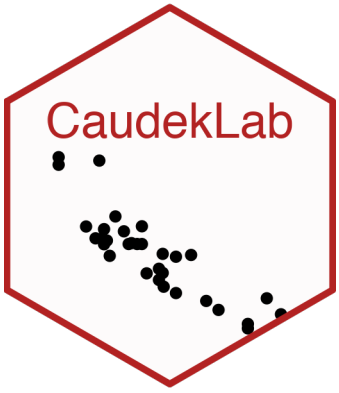
Psicometria

Corrado Caudek

2021-12-27

Questo documento è stato realizzato con:

- LaTeX e la classe memoir (http://www.ctan.org/pkg/memoir);
- R (http://www.r-project.org/) e RStudio (http://www.rstudio.com/);
- bookdown (http://bookdown.org/) e memoiR (https://ericmarcon.github.io/memoiR/).



Nel blog della mia pagina personale sono forniti alcuni approfondimenti degli argomenti qui trattati. https://ccaudek.github.io/caudeklab/

Indice

Copyright © 2022.

Data della versione presente: Dicembre 27, 2021.

Prefazione

Data Science per psicologi contiene il materiale delle lezioni dell'insegnamento di *Psicometria B000286* (A.A. 2021/2022) rivolto agli studenti del primo anno del Corso di Laurea in Scienze e Tecniche Psicologiche dell'Università degli Studi di Firenze. *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

> *It's worth noting, before getting started, that this material is hard. If you find yourself confused at any point, you are normal. Any sense of confusion you feel is just your brain correctly calibrating to the subject matter. Over time, confusion is replaced by comprehension […]*
>
> — Richard McElreath

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

C'è una ragione molto semplice che dovrebbe farci capire perché la Data science è così importante per la psicologia. Infatti, a ben pensarci, la psicologia è una disciplina intrinsecamente statistica, se per statistica intendiamo quella disciplina che studia la variazione delle caratteristiche degli individui nella popolazione. La psicologia studia *gli individui* ed è proprio la variabilità inter- e intra-individuale ciò che vogliamo descrivere e, in certi casi, predire. In questo senso, la psicologia è molto diversa dall'ingegneria, per esempio. Le proprietà di un determinato ponte sotto certe condizioni, ad esempio, sono molto simili a quelle di un altro ponte, sotto le medesime condizioni. Quindi, per un ingegnere la statistica è poco importante: le proprietà dei materiali sono unicamente dipendenti dalla loro composizione e restano costanti. Ma lo stesso non può dirsi degli individui: ogni individuo è unico e cambia nel tempo. E le variazioni tra gli individui, e di un individuo nel tempo, sono l'oggetto di studio proprio della psicologia: è dunque chiaro che i problemi che la psicologia si pone sono molto diversi da quelli affrontati, per esempio, dagli ingegneri. Questa è la ragione per cui abbiamo tanto bisogno della Data science in psicologia: perché la Data science ci consente di descrivere la variazione e il cambiamento. E queste sono appunto le caratteristiche di base dei fenomeni psicologici.

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il "lavoro sporco" (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò che ci interessa, ovvero dei problemi psicologici slegati dai dettagli "tecnici" della Data

science? La risposta a questa domanda è che non è possibile progettare uno studio 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 verrano trattati i temi base della Data science e verrà adottato un punto di vista bayesiano, che corrisponde all'approccio più recente e sempre più diffuso in psicologia.

Come studiare

> *I know quite certainly that I myself have no special talent. Curiosity, obsession and dogged endurance, combined with self-criticism, have brought me to my ideas.*
>
> — Albert Einstein

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

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

Sviluppare un metodo di studio efficace

> *Memorization is not learning.*
>
> — Richard Phillips Feynman

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l'atteggiamento mentale che descriverò qui sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo.

Fornisco qui sotto una breve descrizione del "metodo di studio" che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento (Burger & Starbird, 2012).

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.

- Gli errori che facciamo sono i nostri migliori maestri. Istintivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un'informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.

- C'è ovviamente un aspetto "psicologico" nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: "mi arrendo", "non ho idea di cosa fare!". Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose "migliori" che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c'è qualcosa che non so fare e non ho idea di come affontare, mi dico: "oggi non ho proprio voglia di fare fatica", non ho voglia di mettermi nello stato mentale per cui "in 10 minuti devo risolvere il problema perché dopo devo fare altre cose". Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto "marginale" del problema, che so come affrontare, oppure considero l'aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, opppure lo stesso problema in un altro contesto. Non mi pongo l'obiettivo "risolvi il problema in 10 minuti", ma invece quello di farmi un'idea "generale" del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se "parto per la tangente", ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a "lavorare" al problema. Allora perché non faccio sempre così? C'è ovviamente l'aspetto dei "10 minuti" che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possiamo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.

- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: "se questo è vero, allora deve succedere quest'altra cosa". Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c'era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.

- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare di colmarle. L'atteggiamento naturale, quando non capiamo i dettagli di qualcosa, è quello di pensare: "non importa, ho capito in maniera approssimativa questo punto, non devo preoccuparmi del resto". Ma in realtà non è vero: se la nostra comprensione è superficiale, quando il problema verrà presentato in una nuova forma, non riusciremo a risolverlo. Per cui i dubbi che ci vengono quando studiamo qualcosa sono il nostro alleato più prezioso: ci dicono esattamente quali sono gli aspetti che dobbiamo approfondire per potere migliorare la nostra preparazione.
- È utile sviluppare una visione d'insieme degli argomenti trattati, capire l'obiettivo generale che si vuole raggiungere e avere chiaro il contributo che i vari pezzi di informazione forniscono al raggiungimento di tale obiettivo. Questa organizzazione mentale del materiale di studio facilita la comprensione. È estremamente utile creare degli schemi di ciò che si sta studiando. Non aspettate che sia io a fornirvi un riepilogo di ciò che dovete imparare: sviluppate da soli tali schemi e tali riassunti.
- Tutti noi dobbiamo imparare l'arte di trovare le informazioni, non solo nel caso di questo insegnamento. Quando vi trovate di fronte a qualcosa che non capite, o ottenete un oscuro messaggio di errore da un software, ricordatevi: "Google is your friend".

Corrado Caudek

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

Capitolo 1

Distribuzioni a priori coniugate

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

1.1 Pensare a una proporzione "in termini soggettivi"

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

Consideriamo ora uno scenario diverso: quello in cui non conosciamo le proporzioni delle palline colorate nell'urna. Cioè, nell'esempio precedente, sappiamo solo che ci sono due tipi di palline colorate nell'urna, ma non sappiamo che 40 palline su 60 sono bianche (proporzione di bianco = 2/3) e 20 delle 60 palline sono rosse (proporzione di rosso = 1/3). Come è possibile imparare qualcosa sulle proporzioni di palline bianche e rosse? Poiché contare 60 palline può essere noioso, è possibile invece inferire le proporzioni cercate estraendo un campione di palline dall'urna e osservando i colori delle palline nel campione? Espresso in questo modo, questo diventa un problema di inferenza statistica, perché stiamo cercando di inferire la proporzione π della popolazione, sulla base di un campione della popolazione.

Per continuare con l'esempio precedente: come è possibile inferire π (ad esempio, la proporzione di palline rosse nella popolazione – cioè le 60 palline), in base al numero di palline rosse e bianche che osserviamo nel campione (per esempio, 10 palline)?

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

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

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

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

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

1.2 Il denominatore bayesiano

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

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

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

Definizione 1.1. Una distribuzione di probabilità a priori $p(\theta)$ si dice *coniugata* al modello usato se la distribuzione a priori e la distribuzione a posteriori hanno la stessa forma funzionale. Dunque, le due distribuzioni differiscono solo per il valore dei parametri.

In altre parole, è possibile ottenere la distribuzione posteriore per via analitica solo per alcune specifiche combinazioni di distribuzione a priori e verosimiglianza. Tuttavia, l'uso di distribuzioni coniugate limita considerevolmente la flessibilità della modellizzazione. Per questa ragione, la strada principale che viene seguita nella modellistica bayesiana è quella che porta a determinare la distribuzione a posteriori non per via analitica, ma bensì mediante metodi numerici. La simulazione fornisce dunque la strategia generale del calcolo bayesiano. A questo fine vengono usati i metodi di campionamento detti Monte-Carlo Markov-Chain (MCMC). Tali metodi costituiscono una potente e praticabile alternativa per la costruzione della distribuzione a posteriori per modelli complessi e consentono di decidere quali distribuzioni a priori e quali distribuzioni di verosimiglianza usare sulla base di considerazioni teoriche soltanto, senza dovere preoccuparsi di altri vincoli.[1] Prima di esaminare i metodi di stima della distribuzione a posteriori basati su simulazione numerica, esamineremo qui il caso più semplice, ovvero

[1] Dato che è basata su metodi computazionalmente intensivi, la stima numerica della funzione a posteriori può essere svolta soltanto mediante software. In anni recenti i metodi bayesiani di analisi dei dati sono diventati sempre più popolari proprio perché la potenza di calcolo necessaria per svolgere tali calcoli è ora alla portata di tutti. Questo non era vero solo pochi decenni fa.

quello nel quale è possibile fare inferenza su una proporzione senza dovere ricorrere ai metodi MCMC: se la distribuzione a priori su π è descritta da una distribuzione Beta allora, alla luce dei dati del campione, le proprietà della distribuzione a posteriori sisultano univocamente determinate – e sono facilmente descrivibili. Questa situazione definisce quello che viene chiamato il caso Beta-Binomiale.

1.3 Il modello Beta-Binomiale

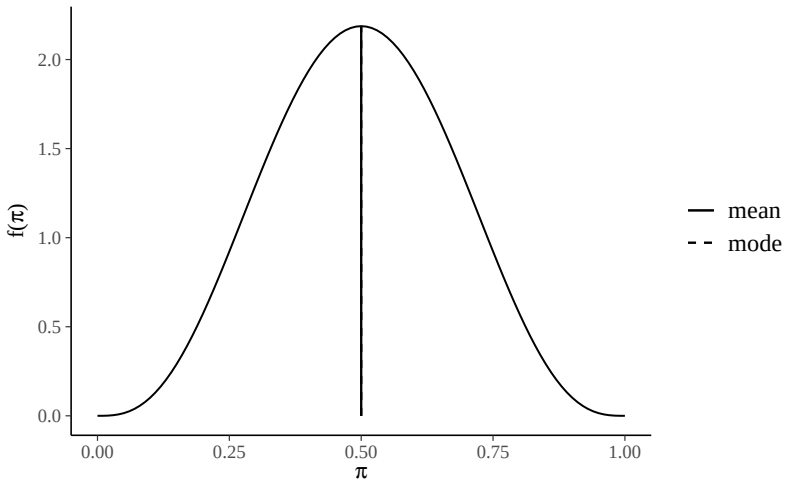
Per fare un esempio concreto, consideriamo nuovamente i dati di Zetsche et al. (2019): nel campione di 30 partecipanti clinici le aspettative future di 23 partecipanti risultano distorte negativamente e quelle di 7 partecipanti risultano distorte positivamente. Nel seguito, indicheremo con θ la probabilità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

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

Parametri della distribuzione Beta

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

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

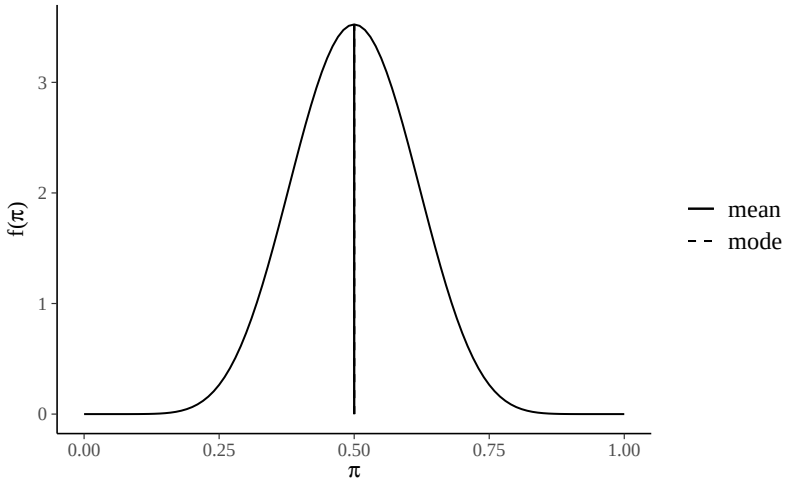


Possiamo quantificare la nostra incertezza calcolando, con un grado di fiducia del 95%, la regione nella quale, in base a tale credenza a priori, si trova il valore del parametro. Per ottenere tale intervallo di credibilità a priori, usiamo la funzione `qbeta()` di R. In `qbeta()` i parametri α e β sono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

```
qbeta(c(0.025, 0.975), shape1 = 4, shape2 = 4)
#> [1] 0.184 0.816
```

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

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



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

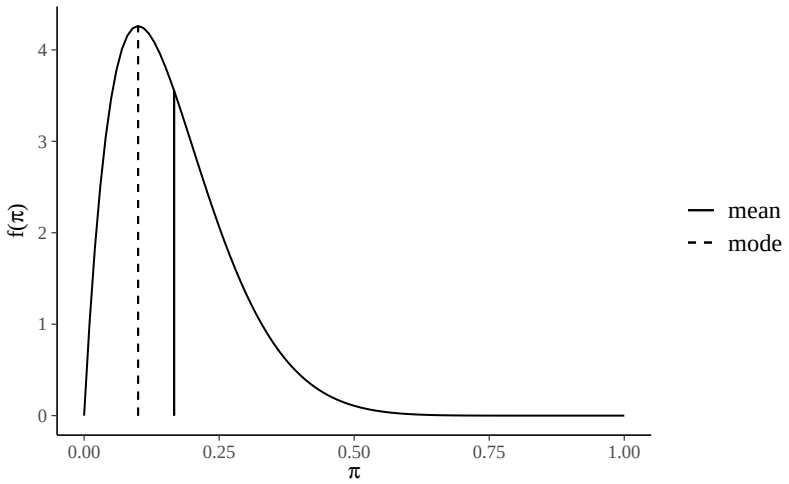
```
qbeta(c(0.025, 0.975), shape1 = 10, shape2 = 10)
#> [1] 0.289 0.711
```

Quale distribuzione a priori dobbiamo scegliere? In un problema concreto di analisi dei dati, la scelta della distribuzione a priori dipende dalle credenze a priori che vogliamo includere nell'analisi dei dati. Se non abbiamo alcuna informazione a priori, potremmo usare $\alpha = 1$ e $\beta = 1$, che produce una distribuzione a priori uniforme. Ma l'uso di distribuzioni a priori uniformi è sconsigliato per vari motivi, inclusa l'instabilità numerica della stima dei parametri. È meglio invece usare una distribuzione a priori poco informativa, come Beta$(2, 2)$.

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

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

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



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

La specificazione della distribuzione a posteriori

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

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

Nel caso presente abbiamo

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

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

Riscriviamo ora l'equazione precedente in termini generali

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

e raccogliendo tutte le costanti otteniamo:

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

Se ignoriamo il termine costante all'interno della parentesi quadra

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

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

Per ottenere una distribuzione di densità, dobbiamo aggiungere una costante di normalizzazione al kernel della distribuzione a posteriori. In base alla definizione della

distribuzione Beta, ed essendo $a' = a+y$ e $b' = b+n-y$, tale costante di normalizzazione sarà uguale a

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

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

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

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

Teorema 1.1. *Sia data la funzione di verosimiglianza* $\text{Bin}(n, y \mid \theta)$ *e sia* $\text{Beta}(\alpha, \beta)$ *una distribuzione a priori. In tali circostanze, la distribuzione a posteriori del parametro* θ *sarà una distribuzione* $\text{Beta}(\alpha+y, \beta+n-y)$.

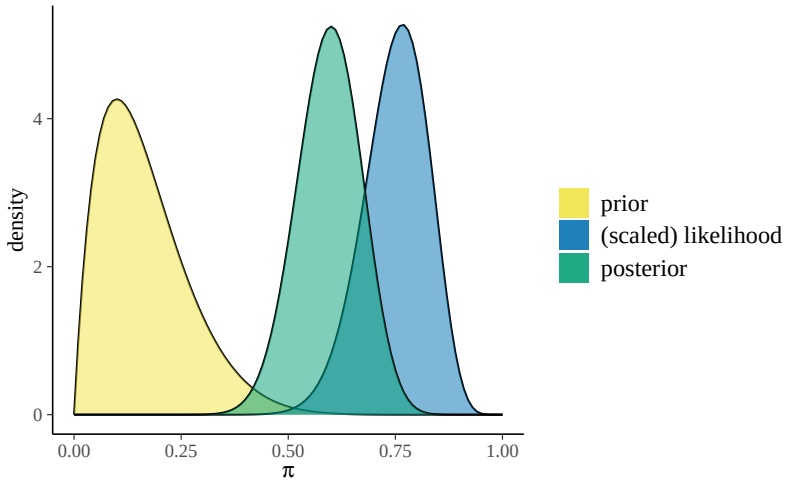
È facile calcolare il valore atteso a posteriori di θ. Essendo $\mathbb{E}[\text{Beta}(\alpha, \beta)] = \frac{\alpha}{\alpha+\beta}$, il risultato cercato diventa

$$\mathbb{E}_{\text{post}}[\text{Beta}(\alpha+y, \beta+n-y)] = \frac{\alpha+y}{\alpha+\beta+n}. \tag{1.1}$$

Esercizio 1.1. Usando le funzione R `plot_beta_binomial()` e `plot_beta_binomial()` del pacchetto `bayesrules`, si rappresenti in maniera grafica e si descriva in forma numerica l'aggiornamento bayesiano Beta-Binomiale per i dati di Zetsche et al. (2019).

Per i dati in discussione, abbiamo:

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



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

```
bayesrules:::summarize_beta_binomial(alpha = 2, beta = 10, y = 23, n = 30)
#>       model alpha beta  mean mode    var     sd
#> 1     prior     2   10 0.167  0.1 0.0107 0.1034
#> 2 posterior    25   17 0.595  0.6 0.0056 0.0749
```

Esercizio 1.2. Per i dati di Zetsche et al. (2019), si trovino la media, la moda, la deviazione standard della distribuzione a posteriori di θ. Si trovi inoltre l'intervallo di credibilità a posteriori del 95% per il parametro θ.

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

```
qbeta(c(0.025, 0.975), shape1 = 25, shape2 = 17)
#> [1] 0.445 0.737
```

La media della distribuzione a posteriori è

```
25 / (25 + 17)
#> [1] 0.595
```

La moda della distribuzione a posteriori è

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

La deviazione standard della distribuzione a priori è

```
sqrt((25 * 17) / ((25 + 17)^2 * (25 + 17 + 1)))
#> [1] 0.0749
```

Esercizio 1.3. Si trovino i parametri e le proprietà della distribuzione a posteriori del parametro θ per i dati dell'esempio relativo alla ricerca di Stanley Milgram discussa da Johnson et al. (2022).

Nel 1963, Stanley Milgram presentò una ricerca sulla propensione delle persone a obbedire agli ordini di figure di autorità, anche quando tali ordini possono danneggiare altre persone (Milgram, 1963). Nell'articolo, Milgram descrive lo studio come

> *consist[ing] of ordering a naive subject to administer electric shock to a victim. A simulated shock generator is used, with 30 clearly marked voltage levels that range from IS to 450 volts. The instrument bears verbal designations that range from Slight Shock to Danger: Severe Shock. The responses of the victim, who is a trained confederate of the experimenter, are standardized. The orders to administer shocks are given to the naive subject in the context of a 'learning experiment' ostensibly set up to study the effects of punishment on memory. As the experiment proceeds the naive subject is commanded to administer increasingly more intense shocks to the victim, even to the point of reaching the level marked Danger: Severe Shock.*

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

Johnson et al. (2022) fanno inferenza sui risultati dello studio di Milgram mediante il modello Beta-Binomiale. Il parametro di interesse è θ, la probabiltà che una persona obbedisca all'autorità (in questo caso, somministrando lo shock più severo), anche se ciò significa recare danno ad altri. Johnson et al. (2022) ipotizzano che, prima di raccogliere dati, le credenze di Milgram relative a θ possano essere rappresentate mediante una Beta$(1, 10)$. Sia $y = 26$ il numero di soggetti che, sui 40 partecipanti allo studio, aveva accettato di infliggere lo shock più severo. Assumendo che ogni partecipante si comporti indipendentemente dagli altri, possiamo modellare la dipendenza di y da θ

usando la distribuzione binomiale. Giungiamo dunque al seguente modello bayesiano Beta-Binomiale:

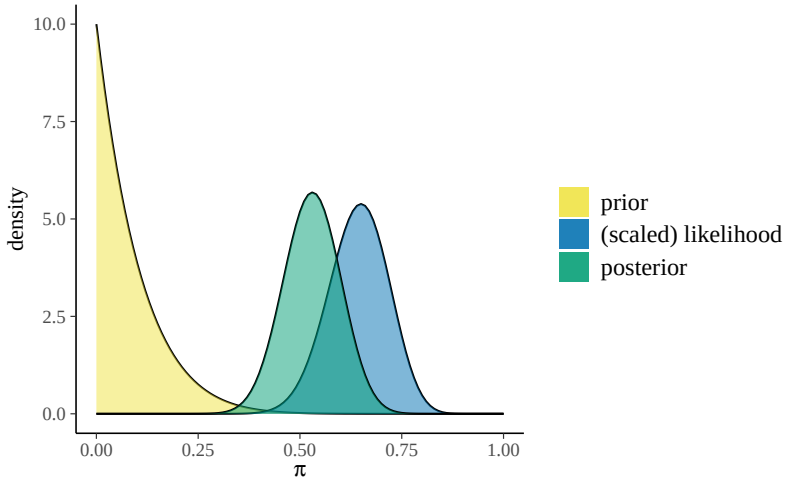
$$\begin{aligned} y \mid \theta &\sim \text{Bin}(n = 40, \theta) \\ \theta &\sim \text{Beta}(1, 10) \ . \end{aligned}$$

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

```
bayesrules:::summarize_beta_binomial(alpha = 1, beta = 10, y = 26, n = 40)
#>       model alpha beta   mean  mode     var     sd
#> 1     prior     1   10 0.0909 0.000 0.00689 0.0830
#> 2 posterior    27   24 0.5294 0.531 0.00479 0.0692
```

Il processo di aggiornamento bayesiano è descritto dalla figura seguente:

```
bayesrules:::plot_beta_binomial(alpha = 1, beta = 10, y = 26, n = 40)
```



1.4 Principali distribuzioni coniugate

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

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

Considerazioni conclusive

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

Bibliografia

Burger, E. B. & Starbird, M. (2012). *The 5 elements of effective thinking*. Princeton University Press. (Cit. a p. ix).

Horn, S. & Loewenstein, G. (2021). Underestimating Learning by Doing. *Available at SSRN 3941441* (cit. a p. x).

Johnson, A. A., Ott, M. & Dogucu, M. (2022). *Bayes Rules! An Introduction to Bayesian Modeling with R*. CRC Press. (Cit. a p. 7).

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378 (cit. a p. 7).

Zetsche, U., Bürkner, P.-C. & Renneberg, B. (2019). Future expectations in clinical depression: Biased or realistic? *Journal of Abnormal Psychology*, *128*(7), 678–688 (cit. alle pp. 3, 6, 7).

Elenco delle figure

Abstract This document contains the material of the lessons of Psicometria B000286 (2021/2022) aimed at students of the first year of the Degree Course in Psychological Sciences and Techniques of the University of Florence, Italy.

Keywords Data science, Bayesian statistics.