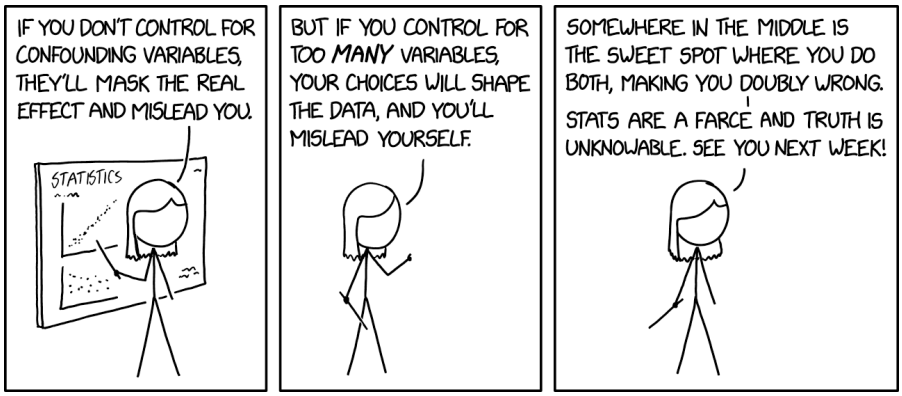
Corrado Caudek

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 bayesiana

1

Flusso di lavoro bayesiano

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

$$(\mathcal{Y}, p(y \mid \theta), p(\theta), \theta \in \Theta), \tag{1.1}$$

dove $\mathcal{Y}$ 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 $p(\theta)$. 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

$$p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{\int_\Theta p(y \mid \theta)p(\theta)\, \mathrm{d}\theta} \quad \theta \in \Theta, \tag{1.2}$$

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

Il denominatore del Teorema di Bayes (1.2), che costituisce la costante di normalizzazione, è la densità marginale dei dati (o verosimiglianza marginale). In ambito bayesiano la distribuzione a posteriori viene utilizzata per calcolare le principali quantità di interesse dell'inferenza, ad esempio la media a posteriori di θ.

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

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, 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 ven-

gono 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; di conseguenza, i parametri assumono una particolare distribuzione che nelle statistica bayesiana viene definita "a priori". I parametri (o il parametro), che possiamo indicare con θ, possono assumere delle distribuzioni a priori differenti: a seconda delle informazioni disponibili bisogna selezionare una distribuzione di θ in modo tale che venga assegnata una probabilità maggiore a quei valori che si ritengono più plausibili per θ. Idealmente, le credenze a priori che portano alla specificazione di una distribuzione a priori dovrebbero essere supportate da una qualche motivazione, come ad esempio i risultati di ricerche precedenti.

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) ad di un

singolo valore del parametro. Poiché questa non è più una distribuzione di probabilità, sebbene ne soddisfi la definizione, in questo caso si parla di una *distribuzione a priori degenerata*. La figura seguente motra alcuni esempi di distribuzioni a priori per il modello Binomiale:

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

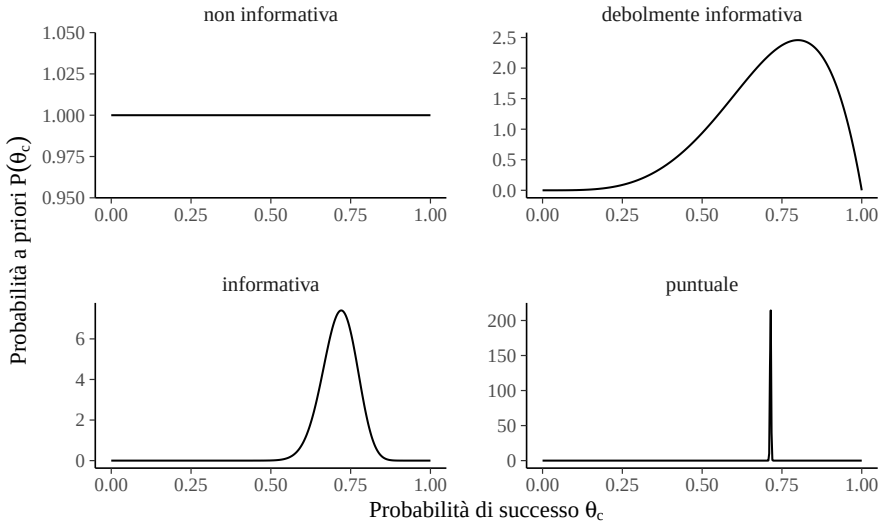


Figura 1.1: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

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 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 svol-

gono due importanti ruoli statistici: quello della "regolarizzazione della stima", ovvero, il processo che porta ad indebolire l'influenza indebita di osservazioni estreme, e quello del miglioramento dell'efficienza della stima, ovvero, la facilitazione dei processi di calcolo numerico di stima della distribuzione a posteriori. L'effetto della distribuzione a priori sulla distribuzione a posteriori verrà discusso in dettaglio nel Capitolo **??**.

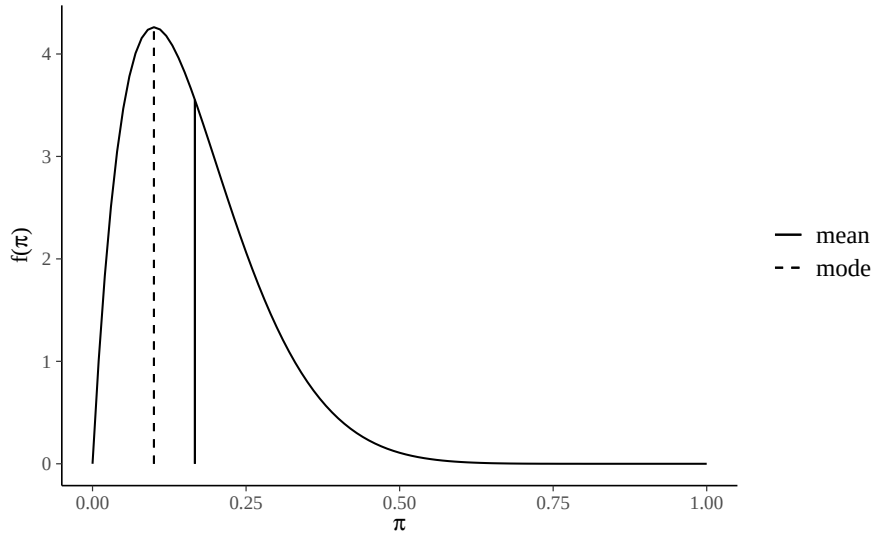
1.2.3 Un esempio concreto

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

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

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

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



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

1.3 La funzione di verosimiglianza

Iniziamo 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 descrive in termini relativi il sostegno empirico che $\theta \in \Theta$ riceve da y. Infatti, la funzione di verosimiglianza assume forme diverse al variare di y. Possiamo dunque pensare alla funzione di verosimiglianza come alla risposta alla seguente domanda: avendo osservato i dati y, quanto risultano (relativamente) credibili i diversi valori del parametro θ? In termini più formali possiamo dire: sulla base dei dati, $\theta_1 \in \Theta$ risulta più credibile di $\theta_2 \in \Theta$ quale indice del modello probabilistico generatore dei dati se $\mathcal{L}(\theta_1) > \mathcal{L}(\theta_1)$.

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

1.3.1 Notazione

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

1.3.2 La log-verosimiglianza

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

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

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 (per un approfondimento, si veda l'Appendice I):

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

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.

1.3.3 Un esempio concreto

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

$$y \sim \text{Bin}(n, \theta), \tag{1.5}$$

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

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

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

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

$$\mathcal{L}(\theta \mid y) = \frac{(23+7)!}{23!7!}\theta^{23} + (1-\theta)^7. \tag{1.6}$$

Per costruire la funzione di verosimiglianza dobbiamo applicare la (1.6) 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)
  )
```

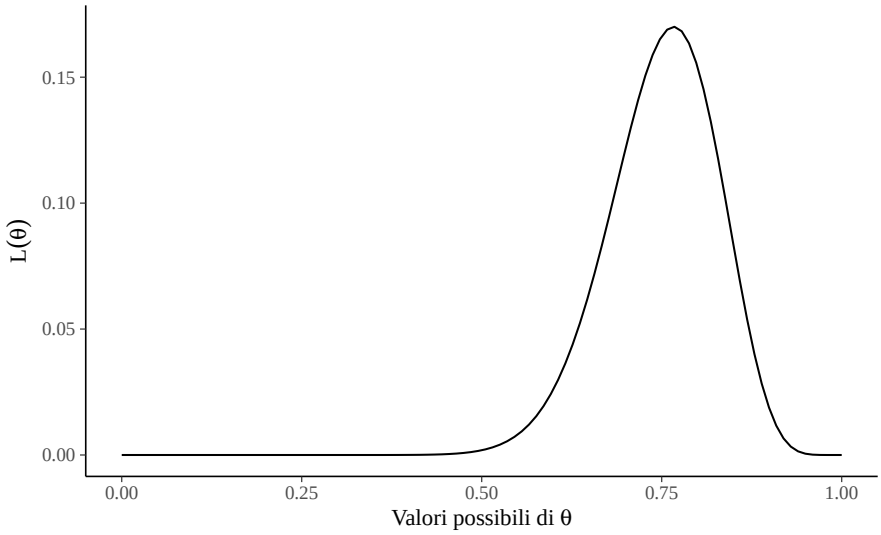


Figura 1.2: *Funzione di verosimiglianza nel caso di 23 successi in 30 prove.*

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 (la moda della funzione di verosimiglianza) è il valore più credibile di tutti.

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

1.4.1 Un esempio concreto

Consideriamo nuovamente i dati di Zetsche et al. (2019). Supponiamo che nel numeratore bayesiano la verosimiglianza sia moltiplicata per una distribuzione uniforme, ovvero Beta(1, 1). In tali circostanze, il prodotto si riduce alla funzione di verosimiglianza. Per i dati di Zetsche et al. (2019), dunque, la costante di normalizzazione si ottiene marginalizzando la funzione di verosimiglianza $p(y = 23, n = 30 \mid \theta)$ sopra θ, ovvero risolvendo l'integrale:

$$p(y = 23, n = 30) = \int_0^1 \binom{30}{23} \theta^{23}(1-\theta)^7 \, \mathrm{d}\theta. \tag{1.7}$$

Una soluzione numerica si trova facilmente usando R:

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

La derivazione analitica è fornita nell'Appendice M.

1.5 Distribuzione a posteriori

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

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

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:

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

Ripeto qui quanto detto sopra: le quantità di interesse della statistica bayesiana(costante di normalizzazione, valore atteso della distribuzione a posteriori, ecc.) contengono integrali che risultano, nella maggior parte dei casi, impossibili da risolvere analiticamente. Per questo motivo, si ricorre a metodi di stima numerici, in particolare a quei metodi Monte Carlo basati sulle proprietà delle catene di Markov (MCMC). Questo argomento verrà discusso nel Capitolo **??**.

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*:

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

La (1.8) descrive la distribuzione prevista dei dati in base al modello (che include la distribuzione a priori e la verosimiglianza), ovvero descrive i dati y^* che ci aspettiamo di osservare, dato il modello, prima di avere osservato i dati del campione.

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

1.7 Distribuzione predittiva a posteriori

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

$$p(\tilde{y} \mid y) = \int_{\Theta} p(\tilde{y} \mid \theta) p(\theta \mid y) \, \mathrm{d}\theta. \tag{1.9}$$

Questa è la distribuzione dei dati attesi futuri $\tilde{y}$ alla luce della distribuzione a posteriori $p(\theta \mid y)$, che a sua volta è una conseguenza del modello adottato (distribuzione a priori e verosimiglianza) e dei dati osservati. In altre parole, questi sono i dati che il modello si aspetta dopo aver osservato i dati de campione. Dalla (1.9) possiamo vedere che le previsioni sui dati attesi futuri sono calcolate integrando (o marginalizzando) sulla distribuzione a posteriori dei parametri. Di conseguenza, le previsioni calcolate in questo modo incorporano l'incertezza relativa alla stima dei parametri del modello.

Commenti e considerazioni finali

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

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 lo schema beta-binomiale.

2.1 Lo schema beta-binomiale

Iniziamo con una definizione.

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 distribuzione a priori è una distribuzione Beta e se la funzione di verosimiglianza è binomiale, allora anche la distribuzione a posteriori sarà una distribuzione Beta.

Da un punto di vista 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 le distribuzioni a priori che risultano matematicamente convenienti. Tuttavia, le famiglie coniugate offrono un utile ausilio didattico nello studio dell'inferenza bayesiana. Questo è il motivo per cui le esamineremo qui. Nello specifico, esamineremo quello che viene chiamato lo schema beta-binomiale.

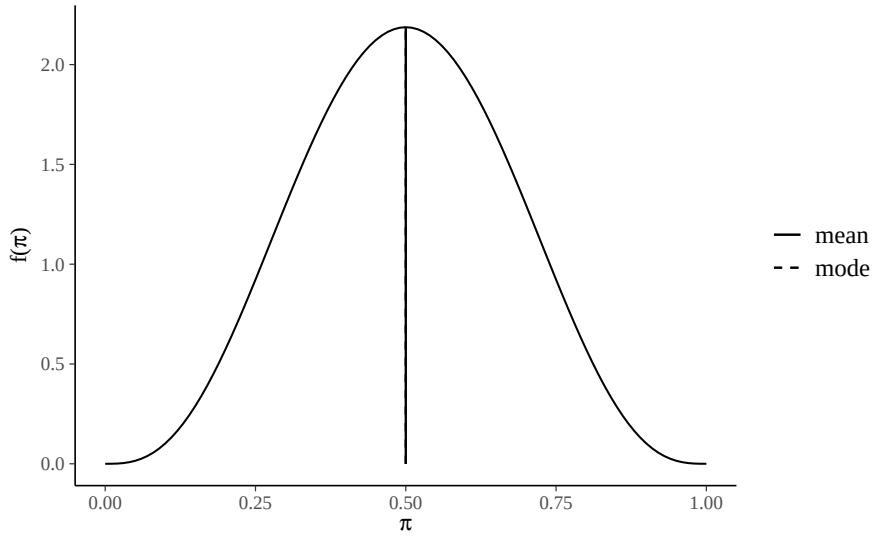
Per fare un esempio concreto, consideriamo nuovamente i dati di Zetsche et al. (2019): nel campione di 30 partecipanti clinici le aspettative future di 23 partecipanti risultano negativamente distorte mentre quelle di 7 partecipanti risultano positivamente distorte. Nel seguito, indicheremo con θ la 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.

2.1.1 La specificazione della distribuzione a priori

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)
```

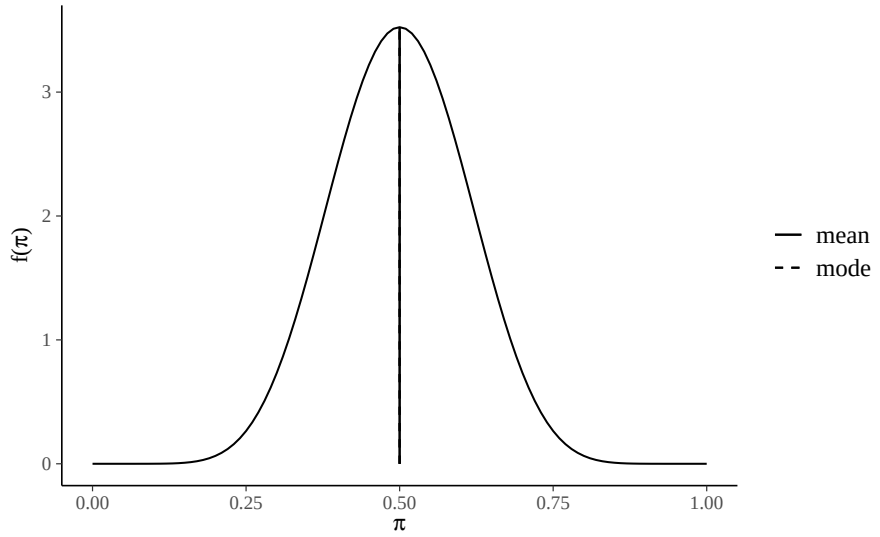


Possiamo quantificare la nostra incertezza calcolando, con un grado di fiducia del 95%, la regione nella quale, in base a tale credenza a priori, si trova il valore del parametro. Per ottenere tale intervallo di credibilità a priori, usiamo la funzione `qbeta()` di R. In `qbeta()` i parametri α e β sono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

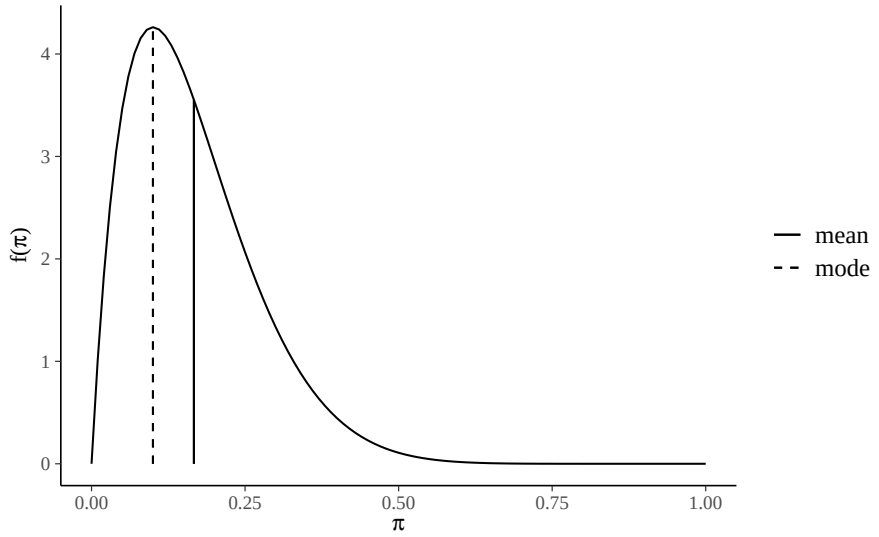
```
qbeta(c(0.025, 0.975), shape1 = 10, shape2 = 10)
#> [1] 0.2886 0.7114
```

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

Nella discussione presente, solo per fare un esempio, useremo quale distribuzione a priori una 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.

2.1.2 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.

Riscriviamo l'equazione precedente in termini più generali:

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

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, nel caso dello schema beta-binomiale, la distribuzione a posteriori è una $\text{Beta}(a + y, b + n - y)$:

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

In sintesi, moltiplicando verosimiglianza $\text{Bin}(n = 30, y = 23 \mid \theta)$ per la la distribuzione a priori $\theta \sim \text{Beta}(2, 10)$ e dividendo per la costante di normalizzazione, abbiamo ottenuto la distribuzione a posteriori $p(\theta \mid n, y) \sim \text{Beta}(25, 17)$. Questo è un esempio di analisi coniugata. La presente combinazione di verosimiglianza e distribuzione a priori è chiamata caso coniugato *beta-binomiale* ed è descritta dal seguente teorema.

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

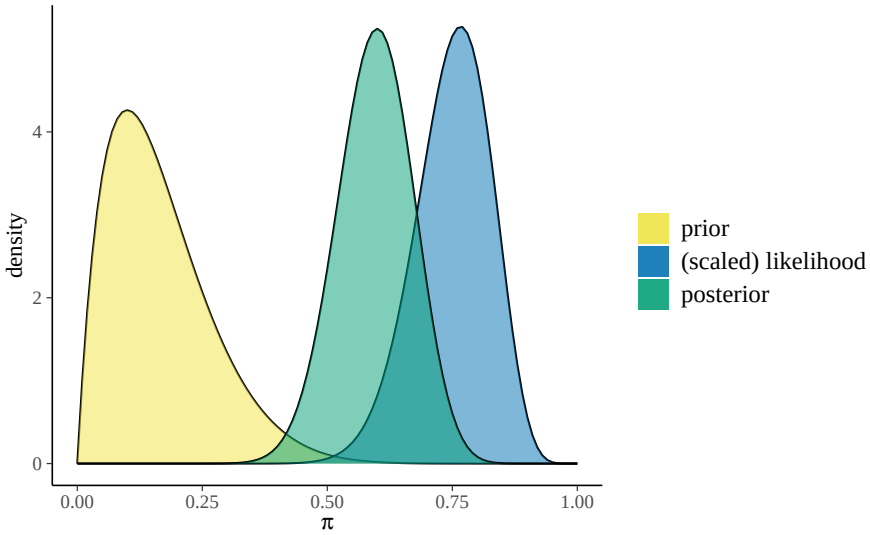
È facile calcolare il valore atteso a posteriori di θ. Essendo $\mathbb{E}[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha+\beta}$, il risultato cercato diventa

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

Esercizio 2.1. Si rappresenti in maniera grafica e si descriva in forma numerica l'aggiornamento bayesiano beta-binomiale per i dati di Zetsche et al. (2019). Si assuma una distribuzione a priori Beta(2, 10).

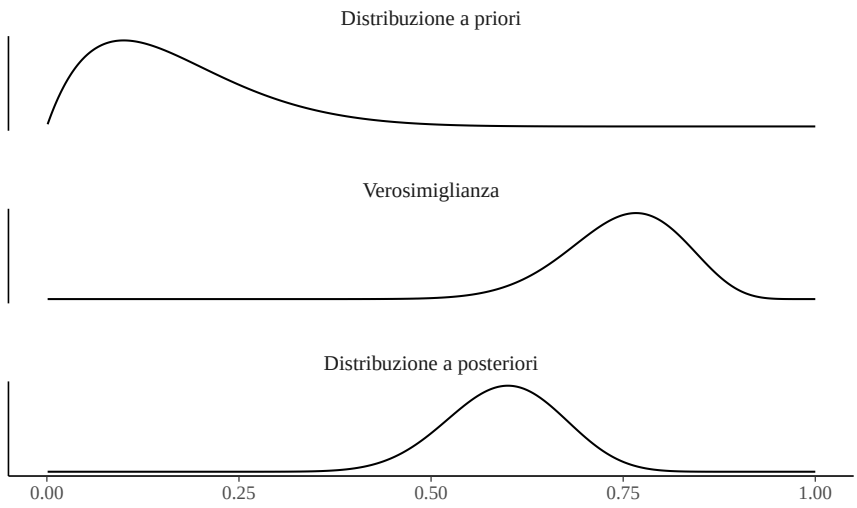
Per i dati in questione, l'aggiornamento bayesiano può essere rappresentato in forma grafica usando la funzione `plot_beta_binomial()` del pacchetto `bayesrules`:

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



Oppure, possiamo scrivere noi stessi una funzione, come ad esempio la funzione `plot_beta_binom()` riportata in Appendice K. Mediante tale la funzione otteniamo

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



Un sommario delle distribuzioni a priori e a posteriori può essere ottenuto, ad esempio, usando la funzione `summarize_beta_binomial()` del pacchetto `bayesrules`:

```
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 il Teorema 2.1, l'intervallo di credibilità a posteriori del 95% per il parametro θ è:

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

Usando la (2.1), la media della distribuzione a posteriori è

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

Per le proprietà della distribuzione Beta, la moda della distribuzione a posteriori è

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

e 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:

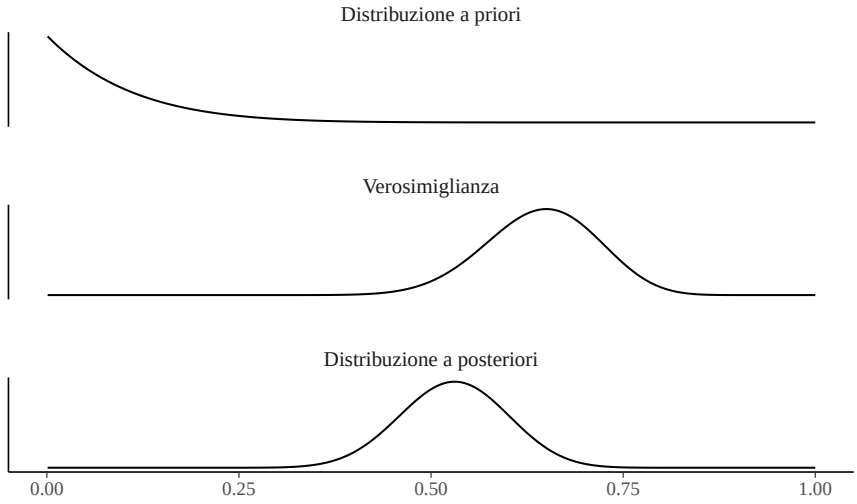
$$\begin{aligned} y \mid \theta &\sim \text{Bin}(n = 40, \theta) \\ \theta &\sim \text{Beta}(1, 10) \ . \end{aligned}$$

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)
```



2.2 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 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. In tali circostanze è 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 è essa stessa una distribuzione Beta – questo è il cosiddetto schema beta-binomiale. Dato che utilizza una distribuzione a priori coniugata, lo schema beta-binomiale rende possibile la determinazione analitica dei parametri della distribuzione a posteriori.

Parte II

Appendici

A

Simbologia di base

Per una scrittura più sintetica possono essere utilizzati alcuni simboli matematici.

- $\log(x)$: il logaritmo naturale di x.
- L'operatore logico booleano $\wedge$ significa "e" (congiunzione forte) mentre il connettivo di disgiunzione $\vee$ significa "o" (oppure) (congiunzione debole).
- Il quantificatore esistenziale $\exists$ vuol dire "esiste almeno un" e indica l'esistenza di almeno una istanza del concetto/oggetto indicato. Il quantificatore esistenziale di unicità $\exists!$ ("esiste soltanto un") indica l'esistenza di esattamente una istanza del concetto/oggetto indicato. Il quantificatore esistenziale $\nexists$ nega l'esistenza del concetto/oggetto indicato.
- Il quantificatore universale $\forall$ vuol dire "per ogni."
- $\mathcal{A}, \mathcal{S}$: insiemi.
- $x \in A$: x è un elemento dell'insieme A.
- L'implicazione logica "$\Rightarrow$" significa "implica" (se …allora). $P \Rightarrow Q$ vuol dire che P è condizione sufficiente per la verità di Q e che Q è condizione necessaria per la verità di P.
- L'equivalenza matematica " $\Longleftrightarrow$ " significa "se e solo se" e indica una condizione necessaria e sufficiente, o corrispondenza biunivoca.
- Il simbolo | si legge "tale che."
- Il simbolo $\triangleq$ (o $:=$) si legge "uguale per definizione."
- Il simbolo Δ indica la differenza fra due valori della variabile scritta a destra del simbolo.
- Il simbolo $\propto$ si legge "proporzionale a."
- Il simbolo $\approx$ si legge "circa."
- Il simbolo $\in$ della teoria degli insiemi vuol dire "appartiene" e indica l'appartenenza di un elemento ad un insieme. Il simbolo $\notin$ vuol dire "non appartiene."
- Il simbolo $\subseteq$ si legge "è un sottoinsieme di" (può coincidere con l'insieme stesso). Il simbolo $\subset$ si legge "è un sottoinsieme proprio di."

- Il simbolo # indica la cardinalità di un insieme.
- Il simbolo $\cap$ indica l'intersezione di due insiemi. Il simbolo $\cup$ indica l'unione di due insiemi.
- Il simbolo $\emptyset$ indica l'insieme vuoto o evento impossibile.
- In matematica, argmax identifica l'insieme dei punti per i quali una data funzione raggiunge il suo massimo. In altre parole, $\text{argmax}_x f(x)$ è l'insieme dei valori di x per i quali $f(x)$ raggiunge il valore più alto.
- a, c, α, γ: scalari.
- x, y: vettori.
- X, Y: matrici.
- $X \sim p$: la variabile casuale X si distribuisce come p.
- $p(\cdot)$: distribuzione di massa o di densità di probabilità.
- $p(y \mid x)$: la probabilità o densità di y dato x, ovvero $p(y = Y \mid x = X)$.
- $f(x)$: una funzione arbitraria di x.
- $f(X; \theta, \gamma)$: f è una funzione di X con parametri θ, γ. Questa notazione indica che X sono i dati che vengono passati ad un modello di parametri θ, γ.
- $\mathcal{N}(\mu, \sigma^2)$: distribuzione gaussiana di media μ e varianza $sigma^2$.
- Beta(α, β): distribuzione Beta di parametri α e β.
- $\mathcal{U}(a, b)$: distribuzione uniforme con limite inferiore a e limite superiore b.
- Cauchy(α, β): distribuzione di Cauchy di parametri α (posizione: media) e β (scala: radice quadrata della varianza).
- $\mathcal{B}(p)$: distribuzione di Bernoulli di parametro p (probabilità di successo).
- Bin(n, p): distribuzione binomiale di parametri n (numero di prove) e p (probabilità di successo).
- $\mathbb{KL}(p \mid\mid q)$: la divergenza di Kullback-Leibler da p a q.

B

Numeri binari, interi, razionali, irrazionali e reali

B.1 Numeri binari

I numeri più semplici sono quelli binari, cioè zero o uno. Useremo spesso numeri binari per indicare se qualcosa è vero o falso, presente o assente. I numeri binari sono molto utili per ottenere facilmente delle statistiche riassuntive in R.Supponiamo di chiedere a 10 studenti "Ti piacciono i mirtilli?" Poniamo che le risposte siano le seguenti:

```
opinion <- c(
  "Yes", "No", "Yes", "No", "Yes", "No", "Yes",
  "Yes", "Yes", "Yes"
)
opinion
#>  [1] "Yes" "No"  "Yes" "No"  "Yes" "No"  "Yes" "Yes"
#>  [9] "Yes" "Yes"
```

Tali risposte possono essere ricodificate nei termini di valori di verità, ovvero, vero e falso, generalmente denotati rispettivamente come 1 e 0. In R tale ricodifica può essere effettuata mediante l'operatore == che è un test per l'uguaglianza e restituisce il valore logico VERO se i due oggetti valutati sono uguali e FALSO se non lo sono:

```
opinion <- opinion == "Yes"
opinion
#>  [1]  TRUE FALSE  TRUE FALSE  TRUE FALSE  TRUE  TRUE
#>  [9]  TRUE  TRUE
```

R considera i valori di verità e i numeri binari in modo equivalente, con TRUE uguale a 1 e FALSE uguale a zero. Di conseguenza, possiamo effettuare operazioni algebriche sui valori logici VERO e FALSO. Nell'esempio, possiamo sommare i valori di verità e dividere per 10

```
sum(opinion) / length(opinion)
#> [1] 0.7
```

in modo tale da calcolare una propozione, il che ci consente di concludere che 7 risposte su 10 sono positive.

B.2 Numeri interi

Un numero intero è un numero senza decimali. Si dicono **naturali** i numeri che servono a contare, come 1, 2, … L'insieme dei numeri naturali si indica con il simbolo $\mathbb{N}$. È anche necessario introdurre i numeri con il segno per poter trattare grandezze negative. Si ottengono così l'insieme numerico dei numeri interi relativi: $\mathbb{Z} = \{0, \pm 1, \pm 2, ... \}$

B.3 Numeri razionali

I numeri razionali sono i numeri frazionari m/n, dove $m, n \in N$, con $n \neq 0$. Si ottengono così i numeri razionali: $\mathbb{Q} = \{\frac{m}{n} \,|\, m, n \in \mathbb{Z}, n \neq 0\}$. È evidente che $\mathbb{N} \subseteq \mathbb{Z} \subseteq \mathbb{Q}$. Anche in questo caso è necessario poter trattare grandezze negative. I numeri razionali non negativi sono indicati con $\mathbb{Q}^+ = \{q \in \mathbb{Q} \,|\, q \geq 0\}$.

B.4 Numeri irrazionali

Tuttavia, non tutti i punti di una retta r possono essere rappresentati mediante i numeri interi e razionali. È dunque necessario introdurre un'altra classe di numeri. Si dicono *irrazionali*, e sono denotati con $\mathbb{R}$, i

numeri che possono essere scritti come una frazione a/b, con a e b interi e b diverso da 0. I numeri irrazionali sono i numeri illimitati e non periodici che quindi non possono essere espressi sotto forma di frazione. Per esempio, $\sqrt{2}$, $\sqrt{3}$ e $\pi = 3,141592...$ sono numeri irrazionali.

B.5 Numeri reali

I punti della retta r sono quindi "di più" dei numeri razionali. Per poter rappresentare tutti i punti della retta abbiamo dunque bisogno dei numeri *reali*. I numeri reali possono essere positivi, negativi o nulli e comprendono, come casi particolari, i numeri interi, i numeri razionali e i numeri irrazionali. Spesso in statisticac il numero dei decimali indica il grado di precisione della misurazione.

B.6 Intervalli

Un intervallo si dice chiuso se gli estremi sono compresi nell'intervallo, aperto se gli estremi non sono compresi. Le caratteristiche degli intervalli sono riportate nella tabella seguente.

Intervallo		
chiuso	$[a, b]$	$a \leq x \leq b$
aperto	(a, b)	$a < x < b$
chiuso a sinistra e aperto a destra	$[a, b)$	$a \leq x < b$
aperto a sinistra e chiuso a destra	$(a, b]$	$a < x \leq b$

C

Insiemi

Un insieme (o collezione, classe, gruppo, …) è un concetto primitivo, ovvero è un concetto che già possediamo. Georg Cantor l'ha definito nel modo seguente: *un insieme è una collezione di oggetti, determinati e distinti, della nostra percezione o del nostro pensiero, concepiti come un tutto unico; tali oggetti si dicono elementi dell'insieme.*

Mentre non è rilevante la natura degli oggetti che costituiscono l'insieme, ciò che importa è distinguere se un dato oggetto appartenga o meno ad un insieme. Deve essere vera una delle due possibilità: il dato oggetto è un elemento dell'insieme considerato oppure non è elemento dell'insieme considerato. Due insiemi A e B si dicono uguali se sono formati dagli stessi elementi, anche se disposti in ordine diverso: $A = B$. Due insiemi A e B si dicono diversi se non contengono gli stessi elementi: $A \neq B$. Ad esempio, i seguenti insiemi sono uguali:

$$\{1, 2, 3\} = \{3, 1, 2\} = \{1, 3, 2\} = \{1, 1, 1, 2, 3, 3, 3\}.$$

Gli insiemi sono denotati da una lettera maiuscola, mentre le lettere minuscole, di solito, designano gli elementi di un insieme. Per esempio, un generico insieme A si indica con

$$A = \{a_1, a_2, \dots, a_n\}, \quad \text{con } n > 0.$$

La scrittura $a \in A$ dice che a è un elemento di A. Per dire che b non è un elemento di A si scrive $b \notin A$.

Per quegli insiemi i cui elementi soddisfano una certa proprietà che li caratterizza, tale proprietà può essere usata per descrivere più sinteticamente l'insieme:

$$A = \{x \mid \text{proprietà posseduta da } x\},$$

che si legge come "A è l'insieme degli elementi x per cui è vera la proprietà indicata." Per esempio, per indicare l'insieme A delle coppie di numeri reali (x, y) che appartengono alla parabola $y = x^2 + 1$ si può scrivere:

$$A = \{(x, y) \mid y = x^2 + 1\}.$$

Dati due insiemi A e B, diremo che A è un *sottoinsieme* di B se e solo se tutti gli elementi di A sono anche elementi di B:

$$A \subseteq B \iff (\forall x \in A \Rightarrow x \in B).$$

Se esiste almeno un elemento di B che non appartiene ad A allora diremo che A è un *sottoinsieme proprio* di B:

$$A \subset B \iff (A \subseteq B, \exists\, x \in B \mid x \notin A).$$

Un altro insieme, detto *insieme delle parti*, o insieme potenza, che si associa all'insieme A è l'insieme di tutti i sottoinsiemi di A, inclusi l'insieme vuoto e A stesso. Per esempio, per l'insieme $A = \{a, b, c\}$, l'insieme delle parti è:

$$\mathcal{P}(A) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{c, b\}, \{a, b, c\}\}.$$

C.1 Operazioni tra insiemi

Si definisce *intersezione* di A e B l'insieme $A \cap B$ di tutti gli elementi x che appartengono ad A e contemporaneamente a B:

$$A \cap B = \{x \mid x \in A \wedge x \in B\}.$$

Si definisce *unione* di A e B l'insieme $A \cup B$ di tutti gli elementi x che appartengono ad A o a B, cioè

$$A \cup B = \{x \mid x \in A \vee x \in B\}.$$

Differenza. Si indica con $A \setminus B$ l'insieme degli elementi di A che non appartengono a B:

$$A \quad B = \{x \mid x \in A \wedge x \notin B\}.$$

Insieme complementare. Nel caso che sia $B \subseteq A$, l'insieme differenza $A \quad B$ è detto insieme complementare di B in A e si indica con B^C.

Dato un insieme S, una *partizione* di S è una collezione di sottoinsiemi di S, $S_1, \ldots, S_k$, tali che

$$S = S_1 \cup S_2 \cup \ldots S_k$$

e

$$S_i \cap S_j, \quad \text{con } i \neq j.$$

La relazione tra unione, intersezione e insieme complementare è data dalle leggi di DeMorgan:

$$(A \cup B)^c = A^c \cap B^c,$$
$$(A \cap B)^c = A^c \cup B^c.$$

C.2 Diagrammi di Eulero-Venn

In molte situazioni è utile servirsi dei cosiddetti diagrammi di Eulero-Venn per rappresentare gli insiemi e verificare le proprietà delle operazioni tra insiemi (si veda la figura C.1. I diagrammi di Venn sono così nominati in onore del matematico inglese del diciannovesimo secolo John Venn anche se Leibnitz e Eulero avevano già in precedenza utilizzato rappresentazioni simili. In tale rappresentazione, gli insiemi sono individuati da regioni del piano delimitate da una curva chiusa. Nel caso di insiemi finiti, è possibile evidenziare esplicitamente alcuni elementi di un insieme mediante punti, quando si possono anche evidenziare tutti gli elementi degli insiemi considerati.

I diagrammi di Eulero-Venn che forniscono una dimostrazione delle leggi di DeMorgan sono forniti nella figura C.2.

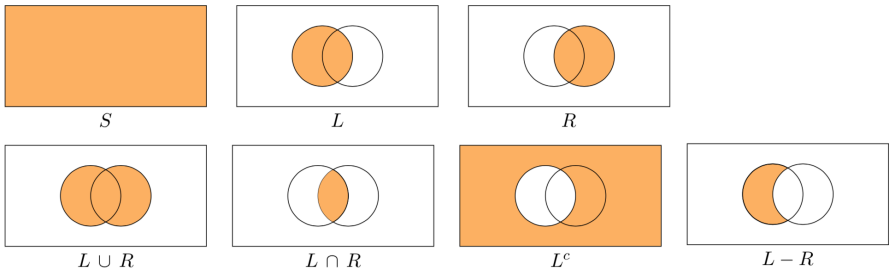


Figura C.1: *In tutte le figure S è la regione delimitata dal rettangolo, L è la regione all'interno del cerchio di sinistra e R è la regione all'interno del cerchio di destra. La regione evidenziata mostra l'insieme indicato sotto ciascuna figura.*

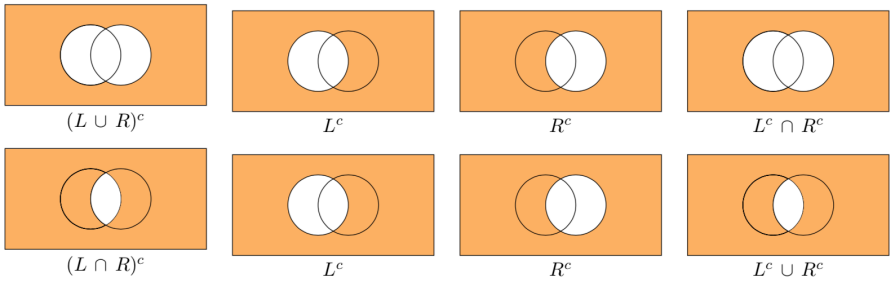


Figura C.2: *Dimostrazione delle leggi di DeMorgan.*

C.3 Coppie ordinate e prodotto cartesiano

Una coppia ordinata (x, y) è l'insieme i cui elementi sono $x \in A$ e $y \in B$ e nella quale x è la prima componente (o prima coordinata), y la seconda. L'insieme di tutte le coppie ordinate costruite a partire dagli insiemi A e B viene detto **prodotto cartesiano**:

$$A \times B = \{(x, y) \mid x \in A \wedge y \in B\}.$$

Ad esempio, sia $A = \{1, 2, 3\}$ e $B = \{a, b\}$. Allora,

$$\{1, 2\} \times \{a, b, c\} = \{(1, a), (1, b), (1, c), (2, a), (2, b), (2, c)\}.$$

C.4 Cardinalità

Si definisce *cardinalità* (o potenza) di un insieme finito il numero degli elementi dell'insieme. Viene indicata con $|A|, \#(A)$ o $\mathrm{c}(A)$.

D

Simbolo di somma (sommatorie)

Le somme si incontrano costantemente in svariati contesti matematici e statistici quindi abbiamo bisogno di una notazione adeguata che ci consenta di gestirle. La somma dei primi n numeri interi può essere scritta come $1 + 2 + \cdots + (n - 1) + n$, dove '...' ci dice di completare la sequenza definita dai termini che vengono prima e dopo. Ovviamente, una notazione come $1 + 7 + \cdots + 73.6$ non avrebbe alcun senso senza qualche altro tipo di precisazione. In generale, nel seguito incontreremo delle somme nella forma

$$x_1 + x_2 + \cdots + x_n,$$

dove x_i è un numero che è stato definito altrove. La notazione precedente, che fa uso dei tre puntini di sospensione, è utile in alcuni contesti ma in altri risulta ambigua. Pertanto la notazione di uso corrente è del tipo

$$\sum_{i=1}^{n} x_i$$

e si legge "sommatoria per i che va da 1 a n di x_i". Il simbolo $\sum$ (lettera sigma maiuscola dell'alfabeto greco) indica l'operazione di somma, il simbolo x_i indica il generico addendo della sommatoria, le lettere 1 ed n indicano i cosiddetti *estremi della sommatoria*, ovvero l'intervallo (da 1 fino a n estremi inclusi) in cui deve variare l'indice i allorché si sommano gli addendi x_i. Solitamente l'estremo inferiore è 1 ma potrebbe essere qualsiasi altri numero $m < n$. Quindi

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n.$$

Per esempio, se i valori x sono $\{3, 11, 4, 7\}$, si avrà

$$\sum_{i=1}^{4} x_i = 3 + 11 + 4 + 7 = 25$$

laddove $x_1 = 3$, $x_2 = 11$, eccetera. La quantità x_i nella formula precedente si dice l'*argomento* della sommatoria, mentre la variabile i, che prende i valori naturali successivi indicati nel simbolo, si dice *indice* della sommatoria.

La notazione di sommatoria può anche essere fornita nella forma seguente

$$\sum_{P(i)} x_i$$

dove $P(i)$ è qualsiasi proposizione riguardante i che può essere vera o falsa. Quando è ovvio che si vogliono sommare tutti i valori di n osservazioni, la notazione può essere semplificata nel modo seguente: $\sum_i x_i$ oppure $\sum x_i$. Al posto di i si possono trovare altre lettere: $k, j, l, ...,$.

D.1 Manipolazione di somme

È conveniente utilizzare le seguenti regole per semplificare i calcoli che coinvolgono l'operatore della sommatoria.

D.1.1 Proprietà 1

La sommatoria di n valori tutti pari alla stessa costante a è pari a n volte la costante stessa:

$$\sum_{i=1}^{n} a = \underbrace{a + a + \cdots + a}_{n \text{ volte}} = na.$$

D.1.2 Proprietà 2 (proprietà distributiva)

Nel caso in cui l'argomento contenga una costante, è possibile riscrivere la sommatoria. Ad esempio con

$$\sum_{i=1}^{n} ax_i = ax_1 + ax_2 + \cdots + ax_n$$

è possibile raccogliere la costante a e fare $a(x_1 + x_2 + \cdots + x_n)$. Quindi possiamo scrivere

$$\sum_{i=1}^{n} ax_i = a \sum_{i=1}^{n} x_i.$$

D.1.3 Proprietà 3 (proprietà associativa)

Nel caso in cui

$$\sum_{i=1}^{n} (a + x_i) = (a + x_1) + (a + x_1) + ... (a + x_n)$$

si ha che

$$\sum_{i=1}^{n} (a + x_i) = na + \sum_{i=1}^{n} x_i.$$

È dunque chiaro che in generale possiamo scrivere

$$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$$

D.1.4 Proprietà 4

Se deve essere eseguita un'operazione algebrica (innalzamento a potenza, logaritmo, ecc.) sull'argomento della sommatoria, allora tale operazione algebrica deve essere eseguita prima della somma. Per esempio,

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2 \neq \left(\sum_{i=1}^{n} x_i \right)^2.$$

D.1.5 Proprietà 5

Nel caso si voglia calcolare $\sum_{i=1}^{n} x_i y_i$, il prodotto tra i punteggi appaiati deve essere eseguito prima e la somma dopo:

$$\sum_{i=1}^{n} x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n,$$

infatti, $a_1 b_1 + a_2 b_2 \neq (a_1 + a_2)(b_1 + b_2)$.

D.2 Doppia sommatoria

È possibile incontrare la seguente espressione in cui figurano una doppia sommatoria e un doppio indice:

$$\sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij}.$$

La doppia sommatoria comporta che per ogni valore dell'indice esterno, i da 1 ad n, occorre sviluppare la seconda sommatoria per j da 1 ad m. Quindi,

$$\sum_{i=1}^{3} \sum_{j=4}^{6} x_{ij} = (x_{1,4}+x_{1,5}+x_{1,6})+(x_{2,4}+x_{2,5}+x_{2,6})+(x_{3,4}+x_{3,5}+x_{3,6}).$$

Un caso particolare interessante di doppia sommatoria è il seguente:

$$\sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j$$

Si può osservare che nella sommatoria interna (quella che dipende dall'indice j), la quantità x_i è costante, ovvero non dipende dall'indice (che è j). Allora possiamo estrarre x_i dall'operatore di sommatoria interna e scrivere

$$\sum_{i=1}^{n} \left(x_i \sum_{j=1}^{n} y_j \right).$$

Allo stesso modo si può osservare che nell'argomento della sommatoria esterna la quantità costituita dalla sommatoria in j non dipende dall'indice i e quindi questa quantità può essere estratta dalla sommatoria esterna. Si ottiene quindi

$$\sum_{i=1}^{n}\sum_{j=1}^{n} x_i y_j = \sum_{i=1}^{n}\left(x_i \sum_{j=1}^{n} y_j\right) = \sum_{i=1}^{n}\ x_i \sum_{j=1}^{n} y_j.$$

Esercizio D.1. Si verifichi quanto detto sopra nel caso particolare di $x = \{2, 3, 1\}$ e $y = \{1, 4, 9\}$, svolgendo prima la doppia sommatoria per poi verificare che quanto così ottenuto sia uguale al prodotto delle due sommatorie.

$$\begin{aligned}\sum_{i=1}^{3}\sum_{j=1}^{3} x_i y_j &= x_1y_1 + x_1y_2 + x_1y_3 + x_2y_1 + x_2y_2 + x_2y_3 + x_3y_1 + x_3y_2 + x_3y_3 \\ &= 2 \times (1 + 4 + 9) + 3 \times (1 + 4 + 9) + 2 \times (1 + 4 + 9) = 84,\end{aligned}$$

ovvero

$$(2 + 3 + 1) \times (1 + 4 + 9) = 84.$$

D.3 Sommatorie (e produttorie) e operazioni vettoriali in R

Si noti che la notazione

$$\sum_{n=0}^{4} 3n$$

non è altro che un ciclo `for`:

```
sum <- 0
for (n in 0:4) {
  sum = sum + 3 * n
}
sum
#> [1] 30
```

In maniera equivalente, e più semplice, possiamo scrivere

```
sum(3 * (0:4))
#> [1] 30
```

Allo stesso modo, la notazione

$$\prod_{n=1}^{4} 2n$$

è anch'essa equivalente al ciclo `for`

```
prod <- 1
for (n in 1:4) {
  prod <- prod * 2 * n
}
prod
#> [1] 384
```

che si può scrivere, più semplicemente, come

```
prod(2 * (1:4))
#> [1] 384
```

In entrambi i casi precedenti, abbiamo sostituito le operazioni aritmetiche eseguite all'interno di un ciclo `for` con le stesse operazioni aritmetiche eseguite sui vettori elemento per elemento.

E

Aggiornamento Bayesiano

Per fornire un esempio di aggiornamento bayesiano, consideriamo il seguente problema. Supponiamo che, per qualche strano errore di produzione, una fabbrica produca due tipi di monete. Il primo tipo di monete ha la caratteristica che, quando una moneta viene lanciata, la probabilità di osservare l'esito "testa" è 0.6. Per semplicità, sia θ la probabilità di osservare l'esito "testa". Per una moneta del primo tipo, dunque, $\theta = 0.6$. Per una moneta del secondo tipo, invece, la probabilità di produrre l'esito "testa" è 0.4. Ovvero, $\theta = 0.4$.

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

Ora immaginiamo di lanciare una moneta due volte e di ottenere il risultato seguente: $\{T, C\}$. Quello che ci chiediamo è: sulla base di questa evidenza, come cambiano le probabilità che associamo alle due ipotesi? In altre parole, ci chiediamo qual è la probabilità di ciascuna ipotesi alla luce dei dati che sono stati osservati: $P(H \mid y)$, laddove y sono i dati osservati. Tale probabilità si chiama probabilità a posteriori. Inoltre, se confrontiamo le due ipotesi, ci chiediamo quale valore assuma il rapporto $\frac{P(H_1|y)}{P(H_2|y)}$. Tale rapporto ci dice quanto è più probabile H_1 rispetto ad H_2, alla luce dei dati osservati. Infine, ci chiediamo come cambia il rapporto definito sopra, quando osserviamo via via nuovi risultati prodotti dal lancio della moneta.

Definiamo il problema in maniera più chiara. Conosciamo le probabilità

a priori, ovvero $P(H_1) = 0.75$ e $P(H_1) = 0.25$. Quello che vogliamo conoscere sono le probabilità a posteriori $P(H_1 \mid y)$ e $P(H_2 \mid y)$. Per trovare le probabilità a posteriori applichiamo il teorema di Bayes:

$$\begin{aligned} P(H_1 \mid y) &= \frac{P(y \mid H_1)P(H_1)}{P(y)} \\ &= \frac{P(y \mid H_1)P(H_1)}{P(y \mid H_1)P(H_1) + P(y \mid H_2)P(H_2)}, \end{aligned}$$

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

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

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

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

$$\begin{aligned} P(y) &= P(y \mid H_1)P(H_1) + P(y \mid H_2)P(H_2) \\ &= 0.24 \cdot 0.75 + 0.24 \cdot 0.25 \\ &= 0.24. \end{aligned}$$

Le probabilità a posteriori diventano:

$$\begin{aligned} P(H_1 \mid y) &= \frac{P(y \mid H_1)P(H_1)}{P(y)} \\ &= \frac{0.24 \cdot 0.75}{0.24} \\ &= 0.75, \end{aligned}$$

$$\begin{aligned} P(H_2 \mid y) &= \frac{P(y \mid H_2)P(H_2)}{P(y)} \\ &= \frac{0.24 \cdot 0.25}{0.24} \\ &= 0.25. \end{aligned}$$

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

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

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

$$\begin{aligned} P(H_1 \mid y) &= \frac{P(y \mid H_1)P(H_1)}{P(y)} \\ &= \frac{0.096 \cdot 0.75}{0.096 \cdot 0.75 + 0.144 \cdot 0.25} \\ &= 0.667, \end{aligned}$$

$$\begin{aligned} P(H_2 \mid y) &= \frac{P(y \mid H_2)P(H_2)}{P(y)} \\ &= \frac{0.144 \cdot 0.25}{0.096 \cdot 0.75 + 0.144 \cdot 0.25} \\ &= 0.333. \end{aligned}$$

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

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

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

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

e le evidenze a favore di H_1 si riducono a 1.33. Se si ottenesse un altro esito croce in un sesto lancio della moneta, l'ipotesi $H2$ diventerebbe più probabile dell'ipotesi H_1.

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

F

Il teorema della probabilità assoluta

Esercizio F.1. Consideriamo un'urna che contiene 5 palline rosse e 2 palline verdi. Due palline vengono estratte, una dopo l'altra. Vogliamo sapere la probabilità dell'evento "la seconda pallina estratta è rossa".

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

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

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

G

Esponenziali e logaritmi

Potenze ad esponente reale

Per un qualsiasi numero razionale $\frac{m}{n}$ (in cui $n > 0$) si ha

$$a^{\frac{m}{n}} = \sqrt[n]{a^m}$$

per numeri a reali positivi.

Proprietà

Se a, b sono reali positivi ed x, y reali qualsiasi, si ha

- $a^0 = 1$ e $a^{-x} = \frac{1}{a^x}$,
- $a^x a^y = a^{x+y}$ e $\frac{a^x}{a^y} = a^{x-y}$,
- $a^x b^x = (ab)^x$ e $\frac{a^x}{b^x} = \left(\frac{a}{b}\right)^x$,
- $(a^x)^y = a^{xy}$.

G.1 Funzione esponenziale

Definizione G.1. La funzione esponenziale con base a è

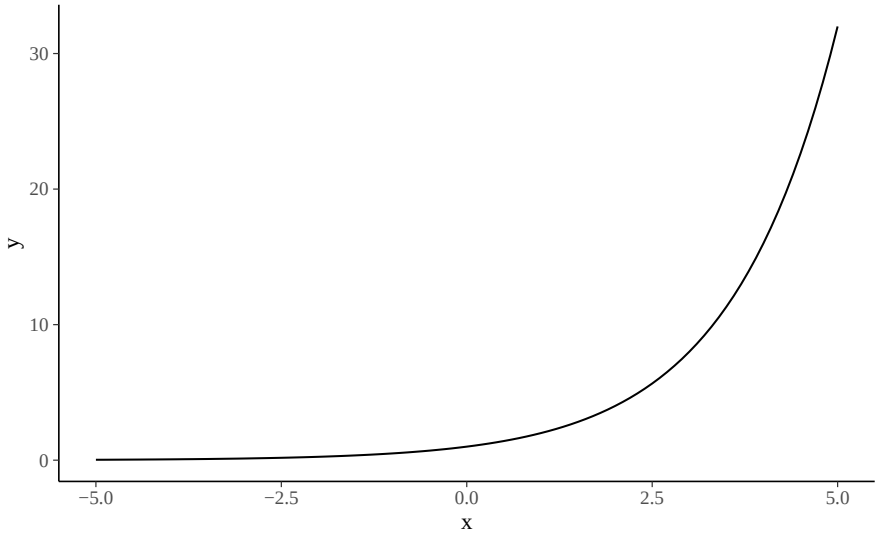
$$f(x) = a^x \tag{G.1}$$

dove $a > 0$, $a \neq 1$ e x è qualsiasi numero reale.

La base $a = 1$ è esclusa perché produce $f(x) = 1^x = 1$, la quale è una costante, non una funzione esponenziale.

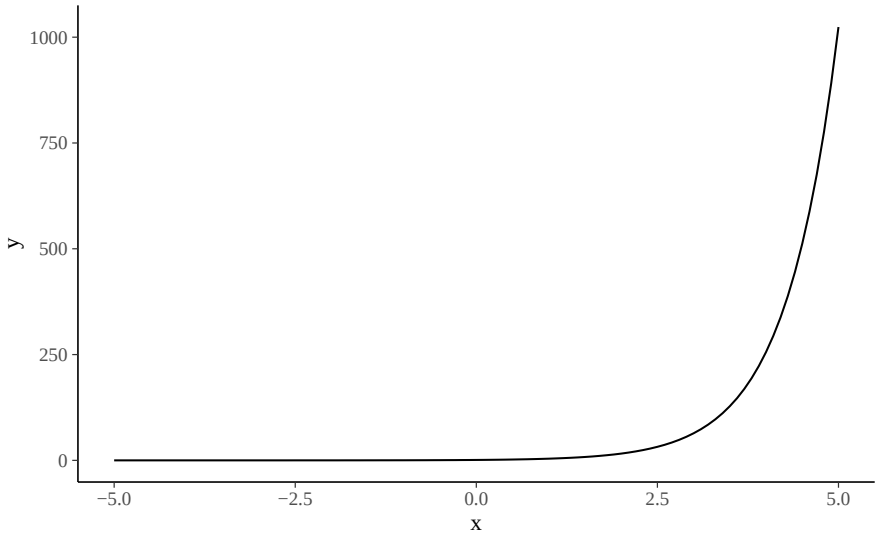
Per esempio, un grafico della funzione esponenziale di base 2 si trova con

```
exp_base2 = function(x){2^x}
tibble(x = c(-5, 5)) %>%
ggplot(aes(x = x)) +
  stat_function(fun = exp_base2)
```



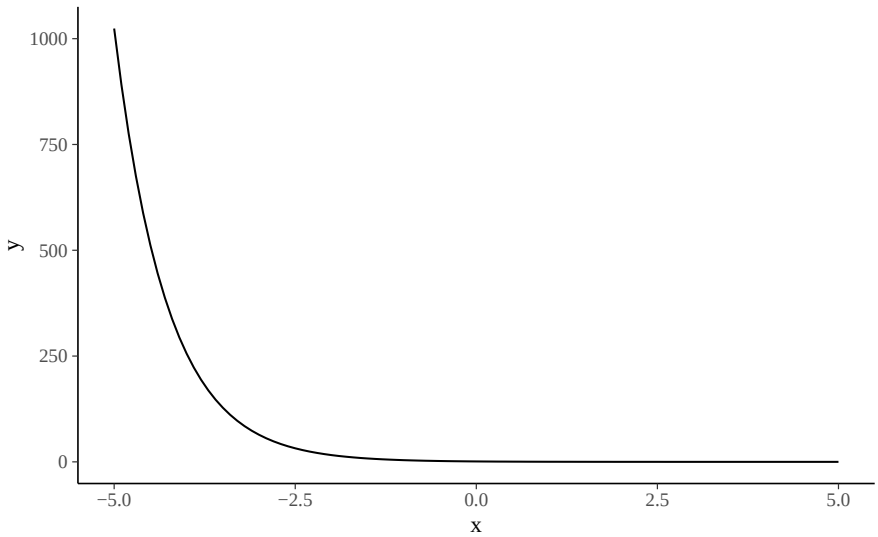
Se usiamo la base 4 troviamo

```
exp_base4 = function(x){4^x}
tibble(x = c(-5, 5)) %>%
ggplot(aes(x = x)) +
  stat_function(fun = exp_base4)
```



Oppure

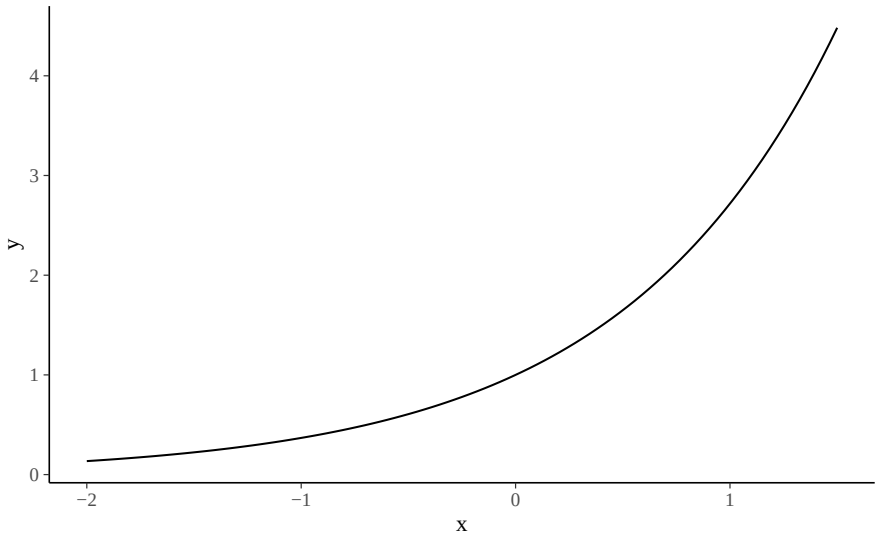
```
exp_base4 = function(x){4^-{x}}
tibble(x = c(-5, 5)) %>%
ggplot(aes(x = x)) +
  stat_function(fun = exp_base4)
```



In molte applicazioni la scelta più conveniente per la base è il numero irrazionale $e = 2.718281828\ldots$. Questo numero è chiamato la *base naturale*. La funzione $f(x) = e^x$ è chiamata *funzione esponenziale naturale*.

Per esempio, abbiamo

```
exp_base_e= function(x){exp(x)}
tibble(x = c(-2, 1.5)) %>%
ggplot(aes(x = x)) +
  stat_function(fun = exp_base_e)
```



Logaritmi

Dati due numeri reali $b > 0$ e $a > 0$ con $a \neq 1$, l'equazione esponenziale $a^x = b$ ammette sempre una ed una sola soluzione. Tale soluzione è detta *logaritmo in base a di b* ed è indicata con la scrittura $\log_a b$, dove b è detto *argomento* del logaritmo. In altri termini, per definizione si ha

$$x = \log_a b \iff a^x = b$$

dove deve essere $a > 0,\ a \neq 1,\ b > 0$.

Quando valutiamo i logaritmi, dobbiamo ricordare che un logaritmo è un esponente: il logaritmo in base a di b, $\log_a b$, è l'esponente da attribuire alla base a per ottenere l'argomento b. Le seguenti equazioni sono dunque equivalenti:

$$y = \log_a x \qquad x = a^y.$$

La prima equazione è in forma logaritmica e la seconda è in forma esponenziale. Ad esempio, l'equazione logaritmica $2 = \log_3 9$ può essere riscritta in forma esponenziale come $9 = 3^2$.

Esempio G.1. Scrivendo l'argomento come potenza della base si ottiene

- $\log_2 8 = \log_2 2^3 = 3$
- $\log_3 \sqrt[7]{3^{20}} = \log_3 3^{\frac{20}{7}} = \frac{20}{7}$
- $\log_{0.1} 0.01 = \log_{\frac{1}{10}} \frac{1}{100} = \log_{\frac{1}{10}} \left(\frac{1}{10}\right)^2 = 2$

Proprietà

Nell'operare con i logaritmi si procede spesso mediante le loro proprietà, che costituiscono una rilettura in termini di logaritmi delle proprietà delle potenze: se a, b sono numeri reali positivi diversi da 1 ed x, y reali positivi qualunque, allora

- $\log_a(xy) = \log_a x + \log_a y$,
- $\log_a \left(\frac{x}{y}\right) = \log_a x - \log_a y$,
- $\log_a (x^\alpha) = \alpha \log_a x, \quad \forall \alpha$ reale,
- $\log_a x = \frac{\log_b x}{\log_b a}$ (cambiamento di base).

Esempio G.2.

$$\begin{aligned} \log_a(x+1) - \log_a x - 2\log_a 2 &= \log_a(x+1) - (\log_a x + \log_a 2^2) \\ &= \log_a(x+1) - \log_a 4x \\ &= \log_a \frac{x+1}{4x}. \end{aligned}$$

G.2 Funzione logaritmica

La funzione logaritmica è la funzione inversa della funzione esponenziale.

Definizione G.2. Siano $a > 0$, $a \neq 1$. Per $x > 0$

$$y = \log_a x \quad \text{se e solo se } x = a^y. \tag{G.2}$$

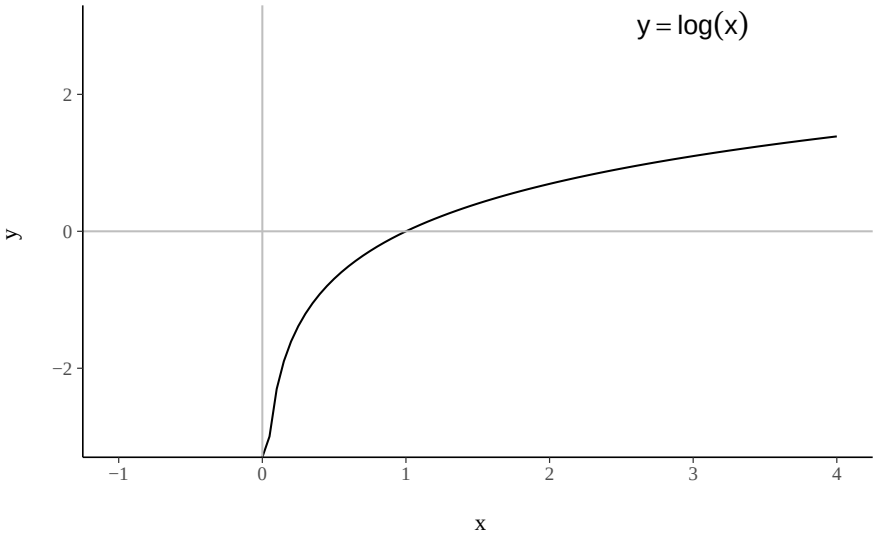
La funzione data da

$$f(x) = \log_a x \tag{G.3}$$

è chiamata funzione logaritmica.

Per esempio, abbiamo

```
log_funct <- function(x){
  log(x)
}
ggplot(tibble(x = c(-0.5, 4)), aes(x = x)) +
  stat_function(fun = log_funct) +
  xlim(c(-1, 4)) +
  ylim(-3, 3) +
  labs(x = "\n x", y = "y \n") +
  annotate("text", x = 3, y = 3, parse = TRUE, size = 5, fontface = "bold",
           label="y == log(x)") +
  geom_hline(yintercept = 0, colour = "gray") +
  geom_vline(xintercept = 0, colour = "gray")
```


y = log(x)
2
0
−2
y
−1
0
1
2
3
4
x

H

La Normale motivata dal metodo dei minimi quadrati

La distribuzione Normale fu scoperta da Gauss nel 1809 e, nella derivazione di Gauss, è intimamente legata al metodo dei minimi quadrati. Vediamo come Gauss arrivò alla definizione della densità Normale.

Tra il 1735 e il 1754 l'Accademia di Francia effettuò quattro misurazioni della lunghezza di un arco di meridiano a latitudini diverse con lo scopo di determinare la figura della Terra.[1] Papa Benedetto XIV volle contribuire a questo progetto e nel 1750 incaricò Roger Joseph Boscovich (1711—1787) e il gesuita inglese Christopher Maire di misurare un arco di meridiana nei pressi di Roma e contemporaneamente di costruire una nuova mappa dello Stato Pontificio. Il loro rapporto fu pubblicato nel 1755.

La relazione tra lunghezza d'arco e latitudine per archi piccoli è approssimativamente $y = \alpha + \beta x$, dove y è la lunghezza dell'arco e $x = sin^2 L$, dove L è la latitudine del punto medio dell'arco. Il problema di Boscovich era quello di stimare α e β da cinque osservazioni di (x, y).

Nel 1757 pubblicò una sintesi del rapporto del 1755 in cui proponeva di risolvere il problema di riconciliare le relazioni lineari inconsistenti mediante la minimizzazione della somma dei valori assoluti dei residui, sotto il vincolo che la somma dei residui fosse uguale a zero. In altre parole, Boscovich propose di minimizzare la quantità $\sum \mid y_i - a - bx_i \mid$ rispetto ad a e b sotto il vincolo $(y_i - a - bx_i) = 0$. Boscovich fu il primo a formulare un metodo per adattare una retta ai dati descritti da un diagramma a dispersione, laddove l'orientamento della retta dipende dalla minimizzazione di una funzione dei residui. La formulazione e la soluzione di Boscovich erano puramente verbali ed era accompagnata da un diagramma che spiegava il metodo di minimizzazione.

[1] L'espressione "figura della Terra" è utilizzata in geodesia per indicare la precisione con cui sono definite la dimensione e la forma della Terra.

Nella *Mécanique Céleste*, Laplace (1749, 1827) ritornò sul problema di Boscovich e mostrò in maniera formale come sia possibile minimizzare la quantità $\sum w_i \mid y_i - a - bx_i \mid$. Il metodo della minimizzazione del valore assoluto degli scarti presentava degli svantaggi rispetto al metodo dei minimi quadrati: (1) la stima della pendenza della retta era complicata da calcolare e (2) il metodo era limitato a una sola variabile indipendente. Il metodo scomparve quindi dalla pratica statistica fino alla seconda metà del XX secolo quando venne riproposto nel contesto della discussione della robustezza delle stime.

In seguito, tale problema venne ripreso da Legendre. Il suo *Nouvelle methods pour la determinazione des orbites des comètes* contiene un'appendice (pp. 72-80) intitolata *Sur la méthode des moindres carrés*, in cui per la prima volta il metodo dei minimi quadrati viene presentato come un metodo algebrico per l'adattamento di un modello lineare ai dati. Legendre scrive *"Tra tutti i princìpi che si possono proporre a questo scopo, credo che non ce ne sia uno più generale, più esatto e più facile da applicare di quello di cui ci siamo serviti nelle precedenti ricerche, e che consiste nel minimizzare la somma dei quadrati degli errori. In questo modo si stabilisce una sorta di equilibrio tra gli errori, che impedisce agli estremi di prevalere e ben si presta a farci conoscere lo stato del sistema più vicino alla verità."*

La somma dei quadrati degli errori è

$$\sum_{i=1}^{n} e_i^2 = (y_i - a - b_1 x_{i1} - \cdots - b_m x_{im})^2.$$

Per trovare il minimo di tale funzione, Legendre pone a zero le derivate della funzione rispetto ad $a, b_1, \ldots, b_m$, il che conduce a quelle che in seguito sono state chiamate le "equazioni normali". Risolvendo il sistema di equazioni normali rispetto $a, b_1, \ldots, b_m$, si determinano le stime dei minimi quadrati dei parametri del modello di regressione.

Tutto questo è rilevante per la derivazione della Normale perché, in questo contesto, Legendre osservò che la media aritmetica, quale caso speciale dei minimi quadrati, si ottiene minimizzando $\sum (y_i - b)^2$. In precedenza, Laplace si era posto il problema di mostrare che la media aritmetica è la migliore stima possibile della tendenza centrale di una distribuzione di errori di misurazione, ma non ci era riuscito perché aveva minimizzato il valore assoluto degli scarti, il che portava ad identificare la mediana

quale migliore stimatore della tendenza centrale della distribuzione degli errori, non la media.

Nel 1809, Gauss riformulò il problema ponendosi le seguenti domande. Che forma deve avere la densità della distribuzione degli errori? Quale quantità deve essere minimizzata per fare in modo che la media aritmetica risulti la miglior stima possibile della tendenza centrale della distribuzione degli errori? *"Si è soliti considerare come un assioma l'ipotesi che se una qualsiasi grandezza è stata determinata da più osservazioni dirette, fatte nelle stesse circostanze e con uguale cura, la media aritmetica dei valori osservati dà il valore più probabile, se non rigorosamente, eppure con una grade approssimazione, così che è sempre più sicuro utilizzare tale valore."*

Basandosi sul risultato di Legendre (ovvero, che è necessario minimizzare il quadrato degli scarti dalla tendenza centrale, non il valore assoluto degli scarti), Gauss derivò la formula della densità Normale quale modello teorico della distribuzione degli errori di misurazione. La Normale ha infatti la proprietà desiderata: il valore atteso della distribuzione corrisponde alla media aritmetica.

La scoperta della distribuzione normale segna l'inizio di una nuova era nella statistica. La distribuzione Normale è importante, in primo luogo, perché molti fenomeni naturali hanno approssimativamente le caratteristiche descritte dall'esempio precedente. In secondo luogo, è importante perché molti modelli statistici assumono che il fenomeno aleatorio di interesse abbia una distribuzione Normale.

Nella derivazione della Normale, Gauss fornì una giustificazione probabilistica al metodo dei minimi quadrati basata sull'ipotesi che le osservazioni siano distribuite normalmente e che la distribuzione a priori del parametro di tendenza centrale sia uniforme. Si noti come la discussione sia formulata in termini bayesiani.

La derivazione formale della Normale è troppo complessa per gli scopi presenti. Il Paragrafo **??** illustra invece come si possa giungere alla Normale mediante una simulazione. La motivazione del presente escursus storico è stata quella di mostrare come la Normale sia fortemente legata, in un contesto storico, al modello lineare e al metodo dei minimi quadrati.

I

La stima di massima verosimiglianza

I.1 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 I.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 $\hat{\theta}$:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \mathcal{L}(\theta). \tag{I.1}$$

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. In base all'approccio bayesiano, invece, il valore più credibile del parametro sconosciuto θ, anziché alla s.m.v., corrisponde invece alla 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 $p(\theta)$.

I.2 La 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 $\widehat{\theta}$) 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.:

$$\widehat{\theta} = \frac{y}{n}, \tag{I.2}$$

ovvero la frequenza relativa dei successi nel campione. □

Calcolo numerico

In maniera più semplice, il risultato descritto nel Paragrafo I.2 può essere ottenuto mediante una simulazione in R. Iniziamo a definire un insieme di valori possibili per il parametro incognito θ:

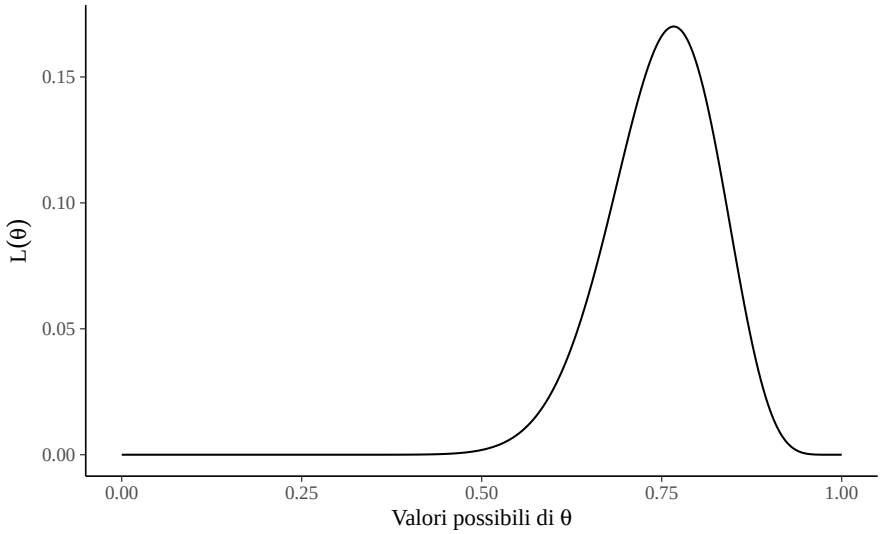
```
theta <- seq(0, 1, length.out = 1e3)
```

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 = 23, size = 30, prob = theta)
```

Si noti che, nell'istruzione precedente, abbiamo passato alla funzione `dbinom()` i dati, ovvero `x = 23` successi in `size = 30` 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] 0.7668
```

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

I.3 La s.m.v. 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.

Dimostrazione. 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 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:

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

laddove la funzione $f(\cdot)$ è la (I.3). Tenendo costanti i dati y, la funzione di verosimiglianza è:

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

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 (I.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

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

e per σ abbiamo

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

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. □

Calcolo numerico

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] 6.607
```

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 (I.3)). 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:

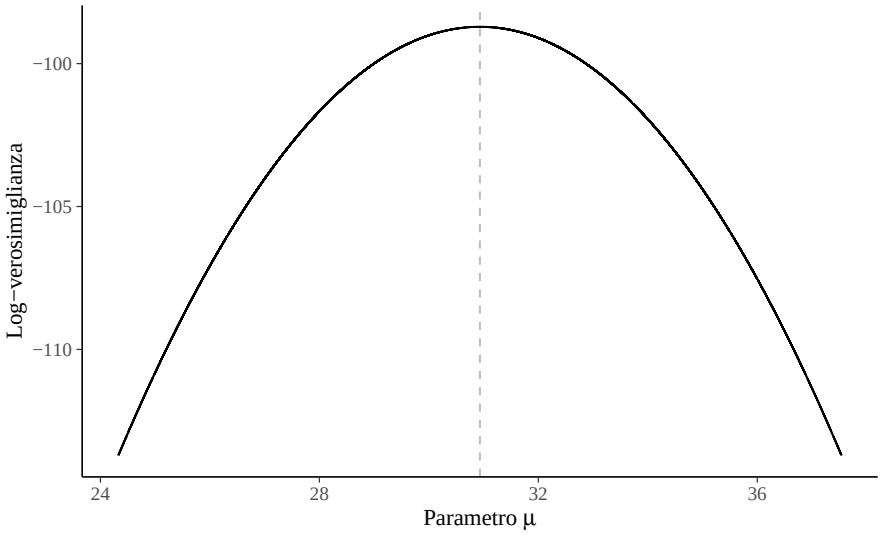
```
nrep <- 1e5
mu <- seq(
  mean(d$y) - sd(d$y),
  mean(d$y) + sd(d$y),
  length.out = nrep
)

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 (I.5).

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.

J

Le aspettative future dei pazienti depressi

J.1 La ricerca di Zetsche et al. (2019)

Per descrivere vari aspetti dell'analisi bayesiana utilizzeremo dei dati reali, nello specifico quelli raccolti da 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 ne-

gativamente". Ci poniamo il problema di ottenere una stima a posteriori di θ avendo osservato 23 "successi" in 30 prove.

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.

K

Modello Beta-binomiale

K.1 Funzione per il modello Beta-binomiale

La seguente funzione può essere usata per rappresentare la distribuzione a priori, la distribuzione a posteriori e la versosimiglianza (normalizzata) nel caso del modello Beta-binomiale. I parametri in input sono, nell'ordine, i parametri α e β della distribuzione a priori Beta, y (numero di successi) e n (numero di prove).

```
plot_beta_bin <- function(a, b, y, n) {
  library("tidyverse")

  df1 <- data.frame(theta = seq(0.001, 1, 0.001))
  prior_un <- dbeta(df1$theta, a, b)
  df1$prior <- prior_un / sum(prior_un)

  # Likelihood
  like_un <- dbinom(y, n, prob = seq(0.001, 1, 0.001))
  df1$like <- like_un / sum(like_un)

  # Posterior
  post_un <- df1$prior * df1$like
  df1$post <- post_un / sum(post_un)

  df2 <- df1 %>%
    pivot_longer(!theta, names_to = "grp", values_to = "val")

  df2$grp <- factor(df2$grp)
  # levels(df2$grp)
  df2$grp <- factor(df2$grp, levels = c("prior", "like", "post"))
```

```
  levels(df2$grp) <-
    c(
      "Distribuzione a priori", "Verosimiglianza",
      "Distribuzione a posteriori"
    )

  p <- ggplot(data = df2) +
    geom_line(aes(theta, val)) +
    facet_wrap(~grp, ncol = 1, scales = "free_y") +
    coord_cartesian(xlim = c(0, 1)) +
    scale_y_continuous(breaks = NULL) +
    labs(x = "", y = "")

  p
}
```

L

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.

M

Verosimiglianza marginale

M.1 Derivazione analitica della costante di normalizzazione

Riportiamo di seguito la derivazione analitica per la costante di normalizzazione discussa nella Sezione 1.4, ovvero dell'integrale (1.6).

Dimostrazione. Sia la distribuzione a priori $\theta \sim \text{Beta}(a, b)$ e sia $y = \{y_1, \dots, y_n\} \sim \text{Binomial}(\theta, n)$. Scrivendo la *funzione beta* come

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

la verosimiglianza marginale diventa

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

in quanto

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

In conclusione, nel caso di una verosimiglianza binomiale $y \mathrel{=\!\sim} \text{Binomial}(\theta, n)$ e di una distribuzione a priori $\theta \sim \text{Beta}(a, b)$, la verosimiglianza marginale diventa uguale alla (M.1). □

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

Per replicare mediante la (M.1) il risultato trovato per via numerica nella Sezione 1.4 assumiamo una distribuzione a priori uniforme, ovvero $\text{Beta}(1, 1)$. 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] 0.03226
```

N

Aspettative degli individui depressi

Per fare pratica, applichiamo il metodo basato su griglia ad un campione di dati reali. Zetsche et al. (2019) 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 usando il metodo basato su griglia.

N.1 La griglia

Fissiamo una griglia di $n = 50$ valori equispaziati nell'intervallo [0, 1] per il parametro θ:

```
n_points <- 50
p_grid <- seq(from = 0, to = 1, length.out = n_points)
p_grid
#>  [1] 0.00000 0.02041 0.04082 0.06122 0.08163 0.10204
#>  [7] 0.12245 0.14286 0.16327 0.18367 0.20408 0.22449
#> [13] 0.24490 0.26531 0.28571 0.30612 0.32653 0.34694
#> [19] 0.36735 0.38776 0.40816 0.42857 0.44898 0.46939
#> [25] 0.48980 0.51020 0.53061 0.55102 0.57143 0.59184
#> [31] 0.61224 0.63265 0.65306 0.67347 0.69388 0.71429
#> [37] 0.73469 0.75510 0.77551 0.79592 0.81633 0.83673
#> [43] 0.85714 0.87755 0.89796 0.91837 0.93878 0.95918
#> [49] 0.97959 1.00000
```

N.2 Distribuzione a priori

Supponiamo di avere scarse credenze a priori sulla tendenza di un individuo clinicamente depresso a manifestare delle aspettative distorte negativamente circa il suo umore futuro. Imponiamo quindi una distribuzione non informativa sulla distribuzione a priori di θ — ovvero, una distribuzione uniforme nell'intervallo [0, 1]. Dato che consideriamo soltanto $n = 50$ valori possibili per il parametro θ, creiamo un vettore di 50 elementi che conterrà i valori della distribuzione a priori scalando ciascun valore del vettore per n in modo tale che la somma di tutti i valori sia uguale a 1.0:

```
prior1 <- dbeta(p_grid, 1, 1) / sum(dbeta(p_grid, 1, 1))
prior1
#>  [1] 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
#> [11] 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
```

```
#> [21] 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
#> [31] 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
#> [41] 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
```

Verifichiamo:

```
sum(prior1)
#> [1] 1
```

La distribuzione a priori così costruita è rappresentata nella figura N.1.

```
p1 <- data.frame(p_grid, prior1) %>%
  ggplot(aes(x=p_grid, xend=p_grid, y=0, yend=prior1)) +
  geom_line() +
  geom_segment() +
  ylim(0, 0.17) +
  labs(
    x = "Parametro \U03B8",
    y = "Probabilità a priori",
    title = "50 punti"
  )
p1
```

N.3 Funzione di verosimiglianza

Calcoliamo ora la funzione di verosimiglianza utilizzando i 50 valori θ definiti in precedenza. A ciascuno dei valori della griglia applichiamo la formula binomiale, tendendo costanti i dati (ovvero 23 "successi" in 30 prove). Ad esempio, in corrispondenza del valore $\theta = 0.816$, l'ordinata della funzione di verosimiglianza diventa

$$\binom{30}{23} \cdot 0.816^{23} \cdot (1 - 0.816)^7 = 0.135.$$

Per $\theta = 0.837$, l'ordinata della funzione di verosimiglianza sarà

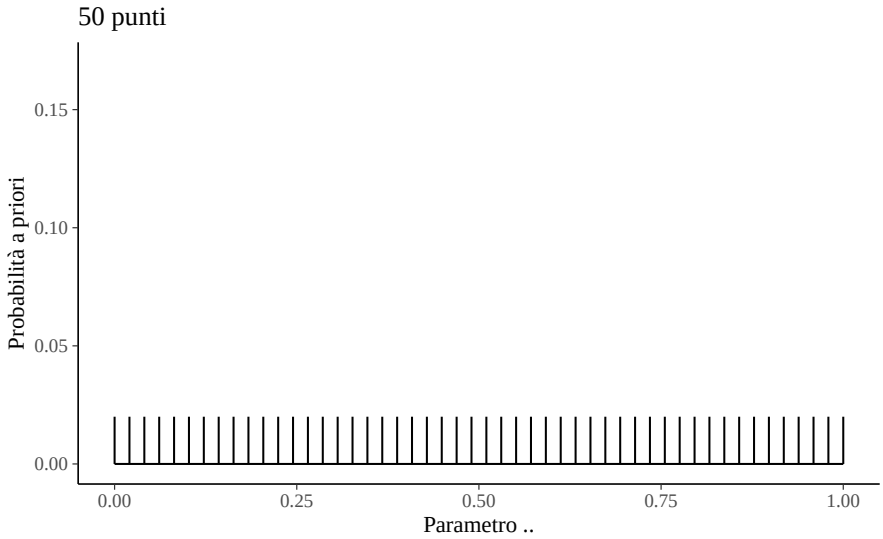


Figura N.1: *Rappresentazione grafica della distribuzione a priori per il parametro heta, ovvero la probabilità di aspettative future distorte negativamente.*

$$\binom{30}{23} \cdot 0.837^{23} \cdot (1 - 0.837)^{7} = 0.104.$$

Dobbiamo svolgere questo calcolo per tutti gli elementi della griglia. Usando R, tale risultato si trova nel modo seguente:

```
likelihood <- dbinom(x = 23, size = 30, prob = p_grid)
likelihood
#>  [1] 0.000e+00 2.353e-33 1.703e-26 1.644e-22 1.054e-19
#>  [6] 1.525e-17 8.602e-16 2.528e-14 4.607e-13 5.819e-12
#> [11] 5.499e-11 4.106e-10 2.520e-09 1.311e-08 5.919e-08
#> [16] 2.362e-07 8.457e-07 2.749e-06 8.197e-06 2.260e-05
#> [21] 5.799e-05 1.393e-04 3.149e-04 6.721e-04 1.359e-03
#> [26] 2.612e-03 4.779e-03 8.340e-03 1.390e-02 2.214e-02
#> [31] 3.372e-02 4.910e-02 6.830e-02 9.068e-02 1.147e-01
#> [36] 1.378e-01 1.568e-01 1.682e-01 1.689e-01 1.575e-01
#> [41] 1.349e-01 1.044e-01 7.133e-02 4.166e-02 1.973e-02
#> [46] 6.937e-03 1.535e-03 1.473e-04 1.868e-06 0.000e+00
```

La funzione `dbinom(x, size, prob)` richiede che vengano specificati tre parametri: il numero di "successi", il numero di prove e la probabilità di successo. Nella chiamata precedente, `x` (numero di successi) e `size` (numero di prove bernoulliane) sono degli scalari e `prob` è il vettore `p_grid`. In tali circostanze, l'output di `dbinom()` è il vettore che abbiamo chiamato `likelihood`. Gli elementi di tale vettore sono stati calcolati applicando la formula della distribuzione binomiale a ciascuno dei 50 elementi della griglia, tenendo sempre costanti i dati [ovvero, `x` (il numero di successi) e `size` (numero di prove bernoulliane)]; ciò che varia è il valore `prob`, che assume valori diversi (`p_grid`) in ciascuna cella della griglia.

La chiamata a `dbinom()` produce dunque un vettore i cui valori corrispondono all'ordinata della funzione di verosimiglianza per per ciascun valore θ specificato in `p_grid`. La verosimiglianza discretizzata così ottenuta è riportata nella figura N.2.

```
p2 <- data.frame(p_grid, likelihood) %>%
  ggplot(aes(x=p_grid, xend=p_grid, y=0, yend=likelihood)) +
  geom_segment() +
  ylim(0, 0.17) +
  labs(
    x = "Parametro \U03B8",
    y = "Verosimiglianza"
  )
p2
```

N.4 Distribuzione a posteriori

L'approssimazione discretizzata della distribuzione a posteriori $p(\theta \mid y)$ si ottiene facendo il prodotto della verosimiglianza e della distribuzione a priori per poi scalare tale prodotto per una costante di normalizzazione. Il prodotto $p(\theta)\mathcal{L}(y \mid \theta)$ produce la distribuzione a posteriori *non standardizzata*.

Nel caso di una distribuzione a priori non informativa (ovvero una distribuzione uniforme), per ottenere la funzione a posteriori non standardizzata è sufficiente moltiplicare ciascun valore della funzione di vero-

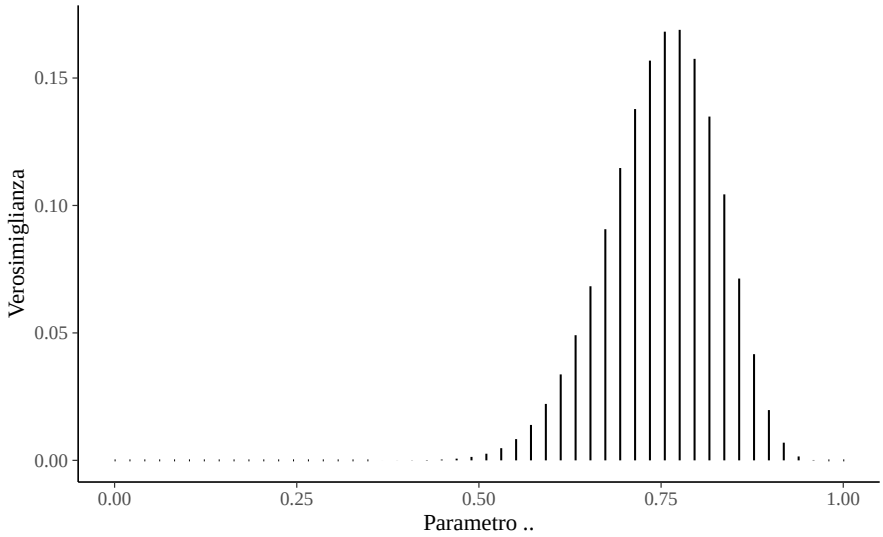


Figura N.2: *Rappresentazione della funzione di verosimiglianza per il parametro θ, ovvero la probabilità di aspettative future distorte negativamente.*

simiglianza per 0.02. Per esempio, per il primo valore della funzione di verosimiglianza usato quale esempio poco sopra, abbiamo $0.135 \cdot 0.02$; per il secondo valore dell'esempio abbiamo $0.104 \cdot 0.02$; e così via. Possiamo svolgere tutti i calcoli usando R nel modo seguente:[1]

```
unstd_posterior <- likelihood * prior1
unstd_posterior
#>  [1] 0.000e+00 4.705e-35 3.406e-28 3.288e-24 2.107e-21
#>  [6] 3.050e-19 1.720e-17 5.057e-16 9.214e-15 1.164e-13
#> [11] 1.100e-12 8.211e-12 5.040e-11 2.622e-10 1.184e-09
#> [16] 4.724e-09 1.691e-08 5.499e-08 1.639e-07 4.519e-07
#> [21] 1.160e-06 2.786e-06 6.297e-06 1.344e-05 2.718e-05
#> [26] 5.224e-05 9.558e-05 1.668e-04 2.780e-04 4.428e-04
#> [31] 6.744e-04 9.820e-04 1.366e-03 1.814e-03 2.294e-03
#> [36] 2.756e-03 3.136e-03 3.363e-03 3.378e-03 3.150e-03
```

[1] Ricordiamo il principio dell'aritmetica vettorializzata: i vettori `likelihood` e `prior1` sono entrambi costituiti da 50 elementi. Se facciamo il prodotto tra i due vettori otteniamo un vettore di 50 elementi, ciascuno dei quali uguale al prodotto dei corrispondenti elementi dei vettori `likelihood` e `prior1`.

```
#> [41] 2.697e-03 2.087e-03 1.427e-03 8.331e-04 3.945e-04
#> [46] 1.387e-04 3.070e-05 2.947e-06 3.736e-08 0.000e+00
```

Avendo calcolato i valori della funzione a posteriori non standardizzata è poi necessario dividere per una costante di normalizzazione. Nel caso discreto, trovare il denominatore del teorema di Bayes è facile: esso è uguale alla somma di tutti i valori della distribuzione a posteriori non normalizzata. Per i dati presenti, tale costante di normalizzazione è uguale a 0.032:

```
sum(unstd_posterior)
#> [1] 0.03161
```

La standardizzazione dei due valori usati come esempio è data da: $0.135 \cdot 0.02/0.032$ e da $0.104 \cdot 0.02/0.032$. Usiamo R per svolgere questo calcolo su tutti i 50 valori di `unstd_posterior` così che la somma dei 50 i valori di `posterior` sia uguale a 1.0:

```
posterior <- unstd_posterior / sum(unstd_posterior)
posterior
#>  [1] 0.000e+00 1.488e-33 1.077e-26 1.040e-22 6.666e-20
#>  [6] 9.649e-18 5.442e-16 1.600e-14 2.915e-13 3.681e-12
#> [11] 3.479e-11 2.597e-10 1.594e-09 8.295e-09 3.745e-08
#> [16] 1.494e-07 5.350e-07 1.739e-06 5.186e-06 1.430e-05
#> [21] 3.669e-05 8.814e-05 1.992e-04 4.252e-04 8.599e-04
#> [26] 1.652e-03 3.023e-03 5.276e-03 8.794e-03 1.401e-02
#> [31] 2.133e-02 3.106e-02 4.321e-02 5.737e-02 7.256e-02
#> [36] 8.719e-02 9.922e-02 1.064e-01 1.069e-01 9.966e-02
#> [41] 8.533e-02 6.602e-02 4.513e-02 2.635e-02 1.248e-02
#> [46] 4.389e-03 9.712e-04 9.321e-05 1.182e-06 0.000e+00
```

Verifichiamo:

```
sum(posterior)
#> [1] 1
```

La distribuzione a posteriori così trovata non è altro che la versione normalizzata della funzione di verosimiglianza: questo avviene perché

la distribuzione a priori uniforme non ha aggiunto altre informazioni oltre a quelle che erano già fornite dalla funzione di verosimiglianza. L'approssimazione discretizzata di $p(\theta \mid y)$ che abbiamo appena trovato è riportata nella figura N.3.

```
p3 <- data.frame(p_grid, posterior) %>%
  ggplot(aes(x=p_grid, xend=p_grid, y=0, yend=posterior)) +
  geom_segment() +
  ylim(0, 0.17) +
  labs(
    x = "Parametro \U03B8",
    y = "Probabilità a posteriori"
  )
p3
```

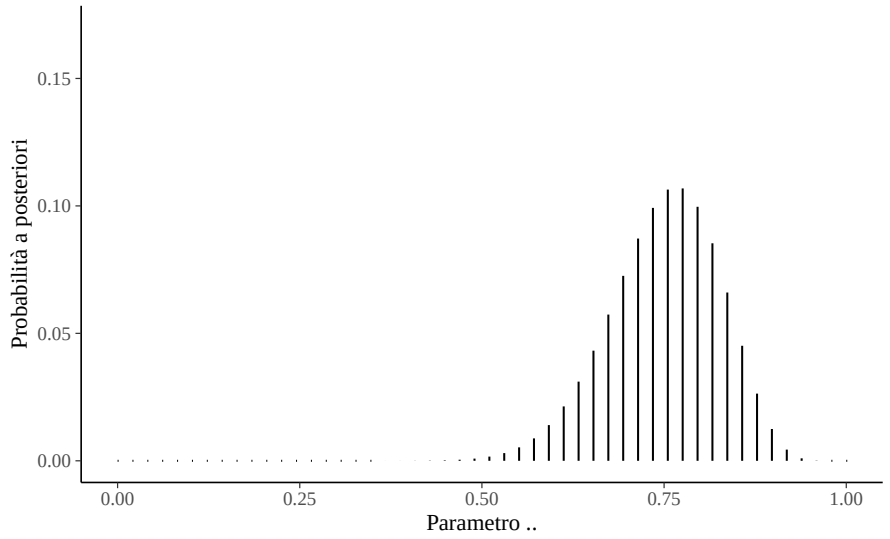


Figura N.3: *Rappresentazione della distribuzione a posteriori per il parametro θ, ovvero la probabilità di aspettative future distorte negativamente.*

I grafici delle figure N.1, N.2 e N.3 sono state calcolati utilizzando una griglia di 50 valori equi-spaziati per il parametro θ. I segmenti verticali rappresentano l'intensità della funzione in corrispondenza di ciascuna modalità parametro θ. Nella figura N.1 e nella figura N.3 la somma delle lunghezze dei segmenti verticali è uguale a 1.0; ciò non si verifica, invece,

nel caso della figura N.3 (la funzione di verosimiglianza non è mai una funzione di probabilità, né nel caso discreto né in quello continuo).

N.5 La stima della distribuzione a posteriori (versione 2)

Continuiamo l'analisi di questi dati esaminiamo l'impatto di una distribuzione a priori informativa sulla distribuzione a posteriori. Una distribuzione a priori informativa riflette un alto grado di certezza a priori sui valori dei parametri del modello. Un ricercatore utilizza una distribuzione a priori informativa per introdurre nel processo di stima informazioni pre-esistenti alla raccolta dei dati, introducendo così delle restrizioni sulla possibile gamma di valori del parametro.

Nel caso presente, supponiamo che la letteratura psicologica fornisca delle informazioni su θ (la probabilità che le aspettative future di un individuo clinicamente depresso siano distorte negativamente). Per fare un esempio, supponiamo (irrealisticamente) che tali conoscenze pregresse possano essere rappresentate da una Beta di parametri $\alpha = 2$ e $\beta = 10$. Tali ipotetiche conoscenze pregresse ritengono molto plausibili valori θ bassi e considerano implausibili valori $\theta > 0.5$. Questo è equivalente a dire che ci aspettiamo che le aspettative relative all'umore futuro siano distorte negativamente solo per pochissimi individui clinicamente depressi — ovvero, ci aspettiamo che la maggioranza degli individui clinicamente depressi sia inguaribilmente ottimista. Questa è, ovviamente, una credenza a priori del tutto irrealistica. La esamino qui, non perché abbia alcun senso nel contesto dei dati di Zetsche et al. (2019), ma soltanto per fare un esempio nel quale risulta chiaro come la distribuzione a posteriori sia una sorta di "compromesso" tra la distribuzione a priori e la verosimiglianza.

Con calcoli del tutto simili a quelli descritti sopra si giunge alla distribuzione a posteriori rappresentata nella figura N.4. Useremo ora una griglia di 100 valori per il parametro θ:

```
n_points <- 100
p_grid <- seq(from = 0, to = 1, length.out = n_points)
```

Per la distribuzione a priori scegliamo una Beta(2, 10):

```
alpha <- 2
beta <- 10
prior2 <- dbeta(p_grid, alpha, beta) / sum(dbeta(p_grid, alpha, beta))
sum(prior2)
#> [1] 1
```

Tale distribuzione a priori è rappresentata nella figura N.4:

```
plot_df <- data.frame(p_grid, prior2)
p4 <- plot_df %>%
  ggplot(aes(x=p_grid, xend=p_grid, y=0, yend=prior2)) +
  geom_segment() +
  ylim(0, 0.17) +
  labs(
    x = "",
    y = "Probabilità a priori"
  )
p4
```

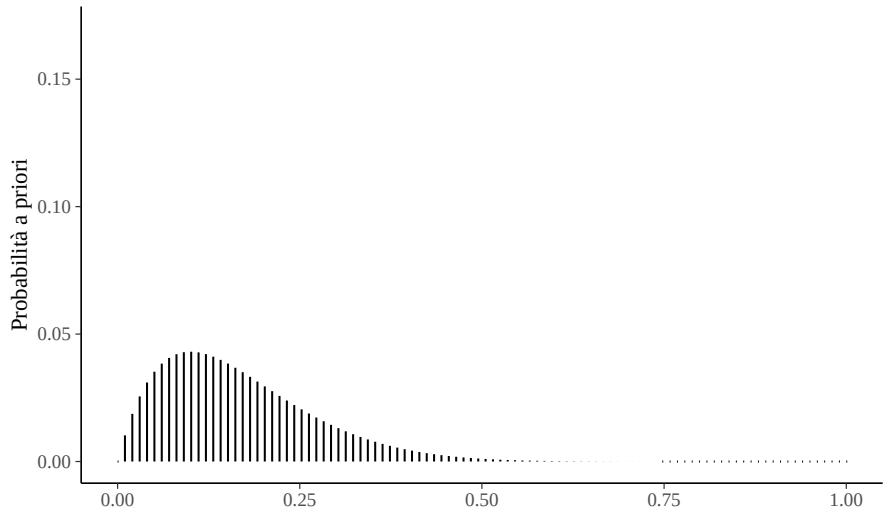


Figura N.4: *Rappresentazione di una funzione a priori informativa per il parametro θ.*

Calcoliamo il valore di verosimiglianza per ciascun punto della griglia:

```
likelihood <- dbinom(23, size = 30, prob = p_grid)
```

Per ciascun punto della griglia, il prodotto tra la verosimiglianza e distribuzione a priori è dato da:

```
unstd_posterior2 <- likelihood * prior2
```

È necessario normalizzare la distribuzione a posteriori discretizzata:

```
posterior2 <- unstd_posterior2 / sum(unstd_posterior2)
```

Verifichiamo:

```
sum(posterior2)
#> [1] 1
```

La nuova funzione a posteriori discretizzata è rappresentata nella figura N.5:

```
plot_df <- data.frame(p_grid, posterior2)
p5 <- plot_df %>%
  ggplot(aes(x = p_grid, xend = p_grid, y = 0, yend = posterior2)) +
  geom_segment() +
  ylim(0, 0.17) +
  labs(
    x = "Parametro \U03B8",
    y = "Probabilità a posteriori"
  )
p5
```

Facendo un confronto tra le figure N.4 e N.5 notiamo una notevole differenza tra la distribuzione a priori e la distribuzione a posteriori. In particolare, la distribuzione a posteriori risulta spostata verso destra su posizioni più vicine a quelle della verosimiglianza [figura N.2]. Si noti inoltre che, a causa dell'effetto della distribuzione a priori, le distribuzioni a posteriori delle figure N.3 e N.5 sono molto diverse tra loro.

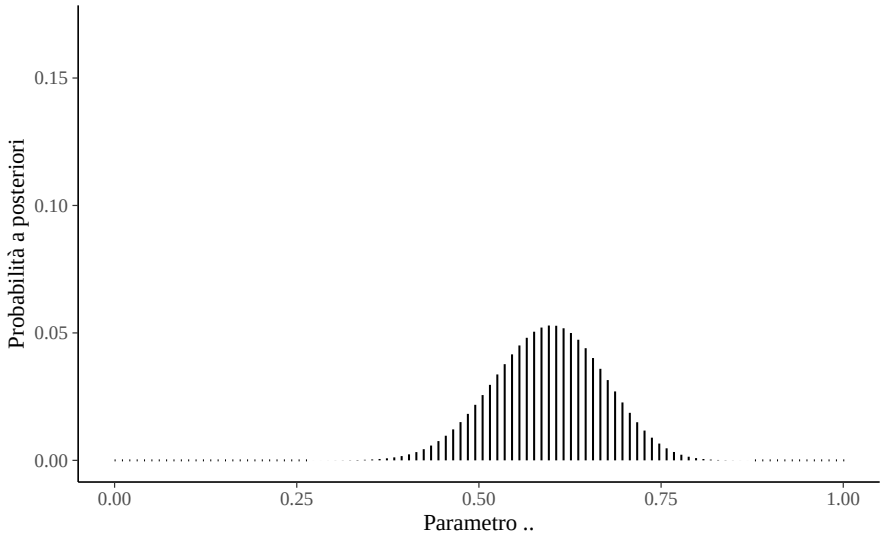


Figura N.5: *Rappresentazione della funzione a posteriori per il parametro θ calcolata utilizzando una distribuzione a priori informativa.*

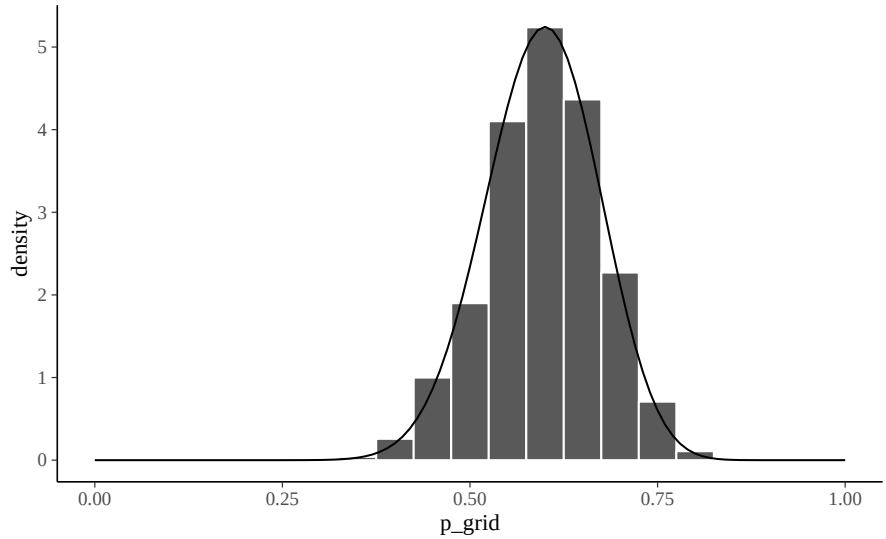
Campioniamo ora 10,000 punti dall'approssimazione discretizzata della distribuzione a posteriori:

```
# Set the seed
set.seed(84735)

df <- data.frame(
  p_grid,
  posterior2
)
# Step 4: sample from the discretized posterior
post_samples <- df %>%
  slice_sample(
  n = 1e5,
  weight_by = posterior2,
  replace = TRUE
)
```

Una rappresentazione grafica del campione casuale estratto dalla distribuzione a posteriori $p(\theta \mid y)$ è data da:

```
post_samples %>%
  ggplot(aes(x = p_grid)) +
  geom_histogram(
    aes(y = ..density..),
    color = "white",
    binwidth = 0.05
  ) +
  stat_function(fun = dbeta, args = list(25, 17)) +
  lims(x = c(0, 1))
```



All'istogramma è stata sovrapposta la corretta distribuzione a posteriori, ovvero una Beta di parametri 25 ($y + \alpha = 23 + 2$) e 17 ($n - y + \beta = 30$ - 23 + 10).

La stima della moda a posteriori si ottiene con

```
df$p_grid[which.max(df$posterior2)]
#> [1] 0.596
```

e corrisponde a

$$\text{Mo} = \frac{\alpha - 1}{\alpha + \beta - 2} = \frac{25 - 1}{25 + 17 - 2} = 0.6.$$

La stima della media a posteriori si ottiene con

```
mean(post_samples$p_grid)
#> [1] 0.5953
```

e corrisponde a

$$\bar{\theta} = \frac{\alpha}{\alpha + \beta} = \frac{25}{25 + 17} \approx 0.5952.$$

La stima della mediana a posteriori si ottiene con

```
median(post_samples$p_grid)
#> [1] 0.596
```

e corrisponde a

$$\text{Me} = \frac{\alpha - \frac{1}{3}}{\alpha + \beta - \frac{2}{3}} \approx 0.5968.$$

N.6 Versione 2

Possiamo semplificare i calcoli precedenti definendo le funzioni `likelihood()`, `prior()` e `posterior()`.

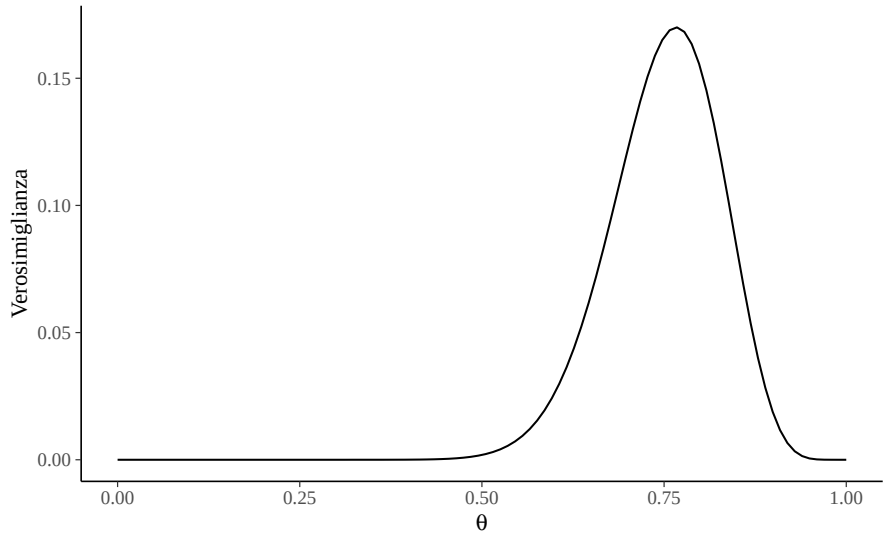
Per calcolare la funzione di verosimiglianza per i 30 valori di Zetsche et al. (2019) useremo la funzione `likelihood()`:

```
x <- 23
N <- 30
param <- seq(0, 1, length.out = 100)

likelihood <- function(param, x = 23, N = 30) {
  dbinom(x, N, param)
}

tibble(
```

```
  x = param,
  y = likelihood(param)
) %>%
  ggplot(aes(x, y)) +
  geom_line() +
  labs(
    x = expression(theta),
    y = "Verosimiglianza"
  )
```



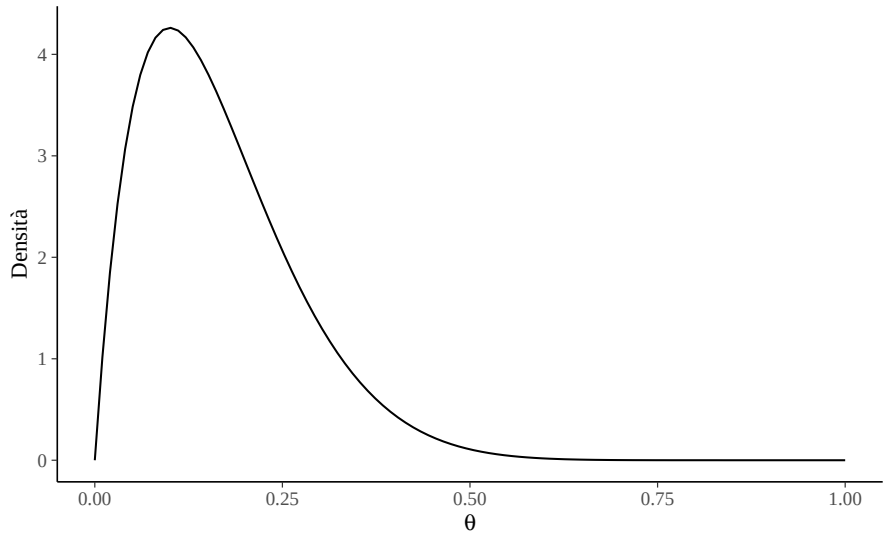
La funzione `likelihood()` ritorna l'ordinata della verosimiglianza binomiale per ciascun valore del vettore `param` in input.

Quale distribuzione a priori utilizzeremo una Beta$(2, 10)$ che è implementata nella funzione `prior()`:

```
prior <- function(param, alpha = 2, beta = 10) {
  param_vals <- seq(0, 1, length.out = 100)
  dbeta(param, alpha, beta) # / sum(dbeta(param_vals, alpha, beta))
}

tibble(
  x = param,
```

```
  y = prior(param)
) %>%
  ggplot(aes(x, y)) +
  geom_line() +
  labs(
    x = expression(theta),
    y = "Densità"
  )
```

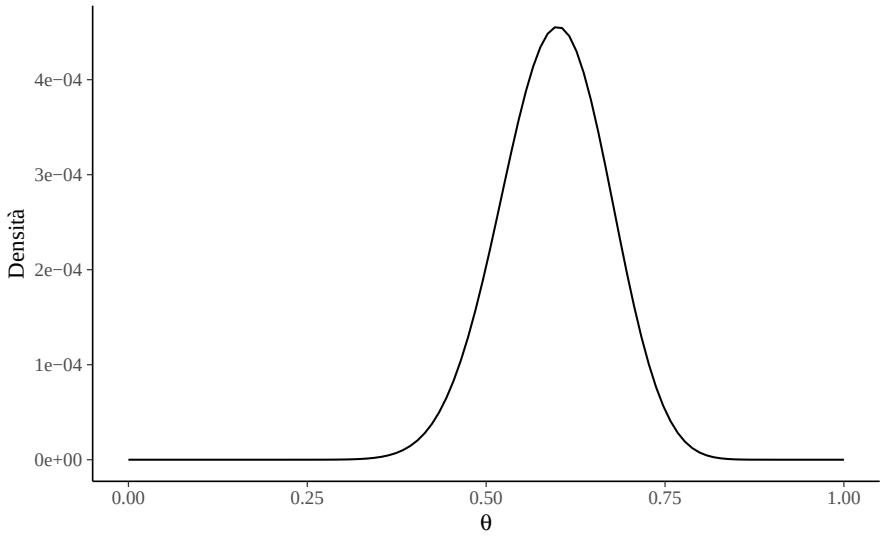


La funzione `posterior()` ritorna il prodotto della densità a priori e della verosimiglianza:

```
posterior <- function(param) {
  likelihood(param) * prior(param)
}

tibble(
  x = param,
  y = posterior(param)
) %>%
  ggplot(aes(x, y)) +
  geom_line() +
```

```
labs(
  x = expression(theta),
  y = "Densità"
)
```



La distribuzione a posteriori non normalizzata mostrata nella figura replica il risultato ottenuto con il codice utilizzato nella prima parte di questo Capitolo. Per l'implementazione dell'algoritmo di Metropolis non è necessaria la normalizzazione 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.