Data Science per psicologi

Corrado Caudek

2021-09-13

Indice

Inferenza Bayesiana

Capitolo 1

Il problema inverso

La statistica descrittiva si occupa della descrizione, sintesi e presentazione delle informazioni contenute nei dati osservati. Essa spesso rappresenta la fase preliminare di uno studio e un ausilio per l'individuazione di possibili modelli da utilizzare nella successiva fase dell'analisi inferenziale. A differenza della statistica descrittiva, quella inferenziale assume che le osservazioni siano il risultato di un campionamento statistico e il suo obiettivo è ricavare informazioni circa l'intera popolazione a partire dall'osservazione di un suo sottoinsieme.

L'aleatorietà del campione è l'aspetto sostanziale che distingue la statistica inferenziale dalla statistica descrittiva. Nel caso più semplice, il campione statistico è trattato come un insieme di realizzazioni di una variabile casuale assunta a modello del fenomeno oggetto di indagine. La natura del campionamento statistico e la distribuzione di tale variabile casuale determinano il modello statistico. La statistica inferenziale consente di stimare, verificare ipotesi, o effettuare previsioni sul fenomeno in esame.

L'inferenza Bayesiana è un approccio all'inferenza statistica in cui le probabilità non sono interpretate come frequenze, proporzioni o concetti analoghi, ma piuttosto come il grado di fiducia che una singola persona attribuisce al verificarsi di un evento sulla base delle proprie conoscenze e delle informazioni di cui dispone. Come suggerito dal nome stesso, la statistica bayesiana è un approccio all'analisi dei dati e alla stima dei parametri basato sul teorema di Bayes.

1.1 Inferenza statistica come un problema inverso

- L'*inferenza deduttiva* procede in maniera deterministica dai fatti verso le conclusioni. Ad esempio, se dico che tutti gli uomini sono mortali e che Socrate è un uomo, posso concludere deduttivamente che Socrate è mortale.

- L'*inferenza induttiva*, invece, procede dalle osservazioni ai fatti. Se pensiamo ai fatti come a ciò che governa o genera le osservazioni, allora l'induzione è una sorta di inferenza inversa.
- L'*inferenza statistica* è un tipo di inferenza induttiva che è specificamente formulata come un problema inverso.

L'inferenza bayesiana è formulata nei termini di un problema inverso che segue la regola di Bayes. Per fissare la notazione, nel seguito y rappresenterà le variabili osservate, ovvero i dati, e θ rappresenterà i parametri incogniti di un modello statistico. Sia y che θ sono concepiti come delle variabili casuali. Con x verranno invece denotate delle quantità note, come i predittori nel modello di regressione.

1.1.1 Funzioni di probabilità

L'inferenza bayesiana utilizza le seguenti distribuzioni di probabilità (o densità di probabilità):

- la *distribuzione a priori* $p(\theta)$ — la credenza iniziale riguardo alla credibilità di ciascun valore θ;
- la *funzione di verosimiglianza* $p(y \mid \theta)$ — la credibilità che il ricercatore assegnerebbe ai dati osservati se conoscesse il parametro di interesse θ;
- la *verosimiglianza marginale* $p(y)$ — quanto sono credibili i dati y alla luce della nostra credenza a priori relativamente a θ. In termini formali:

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

- la *distribuzione a posteriori* $p(\theta \mid y)$ — la nuova credenza a posteriori relativamente alla credibilità di ciascun valore θ alla luce dei dati $Y = y$.

1.1.2 La regola di Bayes

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

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

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

Possiamo dire che la regola di Bayes viene usata per aggiornare le credenze a priori su θ (ovvero, la distribuzione a priori) in modo tale da produrre le nuove credenze a posteriori $p(\theta \mid y)$ che combinano le informazioni fornite dai dati y con le credenze precedenti. La distribuzione a posteriori riflette dunque l'aggiornamento delle credenze del ricercatore alla luce dei dati.

La (1.5) rende evidente che, in ottica bayesiana, la quantità di interesse θ non è fissata come nell'impostazione frequentista, ma è una variabile casuale la cui distribuzione di probabilità è influenzata sia dalle informazioni a priori sia dai dati a disposizione. In altre parole, nell'approccio bayesiano non esiste un valore vero di θ, ma vogliamo fornire invece un giudizio di probabilità. Prima delle osservazioni, sulla base delle nostre conoscenze assegniamo a θ una distribuzione a priori di probabilità. Dopo le osservazioni, correggiamo il nostro giudizio e assegniamo a θ una distribuzione a posteriori di probabilità. La distribuzione a posteriori $p(\theta \mid y)$ contiene tutta l'informazione riguardante il parametro θ e viene utilizzata per produrre indicatori sintetici, per la determinazione di stime puntuali o intervallari, e per la verifica d'ipotesi.

1.2 Un esempio concreto

L'esempio più semplice di inferenza bayesiana è quello nel quale i dati sono rappresentati da una proporzione. Per questo tipo di dati possiamo adottare il seguente modello statistico

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

laddove θ è la probabiltà che una prova Bernoulliana assuma il valore 1 e n corrisponde al numero di prove Bernoulliane. Questo modello statistico 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 ha 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 statistico binomiale, sia possibile assegnare una probabilità a ciascun possibile valore $y \in \{0, 1, \dots, n\}$ *assumendo noto il valore del parametro* θ. Ma ora abbiamo il *problema inverso*, ovvero quello di fare inferenza su θ alla luce dei dati campionari y. In altre parole, riteniamo di conoscere il modello probabilistico che ha generato i dati, ma di tale modello non conosciamo i parametri. Nel caso presente, il modello probabilistico è quello binomiale. Noi vogliamo ottenere informazioni sul valore di θ conoscendo il numero osservato y di successi.

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

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

conto di una tale complessità, un modello come quello che stiamo discutendo qui certamente non sarà sufficiente.

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

1.2.1 Notazione

Per rappresentare in un modo conciso i modelli statistici 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, per l'esempio presente, la verosimiglianza può essere scritta come:

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

1.2.2 Il problema inverso

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

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

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

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

1.2.3 Cos'è un parametro del modello?

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

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

Per comprendere il ruolo del parametro θ, possiamo generare un grafico della verosimiglianza dei dati come funzione di θ. Poniamo $y = 23$ e $n = 30$. La figura mostra, per ogni possibile valore di $\theta \in [0, 1]$ (sull'asse orizzontale), la verosimiglianza dei dati (sull'asse verticale). Dalla figura notiamo che la credibilità dei dati dipende dal valore del parametro θ: i dati risultano più o meno verosimili a seconda del valore di θ.

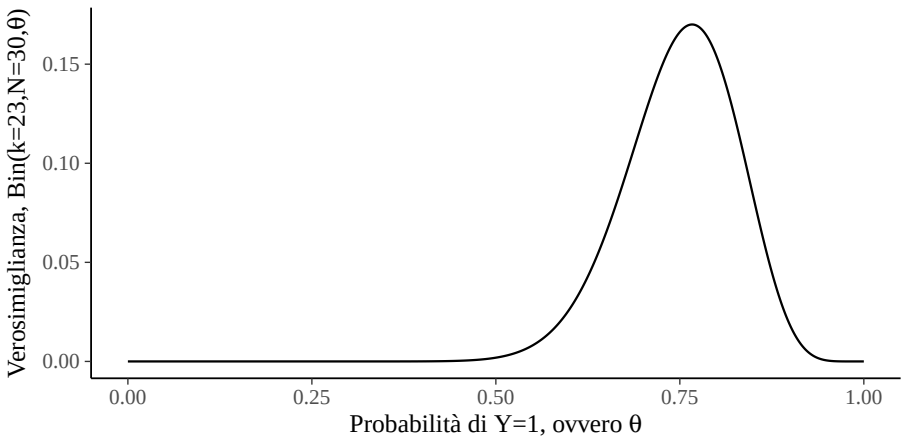

Figura 1.1: *Funzione di verosimiglianza per il modello binomiale con* $y = 23$ *e* $n = 30$.

1.2.4 La distribuzione a priori sui parametri

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

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

una distribuzione a priori dovrebbero essere supportate da una qualche motivazione, come ad esempio i risultati di ricerche precedenti, o altre motivazioni giustificabili. Tuttavia, le credenze soggettive sono solo uno dei possibili modi per giustificare le distribuzioni a priori sui parametri.

Possiamo distinguere tre tipi principali di motivazioni per le distribuzioni a priori $p(\theta)$.

1. Le *distribuzioni a priori soggettive* catturano le credenze del ricercatore nel senso sopra descritto.
2. Le *distribuzioni a priori con finalità pratiche* sono distribuzioni a priori che vengono utilizzate pragmaticamente a causa di una loro utilità specifica, ad esempio, perché semplificano un calcolo matematico o una simulazione al computer, o perché aiutano nel ragionamento statistico, come ad esempio quando vengono formulate gli *skeptical priors* che hanno l'obiettivo di lavorare in senso contrario ad una particolare conclusione.

Oltre alla motivazione che giustifica una distribuzione a priori, possiamo distinguere tra diverse distribuzioni a priori in base a quanto fortemente impegnano il ricercatore a ritenere come plausibile un particolare intervallo di valori dei parametri. Il caso più estremo è quello che rivela una totale assenza di conoscenze a priori, il che conduce alle *distribuzioni a priori non informative*, ovvero quelle che assegnano lo stesso livello di credibilità a tutti i valori dei parametri. Le distribuzioni a priori informative, d'altra parte, possono essere *debolmente informative* o *fortemente informative*, a seconda della forza della credenza che esprimono. Il caso più estremo di credenza a priori è quello che riassume il punto di vista del ricercatore nei termini di un *unico valore* del parametro, il che assegna tutta la probabilità (massa o densità) su di un singolo valore di un parametro. Poiché questa non è più una distribuzione di probabilità, sebbene ne soddisfi la definizione, in questo caso si parla di una *distribuzione a priori degenerata*.

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

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

1.2.5 Scelta 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

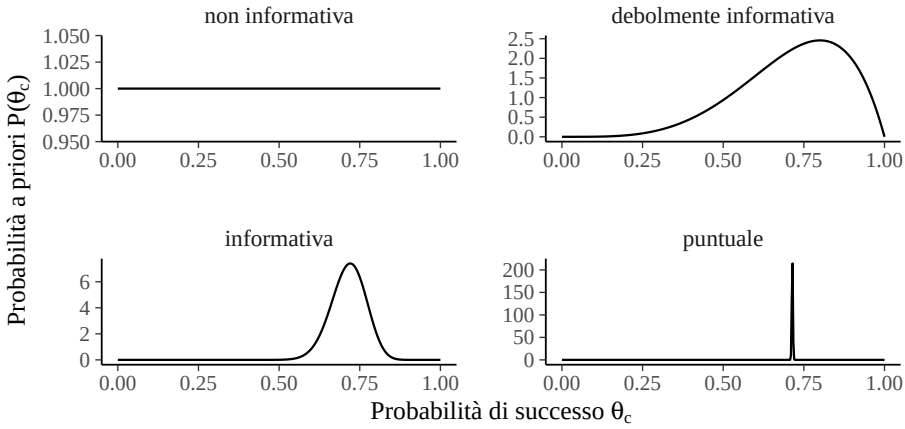

Figura 1.2: *Esempi di distribuzioni a priori per il parametro θ_c nel Modello Binomiale.*

in quanto può avere un impatto sostanziale sui risultati finali. La soggettività delle distribuzioni a priori è evidenziata dai critici come un potenziale svantaggio dei metodi bayesiani. A questa critica, van de Schoot et al. (2021) rispondono dicendo che, al di là della scelta delle distribuzioni a priori, ci sono molti elementi del processo di inferenza statistica che sono soggettivi, ovvero la scelta del modello statistico e le ipotesi sulla distribuzione degli errori. In secondo luogo, van de Schoot et al. (2021) notano come le distribuzioni a priori svolgono due importanti ruoli statistici: quello della "regolarizzazione della stima", ovvero, il processo che porta ad indebolire l'influenza indebita di osservazioni estreme, e quello del miglioramento dell'efficienza della stima, ovvero, la facilitazione dei processi di calcolo numerico di stima della distribuzione a posteriori.

L'effetto della distribuzione a priori sulla distribuzione a posteriori verrà discusso nel Capitolo **??**.

1.3 Verosimiglianza marginale

Al denominatore della regola di Bayes abbiamo la verosimiglianza marginale $p(y)$. Obiettivo di questo Paragrafo è chiarire questo concetto in riferimento al caso più semplice, ovvero quello della distribuzione binomiale.

Sia Y una variabile casuale con funzione di massa di probabilità $p(Y)$. Iniziamo la discussione con un semplice esempio in cui supponiamo che la funzione di massa di probabilità della Y sia definita nei termini del parametro θ e che θ possa assumere solo i valori 0.1, 0.5, 0.9, ciascuno con eguale probabilità. In altre parole, la probabilità che θ sia 0.1, 0.5, o 0.9 è sempre 1/3.

Ponendo $n = 30$ e $y = 23$, ad esempio, la funzione di verosimiglianza diventa

$$p(y = 23, n = 30 \mid \theta) = \binom{30}{23}\theta^{23}(1-\theta)^7.$$

La *verosimiglianza marginale* $p(y = 23, n = 30)$ basata su θ si ottiene marginalizzando rispetto al parametro θ: per ogni possibile valore del parametro θ, calcoliamo il valore della verosimiglianza e lo moltiplichiamo per la probabilità di θ; poi sommiamo tutti i prodotti ottenuti in questo modo. Matematicamente, ciò significa eseguire l'operazione descritta di seguito.

Nell'esempio abbiamo tre possibili valori θ che chiameremo $\theta_1 = 0.1$, $\theta_2 = 0.5$ e $\theta_3 = 0.9$. Ciascuno ha probabilità 1/3, quindi $p(\theta_1) = p(\theta_2) = p(\theta_3) = 1/3$. Date queste informazioni possiamo calcolare la verosimiglianza marginale come segue:

$$\begin{aligned} p(y = 23, n = 30) = &\binom{30}{23}\theta_1^{23}(1-\theta_1)^7 \cdot p(\theta_1) \\ &+ \binom{30}{23}\theta_2^{23}(1-\theta_2)^7 \cdot p(\theta_2) \\ &+ \binom{30}{23}\theta_3^{23}(1-\theta_3)^7 \cdot p(\theta_3), \end{aligned}$$

ovvero

$$\begin{aligned} p(y = 23, n = 30) = &\binom{30}{23}0.1^{23}(1-0.1)^7 \cdot \frac{1}{3} \\ &+ \binom{30}{23}0.5^{23}(1-0.5)^7 \cdot \frac{1}{3} \\ &+ \binom{30}{23}0.9^{23}(1-0.9)^7 \cdot \frac{1}{3}. \end{aligned}$$

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

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

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

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

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

```
BinLik <- function(theta) {
  choose(30, 23) * theta^23 * (1 - theta)^7
}
integrate(BinLik, lower = 0, upper = 1)$value
#> [1] 0.03225806
```

1.3.1 Soluzione analitica

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

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

la verosimiglianza marginale diventa

$$\begin{aligned} p(y) &= \binom{n}{y} \int p(y \mid \theta) p(\theta) d\theta \\ &= \binom{n}{y} \int_0^1 \theta^y (1 - \theta)^{n-y} \frac{1}{B(a, b)} \theta^{a-1} (1 - \theta)^{b-1} d\theta \\ &= \binom{n}{y} \frac{1}{B(a, b)} \int_0^1 \theta^{y+a-1} (1 - \theta)^{n-y+b-1} \\ &= \binom{n}{y} \frac{B(y + a, n - y + b)}{B(a, b)}, \end{aligned} \tag{1.7}$$

in quanto

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

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

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

e dunque

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

Definiamo

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

Il risultato cercato si ottiene con

```
choose(30, 23) * B(alpha, beta) / B(a, b)
#> [1] 0.03225806
```

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

$$\binom{n}{y} \frac{B(y + a, n - y + b)}{B(a, b)}. \tag{1.8}$$

1.4 La distribuzione a posteriori

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

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

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

Considerazioni conclusive

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. Questa nuova distribuzione di probabilità è chiamata "distribuzione a posteriori". L'approccio bayesiano riassumere l'incertezza dell'inferenza fornendo un intervallo di valori sulla distribuzione di probabilità a posteriori che include il 95% della probabilità — questo intervallo è chiamato "intervallo di credibilità del 95%".

Appendice A

Simbologia di base

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

- 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."
- 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 " $\iff$ " 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, $\arg\max$ identifica l'insieme dei punti per i quali una data funzione raggiunge il suo massimo. In altre parole, $\arg\max_x f(x)$ è l'insieme dei valori di x per i quali $f(x)$ raggiunge il valore più alto.

Appendice 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" "Yes"
#> [10] "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  TRUE
#> [10]  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	$(a, b]$	$a < x \leq b$

Appendice 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 $A \ \ B$ l'insieme degli elementi di A che non appartengono a B:

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

Insieme complementare. Nel caso che sia $B \subseteq A$, l'insieme differenza $A \ \ B$ 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.

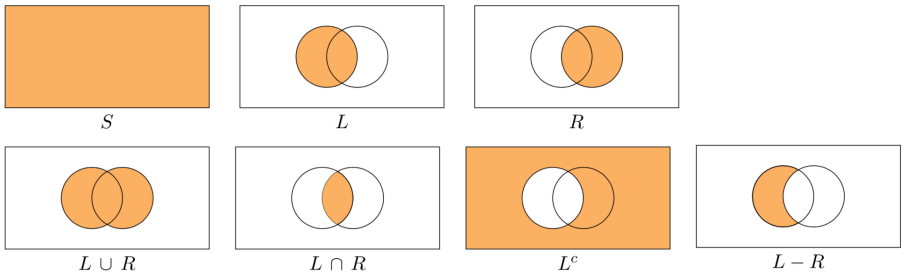

Figura C.1: *In tutte le figure S è la regione delimitata dal rettangolo, L è la regione all'interno del cerchio di sinistra e R è la regione all'interno del cerchio di destra. La regione evidenziata mostra l'insieme indicato sotto ciascuna figura.*

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

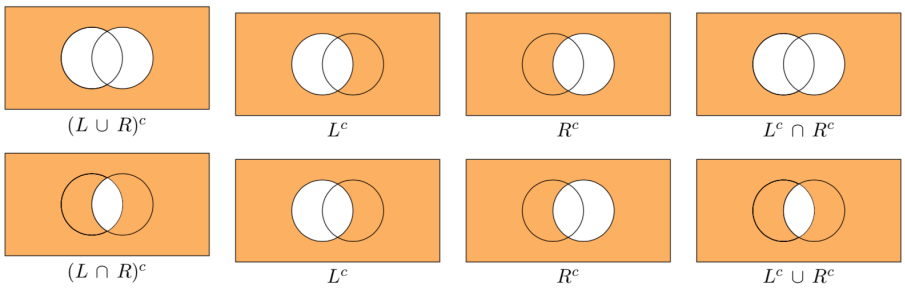

Figura C.2: *Dimostrazione delle leggi di DeMorgan.*

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 $\mathrm{c}(A)$.

Appendice 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

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

Esempio 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

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

Bibliografia

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.