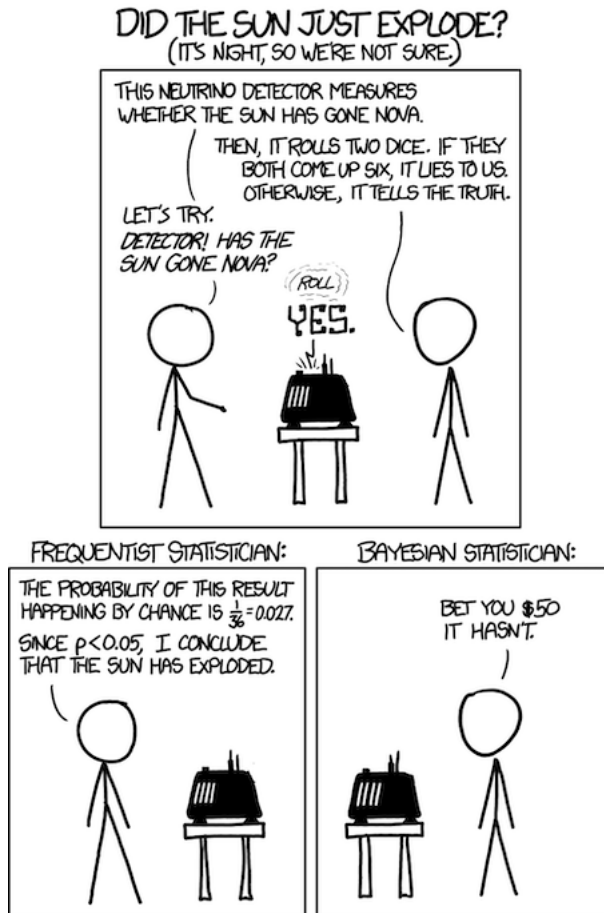


Corrado Caudek

Data Science per psicologi



Psicometria – AA 2021/2022





Indice

Elenco delle figure	ix
Elenco delle tabelle	xiii
Prefazione	xv
I Il calcolo delle probabilità	1
1 La logica dell'incerto	3
1.1 Che cos'è la probabilità?	3
1.2 Variabili casuali e probabilità di un evento	6
1.2.1 Eventi e probabilità	6
1.2.2 Spazio campione e risultati possibili	7
1.3 Variabili casuali	7
1.4 Usare la simulazione per stimare le probabilità	8
1.5 La legge dei grandi numeri	11
1.6 Variabili casuali multiple	14
1.7 Funzione di massa di probabilità	16
2 Probabilità condizionata	21
2.1 Probabilità condizionata su altri eventi	21
2.2 La regola moltiplicativa	24
2.3 L'indipendenza stocastica	26
2.4 Il teorema della probabilità totale	26
2.5 Il teorema della probabilità assoluta	28
2.6 Indipendenza condizionale	30
3 Il teorema di Bayes	33
3.1 Il teorema di Bayes	33
4 Probabilità congiunta	37
4.1 Funzione di probabilità congiunta	37

4.1.1	Proprietà	39
4.1.2	Eventi	40
4.1.3	Funzioni di probabilità marginali	40
4.2	Indipendenza stocastica incondizionata	41
4.3	Indipendenza condizionata tra eventi	42
4.4	Indipendenza di variabili casuali	43
4.5	Marginalizzazione di variabili casuali continue	44
5	La densità di probabilità	47
5.1	Spinner e variabili casuali continue uniformi	47
5.1.1	Il paradosso delle variabili casuali continue	49
5.2	La funzione di ripartizione per una variabile casuale continua	49
5.3	La distribuzione uniforme	52
5.4	Dagli istogrammi alle densità	56
5.5	Funzione di densità di probabilità	59
5.6	La funzione di ripartizione	60
6	Valore atteso e varianza	63
6.1	Valore atteso	63
6.1.1	Interpretazione	64
6.1.2	Proprietà del valore atteso	65
6.1.3	Variabili casuali continue	67
6.2	Varianza	67
6.2.1	Formula alternativa per la varianza	68
6.2.2	Variabili casuali continue	69
6.3	Deviazione standard	69
6.4	Standardizzazione	70
6.5	Momenti di variabili casuali	70
II	Distribuzioni teoriche di probabilità	1
7	Distribuzioni di v.c. discrete	3
7.1	Una prova Bernoulliana	3
7.2	Una sequenza di prove Bernoulliane	4
7.2.1	Valore atteso e deviazione standard	7
7.3	Distribuzione di Poisson	8
7.3.1	Alcune proprietà della variabile di Poisson	10
7.4	Distribuzione discreta uniforme	10
7.4.1	Usiamo R	11

7.5	Distribuzione beta-binomiale	12
8	Distribuzioni di v.c. continue	13
8.1	Distribuzione Normale	13
8.1.1	Limite delle distribuzioni binomiali	13
8.2	La Normale prodotta con una simulazione	16
8.2.1	Concentrazione	20
8.2.2	Funzione di ripartizione	21
8.2.3	Distribuzione Normale standard	21
8.3	Teorema del limite centrale	25
8.4	Distribuzione Chi-quadrato	26
8.4.1	Proprietà	27
8.5	Distribuzione t di Student	28
8.5.1	Proprietà	28
8.6	Funzione beta di Eulero	29
8.7	Distribuzione Beta	30
8.8	Distribuzione di Cauchy	35
8.9	Distribuzione log-normale	35
8.10	Distribuzione di Pareto	36



Elenco delle figure

1.1	Stima della probabilità di successo in funzione del numero di lanci di una moneta.	12
1.2	Stima della probabilità di successo in funzione del numero di lanci di una moneta – scala logaritmica.	13
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.	18
2.1	Rappresentazione ad albero che riporta le frequenze attese dei risultati di una mammografia in un campione di 1,000 donne.	23
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.	28
2.3	Partizione dell'evento certo Ω in tre sottoinsiemi sui quali viene definito l'evento E	29
5.1	Uno spinner che riposa a 36 gradi, o il dieci per cento del percorso intorno al cerchio. La pendenza dello spinner può assumere qualunque valore tra 0 e 360 gradi.	48
5.2	Funzione di distribuzione cumulativa per l'angolo θ (in gradi) risultante da una rotazione di uno spinner simmetrico. La linea tratteggiata mostra il valore a 180 gradi, che corrisponde ad una probabilità di 0.5, e la linea tratteggiata a 270 gradi, che corrisponde ad una probabilità di 0.75.	50

5.3	Grafico della funzione di ripartizione di una variabile casuale Θ che rappresenta il risultato di una rotazione di uno spinner simmetrico. Come previsto, tale funzione è una semplice funzione lineare perché la variabile sottostante Θ ha una distribuzione uniforme.	51
5.4	Istogramma di 10 000 realizzazioni $\Theta \sim \text{Uniform}(0, 1)$. . .	54
5.5	Istogramma di M campioni casuali $\Theta \sim \text{Uniform}(0, 1)$ trasformati in valori $\Phi = \text{logit}(\Theta)$. Il profilo limite dell'istogramma è evidenziato nella figura in basso a destra che è stata costruita usando 1 000 000 di osservazioni. . .	57
5.6	Istogramma di $M = 1\,000\,000$ campioni casuali $\Theta \sim \text{Uniform}(0, 1)$ trasformati in valori $\Phi = \text{logit}(\Theta)$. La spezzata nera congiunge i punti centrali superiori delle barre dell'istogramma. Nel limite, quando il numero di osservazioni e di barre tende all'infinito, tale spezzata approssima la funzione di densità di probabilità della variabile casuale.	59
7.1	Alcune distribuzioni binomiali. Nella figura, il parametro θ è indicato con p	6
7.2	Alcune distribuzioni di Poisson.	9
8.1	Probabilità del numero di successi in $N = 10$ prove bernoulliane indipendenti, ciascuna con una probabilità di successo di 0.90. Il risultato è una distribuzione $\text{Binomial}(y \mid 10, 0.9)$. Con solo dieci prove, la distribuzione è fortemente asimmetrica negativa.	14
8.2	Probabilità del numero di successi in $N = 1000$ prove bernoulliane indipendenti, ciascuna con una probabilità di successo di 0.90. Il risultato è una distribuzione $\text{Binomial}(y \mid 1000, 0.9)$. Con mille prove, la distribuzione è quasi simmetrica a forma campanulare.	15
8.3	Passeggiata casuale di 4, 8 e 16 passi. La spezzata nera indica la media delle distanze dall'origine come funzione del numero di passi.	17
8.4	Kernel density plot dei risultati della passeggiata casuale riportata nella figura precedente, dopo 4, 8 e 16 passi. Nel pannello di destra, una densità Normale di opportuni parametri è sovrapposta all'istogramma lisciato.	19
8.5	Alcune distribuzioni Normali.	20

<i>Elenco delle figure</i>	xi
8.6 Alcune distribuzioni Chi-quadrato.	26
8.7 Alcune distribuzioni t di Student.	29
8.8 Alcune distribuzioni Beta.	32



Elenco delle tabelle

4.1	Spazio campionario dell'esperimento consistente nel lancio di tre monete equilibrate su cui sono state definite le variabili aleatorie X e Y	38
4.2	Distribuzione di probabilità congiunta per i risultati dell'esperimento consistente nel lancio di tre monete equilibrate.	39
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 $P(y)$	41
4.4	Distribuzione di probabilità congiunta $p(y, \theta)$ per due variabili casuali discrete.	44



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 affrontare, 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, oppure 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 carcarle 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

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

xx

Prefazione

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 conoscenza 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 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.

L'incertezza nelle nostre previsioni può sorgere per due ragioni fondamentalmente diverse. Il primo è dovuto alla nostra ignoranza delle cause nascoste sottostanti o dei meccanismi che generano i dati. Questa è appunto un'incertezza *epistemica*, poiché epistemologia è lo studio della conoscenza. Il secondo tipo di incertezza deriva dalla variabilità intrinseca, che non può essere ridotta anche se raccogliamo più dati. Questo è talvolta chiamato incertezza *aleatoria*. Come esempio concreto, considera di lanciare una moneta equilibrata. Potremmo sapere con certezza che la probabilità di testa è $p = 0.5$, quindi non c'è incertezza epistemica, ma non questo non è sufficiente per metterci in grado di prevedere con certezza il risultato.

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 varierà 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 soggettiva di probabilità la quale sembra negare la va-

lità dell'intero costruito 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, la quale può essere vista come un'estensione della logica.

1.2.1 Eventi e probabilità

Nella teoria delle probabilità il risultato “testa” nel lancio di una moneta è chiamato *evento*.¹ Un evento, denotato da una variabile binaria, corrisponde ad uno stato del mondo che si verifica oppure no. Ad esempio, $Y = 1$ può denotare l'evento per cui il lancio di una moneta produce il 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: $P(Y = 1 \wedge Y = 0) = 0$. Gli eventi $Y = 1$ e $Y = 0$ dicono *esaustivi*, nel senso che almeno uno di essi deve verificarsi e nessun altro tipo di evento è possibile. Nella notazione probabilistica, $P(Y = 1 \vee Y = 0) = 1$.

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 $P(A \wedge B) = 0$. 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 $P(A \wedge B) > 0$.

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

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.

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 Ω 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 quello 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]"  
  )  
)
```

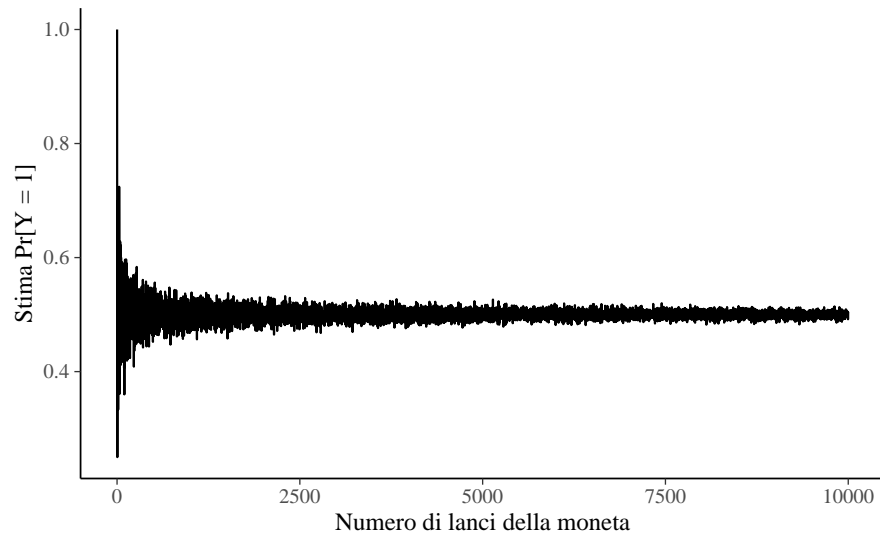


Figura 1.1: Stima della probabilità di successo in funzione del numero di lanci di una moneta.

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]"
  )
)
```

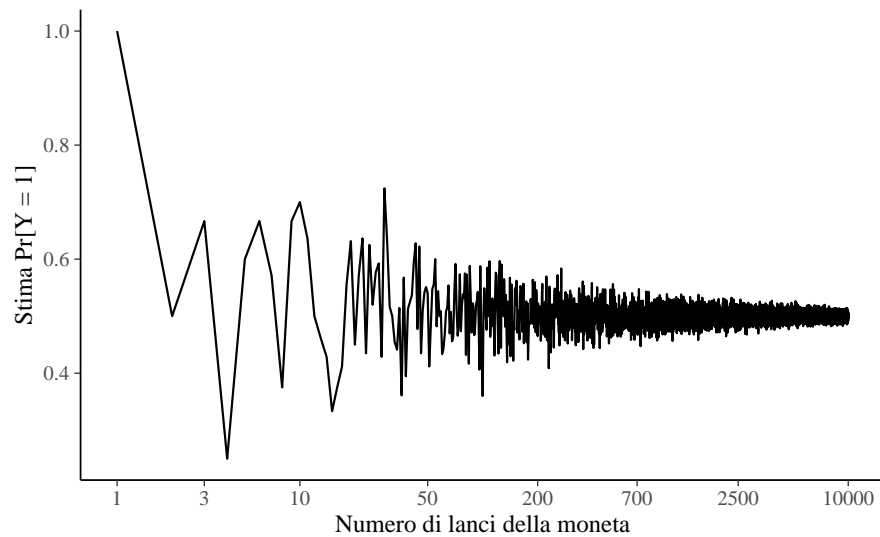


Figura 1.2: Stima della probabilità di successo in funzione del numero di lanci di una moneta – scala logaritmica.

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 equilibrata. 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 $Z = Y_1 + \dots + Y_4$ come il numero di risultati “testa” in 4 lanci della moneta, allora possiamo definire la seguente funzione:

$p_Z(0)$	$=$	$1/16$	TTTT
$p_Z(1)$	$=$	$4/16$	HTTT, THTT, TTHT, TTTH
$p_Z(2)$	$=$	$6/16$	HHTT, HTHT, HTTH, THHT, THTH, TTTH
$p_Z(3)$	$=$	$4/16$	HHHT, HHHT, HTHH, THHH
$p_Z(4)$	$=$	$1/16$	HHHH

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

$$P_Z(z) = P(Z = z).$$

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

```

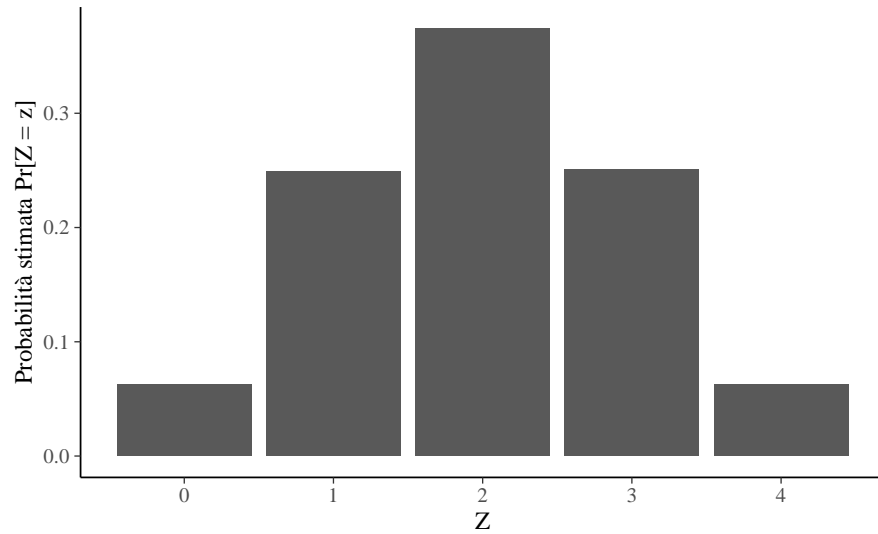


Figura 1.3: Grafico di $M = 100\,000$ simulazioni della funzione di massa di probabilità di una variabile casuale definita come il numero di teste in quattro lanci di una moneta equilibrata.

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

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

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 ridefinizione 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.

Per introdurre la probabilità condizionata, [Albert and Hu \(2019\)](#) utilizzando il famoso paradosso delle tre carte. “Ci sono tre carte, delle quali la prima (A) è rossa su entrambi i lati, la seconda (B) su un lato è rossa e sull'altro è bianca e la terza (C) è bianca su entrambi i lati. Ponendo su un tavolo una delle tre carte, scelta a caso, ottengo che il lato visibile è di colore rosso. Qual è la probabilità che anche il lato non visibile sia di colore rosso? La risposta intuitiva porta solitamente a rispondere che la probabilità ricercata sia pari al 50%, in quanto solo due carte (la A e la B) possono mostrare il colore rosso e solo una di queste (la A) può mostrare anche sull'altro lato il colore rosso; tuttavia si dimostra che la risposta giusta è $2/3$.” (da Wikipedia)

[Albert and Hu \(2019\)](#) propongono di risolvere il problema con una simulazione in R: prima di tutto si sceglie una carta a caso, e poi si sceglie un lato della carta. Ci sono tre carte possibili, che chiamiamo “c_rossa”, “c_bianca”, e “c_entrambi”. Per la carta rossa, ci sono due lati rossi;

per la carta bianca ci sono due lati bianchi e la carta “entrambi” ha un lato rosso e un lato bianco. Ripetiamo l’esperimento 1,000 volte e classifichiamo i risultati per tipo di carta e lato:

```
set.seed(84735)
df <- tibble(
  Carta = c(
    "c_rossa", "c_rossa", "c_bianca", "c_bianca", "c_entrambi",
    "c_entrambi"
  ),
  Lato = c(
    "rosso", "rosso", "bianco", "bianco", "rosso", "bianco"
  )
)
df
#> # A tibble: 6 x 2
#>   Carta      Lato
#>   <chr>    <chr>
#> 1 c_rossa  rosso
#> 2 c_rossa  rosso
#> 3 c_bianca bianco
#> 4 c_bianca bianco
#> 5 c_entrambi rosso
#> 6 c_entrambi bianco
carte <- sample_n(df, 1000, replace = TRUE)
table(carte$Carta, carte$Lato)
#>
#>           bianco rosso
#> c_bianca      353     0
#> c_entrambi    143    160
#> c_rossa         0    344
```

Se si osserva il colore rosso (seconda colonna nella tabella precedente), questo risultato è dovuto ad una carta rossa in 344 casi e ad una carta “entrambi” in 160 casi. Quindi, nella simulazione il colore rosso (344 + 160) è associato ad una carta rossa in circa $2/3$ dei casi – ciò significa che, se il lato visibile è di colore rosso, allora c’è una probabilità di $2/3$ che anche il lato non visibile sia di colore rosso. Questo esempio dimostra come le nostre intuizioni a proposito della probabilità condizionata non

sono sempre corrette. Consideriamo un altro problema più articolato.

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 $\frac{9}{108} = 8\%$.

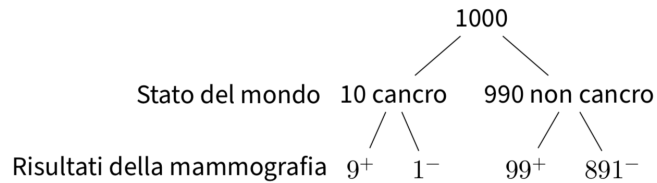


Figura 2.1: Rappresentazione ad albero che riporta le frequenze attese dei risultati di una mammografia in un campione di 1,000 donne.

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 $P(A | B)$ 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

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

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

Concludiamo con un problema veramente semplice per consolidare la nostra comprensione del concetto di probabilità condizionata.

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 La regola moltiplicativa

Dalla definizione di probabilità condizionata è possibile esprimere la probabilità congiunta tramite le condizionate. La regola moltiplicativa (o legge delle probabilità composte, 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:

$$P(A \cap B) = P(B)P(A | B) = P(A)P(B | A). \quad (2.2)$$

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

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

Per esempio, nel caso di quattro eventi abbiamo

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

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)

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

$P(B_1)$ vale 6/10, perché nella prima estrazione Ω è costituito da 10 elementi: 6 palline bianche e 4 nere. La probabilità condizionata $P(B_2 | B_1)$ 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

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

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

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

dove la probabilità $P(B_3 | B_1 \cap B_2)$ 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 $P(B_3 | B_1 \cap B_2) = 1/2$ 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:

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

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:

$$\begin{aligned} P(A | B) &= \frac{P(A)P(B)}{P(B)} = P(A), \\ P(B | A) &= \frac{P(A)P(B)}{P(A)} = P(B). \end{aligned}$$

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

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

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:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \quad (2.5)$$

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:

$$P(A \cup B) = P(A) + P(B) - P(A)P(B). \quad (2.6)$$

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: $A = \{\text{esce un 1 o un 2 nel primo lancio}\}$ e $B = \{\text{il punteggio totale è 8}\}$. Gli eventi A e B sono indipendenti?

Rappresentiamo qui sotto lo spazio campionario dell'esperimento casuale.

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

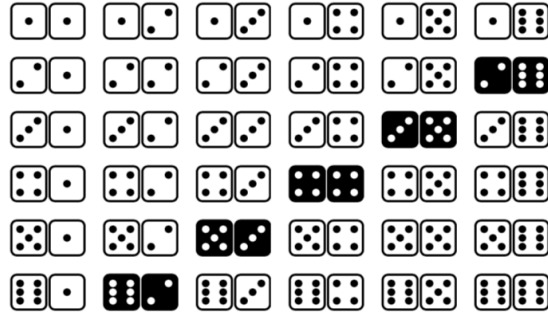


Figura 2.2: Rappresentazione dello spazio campionario dei risultati dell'esperimento casuale corrispondente al lancio di due dadi bilanciati. Sono evidenziati gli eventi elementari che costituiscono l'evento A : esce un 1 o un 2 nel primo lancio.

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

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 Ω , ovvero

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

Nel caso di una partizione dello spazio campione in n sottoinsiemi

abbiamo

$$P(E) = \sum_{i=1}^n P(H_i \cap E) = \sum_{i=1}^n P(H_i)P(E | H_i). \quad (2.7)$$

Consideriamo, ad esempio, una partizione dell'evento certo in tre soli sottoinsiemi.

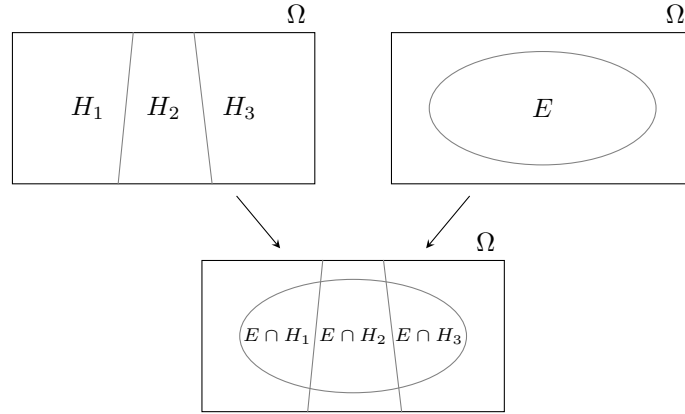


Figura 2.3: Partizione dell'evento certo Ω in tre sottoinsiemi sui quali viene definito l'evento E .

In tali circostanze si ha che

$$P(E) = P(E \cap H_1) + P(E \cap H_2) + P(E \cap H_3),$$

ovvero

$$P(E) = P(E | H_1)P(H_1) + P(E | H_2)P(H_2) + P(E | H_3)P(H_3). \quad (2.8)$$

In base al teorema della probabilità assoluta, dunque, 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 $P(E | H_i)$, 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.

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

$$P(R | U_1) = 0.75, \quad P(R | U_2) = 0.60, \quad P(R | U_3) = 0.45.$$

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, $P(U_1 \cup U_2 \cup U_3) = 1.0$. In base al teorema della probabilità assoluta, la probabilità di estrarre una pallina rossa è dunque

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

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:

$$P(A | B, C) = P(A | C).$$

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 indipendenti 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 condizionalmente 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 Ω e sia $E \subseteq \Omega$ un evento tale che $p(E) > 0$, allora, per $i = 1, \dots, \infty$:*

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

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 $H, E \subseteq \Omega$

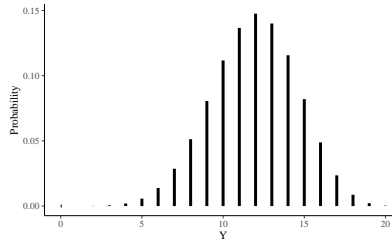
- si definisce *probabilità a priori* la probabilità che viene attribuita al verificarsi di H prima di sapere che si è verificato l'evento E , seguendo l'approccio bayesiano, ovvero tenendo conto delle caratteristiche cognitive del decisore (esperienza, modo di pensare, ecc.);
- si definisce *probabilità 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 : $P(E | H)$.

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 Ω 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 $P(E | H_1)$ e $P(E | H_2)$. Supponendo che si sia verificato l'evento E , vogliamo conoscere le probabilità a posteriori delle ipotesi, ovvero $P(H_1 | E)$ e $P(H_2 | E)$.



Per trovare le probabilità cercate scriviamo:

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

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 $P(E)$ utilizzando il teorema della probabilità assoluta:

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

Sostituendo tale risultato nella formula precedente otteniamo:

$$P(H_1 | E) = \frac{P(E | H_1)P(H_1)}{P(E | H_1)P(H_1) + P(E | H_2)P(H_2)}. \quad (3.2)$$

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^c . L'evidenza E è costituita dal risultato positivo al test, ovvero $+$. Con questa nuova notazione la (3.2) diventa:

$$P(M | +) = \frac{P(+ | M)P(M)}{P(+ | M)P(M) + P(+ | M^c)P(M^c)}$$

Inserendo i dati nella formula, otteniamo

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

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: $P(H_i)$ e $P(E | H_i)$. La probabilità $P(H_i)$ 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 $P(H_i)$ codifica il grado di fiducia che l'agente ripone in H_i precedentemente al verificarsi dell'evidenza E . Nell'interpretazione bayesiana, $P(H_i)$ rappresenta un giudizio personale dell'agente e non esistono criteri esterni che possano determinare se tale giudizio sia corretto o meno. La probabilità condizionata $P(E | H_i)$ 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 $P(H_i | E)$ 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 $P(H_i)$. 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 Ω . 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 .

ω	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 Ω . 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, cioè $P(X = 1, Y = 0) = P(CCT \cup CTC) = 1/8 + 1/8 = 1/4$. 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 Ω , è possibile associare ad ogni evento elementare ω_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, $P(A \cap B) = P(A)P(B | A)$, 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

$$P(A_1, A_2, A_3, A_4) = P(A_1)P(A_2 | A_1)P(A_3 | A_1, A_2)P(A_4 | A_1, A_2, A_3).$$

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

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

dove $P(X = x, Y = y)$ è la distribuzione congiunta di X, Y , mentre $P(X = x | Y = y)$ 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 Ω sono indipendenti si afferma che conoscere qualcosa riguardo al valore di una di esse non apporta alcun-

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

$$P(X, Y) = P_X(x)P_Y(y). \quad (4.2)$$

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

$$P(A \mid B \cap C) = P(A \mid B) \quad (4.3)$$

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

$$P(A \mid B \cap C) = P(A \mid C)P(B \mid C), \quad (4.4)$$

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 $(X \perp\!\!\!\perp Y)$ e diciamo che X e Y sono *marginamente 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

$$P(X, Y | Z) = (X | Z)P(Y | Z).$$

4.5 Marginalizzazione di variabili casuali continue

Nella trattazione della statistica bayesiana useremo spesso il concetto di “marginalizzazione” e vedremo equazioni come la seguente:

$$p(y) = \int_{\theta} p(y, \theta) = \int_{\theta} p(y | \theta)p(\theta), \quad (4.5)$$

laddove y e θ sono due variabili casuali continue – nello specifico, con y denoteremo i dati e con θ i parametri di un modello statistico. Per ora, possiamo pensare a y e θ 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:

$$p(y) = \sum_{\theta} p(y, \theta) = \sum_{\theta} p(y | \theta)p(\theta). \quad (4.6)$$

Esaminiamo un semplice esempio numerico. Siano y e θ 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/θ	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 possibile pensare al caso continuo indicato nella (4.5) come all'estensione dell'esempio presente ad un numero infinito di valori θ .

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 : $P_{X,Y}(x,y) = P_X(x)P_Y(y)$.



5

La densità di probabilità

Finora abbiamo considerato solo variabili casuali discrete, cioè variabili che assumono solo valori interi. Ma cosa succede se vogliamo usare variabili casuali per rappresentare lunghezze o volumi o distanze una qualsiasi delle altre proprietà continue nel mondo fisico (o psicologico)? È necessario generalizzare l'approccio usato finora.

Le variabili casuali continue assumono valori reali. L'insieme dei numeri reali è *non numerabile* perché è più grande dell'insieme degli interi.¹ Le leggi della probabilità sono le stesse per le variabili casuali discrete e quelle continue. La nozione di funzione di massa di probabilità, invece, deve essere sostituita dal suo equivalente continuo, ovvero dalla funzione di densità di probabilità. Lo scopo di questo Capitolo è quello di chiarire il significato di questa nozione, usando un approccio basato sulle simulazioni.

5.1 Spinner e variabili casuali continue uniformi

Consideriamo il seguente esperimento casuale. Facciamo ruotare ad alta velocità uno spinner simmetrico imperniato su un goniometro e osserviamo la posizione in cui si ferma (individuata dall'angolo acuto con segno tra il suo asse e l'asse orizzontale del goniometro). Chiamiamo Θ la variabile casuale "pendenza dello spinner". Nella trattazione seguente useremo i gradi e, di conseguenza, $\Theta \in [0, 360]$.

Cosa implica per Θ dire che lo spinner è simmetrico? Possiamo dire che, in ciascuna prova, la rotazione dello spinner produce un angolo qualunque da 0 a 360 gradi. In altri termini, un valore Θ compreso tra

¹Georg Cantor dimostrò che era impossibile mappare uno a uno i reali negli interi, dimostrando così che l'insieme dei reali è non numerabile.

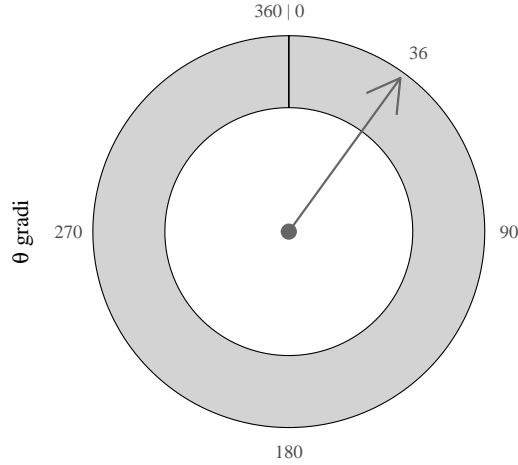


Figura 5.1: Uno spinner che riposa a 36 gradi, o il dieci percento del percorso intorno al cerchio. La pendenza dello spinner può assumere qualunque valore tra 0 e 360 gradi.

0 e 36 gradi ha la stessa probabilità di essere osservato di un valore Θ compreso tra 200 e 236 gradi. Inoltre, poiché 36 gradi è un decimo del percorso intorno al cerchio, la probabilità di ottenere un qualsiasi intervallo di 36 gradi sarà sempre uguale al 10%. Ovvero $P[0 \leq \Theta \leq 36] = \frac{1}{10}$ e $P[200 \leq \Theta \leq 236] = \frac{1}{10}$.

È importante notare che le considerazioni precedenti non si riferiscono al fatto che Θ può assumere uno specifico valore, ma piuttosto alla probabilità di osservare Θ in un particolare intervallo di valori. In generale, la probabilità che la pendenza Θ dello spinner cada in intervallo è la frazione del cerchio rappresentata dall'intervallo, cioè,

$$P[\theta_1 \leq \Theta \leq \theta_2] = \frac{\theta_2 - \theta_1}{360}, \quad 0 \leq \theta_1 \leq \theta_2 \leq 360.$$

La ragione di questo è che le variabili casuali continue non hanno una massa di probabilità. Invece, una massa di probabilità viene assegnata alla realizzazione della variabile casuale in un intervallo di valori.

5.1.1 Il paradosso delle variabili casuali continue

Nel nostro esempio, la pendenza dello spinner è esattamente 36 gradi; ma avrebbe potuto anche essere 36.0376531 gradi o qualunque altro valore in quell'intorno. Qual è la probabilità che la pendenza dello spinner sia esattamente 36? Paradossalmente, la risposta è zero:

$$P[\Theta = 36] = 0.$$

Infatti, se la probabilità di un qualunque valore fosse maggiore di zero, ogni altro possibile valore dovrebbe avere la stessa probabilità, dato che abbiamo assunto che tutti i valori Θ sono egualmente probabili. Ma se poi andiamo a sommare tutte queste probabilità il totale diventerà maggiore di uno, il che non è possibile.

Nel caso delle variabili casuali continue dobbiamo dunque rinunciare a qualcosa, e quel qualcosa è l'idea che, in una distribuzione continua, ciascun valore puntuale della variabile casuale possa avere una massa di probabilità maggiore di zero. Il paradosso sorge perché una realizzazione della variabile casuale continua produce sempre un qualche numero, ma ciascuno di tali numeri ha probabilità nulla.

5.2 La funzione di ripartizione per una variabile casuale continua

Supponiamo che $\Theta \sim \text{uniform}(0, 360)$ sia la pendenza dello spinner. La funzione di ripartizione (ovvero, la distribuzione cumulativa) è definita esattamente come nel caso delle variabili casuali discrete:

$$F_{\Theta}(\theta) = P[\Theta \leq \theta].$$

Cioè, è la probabilità che la variabile casuale Θ assuma un valore minore di o uguale a θ . In questo caso, poiché si presume che lo spinner sia simmetrico, la funzione di distribuzione cumulativa è

$$F_{\Theta}(\theta) = \frac{\theta}{360}.$$

Questa è una funzione lineare di θ , cioè $\frac{1}{360} \times \theta$, come indicato dal grafico della figura 5.2.

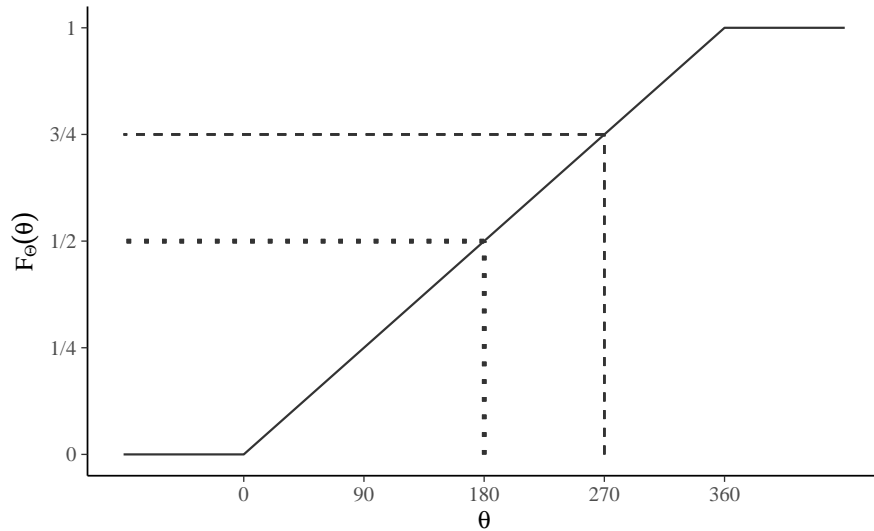


Figura 5.2: Funzione di distribuzione cumulativa per l'angolo θ (in gradi) risultante da una rotazione di uno spinner simmetrico. La linea tratteggiata mostra il valore a 180 gradi, che corrisponde ad una probabilità di 0.5, e la linea tratteggiata a 270 gradi, che corrisponde ad una probabilità di 0.75.

Possiamo verificare questo risultato mediante simulazione. Per stimare la funzione di ripartizione, simuliamo M valori $\theta^{(m)}$ e poi li ordiniamo in ordine crescente.

```
M <- 1000
theta <- runif(M, 0, 360)
theta_asc <- sort(theta)
prob <- (1:M) / M
unif_cdf_df <- data.frame(
  theta = theta_asc,
  prob = prob
)
unif_cdf_plot <-
  unif_cdf_df %>%
  ggplot(aes(x = theta, y = prob)) +
  geom_line() +
```

```
scale_x_continuous(breaks = c(0, 90, 180, 270, 360)) +  
scale_y_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1.0)) +  
xlab(expression(theta)) +  
ylab(expression(F[Theta](theta)))  
unif_cdf_plot
```

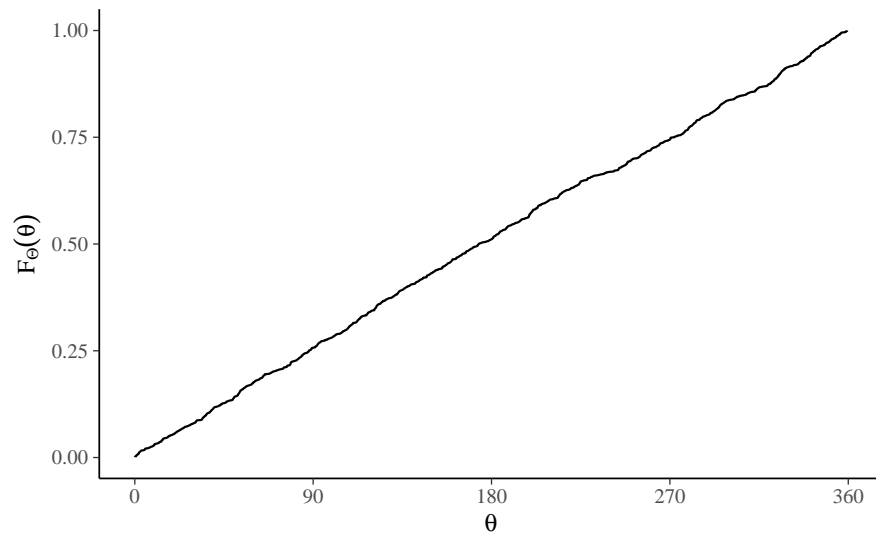


Figura 5.3: Grafico della funzione di ripartizione di una variabile casuale Θ che rappresenta il risultato di una rotazione di uno spinner simmetrico. Come previsto, tale funzione è una semplice funzione lineare perché la variabile sottostante Θ ha una distribuzione uniforme.

Anche con $M = 1000$, tale grafico è praticamente indistinguibile da quello prodotto per via analitica.

Come nel caso delle variabili casuali discrete, la funzione di ripartizione può essere utilizzata per calcolare le probabilità che la variabile casuale assuma valori in un intervallo. Ad esempio

$$\begin{aligned}
 P[180 < \Theta \leq 270] &= P[\Theta \leq 270] - P[\Theta \leq 180] \\
 &= F_{\Theta}(270) - F_{\Theta}(180) \\
 &= \frac{3}{4} - \frac{1}{2} \\
 &= \frac{1}{4}.
 \end{aligned} \tag{5.1}$$

5.3 La distribuzione uniforme

Dopo avere visto come generare numeri casuali uniformi da 0 a 360, consideriamo ora una variabile casuale che assume valori nell'intervallo da 0 a 1. Chiamiamo tale variabile casuale Θ e assumiamo che abbia una distribuzione continua uniforme sull'intervallo $[0, 1]$:

$$\Theta \sim \text{Uniform}(0, 1).$$

Poiché le probabilità assumono valori nell'intervallo $[0, 1]$, possiamo pensare a Θ come ad un valore di probabilità preso a caso in ciascuna realizzazione dell'esperimento casuale.

La distribuzione uniforme è la più semplice delle distribuzioni di densità di probabilità. Per chiarire le proprietà di tale distribuzione, iniziamo con una simulazione e generiamo 10,000 valori casuali di Θ . I primi 10 di tali valori sono stampati qui di seguito:

```

set.seed(1234)
M <- 10000
logit <- function(x) log(x / (1 - x))
theta <- runif(M)
alpha <- logit(theta)
for (m in 1:10) {
  print(alpha[m])
}
#> [1] -2.053
#> [1] 0.4993
#> [1] 0.4443
#> [1] 0.5039

```



```
#> [1] 1.823
#> [1] 0.5767
#> [1] -4.647
#> [1] -1.194
#> [1] 0.6905
#> [1] 0.05702
```

Creiamo ora un istogramma che descrive la distribuzione delle 10,000 realizzazioni Θ che abbiamo trovato:

```
df_prob_unif <- data.frame(theta = theta)
unif_prob_plot <-
  ggplot(df_prob_unif, aes(theta)) +
  geom_histogram(
    binwidth = 1 / 34, center = 1 / 68, color = "black",
    size = 0.25
  ) +
  scale_x_continuous(breaks = c(0, 0.25, 0.5, 0.75, 1)) +
  scale_y_continuous(lim = c(0, 1300), breaks = c(500, 1000)) +
  xlab(expression(paste(Theta, " ~ Uniform(0, 1)")))
unif_prob_plot
```

È chiaro che, all'aumentare del numero delle realizzazioni Θ , il profilo dell'istogramma tenderà a diventare una linea retta. Ciò significa che la funzione di densità di una variabile casuale uniforme continua è una costante. Cioè, se $\Theta \sim \text{Uniform}(a, b)$, allora $p_{\Theta}(\theta) = c$, dove c è una costante.

```
uniform_pdf_df <- data.frame(y = c(0, 1), p_y = c(1, 1))
uniform_pdf_plot <-
  ggplot(uniform_pdf_df, aes(x = y, y = p_y)) +
  geom_line(size = 0.5, color = "#333333") +
  geom_point(size = 1.5, color = "#333333") +
  scale_x_continuous(breaks = c(0, 1), labels = c("a", "b")) +
  scale_y_continuous(
    lim = c(0, 1), breaks = c(0, 1),
    labels = c("0", "c")
  ) +
```

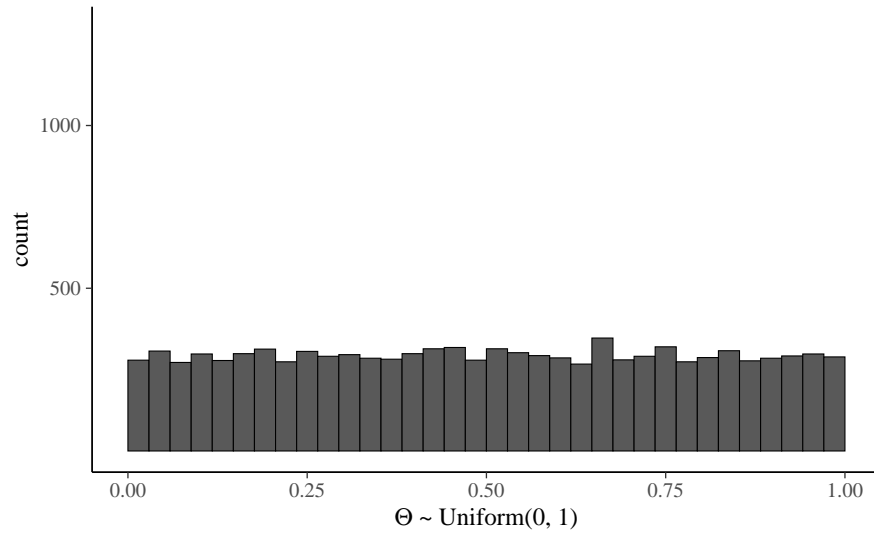


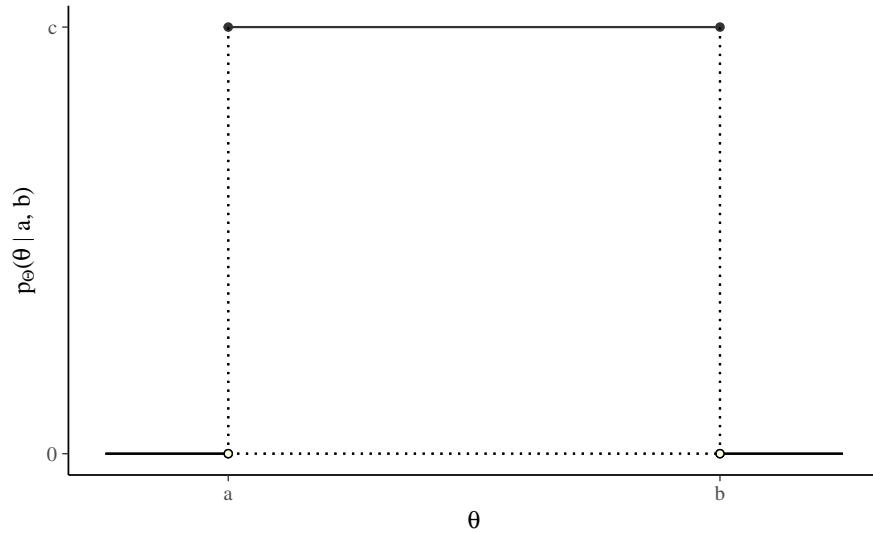
Figura 5.4: Istogramma di 10 000 realizzazioni $\Theta \sim \text{Uniform}(0, 1)$.

```

xlab(expression(theta)) +
ylab(expression(paste(p[Theta], "(", theta, " | a, b)"))) +
geom_segment(aes(x = 0, y = 0, xend = 0, yend = 1),
  linetype = "dotted"
) +
geom_segment(aes(x = 1, y = 0, xend = 1, yend = 1),
  linetype = "dotted"
) +
geom_segment(aes(x = 0, y = 0, xend = 1, yend = 0),
  linetype = "dotted"
) +
geom_segment(aes(x = -0.25, y = 0, xend = 0, yend = 0)) +
geom_segment(aes(x = 1, y = 0, xend = 1.25, yend = 0)) +
geom_point(aes(x = 0, y = 0),
  size = 1.5, shape = 21,
  fill = "#ffffe6"
) +
geom_point(aes(x = 1, y = 0),
  size = 1.5, shape = 21,
  fill = "#ffffe6"

```

```
)
uniform_pdf_plot
```



Dal grafico vediamo che l'area sottesa alla funzione di densità è $(b-a) \times c$. Dato che tale area deve essere unitaria, ovvero, $(b-a) \times c = 1$, possiamo trovare c dividendo entrambi i termini per $b-a$,

$$c = \frac{1}{b-a}.$$

Ovvero, se $\Theta \sim \text{Uniform}(a, b)$, allora

$$p_{\Theta}(\theta) = \text{Uniform}(\theta \mid a, b),$$

laddove

$$\text{Uniform}(\theta \mid a, b) = \frac{1}{b-a}.$$

In conclusione, la densità di una variabile casuale uniforme continua non dipende da θ — è costante e identica per ogni possibile valore θ .²

²Per comodità, possiamo assumere che i valori impossibili di θ abbiano una densità uguale a zero.

Vedremo nel prossimo Paragrafo che, eseguendo una trasformazione su questa variabile casuale uniforme, possiamo creare altre variabili casuali di interesse.

Esercizio 5.1. Si consideri una variabile casuale uniforme X definita sull'intervallo $[0, 100]$. Si trovi la probabilità $P(20 < X < 60)$.

Per trovare la soluzione è sufficiente calcolare l'area di un rettangolo di base $60 - 20 = 40$ e di altezza $1/100$. La probabilità cercata è dunque $P(20 < X < 60) = 40 \cdot 0.01 = 0.4$.

5.4 Dagli istogrammi alle densità

Non esiste l'equivalente di una funzione di massa di probabilità per le variabili casuali continue. Esiste invece una *funzione di densità di probabilità* la quale, nei termini di una simulazione, può essere concepita nel modo seguente: avendo a disposizione un numero enorme di casi, quando l'intervallo Δ di ciascuna classe $\rightarrow 0$, la spezzata tende a diventare una curva continua. Tale curva continua $f(x)$ è detta funzione di densità di probabilità.

Come si trasformano gli istogrammi all'aumentare del numero di osservazioni? Nei grafici seguenti, la numerosità cresce da 10 a 1 000 000.

```
df_log_odds_growth <- data.frame()
for (log10M in 1:6) {
  M <- 10^log10M
  alpha <- logit(runif(M))
  df_log_odds_growth <- rbind(
    df_log_odds_growth,
    data.frame(
      alpha = alpha,
      M = rep(sprintf("M = %d", M), M)
    )
  )
}
log_odds_growth_plot <-
  df_log_odds_growth %>%
```

```

ggplot(aes(alpha)) +
  geom_histogram(color = "black", bins = 75) +
  facet_wrap(~M, scales = "free") +
  scale_x_continuous(
    lim = c(-8.5, 8.5), breaks = c(-5, 0, 5)
  ) +
  xlab(expression(paste(Phi, " = ", logit(Theta)))) +
  ylab("proportion of draws") +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank(),
    panel.spacing.x = unit(2, "lines"),
    panel.spacing.y = unit(2, "lines")
  )
log_odds_growth_plot

```

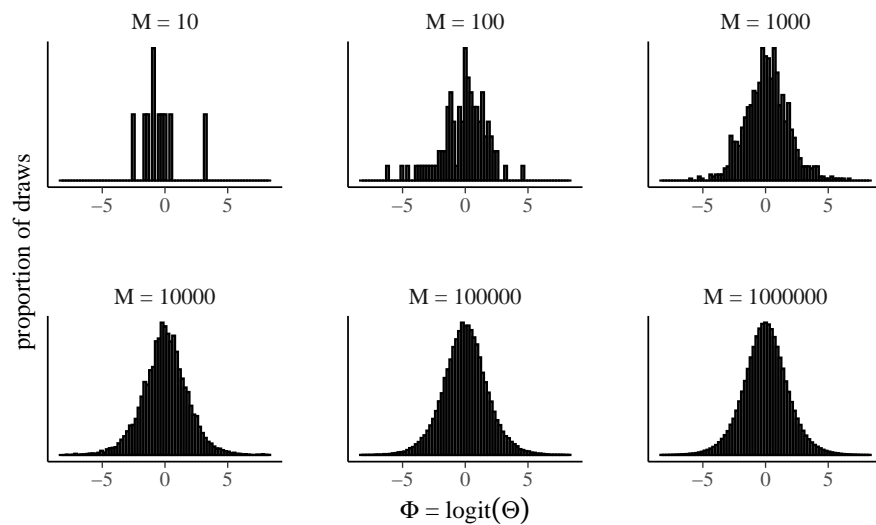


Figura 5.5: Istogramma di M campioni casuali $\Theta \sim \text{Uniform}(0, 1)$ trasformati in valori $\Phi = \text{logit}(\Theta)$. Il profilo limite dell'istogramma è evidenziato nella figura in basso a destra che è stata costruita usando 1 000 000 di osservazioni.

In un istogramma, l'area di ciascuna barra è proporzionale alla frequenza relativa delle osservazioni in quell'intervallo. Perché tutti gli intervalli hanno la stessa ampiezza, anche l'altezza di ciascuna barra sarà

proporzionale alla frequenza relativa delle osservazioni in quell'intervallo.

Nella simulazione, possiamo pensare all'area di ciascuna barra dell'istogramma come alla stima della probabilità che la variabile casuale assuma un valore compreso nell'intervallo considerato. All'aumentare del numero M di osservazioni, le probabilità stimate si avvicinano sempre di più ai veri valori della probabilità. All'aumentare del numero degli intervalli (quando l'ampiezza Δ dell'intervallo $\rightarrow 0$), il profilo dell'istogramma tende a diventare una curva continua. Tale curva continua è la funzione di densità di probabilità della variabile casuale. Per l'esempio presente, con $M = 1\,000\,000$, otteniamo il grafico riportato nella figura 5.6.

```
M <- 1e6
alpha <- logit(runif(M))
density_limit_df <- data.frame(alpha = alpha)
density_limit_plot <-
  density_limit_df %>%
  ggplot(aes(alpha)) +
  geom_histogram(
    stat = "density", n = 75, color = "black", size = 0.15
  ) +
  stat_function(
    fun = dlogis,
    args = list(location = 0, scale = 1),
    col = "black",
    size = 0.3
  ) +
  scale_x_continuous(
    lim = c(-9, 9),
    breaks = c(-6, -4, -2, 0, 2, 4, 6)
  ) +
  xlab(
    expression(paste(Phi, " = ", logit(Theta)))
  ) +
  ylab("Frequenza relativa") +
  theme(
    axis.text.y = element_blank(),
    axis.ticks.y = element_blank()
  )
```

density_limit_plot

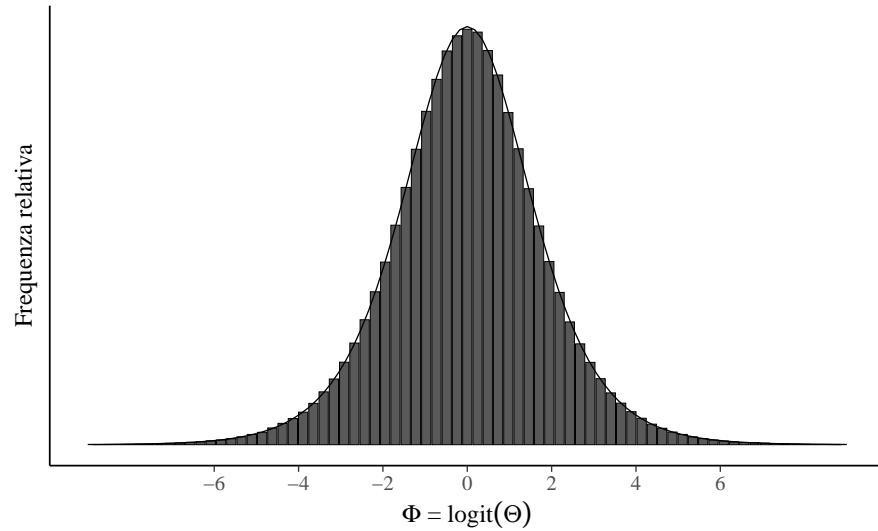


Figura 5.6: Istogramma di $M = 1\,000\,000$ campioni casuali $\Theta \sim \text{Uniform}(0,1)$ trasformati in valori $\Phi = \text{logit}(\Theta)$. La spezzata nera congiunge i punti centrali superiori delle barre dell'istogramma. Nel limite, quando il numero di osservazioni e di barre tende all'infinito, tale spezzata approssima la funzione di densità di probabilità della variabile casuale.

Nella statistica descrittiva abbiamo già incontrato una rappresentazione che ha lo stesso significato della funzione di densità, ovvero il kernel density plot. La stima della densità del kernel (KDE), infatti, è un modo non parametrico per stimare la funzione di densità di probabilità di una variabile casuale.

5.5 Funzione di densità di probabilità

Per descrivere le probabilità che possono essere associate ad una variabile casuale continua X è necessario definire una funzione $p(X)$ che deve soddisfare le seguenti due proprietà:

- $p(x) \geq 0, \forall x$, ovvero, l'ordinata della funzione di densità è 0 o positiva;

- $\int_{-\infty}^{\infty} p(x) \, dx = 1$, ovvero, l'area sottesa dalla $p(x)$ è unitaria³;
- $p(a < x < b) = \int_a^b p(x) \, dx$, se $a \leq b$, ovvero, l'area sottesa dalla $p(x)$ tra due punti a e b corrisponde alla probabilità che la v.c. x assuma un valore compreso tra questi due estremi.

Interpretazione. È possibile che $p(x) > 1$, quindi una densità di probabilità non può essere interpretata come una probabilità. Piuttosto, la densità $p(x)$ può essere utilizzata per confrontare la plausibilità relativa di diversi valori X . Considerata una variabile casuale X di cui è disponibile un insieme di realizzazioni, tanto maggiore è $p(x_k)$ rispetto a $p(x_l)$, tanto più grande sarà la nostra certezza che valori nell'intorno di x_k verranno osservati con maggiore frequenza di valori nell'intorno di x_l .

5.6 La funzione di ripartizione

La funzione di ripartizione $F(X)$ è quella funzione che associa a ogni valore di una variabile aleatoria X la probabilità che la variabile assuma valore minore o uguale a un prefissato valore x_k .

La funzione di ripartizione è sempre non negativa, monotona non decrescente tra 0 e 1, tale che:

$$\lim_{x \rightarrow -\infty} F_x(X) = F_X(-\infty) = 0, \quad \lim_{x \rightarrow +\infty} F_x(X) = F_X(+\infty) = 1.$$

Se X è una variabile aleatoria discreta, tale funzione può essere espressa nel seguente modo:

$$F(x_k) = P(X \leq x_k) = \sum_{i=1}^k P(X = x_i),$$

dove $P(X = x_i)$ rappresenta la probabilità che la variabile aleatoria X assuma la modalità x_i .

Se X è una variabile aleatoria continua allora la funzione di ripartizione è:

³Per quel che riguarda la notazione dell'integrale, ovvero $\int_x dx$, rimando alla discussione di S.P. Thompson: <https://calculusmadeeasy.org/1.html>

$$F(x_k) = P(X \leq x_k) = \int_{-\infty}^{x_k} f(x) \, dx.$$



6

Valore atteso e varianza

Spesso risulta utile fornire una rappresentazione sintetica della distribuzione di una variabile casuale attraverso degli indicatori caratteristici piuttosto che fare riferimento ad una sua rappresentazione completa mediante la funzione di ripartizione, o la funzione di massa o di densità di probabilità. Una descrizione più sintetica di una variabile casuale, tramite pochi valori, ci consente di cogliere le caratteristiche essenziali della distribuzione, quali: la posizione, cioè il baricentro della distribuzione di probabilità; la variabilità, cioè la dispersione della distribuzione di probabilità attorno ad un centro; la forma della distribuzione di probabilità, considerando la simmetria e la curtosi (pesantezza delle code). In questo Capitolo introdurremo quegli indici sintetici che descrivono il centro di una distribuzione di probabilità e la sua variabilità.

6.1 Valore atteso

Quando vogliamo conoscere il comportamento tipico di una variabile casuale spesso vogliamo sapere qual è il suo “valore tipico”. La nozione di “valore tipico”, tuttavia, è ambigua. Infatti, essa può essere definita in almeno tre modi diversi:

- la *media* (somma dei valori divisa per il numero dei valori),
- la *mediana* (il valore centrale della distribuzione, quando la variabile è ordinata in senso crescente o decrescente),
- la *moda* (il valore che ricorre più spesso).

Per esempio, la media di $\{3, 1, 4, 1, 5\}$ è $\frac{3+1+4+1+5}{5} = 2.8$, la mediana è 3 e la moda è 1. Tuttavia, la teoria delle probabilità si occupa di variabili casuali piuttosto che di sequenze di numeri. Diventa dunque necessario precisare che cosa intendiamo per “valore tipico” quando

facciamo riferimento alle variabili casuali. Giungiamo così alla seguente definizione.

Definizione 6.1. Sia Y è una variabile casuale discreta che assume i valori y_1, \dots, y_n con distribuzione $p(y)$, ossia

$$P(Y = y_i) = p(y_i),$$

per definizione il *valore atteso* di Y , $\mathbb{E}(Y)$, è

$$\mathbb{E}(Y) = \sum_{i=1}^n y_i \cdot p(y_i). \quad (6.1)$$

A parole: il valore atteso (o speranza matematica, o aspettazione, o valor medio) di una variabile casuale è definito come la somma di tutti i valori che la variabile casuale può prendere, ciascuno pesato dalla probabilità con cui il valore è preso.

Esercizio 6.1. Calcoliamo il valore atteso della variabile casuale Y corrispondente al lancio di una moneta equilibrata (testa: $Y = 1$; croce: $Y = 0$).

$$\mathbb{E}(Y) = \sum_{i=1}^2 y_i \cdot P(y_i) = 0 \cdot \frac{1}{5} + 1 \cdot \frac{1}{5} = 0.5.$$

Esercizio 6.2. Supponiamo ora che Y sia il risultato del lancio di un dado equilibrato. Il valore atteso di Y diventa:

$$\mathbb{E}(Y) = \sum_{i=1}^6 y_i \cdot P(y_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$

6.1.1 Interpretazione

Che interpretazione può essere assegnata alla nozione di valore atteso? Bruno de Finetti adottò lo stesso termine di *previsione* (e lo stesso simbolo) tanto per la probabilità che per la speranza matematica. Si può pertanto dire che, dal punto di vista bayesiano, la speranza matematica è l'estensione naturale della nozione di probabilità soggettiva.

6.1.2 Proprietà del valore atteso

La proprietà più importante del valore atteso è la linearità: il valore atteso di una somma di variabili casuali è uguale alla somma dei loro rispettivi valori attesi:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y). \quad (6.2)$$

La (6.2) sembra ragionevole quando X e Y sono indipendenti, ma è anche vera quando X e Y sono associati. Abbiamo anche che

$$\mathbb{E}(cY) = c\mathbb{E}(Y). \quad (6.3)$$

La (6.3) ci dice che possiamo estrarre una costante dall'operatore di valore atteso. Tale proprietà si estende a qualunque numero di variabili casuali. Infine, se due variabili casuali X e Y sono indipendenti, abbiamo che

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y). \quad (6.4)$$

Esercizio 6.3. Si considerino le seguenti variabili casuali: X , ovvero il numero che si ottiene dal lancio di un dado equilibrato, e Y , il numero di teste prodotto dal lancio di una moneta equilibrata. Poniamoci il problema di trovare il valore atteso di $X + Y$.

Per risolvere il problema iniziamo a costruire lo spazio campionario dell'esperimento casuale consistente nel lancio di un dado e di una moneta.

x/y	1	2	3	4	5	6
0	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
1	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)

ovvero

x/y	1	2	3	4	5	6
0	1	2	3	4	5	6
1	2	3	4	5	6	7

Il risultato del lancio del dado è indipendente dal risultato del lancio della moneta. Pertanto, ciascun evento elementare dello spazio campionario avrà la stessa probabilità di verificarsi, ovvero $Pr(\omega) = \frac{1}{12}$. Il valore atteso di $X + Y$ è dunque uguale a:

$$\mathbb{E}(X + Y) = 1 \cdot \frac{1}{12} + 2 \cdot \frac{1}{12} + \dots + 7 \cdot \frac{1}{12} = 4.0.$$

Lo stesso risultato si ottiene nel modo seguente:

$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y) = 3.5 + 0.5 = 4.0.$$

Esercizio 6.4. Si considerino le variabili casuali X e Y definite nel caso del lancio di tre monete equilibrate, dove X conta il numero delle teste nei tre lanci e Y conta il numero delle teste al primo lancio. Si calcoli il valore atteso del prodotto delle variabili casuali X e Y .

La distribuzione di probabilità congiunta $P(X, Y)$ è fornita nella tabella seguente.

x/y	0	1	$p(Y)$
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
$p(x)$	4/8	4/8	1.0

Il calcolo del valore atteso di XY si riduce a

$$\mathbb{E}(XY) = 1 \cdot \frac{1}{8} + 2 \cdot \frac{2}{8} + 3 \cdot \frac{1}{8} = 1.0.$$

Si noti che le variabili casuali X e Y non sono indipendenti. Dunque non possiamo usare la proprietà ???. Infatti, il valore atteso di X è

$$\mathbb{E}(X) = 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5$$

e il valore atteso di Y è

$$\mathbb{E}(Y) = 0 \cdot \frac{4}{8} + 1 \cdot \frac{4}{8} = 0.5.$$

Dunque

$$1.5 \cdot 0.5 \neq 1.0.$$

6.1.3 Variabili casuali continue

Nel caso di una variabile casuale continua Y il valore atteso diventa:

$$\mathbb{E}(Y) = \int_{-\infty}^{+\infty} yp(y) \, dy \quad (6.5)$$

Anche in questo caso il valore atteso è una media ponderata della y , nella quale ciascun possibile valore y è ponderato per il corrispondente valore della densità $p(y)$. Possiamo leggere l'integrale pensando che y rappresenti l'ampiezza delle barre infinitamente strette di un istogramma, con la densità $p(y)$ che corrisponde all'altezza di tali barre e la notazione $\int_{-\infty}^{+\infty}$ che corrisponde ad una somma.

Un'altra misura di tendenza centrale delle variabili casuali continue è la moda. La moda della Y individua il valore y più plausibile, ovvero il valore y che massimizza la funzione di densità $p(y)$:

$$\text{Mo}(Y) = \arg \max_y p(y). \quad (6.6)$$

6.2 Varianza

La seconda più importante proprietà di una variabile casuale, dopo che conosciamo il suo valore atteso, è la *varianza*.

Definizione 6.2. Se Y è una variabile casuale discreta con distribuzione $p(y)$, per definizione la varianza di Y , $\mathbb{V}(Y)$, è

$$\mathbb{V}(Y) = \mathbb{E}[(Y - \mathbb{E}(Y))^2]. \quad (6.7)$$

A parole: la varianza è la deviazione media quadratica della variabile dalla sua media.¹ Se denotiamo $\mathbb{E}(Y) = \mu$, la varianza $\mathbb{V}(Y)$ diventa il valore atteso di $(Y - \mu)^2$.

¹Data una variabile casuale Y con valore atteso $\mathbb{E}(Y)$, le “distanze” tra i valori di

Esercizio 6.5. Posta S uguale alla somma dei punti ottenuti nel lancio di due dadi equilibrati, poniamoci il problema di calcolare la varianza di S .

La variabile casuale S ha la seguente distribuzione di probabilità:

s	2	3	4	5	6	7	8	9	10	11	12
$P(S = s)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Essendo $\mathbb{E}(S) = 7$, la varianza diventa

$$\begin{aligned}\mathbb{V}(S) &= \sum (S - \mathbb{E}(S))^2 \cdot P(S) \\ &= (2 - 7)^2 \cdot 0.0278 + (3 - 7)^2 \cdot 0.0556 + \dots + (12 - 7)^2 \cdot 0.0278 \\ &= 5.8333.\end{aligned}$$

6.2.1 Formula alternativa per la varianza

C'è un modo più semplice per calcolare la varianza:

$$\begin{aligned}\mathbb{E}[(X - \mathbb{E}(Y))^2] &= \mathbb{E}(X^2 - 2X\mathbb{E}(Y) + \mathbb{E}(Y)^2) \\ &= \mathbb{E}(Y^2) - 2\mathbb{E}(Y)\mathbb{E}(Y) + \mathbb{E}(Y)^2,\end{aligned}$$

dato che $\mathbb{E}(Y)$ è una costante; pertanto

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2. \quad (6.8)$$

A parole: la varianza è la media dei quadrati meno il quadrato della media.

Y e il valore atteso $\mathbb{E}(Y)$ definiscono la variabile casuale $Y - \mathbb{E}(Y)$ chiamata *scarto*, oppure *deviazione* oppure *variabile casuale centrata*. La variabile $Y - \mathbb{E}(Y)$ equivale ad una traslazione di sistema di riferimento che porta il valore atteso nell'origine degli assi. Si può dimostrare facilmente che il valore atteso della variabile scarto $Y - \mathbb{E}(Y)$ vale zero, dunque la media di tale variabile non può essere usata per quantificare la “dispersione” dei valori di Y relativamente al suo valore medio. Occorre rendere sempre positivi i valori di $Y - \mathbb{E}(Y)$ e tale risultato viene ottenuto considerando la variabile casuale $(Y - \mathbb{E}(Y))^2$.

Esercizio 6.6. Consideriamo la variabile casuale Y che corrisponde al numero di teste che si osservano nel lancio di una moneta truccata con probabilità di testa uguale a 0.8. Il valore atteso di Y è

$$\mathbb{E}(Y) = 0 \cdot 0.2 + 1 \cdot 0.8 = 0.8.$$

Usando la formula tradizionale della varianza otteniamo:

$$\mathbb{V}(Y) = (0 - 0.8)^2 \cdot 0.2 + (1 - 0.8)^2 \cdot 0.8 = 0.16.$$

Lo stesso risultato si trova con la formula alternativa della varianza. Il valore atteso di Y^2 è

$$\mathbb{E}(Y^2) = 0^2 \cdot 0.2 + 1^2 \cdot 0.8 = 0.8.$$

e la varianza diventa

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = 0.8 - 0.8^2 = 0.16.$$

6.2.2 Variabili casuali continue

Nel caso di una variabile casuale continua Y , la varianza diventa:

$$\mathbb{V}(Y) = \int_{-\infty}^{+\infty} [y - \mathbb{E}(Y)]^2 p(y) \, dy \quad (6.9)$$

Come nel caso discreto, la varianza di una v.c. continua y misura approssimativamente la distanza al quadrato tipica o prevista dei possibili valori y dalla loro media.

6.3 Deviazione standard

Quando lavoriamo con le varianze, i termini sono innalzati al quadrato e quindi i numeri possono diventare molto grandi (o molto piccoli). Per trasformare nuovamente i valori nell'unità di misura della scala originaria si prende la radice quadrata. Il valore risultante viene chiamato *deviazione standard* e solitamente è denotato dalla lettera greca σ .

Definizione 6.3. Si definisce scarto quadratico medio (o deviazione standard o scarto tipo) la radice quadrata della varianza:

$$\sigma_Y = \sqrt{\mathbb{V}(Y)}. \quad (6.10)$$

Interpretiamo la deviazione standard di una variabile casuale come nella statistica descrittiva: misura approssimativamente la distanza tipica o prevista dei possibili valori y dalla loro media.

Esercizio 6.7. Per i dadi equilibrati dell'esempio precedente, la deviazione standard della variabile casuale S è uguale a $\sqrt{5.833} = 2.415$.

6.4 Standardizzazione

Definizione 6.4. Data una variabile casuale Y , si dice variabile standardizzata di Y l'espressione

$$Z = \frac{Y - \mathbb{E}(Y)}{\sigma_Y}. \quad (6.11)$$

Solitamente, una variabile standardizzata viene denotata con la lettera Z .

6.5 Momenti di variabili casuali

Definizione 6.5. Si chiama *momento* di ordine q di una v.c. X , dotata di densità $p(x)$, la quantità

$$\mathbb{E}(X^q) = \int_{-\infty}^{+\infty} x^q p(x) dx. \quad (6.12)$$

Se X è una v.c. discreta, i suoi momenti valgono:

$$\mathbb{E}(X^q) = \sum_i x_i^q p(x_i). \quad (6.13)$$

I momenti sono importanti parametri indicatori di certe proprietà di X . I più noti sono senza dubbio quelli per $q = 1$ e $q = 2$. Il momento del primo ordine corrisponde al valore atteso di X . Spesso i momenti di ordine superiore al primo vengono calcolati rispetto al valor medio di X , operando una traslazione $x_0 = x - \mathbb{E}(X)$ che individua lo scarto dalla media. Ne deriva che il momento centrale di ordine 2 corrisponde alla varianza.



Parte II

Distribuzioni teoriche di probabilità



7

Distribuzioni di v.c. discrete

In questo Capitolo verranno esaminate le principali distribuzioni di probabilità delle variabili casuali discrete. Un esperimento casuale che può dare luogo a solo due possibili esiti (successo, insuccesso) è modellabile con una variabile casuale di Bernoulli. Una sequenza di prove di Bernoulli costituisce un processo Bernoulliano. Il numero di successi dopo n prove di Bernoulli corrisponde ad una variabile casuale che segue la legge binomiale. La distribuzione binomiale risulta da un insieme di prove di Bernoulli solo se il numero totale n è fisso per disegno. Se il numero di prove è esso stesso una variabile casuale, allora il numero di successi nella corrispondente sequenza di prove bernoulliane segue la distribuzione di Poisson. Concluderemo con la distribuzione discreta uniforme.

7.1 Una prova Bernoulliana

Se un esperimento casuale ha solo due esiti possibili, allora le repliche indipendenti di questo esperimento sono chiamate “prove Bernoulliane” (il lancio di una moneta è il tipico esempio).

Definizione 7.1. Viene detta variabile di Bernoulli una variabile casuale discreta $Y = \{0, 1\}$ con la seguente distribuzione di probabilità:

$$P(Y | \theta) = \begin{cases} \theta & \text{se } Y = 1, \\ 1 - \theta & \text{se } Y = 0, \end{cases}$$

con $0 \leq \theta \leq 1$. Convenzionalmente l'evento $\{Y = 1\}$ con probabilità θ viene chiamato “successo” mentre l'evento $\{Y = 0\}$ con probabilità $1 - \theta$ viene chiamato “insuccesso”.

Applicando l'operatore di valore atteso e di varianza, otteniamo

$$\mathbb{E}(Y) = 0 \cdot P(Y = 0) + 1 \cdot P(Y = 1) = \theta, \quad (7.1)$$

$$\mathbb{V}(Y) = (0 - \theta)^2 \cdot P(Y = 0) + (1 - \theta)^2 \cdot P(Y = 1) = \theta(1 - \theta). \quad (7.2)$$

Scriviamo $Y \sim \text{Bernoulli}(\theta)$ per indicare che la variabile casuale Y ha una distribuzione Bernoulliana di parametro θ .

Esercizio 7.1. Nel caso del lancio di una moneta equilibrata la variabile casuale di Bernoulli assume i valori 0 e 1. La distribuzione di massa di probabilità è pari a $\frac{1}{2}$ in corrispondenza di entrambi iv valori. La funzione di distribuzione vale $\frac{1}{2}$ per $Y = 0$ e 1 per $Y = 1$.

7.2 Una sequenza di prove Bernoulliane

La distribuzione binomiale è rappresentata dall'elenco di tutti i possibili numeri di successi $Y = \{0, 1, 2, \dots, n\}$ che possono essere osservati in n prove Bernoulliane indipendenti di probabilità θ , a ciascuno dei quali è associata la relativa probabilità. Esempi di una distribuzione binomiale sono i risultati di una serie di lanci di una stessa moneta o di una serie di estrazioni da un'urna (con reintroduzione). La distribuzione binomiale di parametri n e θ è in realtà una famiglia di distribuzioni: al variare dei parametri θ e n variano le probabilità.

Definizione 7.2. La probabilità di ottenere y successi e $n - y$ insuccessi in n prove Bernoulliane è data dalla distribuzione binomiale:

$$\begin{aligned} P(Y = y) &= \binom{n}{y} \theta^y (1 - \theta)^{n-y} \\ &= \frac{n!}{y!(n-y)!} \theta^y (1 - \theta)^{n-y}, \end{aligned} \quad (7.3)$$

dove n = numero di prove Bernoulliane, θ = probabilità di successo in ciascuna prova e y = numero di successi.

Dimostrazione. La (7.3) può essere derivata nel modo seguente. Indichiamo con S il successo e con I l'insuccesso di ciascuna prova. Una sequenza di n prove Bernoulliane darà come esito una sequenza di n elementi S e I . Ad esempio, una sequenza che contiene y successi è la seguente:

$$\overbrace{SS \dots S}^{y \text{ volte}} \overbrace{II \dots I}^{n-y \text{ volte}}$$

Essendo θ la probabilità di S e $1 - \theta$ la probabilità di I , la probabilità di ottenere la specifica sequenza riportata sopra è

$$\overbrace{\theta \theta \dots \theta}^{y \text{ volte}} \overbrace{(1 - \theta)(1 - \theta) \dots (1 - \theta)}^{n-y \text{ volte}} = \theta^y \cdot (1 - \theta)^{n-y}. \quad (7.4)$$

Non siamo però interessati alla probabilità di una *specifica* sequenza di S e I ma, bensì, alla probabilità di osservare una *qualsiasi* sequenza di y successi in n prove. In altre parole, vogliamo la probabilità dell'unione di tutti gli eventi corrispondenti a y successi in n prove.

È immediato notare che una qualsiasi altra sequenza contenente esattamente y successi avrà sempre come probabilità $\theta^y \cdot (1 - \theta)^{n-y}$: il prodotto infatti resta costante anche se cambia l'ordine dei fattori.¹ Per trovare il risultato cercato dobbiamo moltiplicare la (7.4) per il numero di sequenze possibili di y successi in n prove.

Il numero di sequenze che contengono esattamente y successi in n prove. La risposta è fornita dal coefficiente binomiale²:

$$\binom{n}{y} = \frac{n!}{y!(n-y)!}, \quad (7.5)$$

dove il simbolo $n!$ si legge n fattoriale ed è uguale al prodotto di n numeri interi decrescenti a partire da n . Per definizione $0! = 1$.

Essendo la probabilità dell'unione di K elementi incompatibili uguale alla somma delle loro rispettive probabilità, e dato che le sequenze di y successi in n prove hanno tutte la stessa probabilità, per trovare la formula della distribuzione binomiale (7.3) è sufficiente moltiplicare la (7.4) per la (7.5). \square

La distribuzione di probabilità di alcune distribuzioni binomiali, per due valori di n e θ , è fornita nella figura 7.1.

¹Viene detta *scambiabilità* la proprietà per cui l'ordine con cui compiamo le osservazioni è irrilevante per l'assegnazione delle probabilità.

²La derivazione della formula del coefficiente binomiale è fornita nell'Appendice ??.

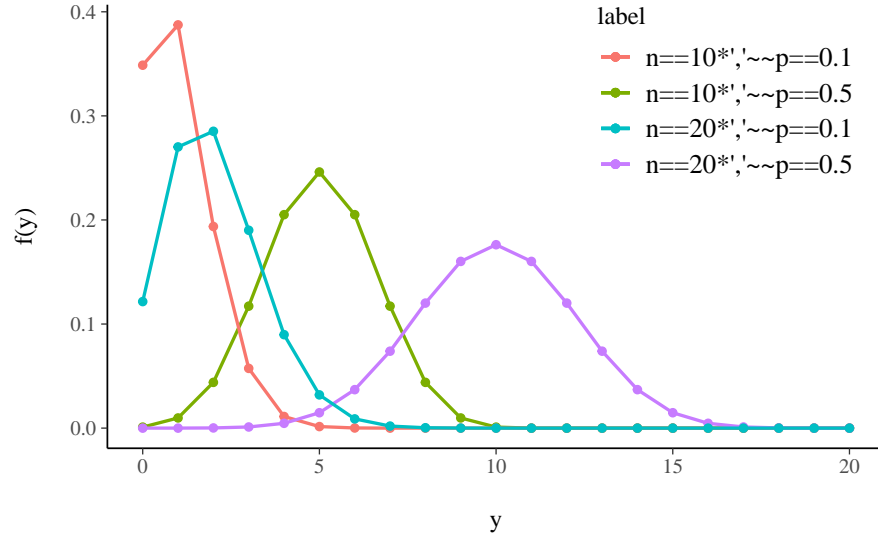


Figura 7.1: Alcune distribuzioni binomiali. Nella figura, il parametro θ è indicato con p .

Esercizio 7.2. Usando la (7.3), si trovi la probabilità di $y = 2$ successi in $n = 4$ prove Bernoulliane indipendenti con $\theta = 0.2$

$$\begin{aligned}
 P(Y = 2) &= \frac{4!}{2!(4-2)!} 0.2^2 (1-0.2)^{4-2} \\
 &= \frac{4 \cdot 3 \cdot 2 \cdot 1}{(2 \cdot 1)(2 \cdot 1)} 0.2^2 0.8^2 = 0.1536.
 \end{aligned}$$

Ripetendo i calcoli per i valori $y = 0, \dots, 4$ troviamo la distribuzione binomiale di parametri $n = 4$ e $\theta = 0.2$:

y	P(Y = y)
0	0.4096
1	0.4096
2	0.1536
3	0.0256
4	0.0016
sum	1.0

Lo stesso risultato si ottiene usando la seguente istruzione R:

```
dbinom(0:4, 4, 0.2)
#> [1] 0.4096 0.4096 0.1536 0.0256 0.0016
```

Esercizio 7.3. Lanciando 5 volte una moneta onesta, qual è la probabilità che esca testa almeno tre volte?

In R, la soluzione si trova con

```
dbinom(3, 5, 0.5) + dbinom(4, 5, 0.5) + dbinom(5, 5, 0.5)
#> [1] 0.5
```

Alternativamente, possiamo trovare la probabilità dell'evento complementare a quello definito dalla funzione di ripartizione calcolata mediante `pbinom()`, ovvero

```
1 - pbinom(2, 5, 0.5)
#> [1] 0.5
```

7.2.1 Valore atteso e deviazione standard

La media (numero atteso di successi in n prove) e la deviazione standard di una distribuzione binomiale sono molto semplici:

$$\begin{aligned}\mu &= n\theta, \\ \sigma &= \sqrt{n\theta(1-\theta)}.\end{aligned}\tag{7.6}$$

Dimostrazione. Essendo Y la somma di n prove Bernoulliane indipendenti Y_i , è facile vedere che

$$\mathbb{E}(Y) = \mathbb{E}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \mathbb{E}(Y_i) = n\theta,\tag{7.7}$$

$$\mathbb{V}(Y) = \mathbb{V}\left(\sum_{i=1}^n Y_i\right) = \sum_{i=1}^n \mathbb{V}(Y_i) = n\theta(1-\theta).\tag{7.8}$$

□

Esercizio 7.4. Si trovino il valore atteso e la varianza del lancio di quattro monete con probabilità di successo pari a $\theta = 0.2$.

Il valore atteso è $\mu = n\theta = 4 \cdot 0.2 = 0.8$. Ciò significa che, se l'esperimento casuale venisse ripetuto infinite volte, l'esito testa verrebbe osservato un numero medio di volte pari a 0.8. La varianza è $n\theta(1-\theta) = 4 \cdot (1-0.2) = 0.8$.³

7.3 Distribuzione di Poisson

La distribuzione di Poisson è una distribuzione di probabilità discreta che esprime le probabilità per il numero di eventi che si verificano successivamente ed indipendentemente in un dato intervallo di tempo, sapendo che mediamente se ne verifica un numero λ . La distribuzione di Poisson serve dunque per contare il numero di volte in cui un evento ha luogo in un determinato intervallo di tempo. La stessa distribuzione può essere estesa anche per contare gli eventi che hanno luogo in una determinata porzione di spazio.

Definizione 7.3. La distribuzione di Poisson può essere intesa come limite della distribuzione binomiale, dove la probabilità di successo θ è pari a $\frac{\lambda}{n}$ con n che tende a ∞ :

$$\lim_{y \rightarrow \infty} \binom{n}{y} \theta^y (1-\theta)^{n-y} = \frac{\lambda^y}{y!} e^{-\lambda}. \quad (7.9)$$

Alcune distribuzioni di Poisson sono riportate nella figura 7.2.

Esercizio 7.5. Supponiamo che un evento accada 300 volte all'ora e si vuole determinare la probabilità che in un minuto accadano esattamente 3 eventi.

Il numero medio di eventi in un minuto è pari a

```
lambda <- 300 / 60
lambda
#> [1] 5
```

³L'eguaglianza di μ e σ è solo una peculiarità di questo esempio.

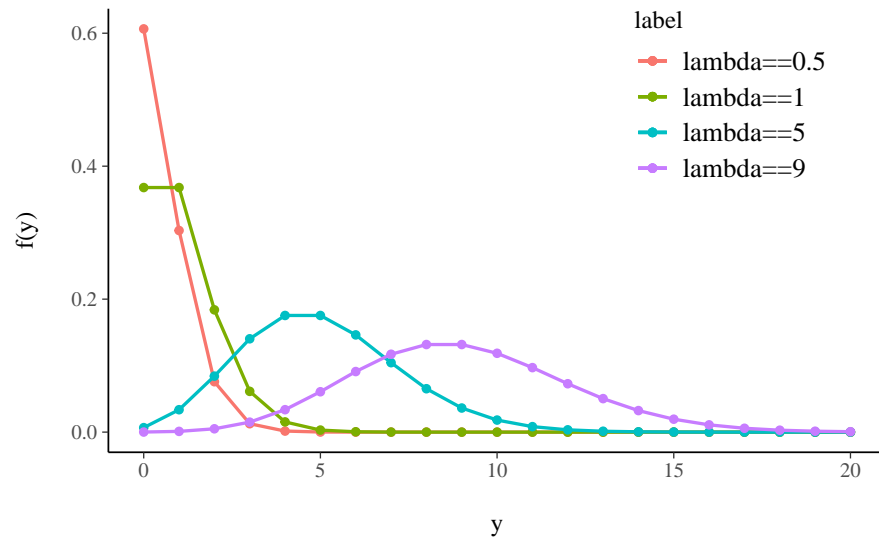


Figura 7.2: Alcune distribuzioni di Poisson.

Quindi la probabilità che in un minuto si abbiano 3 eventi è pari a

```
y <- 3
(lambda^y / factorial(y)) * exp(-lambda)
#> [1] 0.1404
```

Esercizio 7.6. Per i dati dell'esempio precedente, si trovi la probabilità che un evento accada almeno 8 volte in un minuto.

La probabilità cercata è

$$p(y \geq 8) = 1 - p(y \leq 7) = 1 - \sum_{i=0}^7 \frac{\lambda^i}{i!} e^{-\lambda},$$

con $\lambda = 5$.

Svolgendo i calcoli in Rotteniamo:

```
1 - ppois(q = 7, lambda = 5)
#> [1] 0.1334
```

```
ppois(q = 7, lambda = 5, lower.tail = FALSE)
#> [1] 0.1334
```

Esercizio 7.7. Sapendo che un evento avviene in media 6 volte al minuto, si calcoli (a) la probabilità di osservare un numero di eventi uguale o inferiore a 3 in un minuto, e (b) la probabilità di osservare esattamente 2 eventi in 30 secondi.

(a) In questo caso $\lambda = 6$ e la probabilità richiesta è

```
ppois(q = 3, lambda = 6, lower.tail = TRUE)
#> [1] 0.1512
```

(b) In questo caso $\lambda = 6/2$ e la probabilità richiesta è

```
dpois(x = 2, lambda = 3)
#> [1] 0.224
```

7.3.1 Alcune proprietà della variabile di Poisson

- Il valore atteso, la moda e la varianza della variabile di Poisson sono uguali a λ .
- La somma $Y_1 + \dots + Y_n$ di n variabili casuali indipendenti con distribuzioni di Poisson di parametri $\lambda_1, \dots, \lambda_n$ segue una distribuzione di Poisson di parametro $\lambda = \lambda_1 + \dots + \lambda_n$.
- La differenza di due variabili di Poisson non è una variabile di Poisson. Basti infatti pensare che può assumere valori negativi.

7.4 Distribuzione discreta uniforme

Una distribuzione discreta uniforme è una distribuzione di probabilità discreta che è uniforme su un insieme, ovvero che attribuisce ad

ogni elemento dell'insieme discreto e finito S su cui è definita la stessa probabilità p di verificarsi.

Consideriamo la variabile casuale X con supporto $1, 2, \dots, m$. Un esperimento casuale in cui si verifica questa distribuzione è la scelta casuale di un intero compreso tra 1 e m inclusi. Sia X il numero scelto. Allora

$$P(X = x) = \frac{1}{m}, \quad x = 1, \dots, m.$$

Il valore atteso è

$$\mathbb{E}(X) = \sum_{x=1}^m x f_X(x) = \sum_{x=1}^m x \frac{1}{m} = \frac{1}{m} (1 + 2 + \dots + m) = \frac{m+1}{2},$$

dove abbiamo utilizzato l'identità $1 + 2 + \dots + m = m(m+1)/2$.

Per trovare la varianza, prima calcoliamo

$$\mathbb{E}(X^2) = \frac{1}{m} \sum_{x=1}^m x^2,$$

e poi troviamo

$$\mathbb{V}(X) = E(X^2) - [E(X)]^2.$$

7.4.1 Usiamo R

La sintassi generale per simulare una variabile casuale uniforme discreta è `sample(x, size, replace = TRUE)`. L'argomento `x` identifica i numeri da cui campionare casualmente. Se `x` è un numero, il campionamento viene eseguito da 1 a `x`. L'argomento `size` indica quanto dovrebbe essere grande la dimensione del campione e `replace` indica se i numeri devono essere reintrodotti o meno nell'urna dopo essere stati estratti. L'opzione di default è `replace = FALSE` ma per le uniformi discrete i valori estratti devono essere sostituiti. Seguono alcuni esempi.

- Per lanciare un dado equilibrato 3000 volte: `sample(6, size = 3000, replace = TRUE);`
- per scegliere 27 numeri casuali da 30 a 70: `sample(30:70, size = 27, replace = TRUE);`
- per lanciare una moneta equa 1000 volte: `sample(c("H","T"), size = 1000, replace = TRUE).`

7.5 Distribuzione beta-binomiale

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

$$\text{BetaBinomial}(y \mid N, \alpha, \beta) = \binom{N}{y} \frac{B(y + \alpha, N - y + \beta)}{B(\alpha, \beta)}, \quad (7.10)$$

dove la funzione beta è $B(u, v) = \frac{\Gamma(u)\Gamma(v)}{\Gamma(u+v)}$.

Senza entrare nei dettagli, ci accontentiamo di sapere che tale distribuzione è implementata nella funzione `dbbinom()` del pacchetto `extraDistr`.

Commenti e considerazioni finali

La distribuzione binomiale è una distribuzione di probabilità discreta che descrive il numero di successi in un processo di Bernoulli, ovvero la variabile aleatoria $Y = Y_1 + \dots + Y_n$ che somma n variabili casuali indipendenti di uguale distribuzione di Bernoulli $\mathcal{B}(\theta)$, ognuna delle quali può fornire due soli risultati: il successo con probabilità θ e il fallimento con probabilità $1 - \theta$.

La distribuzione binomiale è molto importante per le sue molte applicazioni. Nelle presenti dispense, dedicate all'analisi bayesiana, è soprattutto importante perché costituisce il fondamento del caso più semplice del cosiddetto "aggiornamento bayesiano", ovvero il caso Beta-Binomiale. Il modello Beta-Binomiale ci fornirà infatti un esempio paradigmatico dell'approccio bayesiano all'inferenza e sarà trattato in maniera analitica. È dunque importante che le proprietà della distribuzione binomiale risultino ben chiare.

8

Distribuzioni di v.c. continue

Dopo avere introdotto con una simulazione il concetto di funzione di densità nel Capitolo 5, prendiamo ora in esame alcune delle densità di probabilità più note. La più importante di esse è sicuramente la distribuzione Normale.

8.1 Distribuzione Normale

Non c'è un'unica distribuzione Normale, ma ce ne sono molte. Tali distribuzioni sono anche dette “gaussiane” in onore di Carl Friedrich Gauss (uno dei più grandi matematici della storia il quale, tra le altre cose, scoprì l'utilità di tale funzione di densità per descrivere gli errori di misurazione). Adolphe Quetelet, il padre delle scienze sociali quantitative, fu il primo ad applicare tale densità alle misurazioni dell'uomo. Karl Pearson usò per primo il termine “distribuzione Normale” anche se ammise che questa espressione “ha lo svantaggio di indurre le persone a credere che le altre distribuzioni, in un senso o nell'altro, non siano normali.”

8.1.1 Limite delle distribuzioni binomiali

Iniziamo con un breve excursus storico. Nel 1733, Abraham de Moivre notò che, aumentando il numero di prove in una distribuzione binomiale, la distribuzione risultante diventava quasi simmetrica e a forma campanulare. Con 10 prove e una probabilità di successo di 0.9 in ciascuna prova, la distribuzione è chiaramente asimmetrica.

```
N <- 10
x <- 0:10
y <- dbinom(x, N, 0.9)
```

```
binomial_limit_plot <-  
  tibble(x = x, y = y) %>%  
  ggplot(aes(x = x, y = y)) +  
  geom_bar(  
    stat = "identity", color = "black", size = 0.2  
  ) +  
  xlab("y") +  
  scale_x_continuous(breaks = c(0, 2, 4, 6, 8, 10)) +  
  ylab("Binomial(y | 10, 0.9)")  
binomial_limit_plot
```

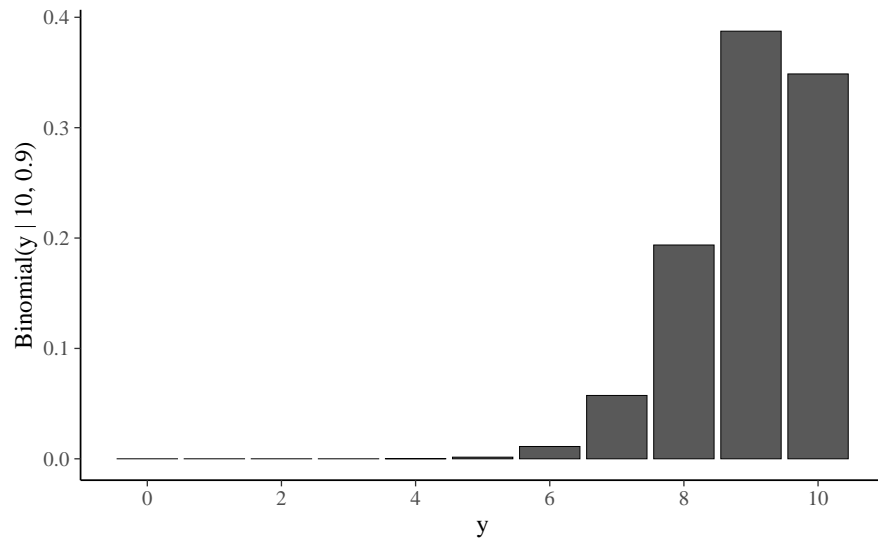


Figura 8.1: Probabilità del numero di successi in $N = 10$ prove bernoulliane indipendenti, ciascuna con una probabilità di successo di 0.90. Il risultato è una distribuzione $\text{Binomial}(y | 10, 0.9)$. Con solo dieci prove, la distribuzione è fortemente asimmetrica negativa.

Ma se aumentiamo il numero di prove di un fattore di 100 a $N = 1000$, senza modificare la probabilità di successo di 0.9, la distribuzione assume una forma campanulare quasi simmetrica. Dunque, de Moivre scoprì che, quando N è grande, la funzione Normale (che introdurremo qui sotto), nonostante sia la densità di v.a. continue, fornisce una buona approssimazione alla funzione di massa di probabilità binomiale.

```
N <- 1000
x <- 0:1000
y <- dbinom(x, N, 0.9)
binomial_limit_plot <-
  tibble(x = x, y = y) %>%
  ggplot(aes(x = x, y = y)) +
  geom_bar(stat = "identity", color = "black", size = 0.2) +
  xlab("y") +
  ylab("Binomial(y | 1000, 0.9)") +
  xlim(850, 950)
binomial_limit_plot
```

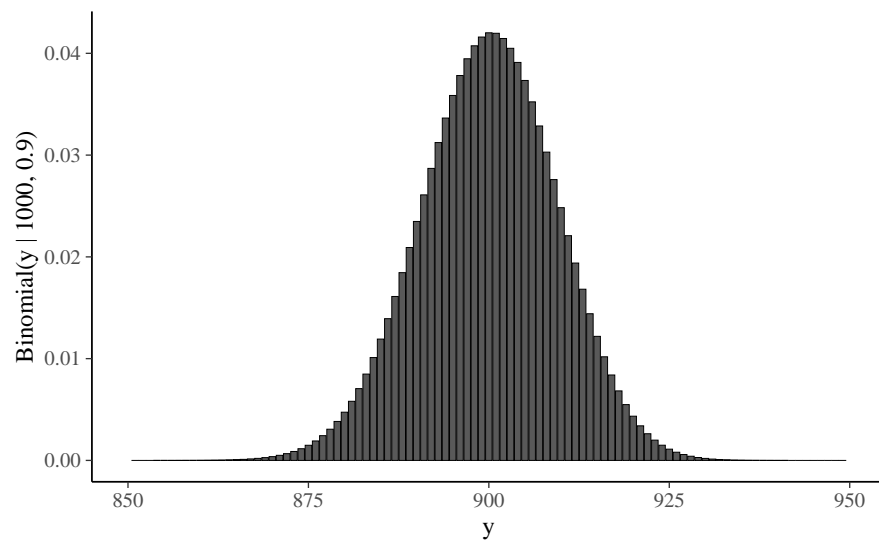


Figura 8.2: Probabilità del numero di successi in $N = 1000$ prove bernoulliane indipendenti, ciascuna con una probabilità di successo di 0.90. Il risultato è una distribuzione $\text{Binomial}(y | 1000, 0.9)$. Con mille prove, la distribuzione è quasi simmetrica a forma campanulare.

La distribuzione Normale fu scoperta da Gauss nel 1809 e, storicamente, è intimamente legata al metodo dei minimi quadrati – si veda l'Appendice ???. Il Paragrafo successivo illustra come si possa giungere alla Normale mediante una simulazione.

8.2 La Normale prodotta con una simulazione

[McElreath \(2020\)](#) presenta un esempio che illustra come sia possibile giungere alla distribuzione Normale mediante una simulazione. Supponiamo che vi siano mille persone tutte allineate su una linea di partenza. Quando viene dato un segnale, ciascuna persona lancia una moneta e fa un passo in avanti oppure all'indietro a seconda che sia uscita testa o croce. Supponiamo che la lunghezza di ciascun passo vari da 0 a 1 metro. Ciascuna persona lancia una moneta 16 volte e dunque compie 16 passi.

Alla conclusione di queste passeggiate casuali (*random walk*) non possiamo sapere con esattezza dove si troverà ciascuna persona, ma possiamo conoscere con certezza le caratteristiche della distribuzione delle mille distanze dall'origine. Per esempio, possiamo predire in maniera accurata la proporzione di persone che si sono spostate in avanti oppure all'indietro. Oppure, possiamo predire accuratamente la proporzione di persone che si troveranno ad una certa distanza dalla linea di partenza (es., a 1.5 m dall'origine).

Queste predizioni sono possibili perché tali distanze si distribuiscono secondo la legge Normale. È facile simulare questo processo usando R. I risultati della simulazione sono riportati nella figura [8.3](#).

```
pos <-
  replicate(100, runif(16, -1, 1)) %>%
  as_tibble() %>%
  rbind(0, .) %>%
  mutate(step = 0:16) %>%
  gather(key, value, -step) %>%
  mutate(person = rep(1:100, each = 17)) %>%
  group_by(person) %>%
  mutate(position = cumsum(value)) %>%
  ungroup()

ggplot(
  data = pos,
  aes(x = step, y = position, group = person)
) +
```

```
geom_vline(xintercept = c(4, 8, 16), linetype = 2) +
geom_line(aes(color = person < 2, alpha = person < 2)) +
scale_color_manual(values = c("gray", "black")) +
scale_alpha_manual(values = c(1 / 5, 1)) +
scale_x_continuous(
  "Numero di passi",
  breaks = c(0, 4, 8, 12, 16)
) +
labs(y = "Posizione") +
theme(legend.position = "none")
```

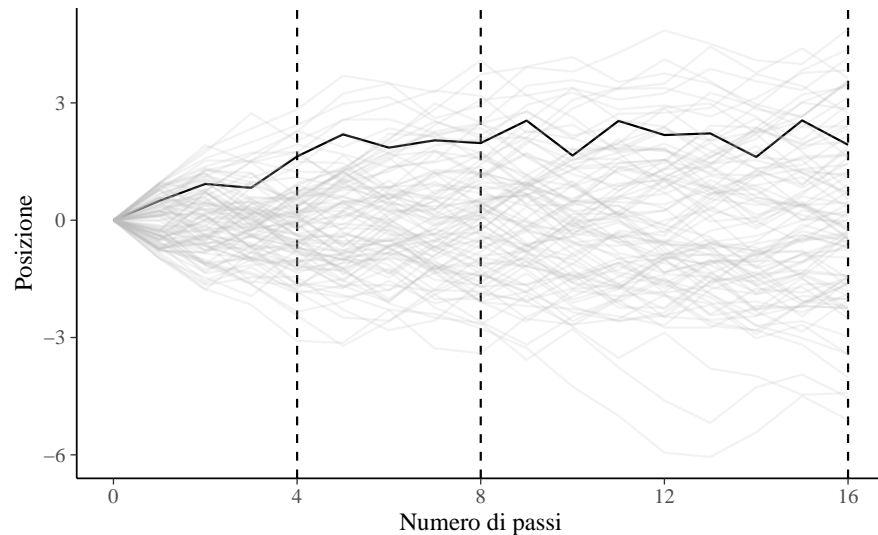


Figura 8.3: Passeggiata casuale di 4, 8 e 16 passi. La spezzata nera indica la media delle distanze dall'origine come funzione del numero di passi.

Un kernel density plot delle distanze ottenute dopo 4, 8 e 16 passi è riportato nella figura 8.4. Nel pannello di destra, al kernel density plot è stata sovrapposta una densità Normale di opportuni parametri (linea tratteggiata).

```
p1 <-
  pos %>%
  filter(step == 4) %>%
```

```

ggplot(aes(x = position)) +
  geom_line(stat = "density", color = "black") +
  labs(title = "4 passi")

p2 <-
  pos %>%
  filter(step == 8) %>%
  ggplot(aes(x = position)) +
  geom_density(color = "black", outline.type = "full") +
  labs(title = "8 passi")

sd <-
  pos %>%
  filter(step == 16) %>%
  summarise(sd = sd(position)) %>%
  pull(sd)

p3 <-
  pos %>%
  filter(step == 16) %>%
  ggplot(aes(x = position)) +
  stat_function(
    fun = dnorm,
    args = list(mean = 0, sd = sd),
    linetype = 2
  ) +
  geom_density(color = "black", alpha = 1 / 2) +
  labs(
    title = "16 passi",
    y = "Densità"
  )

(p1 | p2 | p3) & coord_cartesian(xlim = c(-6, 6))

```

Questa simulazione mostra che qualunque processo nel quale viene sommato un certo numero di valori casuali, tutti provenienti dalla medesima distribuzione, converge ad una distribuzione Normale. Non importa quale sia la forma della distribuzione di partenza: essa può essere uniforme, come nell'esempio presente, o di qualunque altro tipo. La forma del-

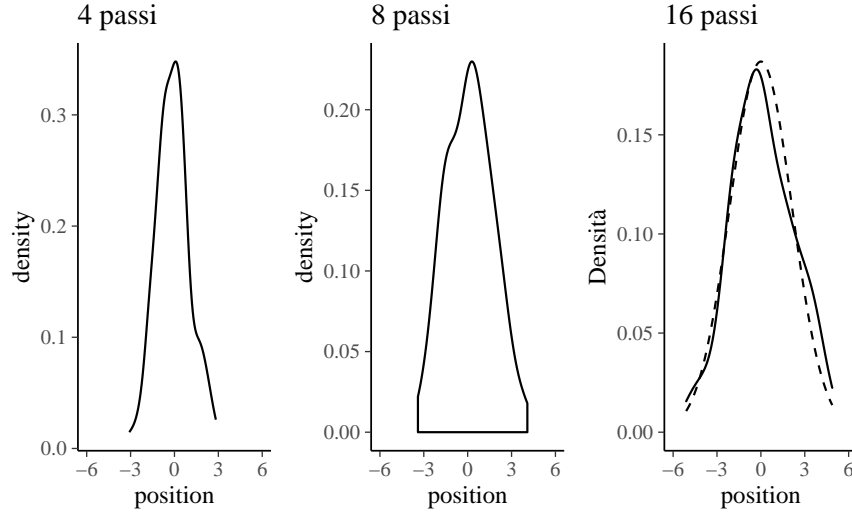


Figura 8.4: Kernel density plot dei risultati della passeggiata casuale riportata nella figura precedente, dopo 4, 8 e 16 passi. Nel pannello di destra, una densità Normale di opportuni parametri è sovrapposta all'istogramma liscio.

la distribuzione da cui viene realizzato il campionamento determina la velocità della convergenza alla Normale. In alcuni casi la convergenza è lenta; in altri casi la convergenza è molto rapida (come nell'esempio presente).

Da un punto di vista formale, diciamo che una variabile casuale continua Y ha una distribuzione Normale se la sua densità è

$$f(y; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{(y - \mu)^2}{2\sigma^2} \right\}, \quad (8.1)$$

dove $\mu \in \mathbb{R}$ e $\sigma > 0$ sono i parametri della distribuzione.

La densità normale è unimodale e simmetrica con una caratteristica forma a campana e con il punto di massima densità in corrispondenza di μ .

Il significato dei parametri μ e σ che appaiono nella (8.1) viene chiarito dalla dimostrazione che

$$\mathbb{E}(X) = \mu, \quad \mathbb{V}(X) = \sigma^2. \quad (8.2)$$

La rappresentazione grafica di quattro densità Normali tutte con media 0 e con deviazioni standard 0.25, 0.5, 1 e 2 è fornita nella figura 8.5.

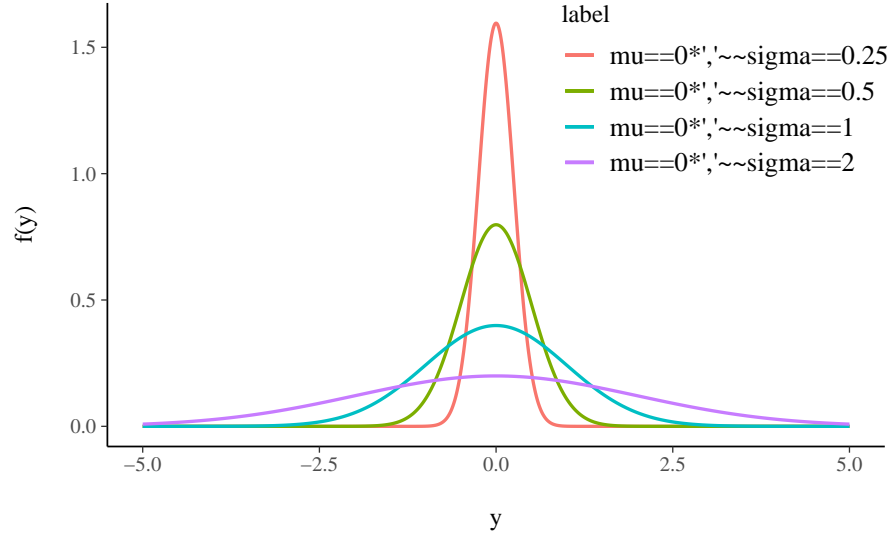


Figura 8.5: Alcune distribuzioni Normali.

8.2.1 Concentrazione

È istruttivo osservare il grado di concentrazione della distribuzione Normale attorno alla media:

$$\begin{aligned} P(\mu - \sigma < X < \mu + \sigma) &= P(-1 < Z < 1) \simeq 0.683, \\ P(\mu - 2\sigma < X < \mu + 2\sigma) &= P(-2 < Z < 2) \simeq 0.956, \\ P(\mu - 3\sigma < X < \mu + 3\sigma) &= P(-3 < Z < 3) \simeq 0.997. \end{aligned}$$

Si noti come un dato la cui distanza dalla media è superiore a 3 volte la deviazione standard presenti un carattere di eccezionalità perché meno del 0.3% dei dati della distribuzione Normale presentano questa caratteristica.

Per indicare la distribuzione Normale si usa la notazione $\mathcal{N}(\mu, \sigma)$.

8.2.2 Funzione di ripartizione

Il valore della funzione di ripartizione di Y nel punto y è l'area sottesa alla curva di densità $f(y)$ nella semiretta $(-\infty, y]$. Non esiste alcuna funzione elementare per la funzione di ripartizione

$$F(y) = \int_{-\infty}^y \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(y-\mu)^2}{2\sigma^2}\right\} dy, \quad (8.3)$$

pertanto le probabilità $P(Y < y)$ vengono calcolate mediante integrazione numerica approssimata. I valori della funzione di ripartizione di una variabile casuale Normale sono dunque forniti da un software.

Esercizio 8.1. Usiamo R per calcolare la funzione di ripartizione della Normale. La funzione `pnorm(q, mean, sd)` restituisce la funzione di ripartizione della Normale con media `mean` e deviazione standard `sd`, ovvero l'area sottesa alla funzione di densità di una Normale con media `mean` e deviazione standard `sd` nell'intervallo $[-\infty, q]$.

Per esempio, in precedenza abbiamo detto che il 68% circa dell'area sottesa ad una Normale è compresa nell'intervallo $\mu \pm \sigma$. Verifichiamo per la distribuzione del QI $\sim \mathcal{N}(\mu = 100, \sigma = 15)$:

```
pnorm(100 + 15, 100, 15) - pnorm(100 - 15, 100, 15)
#> [1] 0.6827
```

Il 95% dell'area è compresa nell'intervallo $\mu \pm 1.96 \cdot \sigma$:

```
pnorm(100 + 1.96 * 15, 100, 15) - pnorm(100 - 1.96 * 15, 100, 15)
#> [1] 0.95
```

Quasi tutta la distribuzione è compresa nell'intervallo $\mu \pm 3 \cdot \sigma$:

```
pnorm(100 + 3 * 15, 100, 15) - pnorm(100 - 3 * 15, 100, 15)
#> [1] 0.9973
```

8.2.3 Distribuzione Normale standard

La distribuzione Normale di parametri $\mu = 0$ e $\sigma = 1$ viene detta *distribuzione Normale standard*. La famiglia Normale è l'insieme avente come

elementi tutte le distribuzioni Normali con parametri μ e σ diversi. Tutte le distribuzioni Normali si ottengono dalla Normale standard mediante una trasformazione lineare: se $Y \sim \mathcal{N}(\mu_Y, \sigma_Y)$ allora

$$X = a + bY \sim \mathcal{N}(\mu_X = a + b\mu_Y, \sigma_X = |b| \sigma_Y). \quad (8.4)$$

L'area sottesa alla curva di densità di $\mathcal{N}(\mu, \sigma)$ nella semiretta $(-\infty, y]$ è uguale all'area sottesa alla densità Normale standard nella semiretta $(-\infty, z]$, in cui $z = (y - \mu_Y)/\sigma_Y$ è il punteggio standard di Y . Per la simmetria della distribuzione, l'area sottesa nella semiretta $[1, \infty)$ è uguale all'area sottesa nella semiretta $(-\infty, 1]$ e quest'ultima coincide con $F(-1)$. Analogamente, l'area sottesa nell'intervallo $[y_a, y_b]$, con $y_a < y_b$, è pari a $F(z_b) - F(z_a)$, dove z_a e z_b sono i punteggi standard di y_a e y_b .

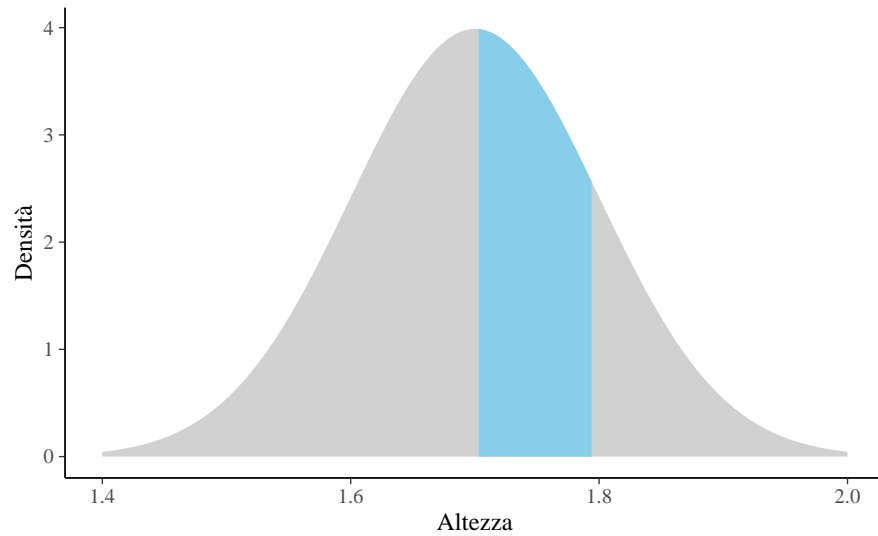
Si ha anche il problema inverso rispetto a quello del calcolo delle aree: dato un numero $0 \leq p \leq 1$, il problema è quello di determinare un numero $z \in \mathbb{R}$ tale che $P(Z < z) = p$. Il valore z cercato è detto *quantile* di ordine p della Normale standard e può essere trovato mediante un software.

Esercizio 8.2. Supponiamo che l'altezza degli individui adulti segua la distribuzione Normale di media $\mu = 1.7$ m e deviazione standard $\sigma = 0.1$ m. Vogliamo sapere la proporzione di individui adulti con un'altezza compresa tra 1.7 e 1.8 m.

Il problema ci chiede di trovare l'area sottesa alla distribuzione $\mathcal{N}(\mu = 1.7, \sigma = 0.1)$ nell'intervallo $[1.7, 1.8]$:

```
df <- tibble(x = seq(1.4, 2.0, length.out = 100)) %>%
  mutate(y = dnorm(x, mean = 1.7, sd = 0.1))

ggplot(df, aes(x, y)) +
  geom_area(fill = "sky blue") +
  gghighlight(x < 1.8 & x > 1.7) +
  labs(
    x = "Altezza",
    y = "Densità"
  )
```



La risposta si trova utilizzando la funzione di ripartizione $F(X)$ della legge $\mathcal{N}(1.7, 0.1)$ in corrispondenza dei due valori forniti dal problema: $F(X = 1.8) - F(X = 1.7)$. Utilizzando la seguente istruzione

```
pnorm(1.8, 1.7, 0.1) - pnorm(1.7, 1.7, 0.1)
#> [1] 0.3413
```

otteniamo il 31.43%.

In maniera equivalente, possiamo standardizzare i valori che delimitano l'intervallo considerato e utilizzare la funzione di ripartizione della normale standardizzata. I limiti inferiore e superiore dell'intervallo sono

$$z_{\text{inf}} = \frac{1.7 - 1.7}{0.1} = 0, \quad z_{\text{sup}} = \frac{1.8 - 1.7}{0.1} = 1.0,$$

quindi otteniamo

```
pnorm(1.0, 0, 1) - pnorm(0, 0, 1)
#> [1] 0.3413
```

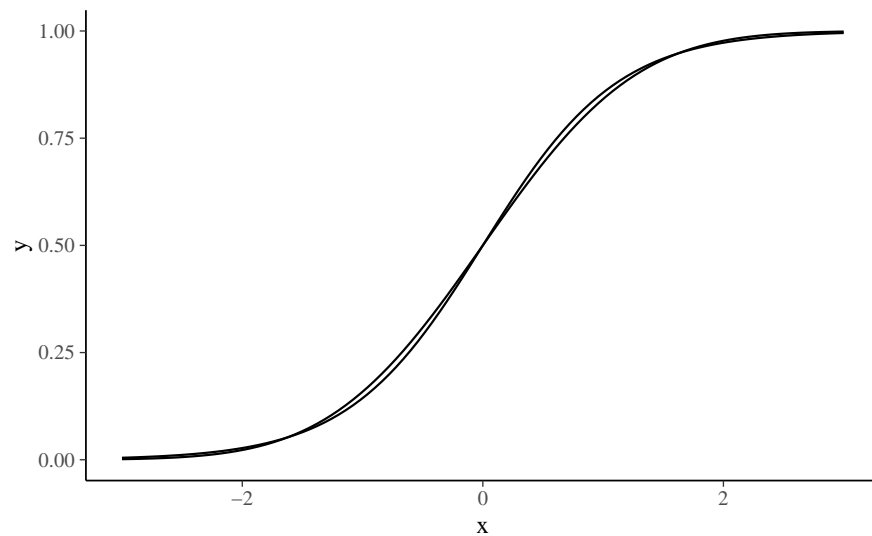
Il modo più semplice per risolvere questo problema resta comunque quello di rendersi conto che la probabilità richiesta non è altro che la metà

dell'area sottesa dalle distribuzioni Normali nell'intervallo $[\mu - \sigma, \mu + \sigma]$, ovvero $0.683/2$.

8.2.3.1 Funzione di ripartizione della normale standard e funzione logistica

Si noti che la funzione logistica (in blu), pur essendo del tutto diversa dalla Normale dal punto di vista formale, assomiglia molto alla Normale standard quando le due cdf hanno la stessa varianza.

```
tibble(x = c(-3, 3)) %>%  
  ggplot(aes(x = x)) +  
    stat_function(fun = pnorm) +  
    stat_function(  
      fun = plogis,  
      args = list(scale = 0.56)  
    )
```



8.3 Teorema del limite centrale

Laplace dimostrò il teorema del limite centrale (TLC) nel 1812. Il TLC ci dice che se prendiamo una sequenza di variabili casuali indipendenti e le sommiamo, tale somma tende a distribuirsi come una Normale. Il TLC specifica inoltre, sulla base dei valori attesi e delle varianze delle v.c. che vengono sommate, quali saranno i parametri della distribuzione Normale così ottenuta.

Teorema 8.1. *Si supponga che $Y = Y_1, Y_2, \dots, Y_N$ sia una sequenza di v.a. i.i.d. con $\mathbb{E}(Y_n) = \mu$ e $\text{SD}(Y_n) = \sigma$. Si definisca una nuova v.c. come la media di Y :*

$$Z = \frac{1}{N} \sum_{n=1}^N Y_n.$$

Con $N \rightarrow \infty$, Z tenderà ad una Normale con lo stesso valore atteso di Y_n e una deviazione standard che sarà più piccola della deviazione standard originaria di un fattore pari a $\sqrt{\frac{1}{N}}$:

$$p_Z(z) \rightarrow \mathcal{N} \left(z \mid \mu, \frac{1}{\sqrt{N}} \cdot \sigma \right). \quad (8.5)$$

Il TLC può essere generalizzato a variabili che non hanno la stessa distribuzione purché siano indipendenti e abbiano aspettative e varianze finite.

Molti fenomeni naturali, come l'altezza dell'uomo adulto di entrambi i sessi, sono il risultato di una serie di effetti additivi relativamente piccoli, la cui combinazione porta alla normalità, indipendentemente da come gli effetti additivi sono distribuiti. In pratica, questo è il motivo per cui la distribuzione normale ha senso come rappresentazione di molti fenomeni naturali.

8.4 Distribuzione Chi-quadrato

Dalla Normale deriva la distribuzione χ^2 . La distribuzione χ^2_k con k gradi di libertà descrive la variabile casuale

$$Z_1^2 + Z_2^2 + \dots + Z_k^2,$$

dove Z_1, Z_2, \dots, Z_k sono variabili casuali i.i.d. con distribuzione Normale standard $\mathcal{N}(0, 1)$. La variabile casuale chi-quadrato dipende dal parametro intero positivo $\nu = k$ che ne identifica il numero di gradi di libertà. La densità di probabilità di χ^2_ν è

$$f(x) = C_\nu x^{\nu/2-1} \exp(-x/2), \quad \text{se } x > 0,$$

dove C_ν è una costante positiva.

La figura 8.6 mostra alcune distribuzioni Chi-quadrato variando il parametro ν .

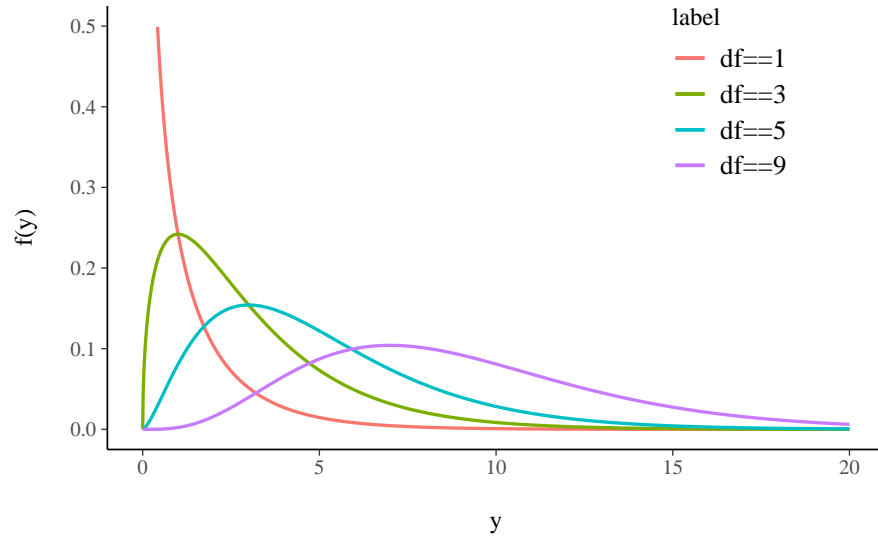


Figura 8.6: Alcune distribuzioni Chi-quadrato.

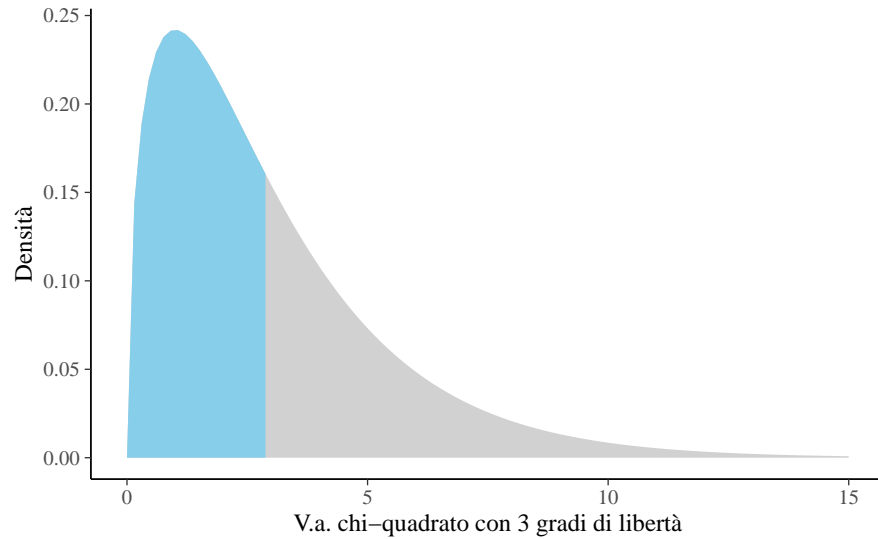
8.4.1 Proprietà

- La distribuzione di densità χ^2_ν è asimmetrica.
- Il valore atteso di una variabile χ^2_ν è uguale a ν .
- La varianza di una variabile χ^2_ν è uguale a 2ν .
- Per $k \rightarrow \infty$, la $\chi^2_\nu \rightarrow \mathcal{N}$.
- Se X e Y sono due variabili casuali chi-quadrato indipendenti con ν_1 e ν_2 gradi di libertà, ne segue che $X + Y \sim \chi^2_m$, con $m = \nu_1 + \nu_2$. Tale principio si estende a qualunque numero finito di variabili casuali chi-quadrato indipendenti.

Esercizio 8.3. Usiamo R per disegnare la densità chi-quadrato con 3 gradi di libertà dividendo l'area sottesa alla curva di densità in due parti uguali.

```
df <- tibble(x = seq(0, 15.0, length.out = 100)) %>%
  mutate(y = dchisq(x, 3))

ggplot(df, aes(x, y)) +
  geom_area(fill = "sky blue") +
  gghighlight(x < 3) +
  labs(
    x = "V.a. chi-quadrato con 3 gradi di libertà",
    y = "Densità"
  )
```



8.5 Distribuzione t di Student

Dalle distribuzioni Normale e Chi quadrato deriva un'altra distribuzione molto nota, la t di Student. Se $Z \sim \mathcal{N}$ e $W \sim \chi^2_\nu$ sono due variabili casuali indipendenti, allora il rapporto

$$T = \frac{Z}{\left(\frac{W}{\nu}\right)^{\frac{1}{2}}} \quad (8.6)$$

definisce la distribuzione t di Student con ν gradi di libertà. Si usa scrivere $T \sim t_\nu$. L'andamento della distribuzione t di Student è simile a quello della distribuzione Normale, ma ha una maggiore dispersione (ha le code più pesanti di una Normale, ovvero ha una varianza maggiore di 1).

La figura 8.7 mostra alcune distribuzioni t di Student variando il parametro ν .

8.5.1 Proprietà

La variabile casuale t di Student soddisfa le seguenti proprietà:

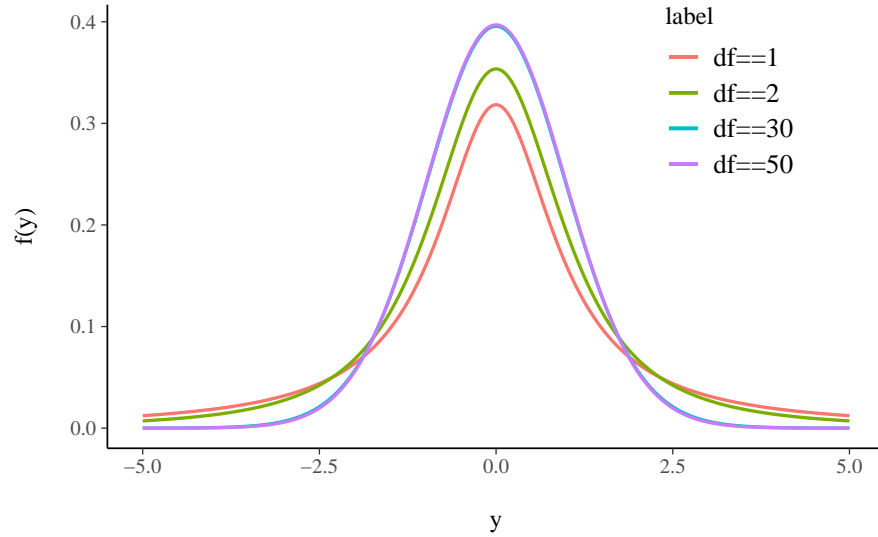


Figura 8.7: Alcune distribuzioni t di Student.

1. Per $\nu \rightarrow \infty$, t_ν tende alla normale standard $\mathcal{N}(0, 1)$.
2. La densità della t_ν è una funzione simmetrica con valore atteso nullo.
3. Per $\nu > 2$, la varianza della t_ν vale $\nu/(\nu-2)$; pertanto è sempre maggiore di 1 e tende a 1 per $\nu \rightarrow \infty$.

8.6 Funzione beta di Eulero

La funzione beta di Eulero è una funzione matematica, *non* una densità di probabilità. La menzioniamo qui perché viene utilizzata nella distribuzione Beta. La funzione beta si può scrivere in molti modi diversi; per i nostri scopi la scriveremo così:

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}, \quad (8.7)$$

dove $\Gamma(x)$ è la funzione Gamma, ovvero il fattoriale discendente, cioè $x(x-1)(x-2)\dots(x-n+1)$.

8.7 Distribuzione Beta

Una distribuzione che viene usata per modellare percentuali e proporzioni è la distribuzione Beta in quanto è definita sull'intervallo $(0; 1)$ – ma non include i valori 0 o 1. La distribuzione Beta è una distribuzione estremamente flessibile e può assumere molti tipi di forme diverse (un'illustrazione è fornita dalla seguente GIF animata¹). Una definizione formale è la seguente.

Definizione 8.1. Sia π una variabile casuale che può assumere qualsiasi valore compreso tra 0 e 1, cioè $\pi \in [0, 1]$. Diremo che π segue la distribuzione Beta di parametri α e β , $\pi \sim \text{Beta}(\alpha, \beta)$, se la sua densità è

$$\begin{aligned} \text{Beta}(\pi \mid \alpha, \beta) &= \frac{1}{B(\alpha, \beta)} \pi^{\alpha-1} (1 - \pi)^{\beta-1} \\ &= \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \pi^{\alpha-1} (1 - \pi)^{\beta-1} \quad \text{per } \pi \in [0, 1], \end{aligned} \quad (8.8)$$

laddove $B(\alpha, \beta)$ è la funzione beta.

I termini α e β sono i parametri della distribuzione Beta e devono essere entrambi positivi. Tali parametri possono essere interpretati come l'espressione delle nostre credenze a priori relative ad una sequenza di prove Bernoulliane. Il parametro α rappresenta il numero di “successi” e il parametro β il numero di “insuccessi”:

$$\frac{\text{Numero di successi}}{\text{Numero di successi} + \text{Numero di insuccessi}} = \frac{\alpha}{\alpha + \beta}.$$

Il rapporto $\frac{1}{B(\alpha, \beta)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}$ è una costante di normalizzazione:

$$\int_0^1 \pi^{\alpha-1} (1 - \pi)^{\beta-1} d\pi = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}. \quad (8.9)$$

Facciamo un esempio con dei valori arbitrari $\alpha = 3$ e $\beta = 6$:

¹https://en.wikipedia.org/wiki/File:PDF_of_the_Beta_distribution.gif

```

a <- 3
b <- 6
integrand <- function(p) {
  p^
  {
    a - 1
  } * (1 - p)^{
    b - 1
  }
}
integrate(integrand, lower = 0, upper = 1)
#> 0.005952 with absolute error < 6.6e-17

```

ovvero

```

1 / (gamma(a + b) / (gamma(a) * gamma(b)))
#> [1] 0.005952

```

Il valore atteso, la moda e la varianza di una distribuzione Beta sono dati dalle seguenti equazioni:

$$\mathbb{E}(\pi) = \frac{\alpha}{\alpha + \beta}, \quad (8.10)$$

$$\text{Mo}(\pi) = \frac{\alpha - 1}{\alpha + \beta - 2}, \quad (8.11)$$

$$\mathbb{V}(\pi) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (8.12)$$

Osservazione. Attenzione alle parole: in questo contesto, il termine “beta” viene utilizzato con tre significati diversi:

- la distribuzione di densità Beta,
- la funzione matematica beta,
- il parametro β .

Al variare di α e β si ottengono molte distribuzioni di forma diversa; per $\alpha = \beta = 1$ si ha la densità uniforme. Vari esempi di distribuzioni Beta sono mostrati nella figura 8.8.

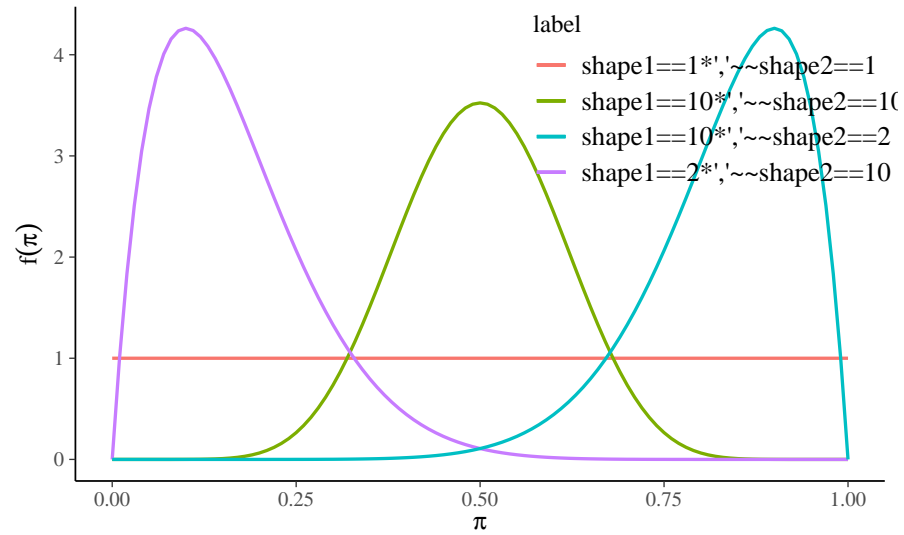
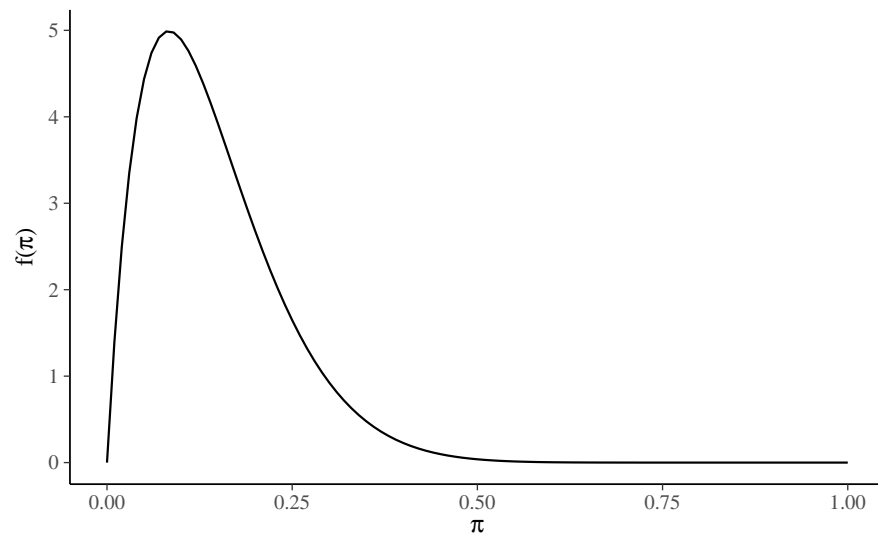


Figura 8.8: Alcune distribuzioni Beta.

Si può ottenere una rappresentazione grafica della distribuzione $\text{Beta}(\pi \mid \alpha, \beta)$ con la funzione `plot_beta()` del pacchetto `bayesrules`. Per esempio:

```
bayesrules::plot_beta(alpha = 2, beta = 12)
```



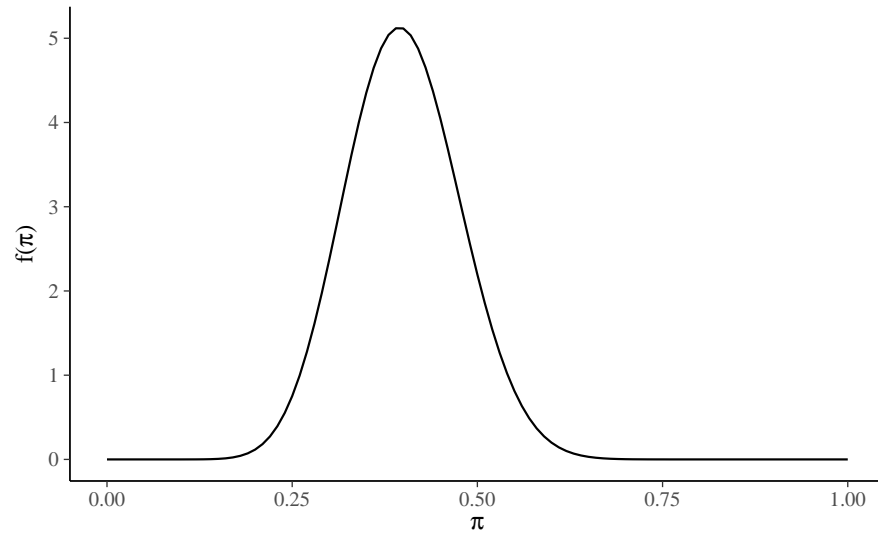
La funzione `bayesrules::summarize_beta()` ci restituisce la media, moda e varianza della distribuzione Beta. Per esempio:

```
bayesrules::summarize_beta(alpha = 2, beta = 12)
#>      mean    mode    var    sd
#> 1 0.1429 0.08333 0.008163 0.09035
```

Esercizio 8.4. Nel disturbo depressivo la recidiva è definita come la comparsa di un nuovo episodio depressivo che si manifesta dopo un prolungato periodo di recupero (6-12 mesi) con stato di eutimia (umore relativamente normale). Supponiamo che una serie di studi mostri una comparsa di recidiva in una proporzione che va dal 20% al 60% dei casi, con una media del 40% (per una recente discussione, si veda [Nuggerud-Galeas et al., 2020](#)). Sulla base di queste ipotetiche informazioni, è possibile usare la distribuzione Beta per rappresentare le nostre credenze a priori relativamente alla probabilità di recidiva. Per fare questo dobbiamo trovare i parametri della distribuzione Beta tali per cui la massa della densità sia compresa tra 0.2 e 0.6, con la media in corrispondenza di 0.4. Procedendo per tentativi ed errori, ed usando la funzione `bayesrules::plot_beta()`, un risultato possibile è $B(16, 24)$.

```
find_pars <- function(ev, n) {
  a <- ev * n
  b <- n - a
  return(c(round(a), round(b)))
}

pars <- find_pars(.4, 40)
pars
#> [1] 16 24
bayesrules::plot_beta(pars[1], pars[2])
```



La media della distribuzione a priori diventa:

```
16 / (16 + 24)
#> [1] 0.4
```

e la moda è

```
(16 - 1) / (16 + 24 - 2)
#> [1] 0.3947
```

Inoltre, la deviazione standard della distribuzione a priori diventa

```
sqrt((16 * 24) / ((16 + 24)^2 * (16 + 24 + 1)))
#> [1] 0.07651
```

uguale a circa 8 punti percentuali. Verifichiamo:

```
bayesrules::summarize_beta(alpha = 16, beta = 24)
#>   mean  mode   var   sd
#> 1  0.4 0.3947 0.005854 0.07651
```

Questo significa che le nostre credenze a priori rispetto la possibilità di recidiva tendono a deviare di circa 8 punti percentuali rispetto alla media della distribuzione a priori che corrisponde circa a 0.40.

8.8 Distribuzione di Cauchy

La distribuzione di Cauchy è un caso speciale della distribuzione di t di Student con 1 grado di libertà. È definita da una densità di probabilità che corrisponde alla funzione, dipendente da due parametri θ e d (con la condizione $d > 0$),

$$f(x; \theta, d) = \frac{1}{\pi d} \frac{1}{1 + \left(\frac{x-\theta}{d}\right)^2}, \quad (8.13)$$

dove θ è la mediana della distribuzione e d ne misura la larghezza a metà altezza.

8.9 Distribuzione log-normale

Sia y una variabile casuale avente distribuzione normale con media μ e varianza σ^2 . Definiamo poi una nuova variabile casuale x attraverso la relazione

$$x = e^y \quad \Longleftrightarrow \quad y = \log x.$$

Il dominio di definizione della x è il semiasse $x > 0$ e la densità di probabilità $f(x)$ è data da

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x} \exp \left\{ -\frac{(\log x - \mu)^2}{2\sigma^2} \right\}. \quad (8.14)$$

Questa funzione di densità si chiama log-normale.

Il valore atteso e la varianza di una distribuzione log-normale sono dati dalle seguenti equazioni:

$$\mathbb{E}(x) = \exp \left\{ \mu + \frac{\sigma^2}{2} \right\}. \quad (8.15)$$

$$\mathbb{V}(x) = \exp \{ 2\mu + \sigma^2 \} (\exp \{ \sigma^2 \} - 1). \quad (8.16)$$

Si può dimostrare che il prodotto di variabili casuali log-normali ed indipendenti segue una distribuzione log-normale.

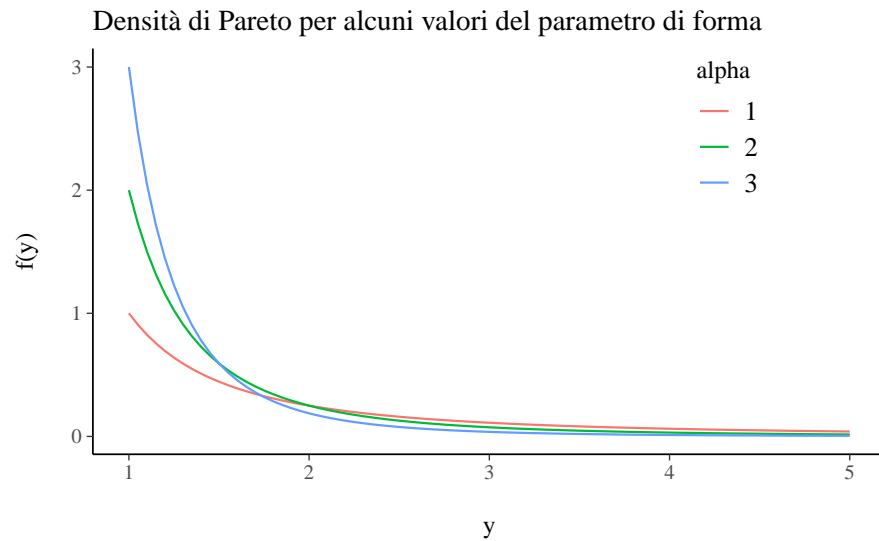
8.10 Distribuzione di Pareto

La distribuzione paretiana (o distribuzione di Pareto) è una distribuzione di probabilità continua e così chiamata in onore di Vilfredo Pareto. La distribuzione di Pareto è una distribuzione di probabilità con legge di potenza utilizzata nella descrizione di fenomeni sociali e molti altri tipi di fenomeni osservabili. Originariamente applicata per descrivere la distribuzione del reddito in una società, adattandosi alla tendenza che una grande porzione di ricchezza è detenuta da una piccola frazione della popolazione, la distribuzione di Pareto è diventata colloquialmente nota e indicata come il principio di Pareto, o “regola 80-20”. Questa regola afferma che, ad esempio, l’80% della ricchezza di una società è detenuto dal 20% della sua popolazione. Viene spesso applicata nello studio della distribuzione del reddito, della dimensione dell’impresa, della dimensione di una popolazione e nelle fluttuazioni del prezzo delle azioni.

La densità di una distribuzione di Pareto è

$$f(x) = (x_m/x)^\alpha,$$

dove x_m (parametro di scala) è il minimo (necessariamente positivo) valore possibile di X e α è un parametro di forma.



La distribuzione di Pareto ha una asimmetria positiva. Il supporto della distribuzione di Pareto è la retta reale positiva. Tutti i valori devono essere maggiori del parametro di scala x_m , che è in realtà un parametro di soglia.

Commenti e considerazioni finali

In questa dispensa le densità continue che useremo più spesso sono la distribuzione gaussiana e la distribuzione Beta. Faremo un uso limitato della distribuzione t di Student e della distribuzione di Cauchy. Le altre distribuzioni qui descritte sono state presentate solo per completezza.



Bibliografía

- Albert, J. and Hu, J. (2019). *Probability and Bayesian Modeling*. Chapman and Hall/CRC.
- de Finetti, B. (1931). Probabilismo. *Logos*, pages 163–219.
- Horn, S. and Loewenstein, G. (2021). Underestimating learning by doing. *Available at SSRN 3941441*.
- McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press, Boca Raton, Florida, 2nd edition edition.
- Nuggerud-Galeas, S., Sáez-Benito Suescun, L., Berenguer Torrijo, N., Sáez-Benito Suescun, A., Aguilar-Latorre, A., Magallón Botaya, R., and Oliván Blázquez, B. (2020). Analysis of depressive episodes, their recurrence and pharmacologic treatment in primary care patients: A retrospective descriptive study. *Plos one*, 15(5):e0233454.