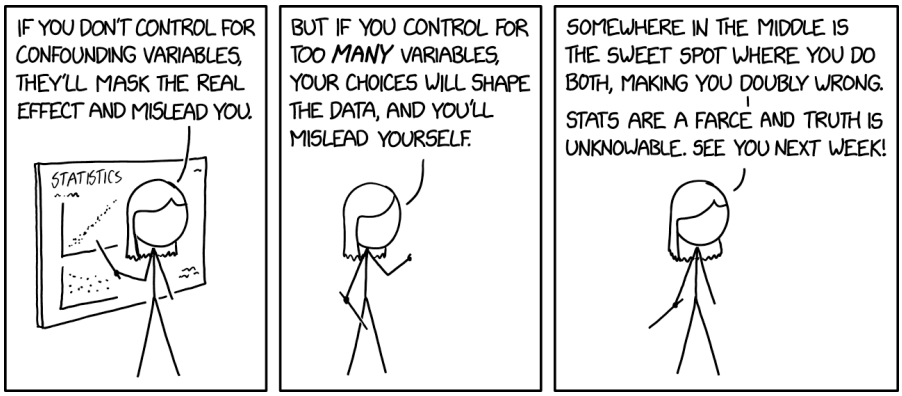
Corrado Caudek

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

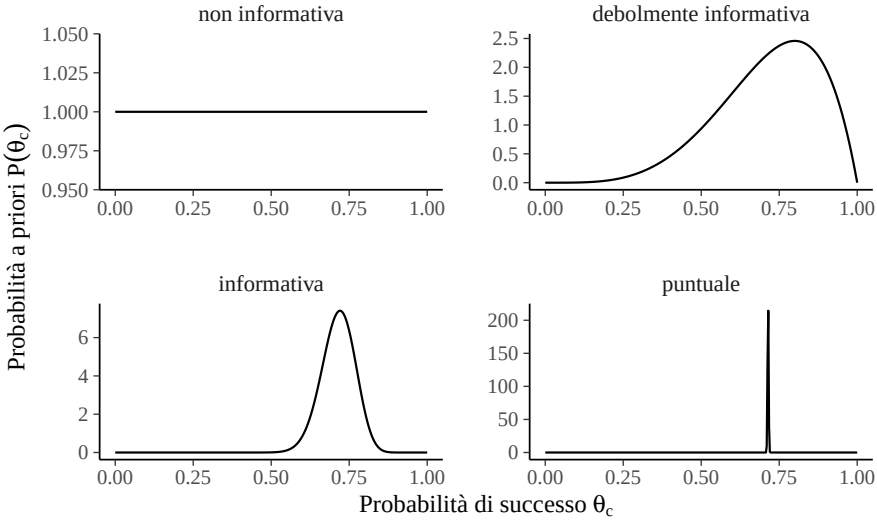


Figura 1.1: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

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 **??**.

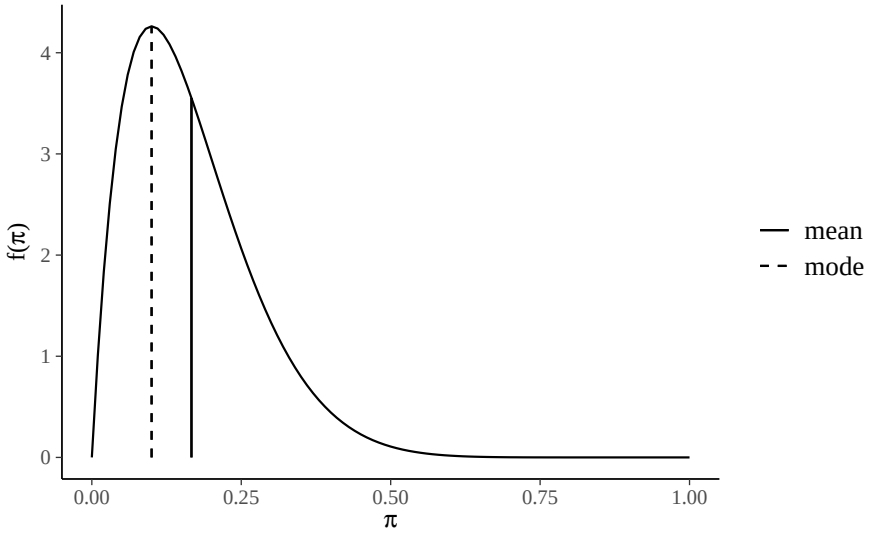
1.2.3 Un'applicazione empirica

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

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

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

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



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

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 **??**):

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

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'applicazione empirica

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

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

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)
  )
```

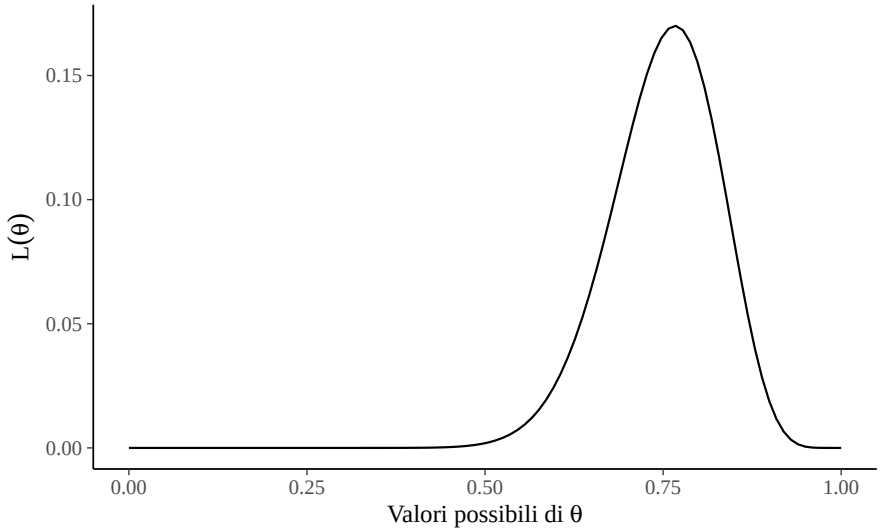


Figura 1.2: *Funzione di verosimiglianza nel caso di 23 successi in 30 prove.*

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'applicazione empirica

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

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

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 **??**.

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

La predizione bayesiana

Oltre ad una sintesi della distribuzione a posteriori attraverso il computo di indici caratteristici e alla verifica di ipotesi, un altro compito dell'analisi bayesiana è la predizione di nuovi dati futuri. Dopo aver osservato i dati di un campione e ottenuto le distribuzioni a posteriori dei parametri, è infatti possibile ottenere una qualche indicazione su come potrebbero essere i dati futuri. L'uso più immediato della stima della distribuzione dei possibili valori futuri della variabile di esito è la verifica del modello. Infatti, il modo più diretto per testare un modello è quello di utilizzare il modello per fare previsioni sui possibili dati futuri per poi confrontare i dati predetti con i dati effettivi. Questa pratica va sotto il nome di controllo predittivo a posteriori. In questo capitolo ci focalizzeremo sul problema della predizione bayesiana esaminando il caso più semplice, ovvero lo schema beta-binomiale. In seguito estenderemo questa discussione al caso generale.

2.1 La distribuzione predittiva

Una volta costruita la distribuzione a posteriori del parametro θ, potremmo essere interessati a utilizzare il nostro modello statistico allo scopo di prevedere la probabilità di risultati futuri basandosi sui dati storici. L'obiettivo è andare oltre la comprensione di cosa è successo per arrivare a una migliore valutazione di quello che accadrà in futuro. Questo tipo di analisi inferenziale va sotto il nome di *analisi predittiva*. L'analisi predittiva utilizza dunque i dati che sono già disponibili per sviluppare un modello che può essere utilizzato per prevedere valori di dati diversi o nuovi.

Consideriamo qui il caso beta-binomiale nel quale la distribuzione a priori per il parametro θ (probabilità di successo) è una distribuzione Beta,

la verosimiglianza è binomiale e i dati sono costituiti dal numero y di successi che è osservato in n prove Bernoulliane indipendenti. Nell'esempio, useremo un'altra volta i dati del campione di pazienti clinici depressi di Zetsche et al. (2019) – si veda l'Appendice **??**. Supponendo di volere esaminare in futuro altri m pazienti clinici, ci chiediamo: quanti di essi manifesteranno una depressione grave?

Siamo interessati a predire i risultati che si potrebbero osservare in nuovi campioni di m osservazioni. Denotiamo con $\tilde{y}$ la manifestazione della variabile casuale $\tilde{Y}$. In un nuovo campione di m osservazioni, $\tilde{y}$ può assumere il valore di $\tilde{y}_1$, in un altro campione il valore di $\tilde{y}_2$, e così via. Siamo interessati a descrivere la *distribuzione predittiva a posteriori* $p(\tilde{Y} = \tilde{y} \mid Y = y)$. Nel caso dell'esempio in discussione, la distribuzione di $\tilde{Y}$ dipende da θ e la nostra conoscenza corrente di θ è fornita dalla distribuzione a posteriori. Usando la regola della catena, possiamo scrivere la distribuzione congiunta di $\tilde{Y}$ e θ nel modo seguente

$$p(\tilde{Y} = \tilde{y}, \theta \mid Y = y) = p(\tilde{Y} = \tilde{y} \mid \theta) p(\theta \mid Y = y). \tag{2.1}$$

Integrando su θ otteniamo la distribuzione predittiva a posteriori:

$$p(\tilde{y} \mid y) = \int_\theta p(\tilde{y} \mid \theta) p(\theta \mid y) \, \mathrm{d}\theta. \tag{2.2}$$

Nel caso dello schema beta-binomiale, la funzione $p(\tilde{y} \mid \theta)$ è binomiale di parametri m e θ, e la distribuzione a posteriori $p(\theta \mid y)$ è $\text{Beta}(\alpha + y, \beta + n - y)$. Risolvendo l'integrale otteniamo:

$$\begin{aligned}
p(\tilde{y} \mid y) &= \int_0^1 p(\tilde{y} \mid \theta) p(\theta \mid y) \, \mathrm{d}\theta \\
&= \int_0^1 \binom{m}{\tilde{y}} \theta^{\tilde{y}} (1-\theta)^{m-\tilde{y}} \text{Beta}(a + y, b + n - y) \, d\theta \\
&= \binom{m}{\tilde{y}} \int_0^1 \theta^{\tilde{y}} (1-\theta)^{m-\tilde{y}} \frac{1}{B(a + y, b + n - y)} \theta^{a+y-1} (1-\theta)^{b+n-y-1} \\
&= \binom{m}{\tilde{y}} \frac{1}{B(a + y, b + n - y)} \int_0^1 \theta^{\tilde{y}+a+y-1} (1-\theta)^{m-\tilde{y}+b+n-y-1} \\
&= \binom{m}{\tilde{y}} \frac{B(\tilde{y} + a + y, b + n - y + m - \tilde{y})}{B(a + y, b + n - y)} \, .
\end{aligned} \tag{2.3}$$

In conclusione, per lo schema beta-binomiale, la distribuzione predittiva a posteriori è

$$f(\tilde{y} \mid y) = \binom{m}{\tilde{y}} \frac{B(a+y+\tilde{y}, b+n-y+m-\tilde{y})}{B(a+y, b+n-y)}, \tag{2.4}$$

ovvero, corrisponde ad una distribuzione di probabilità discreta chiamata *distribuzione beta-binomiale* di parametri m, $\alpha + y$ e $\beta + n - y$.

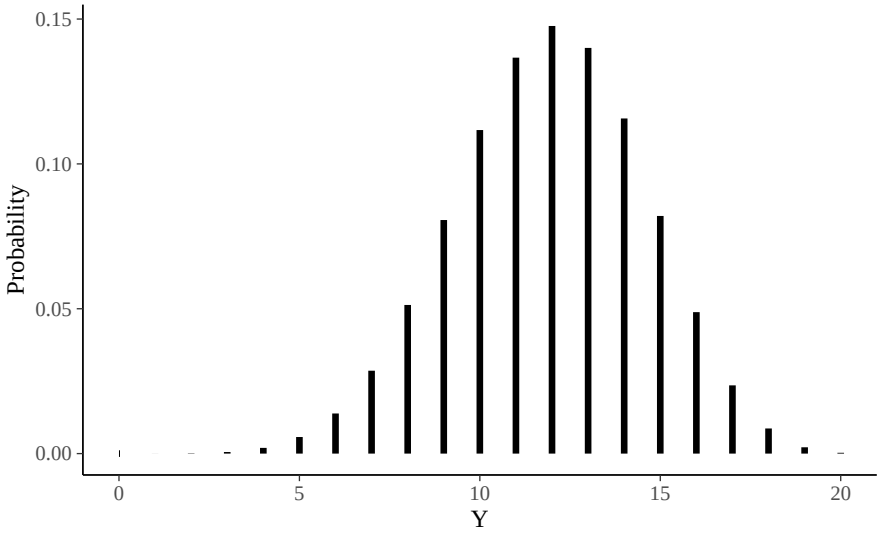
La distribuzione beta-binomiale di parametri N, α e β è una distribuzione discreta con una funzione di massa di probabilità uguale a

$$BetaBinomial(y \mid N, \alpha, \beta) = \binom{N}{y} \frac{B(y+\alpha, N-y+\beta)}{B(\alpha, \beta)},$$

dove la funzione beta è $B(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$. La distribuzione beta-binomiale è una distribuzione che non abbiamo discusso in precedenza e che useremo solo in questo contesto. Senza entrare nei dettagli, ci accontentiamo di sapere che tale distribuzione è implementata nella funzione `dbbinom()` del pacchetto `extraDistr`. Il significato dei parametri è chiarito nell'esempio discusso di seguito.

Nell'esempio che stiamo discutendo, relativo allo studio di Zetsche et al. (2019), la verosimiglianza è binomiale, i dati sono costituiti da 23 successi su 30 prove e la distribuzione a priori su θ è Beta$(2, 10)$; di conseguenza la distribuzione a posteriori è Beta$(25, 17)$. Vogliamo calcolare la distribuzione predittiva a posteriori per un nuovo campione di $m = 20$ osservazioni. Utilizzando il risultato precedente, la distribuzione predittiva sarà una beta-binomiale di parametri m, $\alpha+y$ e $\beta+n-y$, dove m è il numero di prove nel nuovo campione, α e β sono i parametri della distribuzione a priori, e y e n sono le quantità della verosimiglianza. Nel caso dell'esempio, dunque, $m = 20$, $\alpha = 2 + 23 = 25$, $\beta = 10 + 30 - 23 = 17$. Svolgendo i calcoli con la funzione `dbbinom()` possiamo ottenere un grafico della distribuzione predittiva a posteriori nel modo seguente:

```
prob <- extraDistr::dbbinom(0:20, 20, 25, 17)
tibble(Y = 0:20, Probability = prob) %>%
  ProbBayes::prob_plot(Color = "black")
```



La distribuzione predittiva a posteriori illustrata nella figura ci dice qual è la plausibilità di osservare $0, 1, \ldots, 20$ successi su $m = 20$ prove in un futuro campione di osservazioni, alla luce dei dati che abbiamo osservato nel campione corrente (23 successi in 30 prove), e considerate le nostre opinioni a priori sul valore θ (ovvero, Beta$(2, 10)$).

Esaminando la distribuzione predittiva notiamo che, in campioni futuri di 20 osservazioni, il valore più plausibile per $\tilde{y}$ è 12. Tuttavia, $\tilde{y}$ può assumere anche altri valori e la distribuzione predittiva ci informa sulla plausibilità relativa di ciascuno di tali possibili valori futuri $\tilde{y}$.

È desiderabile costruire un intervallo che contiene $\tilde{Y}$ ad un livello specificato di probabilità. Supponiamo che il livello di probabilità sia 0.89. L'intervallo si costruisce aggiungendo valori $\tilde{y}$ all'intervallo fino a che il contenuto di probabilità dell'insieme eccede la soglia di 0.89. La procedura è implementata nella funzione `discint()` del pacchetto `LearnBayes`. Per i dati dell'esempio otteniamo

```
LearnBayes::discint(cbind(0:20, prob), 0.89)
#> $prob
#> [1] 0.9145
#>
```

```
#> $set
#> [1]  8  9 10 11 12 13 14 15 16
```

da cui

$$P(8 \leq \tilde{Y} \leq 16) = 0.9145.$$

2.2 La distribuzione predittiva a posteriori mediante simulazione

In situazioni dove è difficile derivare l'esatta distribuzione predittiva a posteriori è possibile simulare valori estratti da tale distribuzione. Consideriamo un esempio riferito all'esempio che stiamo discutendo. È possibile implementare una simulazione predittiva estraendo prima i valori del parametro (in questo caso, θ) dalla distribuzione a posteriori. Con i valori del parametro così determinati, poi, si possono generare i valori delle possibili osservazioni future (nel caso presente, usando la distribuzione binomiale).

Per l'esempio che stiamo discutendo, la distribuzione a posteriori è una Beta(25, 17). Estaiamo 100,000 valori da tale distribuzione:

```
set.seed(12345)
a <- 2
b <- 10
n <- 30
y <- 23
pred_p_sim <- rbeta(1e5, a + y, b + n - y)
pred_y_sim <- rbinom(1e5, n, pred_p_sim)
```

```
ppd <- table(pred_y_sim) / 1e5
ppd
#> pred_y_sim
#>       3       4       5       6       7       8
#> 0.00002 0.00004 0.00011 0.00036 0.00096 0.00241
```

```
#>       9      10      11      12      13      14
#> 0.00533 0.01000 0.01753 0.02882 0.04290 0.06110
#>      15      16      17      18      19      20
#> 0.07812 0.09476 0.10763 0.11311 0.10821 0.09765
#>      21      22      23      24      25      26
#> 0.07982 0.06185 0.04156 0.02536 0.01299 0.00630
#>      27      28      29      30
#> 0.00224 0.00064 0.00016 0.00002
```

```
LearnBayes::discint(cbind(3:30, ppd), 0.89)
#> $prob
#>     12
#> 0.9155
#>
#> $set
#> 12 13 14 15 16 17 18 19 20 21 22 23
#> 12 13 14 15 16 17 18 19 20 21 22 23
```

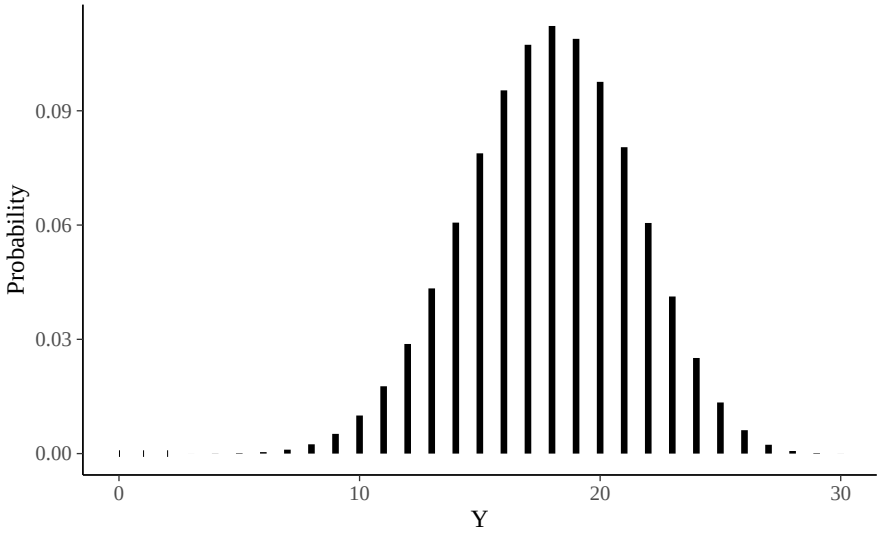
Confrontiamo i valori prodotti dalla simulazione con i valori esatti della distribuzione predittiva a posteriori:

```
prob30 <- extraDistr::dbbinom(0:30, 30, 25, 17)
```

```
LearnBayes::discint(cbind(0:30, prob30), 0.89)
#> $prob
#> [1] 0.9153
#>
#> $set
#>  [1] 12 13 14 15 16 17 18 19 20 21 22 23
```

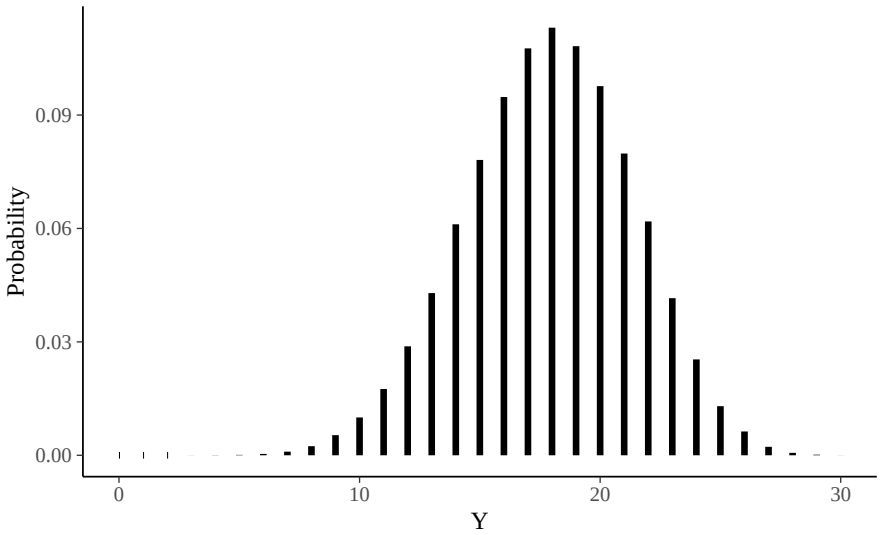
La distribuzione predittiva a posteriori esatta è

```
tibble(Y = 0:30, Probability = prob30) %>%
  ProbBayes::prob_plot(Color = "black")
```



Una rappresentazione della distribuzione a posteriori ottenuta mediante simulazione è

```
tibble(Y = 0:30, Probability = c(0, 0, 0, ppd)) %>%
  ProbBayes::prob_plot(Color = "black")
```



Si noti la somiglianza tra le due distribuzioni.

In conclusione, per il caso che abbiamo discusso, la predizione bayesiana di una nuova osservazione è una distribuzione beta-binomiale di parametri m, $\alpha+y$, e $\beta+n-y$, dove m è il numero di prove nel nuovo campione, α e β sono i parametri della distribuzione a priori, e y e n sono le quantità della verosimiglianza. Ricordiamo che, nello schema beta-binomiale, la distribuzione a posteriori è una Beta di parametri $\alpha + y$ e $\beta + n - y$. Quindi, detto in un altro modo, nello schema beta-binomiale la distribuzione predittiva a posteriori è una distribuzione beta-binomiale i cui tre parametri sono m (la numerosità del nuovo campione) e i due parametri di forma della distribuzione Beta che descrive la distribuzione a posteriori.

2.3 La distribuzione predittiva a posteriori mediante MCMC

Il metodo basato su simulazione che abbiamo discusso nel paragrafo precedente viene utilizzato per ottenere un'approssimazione della distribuzione predittiva a posteriori quando l'inferenza bayesiana viene svolta mediante i metodi MCMC. Le stime delle possibili osservazioni future $p(\tilde{y} \mid y)$, chiamate $p(y^{rep} \mid y)$, si ottengono nel modo seguente:

- campionare $\theta_i \sim p(\theta \mid y)$, ovvero campionare un valore del parametro dalla distribuzione a posteriori;
- campionare $y^{rep} \sim p(y^{rep} \mid \theta_i)$, ovvero campionare il valore di un'osservazione dalla funzione di verosimiglianza condizionata al valore del parametro definito nel passo precedente.

Se i due passaggi descritti sopra vengono ripetuti un numero sufficiente di volte, l'istogramma risultante approssimerà la distribuzione predittiva a posteriori che, in teoria (ma non in pratica) potrebbe essere ottenuta per via analitica.

Esercizio 2.1. Riportiamo qui sotto il codice Stan per generare $p(y^{rep} \mid y)$ nel caso dell'inferenza su una proporzione.

```
modelString <- "
data {
  int<lower=0> N;
  int<lower=0, upper=1> y[N];
}
parameters {
  real<lower=0, upper=1> theta;
}
model {
  theta ~ beta(2, 10);
  y ~ bernoulli(theta);
}
generated quantities {
  int y_rep[N];
  real log_lik[N];
  for (n in 1:N) {
    y_rep[n] = bernoulli_rng(theta);
    log_lik[n] = bernoulli_lpmf(y[n] | theta);
  }
}
"
writeLines(modelString, con = "code/betabin23-30-2-10.stan")
```

Si noti che nel nel blocco `generated quantities` sono state aggiunte le istruzioni necessarie per simulare y^{rep}, ovvero, `y_rep[n] = bernoulli_rng(theta)`. I dati dell'esempio sono:

```
data_list <- list(
  N = 30,
  y = c(rep(1, 23), rep(0, 7))
)
```

Compiliamo il codice Stan

```
file <- file.path("code", "betabin23-30-2-10.stan")
mod <- cmdstan_model(file)
```

ed eseguiamo il campionamento MCMC:

```
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  cores = 4L,
  chains = 4L,
  parallel_chains = 4L,
  refresh = 0,
  thin = 1
)
```

Per comodità, trasformiamo l'oggetto `fit` in un oggetto di classe `stanfit`:

```
stanfit <- rstan::read_stan_csv(fit$output_files())
```

Il contenuto dell'oggetto `stanfit` può essere esaminato nel modo seguente:

```
list_of_draws <- extract(stanfit)
print(names(list_of_draws))
#> [1] "theta"   "y_rep"   "log_lik" "lp__"
```

Dall'oggetto `list_of_draws` recuperiamo `y_rep`:

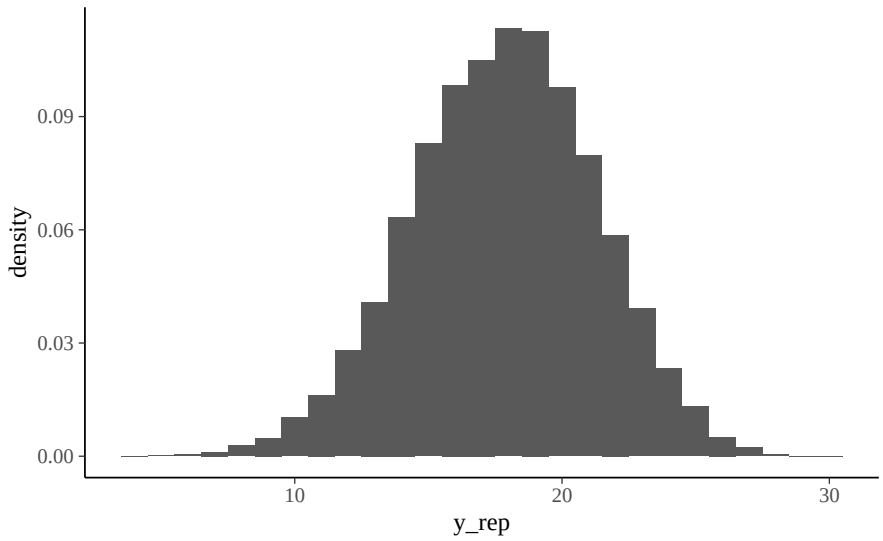
```
y_bern <- list_of_draws$y_rep
dim(y_bern)
#> [1] 16000    30
head(y_bern)
#>
#> iterations [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
#>       [1,]    1    1    1    1    0    1    1    1
#>       [2,]    0    1    0    1    1    1    0    0
#>       [3,]    0    1    0    1    1    1    0    0
#>       [4,]    1    0    0    1    1    0    0    1
#>       [5,]    0    0    0    1    1    0    1    1
#>       [6,]    1    1    1    1    1    1    0    1
```

```
#>
#> iterations [,9] [,10] [,11] [,12] [,13] [,14] [,15]
#>       [1,]    1     1     1     1     1     0     1
#>       [2,]    1     0     0     0     0     1     0
#>       [3,]    1     1     1     0     1     0     0
#>       [4,]    0     1     1     1     0     0     1
#>       [5,]    0     1     0     0     1     0     0
#>       [6,]    0     1     1     1     0     1     1
#>
#> iterations [,16] [,17] [,18] [,19] [,20] [,21] [,22]
#>       [1,]     1     1     1     1     0     0     1
#>       [2,]     0     1     0     1     1     1     0
#>       [3,]     1     0     1     1     0     1     0
#>       [4,]     0     1     0     1     0     0     1
#>       [5,]     1     1     1     1     1     0     1
#>       [6,]     0     1     0     1     1     0     0
#>
#> iterations [,23] [,24] [,25] [,26] [,27] [,28] [,29]
#>       [1,]     0     1     1     1     1     1     1
#>       [2,]     0     0     0     1     1     0     1
#>       [3,]     0     1     1     1     1     1     1
#>       [4,]     0     1     0     1     1     0     0
#>       [5,]     0     1     0     0     0     0     1
#>       [6,]     1     0     0     0     1     0     1
#>
#> iterations [,30]
#>       [1,]     1
#>       [2,]     1
#>       [3,]     0
#>       [4,]     1
#>       [5,]     0
#>       [6,]     1
```

Dato che il codice Stan definisce un modello per i dati grezzi (ovvero, per ciascuna singola prova Bernoulliana del campione), ogni riga di `y_bern` include 30 colonne, ciascuna delle quali corrisponde ad un campione (n = 16000 in questa simulazione) di possibili valori futuri $y_i \in \{0, 1\}$. Per ottenere una stima della distribuzione predittiva a posteriori `p(y_rep)`, ovvero, una stima della probabilità associata a ciascuno dei possibili

numeri di "successi" in $m = 30$ nuove prove future, è sufficiente calcolare la proporzione di valori 1 in ciascuna riga:

```
tibble(y_rep = rowSums(y_bern)) %>%
  ggplot(aes(x = y_rep, after_stat(density))) +
  geom_histogram(binwidth = 1)
```



Si noti che questo istogramma non può essere confrontato con quello ottenuto nella simulazione precedente dato che m è diverso nelle due simulazioni.

2.4 Posterior predictive checks

La distribuzione predittiva a posteriori viene utilizzata per eseguire i cosiddetti *controlli predittivi a posteriori* (*Posterior Predictive Checks*, PPC). Nella distribuzione predittiva a posteriori, viene generato un campione di dati possibili futuri utilizzando le proprietà del modello adattato. È ovvio che tali dati possibili futuri devono almento essere coerenti con i dati del campione presente. I PPC eseguono un confronto grafico tra $p(y^{rep} \mid y)$ e i dati osservati y: confrontando visivamente gli aspet-

ti chiave dei dati previsti futuri y^{rep} e dei dati osservati y è possibile determinare se il modello è adeguato.

Oltre al confronto visivo tra le distribuzioni $p(y)$ e $p(y^{rep})$ è anche possibile un confronto tra la distribuzione di varie statistiche descrittive, i cui valori sono calcolati su diversi campioni y^{rep}, e le corrispondenti statistiche calcolate sui dati osservati. Vengono solitamente considerate statistiche descrittive quali la media, la varianza, la deviazione standard, il minimo o il massimo, ma confronti di questo tipo sono possibili per qualunque altra statistica.

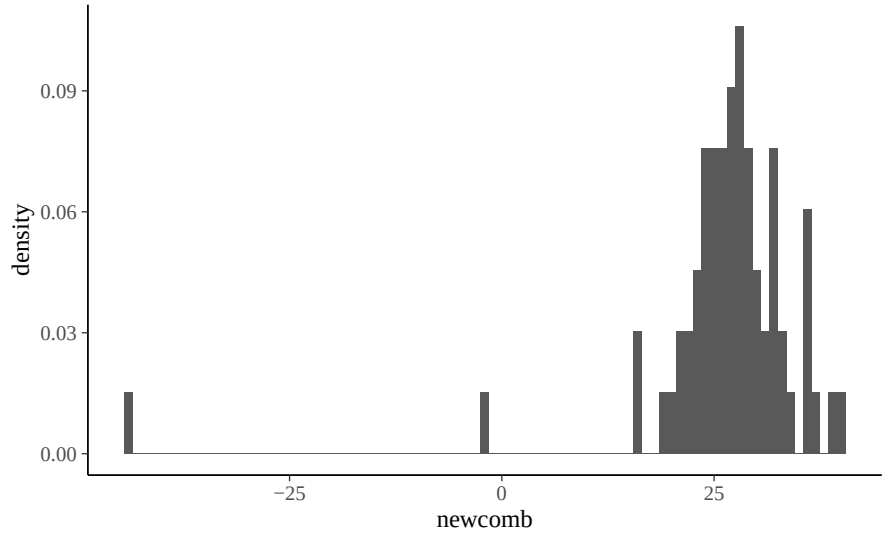
Esercizio 2.2. Esaminiamo ora un set di dati che non seguono la distribuzione normale (Gelman et al., 2020). I dati corrispondono ad una serie di misurazioni prese da Simon Newcomb nel 1882 come parte di un esperimento per stimare la velocità della luce. A questi dati verrà (inappropriatamente) adattata una distribuzione normale. L'obiettivo dell'esempio è quello di mostrare come i PPC possono rivelare la mancanza di adattamento di un modello ai dati.

I PPC mostrano che il modo più semplice per verificare l'adattamento del modello è quello di visualizzare y^{rep} insieme ai dati effettivi. Iniziamo a caricare i dati:

```
library("MASS")
data("newcomb")
```

Visualizziamo la distribuzione dei dati con un istogramma:

```
tibble(newcomb) %>%
  ggplot(aes(x = newcomb, after_stat(density))) +
  geom_histogram(binwidth = 1)
```



Creiamo un oggetto di tipo `list` dove inserire i dati:

```
data_list <- list(
  y = newcomb,
  N = length(newcomb)
)
```

Il codice Stan per il modello normale è il seguente:

```
modelString <- "
data {
  int<lower=0> N;
  vector[N] y;
}
parameters {
  real mu;
  real<lower=0> sigma;
}
model {
  mu ~ normal(25, 10);
  sigma ~ cauchy(0, 10);
  y ~ normal(mu, sigma);
}
```

```
generated quantities {
  vector[N] y_rep;
  for (n in 1:N) {
    y_rep[n] = normal_rng(mu, sigma);
  }
}
"
writeLines(modelString, con = "code/newcomb.stan")
```

Adattando il modello ai dati

```
file <- file.path("code", "newcomb.stan")
mod <- cmdstan_model(file)
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  cores = 4L,
  refresh = 0,
  thin = 1
)
```

otteniamo le seguenti stime dei parametri μ e σ:

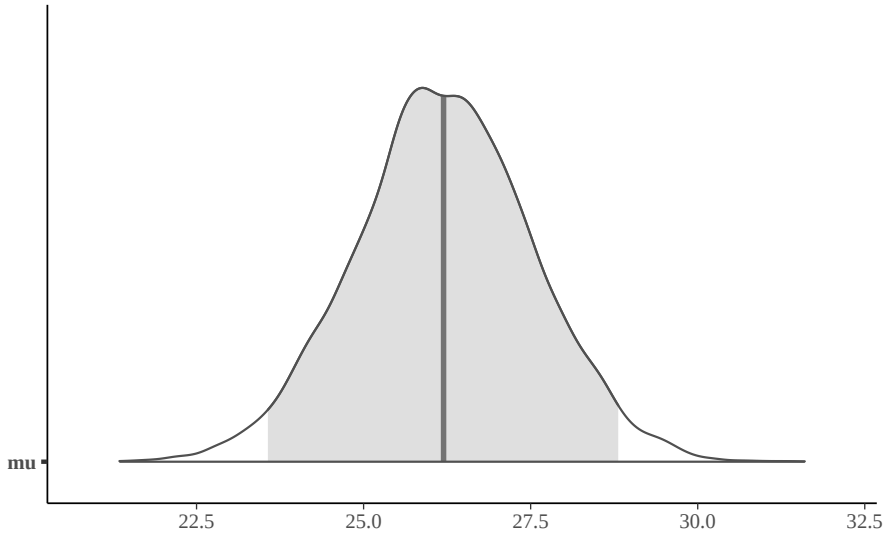
```
fit$summary(c("mu", "sigma"))
#> # A tibble: 2 x 10
#>   variable  mean median    sd   mad    q5   q95  rhat
#>   <chr>    <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 mu        26.2   26.2 1.33  1.30  24.0   28.4  1.00
#> 2 sigma     10.9   10.8 0.958 0.943  9.40  12.5  1.00
#> # ... with 2 more variables: ess_bulk <dbl>,
#> #   ess_tail <dbl>
```

Trasformiamo `fit` in un oggetto `stanfit`:

```
stanfit <- rstan::read_stan_csv(fit$output_files())
```

La distribuzione a posteriori di μ è

```
mu_draws <- as.matrix(stanfit, pars = "mu")
mcmc_areas(mu_draws, prob = 0.95) # color 95% interval
```



Confrontiamo μ con la media di y:

```
mean(newcomb)
#> [1] 26.21
```

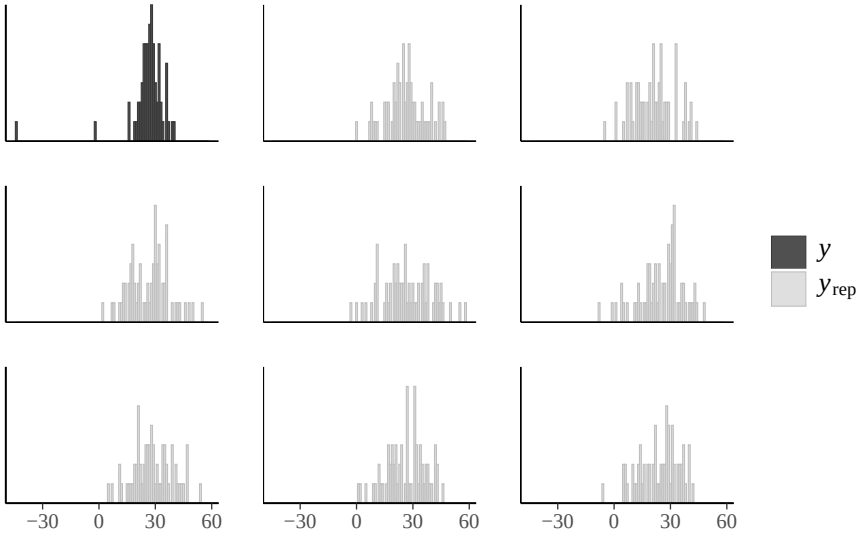
Anche se trova la media giusta, il modello non è comunque adeguato a prevedere le altre proprietà della y. Estraiamo y^{rep} dall'oggetto `stanfit`:

```
y_rep <- as.matrix(stanfit, pars = "y_rep")
dim(y_rep)
#> [1] 16000    66
```

I valori `y_rep` sono i dati della distribuzione predittiva a posteriori che sono stati simulati usando gli stessi valori X dei predittori utilizzati per

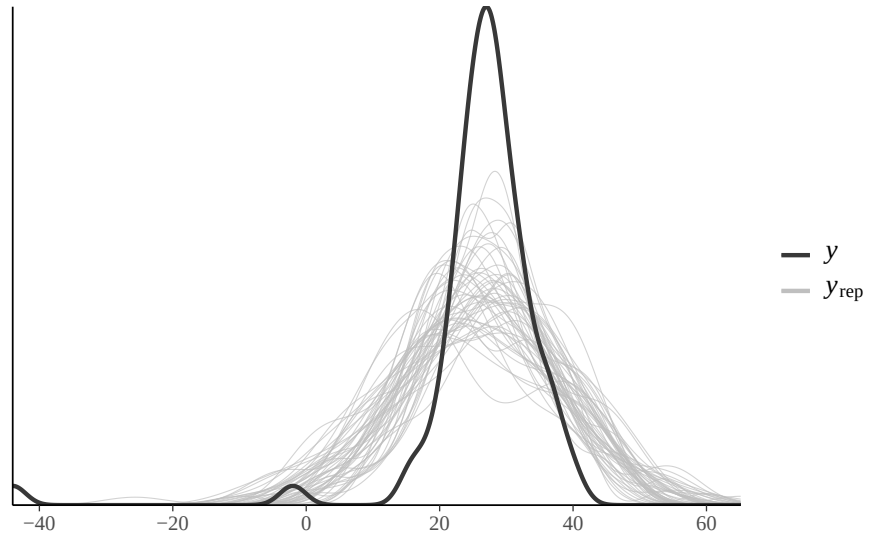
adattare il modello. Il confronto tra l'istogramma della y e gli istogrammi di diversi campioni y^{rep} mostra una scarsa corrispondenza tra i due:

```
ppc_hist(data_list$y, y_rep[1:8, ], binwidth = 1)
```



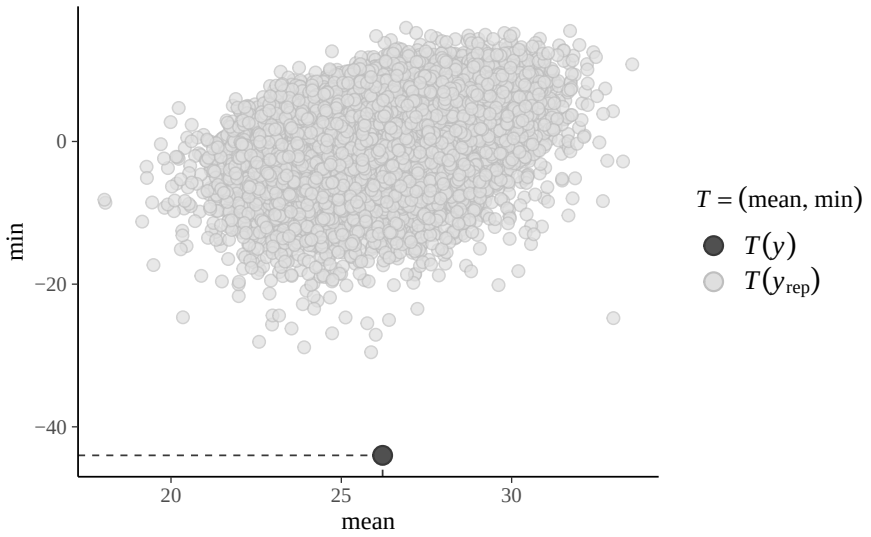
Alla stessa conclusione si giunge tramite un confronto tra la funzione di densità empirica della y e quella di diversi campioni y^{rep}:

```
ppc_dens_overlay(data_list$y, y_rep[1:50, ])
```



Generiamo ora i PPC per la media e il minimo della distribuzione:

```
ppc_stat_2d(data_list$y, y_rep, stat = c("mean", "min"))
```



Mentre la media viene riprodotta accuratamente dal modello (come abbiamo visto sopra), ciò non è vero per il minimo dela distribuzione. L'origine di questa mancanza di adattamento è il fatto che la distribuzione

delle misurazioni della velocità della luce è asimmetrica negativa. Dato che ci sono poche osservazioni nella coda negativa della distribuzione, solo per fare un esempio, utilizzeremo ora un secondo modello che ipotizza una distribuzione t di Student:

```
modelString <- "
data {
  int<lower=0> N;
  vector[N] y;
}
parameters {
  real mu;
  real<lower=0> sigma;
  real<lower=0> nu;
}
model {
  mu ~ normal(25, 10);
  sigma ~ cauchy(0, 10);
  nu ~ cauchy(0, 10);
  y ~ student_t(nu, mu, sigma);
}
generated quantities {
  vector[N] y_rep;
  for (n in 1:N) {
    y_rep[n] = student_t_rng(nu, mu, sigma);
  }
}
"
writeLines(modelString, con = "code/newcomb2.stan")
```

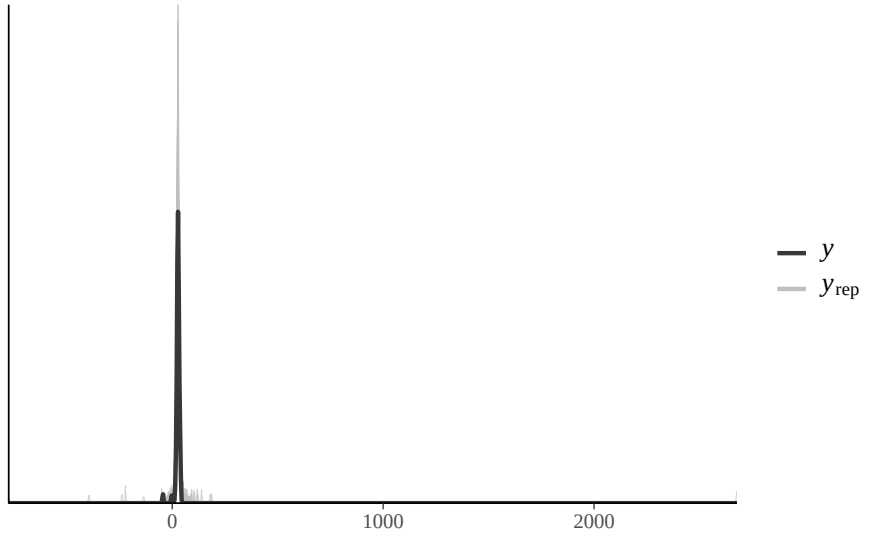
Adattiamo questo secondo modello ai dati.

```
file <- file.path("code", "newcomb2.stan")
mod <- cmdstan_model(file)
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
```

```
  chains = 4L,
  cores = 4L,
  parallel_chains = 2L,
  refresh = 0,
  thin = 1
)
#> Running MCMC with 4 parallel chains...
#>
#> Chain 1 finished in 0.3 seconds.
#> Chain 2 finished in 0.3 seconds.
#> Chain 3 finished in 0.3 seconds.
#> Chain 4 finished in 0.3 seconds.
#>
#> All 4 chains finished successfully.
#> Mean chain execution time: 0.3 seconds.
#> Total execution time: 0.5 seconds.
```

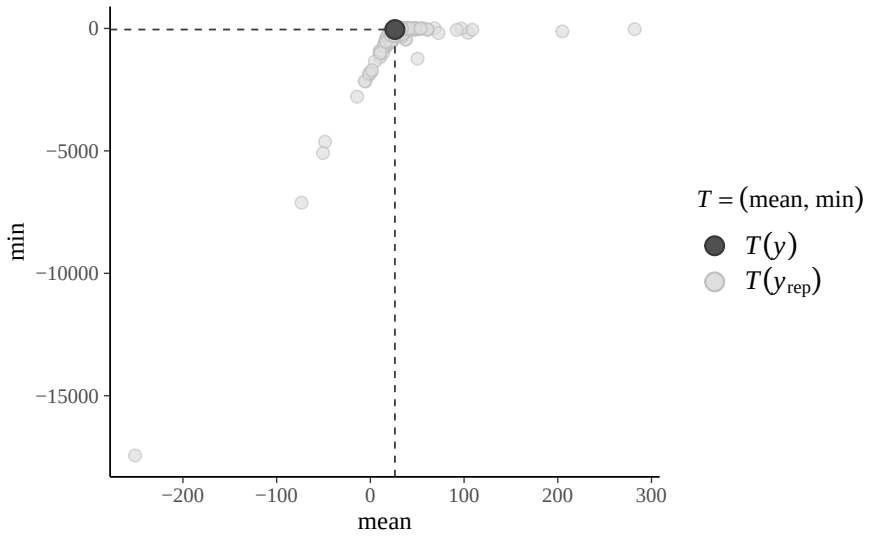
Per questo secondo modello il confronto tra la funzione di densità empirica della y e quella di diversi campioni y^{rep} risulta adeguato:

```
stanfit <- rstan::read_stan_csv(fit$output_files())
y_rep <- as.matrix(stanfit, pars = "y_rep")
ppc_dens_overlay(data_list$y, y_rep[1:50, ])
```



Inoltre, anche la statistica "minimo della distribuzione" viene ben predetta dal modello.

```
ppc_stat_2d(data_list$y, y_rep, stat = c("mean", "min"))
```



In conclusione, per le misurazioni della velocità della luce di Newcomb l'accuratezza predittiva del modello basato sulla distribuzione t

di Student è chiaramente migliore di quella del modello normale.

Commenti e considerazioni finali

Questo capitolo discute la predizione bayesiana e ne mostra un'applicazione nel caso dei controlli predittivi a posteriori. A questo proposito è necessario notare un punto importante: un buona corrispondenza tra y e y^{rep} costituisce una condizione necessaria ma non sufficiente per la validità del modello. Infatti, i PPC non sono in grado di garantire la generalizzabilità del modello a nuovi campioni di dati. D'altra parte, invece, se i PPC mostrano un cattivo adattamento del modello ai dati previsti futuri, questo ci dice chiaramente che il modello è specificato in manera errata.

Bibliografia

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and other stories*. Cambridge University Press.

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

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

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

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