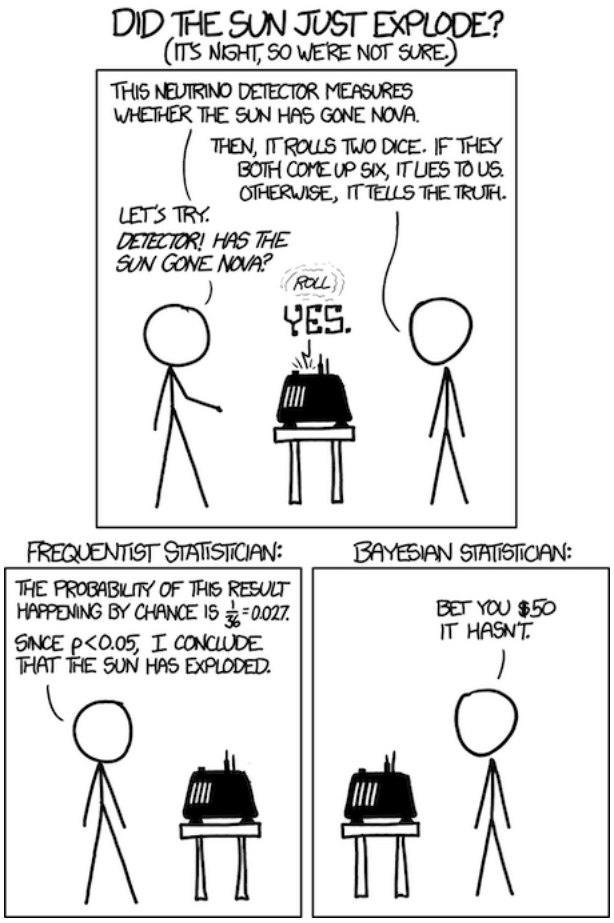
*Corrado Caudek*

# *Data Science per psicologi*

Psicometria – AA 2021/2022



# *Indice*

# *Elenco delle figure*

# *Elenco delle tabelle*

# *Prefazione*

*Data Science per psicologi* contiene il materiale delle lezioni dell'insegnamento di *Psicometria B000286* (A.A. 2021/2022) rivolto agli studenti del primo anno del Corso di Laurea in Scienze e Tecniche Psicologiche dell'Università degli Studi di Firenze. *Psicometria* si propone di fornire agli studenti un'introduzione all'analisi dei dati in psicologia. Le conoscenze/competenze che verranno sviluppate in questo insegnamento sono quelle della Data science, ovvero un insieme di conoscenze/competenze che si pongono all'intersezione tra statistica (ovvero, richiedono la capacità di comprendere teoremi statistici) e informatica (ovvero, richiedono la capacità di sapere utilizzare un software).

## La psicologia e la Data science

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

C'è una ragione molto semplice che dovrebbe farci capire perché la Data science è così importante per la psicologia. Infatti, a ben pensarci, la psicologia è una disciplina intrinsecamente statistica, se per statistica intendiamo quella disciplina che studia la variazione delle caratteristiche degli individui nella popolazione. La psicologia studia *gli individui* ed è proprio la variabilità inter- e intra-individuale ciò che vogliamo descrivere e, in certi casi, predire. In questo senso, la psicologia è molto diversa dall'ingegneria, per esempio. Le proprietà di un determinato ponte sotto certe condizioni, ad esempio, sono molto simili a quelle di un altro pon-

te, sotto le medesime condizioni. Quindi, per un ingegnere la statistica è poco importante: le proprietà dei materiali sono unicamente dipendenti dalla loro composizione e restano costanti. Ma lo stesso non può dirsi degli individui: ogni individuo è unico e cambia nel tempo. E le variazioni tra gli individui, e di un individuo nel tempo, sono l'oggetto di studio proprio della psicologia: è dunque chiaro che i problemi che la psicologia si pone sono molto diversi da quelli affrontati, per esempio, dagli ingegneri. Questa è la ragione per cui abbiamo tanto bisogno della Data science in psicologia: perché la Data science ci consente di descrivere la variazione e il cambiamento. E queste sono appunto le caratteristiche di base dei fenomeni psicologici.

Sono sicuro che, leggendo queste righe, a molti studenti sarà venuta in mente la seguente domanda: perché non chiediamo a qualche esperto di fare il "lavoro sporco" (ovvero le analisi statistiche) per noi, mentre noi (gli psicologi) ci occupiamo solo di ciò che ci interessa, ovvero dei problemi psicologici slegati dai dettagli "tecnici" della Data science? La risposta a questa domanda è che non è possibile progettare uno studio psicologico sensato senza avere almeno una comprensione rudimentale della Data science. Le tematiche della Data science non possono essere ignorate né dai ricercatori in psicologia né da coloro che svolgono la professione di psicologo al di fuori dell'Università. Infatti, anche i professionisti al di fuori dall'università non possono fare a meno di leggere la letteratura psicologica più recente: il continuo aggiornamento delle conoscenze è infatti richiesto dalla deontologia della professione. Ma per potere fare questo è necessario conoscere un bel po' di Data science! Basta aprire a caso una rivista specialistica di psicologia per rendersi conto di quanto ciò sia vero: gli articoli che riportano i risultati delle ricerche psicologiche sono zeppi di analisi statistiche e di modelli formali. E la comprensione della letteratura psicologica rappresenta un requisito minimo nel bagaglio professionale dello psicologo.

Le considerazioni precedenti cercano di chiarire il seguente punto: la Data science non è qualcosa da studiare a malincuore, in un singolo insegnamento universitario, per poi poterla tranquillamente dimenticare. Nel bene e nel male, gli psicologi usano gli strumenti della Data science in tantissimi ambiti della loro attività professionale: in particolare quando costruiscono, somministrano e interpretano i test psicometrici. È dunque chiaro che possedere delle solide basi di Data science è un tassello imprescindibile del bagaglio professionale dello psicologo. In questo insegnamento verranno trattati i temi base della Data science e verrà

adottato un punto di vista bayesiano, che corrisponde all'approccio più recente e sempre più diffuso in psicologia.

## Come studiare

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

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

## Sviluppare un metodo di studio efficace

Avendo insegnato molte volte in passato un corso introduttivo di analisi dei dati ho notato nel corso degli anni che gli studenti con l'atteggiamento mentale che descriverò qui sotto generalmente ottengono ottimi risultati. Alcuni studenti sviluppano naturalmente questo approccio allo studio, ma altri hanno bisogno di fare uno sforzo per maturarlo. Fornisco qui sotto una breve descrizione del "metodo di studio" che, nella mia esperienza, è il più efficace per affrontare le richieste di questo insegnamento.

- Dedicate un tempo sufficiente al materiale di base, apparentemente facile; assicuratevi di averlo capito bene. Cercate le lacune nella vostra comprensione. Leggere presentazioni diverse dello stesso materiale (in libri o articoli diversi) può fornire nuove intuizioni.
- Gli errori che facciamo sono i nostri migliori maestri. Istintivamente cerchiamo di dimenticare subito i nostri errori. Ma il miglior modo di imparare è apprendere dagli errori che commettiamo. In questo senso, una soluzione corretta è meno utile di una soluzione sbagliata. Quando commettiamo un errore questo ci fornisce un'informazione importante: ci fa capire qual è il materiale di studio sul quale dobbiamo ritornare e che dobbiamo capire meglio.
- C'è ovviamente un aspetto "psicologico" nello studio. Quando un esercizio o problema ci sembra incomprensibile, la cosa migliore da fare è dire: "mi arrendo", "non ho idea di cosa fare!". Questo ci rilassa: ci siamo già arresi, quindi non abbiamo niente da perdere, non dobbiamo più preoccuparci. Ma non dobbiamo fermarci qui. Le cose "migliori" che faccio (se ci sono) le faccio quando non ho voglia di lavorare. Alle volte, quando c'è qualcosa che non so fare e non ho idea di come affontare, mi dico: "oggi non ho proprio voglia di fare fatica", non ho voglia di mettermi nello stato mentale per cui "in 10 minuti devo risolvere il problema perché dopo devo fare altre cose". Però ho voglia di *divertirmi* con quel problema e allora mi dedico a qualche aspetto "marginale" del problema, che so come affrontare, oppure considero l'aspetto più difficile del problema, quello che non so come risolvere, ma invece di cercare di risolverlo, guardo come altre persone hanno affrontato problemi simili, opppure lo stesso problema in un altro contesto. Non mi pongo l'obiettivo "risolvi il problema in 10 minuti", ma invece quello di farmi un'idea "generale" del problema, o quello di capire un caso più specifico e più semplice del problema. Senza nessuna pressione. Infatti, in quel momento ho deciso di non lavorare (ovvero, di non fare fatica). Va benissimo se "parto per la tangente", ovvero se mi metto a leggere del materiale che sembra avere poco a che fare con il problema centrale (le nostre intuizioni e la nostra curiosità solitamente ci indirizzano sulla strada giusta). Quando faccio così, molto spesso trovo la soluzione del problema che mi ero posto e, paradossalmente, la trovo in un tempo minore di quello che, in precedenza, avevo dedicato a "lavorare" al problema. Allora perché non faccio sempre così? C'è ovviamente l'aspetto dei "10 minuti" che non è sempre facile da dimenticare. Sotto pressione, possiamo solo agire in maniera automatica, ovvero possia-

mo solo applicare qualcosa che già sappiamo fare. Ma se dobbiamo imparare qualcosa di nuovo, la pressione è un impedimento.

- È utile farsi da soli delle domande sugli argomenti trattati, senza limitarsi a cercare di risolvere gli esercizi che vengono assegnati. Quando studio qualcosa mi viene in mente: "se questo è vero, allora deve succedere quest'altra cosa". Allora verifico se questo è vero, di solito con una simulazione. Se i risultati della simulazione sono quelli che mi aspetto, allora vuol dire che ho capito. Se i risultati sono diversi da quelli che mi aspettavo, allora mi rendo conto di non avere capito e ritorno indietro a studiare con più attenzione la teoria che pensavo di avere capito – e ovviamente mi rendo conto che c'era un aspetto che avevo frainteso. Questo tipo di verifica è qualcosa che dobbiamo fare da soli, in prima persona: nessun altro può fare questo al posto nostro.
- Non aspettatevi di capire tutto la prima volta che incontrate un argomento nuovo.[1] È utile farsi una nota mentalmente delle lacune nella vostra comprensione e tornare su di esse in seguito per carcare di colmarle. L'atteggiamento naturale, quando non capiamo i dettagli di qualcosa, è quello di pensare: "non importa, ho capito in maniera approssimativa questo punto, non devo preoccuparmi del resto". Ma in realtà non è vero: se la nostra comprensione è superficiale, quando il problema verrà presentato in una nuova forma, non riusciremo a risolverlo. Per cui i dubbi che ci vengono quando studiamo qualcosa sono il nostro alleato più prezioso: ci dicono esattamente quali sono gli aspetti che dobbiamo approfondire per potere migliorare la nostra preparazione.
- È utile sviluppare una visione d'insieme degli argomenti trattati, capire l'obiettivo generale che si vuole raggiungere e avere chiaro il contributo che i vari pezzi di informazione forniscono al raggiungimento di tale obiettivo. Questa organizzazione mentale del materiale di studio facilita la comprensione. È estremamente utile creare degli schemi di ciò che si sta studiando. Non aspettate che sia io a fornirvi un riepilogo di ciò che dovete imparare: sviluppate da soli tali schemi e tali riassunti.
- Tutti noi dobbiamo imparare l'arte di trovare le informazioni, non solo nel caso di questo insegnamento. Quando vi trovate di fronte a qualcosa che non capite, o ottenete un oscuro messaggio di errore da

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

un software, ricordatevi: “Google is your friend”!

Corrado Caudek
Marzo 2022

# Parte I

# Il calcolo delle probabilità

# 1

# *La logica dell'incerto*

In questa parte della dispensa verrà introdotta la teoria delle probabilità. Prima di entrare nei dettagli, cerchiamo di capire perché la probabilità sia cruciale per la ricerca scientifica.

La teoria delle probabilità è cruciale per la scienza perché la ricerca procede mediante l'inferenza induttiva. Non siamo mai completamente sicuri della verità di una proposizione (ipotesi, teoria): al valore di verità di una proposizione possiamo solo assegnare un giudizio probabilistico. L'approccio bayesiano è una scuola di pensiero che usa la probabilità per quantificare il grado di fiducia che può essere attribuito ad una proposizione. L'inferenza statistica bayesiana è un tipo di inferenza induttiva che ha lo scopo di quantificare la fiducia che si ha nell'ipotesi $H$ dopo il verificarsi del dato d'evidenza $E$. Per quantificare un tale grado di fiducia l'inferenza statistica bayesiana utilizza la teoria delle probabilità. Una comprensione dell'inferenza statistica bayesiana richiede dunque, preliminarmente, la conoscenze della teoria delle probabilità.

## 1.1 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 ab-

> biamo 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 disponibili. 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: descrive lo stato della nostra incertezza rispetto al contenuto di verità di una proposizione.

Nell'interpretazione frequentista, la probabilità $P(E)$ rappresenta la frequenza relativa a lungo termine di un grande numero di ripetizioni di un esperimento casuale sotto le medesime condizioni. 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 – sono invece esclusi gli eventi unici e irripetibili.

L'interpretazione bayesiana della probabilità fa invece ricorso ad una concezione più ampia, non legata al solo evento in sé ma che include anche il soggetto assegnante la funzione di probabilità. In pratica l'assegnazione di probabilità bayesiana viene effettuata dal decisore, in base alle proprie conoscenze a priori integrate con tutto il generico bagaglio culturale personale. In questo modo, la probabilità non sarà obbligatoriamente la stessa per tutti i soggetti, ma variarierà a seconda delle informazioni a disposizione, dell'esperienza personale e soprattutto del punto di vista proprio di ogni decisore ed è dunque assimilabile al "grado di fiducia" – in inglese *degree of belief* – di un dato soggetto, in un dato istante e con un dato insieme d'informazioni, circa l'accadere dell'evento $E$. "[N]essuna 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" (de Finetti, 1931).

L'impostazione bayesiana, sviluppata da Ramsey e de Finetti, riconduce l'assegnazione di probabilità allo scommettere sul verificarsi di un evento:

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 $E$.

Secondo De Finetti, le valutazioni di probabilità degli eventi devono rispondere ai principi di equità e coerenza. Una scommessa risponde al principio di *equità* se il ruolo di banco e giocatore sono scambiabili in ogni momento del gioco e sempre alle stesse condizioni. Una scommessa risponde al principio di *coerenza* se non vi sono combinazioni di scommesse che consentano (sia al banco che al giocatore) di realizzare perdite o vincite certe.

L'approccio definettiano dell'impostazione della scommessa si basa dunque sulle assunzioni di razionalità e coerenza del decisore, al quale è fatto esplicito divieto di effettuare scommesse a perdita o guadagno certo. Il decisore, proponendo la scommessa, deve essere disposto a scambiare il posto dello scommettitore con quello del banco.

Il metodo della scommessa, oltre che una definizione, fornisce un mezzo operativo di assegnazione della probabilità. Sulla base di questa definizione operativa, che si può ritenere ragionevolmente soddisfatta dal comportamento di un qualunque individuo che agisca in modo razionale in condizioni di incertezza, possono essere agevolmente dimostrate tutte le proprietà classiche della probabilità: essa non può assumere valori negativi, né può essere superiore all'unità; se $E$ è un evento certo, la sua probabilità è 1; se invece $E$ è un evento impossibile, la sua probabilità è 0.

I problemi posti dall'approccio definettiano riguardano l'arbitrarietà dell'assegnazione soggettività di probabilità la quale sembra negare la validità dell'intero costrutto teorico. In risposta a tale critica, i bayesiani sostengono che gli approcci oggettivisti alla probabilità nascondono scelte arbitrarie preliminari e sono basate su assunzioni implausibili. È molto più onesto esplicitare subito tutte le scelte arbitrarie effettuate nel corso dell'analisi in modo da controllarne coerenza e razionalità.

## 1.2 Variabili casuali e probabilità di un evento

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

### 1.2.1 Eventi e probabilità

Nella teoria delle probabilità il risultato "testa" nel lancio di una moneta è chiamato *evento*.[1] Ad esempio, $Y = 1$ denota l'evento in cui il lancio di una moneta produce come risultato testa. Il funzionale $P(\cdot)$ 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 $P(Y = 1) = 0.5$.

Se la moneta è equilibrata dobbiamo anche avere $P(Y = 0) = 0.5$. I due eventi $Y = 1$ e $Y = 0$ sono *mutuamente esclusivi* nel senso che non possono entrambi verificarsi contemporaneamente: $P(Y = 1 \; e \; Y = 0) = 0$. 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, $P(Y = 1 \; o \; Y = 0) = 1$.

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

### 1.2.2 Spazio campione e risultati possibili

Anche se il lancio di una moneta produce sempre uno specifico risultato nel mondo reale, 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.

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

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 $\Omega$ di tutti i risultati possibili è chiamato *spazio campione* (*sample space*). Lo spazio campione 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 campione: 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).

## 1.3 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 i cui valori non possono essere previsti con esattezza. 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 numerici. 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.

Una variabile casuale può essere *discreta* o *continua*. Una variabile casuale discreta può assumere un numero finito di valori $x_1, \dots, x_n$, in corrispondenza degli eventi $E_i, \dots, E_n$ che si verificano con le rispettive probabilità $p_1, \dots, p_n$. Un esempio è il punteggio totale di un test psicometrico costituito da item su scala Likert. Invece un esempio di una variabile casuale continua è la distanza tra due punti, che può assumere infiniti valori all'interno di un certo intervallo. L'insieme $S$ dei valori che la variabile casuale può assumere è detto *spazio dei valori* o *spazio*

*degli stati*. La caratteristica fondamentale di una variabile casuale è data dall'insieme delle probabilità dei suoi valori, detta *distribuzione di probabilità*. Nel seguito useremo la notazione $P(\cdot)$ per fare riferimento alle distribuzioni di probabilità delle variabili casuali discrete e $p(\cdot)$ per fare riferimento alla densità di probabilità delle variabili casuali continue. In questo contesto, l'insieme dei valori che la variabile casuale può assumere è detto *supporto* della sua distribuzione di probabilità. Il supporto di una variabile casuale può essere finito (come nel caso di una variabile casuale uniforme di supporto $[a, b]$) o infinito (nel caso di una variabile causale gaussiana il cui supporto coincide con la retta reale).

## 1.4 Usare la simulazione per stimare le probabilità

I metodi basati sulla simulazione consentono di stimare le probabilità degli eventi in un modo diretto, se siamo in grado di generare molteplici e casuali realizzazioni 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, la stessa operazione si può realizzare 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 $P(Y = 1)$ è data dalla frequenza relativa del numero di volte in cui abbiamo osservato l'evento di interesse $(Y = 1)$:

```
M <- 100
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] = 0.53
```

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(100)
}
#> estimated Pr[Y = 1] = 0.44
#> estimated Pr[Y = 1] = 0.52
#> estimated Pr[Y = 1] = 0.46
#> estimated Pr[Y = 1] = 0.57
#> estimated Pr[Y = 1] = 0.47
#> estimated Pr[Y = 1] = 0.46
#> estimated Pr[Y = 1] = 0.48
#> estimated Pr[Y = 1] = 0.49
```

```
#> estimated Pr[Y = 1] = 0.47
#> estimated Pr[Y = 1] = 0.62
```

Dato che la moneta è equilibrata, la stima delle probabilità dell'evento $Pr[Y = 1]$ è simile a al valore che ci aspettiamo ($P(Y = 1) = 0.5$), 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(1000)
}
#> estimated Pr[Y = 1] = 0.497
#> estimated Pr[Y = 1] = 0.529
#> estimated Pr[Y = 1] = 0.493
#> estimated Pr[Y = 1] = 0.511
#> estimated Pr[Y = 1] = 0.506
#> estimated Pr[Y = 1] = 0.52
#> estimated Pr[Y = 1] = 0.49
#> estimated Pr[Y = 1] = 0.495
#> estimated Pr[Y = 1] = 0.489
#> estimated Pr[Y = 1] = 0.496
```

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] = 0.4885
#> estimated Pr[Y = 1] = 0.4957
#> estimated Pr[Y = 1] = 0.4902
#> estimated Pr[Y = 1] = 0.5032
#> estimated Pr[Y = 1] = 0.5048
#> estimated Pr[Y = 1] = 0.4931
#> estimated Pr[Y = 1] = 0.4965
#> estimated Pr[Y = 1] = 0.499
```

```
#> estimated Pr[Y = 1] = 0.4979
#> estimated Pr[Y = 1] = 0.4973
```

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)

## 1.5 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 qjello che accade all'aumentare del numero $M$ di ripetizioni del lancio della moneta consiste nel registrare la stima della probabilità dell'evento $P(Y = 1)$ in funzione del numero di ripetizioni dell'esperimento casuale per ogni $m \in 1 : M$. Possiamo ottenere un grafico dell'andamento della stima di $P(Y = 1)$ in funzione di $m$ come:

```
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 <- tibble(
  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]"
  )
```

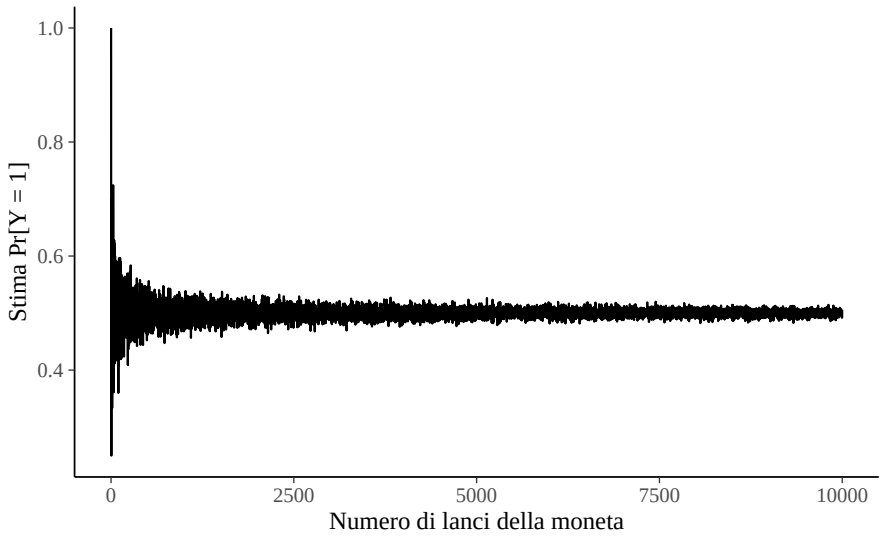



***Figura 1.1:*** *Stima della probabilità di successo in funzione del numero di lanci di una moneta.*

Dato che il grafico 1.1 espresso su una scala lineare non rivela chiaramente l'andamento della simulazione, imponiamo una scala logaritmica sull'asse delle ascisse ($x$). Su scala logaritmica, i valori tra 1 e 10 vengono tracciati all'incirca con la stessa ampiezza che si osserva tra i valori 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]"
)
```

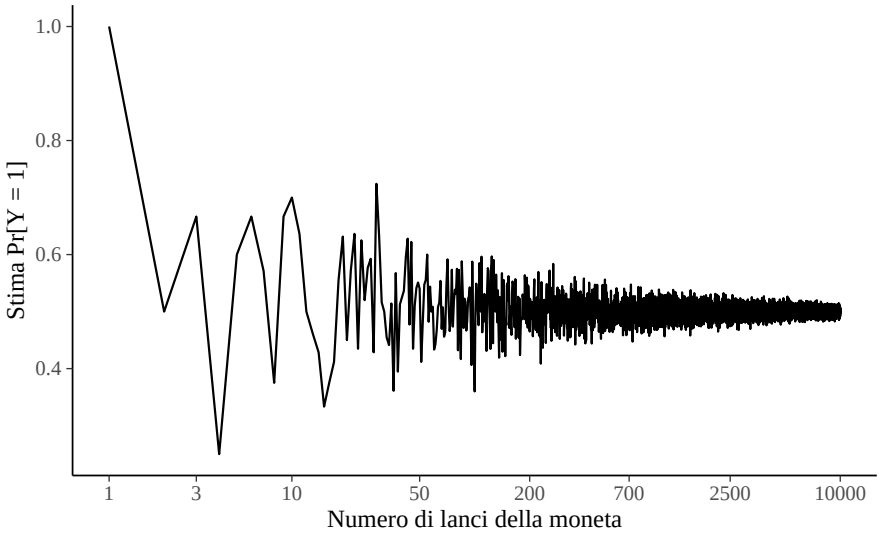



***Figura 1.2:*** *Stima della probabilità di successo in funzione del numero di lanci di una moneta – scala logaritmica.*

La *legge dei grandi numeri* ci dice che, all'aumentare del numero di ripetizioni dell'esperimento casuale, la media dei risultati ottenuti tende al valore atteso, man mano che vengono eseguite più prove. Nella figura 1.2 vediamo infatti che, all'aumentare del numero $M$ di lanci della moneta, la stima di $P(Y = 1)$ converge al valore 0.5.

## 1.6 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. I risultati di ciascuno dei tre lanci possono essere rappresentati da una diversa variabile casuale, ad esempio, $Y_1, Y_2, Y_3$. 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 $P(Y_n = 1) = 0.5$ e $P(Y_n = 0) = 0.5$.

È possibile 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] 1 0 1
z <- sum(c(y1, y2, y3))
cat("z =", z, "\n")
#> z = 2
```

ovvero,

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

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

oppure, ancora più semplicemente:

```
y <- rbinom(3, 1, 0.5)
y
#> [1] 1 0 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
#> 0.1258 0.3750 0.3748 0.1244
```

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
#> 0.06340 0.24917 0.37360 0.25022 0.06361
```

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

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

## 1.7 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

$$P_Z(z) = \mathrm{P}(Z = z).$$

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 1.3.

```
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
```

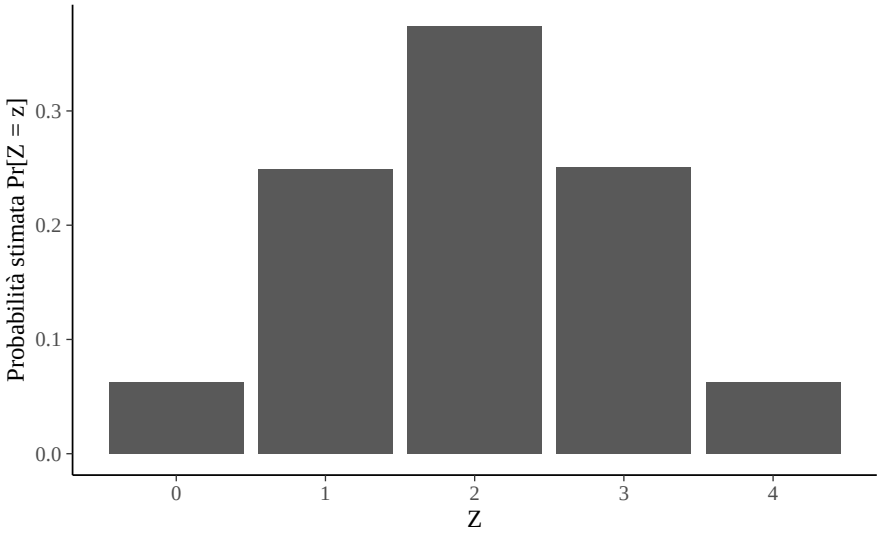



***Figura 1.3:*** *Grafico di $M = 100\,000$ simulazioni della funzione di massa di probabilità di una variabile casuale definita come il numero di teste in quattro lanci di una moneta equilibrata.*

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 1.7, la probabilità che la variabile casuale $Z$ sia un numero dispari è

$$P(\text{Z è un numero dispari}) = P_z(Z = 1) + P_z(Z = 3) = \frac{4}{16} + \frac{4}{16} = \frac{1}{2}.$$

## Commenti e considerazioni finali

In questo capitolo abbiamo visto come si costruisce lo spazio campione di un esperimento casuale, quali sono le proprietà di base della probabilità

e come si assegnano le probabilità agli eventi definiti sopra uno spazio campione 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 $F(X) = P(X < x)$, 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

# *Probabilità condizionata*

Il fondamento della statistica bayesiana è il teorema di Bayes e il teorema di Bayes è una semplice ridescrizione della probabilità condizionata. Esaminiamo dunque la probabilità condizionata.

## 2.1 Probabilità condizionata su altri eventi

L'attribuzione di una probabilità ad un evento è sempre condizionata dalle conoscenze che abbiamo a disposizione. Per un determinato stato di conoscenze, attribuiamo ad un dato evento una certa probabilità di verificarsi; ma se il nostro stato di conoscenze cambia, allora cambierà anche la probabilità che attribuiremo all'evento in questione. Infatti, si può pensare che tutte le probabilità siano probabilità condizionate, anche se l'evento condizionante non è sempre esplicitamente menzionato. Consideriamo il seguente problema.

**Esercizio 2.1.** Supponiamo che lo screening per la diagnosi precoce del tumore mammario si avvalga di test che sono accurati al 90%, nel senso che il 90% delle donne con cancro e il 90% delle donne senza cancro saranno classificate correttamente. Supponiamo che l'1% delle donne sottoposte allo screening abbia effettivamente il cancro al seno. Ci chiediamo: qual è la probabilità che una donna scelta casualmente abbia una mammografia positiva e, se ce l'ha, qual è la probabilità che abbia davvero il cancro?

Per risolvere questo problema, supponiamo che il test in questione venga somministrato ad un grande campione di donne, diciamo a 1000 donne. Di queste 1000 donne, 10 (ovvero, l'1%) hanno il cancro al seno. Per queste 10 donne, il test darà un risultato positivo in 9 casi (ovvero, nel 90% dei casi). Per le rimanenti 990 donne che non hanno il cancro al seno, il test darà un risultato positivo in 99 casi (se la probabilità di un

vero positivo è del 90%, la probabilità di un falso positivo è del 10%). Questa situazione è rappresentata nella figura 2.1.

Combinando i due risultati precedenti, vediamo che il test dà un risultato positivo per 9 donne che hanno effettivamente il cancro al seno e per 99 donne che non ce l'hanno, per un totale di 108 risultati positivi. Dunque, la probabilità di ottenere un risultato positivo al test è $\frac{108}{1000} = 11\%$. Ma delle 108 donne che hanno ottenuto un risultato positivo al test, solo 9 hanno il cancro al seno. Dunque, la probabilità di avere il cancro, dato un risultato positivo al test, è pari a $\frac{9}{108} = 8\%$.

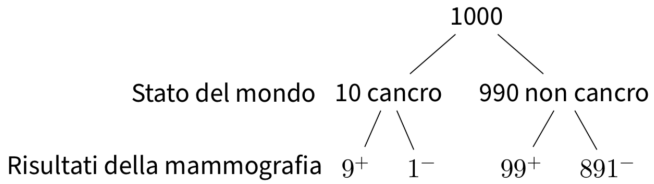



***Figura 2.1:*** *Rappresentazione ad albero che riporta le frequenze attese dei risultati di una mammografia in un campione di 1,000 donne.*

Nell'esercizio precedente, la probabilità dell'evento "ottenere un risultato positivo al test" è una probabilità non condizionata, mentre la probabilità dell'evento "avere il cancro al seno, dato che il test ha prodotto un risultato positivo" è una probabilità condizionata.

In termini generali, la probabilità condizionata $P(A \mid B)$ rappresenta la probabilità che si verifichi l'evento $A$ sapendo che si è verificato l'evento $B$. Ciò ci conduce alla seguente definizione.

**Definizione 2.1.** Dato un qualsiasi evento $A$, si chiama *probabilità condizionata* di $A$ dato $B$ il numero

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{con } P(B) > 0, \tag{2.1}$$

dove $P(A \cap B)$ è la probabilità congiunta dei due eventi, ovvero la probabilità che si verifichino entrambi.

**Esercizio 2.2.** Da un mazzo di 52 carte (13 carte per ciascuno dei 4 semi) ne viene estratta una in modo casuale. Qual è la probabilità che esca una figura di cuori? Sapendo che la carta estratta ha il seme di

cuori, qual è la probabilità che il valore numerico della carta sia 7, 8 o 9?

Ci sono 13 carte di cuori, dunque la risposta alla prima domanda è 1/4 (probabilità non condizionata). Per rispondere alla seconda domanda consideriamo solo le 13 carte di cuori; la probabilità cercata è dunque 3/13 (probabilità condizionata).

## 2.2 Legge delle probabilità composte

Dalla definizione di probabilità condizionata è possibile esprimere la probabilità congiunta tramite le condizionate. La legge delle probabilità composte (o regola moltiplicativa, o regola della catena) afferma che la probabilità che si verifichino due eventi $A$ e $B$ è pari alla probabilità di uno dei due eventi moltiplicato con la probabilità dell'altro evento condizionato al verificarsi del primo:

$$P(A \cap B) = P(B)P(A \mid B) = P(A)P(B \mid A). \tag{2.2}$$

La (2.2) si estende al caso di $n$ eventi $A_1, \ldots, A_n$ nella forma seguente:

$$P\left(\bigcap_{k=1}^{n} A_k\right) = \prod_{k=1}^{n} \left(A_k \,\middle|\, \bigcap_{j=1}^{k-1} A_j\right) \tag{2.3}$$

Per esempio, nel caso di quattro eventi abbiamo

$$P(A_1 \cap A_2 \cap A_3 \cap A_4) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot P(A_4 \mid A_1 \cap A_2 \cap A_3).$$

**Esercizio 2.3.** Da un'urna contenente 6 palline bianche e 4 nere si estrae una pallina per volta, senza reintrodurla nell'urna. Indichiamo con $B_i$ l'evento: "esce una pallina bianca alla $i$-esima estrazione" e con $N_i$ l'estrazione di una pallina nera. L'evento: "escono due palline bianche nelle prime due estrazioni" è rappresentato dalla intersezione $\{B_1 \cap B_2\}$ e la sua probabilità vale, per la (2.2)

$$P(B_1 \cap B_2) = P(B_1)P(B_2 \mid B_1).$$

$P(B_1)$ vale 6/10, perché nella prima estrazione $\Omega$ è costituito da 10 elementi: 6 palline bianche e 4 nere. La probabilità condizionata $P(B_2 \mid B_1)$ vale 5/9, perché nella seconda estrazione, se è verificato l'evento $B_1$, lo spazio campionario consiste di 5 palline bianche e 4 nere. Si ricava pertanto:

$$P(B_1 \cap B_2) = \frac{6}{10} \cdot \frac{5}{9} = \frac{1}{3}.$$

In modo analogo si ha che

$$P(N_1 \cap N_2) = P(N_1)P(N_2 \mid N_1) = \frac{4}{10} \cdot \frac{3}{9} = \frac{4}{30}.$$

Se l'esperimento consiste nell'estrazione successiva di 3 palline, la probabilità che queste siano tutte bianche vale, per la (2.3):

$$P(B_1 \cap B_2 \cap B_3) = P(B_1)P(B_2 \mid B_1)P(B_3 \mid B_1 \cap B_2),$$

dove la probabilità $P(B_3 \mid B_1 \cap B_2)$ si calcola supponendo che si sia verificato l'evento condizionante $\{B_1 \cap B_2\}$. Lo spazio campionario per questa probabilità condizionata è costituito da 4 palline bianche e 4 nere, per cui $P(B_3 \mid B_1 \cap B_2) = 1/2$ e quindi:

$$P(B_1 \cap B_2 \cap B_3) = \frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} = \frac{1}{6}.$$

La probabilità dell'estrazione di tre palline nere è invece:

$$\begin{aligned} P(N_1 \cap N_2 \cap N_3) &= P(N_1)P(N_2 \mid N_1)P(N_3 \mid N_1 \cap N_2) \\ &= \frac{4}{10} \cdot \frac{3}{9} \cdot \frac{2}{8} = \frac{1}{30}. \end{aligned}$$

## 2.3 L'indipendenza stocastica

Un concetto molto importante per le applicazioni statistiche della probabilità è quello dell'indipendenza stocastica. La definizione (2.1) consente di esprimere il concetto di indipendenza di un evento da un altro in forma intuitiva: se $A$ e $B$ sono eventi indipendenti, allora il verificarsi di $A$

non influisce sulla probabilità del verificarsi di $B$, ovvero non la condiziona, e il verificarsi di $B$ non influisce sulla probabilità del verificarsi di $A$. Infatti, per la (2.1), si ha che, se $A$ e $B$ sono due eventi indipendenti, risulta:

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A),$$

$$P(B \mid A) = \frac{P(A)P(B)}{P(A)} = P(B).$$

Possiamo dunque dire che due eventi $A$ e $B$ sono indipendenti se

$$\begin{aligned} P(A \mid B) &= P(A), \\ P(B \mid A) &= P(B). \end{aligned} \tag{2.4}$$

## 2.4 Il teorema della probabilità totale

Dato un insieme finito $A_i$ di eventi, nel calcolo della probabilità dell'unione di tutti gli eventi, se gli eventi considerati non sono a due a due incompatibili, si deve tenere conto delle loro intersezioni. In particolare, la probabilità dell'unione di due eventi $A$ e $B$ è pari alla somma delle singole probabilità $P(A)$ e $P(B)$ diminuita della probabilità della loro intersezione:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{2.5}$$

Nel caso di tre eventi, si ha

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C).$$

La formula per il caso di $n$ eventi si ricava per induzione.

Per il caso di due soli eventi, se $A$ e $B$ sono indipendenti, la (2.5) si modifica nella relazione seguente:

$$P(A \cup B) = P(A) + P(B) - P(A)P(B). \tag{2.6}$$

Nel caso di due eventi $A$ e $B$ incompatibili, se cioè $P(A \cap B) = \emptyset$, si ha che

$$A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B).$$

Si può dimostrare per induzione che ciò vale anche per un insieme finito di eventi $A_n$ a due a due incompatibili, ovvero che:

$$A_i \cap A_j = \emptyset, i \neq j \Rightarrow P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

**Esercizio 2.4.** Nel lancio di due dadi non truccati, si considerino gli eventi: $A = \{$esce un 1 o un 2 nel primo lancio$\}$ e $B = \{$il punteggio totale è 8$\}$. Gli eventi $A$ e $B$ sono indipendenti?

Rappresentiamo qui sotto lo spazio campionario dell'esperimento casuale.

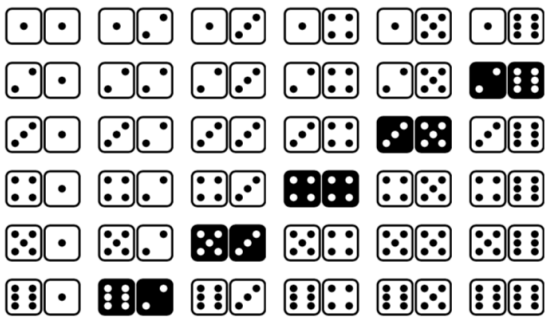


***Figura 2.2:*** *Rappresentazione dello spazio campionario dei risultati dell'esperimento casuale corrispondente al lancio di due dadi bilanciati. Sono evidenziati gli eventi elementari che costituiscono l'evento A: esce un 1 o un 2 nel primo lancio.*

Gli eventi $A$ e $B$ non sono statisticamente indipendenti. Infatti, le loro probabilità valgono $P(\mathrm{A}) = 12/36$ e $P(\mathrm{B}) = 5/36$ e la probabilità della loro intersezione è

$$P(A \cap B) = 1/36 = 3/108 \neq P(A)P(B) = 5/108.$$

*Osservazione.* Il concetto di indipendenza è del tutto differente da quello di incompatibilità. Si noti infatti che due eventi $A$ e $B$ incompatibili (per i quali si ha $A \cap B = \emptyset$) sono statisticamente dipendenti, poiché il verificarsi dell'uno esclude il verificarsi dell'altro: $P(A \cap B) = 0 \neq P(A)P(B)$.

## 2.5 Il teorema della probabilità assoluta

Il teorema della probabilità assoluta consente di calcolare la probabilità di un evento $E$ di cui sono note le probabilità condizionate rispetto ad altri eventi $(H_i)_{i\geq 1}$, a condizione che essi costituiscano una partizione dell'evento certo $\Omega$, ovvero

1. $\bigcup_{i=1}^{\infty} H_i = \Omega$;
2. $H_j \cap H_j = \emptyset, i \neq j$;
3. $P(H_i) > 0, i = 1, \ldots, \infty$.

Nel caso di una partizione dello spazio campionario in tre sottoinsiemi, ad esempio, abbiamo

$$\mathrm{P}(E) = \mathrm{P}(E \cap H_1) + \mathrm{P}(E \cap H_2) + \mathrm{P}(E \cap H_3)$$

ovvero

$$\mathrm{P}(E) = \mathrm{P}(E \mid H_1)\mathrm{P}(H_1) + \mathrm{P}(E \mid H_2)\mathrm{P}(H_2) + \mathrm{P}(E \mid H_3)\mathrm{P}(H_3). \quad (2.7)$$

Il teorema della probabilità assoluta afferma dunque che, se l'evento $E$ è costituito da tutti gli eventi elementari in $E \cap H_1$, $E \cap H_2$ e $E \cap H_3$, allora la sua probabilità è data dalla somma delle probabilità condizionate $P(E \mid H_i)$, ciascuna delle quali pesata per la probabilità dell'evento condizionante $H_i$.

**Esercizio 2.5.** Si considerino tre urne, ciascuna delle quali contiene 100 palline:

- Urna 1: 75 palline rosse e 25 palline blu,
- Urna 2: 60 palline rosse e 40 palline blu,
- Urna 3: 45 palline rosse e 55 palline blu.

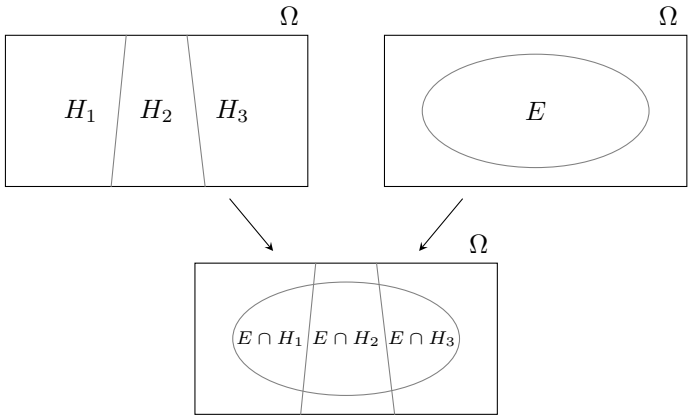



***Figura 2.3:*** *Partizione dell'evento certo $\Omega$ in tre sottoinsiemi sui quali viene definito l'evento $E$.*

Una pallina viene estratta a caso da un'urna anch'essa scelta a caso. Qual è la probabilità che la pallina estratta sia di colore rosso?

Sia $R$ l'evento "la pallina estratta è rossa" e sia $U_i$ l'evento che corrisponde alla scelta dell'$i$-esima urna. Sappiamo che

$$\mathrm{P}(R \mid U_1) = 0.75, \quad \mathrm{P}(R \mid U_2) = 0.60, \quad \mathrm{P}(R \mid U_3) = 0.45.$$

Gli eventi $U_1$, $U_2$ e $U_3$ costituiscono una partizione dello spazio campionario in quanto $U_1$, $U_2$ e $U_3$ sono eventi mutualmente esclusivi ed esaustivi, $\mathrm{P}(U_1 \cup U_2 \cup U_3) = 1.0$. In base al teorema della probabilità assoluta, la probabilità di estrarre una pallina rossa è dunque

$$\begin{aligned}\mathrm{P}(R) &= \mathrm{P}(R \mid U_1)\mathrm{P}(U_1) + \mathrm{P}(R \mid U_2)\mathrm{P}(U_2) + \mathrm{P}(R \mid U_3)\mathrm{P}(U_3) \\ &= 0.75 \cdot \frac{1}{3} + 0.60 \cdot \frac{1}{3} + 0.45 \cdot \frac{1}{3} \\ &= 0.60.\end{aligned}$$

## 2.6 Indipendenza condizionale

Aggiungo qui delle considerazioni sul concetto di indipendenza condizionale a cui si farà riferimento nell'ultima parte della dispensa. L'indipendenza condizionale descrive situazioni in cui un'osservazione è irrilevante

o ridondante quando si valuta la certezza di un'ipotesi. L'indipendenza condizionale è solitamente formulata nei termini della probabilità condizionata, come un caso speciale in cui la probabilità dell'ipotesi data un'osservazione non informativa è uguale alla probabilità senza tale osservazione non informativa.

Se $A$ è l'ipotesi e $B$ e $C$ sono osservazioni, l'indipendenza condizionale può essere espressa come l'uguaglianza:

$$P(A \mid B, C) = P(A \mid C).$$

Dato che $P(A \mid B, C)$ è uguale a $P(A \mid C)$, questa uguaglianza corrisponde all'affermazione che $B$ non fornisce alcun contributo alla certezza di $A$. In questo caso si dice che $A$ e $B$ condizionalmente indipendente dato $C$, scritto simbolicamente come: $(A \perp\!\!\!\perp B \mid C)$.

In maniera equivalente, l'indipendenza condizionale $(A \perp\!\!\!\perp B \mid C)$ si verifica se:

$$P(A, B \mid C) = P(A \mid C)P(B \mid C).$$

Un esempio è il seguente (da Wikipedia). Siano due eventi le probabilità che le persone A e B tornino a casa in tempo per la cena, e il terzo evento è il fatto che una tempesta di neve ha colpito la città. Mentre sia A che B hanno una probabilità più piccola di tornare a casa in tempo per cena (di quando non c'è la neve), tali probabilità sono comunque indipendenti l'una dall'altra. Cioè, sapere che A è in ritardo non ci dice nulla sul fatto che B sia in ritardo o meno. (A e B potrebbero vivere in quartieri diversi, percorrere distanze diverse e utilizzare mezzi di trasporto diversi.) Tuttavia, se sapessimo che A e B vivono nello stesso quartiere, usano lo stesso mezzo di trasporto e lavorano nello stesso luogo, allora i due eventi non sarebbero condizionatamente indipendenti.

## Commenti e considerazioni finali

La probabilità condizionata è importante perché ci fornisce uno strumento per precisare il concetto di indipendenza statistica. Una delle domande più importanti delle analisi statistiche è infatti quella che si

chiede se due variabili sono associate tra loro oppure no. In questo Capitolo abbiamo discusso il concetto di indipendenza (come contrapposto al concetto di associazione – si veda il Capitolo **??**). In seguito vedremo come sia possibile fare inferenza sull'associazione tra variabili.

# 3

# *Il teorema di Bayes*

Il teorema di Bayes assume un ruolo fondamentale nell'interpretazione soggettivista della probabilità perché descrive l'aggiornamento della fiducia che si aveva nel verificarsi di una determinata ipotesi $H$ (identificata con la probabilità assegnata all'ipotesi stessa) in conseguenza del verificarsi dell'evidenza $E$.

## 3.1 Il teorema di Bayes

**Teorema 3.1.** *Sia* $(H_i)_{i\geq 1}$ *una partizione dell'evento certo* $\Omega$ *e sia* $E \subseteq \Omega$ *un evento tale che* $p(E) > 0$*, allora, per* $i = 1, \dots, \infty$*:*

$$P(H_i \mid E) = \frac{P(E \mid H_i)P(H_i)}{\sum_{j=1}^{\infty} P(H_j)P(E \mid H_j)}. \tag{3.1}$$

La formula di Bayes contiene tre concetti fondamentali. I primi due distinguono il grado di fiducia precedente al verificarsi dell'evidenza $E$ da quello successivo al verificarsi dell'evidenza $E$. Pertanto, dati gli eventi $H, E \subseteq \Omega$

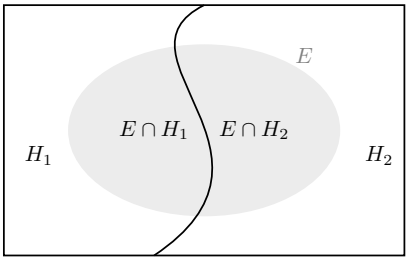
- si definisce *probabilità a priori* la probabilità che viene attribuita al verificarsi di $H$ prima di sapere che si è verificato l'evento $E$, seguendo l'approccio bayesiano, ovvero tenendo conto delle caratteristiche cognitive del decisore (esperienza, modo di pensare, ecc.);
- si definisce *probabilità a posteriori* la probabilità assegnata ad $H$, una volta che sia noto $E$, ovvero l'aggiornamento della probabilità a priori alla luce della nuova evidenza $E$.

Il terzo concetto definisce la probabilità che ha l'evento $E$ di verificarsi quando è vera l'ipotesi $H$, ovvero la probabilità dell'evidenza in base all'ipotesi. Pertanto, dati gli eventi $H, E \subseteq \Omega$

- si definisce *verosimiglianza* di $H$ dato $E$, la probabilità condizionata che si verifichi $E$, se è vera $H$: $P(E \mid H)$.

Si noti che per il calcolo della quantità a denominatore si ricorre al teorema della probabilità assoluta.

Per fare un esempio concreto, considerando una partizione dell'evento certo $\Omega$ in due soli eventi che chiamiamo ipotesi $H_1$ e $H_2$. Supponiamo conosciute le probabilità a priori $P(H_1)$ e $P(H_2)$. Consideriamo un terzo evento $E \subseteq \Omega$ con probabilità non nulla di cui si conosce la verosimiglianza, ovvero si conoscono le probabilità condizionate $\mathrm{P}(E \mid H_1)$ e $P(E \mid H_2)$. Supponendo che si sia verificato l'evento $E$, vogliamo conoscere le probabilità a posteriori delle ipotesi, ovvero $P(H_1 \mid E)$ e $P(H_2 \mid E)$.




Per trovare le probabilità cercate scriviamo:

$$\begin{aligned}\mathrm{P}(H_1 \mid E) &= \frac{\mathrm{P}(E \cap H_1)}{\mathrm{P}(E)} \\ &= \frac{\mathrm{P}(E \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(E)}.\end{aligned}$$

Sapendo che $E = (E \cap H_1) \cup (E \cap H_2)$ e che $H_1$ e $H_2$ sono eventi disgiunti, ovvero $H_1 \cap H_2 = \emptyset$, ne segue che possiamo calcolare $\mathrm{P}(E)$ utilizzando il teorema della probabilità assoluta:

$$\begin{aligned}\mathrm{P}(E) &= \mathrm{P}(E \cap H_1) + \mathrm{P}(E \cap H_2) \\ &= \mathrm{P}(E \mid H_1)P(H_1) + \mathrm{P}(E \mid H_2)\mathrm{P}(H_2).\end{aligned}$$

Sostituendo tale risultato nella formula precedente otteniamo:

$$\mathrm{P}(H_1 \mid E) = \frac{\mathrm{P}(E \mid H_1)\mathrm{P}(H_1)}{\mathrm{P}(E \mid H_1)\mathrm{P}(H_1) + \mathrm{P}(E \mid H_2)P(H_2)}. \tag{3.2}$$

Un lettore attento si sarà reso conto che, in precedenza, abbiamo già applicato il teorema di Bayes quando abbiamo risolto l'esercizio riportato nella Sezione 2.1. In quel caso, le due ipotesi erano "malattia presente", che possiamo denotare con $M$, e "malattia assente", $M^{\mathrm{C}}$. L'evidenza $E$ è costituita dal risultato positivo al test, ovvero $+$. Con questa nuova notazione la (3.2) diventa:

$$\mathrm{P}(M \mid +) = \frac{\mathrm{P}(+ \mid M)\mathrm{P}(M)}{\mathrm{P}(+ \mid M)\mathrm{P}(M) + \mathrm{P}(+ \mid M^{\mathrm{C}})\mathrm{P}(M^{\mathrm{C}})}$$

Inserendo i dati nella formula, otteniamo

$$\begin{aligned}\mathrm{P}(M \mid +) &= \frac{0.9 \cdot 10/1000}{0.9 \cdot 10/1000 + 99/990 \cdot 990/1000} \\ &= \frac{9}{108}.\end{aligned}$$

### Commenti e considerazioni finali

Il teorema di Bayes rende esplicito il motivo per cui la probabilità non possa essere pensata come uno stato oggettivo, quanto piuttosto come un'inferenza soggettiva e condizionata. Il denominatore del membro di destra della (3.1) è un semplice fattore di normalizzazione. Nel numeratore compaiono invece due quantità: $\mathrm{P}(H_i)$ e $\mathrm{P}(E \mid H_i)$. La probabilità $\mathrm{P}(H_i)$ è la probabilità *probabilità a priori* (*prior*) dell'ipotesi $H_i$ e rappresenta l'informazione che l'agente bayesiano possiede a proposito dell'ipotesi $H_i$. Diremo che $\mathrm{P}(H_i)$ codifica il grado di fiducia che l'agente ripone in $H_i$ precedentemente al verificarsi dell'evidenza $E$. Nell'interpretazione bayesiana, $\mathrm{P}(H_i)$ rappresenta un giudizio personale dell'agente e non esistono criteri esterni che possano determinare se tale giudizio sia coretto o meno. La probabilità condizionata $\mathrm{P}(E \mid H_i)$ rappresenta invece la verosimiglianza di $H_i$ dato $E$ e descrive la plausibilità che si verifichi l'evento $E$ se è vera l'ipotesi $H_i$. Il teorema di Bayes descrive la regola che l'agente deve seguire per aggiornare il suo grado di fiducia nell'ipotesi $H_i$ alla luce del verificarsi dell'evento $E$. La $\mathrm{P}(H_i \mid E)$ è chiamata probabilità a posteriori dato che rappresenta la nuova probabilità che l'agente assegna all'ipotesi $H_i$ affinché rimanga consistente con le nuove informazioni fornitegli da $E$.

La probabilità a posteriori dipende sia dall'evidenza $E$, sia dalla conoscenza a priori dell'agente $\mathrm{P}(H_i)$. È dunque chiaro come non abbia senso parlare di una probabilità oggettiva: per il teorema di Bayes la probabilità è definita condizionatamente alla probabilità a priori, la quale a sua volta, per definizione, è un'assegnazione soggettiva. Ne segue pertanto che ogni probabilità deve essere considerata come una rappresentazione del grado di fiducia soggettiva dell'agente. Dato che ogni assegnazione probabilistica rappresenta uno stato di conoscenza e che ciascun particolare stato di conoscenza è arbitrario, un accordo tra agenti diversi non è richiesto. Tuttavia, la teoria delle probabilità ci fornisce uno strumento che, alla luce di nuove evidenze, consente di aggiornare in un modo razionale il grado di fiducia che attribuiamo ad un'ipotesi, via via che nuove evidenze vengono raccolte, in modo tale da formulare un'ipotesi a posteriori la quale non è mai definitiva, ma può sempre essere aggiornata in base alle nuove evidenze disponibili. Questo processo si chiama *aggiornamento bayesiano*. Vedremo nel Capitolo **??** come estendere la (3.1) al caso continuo.

# 4

# *Probabilità congiunta*

La probabilità congiunta è la probabilità che due o più eventi si verifichino contemporaneamente. In questo Capitolo verrà esaminato in dettaglio il caso discreto.

## 4.1 Funzione di probabilità congiunta

Dopo aver trattato della distribuzione di probabilità di una variabile casuale, la quale associa ad ogni evento elementare dello spazio campionario uno ed un solo numero reale, è naturale estendere questo concetto al caso di due o più variabili casuali.

Iniziamo a descrivere il caso discreto con un esempio. Consideriamo l'esperimento casuale corrispondente al lancio di tre monete equilibrate. Lo spazio campionario è

$$\Omega = \{TTT, TTC, TCT, CTT, CCT, CTC, TCC, CCC\}.$$

Dato che i tre lanci sono tra loro indipendenti, non c'è ragione di aspettarsi che uno degli otto risultati possibili dell'esperimento sia più probabile degli altri, dunque possiamo associare a ciascuno degli otto eventi elementari dello spazio campionario la stessa probabilità, ovvero 1/8.

Siano $X \in \{0, 1, 2, 3\}$ = "numero di realizzazioni con il risultato testa nei tre lanci" e $Y \in \{0, 1\}$ = "numero di realizzazioni con il risultato testa nel primo lancio" due variabili casuali definite sullo spazio campionario $\Omega$. Indicando con T = 'testa' e C = 'croce', si ottiene la situazione riportata nella tabella 4.1.

***Tabella 4.1:*** *Spazio campionario dell'esperimento consistente nel lancio di tre monete equilibrate su cui sono state definite le variabili aleatorie $X$ e $Y$.*

| $\omega$ | $X$ | $Y$ | $P(\omega)$ |
|---|---|---|---|
| $\omega_1 =$ TTT | 3 | 1 | 1/8 |
| $\omega_2 =$ TTC | 2 | 1 | 1/8 |
| $\omega_3 =$ TCT | 2 | 1 | 1/8 |
| $\omega_4 =$ CTT | 2 | 0 | 1/8 |
| $\omega_5 =$ CCT | 1 | 0 | 1/8 |
| $\omega_6 =$ CTC | 1 | 0 | 1/8 |
| $\omega_7 =$ TCC | 1 | 1 | 1/8 |
| $\omega_8 =$ CCC | 0 | 0 | 1/8 |

Ci poniamo il problema di associare un livello di probabilità ad ogni coppia $(x, y)$ definita su $\Omega$. La coppia $(X = 0, Y = 0)$ si realizza in corrispondenza di un solo evento elementare, ovvero CCC; avrà dunque una probabilità pari a $P(X = 0, Y = 0) = P(CCC) = 1/8$. Nel caso della coppia $(X = 1, Y = 0)$ ci sono due eventi elementari che danno luogo al risultato considerato, ovvero, CCT e CTC; la probabilità $P(X = 1, Y = 0)$ sarà dunque data dalla probabilità dell'unione dei due eventi elementari, cioé $P(X = 1, Y = 0) = P(CCT \cup CTC) = 1/8+1/8 = 1/4$. Sono riportati qui sotto i calcoli per tutti i possibili valori di $X$ e $Y$.

$$
\begin{aligned}
&P(X = 0, Y = 0) = P(\omega_8 = CCC) = 1/8;\\
&P(X = 1, Y = 0) = P(\omega_5 = CCT) + P(\omega_6 = CTC) = 2/8;\\
&P(X = 1, Y = 1) = P(\omega_7 = TCC) = 1/8;\\
&P(X = 2, Y = 0) = P(\omega_4 = CTT) = 1/8;\\
&P(X = 2, Y = 1) = P(\omega_3 = TCT) + P(\omega_2 = TTC) = 2/8;\\
&P(X = 3, Y = 1) = P(\omega_1 = TTT) = 1/8;
\end{aligned}
$$

Le probabilità così trovate sono riportate nella tabella 4.2 la quale descrive la distribuzione di probabilità congiunta delle variabili casuali $X =$ "numero di realizzazioni con il risultato testa nei tre lanci" e $Y =$ "numero di realizzazioni con il risultato testa nel primo lancio" per l'esperimento casuale consistente nel lancio di tre monete equilibrate.

***Tabella 4.2:*** *Distribuzione di probabilità congiunta per i risultati dell'esperimento consistente nel lancio di tre monete equilibrate.*

| $x/y$ | 0 | 1 |
|---|---|---|
| 0 | 1/8 | 0 |
| 1 | 2/8 | 1/8 |
| 2 | 1/8 | 2/8 |
| 3 | 0 | 1/8 |

In generale, possiamo dire che, dato uno spazio campionario discreto $\Omega$, è possibile associare ad ogni evento elementare $\omega_i$ dello spazio campionario una coppia di numeri reali $(x, y)$, essendo $x = X(\omega)$ e $y = Y(\omega)$, il che ci conduce alla seguente definizione.

**Definizione 4.1.** Siano $X$ e $Y$ due variabili casuali. La funzione che associa ad ogni coppia $(x, y)$ un livello di probabilità prende il nome di funzione di probabilità congiunta:

$$P(x, y) = P(X = x, Y = y).$$

Il termine "congiunta" deriva dal fatto che questa probabilità è legata al verificarsi di una coppia di valori, il primo associato alla variabile casuale $X$ ed il secondo alla variabile casuale $Y$. Nel caso di due sole variabili casuali si parla di distribuzione bivariata, mentre nel caso di più variabili casuali si parla di distribuzione multivariata.

La regola della catena, $P(A \cap B) = P(A)P(B \mid A)$, permette il calcolo di qualsiasi membro della distribuzione congiunta di un insieme di variabili casuali utilizzando solo le probabilità condizionate. Nel caso di 4 eventi, per esempio, la regola della catena diventa

$$P(A_1, A_2, A_3, A_4) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2)P(A_4 \mid A_1, A_2, A_3).$$

### 4.1.1 Proprietà

Una distribuzione di massa di probabilità congiunta bivariata deve soddisfare due proprietà:

1. $0 \leq P(x_i, y_j) \leq 1$;

2. la probabilità totale deve essere uguale a 1.0. Tale proprietà può essere espressa nel modo seguente

$$\sum_i \sum_j P(x_i, y_j) = 1.0.$$

### 4.1.2 Eventi

Si noti che dalla probabilità congiunta possiamo calcolare la probabilità di qualsiasi evento definito in base alle variabili aleatorie $X$ e $Y$. Per capire come questo possa essere fatto, consideriamo nuovamente l'esperimento casuale discusso in precedenza.

**Esercizio 4.1.** Per la distribuzione di massa di probabilità congiunta riportata nella tabella precedente si trovi la probabilità dell'evento $X + Y \leq 1$.

Per trovare la probabilità richiesta dobbiamo semplicemente sommare le probabilità associate a tutte le coppie $(x, y)$ che soddisfano la condizione $X + Y \leq 1$, ovvero

$$P_{XY}(X + Y \leq 1) = P_{XY}(0, 0) + P_{XY}(1, 0) = 3/8.$$

### 4.1.3 Funzioni di probabilità marginali

Nel caso di due variabili casuali discrete $X$ e $Y$ di cui conosciamo la distribuzione congiunta, la distribuzione marginale di $X$ è calcolata sommando la distribuzione di probabilità congiunta sopra la variabile da "scartare", in questo caso la $Y$. La funzione di massa di probabilità marginale $P(X = x)$ è

$$P(X = x) = \sum_y P(X, Y = y) = \sum_y P(X \mid Y = y)P(Y = y), \quad (4.1)$$

dove $P(X = x, Y = y)$ è la distribuzione congiunta di $X, Y$, mentre $P(X = x \mid Y = y)$ è la distribuzione condizionata di $X$ dato $Y$. Se esaminiamo $P(X = x)$, diciamo che la variabile $Y$ è stata marginalizzata. Le probabilità bivariate marginali e congiunte per variabili casuali discrete sono spesso mostrate come tabelle di contingenza.

Si noti che $P(X = x)$ e $P(Y = y)$ sono normalizzate:

$$\sum_x P(X = x) = 1.0, \quad \sum_y P(Y = y) = 1.0.$$

Nel caso continuo si sostituisce l'integrazione alla somma – si veda la Sezione 4.5.

**Esercizio 4.2.** Per l'esperimento casuale consistente nel lancio di tre monete equilibrate, si calcolino le probabilità marginali di $X$ e $Y$.

Nell'ultima colonna a destra e nell'ultima riga in basso della tabella 4.3 sono riportate le distribuzioni di probabilità marginali di $X$ e $Y$. $P_X$ si ottiene sommando su ciascuna riga fissata la colonna $j$, $P_X(X = j) = \sum_y p_{xy}(x = j, y)$. $P_Y$ si trova sommando su ciascuna colonna fissata la riga $i$, $P_Y(Y = i) = \sum_x p_{xy}(x, y = i)$.

***Tabella 4.3:*** *Distribuzione di probabilità congiunta $p(x, y)$ per i risultati dell'esperimento consistente nel lancio di tre monete equilibrate e probabilità marginali $P(x)$ e $P(y)$.*

| $x/y$ | 0 | 1 | $P(x)$ |
|---|---|---|---|
| 0 | 1/8 | 0 | 1/8 |
| 1 | 2/8 | 1/8 | 3/8 |
| 2 | 1/8 | 2/8 | 3/8 |
| 3 | 0 | 1/8 | 1/8 |
| $P(y)$ | 4/8 | 4/8 | 1.0 |

## 4.2 Indipendenza stocastica incondizionata

In precedenza abbiamo visto come l'indipendenza stocastica di due eventi $A$ e $B$ si ha quando il verificarsi di uno non modifica la probabilità di verificarsi dell'altro, ovvero quando $P(A \mid B = P(A)$ e $P(B \mid A) = P(B)$. Queste due condizioni si possono sintetizzare con la formula $P(A \cap B) = P(A)P(B)$.

Analogamente, quando si afferma che due variabili casuali $X$ e $Y$ definite sullo stesso spazio campionario $\Omega$ sono indipendenti si afferma che conoscere qualcosa riguardo al valore di una di esse non apporta alcu-

na informazione circa il valore dell'altra. Formalmente, questo si verifica quando

$$P(X, Y) = P_X(x)P_Y(y). \tag{4.2}$$

Nel caso discreto, dunque, l'indipendenza implica che la probabilità riportata in ciascuna cella della tabella di probabilità congiunta deve essere uguale al prodotto delle probabilità marginali di riga e di colonna:

$$P(x_i, y_i) = P_X(x_i)P_Y(y_i).$$

**Esercizio 4.3.** Per la situazione rappresentata nella tabella 4.3 le variabili casuali $X$ e $Y$ sono indipendenti?

Nella tabella le variabili casuali $X$ e $Y$ non sono indipendenti: le probabilità congiunte non sono ricavabili dal prodotto delle marginali. Per esempio, nessuna delle probabilità marginali è uguale a 0 per cui nessuno dei valori dentro la tabella (probabilità congiunte) che risulta essere uguale a 0 può essere il prodotto delle probabilità marginali.

## 4.3 Indipendenza condizionata tra eventi

Sebbene l'indipendenza incondizionata sia una proprietà utile, non capita spesso di incontrare due eventi indipendenti. Una situazione più comune è quando due eventi sono indipendenti dato un terzo evento. Ad esempio, supponiamo di voler ragionare sulla possibilità che uno studente che è in possesso di un titolo di laurea triennale venga accettato al Corso di Laurea Magistrale (CdL) $A$ o al CdL Magistrale $B$. Nella maggior parte dei casi, questi due eventi non sono indipendenti. Se apprendiamo che lo studente è stato accettato al CdL $A$, la nostra stima della sua probabilità che venga accettato al CdL $B$ è ora più alta, poiché è aumentata la nostra credenza che lo studente in questione sia uno studente "promettente".

Ora, supponiamo che entrambi i CdL basino le loro decisioni unicamente sul voto di laurea triennale (chiamiamolo $C$) dello studente e supponiamo di sapere che, per lo studente in questione, $C = 105/110$. In questo caso, apprendere che lo studente è stato ammesso al CdL $A$ non cambia la

probabilità che venga ammesso al CdL $B$: il suo voto di laurea $V$ fornisce tutte le informazioni rilevanti circa la possibilità che lo studente venga ammesso al CdL $A$; sapere che è stato ammesso al CdL $B$ non aggiunge niente a tutto ciò. Formalmente, possiamo scrivere

$$P(A \mid B \cap C) = P(A \mid B) \tag{4.3}$$

Se la condizione precedente si verifica, gli eventi $A$ e $B$ si dicono condizionatamente indipendenti dall'evento $C$.

Alternativamente, possiamo dire che gli eventi $A$ e $B$ sono condizionatamente indipendenti dall'evento $C$ se e solo se

$$P(A \mid B \cap C) = P(A \mid C)P(B \mid C), \tag{4.4}$$

oppure, maniera equivalente se

$$P(A \mid B, C) = P(A \mid C).$$

Poiché la probabilità di $A$ dato $C$ è uguale alla probabilità di $A$ dati sia $B$ che $C$, questa uguaglianza esprime il fatto che $B$ non aggiunge nulla alla nostra conoscenza della probabilità di $A$.

Solitamente, l'indipendenza condizionata viene indicata utilizzando la notazione $(A \perp\!\!\!\perp B \mid C)$.

## 4.4 Indipendenza di variabili casuali

Siano $X, Y, Z$ tre variabili casuali. Diciamo che $X$ è condizionatamente indipendente da $Y$ data $Z$ in una distribuzione $P$ se $P$ soddisfa $(X = x \perp\!\!\!\perp Y = y \mid Z = z)$ per tutti i valori $x \in X$, $y \in Y$ e $z \in Z$. Se l'insieme $Z$ è vuoto, invece di scrivere $(X \perp\!\!\!\perp Y \mid \emptyset)$, scriviamo $(X \mid Y)$ e diciamo che $X$ e $Y$ sono *marginalmente indipendenti*.

Da ciò segue la seguente definizione alternativa di indipendenza condizionata.

**Definizione 4.2.** La distribuzione $P$ soddisfa $(X \perp\!\!\!\perp Y \mid Z)$ se e solo se

$$P(X, Y \mid Z) = (X \mid Z)P(Y \mid Z).$$

## 4.5 Marginalizzazione di variabili casuali continue

Nella trattazione della statistca bayesiana useremo spesso il concetto di "marginalizzazione" e vedremo equazioni come la seguente:

$$p(y) = \int_\theta p(y, \theta) = \int_\theta p(y \mid \theta)p(\theta), \tag{4.5}$$

laddove $y$ e $\theta$ sono due variabili casuali continue – nello specifico, con $y$ denoteremo i dati e con $\theta$ i parametri di un modello statistico. Per ora, possiamo pensare a $y$ e $\theta$ come a due variabili casuali qualsiasi. La (4.5) descrive la distribuzione marginale di $y$.

Per meglio comprendere la (4.5) possiamo esaminare il corrispondente caso discreto nel quale sostituiamo semplicemente l'integrale con una somma, il che ci riporta alla situazione descritta nella Sezione 4.1.3. Possiamo dunque scrivere:

$$p(y) = \sum_\theta p(y, \theta) = \sum_\theta p(y \mid \theta)p(\theta). \tag{4.6}$$

Esaminiamo un semplice esempio numerico. Siano $y$ e $\theta$ due variabili discrete aventi la distribuzione di massa di probabilità congiunta riportata nella tabella 4.4.

***Tabella 4.4:*** *Distribuzione di probabilità congiunta $p(y, \theta)$ per due variabili casuali discrete.*

| $y/\theta$ | 0 | 1 | $p(y)$ |
|---|---|---|---|
| 0 | 0.1 | 0.2 | 0.3 |
| 1 | 0.3 | 0.4 | 0.7 |
| $p(\theta)$ | 0.4 | 0.6 | 1.0 |

Applicando la (4.6), la distribuzione marginale $p(y) = \{0.3, 0.7\}$ può essere trovata nel modo seguente:

$$\begin{pmatrix} 0.1/0.4 \\ 0.3/0.4 \end{pmatrix} \cdot 0.4 + \begin{pmatrix} 0.2/0.6 \\ 0.4/0.6 \end{pmatrix} \cdot 0.6 = \begin{pmatrix} 0.3 \\ 0.7 \end{pmatrix}.$$

È possibiile pensare al caso continuo indicato nella (4.5) come all'estensione dell'esempio presente ad un numero infinito di valori $\theta$.

## Commenti e considerazioni finali

La funzione di probabilità congiunta tiene simultaneamente conto del comportamento di due variabili casuali $X$ e $Y$ e di come esse si influenzano reciprocamente. In particolare, si osserva che se le due variabili discrete $X$ e $Y$ non si influenzano, cioè se sono statisticamente indipendenti, allora la distribuzione di massa di probabilità congiunta si ottiene come prodotto delle funzioni di probabilità marginali di $X$ e $Y$: $P_{X,Y}(x, y) = P_X(x)P_Y(y)$.

# Parte II

# Il confronto bayesiano di modelli

# 5

# *Entropia*

Il principio base del metodo scientifico è la *replicabilità* delle osservazioni: le osservazioni che non possono essere replicate sono poco interessanti. Parallelamente, una caratteristica fondamentale di un modello scientifico è la *generalizzabilità*: se un modello è capace di descrivere soltanto le proprietà di uno specifico campione di osservazioni, allora è poco utile. Ma come è possibile valutare la generalizzabilità di un modello statistico? Questa è la domanda a cui cercheremo di rispondere in questa parte della dispensa. In questo Capitolo inizieremo questa discussione introducendo il concetto di entropia.

## 5.1 La generalizzabilità dei modelli

Secondo Johnson et al. (2022), nel valutare un modello, il ricercatore deve porsi tre domande critiche.

- Quali conseguenze più ampie derivano dall'inferenza? Come e chi ha raccolto i dati? Colui che svolge la ricerca otterrebbe di benefici manipolando i dati (escludendo delle osservazioni; selezionando il campione)? Che impatto hanno inferenze che vengono tratte dai dati sugli individui e sulla società? Quali pregiudizi o strutture di potere possono essere coinvolti in questa analisi?
- Che tipo di distorsioni sistematiche potrebbero essere presenti nell'analisi statistica? Ricordiamo la famosa citazione di George Box: "Tutti i modelli sono sbagliati, ma alcuni sono utili". È dunque importante sapere quanto è sbagliato il modello. Le assunzioni che stanno alla base del modello sono ragionevoli? Il meccanismo generatore dei dati che è stato ipotizzato è adeguato per il fenomeno in esame?
- Quanto è accurato il modello? Quanto sono lontane dalla realtà le

previsioni del modello?

Per approfondire questi temi, si rinvia al testo di Johnson et al. (2022). Qui ci concentreremo su uno dei temi critici relativa alla validità di un modello, ovvero sul tema della generalizzabilità del modello.

Nella scienza l'utilità di una teoria viene verificata esaminando la corrispondenza tra predizioni teoriche e osservazioni. Se vi sono discrepanze significative tra predizioni e osservazioni ciò suggerisce che la teoria, o nella nostra visione più ristretta, il modello statistico, è poco utile. Il problema della capacità predittiva del modello non riguarda soltanto l'adeguatezza del modello in riferimento ad uno specifico campione di dati, ma riguarda anche la capacità di un modello statistico sviluppato in un campione di dati di ben adattarsi ad altri campioni della stessa popolazione.

In generale, i modelli statistici tendono a non generalizzarsi bene a un nuovo campione; questo perché sfruttano le caratteristiche specifiche dei dati del campione e tendono a produrre risultati eccessivamente ottimistici (cioè le dimensioni dell'effetto) che sovrastimano la dimensione dell'effetto atteso sia nella popolazione che in nuovi campioni. Benché i problemi della generalizzabilità dei modelli e il metodo chiave per valutarli – ovvero, la convalida incrociata (*cross-validation*) – siano stati discussi sin dagli esordi della letteratura psicometrica (Lord, 1950), tali temi sono stati sottovalutati nella formazione psicologica contemporanea e nella ricerca. Tuttavia, questi concetti diventeranno sempre più importanti considerata l'enfasi corrente sulla necessità di condurre ricerche replicabili. Un'introduzione a questi temi è fornita, da esempio, da Song et al. (2021). Nello specifico, Song et al. (2021) mostrano che un modello che viene adattato a un campione (*campione di calibrazione*) non si generalizza bene a un altro campione (*campione di convalida*): la capacità predittiva del modello è minore quando il modello viene applicato al campione di convalida piuttosto che al campione di calibrazione. Questo problema è detto *sovra-adattamento* (*overfitting*). In generale, Song et al. (2021) mostrano come la capacità di generalizzazione del modello diminuisce (a) all'aumentare della complessità del modello, (b) al diminuire dell'ampiezza del campione di calibrazione, e (c) al diminuire della dimensione dell'effetto nella popolazione.

Sebbene i modelli statistici producono comunemente un sovra-adattamento, è anche possibile che essi producano un *sotto-adattamento* (*underfitting*) dei dati. Tale mancanza di adattamento è dovuta dal-

la variabilità campionaria e dalla complessità del modello. Il sotto-adattamento porta ad un $R^2$ basso e ad un *MSE* alto, sia nei campioni di calibrazione che in quelli di convalida. Per questo motivo, la scarsa generalizzabilità del modello può essere dovuta sia al sovra-adattamento che al sotto-adattamento del modello.

Per aumentarne la capacità di generalizzazione del modello devono essere soddisfatte tre condizioni: (a) campioni di calibrazione grandi, (b) dimensioni dell'effetto non piccole nella popolazione, e (c) modelli che non siano inutilmente complessi. Tuttavia, nella ricerca psicologica queste tre condizioni sono difficili da soddisfare: l'aumento della dimensione del campione spesso richiede l'utilizzo di maggiori risorse, la dimensione di un dato effetto nella popolazione non è soggetta alla discrezione dei ricercatori e la complessità del modello è spesso guidata da motivazioni teoriche. Pertanto, negli studi psicologici la generalizzabilità dei modelli è spesso problematica. Ciò rende necessario che il ricercatore fornisca informazioni aggiuntive relative alla capacità del modello di generalizzarsi a nuovi campioni. L'obiettivo di questa parte della dispensa è di descrivere come questo possa essere fatto utilizzando l'approccio bayesiano.

## 5.2 Capacità predittiva

Nel framework bayesiano il problema della generalizzabilità di un modello viene affrontato valutando la capacità predittiva del modello, laddove per capacità predittiva si intende la capacità di un modello, i cui parametri sono stati stimati usando le informazioni di un campione, di ben adattarsi ad un campione di osservazioni future. In questo Capitolo cercheremo di rispondere a tre domande.

1. Quali criteri consentono di valutare la capacità predittiva di un modello?
2. Come quantificare la capacità predittiva di un modello usando solo un campione di osservazioni?
3. Come confrontare le capacità predittive di modelli diversi?

## 5.3 Il rasoio di Ockham

Il problema di scegliere il modello più adatto a spiegare un fenomeno di interesse è uno dei più importanti problemi in campo scientifico. I ricercatori si chiedono: il modello è completo? È necessario aggiungere un nuovo parametro al modello? Come può essere migliorato il modello? Se ci sono modelli diversi, qual'è il modello migliore?

Per rispondere a queste domande è possibile usare il rasoio di Ockham: *frustra fit per plura quod potest fieri per pauciora* ("si fa inutilmente con molte cose ciò che si può fare con poche cose"). Parafrasando la massima si potrebbe dire: se due modelli descrivono i dati egualmente bene, viene sempre preferito il modello più semplice. Questo è il principio che sta alla base della ricerca scientifica.

Il rasoio di Ockham, però, non consente sempre di scegliere tra modelli alternativi. Se due modelli fanno le stesse predizioni ma differiscono in termini di complessità — per esempio, relativamente al numero di parametri di cui sono costituiti — allora è facile decidere: viene preferito il modello più semplice, anche perché, pragmaticamente, è il più facile da usare. Tuttavia, in generale, i modelli differiscono sia per complessità (ovvero, per il numero di parametri) che per accuratezza (ovvero, per la grandezza degli errori di predizione). In tali circostanze il rasoio di Ockham non è sufficiente: non consente infatti di trovare un equilibrio tra accuratezza e semplicità.

In questo Capitolo ci chiederemo come sia possibile misurare l'accuratezza predittiva di un modello. Ciò ci consentirà, in seguito, di usare il rasoio di Ockham: a parità di accuratezza, sarà possibile scegliere il modello più semplice. Ma nella pratica scientifica non si sacrifica mai l'accuratezza per la semplicità: il criterio prioritario è sempre l'accuratezza.

### 5.3.1 Sovra-adattamento e sotto-adattamento

Secondo McElreath (2020), la selezione tra modelli deve evitare due opposti errori: il sovra-adattamento e il sotto-adattamento. Tale problema va sotto il nome di *bias-variance trade-off*: il sotto-adattamento, infatti, porta a distorsioni (*bias*) nella stima dei parametri, mentre il sovra-adattamento porta a previsioni scadenti in campioni futuri. Spesso l'incertezza relativa alla scelta del modello (sotto-adattamento ver-

sus sovra-adattamento) passa inosservata ma il suo impatto può essere drammatico. Secondo Hoeting et al. (1999), *"Standard statistical practice ignores model uncertainty. Data analysts typically select a model from some class of models and then proceed as if the selected model had generated the data. This approach ignores the uncertainty in model selection, leading to over-confident inferences and decisions that are more risky than one thinks they are.*

In questo Capitolo esamineremo alcune tecniche bayesiane che possono essere utilizzate per operare una selezione tra modelli alternativi, tenendo sotto controllo i pericoli del sovra-adattamento e del sotto-adattamento. In particolare, ci chiederemo quale, tra due o più modelli, sia quello da preferire in base al criterio della capacità predittiva.

### 5.3.2 Stargazing

Nella pratica concreta della ricerca, il metodo più comune per la selezione tra modelli alternativi utilizza i test di ipotesi statistiche di stampo frequentista. Questo metodo viene chiamato *stargazing*, poiché richiede soltanto l'esame degli asterischi (∗∗) che si trovano nell'output di un software statistico (gli asterischi marcano i coefficienti del modello che sono "statisticamente significativi"): alcuni ricercatori ritengono che il modello con più stelline sia anche il modello migliore. Questo però non è vero. Al di là dei problemi legati ai test dell'ipotesi nulla, è sicuramente un errore usare i test di significatività per la selezione di modelli: i valori-$p$ non consentono di trovare un equilibrio tra *underfitting* e *overfitting*. Infatti, le variabili che migliorano la capacità predittiva di un modello non sono sempre statisticamente significative; d'altra parte, le variabili statisticamente significative non sempre migliorano la capacità predittiva di un modello.

Quando ci chiediamo quale, tra modelli alternativi, è il modello che meglio rappresenta il "vero" processo di generazione dei dati, ci troviamo di fronte al problema di quantificare il grado di "vicinanza" di un modello al "vero" processo di generazione dei dati. Si noti che, in tale confronto, facciamo riferimento sia alla famiglia distributiva così come ai valori dei parametri. Ad esempio, il modello $y_i \sim \mathcal{N}(5, 3)$ è diverso dal modello $y_i \sim \mathcal{N}(5, 6)$, ed è anche diverso dal modello $y_i \sim \Gamma(2, 2)$. I primi due modelli appartengono alla stessa famiglia distributiva ma differiscono nei termini dei valori dei parametri; gli ultimi due modelli appartengono a famiglie distributive diverse (gaussiano vs. Gamma). Per misurare il

grado di "vicinanza" tra due modelli, $\mathcal{M}_1$ e $\mathcal{M}_2$, la metrica di gran lunga più popolare è la *divergenza di Kullback-Leibler*. Per chiarire questo concetto è però prima necessario introdurre la nozione di entropia.

## 5.4 La misura del disordine

Se vogliamo ottenere una comprensione intuitiva del concetto di entropia[1] possiamo pensare a quant'è informativa una distribuzione. Maggiore è l'entropia di una distribuzione, meno informativa sarà quella distribuzione e più uniformemente verranno assegnate le probabilità agli eventi. In altri termini, ottenere la risposta di "42" è più informativo della risposta "42 $\pm$ 5", che a sua volta è più informativo della risposta "un numero qualsiasi". L'entropia quantifica questa osservazione qualitativa.

Il concetto di entropia si applica sia alle distribuzioni continue sia a quelle discrete, ma è più facile da capire usando le distribuzioni discrete. Negli esempi successivi vedremo alcuni esempi applicati al caso discreto, ma gli stessi concetti si applicano al caso continuo.

### 5.4.1 Entropia di un singolo evento

Il concetto di entropia può essere usato per descrivere la quantità di informazione fornita da un evento. L'intuizione che sta alla base del concetto di entropia è che l'informazione fornita da un evento descrive la sorpresa suscitata dall'evento: gli eventi rari (a bassa probabilità) sono più sorprendenti – e quindi forniscono più informazione – degli eventi comuni (ad alta probabilità). In altre parole,

- un evento a bassa probabilità è sorprendente e fornisce molta informazione;
- un evento ad alta probabilità è poco o per niente sorprendente e fornisce poca (o nessuna) informazione.

[1] La nozione di entropia fu introdotta agli inizi del XIX secolo nel campo della termodinamica classica; il secondo principio della termodinamica è infatti basato sul concetto di entropia che, in generale, è assunto come una misura del disordine di un sistema fisico. Successivamente Boltzmann fornì una definizione statistica di entropia. Nel 1948 Shannon impiegò la nozione di entropia nell'ambito della teoria delle comunicazioni.

È dunque possibile quantificare l'informazione fornita dal verificarsi di un evento usando la probabilità di quell'evento. Una tale *quantità di informazione* è chiamata "informazione di Shannon", "auto-informazione" o semplicemente "informazione" e, per un evento discreto $x$, può essere calcolata come:

$$\text{informazione}(x) = -\log_2 p(x),$$

dove $\log_2$ è il logaritmo in base 2 e $p(x)$ è la probabilità dell'evento $x$.

La scelta del logaritmo in base 2 significa che l'unità di misura dell'informazione è il bit (cifre binarie). Questo può essere interpretato dicendo che l'informazione misura il numero di bit richiesti per rappresentare un evento.[2] Solitamente, si denota la quantità di informazione con $h()$:

$$h(x) = -\log p(x).$$

Il segno negativo garantisce che il risultato sia sempre positivo o zero. L'informazione è zero quando la probabilità dell'evento è 1.0, ovvero quando l'evento è certo (assenza di sorpresa).

**Esempio 5.1.** Consideriamo il lancio di una moneta equilibrata. La probabilità di testa (e croce) è 0.5. La quantità di informazione di ottenere "testa" è dunque

```
-log2(0.5)
#> [1] 1
```

Per rappresentare questo evento abbiamo bisogno di 1 bit di informazione. Se la stessa moneta venisse lanciata $n$ volte, la quantità di informazione necessaria per rappresentare questo evento (ovvero, questa

[2] È possibile pensare all'entropia nei termini del numero di domande sì/no che devono essere poste per ridurre l'incertezza. Per esempio, se in un certo giorno ci può essere solo sole o pioggia, per ridurre l'incertezza, a fine giornata chiediamo: "ha piovuto?" La risposta (sì/no) ad una singola domanda elimina l'incertezza, e quindi l'informazione ottenuta (ovvero, la riduzione dell'incertezza) è uguale ad 1 bit. Se in una certa giornata ci potrebbero essere sole, pioggia o neve, per ridurre l'incertezza sono necessarie due domande: "c'era sole?"; "ha piovuto?" In questo secondo caso, l'informazione ottenuta (ovvero, la riduzione dell'incertezza) è uguale ad 2 bit. Usando un logaritmo in base 2, dunque, l'entropia può essere interpretata come il numero minimo di bit necessari per codificare la quantità di informazione nei dati.

sequenza di lanci) sarebbe pari a $n$ bit. Se la moneta non è equilibrata e la probabilità di testa è 0.1, allora l'evento "testa" è più raro e richiede più di 3 bit di informazione:

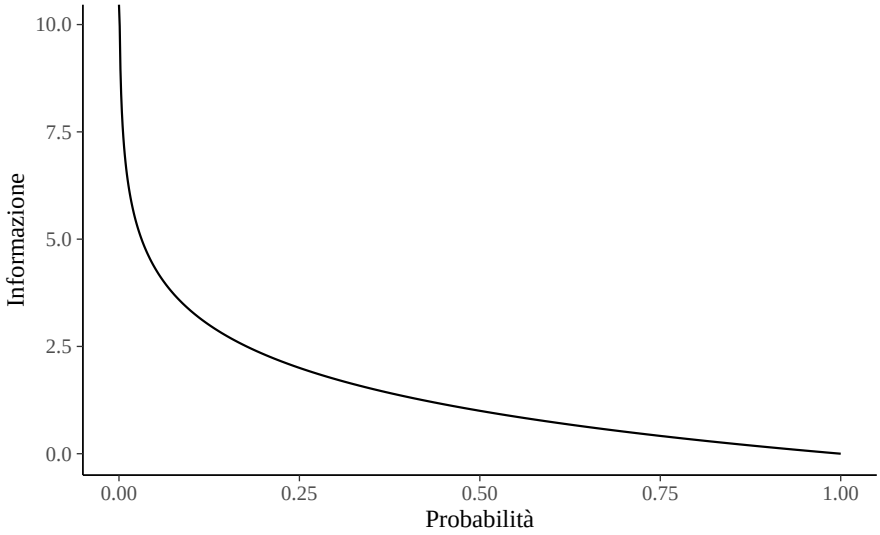
```
-log2(0.1)
#> [1] 3.322
```

Consideriamo ora il lancio di un dado. Quanta informazione viene fornita, ad esempio, dall'evento "esce il numero 6"? Dato che la probabilità di ottenere un 6 nel lancio di un dado è più piccola della probabilità di ottenere "testa" nel lancio di una moneta, il risultato del lancio di un dado deve produrre una sorpresa maggiore del risultato del lancio di una moneta. Per cui, la quantità di informazione associata all'evento "è uscito 6", dovrà essere maggiore di quella associata all'evento "testa". Infatti, la quantità di informazione dell'evento "è uscito un 6" è più che doppia rispetto alla quantità di informazione dell'evento "testa":

```
-log2(1 / 6)
#> [1] 2.585
```

**Esempio 5.2.** Nella figura successiva viene esaminata la relazione tra probabilità e informazione, per valori di probabilità nell'intervallo tra 0 e 1.

```
p <- seq(0, 1, length.out = 1000)
h <- -log2(p)
ggplot(tibble(p, h), aes(p, h)) +
  geom_line() +
  labs(
    x = "Probabilità",
    y = "Informazione"
  )
```




La figura mostra che questa relazione non è lineare, è infatti leggermente sublineare. Questo ha senso dato che abbiamo usato una funzione logaritmica.

### 5.4.2 Entropia di una variabile casuale

Possiamo estendere questa discussione pensando ad un insieme di eventi, ovvero ad una distribuzione. Nella teoria della probabilità usiamo la nozione di variabile casuale per fare riferimento ad un insieme di eventi e alle probabilità associate a tali eventi. L'entropia quantifica l'informazione che viene fornita da una variabile casuale.

**Definizione 5.1.** Sia $Y = y_1, \dots, y_n$ una variabile casuale e $p_t(y)$ una distribuzione di probabilità su $Y$. Si definisce la sua entropia (detta di Shannon) come:

$$H(Y) = -\sum_{i=1}^{n} p_t(y_i) \cdot \log_2 p_t(y_i). \tag{5.1}$$

Per interpretare la (5.1), consideriamo un esempio discusso da Martin et al. (2022).

Nella figura 5.1 sono rappresentate sei distribuzioni. viene anche riportato il valore di entropia di ciascuna distribuzione. La distribuzione con il picco più pronunciato o con la dispersione minore è `q`, e questa è la

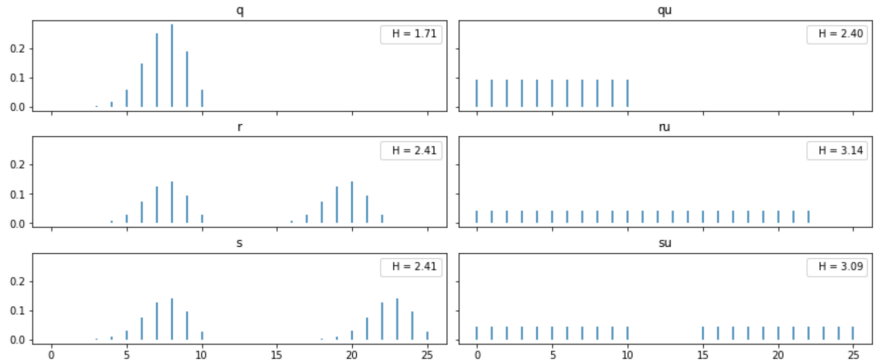



***Figura 5.1:*** *Funzioni di massa di probabilità e associata entropia.*

distribuzione con il valore di entropia più basso tra le sei distribuzioni considerate. Per `q` la distribuzione è `q ~ binom(n = 10, p = 0.75)`; quindi ci sono 11 possibili eventi. `qu` ha una distribuzione uniforme sugli stessi 11 possibili eventi. L'entropia di `qu` è maggiore dell'entropia di `q`. Infatti, se calcoliamo l'entropia di distribuzioni binomiali con $n = 10$ (con valori diversi di $p$) ci rendiamo conto che nessuna di tali distribuzioni ha un'entropia maggiore di `qu`. Dobbiamo aumentare $n \approx 3$ volte per trovare la prima distribuzione binomiale con entropia maggiore di `qu`. Passiamo alla riga successiva. Generiamo la distribuzione `r` spostando a destra `q` e normalizzando (per garantire che la somma di tutte le probabilità sia 1). Poiché `r` ha una dispersione maggiore di `q`, la sua entropia è maggiore. `ru` è una distribuzione uniforme con lo stesso numero di eventi possibili come `r` (ovvero 22) – si noti che sono stati inclusi come valori possibili anche quelli nella "valle" tra i due picchi. Ancora una volta, la distribuzione uniforme ha l'entropia più grande.

Gli esempi discussi finora sembrano suggerire che l'entropia è proporzionale alla varianza della distribuzione. Verifichiamo questa intuizione esaminiamo le ultime due distribuzioni della figura 5.1. La distribuzione `s` è simile a `r` ma presenta una separazione maggiore tra i due picchi della distribuzione – dunque, ha una varianza più grande. Ciò nonostante, l'entropia non varia. Quindi la relazione tra entropia e varianza non è così semplice come ci sembrava. Il risultato che abbiamo trovato può essere spiegato dicendo che, nel calcolo dell'entropia, non vengono considerati gli eventi con probabilità nulla (per questa ragione, nell'esempio, è stato possibile aumentare la varianza senza cambiare l'entropia). La distribuzione `su` è stata costruita sostituendo i due picchi in `s` con `qu`

(e normalizzando). Possiamo vedere che `su` ha un'entropia minore di `ru`, anche se `su` ha una dispersione maggiore di `ru`. Questo è dovuto al fatto che `su` distribuisce la probabilità totale tra un numero minore di eventi (22) di `ru` (che ne conta 23); quindi è sensato attribuire a `su` un'entropia minore di `ru`.

**Esempio 5.3.** Consideriamo ora un esempio riguardante le previsioni del tempo. Supponiamo che le probabilità di pioggia e sole siano, rispettivamente, $p_1 = 0.3$ e $p_2 = 0.7$. Quindi

$$H(p) = -[p(y_1) \log_2 p(y_1) + p(y_2) \log_2 p(y_2)] \approx 0.61.$$

Se però viviamo a Las Vegas, allora le probabilità di pioggia e sole saranno simili a $p(y_1) = 0.01$ e $p(y_2) = 0.99$. In questo secondo caso, l'entropia è 0.06, ovvero, molto minore di prima. Infatti, a Las Vegas non piove quasi mai, per cui quando abbiamo imparato che, in un certo giorno, non ha piovuto, abbiamo imparato molto poco rispetto a quello che già sapevamo in precedenza.

**Esempio 5.4.** Nell'esempio precedente abbiamo visto che, se gli esiti possibili sono pioggia o sole con $p(y_1) = 0.7$, $p(y_2) = 0.3$, allora l'entropia è

```
-(0.7 * log(0.7) + 0.3 * log(0.3))
#> [1] 0.6109
```

Ma se gli esiti possibili sono pioggia, neve o sole con $p(y_1) = 0.7$, $p(y_2) = 0.15$ e $p(y_3) = 0.15$, rispettivamente, allora l'entropia cresce:

```
-(0.7 * log(0.7) + 0.15 * log(0.15) + 0.15 * log(0.15))
#> [1] 0.8188
```

## Commenti e considerazioni finali

In questo Capitolo abbiamo visto come sia possibile quantificare l'incertezza tramite l'entropia. Ma come è possibile usare l'entropia dell'infor-

mazione per specificare la "distanza" tra un modello e il vero meccanismo generatore dei dati? La risposta a questa domanda è fornita dalla divergenza di Kullback-Leibler che verrà discussa nel Capitolo 6.

# 6

# *La divergenza di Kullback-Leibler*

È comune in statistica utilizzare una distribuzione di probabilità $q$ per approssimare un'altra distribuzione $p$ – generalmente, questo viene fatto se $p$ non è conosciuta o è troppo complessa. In questi casi possiamo chiederci quanta informazione venga perduta usando $q$ al posto di $p$, o equivalentemente, quanta incertezza aggiuntiva viene introdotta nell'analisi statistica. La quantificazione di questo incremento di incertezza è fornita dalla divergenza di Kullback-Leibler.

## 6.1 La perdita di informazione

Intuitivamente, per quantificare l'informazione che si perde quando una distribuzione approssimata $q$ viene usata in luogo della distribuzione corretta $p$ sembra necessaria una quantità che ha valore zero quando $q = p$, e un valore positivo altrimenti. Seguendo la definizione (5.1) di entropia, possiamo quantificare una tale perdita di informazione mediante il valore atteso della differenza tra $\log(p)$ e $\log(q)$. Questa quantità è chiamata *entropia relativa* o *divergenza di Kullback-Leibler*:

$$\mathbb{KL}(p \,||\, q) = \mathbb{E}(\log p - \log q). \tag{6.1}$$

La divergenza $\mathbb{KL}(p \,||\, q)$ corrisponde alla differenza media nelle probabilità logaritmiche quando $q$ viene usato per approssimare $p$. Poiché gli eventi si manifestano secondo $p$, è necessario calcolare il valore atteso rispetto a $p$. Per distribuzioni discrete dunque abbiamo:

$$\mathbb{KL}(p \,||\, q) = \sum_i^n p_i(\log p_i - \log q_i) = \sum_i^n p_i \log \frac{p_i}{q_i}. \tag{6.2}$$

Riarrangiando i termini otteniamo:

$$\mathbb{KL}(p \mid\mid q) = -\sum_{i}^{n} p_i(\log q_i - \log p_i), \tag{6.3}$$

ovvero,

$$\mathbb{KL}(p \mid\mid q) = \underbrace{-\sum_{i}^{n} p_i \log q_i}_{h(p,q)} - \underbrace{\left(-\sum_{i}^{n} p_i \log p_i\right)}_{h(p)}, \tag{6.4}$$

laddove $h(p)$ è l'entropia di $p$ e $h(p, q) = -\mathbb{E}[\log q]$ può essere intesa come l'entropia di $q$, ma valutata secondo i valori di probabilità $p$.

Riarrangiando l'equazione precedente otteniamo:

$$h(p, q) = h(p) + \mathbb{KL}(p \mid\mid q), \tag{6.5}$$

il che mostra come la divergenza $\mathbb{KL}$ possa essere interpretata come l'incremento di entropia, rispetto a $h(p)$, quando $q$ viene usata per rappresentare $p$.

**Esempio 6.1.** (da McElreath, 2020) Sia la distribuzione target $p = \{0.3, 0.7\}$. Supponiamo che la distribuzione approssimata $q$ possa assumere valori da $q = \{0.01, 0.99\}$ a $q = \{0.99, 0.01\}$. Calcoliamo la divergenza KL.

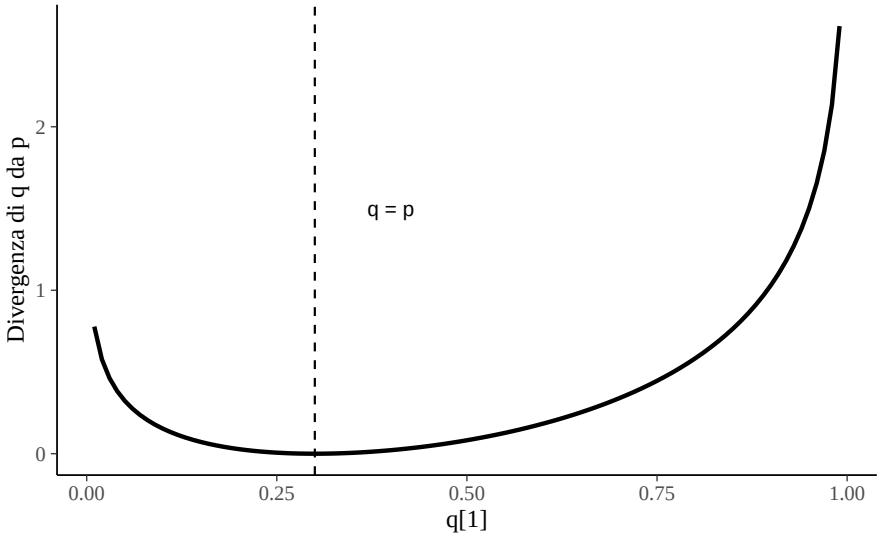
Le istruzioni R sono le seguenti:

```
t <-
  tibble(
    p_1 = .3,
    p_2 = .7,
    q_1 = seq(from = .01, to = .99, by = .01)
  ) %>%
  mutate(
    q_2 = 1 - q_1
  ) %>%
  mutate(
    d_kl = (p_1 * log(p_1 / q_1)) + (p_2 * log(p_2 / q_2))
  )
```

```
head(t)
#> # A tibble: 6 x 5
#>     p_1   p_2   q_1   q_2  d_kl
#>   <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1   0.3   0.7  0.01  0.99 0.778
#> 2   0.3   0.7  0.02  0.98 0.577
#> 3   0.3   0.7  0.03  0.97 0.462
#> 4   0.3   0.7  0.04  0.96 0.383
#> 5   0.3   0.7  0.05  0.95 0.324
#> 6   0.3   0.7  0.06  0.94 0.276
```

Nella figura seguente sull'asse delle ascisse sono rappresentati i valori $q$ e sull'asse delle ordinante sono riportati i corrispondenti valori $\mathbb{KL}$.

```
t %>%
  ggplot(aes(x = q_1, y = d_kl)) +
  geom_vline(xintercept = .3, linetype = 2) +
  geom_line(size = 1) +
  annotate(
    geom = "text", x = .4, y = 1.5, label = "q = p",
    size = 3.5
  ) +
  labs(
    x = "q[1]",
    y = "Divergenza di q da p"
  )
```




Tanto meglio la distribuzione $q$ approssima la distribuzione target tanto più piccolo è il valore di divergenza KL.

**Esempio 6.2.** Sia $p$ una distribuzione binomiale di parametri $\theta = 0.2$ e $n = 5$

```
n <- 4
p <- 0.2
true_py <- dbinom(0:n, n, 0.2)
true_py
#> [1] 0.4096 0.4096 0.1536 0.0256 0.0016
```

Sia $q_1$ una approssimazione a $p$:

```
q1 <- c(0.46, 0.42, 0.10, 0.01, 0.01)
q1
#> [1] 0.46 0.42 0.10 0.01 0.01
```

Sia $q_2$ una distribuzione uniforme:

```
q2 <- rep(0.2, 5)
q2
#> [1] 0.2 0.2 0.2 0.2 0.2
```

La divergenza KL di $q_1$ da $p$ è

```
sum(true_py * log(true_py / q1))
#> [1] 0.02925
```

La divergenza KL di $q_2$ da $p$ è:

```
sum(true_py * log(true_py / q2))
#> [1] 0.4864
```

È chiaro che perdiamo una quantità maggiore di informazioni se, per descrivere la distribuzione binomiale $p$, usiamo la distribuzione uniforme $q_2$ anziché $q_1$.

## 6.2 La divergenza dipende dalla direzione

La divergenza KL non è una vera e propria metrica: per esempio, non è simmetrica. In generale, $\mathbb{KL}(p \mid\mid q) \neq \mathbb{KL}(q \mid\mid p)$, ovvero la $\mathbb{KL}$ da $p$ a $q$ è diversa dalla $\mathbb{KL}$ da $q$ a $p$.

**Esempio 6.3.** Usando le seguenti istruzioni R otteniamo:

```
tibble(
  direction = c("Da q a p", "Da p a q"),
  p_1 = c(.01, .7),
  q_1 = c(.7, .01)
) %>%
  mutate(
    p_2 = 1 - p_1,
    q_2 = 1 - q_1
  ) %>%
  mutate(d_kl = (p_1 * log(p_1 / q_1)) + (p_2 * log(p_2 / q_2)))
#> # A tibble: 2 x 6
#>   direction   p_1   q_1   p_2   q_2  d_kl
```

```
#>   <chr>     <dbl> <dbl> <dbl> <dbl> <dbl>
#> 1 Da q a p  0.01  0.7   0.99  0.3   1.14
#> 2 Da p a q  0.7   0.01  0.3   0.99  2.62
```

## 6.3 Confronto tra modelli

La divergenza KL viene utilizzata nel confronto tra modelli, ovvero ci consente di quantificare l'informazione che viene perduta quando utilizziamo la distribuzione di probabilità ipotizzata da un modello, chiamiamola $p_{\mathcal{M}}$, per approssimare la distribuzione di probabilità del vero modello generatore dei dati, $p_t$.

Nel Capitolo **??** abbiamo introdotto il concetto di distribuzione predittiva a posteriori:

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

La distribuzione predittiva a posteriori descrive il tipo di dati che ci aspettiamo vengano prodotti dal modello generativo $\mathcal{M}$, alla luce delle nostre credenze iniziali, $p(\theta)$ e dei dati osservati $y$. Quando valutiamo un modello ci chiediamo in che misura $p_{\mathcal{M}}(\tilde{y} \mid y)$ approssimi $p_t(\tilde{y})$. Cioè, ci chiediamo quanto siano simili i dati $p_{\mathcal{M}}(\cdot)$ prodotti dal modello $\mathcal{M}$ ai dati prodotti dal vero processo generatore dei dati $p_t(\cdot)$.

Una misura della "somiglianza" tra la distribuzione $q_{\mathcal{M}}$ ipotizzata dal modello $\mathcal{M}$ e la distribuzione $p_t$ del vero modello generatore dei dati è fornita dalla divergenza di Kullback-Leibler $\mathbb{KL}(p_t \mid\mid q_{\mathcal{M}})$. Supponendo di avere $k$ modelli della distribuzione a posteriori, $\{q_{\mathcal{M}_1}, q_{\mathcal{M}_2}, \dots, q_{\mathcal{M}_k}\}$, e di conoscere il vero modello generatore dei dati, possiamo scrivere

$$\begin{aligned}
\mathbb{KL}(p_t \mid\mid q_{\mathcal{M}_1}) &= \mathbb{E}(\log p_{\mathcal{M}_0}) - \mathbb{E}(\log q_{\mathcal{M}_1}) \\
\mathbb{KL}(p_t \mid\mid q_{\mathcal{M}_2}) &= \mathbb{E}(\log p_t) - \mathbb{E}(\log q_{\mathcal{M}_2}) \\
&\dots \\
\mathbb{KL}(p_t \mid\mid q_{\mathcal{M}_k}) &= \mathbb{E}(\log p_{\mathcal{M}_0}) - \mathbb{E}(\log q_{\mathcal{M}_k}).
\end{aligned} \tag{6.6}$$

La (6.6) può sembrare un esercizio futile poiché nella vita reale non conosciamo il vero modello generatore dei dati. È però facile rendersi conto che, poiché $p_t$ è la stessa per tutti i confronti, diventa possibile costruire un ordinamento dei modelli basato unicamente sul secondo termine della (6.6), ovvero senza nessun riferimento al vero modello generatore dei dati. Per un generico modello $\mathcal{M}$, il secondo termine della (6.6) può essere scritto come:

$$\mathbb{E} \log p_{\mathcal{M}}(y) = \int_{-\infty}^{+\infty} p_t(y) \log p_{\mathcal{M}}(y) \, \mathrm{d}y. \tag{6.7}$$

## 6.4 Expected log predictive density

Le previsioni del modello $\mathcal{M}$ sui nuovi dati futuri sono date dalla distribuzione predittiva a posteriori. Possiamo dunque riscrivere la (6.7) come

$$\text{elpd} = \int_{\tilde{y}} p_t(\tilde{y}) \log p(\tilde{y} \mid y) \, \mathrm{d}\tilde{y}. \tag{6.8}$$

La (6.8) è chiamata *expected log predictive density* (elpd) e fornisce la risposta al problema che ci eravamo posti: nel confronto tra modelli, come è possibile scegliere il modello più simile al vero meccanismo generatore dei dati? Possiamo pensare alla (6.8) dicendo che descrive la distribuzione predittiva a posteriori del modello ponderando la verosimiglianza dei possibili (sconosciuti) dati futuri ($\tilde{y}$) con la vera distribuzione $p_t$. Di conseguenza, valori elpd più grandi identificano il modello che risulta più simile al vero meccanismo generatore dei dati.

Non dobbiamo preoccuparci di trovare una formulazione analitica della distribuzione predittiva a posteriori $p(\tilde{y} \mid y)$ perché, come abbiamo visto nel Capitolo **??**, è possibile approssimare tale distribuzione mediante simulazione. Notiamo però che la (6.8) include un termine, $p_t(\tilde{y})$, il quale descrive la distribuzione dei dati futuri $\tilde{y}$ secondo il vero modello generatore dei dati. Il termine $p_t$, ovviamente, è ignoto.[1] Di conseguenza, la quantità elpd non può mai essere calcolata in maniera esatta, ma può

[1] Se il modello sottostante i dati fosse noto non avremmo bisogno di cercare il modello migliore, perché $p_t$ è il modello migliore.

solo essere stimata. Il secondo problema di questo Capitolo è capire come la (6.8) possa essere stimata utilizzando un campione di osservazioni.

### 6.4.1 Log pointwise predictive density

Ingenuamente, potremmo pensare di stimare la (6.8) ipotizzando che la distribuzione del campione coincida con $p_t$. Usare la distribuzione del campione come proxy del vero modello generatore dei dati (ovvero, ipotizzare che la distribuzione del campione rappresenti fedelmente $p_t$) comporta due conseguenze:

- non è necessario ponderare per $p_t$, in quanto assumiamo che la distribuzione empirica del campione corrisponda a $p_t$ (ciò significa assumere che i valori più comunemente osservati nel campione siano anche quelli più verosimili nella vera distribuzione $p_t$);
- dato che il campione è finito, anziché eseguire un'operazione di integrazione possiamo semplicemente sommare la densità predittiva a posteriori delle osservazioni.

Questo conduce alla seguente equazione:[2]

$$\frac{1}{n}\sum_{i=1}^{n}\log p(y_i^{rep} \mid y). \tag{6.9}$$

La quantità (6.9), senza il passaggio finale della divisione per il numero di osservazioni, è chiamata *log pointwise predictive density* (lppd)

$$\text{lppd} = \sum_{i=1}^{n}\log p(y_i^{rep} \mid y) \tag{6.10}$$

e corrisponde alla somma delle densità predittive logaritmiche delle $n$ osservazioni. Valori più grandi della (6.10) sono da preferire perché indicano una maggiore accuratezza media. È anche comune vedere espressa la quantità precedente nei termini della *devianza*, ovvero alla lppd moltiplicata per -2. In questo secondo caso sono da preferire valori piccoli.

[2] In riferimento alla notazione, ricordiamo che Gelman et al. (2014) distinguono tra $y^{rep}$ e $\tilde{y}$. I valori $y^{rep}$ corrispondono ad un'altra possibile realizzazione del medesimo modello statistico che ha prodotto $y$ mediante determinati valori dei parametri $\theta$ (repliche sotto lo stesso modello statistico). I valori $\tilde{y}$ corrispondono invece ad un campione empirico di dati osservato in qualche futura occasione.

È importante notare che lppd fornisce una *sovrastima* della (6.8). Tale sovrastima è dovuta al fatto che, nel calcolo della (6.10), abbiamo usato $p(y^{rep} \mid y)$ al posto di $p(\tilde{y} \mid y)$: in altri termini, abbiamo considerato le osservazioni del campione come se fossero un nuovo campione di dati. In una serie di simulazioni, McElreath (2020) esamina il significato di questa sovrastima. Nelle simulazioni la devianza viene calcolata come funzione della complessità (ovvero, il numero di parametri) del modello. La simulazione mostra che lppd aumenta al crescere del numero di parametri del modello. Ciò significa che lppd mostra lo stesso limite del coefficiente di determinazione: aumenta all'aumentare della complessità del modello.

**Esempio 6.4.** Esaminiamo un esempio tratto da Bayesian Data Analysis for Cognitive Science[3] nel quale la elpd viene calcolata in forma esatta oppure mediante approssimazione. Supponiamo di disporre di un campione di $n$ osservazioni. Supponiamo inoltre di conoscere il vero processo generativo dei dati (qualcosa che in pratica non è mai possibile), ovvero:

$$p_t(y) = \mathrm{B}(1, 3).$$

I dati sono

```
set.seed(75)
n <- 10000
y_data <- rbeta(n, 1, 3)
head(y_data)
#> [1] 0.55062 0.13346 0.80251 0.21431 0.01913 0.08677
```

Supponiamo inoltre di avere adattato ai dati un modello bayesiano $\mathcal{M}$ e di avere ottenuto la distribuzione a posteriori per i parametri del modello. Inoltre, supponiamo di avere derivato la forma analitica della distribuzione predittiva a posteriori per il modello:

$$p(y^{rep} \mid y) \sim \mathrm{B}(2, 2).$$

Questa distribuzione ci dice quanto sono credibili i possibili dati futuri.

[3] https://vasishth.github.io/bayescogsci/book/expected-log-predictive-density-of-a-model.html

Conoscendo la vera distribuzione dei dati $p_t(y)$ possiamo calcolare in forma esatta la quantità elpd, ovvero

$$\text{elpd} = \int_{y^{rep}} p_t(y^{rep}) \log p(y^{rep} \mid y) \, \mathrm{d}y^{rep}.$$

Svolgiamo i calcoli in R otteniamo:

```
# True distribution
p_t <- function(y) dbeta(y, 1, 3)
# Predictive distribution
p <- function(y) dbeta(y, 2, 2)
# Integration
integrand <- function(y) p_t(y) * log(p(y))
integrate(f = integrand, lower = 0, upper = 1)
#> -0.3749 with absolute error < 6.8e-07
```

Tuttavia, in pratica non conosciamo mai $p_t(y)$. Quindi approssimiamo elpd usando la (6.8):

$$\frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid y).$$

Così facendo, e svolgendo i calcoli in R, otteniamo un valore diverso da quello trovato in precedenza:

```
1 / n * sum(log(p(y_data)))
#> [1] -0.3639
```

## Commenti e considerazioni finali

Dato che non conosciamo il vero meccanismo generatore dei dati $p$, possiamo usare la distribuzione dei dati osservata come proxy per la vera distribuzione $p$. Quindi, invece di ponderare la distribuzione predittiva in base alla densità reale di tutti i possibili dati futuri, utilizziamo semplicemente le $n$ osservazioni che abbiamo. Possiamo farlo perché assumiamo

che le nostre osservazioni costituiscano un campione dalla vera distribuzione dei dati: in base a questa ipotesi, nel campione ci aspettiamo di osservare più frequentemente quelle osservazioni che hanno una maggiore verosimiglianza nella vera distribuzione $p$. È così possibile giungere ad una stima numerica della elpd chiamata *log pointwise predictive density* (lppd).

# *Bibliografia*


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

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417.

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

Johnson, A. A., Ott, M., and Dogucu, M. (2022). *Bayes Rules! An Introduction to Bayesian Modeling with R*. CRC Press.

Lord, F. M. (1950). Efficiency of prediction when a regression equation from one sample is used in a new sample. *ETS Research Bulletin Series*, 1950(2):1–6.

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

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press, Boca Raton, Florida, 2nd edition edition.

Song, Q. C., Tang, C., and Wee, S. (2021). Making sense of model generalizability: A tutorial on cross-validation in r and shiny. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920947067.