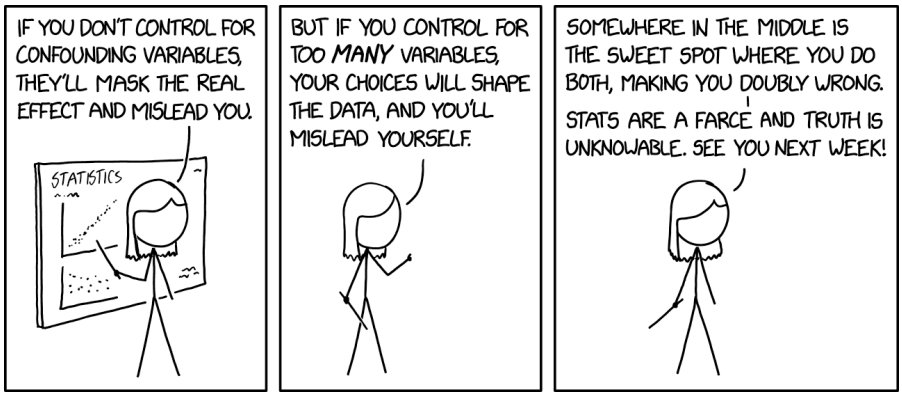
Corrado Caudek

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

Nozioni di base

1

Il calcolo delle probabilità

Una possibile definizione della teoria delle probabilità è la seguente: la teoria delle probabilità ci fornisce gli strumenti per prendere decisioni razionali in condizioni di incertezza, ovvero per formulare le migliori congetture possibili.

1.1 La probabilità come la logica della scienza

La figura 1.1 fornisce una rappresentazione schematica del processo dell'indagine scientifica. Possiamo pensare al progresso scientifico come alla ripetizione di questo ciclo, laddove i fenomeni naturali (e, ovviamente psicologici) vengono esplorati e i ricercatori imparano sempre di più sul loro funzionamento. Le caselle della figura descrivono le varie fasi del processo di ingagine scientifica, mentre lungo le frecce sono riportati i compiti che conducono i ricercatori da una fase alla successiva.

Consideriamo i compiti e le fasi dell'indagine scientifica. Iniziamo in basso a sinistra.

- *Invenzione e perfezionamento delle ipotesi.* In questa fase del processo scientifico, i ricercatori pensano ai fenomeni naturali, a ciò che è presente nella letteratura scientifica, ai risultati dei loro esperimenti, e formulano ipotesi o teorie che possono essere valutare mediante esperimenti empirici. Questo passaggio richiede innovazione e creatività.
- L'*inferenza deduttiva* procede in maniera deterministica dai fatti alle conclusioni. Ad esempio, se dico che tutti gli uomini sono mortali e che Socrate è un uomo, allora posso concludere deduttivamente che Socrate è mortale. Quando i ricercatori progettano gli esperimenti in base alle teorie, usano la logica deduttiva per dire: "Se A è vero, allora

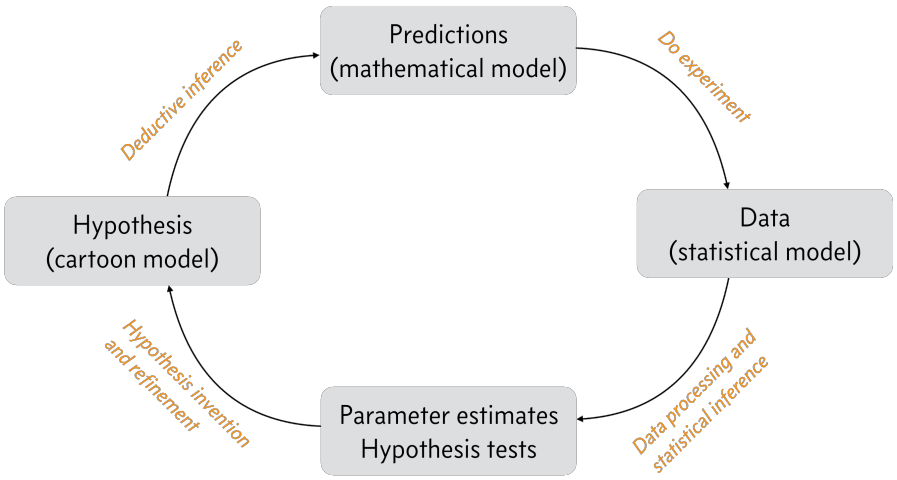


Figura 1.1: *Rappresentazione schematica del processo scientifico (figura adattata dalla Fig. 1.1 di P. Gregory, Bayesian Logical Data Analysis for the Physical Sciences, Cambridge, 2005).*

B deve essere vero", dove A è l'ipotesi teorica e B è l'osservazione sperimentale.

- *Esecuzione degli esperimenti.* Questa fase richiede molte risorse (tempo e denaro). Richiede anche innovazione e creatività. Nello specifico, i ricercatori devono pensare attentamente a come costruire l'esperimento necessario per verificare la teoria di interesse. Quale risultato dell'esperimento si ottengono i dati.
- L'*inferenza induttiva* 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. Supponiamo di avere osservato B. Questo rende A vero? Non necessariamente. Ma può rendere A più plausibile. Questo è un sillogismo debole. Ad esempio, si consideri la seguente coppia ipotesi/osservazioni.
  - A = L'iniezione di acque reflue dopo la fratturazione idraulica, nota come fracking, può portare a una maggiore frequenza di terremoti.
  - B = La frequenza dei terremoti in Oklahoma è aumentata di 100 volte dal 2010, quando il fracking è diventato una pratica comune.

  - Poiché B è stato osservato, A è più plausibile. A non è necessariamente vero, ma è più plausibile.

- L'*inferenza statistica* è un tipo di inferenza induttiva che è specificamente formulata come un problema inverso. L'inferenza statistica è quell'insieme di procedure che hanno lo scopo di quantificare quanto più plausibile sia A dopo aver osservato B. Per svolgere l'inferenza statistica è dunque necessario quantificare tale plausibilità. Lo strumento che ci consente di fare questo è la teoria delle probabilità.

L'inferenza statistica è l'aspetto del processo dell'indagine scientifica che costituisce il tema centrale di questo insegnamento. Il risultato dell'inferenza statistica è la conoscenza di quanto siano plausibili le ipotesi e le stime dei parametri sotto le ipotesi considerate. Ma l'inferenza statistica richiede una teoria delle probabilità, laddove la teoria delle probabilità può essere vista come una generalizzazione della logica. A causa di questa connessione con la logica, e del suo ruolo cruciale nella scienza, E. T. Jaynes ha dichiarato che "la probabilità è la logica della scienza". Per potere trattare i temi di base dell'inferenza statistica è dunque necessario esaminare preliminarmente alcune nozioni della teoria delle probabilità.

1.2 Che cos'è la probabilità?

La definizione della probabilità è un problema estremamente dibattuto ed aperto. Sono state fornite due possibili soluzioni al problema di definire il concetto di probabilità.

(a) La natura della probabilità è "ontologica" (ovvero, basata sulla metafisica): la probabilità è una proprietà della della realtà, del mondo, di come sono le cose, indipendentemente dalla nostra esperienza. È una visione che qualcuno chiama "oggettiva".

(b) La natura della probabilità è "epistemica" (ovvero, basata sulla conoscenza): la probabilità si riferisce alla conoscenza che abbiamo del mondo, non al mondo in sé. Di conseguenza è detta, in contrapposizione alla precedente definizione, "soggettiva".

In termini epistemici, la probabilità fornisce una misura della nostra incertezza sul verificarsi di un fenomeno, alla luce delle informazioni di-

sponibili. Potremmo dire che c'è una "scala" naturale che ha per estremi il vero (1: evento certo) da una parte ed il falso (0: evento impossibile) dall'altra. La probabilità è la quantificazione di questa scala: quantifica lo stato della nostra incertezza rispetto al contenuto di verità di una proposizione (ovvero, quantifica la plausibilità di una proposizione).

- Nell'interpretazione frequentista della probabilità, la probabilità $P(A)$ rappresenta la frequenza relativa a lungo termine nel caso di un grande numero di ripetizioni di un esperimento casuale sotto le medesime condizioni. L'evento A deve essere una proposizione relativa alle variabili casuali[1].
- Nell'interpretazione bayesiana della probabilità $P(A)$ rappresenta il grado di credenza, o plausibilità, a proposito di A, dove A può essere qualsiasi proposizione logica.

In questo insegnamento utilizzeremo l'interpretazione bayesiana della probabilità. Possiamo citare De Finetti, ad esempio, il quale ha formulato la seguente definizione "soggettiva" di probabilità la quale risulta applicabile anche ad esperimenti casuali i cui eventi elementari non siano ritenuti ugualmente possibili e che non siano necessariamente ripetibili più volte sotto le stesse condizioni:

Definizione 1.1. La probabilità di un evento E è la quota $p(E)$ che un individuo reputa di dover pagare ad un banco per ricevere "1" ovvero "0" verificandosi o non verificandosi E. Le valutazioni di probabilità degli eventi devono rispondere ai pricipi di equità e coerenza.

I principi di equità e coerenza sono definiti come segue.

Definizione 1.2. Una scommessa risponde ai pricipi di *equità* se il ruolo di banco e giocatore sono scambiabili in ogni momento del gioco e sempre alle stesse condizioni; *coerenza* se non vi sono combinazioni di scommesse che consentano (sia al banco che al giocatore) di realizzare perdite o vincite certe.

[1] Viene stressata qui l'idea che ciò di cui parliamo è qualcosa che emerge nel momento in cui è possibile ripetere l'esperimento casuale tante volte sotto le medesime condizioni. Le variabili casuali, infatti, forniscono una quantificazione dei risultati che si ottengono ripetendo tante volte l'esperimento casuale sotto le medesime condizioni.

Secondo de Finetti (1931), *"nessuna scienza ci permetterà di dire: il tale fatto accadrà, andrà così e così, perché ciò è conseguenza di tale legge, e tale legge è una verità assoluta, ma tanto meno ci condurrà a concludere scetticamente: la verità assoluta non esiste, e quindi tale fatto può accadere e può non accadere, può andare così e può andare in tutt'altro modo, nulla io ne so. Quel che si potrà dire è questo: io prevedo che il tale fatto avverrà, e avverrà nel tal modo, perché l'esperienza del passato e l'elaborazione scientifica cui il pensiero dell'uomo l'ha sottoposta mi fanno sembrare ragionevole questa previsione."*

In altri termini, de Finetti ritiene che la probabilità debba essere concepita non come una proprietà "oggettiva" dei fenomeni ("la probabilità di un fenomeno ha un valore determinato che dobbiamo solo scoprire"), ma bensì come il "grado di fiducia – in inglese *degree of belief* – di un dato soggetto, in un dato istante e con un dato insieme d'informazioni, riguardo al verificarsi di un evento". Per denotare sia la probabilità (soggettiva) di un evento sia il concetto di *valore atteso* (che descriveremo in seguito), de Finetti (1970) utilizza il termine "previsione" (e lo stesso simbolo P): *"la previsione [...] consiste nel considerare ponderatamente tutte le alternative possibili per ripartire fra di esse nel modo che parrà più appropriato le proprie aspettative, le proprie sensazioni di probabilità."*

1.3 Variabili casuali e probabilità di un evento

Esaminiamo qui di seguito alcuni concetti di base della teoria delle probabilità.

1.3.1 Variabili casuali

Sia Y il risultato del lancio di moneta equilibrata, non di un generico lancio di una moneta, ma un'istanza specifica del lancio di una specifica moneta in un dato momento. Definita in questo modo, Y è una *variabile casuale*, ovvero una variabile che assume valori diversi con probabilità diverse. Se la moneta è equilibrata, c'è una probabilità del 50% che il lancio della moneta dia come risultato "testa" e una probabilità del 50% che dia come risultato "croce".

Per facilitare la trattazione, le variabili casuali assumono solo valori nu-

merici. Per lo specifico lancio della moneta in questione, diciamo, ad esempio, che la variabile casuale Y assume il valore 1 se esce testa e il valore 0 se esce croce.

1.3.2 Eventi e probabilità

Nella teoria delle probabilità il risultato "testa" nel lancio di una moneta è chiamato *evento*.[2] Ad esempio, $Y = 1$ denota l'evento in cui il lancio di una moneta produce come risultato testa.

Il funzionale $Pr[\cdot]$ definisce la probabilità di un evento. Ad esempio, per il lancio di una moneta equilibrata, la probabilità dell'evento "il risultato del lancio della moneta è testa" è scritta come

$$Pr[Y = 1] = 0.5.$$

Se la moneta è equilibrata dobbiamo anche avere $Pr[Y = 0] = 0.5$. I due eventi $Y = 1$ e $Y = 0$ sono *mutuamente esclusivi* nel senso che non possono entrambi verificarsi contemporaneamente. Nella notazione probabilistica,

$$Pr[Y = 1 \; e \; Y = 0] = 0.$$

Gli eventi $Y = 1$ e $Y = 0$ di dicono *esaustivi*, nel senso che almeno uno di essi deve verificarsi e nessun altro tipo di evento è possibile. Nella notazione probabilistica,

$$Pr[Y = 1 \; o \; Y = 0] = 1.$$

Il connettivo logico "e" specifica eventi *congiunti*, ovvero eventi che possono verificarsi contemporaneamente (eventi *compatibili*) e per i quali, perciò, la probabilità della loro congiunzione è $Pr(A \; e \; B) > 0$. Il connettivo logico "o" specifica eventi *disgiunti*, ovvero eventi che non possono verificarsi contemporaneamente (eventi *incompatibili*) e per i quali, perciò, la probabilità della loro congiunzione è $P(A \; e \; B) = 0$.

[2] Per un ripasso delle nozioni di base della teoria degli insiemi, si veda l'Appendice **??**.

1.4 Spazio campionario e risultati possibili

Anche se il lancio di una moneta produce sempre uno specifico risultato nel mondo reale, noi possiamo anche immaginare i possibili risultati alternativi che si sarebbero potuti osservare. Quindi, anche se in uno specifico lancio la moneta dà testa ($Y = 1$), possiamo immaginare la possibilità che il lancio possa avere prodotto croce ($Y = 0$). Tale ragionamento controfattuale è la chiave per comprendere la teoria delle probabilità e l'inferenza statistica.

I risultati possibili che si possono osservare come conseguenza del lancio di una moneta determinano i valori possibili che la variabile casuale può assumere. L'insieme di tutti i risultati possibili è chiamato *spazio campionario*. Lo spazio campionario può essere concettualizzato come un'urna contenente una pallina per ogni possibile risultato del lancio della moneta. Su ogni pallina è scritto il valore della variabile casuale. Uno specifico lancio di una moneta – ovvero, l'osservazione di uno specifico valore di una variabile casuale – è chiamato *esperimento casuale*.

Il lancio di un dado ci fornisce l'esempio di un altro esperimento casuale. Supponiamo di essere interessati all'evento "il lancio del dado produce un numero dispari". Un *evento* seleziona un sottoinsieme dello spazio campionario: in questo caso, l'insieme dei risultati $\{1, 3, 5\}$. Se esce 3, per esempio, diciamo che si è verificato l'evento "dispari" (ma l'evento "dispari" si sarebbe anche verificato anche se fosse uscito 1 o 5).

1.5 Usare la simulazione per stimare le probabilità

I metodi basati sulla simulazione ci consentono di stimare le probabilità degli eventi in un modo diretto se siamo in grado di generare realizzazioni molteplici e casuali delle variabili casuali coinvolte nelle definizioni degli eventi. Per simulare il lancio di una moneta equilibrata in R iniziamo a definire un vettore che contiene i possibili risultati del lancio della moneta (ovvero i possibili valori della variabile casuale Y):

```
coin <- c(0, 1)
```

L'estrazione casuale di uno di questi due possibili valori (ovvero, la simulazione di uno specifico lancio di una moneta) si realizza con la funzione `sample()`:

```
sample(coin, size = 1)
#> [1] 0
```

In maniera equivalente, lo stesso risultato si ottiene mediante l'istruzione

```
rbinom(1, 1, 0.5)
#> [1] 1
```

Supponiamo di ripetere questo esperimento casuale 100 volte e di registrare i risultati così ottenuti. La stima della probabilità dell'evento $Pr[Y = 1]$ è data dalla frequenza relativa del numero di volte in cui abbiamo osservato l'evento di interesse $(Y = 1)$:

```
M <- 10
y <- rep(NA, M)
for (m in 1:M) {
  y[m] = rbinom(1, 1, 0.5)
}
estimate = sum(y) / M

cat("estimated Pr[Y = 1] =", estimate)
#> estimated Pr[Y = 1] = 0.5
```

Ripetiamo questa procedura 10 volte.

```
flip_coin <- function(M) {
  y <- rep(NA, M)
  for (m in 1:M) {
    y[m] = rbinom(1, 1, 0.5)
  }
  estimate <- sum(y) / M
```

```
  cat("estimated Pr[Y = 1] =", estimate, "\n")
}
```

```
for(i in 1:10) {
  flip_coin(10)
}
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.3
#> estimated Pr[Y = 1] = 0.7
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.6
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.8
#> estimated Pr[Y = 1] = 0.4
#> estimated Pr[Y = 1] = 0.5
```

Dato che la moneta è equilibrata, la stima delle probabilità dell'evento $Pr[Y = 1]$ è simile a al valore che ci aspettiamo ($Pr[Y = 1] = 0.5$), ma il risultato ottenuto nelle varie simulazioni non è sempre esatto. Proviamo ad aumentare il numero di lanci in ciascuna simulazione:

```
for(i in 1:10) {
  flip_coin(100)
}
#> estimated Pr[Y = 1] = 0.44
#> estimated Pr[Y = 1] = 0.53
#> estimated Pr[Y = 1] = 0.43
#> estimated Pr[Y = 1] = 0.58
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.41
#> estimated Pr[Y = 1] = 0.51
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.5
#> estimated Pr[Y = 1] = 0.57
```

In questo secondo caso, gli errori tendono ad essere più piccoli della simulazione precedente. Cosa succede se in ciascuna simulazione esaminiamo

i risultati di 10,000 lanci della moneta?

```
for(i in 1:10) {
  flip_coin(1e4)
}
#> estimated Pr[Y = 1] = 0.5029
#> estimated Pr[Y = 1] = 0.4886
#> estimated Pr[Y = 1] = 0.4956
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.5032
#> estimated Pr[Y = 1] = 0.5051
#> estimated Pr[Y = 1] = 0.4928
#> estimated Pr[Y = 1] = 0.4968
#> estimated Pr[Y = 1] = 0.4991
#> estimated Pr[Y = 1] = 0.4976
```

Ora le stime ottenute sono molto vicine alla vera probabilità che vogliamo stimare (cioè 0.5, perché la moneta è equilibrata). I risultati delle simulazioni precedenti pongono dunque il problema di determinare quale sia il numero di lanci di cui abbiamo bisogno per assicurarci che le stime siano accurate (ovvero, vicine al valore corretto della probabilità)

1.6 La legge dei grandi numeri

La visualizzazione mediante grafici contribuisce alla comprensione dei concetti della statistica e della teoria delle probabilità. Un modo per descrivere ciò che accade all'aumentare del numero M di ripetizioni del lancio della moneta consiste nel registrare la stima della probabilità dell'evento $Pr[Y = 1]$ in funzione del numero di ripetizioni dell'esperimento casuale per ogni $m \in 1 : M$. Un grafico dell'andamento della stima di $Pr[Y = 1]$ in funzione di m si ottiene nel modo seguente.

```
nrep <- 1e4
estimate <- rep(NA, nrep)
flip_coin <- function(m) {
  y <- rbinom(m, 1, 0.5)
```

```
  phat <- sum(y) / m
  phat
}
for(i in 1:nrep) {
  estimate[i] <- flip_coin(i)
}
d <- data.frame(
  n = 1:nrep,
  estimate
)
d %>%
  ggplot(
    aes(x = n, y = estimate)
  ) +
  geom_line() +
  theme(legend.title = element_blank()) +
  labs(
    x = "Numero di lanci della moneta",
    y = "Stima Pr[Y = 1]"
)
```

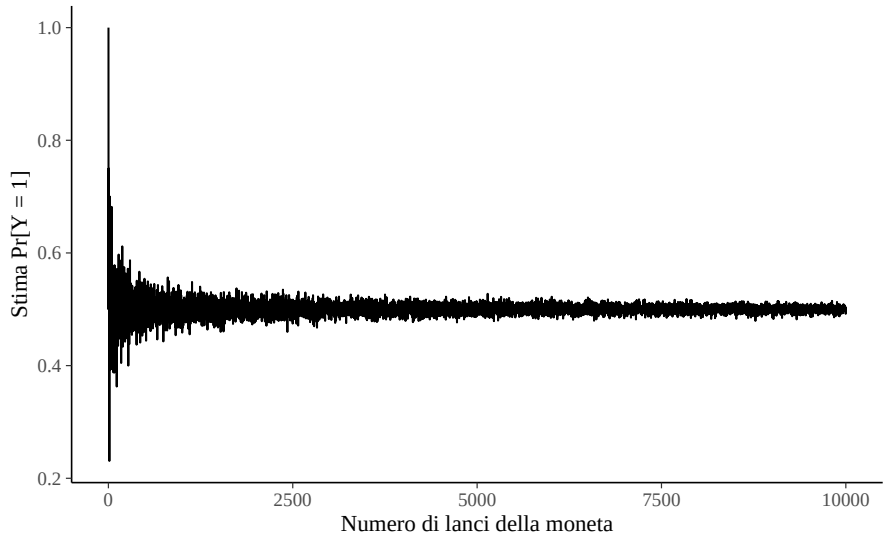


Figura 1.2: *Stima della probabilità di successo in funzione del numero di lanci di una moneta.*

Dato che il grafico 1.2 su una scala lineare non rivela chiaramente l'andamento della simulazione, utilizzeremo invece un grafico in cui sull'asse x è stata imposta una scala logaritmica. Con l'asse x su scala logaritmica, i valori tra 1 e 10 vengono tracciati all'incirca con la stessa ampiezza come nel caso dei valori tra 50 e 700, eccetera.

```
d %>%
  ggplot(
    aes(x = n, y = estimate)
  ) +
  geom_line() +
  scale_x_log10(
    breaks = c(1, 3, 10, 50, 200,
               700, 2500, 10000)
  ) +
  theme(legend.title = element_blank()) +
  labs(
    x = "Numero di lanci della moneta",
    y = "Stima Pr[Y = 1]"
)
```

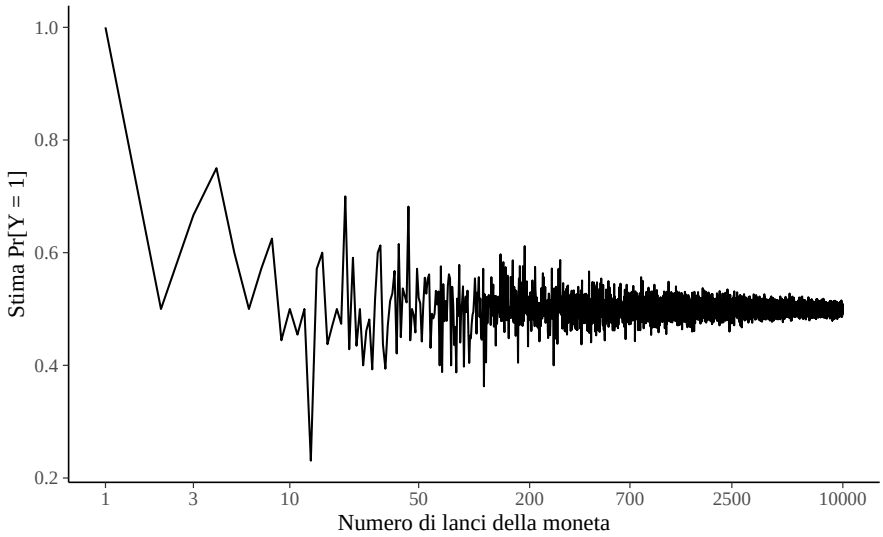


Figura 1.3: *Stima della probabilità di successo in funzione del numero di lanci di una moneta – scala logaritmica.*

La *legge dei grandi numeri* ci dice che all'aumentare del numero di ripetizioni dell'esperimento casuale la media dei risultati ottenuti tenderà ad avvicinarsi al valore atteso man mano che verranno eseguite più prove. Nel caso presente, la figura 1.3 mostra appunto che, all'aumentare del numero M di lanci della moneta, la stima di $Pr[Y = 1]$ tende a convergere al vero valore di 0.5.

1.7 Variabili casuali multiple

Le variabili casuali non esistono isolatamente. Abbiamo iniziato con una singola variabile casuale Y che rappresenta il risultato di un singolo, specifico lancio di una moneta equlibrata. Ma supponiamo ora di lanciare la moneta tre volte. Ciò suggerisce che possiamo avere le variabili casuali Y_1, Y_2, Y_3 che rappresentano i risultati di ciascuno dei lanci. Possiamo assumere che ogni lancio sia indipendente, ovvero che non dipenda dal risultato degli altri lanci. Ognuna di queste variabili Y_n per $n \in 1:3$ ha $Pr[Y_n = 1] = 0.5$ e $Pr[Y_n = 0] = 0.5$. Possiamo combinare più variabili casuali usando le operazioni aritmetiche. Se Y_1, Y_2, Y_3 sono variabili casuali che rappresentano tre lanci di una moneta equilibrata (o un lancio di tre monete equilibrate), possiamo definire la somma di tali variabili casuali come

$$Z = Y_1 + Y_2 + Y_3.$$

Possiamo simulare i valori assunti dalla variabile casuale Z simulando i valori di Y_1, Y_2, Y_3 per poi sommarli.

```
y1 <- rbinom(1, 1, 0.5)
y2 <- rbinom(1, 1, 0.5)
y3 <- rbinom(1, 1, 0.5)
c(y1, y2, y3)
#> [1] 0 0 1
z <- sum(c(y1, y2, y3))
cat("z =", z, "\n")
#> z = 1
```

ovvero,

```
y <- rep(NA, 3)
for (i in 1:3) {
  y[i] <- rbinom(1, 1, 0.5)
}
y
#> [1] 1 0 0
z <- sum(y)
cat("z =", z, "\n")
#> z = 1
```

oppure, ancora più semplicemente:

```
y <- rbinom(3, 1, 0.5)
y
#> [1] 0 1 1
z <- sum(y)
cat("z =", z, "\n")
#> z = 2
```

Possiamo ripetere questa simulazione $M = 1e5$ volte:

```
M <- 1e5
z <- rep(NA, M)
for(i in 1:M) {
  y <- rbinom(3, 1, 0.5)
  z[i] <- sum(y)
}
```

e calcolare una stima della probabilità che la variabile casuale Z assuma i valori 0, 1, 2, 3:

```
table(z) / M
#> z
#>      0      1      2      3
#> 0.1256 0.3750 0.3749 0.1245
```

Nel caso di 4 monete equilibrate, avremo:

```
M <- 1e5
z <- rep(NA, M)
for(i in 1:M) {
  y <- rbinom(4, 1, 0.5)
  z[i] <- sum(y)
}
table(z) / M
#> z
#>       0       1       2       3       4
#> 0.06213 0.25019 0.37400 0.25097 0.06271
```

Viene detta *variabile casuale discreta* una variabile casuale le cui modalità possono essere costituite solo da numeri interi:

$$\mathbb{Z} = \dots, -2, -1, 0, 1, 2, \dots$$

1.8 Funzione di massa di probabilità

È conveniente avere una funzione che associa ogni possibile valore di una variabile casuale alla sua probabilità. In generale, ciò è possibile se e solo se la variabile casuale è discreta, così com'è stata definita nel Paragrafo precedente.

Ad esempio, se consideriamo $Z = Y_1 + \cdots + Y_4$ come il numero di risultati "testa" in 4 lanci della moneta, allora possiamo definire la seguente funzione:

$$\begin{array}{lcll} p_Z(0) & = & 1/16 & \text{TTTT} \\ p_Z(1) & = & 4/16 & \text{HTTT, THTT, TTHT, TTTH} \\ p_Z(2) & = & 6/16 & \text{HHTT, HTHT, HTTH, THHT, THTH, TTTH} \\ p_Z(3) & = & 4/16 & \text{HHHT, HHTH, HTHH, THHH} \\ p_Z(4) & = & 1/16 & \text{HHHH} \end{array}$$

Il lancio di quattro monete può produrre sedici possibili risultati. Dato che i lanci sono indipendenti e le monete sono equilibrate, ogni possibile risultato è ugualmente probabile. Nella tabella in alto, le sequenze dei risultati possibili del lancio delle 4 monete sono riportate nella colonna

più a destra. Le probabilità si ottengono dividendo il numero di sequenze che producono lo stesso numero di eventi testa per il numero dei risultati possibili.

La funzione p_Z è stata costruita per mappare un valore u per Z alla probabilità dell'evento $Z = u$. Convenzionalmente, queste probabilità sono scritte come

$$p_Z(z) = \Pr[Z = z].$$

La parte a destra dell'uguale si può leggere come: "la probabilità che la variabile casuale Z assuma il valore z".

Una funzione definita come sopra è detta *funzione di massa di probabilità* della variabile casuale Z. Ad ogni variabile casuale discreta è associata un'unica funzione di massa di probabilità.

Una rappresentazione grafica della stima della funzione di massa di probabilità per l'esperimento casuale del lancio di quattro monete equilibrate è fornita nella figura 1.4.

```
set.seed(1234)
M <- 1e5
nflips <- 4
u <- rbinom(M, nflips, 0.5)
x <- 0:nflips
y <- rep(NA, nflips+1)
for (n in 0:nflips)
  y[n + 1] <- sum(u == n) / M
bar_plot <-
  data.frame(Z = x, count = y) %>%
  ggplot(
    aes(x = Z, y = count)
  ) +
  geom_bar(stat = "identity") +
  scale_x_continuous(
    breaks = 0:4,
    labels = c(0, 1, 2, 3, 4)
  ) +
  labs(
```

```
    y = "Probabilità stimata Pr[Z = z]"
)
bar_plot
```

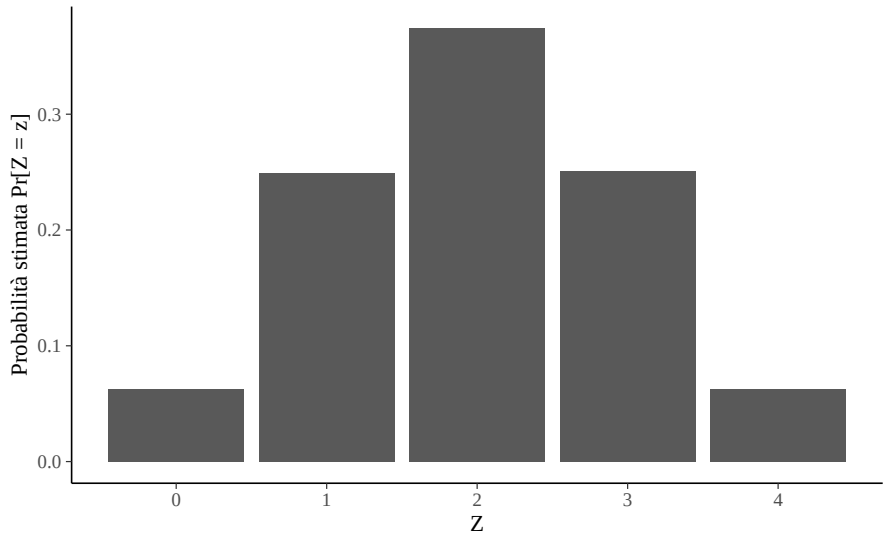


Figura 1.4: *Grafico di $M = 100\,000$ simulazioni della funzione di massa di probabilità di una variabile casuale definita come il numero di teste in quattro lanci di una moneta equilibrata.*

Se A è un sottoinsieme della variabile casuale Z, allora denotiamo con $P_z(A)$ la probabilità assegnata ad A dalla distribuzione P_z. Mediante una distribuzione di probabilità P_z è dunque possibile determinare la probabilità di ciascun sottoinsieme $A \subset Z$ come

$$P_z(A) = \sum_{z \in A} P_z(Z).$$

Esempio 1.1. Nel caso dell'esempio discusso nella Sezione 1.8, la probabilità che la variabile casuale Z sia un numero dispari è $Pr(\text{Z è un numero dispari}) = P_z(Z = 1) + P_z(Z = 3) = \frac{4}{16} + \frac{4}{16} = \frac{1}{2}$.

Considerazioni conclusive

In questo capitolo abbiamo visto come si costruisce lo spazio campionario di un esperimento casuale, quali sono le proprietà di base della probabilità e come si assegnano le probabilità agli eventi definiti sopra uno spazio campionario discreto. Abbiamo anche introdotto le nozioni di "variabile casuale", ovvero di una variabile che prende i suoi valori casualmente. E abbiamo descritto il modo di specificare la probabilità con cui sono presi i differenti valori, ovvero la funzione di distribuzione probabilistica $F(X) = Pr(X < x)$, e la funzione di massa di probabilità. Le procedure di analisi dei dati psicologici che discuteremo in seguito faranno un grande uso di questi concetti e della notazione qui introdotta.

2

Modello lineare in Stan

Obiettivo di questo Capitolo è illustrare come svolgere l'analisi bayesiana del modello lineare usando il linguaggio Stan.[1]

2.1 Il modello lineare in linguaggio Stan

Leggiamo in R il dataset `kidiq`:

```
library("rio")
df <- rio::import(here::here("data", "kidiq.dta"))
head(df)
#>   kid_score mom_hs mom_iq mom_work mom_age
#> 1        65      1 121.12        4      27
#> 2        98      1  89.36        4      25
#> 3        85      1 115.44        4      27
#> 4        83      1  99.45        3      25
#> 5       115      1  92.75        4      27
#> 6        98      0 107.90        1      18
```

Vogliamo descrivere l'associazione tra il QI dei figli e il QI delle madri mediante un modello lineare. Per farci un'idea del valore dei parametri, adattiamo il modello lineare ai dati mediante la procedura di massima verosimiglianza:

```
summary(lm(kid_score ~ mom_iq, data = df))
#>
#> Call:
```

[1] Una descrizione dell'approccio frequentista è fornita nell'Appendice **??**.

```
#> lm(formula = kid_score ~ mom_iq, data = df)
#>
#> Residuals:
#>    Min     1Q Median     3Q    Max
#> -56.75 -12.07   2.22  11.71  47.69
#>
#> Coefficients:
#>             Estimate Std. Error t value Pr(>|t|)
#> (Intercept)  25.7998     5.9174    4.36  1.6e-05 ***
#> mom_iq        0.6100     0.0585   10.42  < 2e-16 ***
#> ---
#> Signif. codes:
#> 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#>
#> Residual standard error: 18.3 on 432 degrees of freedom
#> Multiple R-squared:  0.201,  Adjusted R-squared:  0.199
#> F-statistic:  109 on 1 and 432 DF,  p-value: <2e-16
```

La formulazione bayesiana del modello lineare è:

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\mu_i, \sigma) \\
\mu_i &= \beta_0 + \beta_1 x_i \\
\beta_0 &\sim \mathcal{N}(25, 10) \\
\beta_1 &\sim \mathcal{N}(0, 1) \\
\sigma &\sim \text{Cauchy}(18, 5)
\end{aligned}
$$

La prima riga definisce la funzione di verosimiglianza e le righe successive definiscono le distribuzioni a priori dei parametri. Il segno $\sim$ (tilde) si può leggere "si distribuisce come". La prima riga ci dice che ciascuna osservazione y_i è una variabile casuale che segue la distribuzione gaussiana di parametri μ_i e σ. La seconda riga specifica, in maniera deterministica, che ciascun μ_i è una funzione lineare di x_i, con parametri β_0 e β_1. Le due righe successive specificano le distribuzioni a priori per β_0 e β_1. La distribuzione a priori di β_0 è una distribuzione gaussiana di parametri $\mu_\alpha = 25$ e deviazione standard $\sigma_\alpha = 10$; la distribuzione a priori di β_1 è una distribuzione gaussiana standardizzata. L'ultima riga definisce la distribuzione a priori di σ, ovvero una Cauchy di parametri 18 e 5.

Poniamoci ora il problema di specificare il modello bayesiano descritto sopra in linguaggio Stan[2]. Il codice Stan viene eseguito più velocemente se l'input è standardizzato così da avere una media pari a zero e una varianza unitaria.[4] Ponendo $y = (y_1, \dots, y_n)$ e $x = (x_1, \dots, x_n)$, il modello lineare può essere scritto come

$$y_i = \alpha + \beta x_i + \varepsilon_i,$$

dove

$$\varepsilon_i \sim \mathcal{N}(0, \sigma).$$

Seguendo la notazione del manuale Stan, i parametri del modello lineare sono qui denotati da α e β.

È necessario prima centrare i dati, sottraendo da essi la media campionaria, per poi scalarli dividendo per la deviazione standard campionaria. Una singola osservazione u viene standardizzata dalla funzione z definita da

$$z_y(u) = \frac{u - \bar{y}}{\texttt{sd}(y)}$$

dove la media $\bar{y}$ è

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

e la deviazione standard è

[2]Nella discussione che segue ripeto pari pari ciò che è riportato nel manuale del linguaggio Stan[3].

[4]Si noti un punto importante. Il fatto di standardizzare i dati fa in modo che le distribuzioni a priori sui parametri vadano espresse sulla scala delle v.c. normali standardizzate. Se centriamo sullo 0 tali distribuzioni a priori, con una deviazione standard dell'ordine di grandezza dell'unità, i discorsi sull'arbitrarietà delle distribuzioni a priori perdono di significato: nel caso di dati standardizzati le distribuzioni a priori formulate come indicato sopra sono distribuzioni debolmente informative il cui unico scopo è la regolarizzazione dei dati, ovvero di mantenere le inferenze in una gamma ragionevole di valori; ciò contribuisce nel contempo a limitare l'influenza eccessiva delle osservazioni estreme (valori anomali) — quello che è importante è che tali distribuzioni a priori non introducono alcuna distorsione sistematica nella stima a posteriori.

$$\texttt{sd} = \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \right)^{-\frac{1}{2}}.$$

La trasformata inversa è definita invertendo i due passaggi precedenti: la deviazione standard è usata per scalare i valori u e la media campionaria è usata per traslare la distribuzione dei valori u scalati:

$$z_y^{-1}(u) = \texttt{sd}(y)u + \bar{y}.$$

Consideriamo il seguente modello iniziale in linguaggio Stan:

```
modelString = "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
parameters {
  real alpha;
  real beta;
  real<lower=0> sigma;
}
model {
  // priors
  alpha ~ normal(25, 10);
  beta ~ normal(0, 1);
  sigma ~ cauchy(18, 5);
  // likelihood
  for (n in 1:N)
    y[n] ~ normal(alpha + beta * x[n], sigma);
}
"
writeLines(modelString, con = "code/simpleregkidiq.stan")
```

La funzione `modelString()` registra una stringa di testo mentre `writeLines()` crea un file nell'indirizzo specificato. Tale file deve avere l'estensione `.stan`.

Modificando il codice precedente otteniamo il modello Stan per dati standardizzati. Il blocco `data` è identico a quello del caso precedente. I predittori e la risposta standardizzati sono definiti nel blocco `transformed data`. Per semplificare la notazione (e velocizzare l'esecuzione), nel blocco `model` l'istruzione di campionamento è espressa in forma vettorializzata: `y_std ~ normal(alpha_std + beta_std * x_std, sigma_std);`.

```
modelString = "
data {
  int<lower=0> N;
  vector[N] y;
  vector[N] x;
}
transformed data {
  vector[N] x_std;
  vector[N] y_std;
  x_std = (x - mean(x)) / sd(x);
  y_std = (y - mean(y)) / sd(y);
}
parameters {
  real alpha_std;
  real beta_std;
  real<lower=0> sigma_std;
}
transformed parameters {
  vector[N] mu_std = alpha_std + beta_std * x_std;
}
model {
  alpha_std ~ normal(0, 1);
  beta_std ~ normal(0, 1);
  sigma_std ~ normal(0, 1);
  y_std ~ normal(mu_std, sigma_std);
}
generated quantities {
  // transform to the original data scale
  real alpha;
  real beta;
  real<lower=0> sigma;
  alpha = sd(y) * (alpha_std - beta_std * mean(x) / sd(x)) + mean(y);
```

```
  beta = beta_std * sd(y) / sd(x);
  sigma = sd(y) * sigma_std;
}
"
writeLines(modelString, con = "code/simpleregstd.stan")
```

Si noti che i parametri vengono rinominati per indicare che non sono i parametri "naturali", ma per il resto il modello è identico. Sono qui utilizzate distribuzioni a priori debolmente informative per i parametri `alpha` e `beta`.

I valori dei parametri sulla scala originale dei dati vengono calcolati nel blocco `generated quantities` e possono essere recuperati con un po' di algebra.

$$\begin{aligned} y_n &= \mathrm{z}_y^{-1}(\mathrm{z}_y(y_n)) \\ &= \mathrm{z}_y^{-1}\left(\alpha' + \beta' \mathrm{z}_x(x_n) + \epsilon'_n\right) \\ &= \mathrm{z}_y^{-1}\left(\alpha' + \beta'\left(\frac{x_n - \bar{x}}{\mathtt{sd}(x)}\right) + \epsilon'_n\right) \\ &= \mathtt{sd}(y)\left(\alpha' + \beta'\left(\frac{x_n - \bar{x}}{\mathtt{sd}(x)}\right) + \epsilon'_n\right) + \bar{y} \\ &= \left(\mathtt{sd}(y)\left(\alpha' - \beta'\frac{\bar{x}}{\mathtt{sd}(x)}\right) + \bar{y}\right) + \left(\beta'\frac{\mathtt{sd}(y)}{\mathtt{sd}(x)}\right) x_n + \mathtt{sd}(y)\epsilon'_n, \end{aligned} \tag{2.1}$$

da cui

$$\alpha = \mathtt{sd}(y)\left(\alpha' - \beta'\frac{\bar{x}}{\mathtt{sd}(x)}\right) + \bar{y}; \qquad \beta = \beta'\frac{\mathtt{sd}(y)}{\mathtt{sd}(x)}; \qquad \sigma = \mathtt{sd}(y)\sigma'.$$

Per svolgere l'analisi bayesiana sistemiamo i dati nel formato appropriato per Stan:

```
data_list <- list(
  N = length(df$kid_score),
  y = df$kid_score,
  x = df$mom_iq
)
```

La funzione `file.path()` ritorna l'indirizzo del file con il codice Stan:

```
file <- file.path("code", "simpleregstd.stan")
```

Il codice Stan può essere stampato usando il metodo `$print()`:

```
mod$print()
```

Prendendo come input un file contenente un programma Stan, la funzione `cmdstan_model()` ritorna un oggetto di classe `CmdStanModel`. In pratica, `CmdStan` traduce un programma Stan in C++ e crea un eseguibile compilato.

```
mod <- cmdstan_model(file)
```

L'indirizzo dell'eseguibile compilato viene ritornato da `$exe_file()`:

```
mod$exe_file()
```

Applicando il metodo `$sample()` ad un oggetto `CmdStanModel` eseguiamo il campionamento MCMC:

```
fit <- mod$sample(
  data = data_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  parallel_chains = 2L,
  refresh = 0,
  thin = 1
)
```

Al metodo `$sample()` possono essere passati molti argomenti. La pagina di documentazione è disponibile al seguente link[5].

[5] https://mc-stan.org/cmdstanr/reference/model-method-sample.html

Un sommario della distribuzione a posteriori per i parametri stimati si ottiene con il metodo `$summary()`, il quale chiama la funzione `summarise_draws()` del pacchetto `posterior`:

```
fit$summary(c("alpha", "beta", "sigma"))
#> # A tibble: 3 x 10
#>   variable   mean median     sd    mad     q5    q95
#>   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
#> 1 alpha    25.9   25.8   6.02   6.02   16.0   35.8
#> 2 beta      0.609  0.609 0.0596 0.0603  0.511  0.707
#> 3 sigma    18.3   18.3   0.634  0.644  17.3   19.4
#> # ... with 3 more variables: rhat <dbl>,
#> #   ess_bulk <dbl>, ess_tail <dbl>
```

Da questo output possiamo valutare rapidamente la convergenza del modello osservando i valori di Rhat per ciascun parametro. Quando questi sono pari o vicini a 1, le catene hanno realizzato la convergenza. Ci sono molti altri test diagnostici, ma questo test è importante per Stan. Oppure è possibile usare:

```
fit$cmdstan_summary()
```

Le statistiche diagnostiche sono fornite dal metodo `$cmdstan_diagnose()`:

```
fit$cmdstan_diagnose()
#> Processing csv files: /var/folders/hl/dt523djx7_q7xjrthzjpdvc40000gn/T/RtmpmXQvl8/simpler
#>
#> Checking sampler transitions treedepth.
#> Treedepth satisfactory for all transitions.
#>
#> Checking sampler transitions for divergences.
#> No divergent transitions found.
#>
#> Checking E-BFMI - sampler transitions HMC potential energy.
#> E-BFMI satisfactory.
#>
#> Effective sample size satisfactory.
#>
```

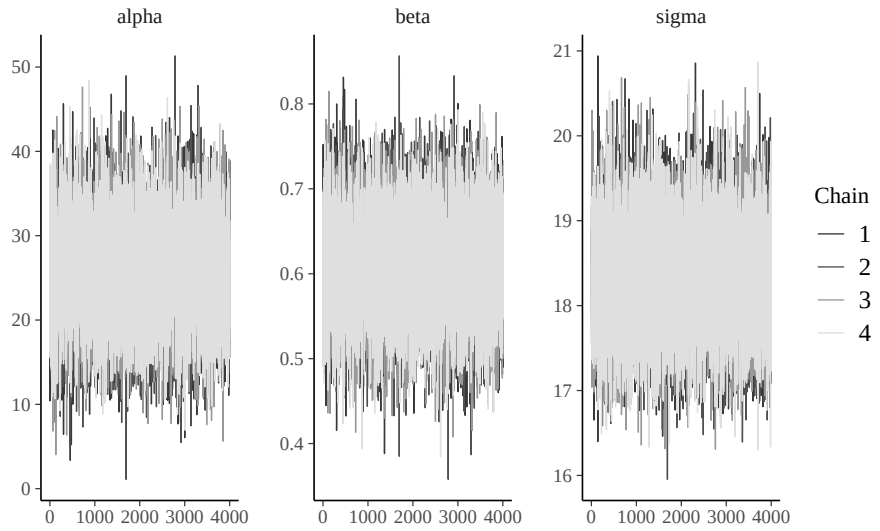
```
#> Split R-hat values satisfactory all parameters.
#>
#> Processing complete, no problems detected.
```

È possibile creare un oggetto di classe `stanfit`

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

per poi utilizzare le funzioni del pacchetto `bayesplot`. Ad esempio:

```
stanfit %>%
  mcmc_trace(pars = c("alpha", "beta", "sigma"))
```



Infine, eseguendo la funzione `launch_shinystan(fit)`, è possibile analizzare oggetti di classe `stanfit` mediante le funzionalità del pacchetto `ShinyStan`.

2.2 Interpretazione dei parametri

Assegnamo ai parametri la seguente interpretazione.

- L'intercetta pari a 25.9 indica il QI medio dei bamini la cui madre ha un QI = 0. Ovviamente questo non ha alcun significato. Vedremo nel modello successivo come trasformare il modello in modo da potere assegnare all'intercetta un'interpretazione sensata.
- La pendenza di 0.61 indica che, all'aumentare di un punto del QI delle madri, il QI medio dei loro bambini aumenta di 0.61 unità. Se consideriamo la gamma di variazione del QI delle madri nel campione, il QI medio dei bambini cambia di 41 punti. Questo indica un sostanziale effetto del QI delle madri sul QI dei loro bambini: $(138.89 - 71.04) * 0.61 = 41.39$.
- Il parametro $\sigma = 18.3$ fornisce una stima della dispersione delle osservazioni attorno al valore predetto dal modello lineare, ovvero fornisce una stima della deviazione standard dei residui attorno al valore atteso del modello lineare.

2.2.1 Centrare i predittori

Per migliorare l'interpretazione dell'intercetta possiamo "centrare" la x, ovvero esprimere la x nei termini degli scarti dalla media: $x - \bar{x}$. In tali circostanze, la pendenza della retta specificata dal modello lineare resta immutata, ma l'intercetta corrisponde a $\mathbb{E}(y \mid x = \bar{x})$. Per ottenere questo risultato, modifichiamo i dati da passare a Stan:

```
data2_list <- list(
  N = length(df$kid_score),
  y = df$kid_score,
  x = df$mom_iq - mean(df$mom_iq)
)
```

Adattiamo il modello:

```
fit2 <- mod$sample(
  data = data2_list,
  iter_sampling = 4000L,
  iter_warmup = 2000L,
  seed = SEED,
  chains = 4L,
  parallel_chains = 2L,
```

```
  refresh = 0,
  thin = 1
)
```

Trasformiamo l'oggetto `fit` in un oggetto di classe `stanfit`:

```
stanfit <- rstan::read_stan_csv(fit2$output_files())
```

Le stime a posteriori dei parametri si ottengono con

```
fit2$summary(c("alpha", "beta", "sigma"))
#> # A tibble: 3 x 10
#>   variable   mean median     sd    mad     q5    q95
#>   <chr>     <dbl>  <dbl>  <dbl>  <dbl>  <dbl>  <dbl>
#> 1 alpha    86.8   86.8   0.876  0.871  85.4   88.2
#> 2 beta      0.609  0.609 0.0591 0.0589  0.513  0.707
#> 3 sigma    18.3   18.3   0.630  0.624  17.3   19.4
#> # ... with 3 more variables: rhat <dbl>,
#> #   ess_bulk <dbl>, ess_tail <dbl>
```

Si noti la nuova intercetta, ovvero 86.8. Questo valore indica il QI medio dei bambini le cui madri hanno un QI pari alla media del campione. Centrare i dati consente dunque di assegnare all'intercetta un'interpretazione utile.

Considerazioni conclusive

La presente discussione suggerisce che, prima di procedere con l'analisi, è conveniente standardizzare i dati. Ciò può essere fatto all'interno del codice Stan (come negli esempi di questo Capitolo), oppure prima di passare i dati a Stan. Se vengono usati dati standardizzati diventa poi facile utilizzare distribuzioni a priori debolmente informative per i parametri. Tali distribuzioni a priori hanno, come unico scopo, quello di regolarizzare i dati e di facilitare la stima dei parametri mediante MCMC.

Bibliografia

de Finetti, B. (1931). Probabilismo. *Logos*, pages 163–219.

de Finetti, B. (1970). *Teoria delle probabilità: sintesi introduttiva con appendice critica*. Einaudi.

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