedente, sappiamo solo che ci sono due tipi di palline colorate nell'urna, ma non sappiamo che 40 palline su 60 sono bianche (proporzione di bianco = 2/3) e 20 delle 60 palline sono rosse (proporzione di rosso = 1/3). Ci poniamo la seguente domanda: è possibile inferire le proporzioni cercate estraendo un campione di palline dall'urna e osservando i colori delle palline nel campione? Espresso in questo modo, questo diventa un problema di inferenza statistica, perché stiamo cercando di inferire la proporzione π della popolazione sulla base di un campione casuale della popolazione. Per continuare con l'esempio precedente, quello che ci chiediamo è: come è possibile inferire π, la proporzione di palline rosse nella popolazione, in base al numero (per esempio, 10) di palline rosse e bianche che osserviamo nel campione?

Le proporzioni assomigliano alle probabilità. Ricordiamo che sono state proposte tre diverse interpretazioni del concetto di una probabilità.

- Il punto di vista classico: è necessario enumerare tutti gli eventi elementari dello spazio campionario in cui ogni risultato è ugualmente probabile.
- Il punto di vista frequentista: è necessario ripetere l'esperimento esperimento casuale (cioè l'estrazione del campione) molte volte in condizioni identiche.
- La visione soggettiva: è necessario esprimere la propria opinione sulla probabilità di un evento unico e irripetibile.

La visione classica non sembra potere funzionare qui, perché sappiamo solo che ci sono due tipi di palline colorate e il numero totale di palline è 60. Anche se estraiamo un campione di 10 palline, possiamo solo osservare la proporzione di palline rosse palline nel campione. Non c'è modo per stabilire quali sono le proprietà dello spazio campionario in cui ogni risultato è ugualmente probabile.

La visione frequentista potrebbe funzionare nel caso presente. Possiamo considerare il processo del campionamento (cioè l'estrazione di un campione casuale di 10 palline dall'urna) come un esperimento casuale che produce una proporzione campionaria p. Potremmo quindi pensare di ripetere l'esperimento molte volte nelle stesse condizioni, ottenere molte proporzioni campionarie p e riassumere poi in qualche modo questa distribuzione di statistiche campionarie. Ripetendo l'esperimento casuale tante volte è possibile ottenere una stima abbastanza accurata della proporzione π di palline rosse nell'urna. Questo processo è fattibile, ma è però noioso, dispendioso in termini di tempo e soggetto a errori.

La visione soggettivista concepisce invece la probabilità sconosciuta π come un'opinione soggettiva di cui possiamo essere più o meno sicuri. Abbiamo visto in precedenza come questa opinione soggettiva dipende da due fonti di evidenza: le nostre credenze iniziali e le nuove informazioni fornite dai dati che abbiamo osservato. Vedremo in questo capitolo come sia possibile combinare le credenze iniziali rispetto al possibile valore π con le evidenza fornite dai dati per giungere ad una credenza a posteriori su π. Se le nostre credenze a priori sono espresse nei termini di una distribuzione Beta, allora è possibile derivare le proprietà della distribuzione a priori per via analitica. Questo capitolo ha lo scopo di mostrare come questo possa essere fatto.

2.2 Il denominatore bayesiano

In termini generali possiamo dire che, in un problema bayesiano, i dati y provengono da una distribuzione $p(y \mid \theta)$ e al parametro θ viene assegnata una distribuzione a priori $p(\theta)$. La scelta della distribuzione a priori ha importanti conseguenze di tipo computazionale. Infatti, a meno di non utilizzare particolari forme analitiche, risulta impossibile ottenere espressioni esplicite per la distribuzione a posteriori. Ciò dipende dall'espressione a denominatore della formula di Bayes

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

il cui calcolo, in generale, non è eseguibile in modo analitico in forma chiusa. Una soluzione analitica dell'integrale al denominatore della regola di Bayes è possibile solo se vengono usate distribuzioni provenienti da famiglie coniugate.

Definizione 2.1. Una distribuzione di probabilità a priori $p(\theta)$ si dice *coniugata* al modello usato se la distribuzione a priori e la distribuzione a posteriori hanno la stessa forma funzionale. Dunque, le due distribuzioni differiscono solo per il valore dei parametri.

Ad esempio, se la verosimiglianza è Poisson e la distribuzione a priori è Gamma, allora anche la distribuzione a posteriori sarà una distribuzione Gamma. Da un punto di vista puramente matematico, le distribuzioni a priori coniugate sono la scelta più conveniente in quanto ci consentono di calcolare analiticamente la distribuzione a posteriori con "carta e penna", senza la necessità di ricorrere a calcoli complessi. Da una prospettiva computazionale moderna, però, le distribuzioni a priori coniugate generalmente non sono migliori delle alternative, dato che i moderni metodi computazionali ci consentono di eseguire l'inferenza praticamente con qualsiasi scelta delle distribuzioni a priori, e non solo con quelle che sono matematicamente convenienti. Tuttavia, le famiglie coniugate offrono un utile ausilio didattico nello studio dell'inferenza bayesiana (e anche in alcune situazioni in cui è necessario utilizzare espressioni analitiche per la distribuzione a posteriori). Questo è il motivo per cui le esamineremo qui. Nello specifico, esamineremo quello che viene chiamato il caso Beta-Binomiale.

2.3 Il modello Beta-Binomiale

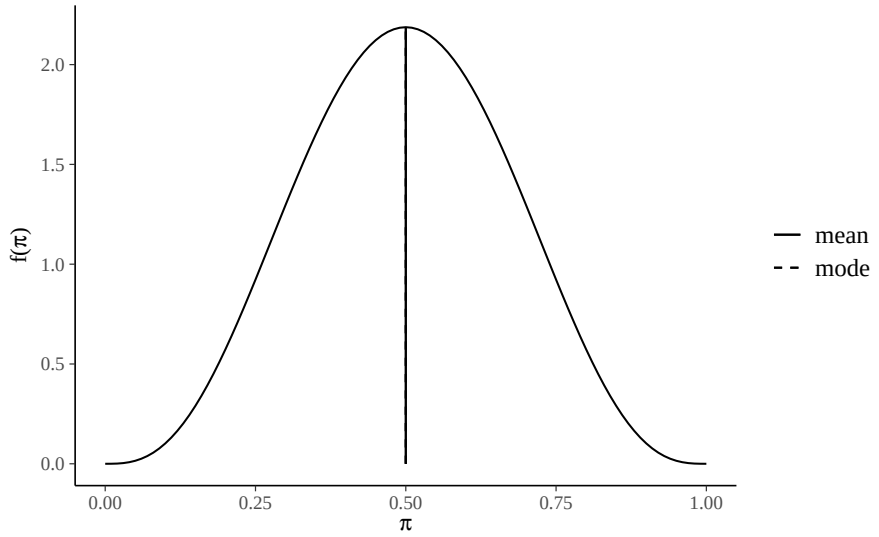
Per fare un esempio concreto, consideriamo nuovamente i dati di Zetsche et al. (2019): nel campione di 30 partecipanti clinici le aspettative future di 23 partecipanti risultano distorte negativamente e quelle di 7 partecipanti risultano distorte positivamente. Nel seguito, indicheremo con θ la probabilità che le aspettative di un paziente clinico siano distorte negativamente. Ci poniamo il problema di ottenere una stima a posteriori di θ avendo osservato 23 "successi" in 30 prove.

I dati osservati ($y = 23$) possono essere considerati la manifestazione di una variabile casuale Bernoulliana. In tali circostanze, esiste una famiglia di distribuzioni che, qualora venga scelta per la distribuzione a priori, fa sì che la distribuzione a posteriori abbia la stessa forma funzionale della distribuzione a priori. Questo consente una soluzione analitica dell'integrale che compare a denominatore nella formula di Bayes. Nel caso presente, la famiglia di distribuzioni che ha questa proprietà è la distribuzione Beta.

2.3.1 Parametri della distribuzione Beta

È possibile esprimere diverse credenze iniziali rispetto a θ mediante la distribuzione Beta. Ad esempio, la scelta di una Beta($\alpha = 4, \beta = 4$) quale distribuzione a priori per il parametro θ corrisponde alla credenza a priori che associa all'evento "presenza di una aspettativa futura distorta negativamente" una grande incertezza: il valore 0.5 è il valore di θ più plausibile, ma anche gli altri valori del parametro (tranne gli estremi) sono ritenuti piuttosto plausibili. Questa distribuzione a priori esprime la credenza che sia egualmente probabile per un'aspettativa futura essere distorta negativamente o positivamente.

```
library("bayesrules")
plot_beta(alpha = 4, beta = 4, mean = TRUE, mode = TRUE)
```

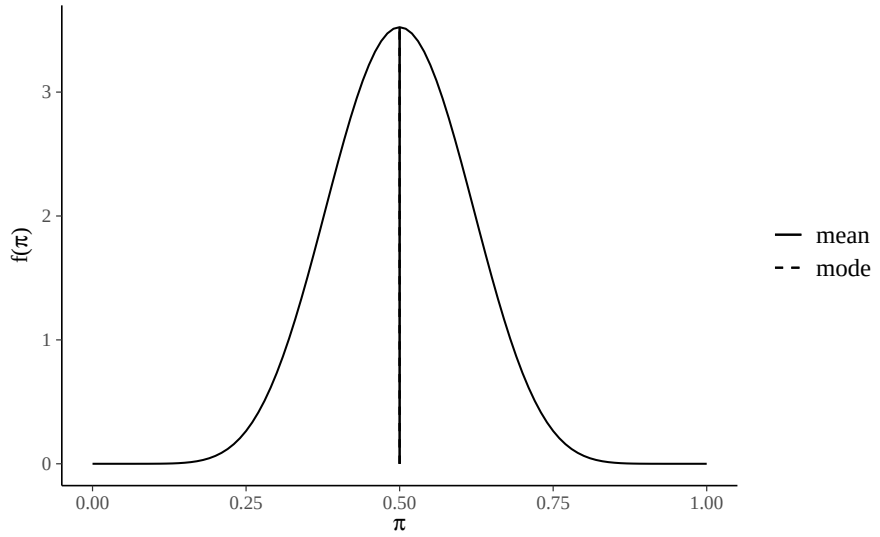


Possiamo quantificare la nostra incertezza calcolando, con un grado di fiducia del 95%, la regione nella quale, in base a tale credenza a priori, si trova il valore del parametro. Per ottenere tale intervallo di credibilità a priori, usiamo la funzione `qbeta()` di R. In `qbeta()` i parametri α e β sono chiamati `shape1` e `shape2`:

```
qbeta(c(0.025, 0.975), shape1 = 4, shape2 = 4)
#> [1] 0.1841 0.8159
```

Se poniamo $\alpha = 10$ e $\beta = 10$, questo corrisponde ad una credenza a priori che sia egualmente probabile per un'aspettativa futura essere distorta negativamente o positivamente,

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



ma ora la nostra certezza a priori sul valore del parametro è maggiore, come indicato dall'intervallo al 95%:

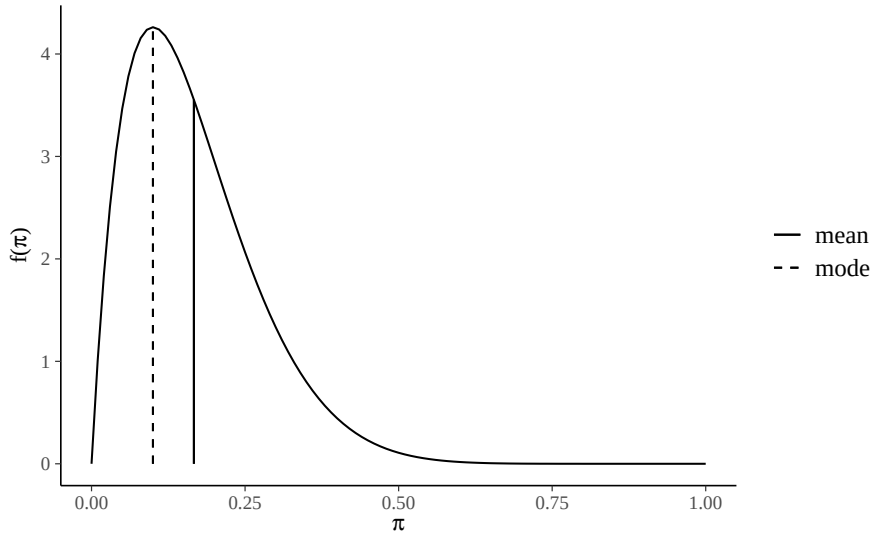
```
qbeta(c(0.025, 0.975), shape1 = 10, shape2 = 10)
#> [1] 0.2886 0.7114
```

Quale distribuzione a priori dobbiamo scegliere? 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 usare $\alpha = 1$ e $\beta = 1$, che produce una distribuzione a priori uniforme. Ma l'uso di distribuzioni a priori uniformi è sconsigliato per vari motivi, inclusa l'instabilità numerica della stima dei parametri. È meglio invece usare una distribuzione a priori poco informativa, come $\text{Beta}(2, 2)$.

Nella discussione successiva, solo per fare un esempio, useremo quale distribuzione a priori una $\text{Beta}(2, 10)$, ovvero:

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

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



La Beta$(2, 10)$ esprime la credenza che $\theta < 0.5$, con il valore più plausibile pari a cicrca 0.1.

2.3.2 La specificazione della distribuzione a posteriori

Una volta scelta una distribuzione a priori di tipo Beta, i cui parametri rispecchiano le nostre credenze iniziali su θ, la distribuzione a posteriori viene specificata dalla formula di Bayes:

$$\text{distribuzione a posteriori} = \frac{\text{verosimiglianza} \cdot \text{distribuzione a priori}}{\text{verosimiglianza marginale}}.$$

Nel caso presente abbiamo

$$p(\theta \mid n = 30, y = 23) = \frac{\left[\binom{30}{23}\theta^{23}(1-\theta)^{30-23}\right]\left[\frac{\Gamma(12)}{\Gamma(2)\Gamma(10)}\theta^{2-1}(1-\theta)^{10-1}\right]}{p(y = 23)},$$

laddove $p(y = 23)$, ovvero la verosimiglianza marginale, è una costante di normalizzazione che fa sì che l'area sottesa alla densità a posteriori sia unitaria.

Riscriviamo ora l'equazione precedente in termini generali

$$p(\theta \mid n, y) = \frac{\left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right]}{p(y)}$$

e raccogliendo tutte le costanti otteniamo:

$$p(\theta \mid n, y) = \left[\frac{\binom{n}{y}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}}{p(y)}\right]\theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1}.$$

Se ignoriamo il termine costante all'interno della parentesi quadra

$$\begin{aligned} p(\theta \mid n, y) &\propto \theta^y(1-\theta)^{n-y}\theta^{a-1}(1-\theta)^{b-1}, \\ &\propto \theta^{a+y-1}(1-\theta)^{b+n-y-1}, \end{aligned}$$

il termine di destra dell'equazione precedente identifica il *kernel* della distribuzione a posteriori e corrisponde ad una Beta *non normalizzata* di parametri $a+y$ e $b+n-y$.

Per ottenere una distribuzione di densità, dobbiamo aggiungere una costante di normalizzazione al kernel della distribuzione a posteriori. In base alla definizione della distribuzione Beta, ed essendo $a' = a+y$ e $b' = b+n-y$, tale costante di normalizzazione sarà uguale a

$$\frac{\Gamma(a'+b')}{\Gamma(a')\Gamma(b')} = \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}.$$

In altri termini, la distribuzione a posteriori diventa una Beta$(a+y, b+n-y)$:

$$\text{Beta}(a+y, b+n-y) = \frac{\Gamma(a+b+n)}{\Gamma(a+y)\Gamma(b+n-y)}\theta^{a+y-1}(1-\theta)^{b+n-y-1}.$$

Possiamo concludere dicendo che siamo partiti da una verosimiglianza Bin$(n=30, y=23 \mid \theta)$. Moltiplicando la verosimiglianza per la distribuzione a priori $\theta \sim \text{Beta}(2, 10)$, abbiamo ottenuto la distribuzione a posteriori $p(\theta \mid n, y) \sim \text{Beta}(25, 17)$. Questo è un esempio di analisi coniugata: la distribuzione a posteriori del parametro ha la stessa forma funzionale della distribuzione a priori. La presente combinazione di verosimiglianza e distribuzione a priori è chiamata caso coniugato *Beta-Binomiale* ed è descritto dal seguente teorema.

Teorema 2.1. *Sia data la funzione di verosimiglianza* $\text{Bin}(n, y \mid \theta)$ *e sia* $Beta(\alpha, \beta)$ *una distribuzione a priori. In tali circostanze, la distribuzione a posteriori del parametro* θ *sarà una distribuzione* $Beta(\alpha+y, \beta+n-y)$.

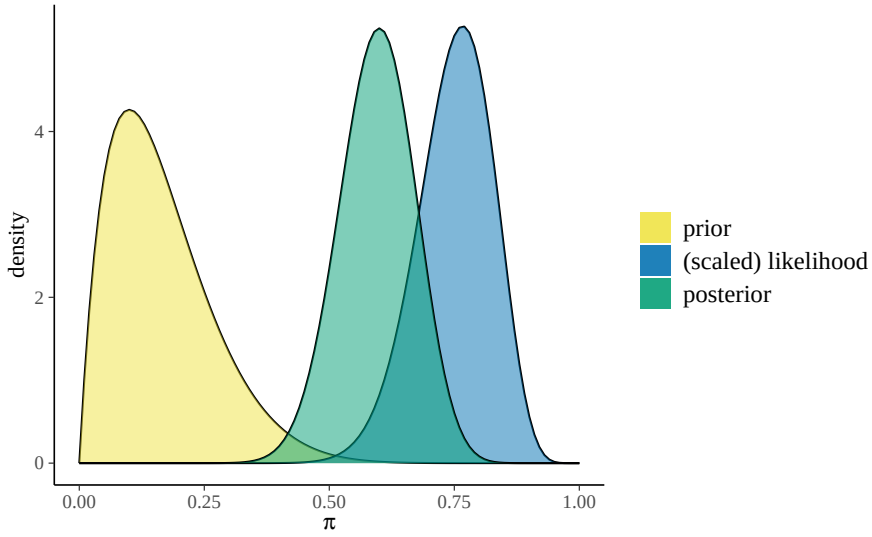
È facile calcolare il valore atteso a posteriori di θ. Essendo $\mathbb{E}[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha+\beta}$, il risultato cercato diventa

$$\mathbb{E}_{\text{post}}[\text{Beta}(\alpha + y, \beta + n - y)] = \frac{\alpha + y}{\alpha + \beta + n}. \tag{2.1}$$

Esercizio 2.1. Usando le funzione R `plot_beta_binomial()` e `plot_beta_binomial()` del pacchetto `bayesrules`, si rappresenti in maniera grafica e si descriva in forma numerica l'aggiornamento bayesiano Beta-Binomiale per i dati di Zetsche et al. (2019).

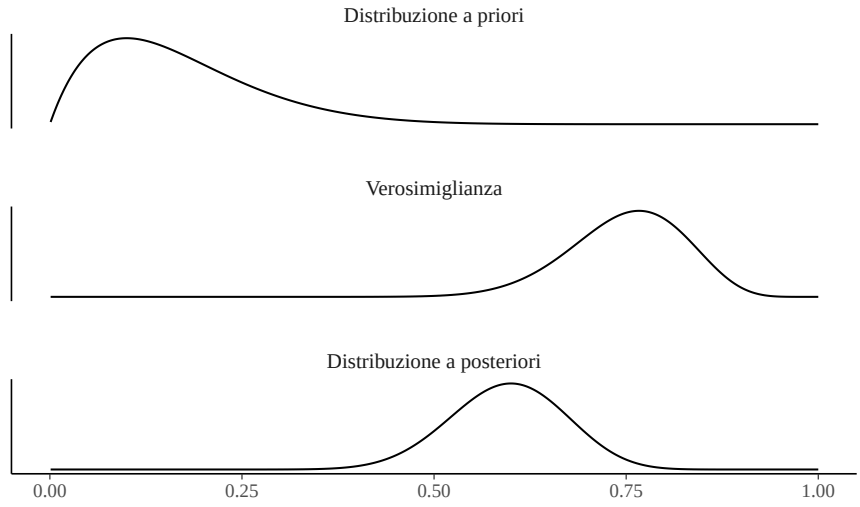
Per i dati in discussione, abbiamo:

```
bayesrules::plot_beta_binomial(
  alpha = 2, beta = 10, y = 23, n = 30
  )
```



È facile replicare il grafico precedente scrivendo noi stessi una funzione, come descritto in Appendice **??**. Con la funzione `plot_beta_binom()` e i dati usati in precedenza otteniamo

```
plot_beta_bin(2, 10, 23, 30)
```



Un sommario delle distribuzioni a priori e a posteriori si ottiene usando la funzione `summarize_beta_binomial()`:

```
bayesrules:::summarize_beta_binomial(
  alpha = 2, beta = 10, y = 23, n = 30
)
#>       model alpha beta   mean mode      var      sd
#> 1     prior     2   10 0.1667  0.1 0.010684 0.10336
#> 2 posterior    25   17 0.5952  0.6 0.005603 0.07485
```

Esercizio 2.2. Per i dati di Zetsche et al. (2019), si trovino la media, la moda, la deviazione standard della distribuzione a posteriori di θ. Si trovi inoltre l'intervallo di credibilità a posteriori del 95% per il parametro θ.

Usando la **??**, possiamo ottenere l'intervallo di credibilità a posteriori del 95% per il parametro θ come segue:

```
qbeta(c(0.025, 0.975), shape1 = 25, shape2 = 17)
#> [1] 0.4450 0.7368
```

La media della distribuzione a posteriori è

```
25 / (25 + 17)
#> [1] 0.5952
```

La moda della distribuzione a posteriori è

```
(25 - 1) / (25 + 17 - 2)
#> [1] 0.6
```

La deviazione standard della distribuzione a priori è

```
sqrt((25 * 17) / ((25 + 17)^2 * (25 + 17 + 1)))
#> [1] 0.07485
```

Esercizio 2.3. Si trovino i parametri e le proprietà della distribuzione a posteriori del parametro θ per i dati dell'esempio relativo alla ricerca di Stanley Milgram discussa da Johnson et al. (2022).

Nel 1963, Stanley Milgram presentò una ricerca sulla propensione delle persone a obbedire agli ordini di figure di autorità, anche quando tali ordini possono danneggiare altre persone (Milgram, 1963). Nell'articolo, Milgram descrive lo studio come *"consist[ing] of ordering a naive subject to administer electric shock to a victim. A simulated shock generator is used, with 30 clearly marked voltage levels that range from IS to 450 volts. The instrument bears verbal designations that range from Slight Shock to Danger: Severe Shock. The responses of the victim, who is a trained confederate of the experimenter, are standardized. The orders to administer shocks are given to the naive subject in the context of a 'learning experiment' ostensibly set up to study the effects of punishment on memory. As the experiment proceeds the naive subject is commanded to administer increasingly more intense shocks to the victim, even to the point of reaching the level marked Danger: Severe Shock."*

All'insaputa del partecipante, gli shock elettrici erano falsi e l'attore stava solo fingendo di provare il dolore dello shock.

Johnson et al. (2022) fanno inferenza sui risultati dello studio di Milgram mediante il modello Beta-Binomiale. Il parametro di interesse è θ, la probabiltà che una persona obbedisca all'autorità (in questo caso,

somministrando lo shock più severo), anche se ciò significa recare danno ad altri. Johnson et al. (2022) ipotizzano che, prima di raccogliere dati, le credenze di Milgram relative a θ possano essere rappresentate mediante una Beta$(1, 10)$. Sia $y = 26$ il numero di soggetti che, sui 40 partecipanti allo studio, aveva accettato di infliggere lo shock più severo. Assumendo che ogni partecipante si comporti indipendentemente dagli altri, possiamo modellare la dipendenza di y da θ usando la distribuzione binomiale. Giungiamo dunque al seguente modello bayesiano Beta-Binomiale:

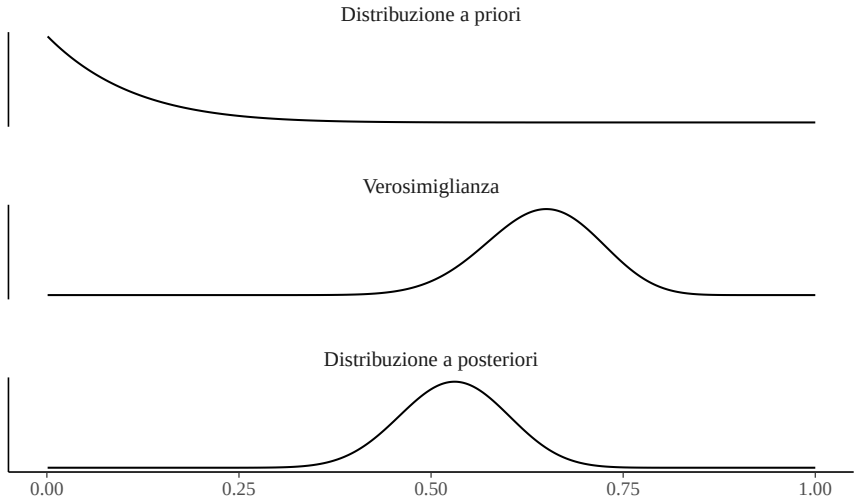
$$\begin{aligned} y \mid \theta &\sim \text{Bin}(n = 40, \theta) \\ \theta &\sim \text{Beta}(1, 10) \ . \end{aligned}$$

Usando le funzioni di `bayesrules` possiamo facilmente calcolare i parametri e le proprietà della distribuzione a posteriori:

```
bayesrules:::summarize_beta_binomial(
  alpha = 1, beta = 10, y = 26, n = 40
)
#>       model alpha beta    mean   mode      var      sd
#> 1     prior     1   10 0.09091 0.0000 0.006887 0.08299
#> 2 posterior    27   24 0.52941 0.5306 0.004791 0.06922
```

Il processo di aggiornamento bayesiano è descritto dalla figura seguente:

```
plot_beta_bin(1, 10, 26, 40)
```



2.4 Principali distribuzioni coniugate

Esistono molte altre combinazioni simili di verosimiglianza e distribuzione a priori le quali producono una distribuzione a posteriori che ha la stessa densità della distribuzione a priori. Sono elencate qui sotto le più note coniugazioni tra modelli statistici e distribuzioni a priori.

- Per il modello Normale-Normale $\mathcal{N}(\mu, \sigma_0^2)$, la distribizione iniziale è $\mathcal{N}(\mu_0, \tau^2)$ e la distribuzione finale è $\mathcal{N}\left(\frac{\mu_0\sigma^2+\bar{y}n\tau^2}{\sigma^2+n\tau^2}, \frac{\sigma^2\tau^2}{\sigma^2+n\tau^2}\right)$.
- Per il modello Poisson-gamma $\text{Po}(\theta)$, la distribizione iniziale è $\Gamma(\lambda, \delta)$ e la distribuzione finale è $\Gamma(\lambda + n\bar{y}, \delta + n)$.
- Per il modello esponenziale $\text{Exp}(\theta)$, la distribizione iniziale è $\Gamma(\lambda, \delta)$ e la distribuzione finale è $\Gamma(\lambda + n, \delta + n\bar{y})$.
- Per il modello uniforme-Pareto $\text{U}(0, \theta)$, la distribizione iniziale è $\text{Pa}(\alpha, \varepsilon)$ e la distribuzione finale è $\text{Pa}(\alpha + n, \max(y_{(n)}, \varepsilon))$.

Commenti e considerazioni finali

Lo scopo di questa discussione è stato quello di mostrare come sia possibile combinare le nostre conoscenze a priori (espresse nei termini di una densità di probabilità) con le evidenze fornite dai dati (espresse nei termini della funzione di verosimiglianza), così da determinare, mediante il teorema di Bayes, una distribuzione a posteriori, la quale condensa l'incertezza che abbiamo sul parametro θ. Per illustrare tale problema, abbiamo considerato una situazione nella quale θ corrisponde alla probabilità di successo in una sequenza di prove Bernoulliane. Abbiamo visto come, in queste circostanze, sia ragionevole esprimere le nostre credenze a priori mediante la densità Beta, con opportuni parametri. L'inferenza rispetto ad una proporzione rappresenta un caso particolare, ovvero un caso nel quale la distribuzione a priori è Beta e la verosimiglianza è Binomiale. In tali circostanze, la distribuzione a posteriori diventa una distribuzione Beta – questo è il cosiddetto modello Beta-Binomiale. Dato che utilizza una distribuzione a priori coniugata, dunque, il modello Beta-Binomiale rende possibile la determinazione analitica dei parametri della distribuzione a posteriori.

Bibliografia

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

Johnson, A. A., Ott, M., and Dogucu, M. (2022). *Bayes Rules! An Introduction to Bayesian Modeling with R*. CRC Press.

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

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, 67(4):371–378.

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

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