Probabilità e Statistica

Note del corso 2020/21

LUIGI AMEDEO BIANCHI

INDICE

Lezione 1

INTRODUZIONE

Queste note coprono ed espandono quanto presentato nel corso di Probabilità e Statistica tenuto nel secondo semestre dell'anno accademico 2019/20 e dell'anno accademico 2020/21 al Corso di Laurea triennale in Informatica.

Parte di queste note (le prime lezioni) è stata pubblicata dalla casa editrice Scienza Express nella collana UMath. Per parte di queste note ho preso ispirazione dalle note del corso del prof. Claudio Agostinelli. Tra le altre fonti devo ringraziare il prof. Francesco Morandin dell'Università di Parma e il libro *Probabilità e Statistica per l'ingegneria e le scienze* di Sheldon Ross, pubblicato da Apogeo.

Le note sono scritte in TEXMACS (https://www.texmacs.org). Alcune delle immagini sono realizzate in TikZ, altre in R.

COSA È LA PROBABILITÀ?

Possiamo vedere la probabilità come uno strumento per misurare l'incertezza, pensata in diverse accezioni: tra casi o soggetti, nel tempo, nello spazio, nella misurazione…

Abbiamo tutti un'idea intuitiva di cosa intendiamo parlando di probabilità nel linguaggio informale di tutti i giorni: qualcosa è tanto più probabile quanto meno ci sorprenderebbe vederla accadere. Tuttavia abbiamo bisogno di passare a un linguaggio più formale, per mezzo della matematica, per assicurarci che queste valutazioni siano coerenti. Il linguaggio naturale, infatti, lascia aperti spiragli a possibili fraintendimenti, dovuti in parte all'uso di "categorie mentali" diverse.

Esempio 1. Linda è una giovane donna che ha studiato Scienze Sociali a Pisa. Negli anni dell'università ha partecipato a numerose manifestazioni contro la discriminazione delle minoranze, anche nel mondo accademico. Durante la visita del Presidente della Repubblica all'Ateneo di Pisa si è fatta portavoce delle richieste degli studenti, chiedendo pubblicamente al Presidente di intervenire per ampliare gli strumenti finanziari a sostegno degli studenti in difficoltà economiche. Si è laureata con una tesi critica dell'impatto negativo del mondo della finanza sulla società.

È più probabile che oggi Linda sia impiegata in banca o che sia responsabile delle pari opportunità in Banca Etica?

L'esempio precedente è stato proposto, in versioni leggermente diverse, dagli psicologi israeliani Kahneman e Tverski in alcuni loro studi. Dei volontari intervistati una considerevole maggioranza assegnava una probabilità maggiore alla seconda opzione. È una storia che ci tenta: ci sembra più in linea con il racconto precedente, si sposa meglio con l'idea che ci siamo fatti di Linda. Eppure, da un punto di vista della coerenza logica, è la risposta sbagliata.

Infatti se Linda lavora come responsabile delle pari opportunità in Banca Etica, allora è un'impiegata in una banca e di conseguenza è il primo evento ad avere una probabilità maggiore. Se Linda non lavora in banca, allora non lavora nemmeno in Banca Etica, ma potrebbe aver accettato un lavoro in un'altra banca (per necessità, perché ha cambiato le proprie idee o magari solo per caso), quindi ci sono più modi in cui Linda sta lavorando in una banca qualunque che non modi in cui è in Banca Etica e addirittura specificamente come responsabile delle pari opportunità.

Anche per risolvere problemi di questo tipo, negli anni Trenta del secolo scorso è stata sviluppata la cosiddetta *teoria assiomatica della probabilità*, principalmente da A. Kolmogorov. Questa teoria, su cui si baserà la prima parte di questo corso, identifica alcuni assiomi e alcune proprietà che una probabilità deve avere per essere coerente. Dice inoltre come è possibile manipolare matematicamente le probabilità, da cui il nome di *calcolo delle probabilità*.

La teoria assiomatica, però, non dà un teorema di unicità della probabilità: non garantisce che esista una e una sola scelta di probabilità che soddisfa gli assiomi. Gli assiomi danno dei vincoli, ma lasciano anche libertà di scelta: certi aspetti di una probabilità sono una scelta di modello, dipendono da quello che vogliamo rappresentare, ma anche da posizioni filosofiche. Possiamo parlare di probabilità frequentista o soggettivista (detta anche bayesiana), principalmente, ma ci sono poi numerose sfumature e interpretazioni, come la probabilità logica, la probabilità comparativa e così via. Non approfondiremo questi aspetti, anche se vedremo qualche accenno in seguito, perché vale quanto detto prima: qualunque sia il nostro approccio filosofico, la nostra scelta di probabilità deve soddisfare gli assiomi e di conseguenza godrà delle proprietà che studieremo come conseguenza degli assiomi stessi.

Può essere difficile mettere assieme la nostra idea di matematica come strumento deterministico (e assoluto) per eccellenza con il concetto di probabilità e l'incertezza che le associamo. Possiamo però cercare di capire la situazione con un'analogia: la matematica è il sistema operativo, mentre la probabilità è un programma lato utente, un'interfaccia tra il mondo reale e le sue incertezze e la matematica. La probabilità traduce (o rappresenta) l'incertezza in termini matematici permettendoci così di usare questo potente linguaggio formale per studiare situazioni non deterministiche.

Il corso, oltre alla probabilità, ha nel nome anche la statistica. Possiamo considerare la statistica come se fosse divisa in due: statistica descrittiva e statistica inferenziale.

La statistica descrittiva lavora su un'intera popolazione e cerca di *descriverla* in termini numerici, sintetizzando alcune caratteristiche della popolazione attraverso dei numeri. Tuttavia spesso non è possibile avere dati sull'intera popolazione di interesse, ma si ha accesso solamente a un campione casuale (ecco il primo collegamento con la probabilità) della popolazione stessa. La statistica inferenziale ci dà strumenti per *dedurre* o *inferire* caratteristiche della popolazione intera a partire da misurazioni fatte sul solo campione. Dal momento che il campione è casuale, questa descrizione dedotta non può essere certa, ma contiene al suo interno una misura di incertezza. Dietro alla statistica inferenziale abbiamo modelli probabilistici, per studiare i quali avremo bisogno della probabilità.

La probabilità e la statistica sono importanti in generale, dal momento che ci permettono di descrivere in termini matematici situazioni di incertezza, come quelle che incontriamo tutti i giorni, e di prendere decisioni che tengano opportunamente conto di tale incertezza. Nel campo specifico dell'informatica, poi, ci sono alcuni temi che si appoggiano alla probabilità e alla statistica, ad esempio il machine learning e lo statistical learning, ma anche gli algoritmi casuali, alcune strutture dati (tabelle hash ottimizzate), i processi di assegnazione delle risorse (Random Access Memory, ma anche cloud), la teoria dei segnali, in particolare nel canali con rumore, gli algoritmi di compressione e così via.

Parte I

Probabilità

CAPITOLO 1

COMBINATORIA

Cominciamo a parlare di probabilità, in una situazione speciale in cui tutti i casi sono *equiprobabili*. Possiamo calcolare la probabilità di qualcosa semplicemente contando tutti i casi favorevoli (cioè i casi in cui si verifica il qualcosa che cerchiamo) e dividere questo numero per quello di tutti i casi possibili.

È chiaro però che, se da un punto di vista intuitivo questa definizione ci può andare bene, da un punto di vista rigoroso lascia molto a desiderare: se non abbiamo ancora definito cosa significhi *probabilità*, come possiamo parlare di casi equiprobabili? Al tempo stesso questo approccio è molto naturale: a ben pensarci tutte le misurazioni iniziano usando un riferimento. Non solo, anche storicamente questo è stato uno dei primi modi di avvicinarsi alla probabilità, seppur al prezzo di rischiare qualche errore in più.

Lasciamo per il momento da parte questa perplessità e abbracciamo l'approccio intuitivo: possiamo comunque vedere numerosi esercizi ed esempi interessanti. Il punto cruciale è che trasformiamo il problema di calcolare la probabilità di qualcosa in un conteggio: vogliamo contare i casi favorevoli e i casi totali. La branca della matematica che si occupa di questo tipo di problemi si chiama *combinatoria*.

Per prima cosa vogliamo introdurre i tre principi fondamentali della combinatoria. Per fare questo usiamo il linguaggio della teoria elementare degli insiemi. Chi avesse bisogno di un ripasso, troverà un po' di risultati in Appendice A.1.

1.1. I TRE PRINCIPI DELLA COMBINATORIA

Dopo questa breve escursione nella teoria elementare degli insiemi, torniamo alla combinatoria, in particolare ai tre principi che avevamo menzionato prima.

Il *primo principio della combinatoria* sostituisce il conteggio degli elementi di un insieme con il conteggio degli elementi di una sua partizione, ossia con una rappresentazione dell'insieme come unione disgiunta di suoi sottoinsiemi. È il principio che ci apre la via al paradigma del *divide et impera*: spezzare un problema in parti più piccole e mutualmente esclusive affrontandole separatamente e combinando alla fine i risultati.

PROPOSIZIONE 1.1. *Siano A un insieme e $\{E_i\}_{i=1}^n$ una partizione di A. Allora $\#A = \sum_{i=1}^n \#E_i$.*

Cosa c'entra questo con la combinatoria? Proviamo a fare un paio di esempi.

Esempio 1.2. Con un buono regalo possiamo decidere se avere o un film o un videogioco. Sapendo che ci sono 10 film e 6 videogiochi disponibili, in tutto abbiamo $10 + 6$ omaggi diversi tra cui scegliere quale portarci a casa. In questo caso A è l'insieme di tutti gli omaggi tra cui possiamo scegliere e la sua partizione è data da E_1, insieme dei film disponibili, ed E_2, insieme dei videogiochi disponibili.

Esempio 1.3. In una scuola ci sono 28 studentesse e studenti del primo anno, 25 del secondo, 21 del terzo, 26 del quarto e 26 del quinto. In tutto, nella scuola ci sono allora $28 + 25 + 21 + 26 + 26 = 126$ studentesse e studenti; infatti ognuno di loro non può che appartenere a uno e un solo anno di corso. Qui A è l'insieme di tutti gli studenti della scuola, ed E_i, per i da 1 a 5, l'insieme di quelli dell'i-esimo anno.

Per introdurre il *secondo principio della combinatoria*, ossia il principio del prodotto, dobbiamo prima richiamare brevemente il prodotto cartesiano di insiemi.

DEFINIZIONE 1.4. *Dati due insiemi A e B, il loro* prodotto cartesiano, *indicato con $A \times B$, è l'insieme delle coppie ordinate (a,b) tali che $a \in A$ e $b \in B$.*

Notiamo l'aggettivo che compare nella precedente definizione: le coppie che consideriamo sono *ordinate*. Questo significa che una coppia non è determinata solamente dagli elementi che la compongono, ma anche dall'ordine in cui compaiono: le due coppie $(1,3)$ e $(3,1)$ sono coppie ordinate distinte. Vedremo che è importante non dimenticarsi se stiamo considerando coppie (o terne, o n-uple) ordinate o no.

Ora che sappiamo che cosa è il prodotto cartesiano, andiamo a vedere perché ci interessa in combinatoria. In questo caso vogliamo (poco sorprendentemente) contare le coppie ordinate.

PROPOSIZIONE 1.5. *Dati due insiemi A e B e il loro prodotto cartesiano $A \times B$, vale la seguente uguaglianza: $\#(A \times B) = \#A \cdot \#B$.*

Dobbiamo fare attenzione: in generale i due insiemi $A \times B$ e $B \times A$, pur avendo la stessa cardinalità, sono diversi, perché formati da coppie ordinate diverse. D'altra parte, anche se gli insiemi sono distinti, possiamo mostrare una relazione biunivoca tra essi. In particolare la mappa che scambia le due componenti soddisfa questa condizione. In effetti, se pensiamo a quello che significano le varie quantità, stiamo dicendo che anche se i due insiemi hanno lo stesso numero di elementi, non necessariamente sono uguali.

Esempio 1.6. Per fare un esempio più che classico, pensiamo a un pasto in una mensa o in una tavola calda: il pasto consiste di un primo a scelta tra minestra, pasta e riso e di un secondo a scelta tra carne, pesce, formaggio, uova e sformato di verdure. In quanti modi diversi possiamo comporre un pasto?

Vogliamo contare le coppie ordinate in cui alla prima componente abbiamo un primo e alla seconda un secondo (molto appropriatamente). I modi che abbiamo sono in questo caso $3 \cdot 5$. Possiamo leggere il risultato così: per ogni scelta del primo tra i 3 disponibili, abbiamo 5 possibili secondi (e viceversa: visto che la moltiplicazione è commutativa, possiamo anche fissare prima il secondo, scegliendolo fra i 5 a nostra disposizione e, in seguito, determina uno dei 3 primi).

Possiamo definire in modo del tutto simile il prodotto cartesiano tra più di due insiemi, a patto che siano in numero finito, e vale un risultato analogo per la sua cardinalità.

PROPOSIZIONE 1.7. *Data una famiglia finita di insiemi $\{A_i\}_{i=1}^n$, prendiamo il loro prodotto cartesiano $A_1 \times \cdots \times A_n$ che denotiamo anche con $\bigotimes_{i=1}^n A_i$. Vale la seguente uguaglianza: $\#(\bigotimes_{i=1}^n A_i) = \prod_{i=1}^n \#A_i$.*

Esempio 1.8. Torniamo alla nostra mensa: è cambiata la gestione e ora, oltre a un primo e a un secondo come prima, possiamo scegliere anche un contorno, tra patate, carote, spinaci e piselli e un dessert tra budino, crème caramel e gelato. In quanti modi diversi possiamo ora comporre un pasto?

Ora vogliamo contare le 4-uple ordinate, in cui compaiono, nell'ordine, un primo, un secondo, un contorno e un dessert. Le scelte sono, nel medesimo ordine, 3, 5, 4 e 3, per un numero totale di $3 \cdot 5 \cdot 4 \cdot 3 = 180$ modi differenti di comporre un pasto.

Gli insiemi A_i che andiamo a moltiplicare non devono necessariamente essere disgiunti e, in realtà, nemmeno distinti. In particolare nulla ci impedisce di considerare n copie dello stesso insieme. In questo caso abbiamo semplicemente l'insieme $\bigotimes_{i=1}^n A = A^n$, la cui cardinalità è $\#(A^n) = (\#A)^n$.

Esempio 1.9. Quanti sono i possibili PIN a 6 cifre?

In questo caso abbiamo $A = \{0,1,\ldots,9\}$ come insieme nel quale peschiamo ciascuna delle 6 cifre del PIN. Stiamo quindi cercando la cardinalità dell'insieme A^6, cioè il numero 10^6.

Esempio 1.10. Se invece volessimo i PIN a 6 cifre in cui non ci sono cifre consecutive uguali?

Come prima cosa notiamo che non siamo più nel caso precedente; in particolare ci aspettiamo di ottenere un numero più basso, visto che stiamo considerando un sottoinsieme di tutti i PIN possibili. Per la prima cifra[1.1] abbiamo 10 possibili valori (tutti i numeri tra 0 e 9). Quando passiamo alla seconda cifra, adiacente alla prima, uno dei valori non è più a nostra disposizione (quello scelto per la prima cifra). Ma solamente quel valore va escluso, quindi ci restano 9 scelte possibili. Similmente per le cifre successive, per un totale di $10 \cdot 9^5$ possibili PIN che soddisfano la nostra condizione.

Osservazione 1.11. Riguardiamo ancora l'esempio precedente: potremmo pensare di procedere in un modo diverso, non necessariamente passando alle cifre vicine. Ad esempio potremmo cominciare scegliendo la prima, la terza e la quinta. Siccome esse non si toccano, possiamo scegliere ciascuna di esse in 10 modi. Quando però andiamo a considerare la seconda cifra del nostro PIN, dobbiamo distinguere due casi, per sapere quante scelte siano possibili: se la prima e la terza cifra sono uguali, allora la seconda può essere scelta in 9 modi. Se invece sono diverse tra loro, la seconda può essere scelta solamente in 8 modi. Questo modo di conteggiare, per quanto corretto e possibile, è quindi più a rischio per quanto riguarda gli errori di conto.

Nell'esempio, infatti, stiamo sfruttando un approccio (un algoritmo) che sfrutta una proprietà particolare: non ci interessa quale cifra estraiamo in un dato punto, perché stiamo considerando qualcosa di invariante rispetto alla scelta specifica della cifra, ossia la cardinalità delle cifre che ci restano da scegliere al passaggio successivo. È per questo che fissare cifre del PIN saltando qua e là non è altrettanto efficace: non abbiamo un invariante analogo.

Dopo questa breve divagazione torniamo alla combinatoria e, per concludere questa sezione, vediamo il *terzo principio della combinatoria*. Per farlo, ci mettiamo in una situazione simile a quella vista per il primo principio: vogliamo ottenere la cardinalità dell'unione di alcuni insiemi, lasciando però cadere l'ipotesi che siano disgiunti.

Esempio 1.12. Alcuni eventi della Coppa del Mondo di arrampicata hanno gare di due diverse specialità: *boulder* e *lead*. Sapendo che a uno di questi hanno partecipato 37 atleti nel *boulder*, 33 nel *lead* e 14 a entrambe le specialità, quanti erano gli atleti presenti all'evento?

In analogia a quanto visto per il primo principio, la prima idea che ci viene è quella di andare a sommare i partecipanti al *boulder* con quelli al *lead*, ottenendo $37+33=70$ atleti. Tuttavia sappiamo che il primo principio richiede che gli insiemi siano disgiunti, mentre qui sappiamo che questa ipotesi non è verificata. Cosa cambia? Pensiamo ai 14 atleti che hanno preso parte a entrambe le gare di specialità: li abbiamo contati due volte, sia nel *boulder*, sia nel *lead*, quindi per avere il numero totale di atleti presenti dobbiamo sottrarre 14 da 70, ottenendo in tutto 56 partecipanti.

In generale, possiamo enunciare il terzo principio come segue.

PROPOSIZIONE 1.13. *Se abbiamo due insiemi A_1 e A_2, la cardinalità della loro unione sarà*

$$\#(A_1 \cup A_2) = \#A_1 + \#A_2 - \#(A_1 \cap A_2).$$

Dimostrazione. Possiamo dimostrare questo risultato riconducendoci al primo principio, scrivendo l'unione come unione disgiunta:

$$A_1 \cup A_2 = (A_1 \setminus A_2) \cup (A_2 \setminus A_1) \cup (A_1 \cap A_2).$$

1.1. La prima cifra che inseriamo. Infatti si può osservare abbastanza facilmente che il ragionamento non cambia se andiamo a scegliere per prima la cifra in una qualunque posizione (ad esempio in posizione 3), muovendoci poi in entrambe le direzioni passando ogni volta a una cifra adiacente a una già scelta.

Ora non ci resta che osservare che $A_1 = (A_1 \setminus A_2) \cup (A_1 \cap A_2)$ (e analogamente per A_2) e mettere assieme i vari pezzi per ottenere quanto cercato. □

Anche qui, come in precedenza, nulla ci costringe a considerare solamente due insiemi. Se passiamo all'unione di tre insiemi A_1, A_2 e A_3, iniziamo come prima, sommando le cardinalità dei tre insiemi e togliendo le (tre) intersezioni degli insiemi a due a due. In questo modo abbiamo contato una sola volta tutti gli elementi, tranne quelli di un sottoinsieme: l'intersezione di A_1, A_2 e A_3. Guardiamo un elemento di questo sottoinsieme: lo abbiamo contato una volta in ciascuno dei tre insiemi, lo abbiamo poi tolto una volta per ciascuna delle tre intersezioni a due a due, col risultato che lo abbiamo contato zero volte. Dobbiamo quindi andare ad aggiungere l'intersezione a tre,

$$\begin{aligned}\#(A_1 \cup A_2 \cup A_3) \;=\; & \#A_1 + \#A_2 + \#A_3 \\ & -\#(A_1 \cap A_2) - \#(A_1 \cap A_3) - \#(A_2 \cap A_3) \\ & +\#(A_1 \cap A_2 \cap A_3).\end{aligned}$$

Come conseguenza di questo aggiungere e togliere elementi, il terzo principio prende anche il nome di *principio di inclusione-esclusione*.

PROPOSIZIONE 1.14. *Con n insiemi $A_1, \ldots, A_n$ abbiamo l'uguaglianza*

$$\#\bigcup_{i=1}^{n} A_i = \sum_{i=1}^{n} \#A_i - \sum_{i<j} \#(A_i \cap A_j) + \sum_{i<j<k} \#(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1} \#\bigcap_{i=1}^{n} A_i.$$

Osservazione 1.15. Osserviamo che se non consideriamo l'intera somma, ma solo i primi addendi, possiamo avere una stima del totale. È una stima dal basso nel caso in cui il primo termine della somma che ignoriamo ha segno positivo (cioè è in posizione dispari), dall'alto se ha segno negativo (ossia è in posizione pari).

1.2. PERMUTAZIONI E ANAGRAMMI

Pensiamo ora alla seguente situazione: abbiamo un insieme A che contiene n oggetti distinti. Ci chiediamo quante siano le permutazioni di questi oggetti, ossia i modi di disporli in fila.

Iniziamo dal primo oggetto della fila: lo possiamo scegliere a piacere tra tutti gli elementi di A, cioè abbiamo n modi per sceglierlo. Passiamo ora al secondo. Anche senza sapere quale elemento di A abbiamo messo al primo posto, sappiamo che ce ne sono rimasti altri $n-1$ tra cui scegliere: tutti gli elementi di A, tranne quello già usato. Possiamo continuare in questo modo: a ogni passo avanti nella fila di oggetti, avremo un elemento in meno tra cui scegliere, fino ad arrivare all'ultimo posto, per il quale non ci sarà rimasto che un solo elemento.

Scrivendo tutto questo abbiamo che le *permutazioni*, o *riordinamenti*, di A sono $n \cdot (n-1) \cdot (n-2) \cdot \cdots \cdot 3 \cdot 2 \cdot 1$, cioè il prodotto di tutti i numeri interi positivi minori o uguali di n. Questo prodotto è talmente importante in matematica che viene denotato con un simbolo, $n!$, detto n fattoriale. Per il caso limite $n = 0$, poniamo $0! = 1$, con l'idea che abbiamo un solo modo per ordinare l'insieme vuoto.

Come nel prodotto cartesiano, anche qui l'ordine è importante. E in un certo senso siamo ancora nel caso del prodotto cartesiano: semplicemente partiamo con l'insieme al completo e, a ogni passo, lo moltiplichiamo (cartesianamente) con una versione sempre più piccola, che ha perso l'elemento appena scelto. Anche se non sappiamo con precisione quale sia l'elemento che abbiamo scelto, a ogni passo ce ne sarà rimasto uno in meno rispetto a quelli che avevamo in precedenza.

Tra gli insiemi da riordinare un ruolo speciale è costituito dalle parole, intese come insiemi di lettere. In questo caso indichiamo i riordinamenti col nome *anagrammi*. Attenzione, vogliamo contare tutti gli anagrammi, non stiamo chiedendo che abbiano senso in qualche lingua.

Esempio 1.16. Prendiamo ora una parola, ad esempio ``PRENDIAMO'': quanti sono i suoi anagrammi?

Siamo nello stesso caso visto sopra: il nostro insieme A è ora

$$A = \{\mathrm{P}, \mathrm{R}, \mathrm{E}, \mathrm{N}, \mathrm{D}, \mathrm{I}, \mathrm{A}, \mathrm{M}, \mathrm{O}\}$$

e in particolare ha 9 elementi, tutti distinti tra loro. I loro riordinamenti, cioè gli anagrammi di ``PRENDIAMO'', sono quindi $9! = 362880$.

Prima di continuare con altri esempi di anagrammi, parliamo per un momento del fattoriale. Una delle prime cose che possiamo osservare incontrandolo è quanto velocemente cresce: nell'esempio precedente abbiamo visto che $9! = 362880$, mentre $10!$ è 10 volte più grande. Insomma, diventa rapidamente complicato scriverlo per esteso e conviene lasciarlo indicato con il suo simbolo finché si può. Non solo, nel momento in cui volessimo semplificarlo, ci conviene sfruttare la fattorizzazione naturale nascosta nella sua definizione, cioè la scrittura come prodotto dei primi n interi positivi, per semplificare tutto il semplificabile. Vedremo alcuni esempi di queste semplificazioni più avanti, perché il fattoriale salterà fuori spesso (cosa che rende ancora più comoda la notazione col punto esclamativo).

Consideriamo una variante della situazione precedente, molto comune quando stiamo anagrammando parole: cosa succede se abbiamo delle ripetizioni? Nel caso delle parole: cosa succede se una lettera compare più volte?

Esempio 1.17. Consideriamo la parola ``ANAGRAMMI'': quanti sono i suoi anagrammi?

Sicuramente sono al più $9!$, cioè tutte le permutazioni delle sue lettere. Però questo non tiene conto del fatto che abbiamo alcune lettere che si ripetono: A compare tre volte, M due. Se contassimo solamente le permutazioni, come fatto prima, staremmo contando come distinti due anagrammi ottenuti scambiando tra loro due lettere uguali (ad esempio le due M). Tuttavia questi sono indistinguibili tra loro:

$$\mathrm{ANAGRAM_1M_2I} = \mathrm{ANAGRAM_2M_1I}.$$

Dobbiamo allora contare in quanti modi possiamo permutare tra loro le lettere uguali. In questo esempio possiamo riordinare le M tra loro in $2! = 2$ modi e le A in $3! = 6$ modi. Dividiamo allora il fattoriale della lunghezza della parola per il numero di permutazioni di ciascun gruppo di lettere uguali, cioè per il fattoriale del numero delle loro occorrenze. In questo caso le permutazioni distinte di ``ANAGRAMMI'' sono

$$\frac{9!}{3! \cdot 2!} = \frac{362880}{12} = 30240.$$

Possiamo seguire questo approccio in generale, non solo per insiemi di lettere: se abbiamo un insieme A costituito da n elementi di m tipi diversi (necessariamente deve essere $m \leqslant n$), ciascun tipo $i \in \{1, \dots m\}$ presente in k_i copie, le permutazioni possibili di tutti gli elementi di A, non distinguendo elementi di uno stesso tipo, sono

$$\frac{n!}{k_1! \cdot k_2! \cdot \dots \cdot k_m!}.$$

Esempio 1.18. In una famiglia è consuetudine, per le festività invernali, decorare la ringhiera del balcone. Per farlo, mettono in fila palline luminose di tre colori: 8 sono rosse, 6 sono verdi e 4 sono azzurre. Ogni anno vogliono avere una decorazione diversa da quelle degli anni precedenti: dopo quanti anni dovranno necessariamente ripetersi?

Ci sono

$$\frac{18!}{8! \cdot 6! \cdot 4!} = \frac{9 \cdot 10 \cdot \dots \cdot 17 \cdot 18}{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 2 \cdot 3 \cdot 4} = 11 \cdot 13 \cdot 14 \cdot 15 \cdot 17 \cdot 18 = 9189180$$

possibili anagrammi delle lampadine a loro disposizione, quindi molto probabilmente ci saremo già estinti da un po'. Prima di continuare, notiamo come abbiamo semplificato tutto quello che potevamo, prima di fare il conto conclusivo[1.2].

Esempio 1.19. Quanti sono gli anagrammi di ``ANAGRAMMI'' in cui le due ``M'' sono adiacenti?

Se le due ``M'' devono essere adiacenti, possiamo considerarle come un'unica lettera ``X'' e contare gli anagrammi della parola ``ANAGRAXI''.

A questo punto abbiamo $\frac{8!}{3!} = 6720$ anagrammi possibili, avendo 8 lettere di cui una, la ``A'', ripetuta 3 volte.

Esempio 1.20. Goffredo ha recentemente avuto una delusione in amore, quindi odia tutto quello che gli ricorda il tema. Nel fare gli anagrammi di ``ANAGRAMMI'' esclude tutti quelli in cui compaiono le stringhe ``AMA'' o ``AMI''. Quanti anagrammi gli rimangono?

Ci conviene contare quanti sono in tutto gli anagrammi, quanti sono quelli con una delle stringhe incriminate e sottrarre il secondo numero dal primo. Dobbiamo anche prestare attenzione al fatto che, essendoci più ``A'' e ``M'', potremmo avere più stringhe incriminate in un medesimo anagramma. Ci servirà allora il principio di inclusione-esclusione.

Cominciamo a codificare ``X''=``AMA'' e ``Y''=``AMI''. Contiamo gli anagrammi che contengono ``AMA'': sono i riordinamenti di ``NGRAMIX'', che sono 7!. Ce ne sono però alcuni che stiamo contando due volte: quelli in cui compare la stringa ``AMAMA''. Se la chiamiamo ``W'', per sapere quanti ne abbiamo contati di troppo, ci basta contare gli anagrammi di ``NGRIW'', che sono 5!.

Passiamo ora agli anagrammi di ``ANAGRAMMI'' che contengono ``AMI'': sono quelli di ``NAGRAMY'' cioè $\frac{7!}{2!}$. Anche in questo caso, però, ci sono alcuni anagrammi che contengono sia ``AMA'' sia ``AMI'' e che quindi abbiamo già contato in precedenza. Sono quelli della parola ``NGRXY'', 5!, ma anche quelli della parola ``NGRAZ'', in cui ``Z''=``AMAMI'', anche questi 5!. Quindi gli anagrammi di ``ANAGRAMMI'' che contengono ``AMA'' o ``AMI'' sono

$$7! - 5! + \frac{7!}{2!} - 5! - 5! = 5! \cdot (6 \cdot 7 + 3 \cdot 7 - 3) = 5! \cdot 60.$$

Ora dobbiamo sottrarre questo numero, che ci dice quanti riordinamenti non vanno bene a Goffredo, dal numero di tutte le possibili permutazioni di ``ANAGRAMMI'', che sono $\frac{9!}{3!2!}$. La risposta è quindi

$$\frac{9!}{3!2!} - 5! \cdot 60 = 5!(7 \cdot 4 \cdot 9 - 60) = 120 \cdot 192 = 23040.$$

Torniamo ora al caso in cui tutti gli n elementi del nostro insieme sono distinti tra loro. Questa volta, però, vogliamo contare quante sono le possibili disposizioni di un numero $k \leqslant n$ di suoi elementi.

Esempio 1.21. Dodici amici hanno organizzato tra loro una lotteria, per la quale hanno 5 premi di valore decrescente. Quanti sono i diversi modi di distribuire i premi?

In un certo senso ci stiamo chiedendo nuovamente quanti siano i modi di mettere in fila i 12 amici (al primo della fila daremo il primo premio e così via), con la differenza che non ci interessa davvero sapere come sono disposti dalla sesta posizione in poi, perché uno scambio tra due persone oltre la quinta posizione non ha influenza sulla distribuzione dei premi. Abbiamo quindi, in questo caso, 12 scelte per il vincitore del primo premio, 11 per il secondo, fino a 8 scelte per il vincitore del quinto premio. La risposta è quindi $12 \cdot 11 \cdot 10 \cdot 9 \cdot 8 = 95040$.

Possiamo però osservare che questo è il prodotto dei numeri consecutivi da 8 a 12, una quantità che sappiamo esprimere come un rapporto di fattoriali:

$$\frac{12!}{7!} = 8 \cdot 9 \cdot 10 \cdot 11 \cdot 12.$$

1.2. Non dobbiamo pensare che siano semplificazioni inutili, anche quando abbiamo a portata di mano una calcolatrice o un computer: i fattoriali crescono talmente in fretta che, anche se il risultato finale è alla loro portata, gli strumenti di calcolo possono dare errori di approssimazione prima di arrivare in fondo.

Questo ci suggerisce un altro modo di vedere lo stesso risultato: le ultime 7 posizioni sono uguali tra loro, nel senso che sono tutte non vincenti, quindi stiamo contando le permutazioni di un insieme con 5 elementi tutti distinti tra loro e altri 7 tutti dello stesso tipo. Possiamo vederlo come l'insieme dei premi: primo, secondo, …, quinto, niente, niente, …, niente.

In casi come questo si parla di *permutazioni incomplete* o *k-permutazioni*. Il numero di permutazioni di k elementi in un insieme di n elementi distinti è $\frac{n!}{(n-k)!}$.

Esempio 1.22. A ogni Gran Premio di Formula E partecipano 24 piloti. Al termine della gara vengono assegnati punti (diversi per ciascun piazzamento) ai primi 10 piloti. In quanti modi diversi è possibile assegnare i punteggi?

Abbiamo in tutto 24! riordinamenti possibili dei piloti. Ai fini della classifica, però, contano solamente le prime 10 posizioni, quindi non sono diversi tra loro quei riordinamenti che differiscono solo per permutazioni delle ultime 14 posizioni. Quindi la risposta è $\frac{24!}{14!} = 7117005772800$.

Esempio 1.23. Nelle gare di Coppa del Mondo di arrampicata, vengono assegnati punti ai primi 30 classificati. Uomini e donne gareggiano in competizioni separate. Se alla gara di Garmisch-Partenkirchen hanno preso parte 50 uomini e 48 donne, quanti sono i modi diversi di assegnare i punteggi?

Cominciamo considerando separatamente la classifica maschile e quella femminile. Come visto nell'Esempio 1.22, abbiamo $\frac{50!}{(50-30)!}$ modi di assegnare punti nella gara maschile e $\frac{48!}{(48-30)!}$ modi per la gara femminile.

Dobbiamo ora combinare questi risultati, per avere il numero delle possibili classifiche dell'intero evento. Per il secondo principio della combinatoria, siccome i due ambiti sono distinti, dobbiamo moltiplicare i due risultati parziali, per avere quello totale: $\frac{50!}{20!} \cdot \frac{48!}{18!}$. Questo numero si può semplificare un po', ma è dell'ordine di 10^{91}. Sconsiglio di provare a calcolarlo.

Resta per il momento in sospeso il caso delle k-permutazioni con ripetizioni. Le idee non sono molto diverse da quelle viste finora, ma diventano più semplici se viste sotto una lente diversa, quella delle combinazioni, che vedremo ora.

1.3. COMBINAZIONI E COEFFICIENTE BINOMIALE

Passiamo a un problema diverso, anche se l'ambientazione è analoga a quella dell'Esempio 1.22.

Esempio 1.24. Le qualifiche di Formula E per stabilire l'ordine di partenza in un Gran Premio sono divise in due fasi: nella prima concorrono tutti i partecipanti, dopodiché i 6 più veloci nella prima fase competono tra loro nella Super Pole per determinare le prime 6 posizioni. In quanti modi diversi possiamo scegliere i 6 piloti (tra i 24 totali) che parteciperanno alla Super Pole?

Osserviamo che non siamo nella situazione già vista delle permutazioni incomplete, perché non ci interessa in che ordine siano i primi 6: tutti i risultati delle qualifiche (cioè tutti gli ordinamenti dei 24 piloti) che differiscono tra loro per riordinamenti dei primi 6 o degli ultimi 18 sono equivalenti. Quindi possiamo prendere tutti gli ordinamenti, dividerli per i riordinamenti degli ultimi 18, ottenendo le permutazioni incomplete viste prima, e dividere ancora una volta per i riarrangiamenti dei primi 6: abbiamo allora $\frac{24!}{18! \cdot 6!}$.

Quello che abbiamo fatto in questo esempio è semplicemente contare i modi di scegliere 6 piloti tra 24. Possiamo generalizzarlo a n e k qualunque tra i numeri naturali, contando i modi di scegliere k oggetti tra n disponibili (ovviamente ci aspettiamo di farlo per $0 \leqslant k \leqslant n$) o, equivalentemente, di dividere gli n elementi di un insieme in k sottoinsiemi: essi prendono il nome di *combinazioni* di k oggetti scelti tra n. Per quanto appena detto, tali combinazioni saranno $\frac{n!}{k!(n-k)!}$, quantità per cui introduciamo la notazione $\binom{n}{k}$, detta *coefficiente binomiale*.

Esempio 1.25. Un professore prepara 13 problemi per un esame orale, in modo da poterne assegnare uno diverso a ciascun partecipante. All'esame, però, si presentano solo in 3. In quanti modi può scegliere 3 problemi da assegnare ai presenti?

Il professore deve scegliere 3 problemi tra i 13 che ha. Può farlo in $\binom{13}{3} = 286$ modi.

Fino a qui può sembrare che il coefficiente binomiale sia solo una comoda scrittura. Ma oltre a essere comodo è anche importante, perché tende a saltare fuori molto spesso nei problemi di combinatoria, anche più difficili di quelli appena visti. Prima di passare ad altri esempi più interessanti, tuttavia, vediamo alcune proprietà del coefficiente binomiale.

PROPOSIZIONE 1.26. *Siano k e n numeri naturali tali che $0 \leqslant k \leqslant n$. Valgono le seguenti proprietà:*

1. $\binom{n}{k} = \binom{n}{n-k}$
2. $\binom{n}{0} = \binom{n}{n} = 1$
3. $\sum_{k=0}^{n} \binom{n}{k} = 2^n$
4. $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$.

Dimostrazione. Lasciata come Problema 1.4. □

Anche nel caso del coefficiente binomiale, come per il fattoriale e per la combinatoria in generale, non possiamo andare a fondo e studiare tutte le sue proprietà. Accenniamo solamente alla rappresentazione dei coefficienti binomiali in forma grafica, con il triangolo di Tartaglia[1.3] (o di Pascal[1.4]), del quale si può scoprire di più cercando online o consultando altri libri dedicati alla combinatoria.

Ci sono però problemi in cui il coefficiente binomiale entra in gioco in maniera non ovvia, come possiamo vedere nel prossimo esempio.

Esempio 1.27. Sul Lungarno a Pisa ci sono 18 palazzi, l'uno accanto all'altro. Il nuovo sindaco vuole ridipingerli in modo che siano soddisfatte le seguenti condizioni:

1. devono essere usati tutti e 7 i colori dell'arcobaleno;
2. tutti i palazzi del medesimo colore devono essere adiacenti.

In quanti modi diversi può farlo?

Cominciamo subito spezzando il problema in due parti: siccome tutti i palazzi del medesimo colore sono adiacenti, possiamo separare la scelta dell'ordine dei colori e i modi di colorare i palazzi una volta fissato l'ordine dei colori. In particolare nella soluzione comparirà un fattore 7! a contare i possibili riordinamenti dei colori.

Supponiamo ora fissato l'ordine dei colori. La seconda parte del problema è scegliere in quanti modi possiamo raggruppare i 18 palazzi in 7 sottoinsiemi, tenendo conto dei vincoli. Sentiamo puzza di coefficiente binomiale, ma non possiamo usarlo direttamente. Proviamo allora a cambiare punto di vista: mettiamoci sul Lungarno anche noi e guardiamo i palazzi che abbiamo accanto. Cominciamo a camminare: prima ne abbiamo un po' di un colore, poi passano al secondo, al terzo e così via, fino al passaggio dal sesto al settimo colore. Ehi! Abbiamo 6 cambi di colore, per via delle due condizioni. Quanti sono i posti in cui possiamo avere questi cambi di colore? Sono possibili a ogni confine tra due palazzi, quindi ne abbiamo uno dopo il primo palazzo, uno dopo il secondo e così via fino all'ultimo confine, dopo il penultimo (diciassettesimo) palazzo e prima dell'ultimo (diciottesimo). Quindi dobbiamo piazzare 6 cambi di colore in 17 posti possibili, per un contributo di $\binom{17}{6}$. In generale, con p palazzi e c colori avremmo $\binom{p-1}{c-1}$ possibilità.

1.3. Niccolò Fontana detto Tartaglia (1499 circa – 1557).

1.4. Blaise Pascal (1623 – 1662).

Mettendo assieme i due pezzi del problema, abbiamo allora $7! \cdot \binom{17}{6}$.

1.4. UN PO' DI PROBABILITÀ

Abbiamo detto che ci interessavamo ai conteggi e alla combinatoria per poter parlare di probabilità. Vediamo allora qualche esempio in cui abbiamo casi equiprobabili per cui possiamo usare come definizione di probabilità il rapporto tra il numero di casi favorevoli e quello di casi totali.

Esempio 1.28. Se nel Dipartimento di Informatica ci sono 22 docenti che possono essere in commissione di laurea e una commissione di laurea è costituita da 5 docenti, con che probabilità la prossima commissione sarà composta dai prof. Bianchi, Ronchetti, Ghiloni, Montresor e Kupfer?

C'è una sola commissione con quei 5 professori, quindi il numero di casi favorevoli è uguale a 1. Quante sono invece le possibili commissioni? Sono $\binom{22}{5} = \frac{22!}{17!5!} = 26334$. La probabilità di avere proprio quella commissione, allora è $\frac{1}{26334} \approx 0.00004$.

Esempio 1.29. Giocando al Superenalotto con una scheda normale, cioè scegliendo 6 dei 90 numeri possibili, qual è la probabilità dei seguenti risultati?

i. Fare 6.
ii. Fare esattamente 5.
iii. Fare almeno 5.
iv. Fare esattamente 3.

Quante sono le possibili sestine? Sono

$$\binom{90}{6} = \frac{90!}{6!84!} = \frac{85 \cdot 86 \cdot 87 \cdot 88 \cdot 89 \cdot 90}{2 \cdot 3 \cdot 4 \cdot 5 \cdot 6} \approx 6 \cdot 10^8.$$

A questo punto dobbiamo solamente contare i casi favorevoli. Nel primo caso abbiamo un solo caso favorevole. Nel secondo ne abbiamo $\binom{6}{5}\binom{84}{1} = 6 \cdot 84$ e la probabilità cercata è quindi dell'ordine di 10^{-6}. Il terzo caso è dato dalla somma dei primi due casi, perché almeno 5 significa o esattamente 5 o esattamente 6. I modi di fare esattamente 3 sono $\binom{6}{3}\binom{84}{3} = 1905680$ e la probabilità associata è circa lo 0.3%.

Esempio 1.30. Qual è la probabilità che in un'aula con 70 studenti almeno 2 abbiano lo stesso compleanno?

Calcoliamo la probabilità del caso opposto (o *complementare*), ossia che abbiano tutti compleanni in giorni diversi. Dalla definizione usata finora di probabilità abbiamo infatti che la probabilità del complementare è il numero di casi non favorevoli diviso il numero di casi totali, quindi è uguale a 1 meno il numero di casi favorevoli diviso il numero di casi totali, dal momento che il numero di casi totali è la somma dei casi favorevoli e di quelli non favorevoli. Calcolando una delle due probabilità, possiamo ottenere l'altra.

Andiamo in ordine nell'aula e guardiamo la stringa di 70 giorni di compleanno. Se vogliamo che siano tutte diverse tra loro, abbiamo $\frac{365!}{(365-70)!}$ modi di sceglierle. Tutte le stringhe possibili (con ripetizioni) sono 365^{70}. Facendo il rapporto abbiamo che la probabilità che non ci siano due persone con il medesimo compleanno è circa 0.0008, ossia la probabilità che almeno due abbiano lo stesso compleanno è maggiore di 99.9%.

1.5. L'IMPORTANZA DELLA CLAUSOLA "EQUIPROBABILI"

Dobbiamo però controllare che le ipotesi di equiprobabilità siano verificate, altrimenti rischiamo di sbagliare, anche grossolanamente. Il classico controesempio alla formuletta mnemonica "casi favorevoli su casi totali" è quello della lotteria: ci sono due casi possibili, vincere e non vincere, di cui uno solo è a noi favorevole, quindi la probabilità di vittoria è $\frac{1}{2}$.

Ce ne sono però anche di più subdoli, in cui il risultato con un'interpretazione errata non è così lontano da quello corretto. In questi casi non possiamo sfruttare l'implausibilità della probabilità che otteniamo per accorgerci di aver sbagliato.

Esempio 1.31. Lanciamo due normali dadi a 6 facce. Qual è la probabilità di ottenere almeno un 4?

Quanti sono i possibili risultati, visti come coppie non ordinate? Ne abbiamo sei in cui compare almeno un 1, cinque in cui compare almeno un 2 e non compaiono 1 (abbiamo già contato $\{1,2\}$) e così via. Le possibili coppie non ordinate sono $\frac{6\cdot 7}{2}=21$. Quelle in cui compare almeno un 4 sono sei. Quindi potremmo dire che la probabilità di vedere almeno un 4 sia $\frac{6}{21}=\frac{2}{7}\approx 29\%$.

Come però si può notare, queste coppie non ordinate non sono tra loro equiprobabili. E infatti se andiamo a contare le coppie ordinate, in cui il primo elemento rappresenta il risultato del primo dado e il secondo elemento quello del secondo dado, abbiamo 36 casi possibili, di cui 11 favorevoli, per una probabilità di vedere almeno un 4 uguale a $\frac{11}{36}\approx 31\%$.

Come avevamo detto prima, considerare le coppie come ordinate oppure no cambia le carte in tavola. Alle volte è l'ambientazione del problema a complicare le cose, ad esempio quando ci presenta (come singoli) oggetti che siamo abituati a considerare a coppie.

Esempio 1.32. In una scarpiera, Andrea ha n paia di scarpe. Se prende a caso un numero pari di scarpe inferiore alla metà del totale, con che probabilità non avrà un paio completo?

Dobbiamo fare attenzione a non confondere scarpe e paia. Nella scarpiera ci sono $2n$ scarpe e Andrea ne prende $2s$, con $2s<n$. In quanti modi può farlo? È un coefficiente binomiale: può scegliere le scarpe in $\binom{2n}{2s}$ modi.

Passiamo allora al secondo conteggio, quello dei casi favorevoli[1.5]. Cosa vuol dire che non ha alcun paio completo? Significa che ha scelto al più una scarpa per ogni paio disponibile e, in particolare, ha scelto $2s$ *tipi* di scarpa (cioè tipi di paia) tra gli n disponibili e per ciascuno di essi (cioè per $2s$ volte) ha scelto una delle due scarpe. In altre parole lo può fare in $\binom{n}{2s}\cdot\binom{2}{1}^{2s}=\binom{n}{2s}\cdot 2^{2s}$ modi diversi. La probabilità cercata è allora

$$\frac{\binom{n}{2s}\cdot 2^{2s}}{\binom{2n}{2s}}=\frac{n!}{(2n)!}\cdot\frac{(2n-2s)!}{(n-2s)!}\cdot 2^{2s}.$$

Se questo ragionamento non ci convince del tutto, magari perché ci confondiamo nel passare da scarpe a paia, possiamo provare a calcolare il tutto in altri modi.

Supponiamo che le scarpe siano tutte in fila e che quelle scelte da Andrea siano le prime $2s$. I casi totali sono allora $(2n)!$.

Passiamo allora ai casi favorevoli. Possiamo scegliere la prima scarpa come vogliamo, quindi abbiamo $2n$ modi di farlo. Per la seconda, non volendo avere paia complete, abbiamo $2n-2$ scelte, per la terza $2n-4$ e così via, fino alla scarpa in posizione $2s$ che possiamo scegliere in $2n-2\cdot(2s-1)=2n-4s+2$ modi. A questo punto tutti i possibili ordinamenti delle successive $2n-2s$ scarpe ci vanno bene, quindi abbiamo un fattore $(2n-2s)!$.

I casi favorevoli sono in tutto $2n\cdot(2n-2)\cdot(2n-4)\cdot\dots\cdot[2n-2\cdot(2s-1)]\cdot(2n-2s)!$ e la probabilità cercata è

$$\frac{2n\cdot(2n-2)\cdot\dots\cdot[2n-2\cdot(2s-1)]\cdot(2n-2s)!}{(2n)!}=2^{2s}\cdot n\cdot(n-1)\cdot\dots\cdot[n-(2s-1)]\cdot\frac{(2n-2s)!}{(2n)!},$$

cioè lo stesso risultato ottenuto prima (per fortuna).

E se volessimo ragionare per probabilità sulle singole scarpe, potremmo osservare che la prima ci va bene in $2n$ casi su $2n$, cioè con probabilità $\frac{2n}{2n}=1$, la seconda con probabilità $\frac{2n-2}{2n-1}$, e così via fino alla scarpa numero $2s$ che ci va bene con probabilità $\frac{2n-4s+2}{2n-2s+1}$. Mettendo il tutto assieme,

$$\frac{2n}{2n}\cdot\frac{2n-2}{2n-1}\cdot\dots\cdot\frac{2n-2\cdot(2s-1)}{2n-(2s-1)}=2^{2s}\cdot\frac{n!}{(n-2s)!}\cdot\frac{(2n-2s)!}{(2n)!}.$$

1.5. In realtà per Andrea non sono molto favorevoli.

Come riscaldamento, possiamo fermarci qui: con un po' di creatività e di attenzione, sono tantissimi i problemi di probabilità che si possono scrivere in termini di casi equiprobabili. Ma come detto, la definizione data non ci soddisfa del tutto. Non solo, ci restringe a casi in cui possiamo contare cose, quindi in numero finito. E ci obbliga a prestare attenzione al fatto che tutti i casi sono equiprobabili (come possiamo ad esempio considerare una moneta truccata?). Non è impossibile, ma come vedremo più avanti, con poca fatica e un po' di astrazione in più, riusciremo affrontare problemi e situazioni molto più generali.

1.6. PROBLEMI

Problema 1.1. Quante sono le partizioni di un insieme di cardinalità $n=10$?

Soluzione. Indichiamo con B_n il numero di partizioni distinte di un insieme di cardinalità n. Sappiamo che $B_0=1$, perché l'insieme vuoto ha una sola partizione possibile. Se passiamo al caso $n=1$, abbiamo nuovamente $B_1=1$. Possiamo provare a continuare ancora per qualche passo, ma è meglio provare a caratterizzare i B_n ricorsivamente.

Supponiamo di avere un insieme E di n elementi numerati da 1 a n e di andare a prendere, in una sua partizione, l'insieme $S\subseteq E$ a cui appartiene l'elemento n. A questo punto ci rimangono un certo numero di insiemi nella partizione che, tutti assieme, hanno un numero k di elementi (di E), con $0\leqslant k\leqslant n-1$. Per ciascuno di questi valori di k, possiamo scegliere questi k elementi in $\binom{n-1}{k}$ modi, dal momento che abbiamo già usato l'elemento n, e per ciascuna scelta abbiamo B_k partizioni possibili dei k elementi rimasti.

Abbiamo allora che

$$B_n=\sum_{k=0}^{n-1}\binom{n-1}{k}B_k,\qquad n\geqslant 1,$$

i cui primi valori sono $B_0=B_1=1$, $B_2=2$, $B_3=5$, $B_4=15$, $B_5=52$, $B_6=203$, $B_7=877$, $B_8=4140$, $B_9=21147$, $B_{10}=115975$.

Problema 1.2. Quante sono le funzioni iniettive da un insieme A a un insieme B, entrambi di cardinalità finita?

Soluzione. Innanzitutto se vogliamo che esista una funzione iniettiva da A a B è necessario che B abbia almeno tanti elementi quanti A, cioè $\#A\leqslant\#B$. A questo punto osserviamo che possiamo scegliere l'immagine del primo elemento di A in $\#B$ modi, quella del secondo in $\#B-1$ modi e così via, fino all'immagine dell'ultimo elemento di A, che potrà essere scelta in $\#B-(\#A-1)$ modi. Quindi abbiamo

$$\#B\cdot(\#B-1)\cdot\dots\cdot(\#B-(\#A-1))=\frac{(\#B)!}{(\#B-\#A)!}.$$

Possiamo vedere questo risultato anche in un altro modo: ci interessano i riordinamenti degli elementi di B, trascurando però tutti gli elementi oltre quello in posizione $\#A$.

Problema 1.3. Quante sono le funzioni suriettive da un insieme A a un insieme B, entrambi di cardinalità finita?

Soluzione. Siccome stiamo chiedendo che le funzioni siano suriettive, devono esserci in A almeno tanti elementi quanti ce ne sono in B, cioè $\#A\geqslant\#B$. Sappiamo che le funzioni da A a B sono $\#B^{\#A}$, quindi questo impone un limite superiore al numero di funzioni suriettive: tra tutte le funzioni, infatti, ci sono ad esempio quelle che hanno nell'immagine tutto B tranne un unico elemento. Vogliamo quindi toglierle dal conteggio di tutte le funzioni. Quante sono le funzioni che escludono un solo elemento? Possiamo scegliere l'elemento escluso in $\#B$ modi e, per ciascuna scelta, ci sono $(\#B-1)^{\#A}$ funzioni, quindi dobbiamo sottrarre $\#B\cdot(\#B-1)^{\#A}$ a $\#B^{\#A}$.

A questo punto dobbiamo considerare le funzioni che escludono due elementi di B dall'immagine, infatti le abbiamo sottratte due volte, una volta per ciascuno dei due elementi. Insomma, dobbiamo continuare con il principio di inclusione-esclusione.

Vediamo esplicitamente quante ne dobbiamo aggiungere: dobbiamo contare in quanti modi possiamo scegliere due elementi tra $\#B$, cioè $\binom{\#B}{2}$ e, per ciascuno di questi modi, quante sono le funzioni da A all'insieme B tranne questi due elementi, cioè $(\#B-2)^{\#A}$. Ricordiamo anche che queste vanno aggiunte, perché sottratte due volte, poi sarà il turno di quelle con tre elementi esclusi, che sono state sottratte tre volte, ma anche ri-aggiunte tre volte e devono quindi essere sottratte di nuovo.

Mettendo tutto assieme abbiamo l'uguaglianza

$$\#B^{\#A}-\#B\cdot(\#B-1)^{\#A}+\binom{\#B}{2}\cdot(\#B-2)^{\#A}-\dots=\sum_{i=0}^{\#B}(-1)^i\cdot\binom{\#B}{i}\cdot(\#B-i)^{\#A}.$$

Problema 1.4. (PROPOSIZIONE 1.26) Siano k e n numeri naturali tali che $0\leqslant k\leqslant n$. Verificare le seguenti proprietà:

1. $\binom{n}{k}=\binom{n}{n-k}$
2. $\binom{n}{0}=\binom{n}{n}=1$

3. $\sum_{k=0}^{n} \binom{n}{k} = 2^n$
4. $\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}$.

Soluzione. Vediamole in ordine.

1. In questo caso ci basta scrivere la definizione di coefficiente binomiale e sfruttare la commutatività del prodotto,
$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} = \frac{n!}{(n-k)! \cdot [n-(n-k)]!} = \binom{n}{n-k}.$$
Oltre a svolgere il conto, avremmo anche potuto osservare che selezionare k elementi tra n è equivalente a scegliere $n-k$ elementi da scartare tra n.
2. Questo è il caso limite del precedente, ma merita qualche parola in più: in quanti modi possiamo scegliere n oggetti tra n disponibili? Solamente in 1 modo. Viceversa, potremmo discutere sul fatto che ci sia solo un modo di scegliere 0 oggetti tra n (non scegliere alcun oggetto), ma è quello che esce sostituendo 0 a k nella definizione data, poiché $0! = 1$, ed è anche consistente con la proprietà vista sopra.
3. Questa proprietà è molto interessante: osservando il triangolo di Tartaglia, notiamo che la somma sulla riga n-esima è uguale a 2^n, cioè che sommando tutti i coefficienti binomiali che hanno n nella posizione superiore otteniamo 2^n. Per capire come mai, ci appoggiamo al binomio di Newton, ossia allo sviluppo di un binomio elevato a potenza n. Sappiamo che
$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} \cdot a^k \cdot b^{(n-k)}.$$
Possiamo vederlo in modo combinatorio come segue: quando andiamo a fare l'elevamento a potenza, stiamo svolgendo il prodotto tra n fattori, ciascuno dei quali è una copia della somma $a+b$. Da ogni copia possiamo prendere una a o una b. In quanti modi possiamo avere k fattori a e $n-k$ fattori b?

   Dobbiamo solamente scegliere in quali k delle n copie di $a+b$ peschiamo le a, cosa che possiamo fare in $\binom{n}{k}$ modi. Tornando al quesito iniziale, possiamo a questo punto prendere $a = b = 1$ e abbiamo nel secondo membro la somma cercata, con il primo membro che diventa uguale a 2^n.
4. Questa è la proprietà alla base della costruzione del triangolo di Tartaglia. Immaginiamo di avere $n+1$ oggetti in fila e di sceglierne $k+1$ tra di essi.

   Consideriamo due casi possibili (disgiunti): tutti quelli in cui prendiamo l'ultimo oggetto e tutti quelli in cui non lo prendiamo. Essi sono, rispettivamente, $\binom{n}{k}$, perché dobbiamo scegliere altri k oggetti tra gli n rimanenti, e $\binom{n}{k+1}$, perché avendo escluso l'ultimo oggetto, dobbiamo scegliere tutti i $k+1$ oggetti tra i rimanenti n.

   Se questa dimostrazione non ci piace, possiamo sempre usare le definizioni:
$$\begin{aligned}
\binom{n}{k} + \binom{n}{k+1} &= \frac{n!}{k!(n-k)!} + \frac{n!}{(k+1)!(n-k+1)!} \\
&= \frac{n!(k+1+n-k)}{(k+1)!(n-k)!} \\
&= \frac{n!(n+1)}{(k+1)![(n+1)-(k+1)]!} \\
&= \binom{n+1}{k+1}.
\end{aligned}$$

CAPITOLO 2

UNA NUOVA PROBABILITÀ

Esempio 2.1. Abbiamo un sacchetto che contiene 60 monete, tutte delle stesse dimensioni. Di queste, 20 sono d'ottone e hanno una densità di 8.5 $g \cdot cm^{-3}$, 20 sono d'acciaio, con densità 7.8 $g \cdot cm^{-3}$ e 20 sono d'oro, con densità 19.2 $g \cdot cm^{-3}$. Qual è la probabilità di estrarre una moneta d'oro?

Possiamo fare una domanda su questa domanda: cosa significa "qual è la probabilità di estrarre una moneta d'oro?" Ci sono diversi modi di estrarre una moneta dal sacchetto:

i. posso pescare senza guardare né soppesare;

ii. posso rovesciare il sacchetto e prendere la moneta che cade per ultima e rimane più in alto nella pila delle monete rovesciate;

iii. posso soppesare tutte le monete e controllarne il colore, prima di scegliere una che ritengo essere d'oro;

iv. posso estrarre un certo numero di monete e fermarmi quando penso di averne una d'oro in mano.

Cosa possiamo dire, a livello ancora intuitivo, sulla probabilità che la moneta estratta in questi modi sia effettivamente d'oro? Vediamolo caso per caso.

i. Questa situazione ci ricorda quanto (forse) visto a scuola: abbiamo 20 casi favorevoli, le 20 monete d'oro, e 60 casi totali. La probabilità di estrarre una moneta d'oro è quindi $1/3$. Osserviamo che stiamo dicendo che la probabilità di estrarre una particolare moneta delle 60 nel sacchetto è uguale a $1/60$.

ii. Lasciando cadere le monete, possiamo aspettarci che le differenze fisiche (in particolare la diversa densità e di conseguenza la diversa massa) influiscano sull'ordine di caduta. Non è però detto che siamo in grado di descrivere matematicamente (cioè di modellizzare) con precisione come ciò avviene. Sia che lo sappiamo, sia che non lo sappiamo fare avremo dei margini di incertezza (diversi nei due casi). La nostra miglior stima sarà la probabilità.

iii. In questo caso potremmo pensare di avere la certezza di prendere una moneta d'oro. Tuttavia, anche se ne abbiamo la certezza pratica, non possiamo escludere un piccolo margine d'errore. La probabilità sarà quindi $1 - \varepsilon$, con ε positivo e tanto più piccolo quanto meno riteniamo plausibile un errore.

iv. Rispetto al caso precedente abbiamo molta più incertezza: non stiamo più confrontando tutte le monete. Infatti se, dopo aver considerato una moneta, scegliamo di proseguire con una nuova estrazione, la moneta sarà persa per sempre, non potremo più sceglierla. La probabilità, dunque, dipenderà da chi estrae e dalla strategia decisionale che sceglie. Come possiamo descrivere questa situazione? Quale può essere la probabilità?

Qual è la morale di questo esempio? Abbiamo bisogno di definire in modo più chiaro la situazione che consideriamo. Abbiamo visto che alcuni esperimenti, come quello di questo esempio, hanno risultati che possiamo prevedere solo in parte. Incertezza e previsione sono i punti di partenza per parlare di probabilità.

DEFINIZIONE 2.2. *Un esperimento si dice* aleatorio *o* casuale *se, coi dati iniziali a disposizione, il suo risultato è incerto. In altre parole, se non possiamo prevederne con certezza l'esito.*

Possiamo osservare che abbiamo preso una definizione abbastanza ampia di esperimento aleatorio. L'incertezza, infatti, può essere nei dai iniziali, nella "legge" che governa il fenomeno o nella nostra comprensione. Una conseguenza di questo è che un esperimento quale ad esempio il lancio di una moneta può dare origine a esperimenti aleatori (intesi come oggetti matematici) distinti: cambiare lo sperimentatore, il tempo o lo spazio può portare a livelli di incertezza diversi. Quando dichiareremo un esperimento aleatorio, sarà importante essere il più precisi possibile sulle sue caratteristiche rilevanti. Vedremo più avanti che sottovalutare questo aspetto può avere conseguenze significative.

È curioso il fatto che la probabilità ha preso il ruolo di linguaggio della scienza nel momento in cui l'ideale del determinismo è andato in pezzi con la teoria dei quanti. Già prima, con la meccanica statistica, la probabilità aveva mostrato di poter descrivere e predire fenomeni complessi e in particolare di poter rappresentare la nostra incertezza (o ignoranza) nello studio di un fenomeno. Tuttavia la teoria dei quanti ha mostrato che l'incertezza è intrinseca in certi fenomeni, quindi la probabilità non è più una stampella temporanea in attesa di conoscere il modello deterministico, ma è la descrizione corretta.

Vogliamo descrivere con precisione, in termini matematici, un esperimento aleatorio. Dobbiamo allora dichiarare tutto quello che lo caratterizza. Come prima cosa, ne consideriamo i possibili risultati.

DEFINIZIONE 2.3. *I risultati, a due a due incompatibili, di un esperimento aleatorio prendono il nome di* esiti. *Matematicamente possiamo rappresentarli come elementi di un insieme, detto* spazio campionario, spazio degli esiti, popolazione *o* insieme universo *e denotato con* Ω *o* U. *Questo insieme contiene tutti e soli i possibili risultati dell'esperimento aleatorio.*

Esempio 2.4. Consideriamo un contenitore, detto *urna*, in cui ci sono un certo numero di oggetti, detti *biglie*, indistinguibili al tatto, ma di diverso colore, ad esempio bianco e nero, oppure bianco, rosso e nero. Un esperimento aleatorio può essere l'estrazione di una biglia dall'urna: infiliamo la mano nell'urna senza guardare e prendiamo una biglia. Se l'urna contiene biglie bianche, rosse e nere, gli esiti possibili sono tre: la biglia estratta è bianca, oppure è rossa, oppure è nera.

Un diverso esperimento aleatorio può prevedere l'estrazione di due biglie dalla stessa urna, con reimmissione, ossia rimettendo la biglia pescata per prima nell'urna (dopo averla guardata) e rimescolando le biglie nell'urna, prima di estrarre la seconda biglia. In questo caso lo spazio degli esiti, abbreviando i colori con le loro iniziali, è costituito dalle coppie ordinate

$$\Omega = \{(B,B),(B,R),(B,N),(R,B),(R,R),(R,N),(N,B),(N,R),(N,N)\}.$$

Un terzo esperimento casuale, sempre con la stessa urna, può essere l'estrazione di due biglie senza reimmissione. In questo caso, però, abbiamo bisogno di qualche informazione in più sul numero di biglie nell'urna. Infatti nei casi precedenti bastava sapere che c'erano biglie di tre colori, ossia che c'era almeno una biglia di ciascun colore. In questo caso, invece, se ci fosse, per esempio, una sola biglia bianca, la coppia ordinata (B,B) non sarebbe più un esito, poiché non è un risultato possibile[2.1].

Osservazione 2.5. Capita spesso che, quando ci si avvicina per la prima volta alla probabilità, ci si chieda come mai i problemi e gli esempi siano popolati da urne. Non sembrano essere qualcosa di cui ci interessiamo spesso, nel mondo reale, quindi perché usarli come esempi?

2.1. In realtà la questione può essere più sfumata: infatti potremmo non sapere quante biglie bianche ci sono nell'urna. In questo caso la coppia (B,B) è a priori un risultato possibile. Potremo codificare l'informazione sull'assenza di una seconda biglia bianca nella probabilità, come vedremo più avanti. In generale è fondamentale che l'insieme universo contenga tutti gli esiti, ma abbiamo più flessibilità sul fatto che siano i soli elementi dell'insieme.

La risposta è che le urne, come altri esempi, sono un buon compromesso tra l'astrazione delle caratteristiche cruciali dell'esperimento aleatorio, pur lasciando un'immagine sensoriale che sia di supporto alla rappresentazione mentale. Ciascuno può scegliere tra le rappresentazioni di esperimenti aleatori equivalenti[2.2] quella che genera l'immagine mentale più forte, non necessariamente legata al senso della vista o del tatto, ad esempio contenitori con oggetti indistinguibili al tatto e alla vista, ma con odori diversi.

In un esperimento aleatorio, però, possiamo osservare altre cose, oltre ai risultati specifici finali. Chiamiamo, per ora informalmente, *evento* un'osservabile dell'esperimento aleatorio, ossia un fatto che, al termine dell'esperimento, possiamo dire essere vero o falso, a seconda del risultato dell'esperimento stesso. Nell'estrazione con reimmissione vista nell'Esempio 2.4 un evento è "è stata estratta almeno una biglia bianca". A seconda dell'esito dell'esperimento potremo dire se questo evento è vero, oppure falso.

Sembra ragionevole, allora, pensare a un evento come a un insieme di risultati per cui l'evento è vero. Una possibile rappresentazione di un evento è quindi come insieme di esiti, ossia come sottoinsieme di Ω. Diciamo che un vento si verifica o si realizza se il risultato dell'esperimento aleatorio è (come esito) un elemento dell'evento.

Esempio 2.6. Lanciamo un dado a 6 facce[2.3], alcuni possibili eventi sono:

- esce una faccia con un numero pari, $E_1 = \{2,4,6\}$;
- esce una faccia con un numero minore o uguale a 4, $E_2 = \{1,2,3,4\}$;
- esce una faccia con un numero maggiore di 6, $E_3 = \emptyset$;
- esce una faccia con il numero 3, $E_4 = \{3\}\ldots$

Vedremo che serve in generale qualche accortezza in più nell'identificare gli eventi coi sottoinsiemi dello spazio degli esiti.

Ricordiamo che il numero di elementi di un insieme A si chiama *cardinalità* di A e lo indichiamo con la notazione $\#A$. Finché abbiamo a che fare con insiemi finiti, non ci sono troppi problemi; ma nel momento in cui passiamo a insiemi infiniti, abbiamo bisogno di un po' più di precisione. Diciamo allora che due insiemi A e B hanno la stessa cardinalità, cioè sono *equipotenti*, se esiste una funzione biettiva $f: A \to B$. In particolare un insieme equipotente all'insieme $\mathbb{N}$ dei numeri naturali ha cardinalità (infinita) numerabile e questa quantità è denotata con $\aleph_0$, il primo dei numeri cardinali (cioè usati per indicare le cardinalità) infiniti[2.4].

L'insieme universo Ω in un esperimento aleatorio non ha necessariamente un numero finito di elementi: possono essere in numero finito o anche infiniti, numerabili o più che numerabili. Vogliamo poter considerare situazioni in cui il numero di esiti è infinito. Questo ci creerebbe qualche problema, se volessimo usare la definizione "casi favorevoli su casi totali", perché dovremmo confrontare cardinali infiniti, ma per essi non vale la legge di cancellazione. Per risolvere questo problema abbiamo bisogno di rendere più robusta la nostra teoria.

Prima di proseguire, vediamo alcuni esempi di esperimenti casuali e dei corrispondenti insiemi degli esiti.

- Il lancio di una moneta: in questo caso abbiamo $\Omega = \{\text{testa}, \text{croce}\}$ che è un insieme finito.
- Il numero di tentativi prima di colpire il centro a freccette: qui $\Omega = \mathbb{N}$ e ha cardinalità numerabile. Infatti non è possibile stabilire a priori quanti lanci saranno sufficienti (e dare quindi un limite superiore a Ω).

2.2. L'equivalenza in termini probabilistici di esperimenti aleatori verrà affrontata più avanti, nel Capitolo 5

2.3. In generale ci sono dadi "fisici" a 2, 4, 6, 8, 10, 12, 20 facce: il primo si chiama anche "moneta", quelli a 4, 6, 8, 12, 20 facce sono i solidi platonici, mentre quello a 10 facce è un solido non platonico. Possiamo però considerare anche dadi con altri numeri di facce, ad esempio 3, 30 o 100. Un dado con n facce è spesso indicato come dn.

2.4. Il fatto che $\aleph_0$ sia il primo cardinale infinito suggerisce che ce ne siano degli altri, più grandi. Così è, in effetti, e vedremo un esempio nelle prossime pagine.

- Le possibili lunghezze di un segmento contenuto nell'intervallo reale $[0,1]$: il corrispondente insieme degli esiti è $\Omega = (0,1]$, un intervallo di cardinalità pari al continuo.

2.1. ALGEBRE E TRIBÙ

Considerare tutti i sottoinsiemi di Ω significa considerarne l'insieme potenza (o delle parti), che ha cardinalità $2^{\#\Omega}$, che in particolare è la cardinalità del continuo, se Ω ha cardinalità numerabile e addirittura strettamente maggiore della cardinalità del continuo, se Ω è equipotente all'insieme $\mathbb{R}$ dei numeri reali. Ulteriori dettagli su questi risultati sono in Appendice A.1.

Avendo richiamato i risultati precedenti sulla cardinalità degli insiemi potenza, sorge spontaneo un pensiero: sarebbe bello poter considerare solo una parte dei sottoinsiemi, qualora non ci interessassero proprio tutti. Ad esempio, se stessimo scegliendo un numero tra tutti i naturali, ma ci interessasse solo sapere se il numero è pari o no, ci farebbe comodo poter considerare solo i due sottoinsiemi "numeri pari" e "numeri dispari", invece che tutti i sottoinsiemi di $\mathbb{N}$.

Pensiamo infatti al nostro obiettivo: vogliamo definire una probabilità, ma vorremmo farlo solo su alcuni insiemi, quelli che ci interessano, e non necessariamente su tutti quanti, perché sarebbe un po' uno spreco. Possiamo pensare che definire la probabilità di un evento abbia un costo non trascurabile, dal momento che, come vedremo, dobbiamo scegliere tale probabilità con attenzione in modo che soddisfi certe importanti proprietà. È un prezzo che non vogliamo pagare inutilmente.

Il nostro piano è quindi quello di considerare in generale una famiglia $\mathcal{F}$ di sottoinsiemi, quindi $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, ma non necessariamente tutto $\mathcal{P}(\Omega)$. Ancora una volta, pensiamo al nostro traguardo a lungo termine: definire una probabilità su questi sottoinsiemi. Abbiamo quindi bisogno che questa famiglia sia, in un qualche senso, "stabile".

Per capire meglio cosa intendiamo, pensiamo di nuovo al nostro obiettivo: vogliamo definire una probabilità in modo sensato e vogliamo definirla su questa collezione di insiemi. Vorremmo in particolare che questa collezione contenesse l'insieme Ω e che fosse chiusa rispetto alle operazioni di unione, intersezione e complementare. In altre parole: se due insiemi appartengono alla collezione, vorremmo che ci appartenessero anche la loro unione, la loro intersezione e i loro complementari.

DEFINIZIONE 2.7. *Una famiglia $\mathcal{F}$ di sottoinsiemi di un insieme Ω è un'*algebra *se valgono tutte le seguenti proprietà:*

i. *$\Omega \in \mathcal{F}$;*

ii. *se $A \in \mathcal{F}$, allora anche il suo complementare $A^c \in \mathcal{F}$;*

iii-finita. *se $A, B \in \mathcal{F}$, allora $A \cup B \in \mathcal{F}$.*

Osserviamo che, per come è scritta, la proprietà **iii-finita** della Definizione 2.7 dovrebbe essere chiamata **iii-binaria**. Possiamo però estenderla al caso più generale dell'unione finita: se abbiamo una famiglia finita $(A_i)_{i=1}^n$ di sottoinsiemi di Ω tali che $(A_i)_{i=1}^n \subseteq \mathcal{F}$, allora per la proprietà associativa dell'unione $\bigcup_{i=1}^n A_i \in \mathcal{F}$.

Qualche riga più in alto parlavamo di avere una collezione chiusa anche rispetto all'intersezione. La Definizione 2.7 non menziona esplicitamente l'intersezione e parla solo di unione e complementare. Tuttavia ci garantisce anche che un'algebra sia chiusa rispetto all'intersezione, assieme ad altre proprietà, come mostrato nel seguente risultato.

PROPOSIZIONE 2.8. *Data un'algebra $\mathcal{F}$ su Ω, valgono le seguenti proprietà:*

1. *$\emptyset \in \mathcal{F}$;*
2. *se $A, B \in \mathcal{F}$, allora $A \cap B \in \mathcal{F}$;*

3. *se* $(A_i)_{i=1}^n \subseteq \mathcal{F}$, *allora* $\bigcap_{i=1}^n A_i \in \mathcal{F}$;
4. *se* $A, B \in \mathcal{F}$, *allora* $A \setminus B \in \mathcal{F}$;
5. *se* $A, B \in \mathcal{F}$, *allora* $A \triangle B \in \mathcal{F}$.

Dimostrazione. Procediamo in ordine.

1. Sappiamo che $\Omega \in \mathcal{F}$, per la prima proprietà, e che anche il suo complementare appartiene a $\mathcal{F}$, per la seconda. Ma $\Omega^c = \emptyset$, che quindi appartiene a $\mathcal{F}$.
2. Osserviamo che $A \cap B = (A^c \cup B^c)^c$. Ora, sia A^c sia B^c appartengono a $\mathcal{F}$, dunque anche $A^c \cup B^c$ e il suo complementare.
3. Possiamo iterare il ragionamento visto al punto precedente, sfruttando l'associatività dell'intersezione.
4. Anche qui il trucco è riscrivere l'insieme in una forma più comoda della precedente: $A \setminus B = A \cap B^c$. A questo punto ci basta usare la seconda proprietà di algebra e la chiusura rispetto all'intersezione mostrata sopra.
5. Riscriviamo $A \triangle B = (A \cup B) \cap (A^c \cup B^c)$ (come fatto in dettaglio nella Proposizione A.2) e concludiamo usando le proprietà già mostrate. □

Esempio 2.9. Prendiamo $\Omega = \{0,1,2\}$. Allora $\mathcal{F} = \{\emptyset, \{0\}, \{1,2\}, \Omega\}$ è un'algebra su Ω. In particolare quest'algebra è diversa dall'insieme potenza $\mathcal{P}(\Omega)$.

Questo ci suggerisce in particolare che, dato un insieme Ω (con almeno due elementi), esiste più di un'algebra su di esso. Quante ne possiamo avere? Nel caso di Ω finito, le algebre sono tante quante le partizioni di Ω, cioè $B_{\#\Omega}$, come abbiamo visto nel Problema 1.1.

Esempio 2.10. Prendiamo ora $\Omega = \{a,b,c,d,e,f,g\}$. Le seguenti famiglie di insiemi non sono algebre:

- $\mathcal{F}_1 = \{\emptyset, \{a,b,c,d,e\}, \Omega\}$. Infatti manca il complementare di $\{a,b,c,d,e\}$; il completamento di $\mathcal{F}_1$ a un'algebra è $\{\emptyset, \{a,b,c,d,e\}, \{f,g\}, \Omega\}$.
- $\mathcal{F}_2 = \{\{a\}, \{b,c,d\}, \{e,f,g\}, \Omega\}$, poiché manca un complementare, l'insieme vuoto.
- $\mathcal{F}_3 = \{\emptyset, \{a\}, \{b\}, \{b,c,d,e,f,g\}, \{a,c,d,e,f,g\}, \Omega\}$, siccome mancano i due insiemi $\{a,b\}$, $\{c,d,e,f,g\}$, l'unione di $\{a\}$ e $\{b\}$ e il suo complementare.

Nella Definizione 2.7 la terza proprietà è chiamata "iii-finita" e non "iii": questo potrebbe farci sospettare che esistano famiglie di sottoinsiemi per cui la proprietà è sostituita da una sua variante infinita. Così è, ma è un infinito "controllato".

DEFINIZIONE 2.11. *Una famiglia* $\mathcal{F}$ *di sottoinsiemi di un insieme* Ω *è una* tribù *(o* σ-algebra[2.5]*) se valgono tutte le seguenti proprietà:*

i. $\Omega \in \mathcal{F}$;

ii. *per ogni* $A \subseteq \Omega$, *se* $A \in \mathcal{F}$, *allora* $A^c \in \mathcal{F}$;

iii. *per ogni famiglia numerabile* $(A_i)_{i=1}^{+\infty}$ *di insiemi di* Ω, *se tutti gli insiemi* A_i *della famiglia appartengono a* $\mathcal{F}$, *allora* $\bigcup_{i=1}^{+\infty} A_i \in \mathcal{F}$.

Esempio 2.12. L'insieme delle parti di Ω è esso stesso una tribù. Possiamo osservare, infatti, che soddisfa tutte le proprietà richieste. Dal momento che include tutti i possibili sottoinsiemi di Ω, contiene Ω stesso, il complementare di ogni sottoinsieme di Ω e anche ogni unione numerabile di sottoinsiemi di Ω.

2.5. In realtà questo termine non è del tutto corretto: bisognerebbe parlare di *σ-algebre (o σ-campi) di insiemi*, che sono un particolare caso di σ-algebre booleane. Nella pratica tra i probabilisti il termine σ-algebra è sdoganato.

Rispetto alla definizione di algebra, stiamo chiedendo che anche l'unione numerabile sia un'operazione interna. Osserviamo inoltre che, se $\mathcal{F}$ è una tribù, è in particolare un'algebra, ma il viceversa non è vero in generale. Vale tuttavia il risultato seguente.

PROPOSIZIONE 2.13. *Sia $\mathcal{F}$ un'algebra finita su un insieme Ω. Allora $\mathcal{F}$ è una tribù.*

Dimostrazione. La differenza tra un'algebra e una tribù sta nella proprietà **iii** della Definizione iii: dobbiamo mostrare che ogni unione numerabile di elementi di $\mathcal{F}$ sta in $\mathcal{F}$. Siccome $\mathcal{F}$ è finita, contiene solamente un numero finito di elementi, cioè di sottoinsiemi di Ω. Di conseguenza ogni unione numerabile di elementi di $\mathcal{F}$ sarà in realtà un'unione finita, dal momento che abbiamo solo un numero finito di possibili elementi. Tale unione finita appartiene a $\mathcal{F}$, poiché $\mathcal{F}$ è un'algebra. □

Quindi, finché abbiamo a che fare con insiemi finiti, non abbiamo davvero bisogno di parlare di tribù: ci basta controllare che la nostra famiglia di sottoinsiemi sia un'algebra. Questa proprietà ci dà anche un'interessante condizione necessaria affinché una famiglia finita di sottoinsiemi sia una tribù: deve avere un numero di elementi uguale a una potenza di 2. Lasciamo da parte la dimostrazione (che si può fare per induzione, con qualche accortezza), ma osserviamo che grazie a questa condizione abbiamo un modo rapido per dire che una famiglia di sottoinsiemi non è una tribù. Infatti, se una collezione di insiemi ha cardinalità diversa da una potenza di 2, sappiamo che sicuramente non può essere una tribù.

Proseguiamo con altre proprietà di algebre e tribù.

PROPOSIZIONE 2.14. *Date su Ω due algebre $\mathcal{F}_1$ ed $\mathcal{F}_2$, la loro intersezione $\mathcal{F}_1 \cap \mathcal{F}_2$ è a sua volta un'algebra su Ω. Lo stesso vale se sostituiamo ``algebra'' con ``tribù''.*

Dimostrazione. Dimostriamo questa proposizione per due tribù: in questo modo abbiamo il risultato anche per le algebre.

Dobbiamo far vedere che $\mathcal{F}_1 \cap \mathcal{F}_2$ soddisfa le proprietà di una tribù. Procediamo punto per punto.

i. Siccome $\mathcal{F}_1$ ed $\mathcal{F}_2$ sono tribù, $\Omega \in \mathcal{F}_1$ e $\Omega \in \mathcal{F}_2$, quindi $\Omega \in \mathcal{F}_1 \cap \mathcal{F}_2$.

ii. Sia $E \in \mathcal{F}_1 \cap \mathcal{F}_2$, allora $E \in \mathcal{F}_1$ ed $E \in \mathcal{F}_2$. Siccome $\mathcal{F}_1$ ed $\mathcal{F}_2$ sono due tribù, $E^c \in \mathcal{F}_1$ ed $E^c \in \mathcal{F}_2$, quindi $E^c \in \mathcal{F}_1 \cap \mathcal{F}_2$.

iii. Prendiamo $(E_i)_{i=1}^{+\infty} \subset \mathcal{F}_1 \cap \mathcal{F}_2$. Allora la successione sarà in entrambe le tribù, $(E_i)_{i=1}^{+\infty} \subset \mathcal{F}_1$ ed $(E_i)_{i=1}^{+\infty} \subset \mathcal{F}_2$. Di conseguenza, $\bigcup_{i=1}^{+\infty} E_i \in \mathcal{F}_1$ e $\bigcup_{i=1}^{+\infty} E_i \in \mathcal{F}_2$ e quindi anche $\bigcup_{i=1}^{+\infty} E_i \in \mathcal{F}_1 \cap \mathcal{F}_2$.

Questo conclude la dimostrazione. □

Avendo preso familiarità con le tribù, ripensiamo al motivo per cui le abbiamo definite. Dato un insieme Ω e una tribù $\mathcal{F}$ su di esso, ci interessano gli elementi di $\mathcal{F}$. Andiamo quindi a dar loro un nome.

DEFINIZIONE 2.15. *Sia $\mathcal{F}$ una tribù su Ω. Ogni elemento $E \in \mathcal{F}$ prende il nome di* evento*. I singoletti in $\mathcal{F}$ prendono il nome di* eventi elementari*. Si dice che un evento E si* verifica *se il risultato osservato dell'esperimento casuale è un esito appartenente a E.*

Un evento è un elemento di una famiglia di insiemi, quindi è lui stesso un insieme. Questo può alle volte causare un po' di confusione di terminologia con uno scontro tra elementi e insiemi. È per questo che chiamiamo esiti gli elementi di Ω, eventi gli elementi di $\mathcal{F}$ e universo l'insieme Ω.

Esempio 2.16. Prendiamo un insieme $\Omega = \{a, b, c, d\}$, e su di esso la tribù $\mathcal{F} = \{\emptyset, \{a\}, \{d\}, \{a, d\}, \{b, c\}, \{a, b, c\}, \{b, c, d\}, \Omega\}$. Allora $A = \{a\}$ è un evento, in particolare un evento elementare, ma è anche un sottoinsieme di Ω, quindi un insieme, e un elemento di $\mathcal{F}$. A sua volta $E = \{a, b, c\}$ è un evento, ma non elementare, mentre $N = \{b\}$ non è un evento, poiché non compare in $\mathcal{F}$.

Non sempre, come vedremo, viene assegnata esplicitamente una tribù e non sempre abbiamo una sola scelta possibile, anche quando sappiamo quali eventi vogliamo che siano al suo interno. In questi casi può venir comoda la seguente definizione.

DEFINIZIONE 2.17. *Data una famiglia $\mathcal{G}$ di sottoinsiemi di Ω, definiamo $\sigma(\mathcal{G})$, detta* tribù generata da *$\mathcal{G}$, la più piccola tribù che contiene $\mathcal{G}$, cioè*

$$\sigma(\mathcal{G}) = \bigcap \{\mathcal{F} : \mathcal{F} \text{ è una tribù e } \mathcal{G} \subseteq \mathcal{F}\}.$$

Se vogliamo, questo è un modo per semplificarci la vita: sappiamo quali sono gli eventi che vogliamo avere e generiamo a partire da essi una famiglia che li contenga e che sia anche una tribù. Per farlo possiamo pensare di aggiungere a $\mathcal{G}$ i complementari di insiemi di $\mathcal{G}$, poi unioni numerabili, poi ancora complementari e così via. Prendiamo la più piccola possibile perché non vogliamo doverci occupare di più eventi di quanto non sia strettamente necessario. Il perché di questo essere un po' avari, già menzionato in precedenza, sarà chiaro nella prossima sezione.

Problema 2.1. Consideriamo $\Omega = \mathbb{R}$ ed $\mathcal{F} = \left\{\{0,1\}, \left[\frac{1}{2^{n+1}}, \frac{1}{2^n}\right), n \in \mathbb{N}\right\}$. Possiamo osservare che $\mathcal{F}$ è una famiglia di sottoinsiemi di Ω, ma non è né un'algebra né una tribù. Quali dei seguenti insiemi sono nella tribù $\sigma(\mathcal{F})$ generata da $\mathcal{F}$?

1. $\{0\}$
2. $\{1\}$
3. $\left\{\frac{1}{2}\right\}$
4. $\left\{\frac{1}{3}\right\}$
5. $[0,1]$
6. $\left[\frac{1}{4}, 1\right]$
7. $\left[0, \frac{1}{2}\right]$
8. $\left[\frac{1}{4}, 1\right)$
9. $\left(0, \frac{1}{2}\right)$
10. $(0,1)$

Soluzione. TBA

2.2. SPAZI DI PROBABILITÀ

Abbiamo fatto tutto questo lavoro di teoria degli insiemi per poter introdurre le prossime tre definizioni sulla probabilità. Cominciamo mettendo assieme due oggetti che abbiamo già definito.

DEFINIZIONE 2.18. *Dati un insieme Ω e una*[2.6] *tribù $\mathcal{F}$ su di esso, la coppia $(\Omega, \mathcal{F})$ prende il nome di* spazio probabilizzabile.

Il nome ci suggerisce che siamo quasi arrivati al nostro obiettivo: abbiamo le fondamenta su cui costruire o definire la probabilità, anche se siamo ancora a una probabilità "in potenza". Ricordiamo che vogliamo far sì che ogni evento abbia una probabilità, quindi dobbiamo definire una funzione che abbia come dominio $\mathcal{F}$.

Qui entra in gioco Kolmogorov, che ci dice quali sono le proprietà che deve soddisfare una funzione per essere accettabile come funzione di probabilità.

DEFINIZIONE 2.19. *Assegnato uno spazio probabilizzabile $(\Omega, \mathcal{F})$, una funzione $P\colon \mathcal{F} \to \mathbb{R}$ si dice* funzione o misura[2.7] di probabilità *se soddisfa le seguenti proprietà (dette* assiomi di Kolmogorov*):*

1. *per ogni evento E, $P(E) \geqslant 0$ (*non negatività*);*
2. *$P(\Omega) = 1$ (*normalizzazione*);*
3. *data una famiglia numerabile $(E_i)_{i=1}^{+\infty}$ di eventi a due a due disgiunti (cioè $E_i \cap E_j = \emptyset$ se $i \neq j$) allora $P(\bigcup_{i=1}^{+\infty} E_i) = \sum_{i=1}^{+\infty} P(E_i)$ (σ-*additività*).*

Il valore $P(E)$ della funzione in un evento E si dice probabilità di *E.*

2.6. Abbiamo già visto, ma lo sottolineiamo ancora una volta, che dato Ω, in genere $\mathcal{F}$ non è unica. Scegliere una particolare tribù tra quelle disponibili è una scelta di modello: a priori non esiste una scelta giusta, dipende dal problema che stiamo considerando. Di volta in volta, sceglieremo $\mathcal{F}$ in modo che sia adatta ai nostri scopi.

2.7. Il nome *misura* viene dal fatto che questa funzione misura la grandezza dell'evento in termini di probabilità. Vedremo più avanti che prendendo come Ω l'intervallo $[0,1]$, una particolare misura di probabilità è quella che restituisce la lunghezza dei segmenti, cioè la loro misura.

Possiamo considerare una versione finita del terzo assioma: se abbiamo una famiglia finita di eventi disgiunti $(E_i)_{i=1}^n$, allora $P(\bigcup_{i=1}^n E_i) = \sum_{i=1}^n P(E_i)$. Chiaramente il terzo assioma implica questa versione finita, ma in genere non vale il viceversa, a meno che $\mathcal{F}$ non sia un'algebra finita, cosa che sappiamo essere vera ogni volta che Ω è finito.

Possiamo ora dare la definizione cui stavamo puntando dall'inizio di questo capitolo.

DEFINIZIONE 2.20. *Siano Ω un insieme, $\mathcal{F}$ una tribù su Ω e P una funzione di probabilità su $\mathcal{F}$. La tripla $(\Omega, \mathcal{F}, P)$ prende il nome di* spazio di probabilità.

Esempio 2.21. Se prendiamo $\Omega = \{0,1\}$, $\mathcal{F} = \{\emptyset, \{0\}, \{1\}, \{0,1\}\} = \mathcal{P}(\Omega)$ e P tale che

$$P(\emptyset) = 0, \qquad P(\{0\}) = P(\{1\}) = \frac{1}{2}, \qquad P(\{0,1\}) = 1,$$

abbiamo uno spazio di probabilità. Possiamo mostrare che tutte le proprietà sono soddisfatte: $\mathcal{F}$ è una tribù, $P(E) \geqslant 0$ per ogni $E \in \mathcal{F}$, $P(\Omega) = 1$, e $P(\{0\}) + P(\{1\}) = 1 = P(\{0,1\})$.

In particolare se identifichiamo 0 con ``testa'' e 1 con ``croce'', questo è un modo di rappresentare il lancio di una moneta bilanciata come spazio di probabilità.

Esempio 2.22. Prendiamo ora $\Omega = \{\clubsuit, \blacklozenge, \heartsuit, \spadesuit\}$ e $\mathcal{F} = \mathcal{P}(\Omega)$. Della probabilità P sappiamo quanto segue:

$$\begin{aligned} P(\emptyset) &= 0 \quad P(\{\clubsuit\}) = P(\{\blacklozenge\}) = \frac{1}{3} \quad P(\{\clubsuit, \blacklozenge, \spadesuit\}) = \frac{7}{9} \\ P(\{\spadesuit\}) &= q \quad P(\{\heartsuit\}) = p. \end{aligned}$$

Possiamo determinare p e q tali per cui P può essere una probabilità?

Se vogliamo che P sia una probabilità, $P(\Omega) = 1$ e quindi

$$1 = P(\Omega) = P(\{\clubsuit, \blacklozenge, \spadesuit\} \cup \{\heartsuit\}) = P(\{\clubsuit, \blacklozenge, \spadesuit\}) + P(\{\heartsuit\}) = \frac{7}{9} + p,$$

da cui $p = \frac{2}{9}$. A questo punto possiamo ricavare q, in modo del tutto simile,

$$\begin{aligned} 1 &= P(\Omega) = P(\{\clubsuit\} \cup \{\blacklozenge\} \cup \{\heartsuit\} \cup \{\spadesuit\}) \\ &= P(\{\clubsuit\}) + P(\{\blacklozenge\}) + P(\{\heartsuit\}) + P(\{\spadesuit\}) \\ &= \frac{1}{3} + \frac{1}{3} + \frac{2}{9} + q \end{aligned}$$

da cui $q = \frac{1}{3} - \frac{2}{9} = \frac{1}{9}$.

Perché questo ci dice che P *può* essere una probabilità (e non che P *è* una probabilità)? Perché non sappiamo quanto valga, ad esempio, $P(\{\clubsuit, \blacklozenge\})$. Se fosse $P(\{\clubsuit, \blacklozenge\}) \neq \frac{2}{3}$, P non potrebbe essere una probabilità, perché avremmo una contraddizione con il terzo assioma.

Prima di continuare, vale la pena fare un'osservazione: gli assiomi di Kolmogorov non ci dicono come definire *la* probabilità sul nostro spazio probabilizzabile, ma ci permettono di dire se una funzione definita su $(\Omega, \mathcal{F})$ sia o meno una misura di probabilità. C'è un buon motivo per cui gli assiomi non ci garantiscono l'unicità della probabilità: questa unicità non c'è! Una volta fissato lo spazio probabilizzabile, possiamo definire più probabilità non equivalenti tra loro.

Esempio 2.23. Prendiamo $\Omega = \{0,1\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, cioè lo stesso spazio probabilizzabile visto nell'Esempio 2.21. Possiamo definire $Q\colon \mathcal{F} \to [0,1]$ come segue:

$$Q(\emptyset) = 0, \quad Q(\{0\}) = \frac{3}{5}, \quad Q(\{1\}) = \frac{2}{5}, \quad Q(\{0,1\}) = 1.$$

Anche la funzione Q appena definita è una probabilità, ma è diversa dalla probabilità P vista prima. In particolare, possiamo vedere questo spazio di probabilità come un modello matematico per una moneta sbilanciata in cui la ``testa'' è una volta e mezza più probabile della ``croce''.

Quello che abbiamo visto in quest'ultimo esempio non è un caso isolato: vedremo più avanti come costruire probabilità in modo che soddisfino gli assiomi di Kolmogorov, ma siano anche buoni modelli per i problemi che considereremo volta per volta.

Lezione 3

2.3. PROPRIETÀ DELLA (MISURA DI) PROBABILITÀ

Esempio 2.24. Vogliamo descrivere un esperimento aleatorio in cui un individuo lancia delle freccette ad un bersaglio costituito da tre cerchi concentrici, di raggi r, $2r$ e $3r$. A seconda della corona circolare che la freccia colpisce, il punteggio è, dall'interno all'esterno, 25, 10 e 5 punti.

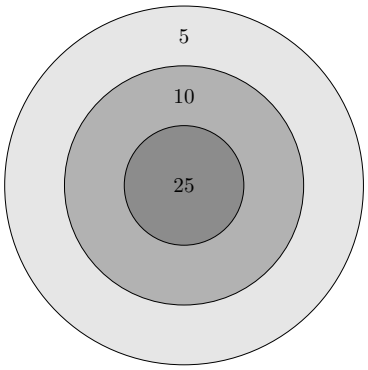


Figura 2.1. Freccette

Un possibile modo di farlo è il seguente: scegliamo $\Omega = \{1,2,3\}$, con le tre aree indicizzate dal loro raggio (o meglio dal rapporto tra il loro raggio ed r. Osserviamo che con questa scelta stiamo escludendo l'eventualità che la freccetta manchi il bersaglio (o, con una terminologia che vedremo in seguito, stiamo condizionando all'aver colpito il bersaglio). Con questa scelta, come tribù è ragionevole scegliere $\mathcal{F} = \sigma\{\{1\},\{2\},\{3\}\} = \mathcal{P}(\Omega)$.

Arriviamo alla scelta della probabilità P: se non sappiamo nulla delle abilità di lancio del giocatore, una possibile descrizione dell'esperimento è ritenere la probabilità di colpire una delle tre aree proporzionale alla superficie dell'area stessa. Abbiamo allora

$$\begin{cases} P(\{1\}) &= \frac{\pi r^2}{\pi (3r)^2} = \frac{1}{9} \\ P(\{2\}) &= \frac{\pi (2r)^2 - \pi r^2}{\pi (3r)^2} = \frac{1}{3} \\ P(\{3\}) &= \frac{\pi (3r)^2 - \pi (2r)^2}{\pi (3r)^2} = \frac{5}{9} \end{cases}$$

Possiamo osservare che queste probabilità non dipendono dal raggio e quindi dalla superficie delle aree, ma solo dai rapporti tra le superfici.

Inoltre, la probabilità di colpire l'area centrale è 5 volte più piccola di quella di colpire la corona circolare più esterna, quindi è sensato che il punteggio sia 5 volte maggiore, mentre per la corona circolare centrale un punteggio più equo (nel senso di proporzionale alla probabilità) sarebbe $8.\bar{3}$. Torneremo a parlare di "equità" più avanti nel corso, quando parleremo di speranza matematica.

Quale potrebbe essere una scelta diversa per Ω? Di quali altri fattori potremmo tenere conto nello scegliere P? Quali limitazioni imposte dalle nostre scelte di modello sono quelle di cui vorremmo fare a meno?

Le proprietà viste sopra sono quelle essenziali per caratterizzare una probabilità. Tuttavia ce ne sono molte altre, che possiamo dedurre da quelle enunciate nella Definizione 2.19 e dalle proprietà delle tribù. Nelle prossime pagine ne vedremo un po', alcune ovvie, altre meno. Tutte quante, però, importanti per manipolare le probabilità, come vedremo negli esempi.

PROPOSIZIONE 2.25. *La probabilità dell'evento $\emptyset$ è sempre uguale a 0.*

Dimostrazione. Osserviamo che $\Omega \cup \emptyset = \Omega$ e che allo stesso tempo $\Omega \cap \emptyset = \emptyset$. Allora abbiamo

$$1 = P(\Omega) = P(\Omega \cup \emptyset) = P(\Omega) + P(\emptyset) = 1 + P(\emptyset),$$

in cui la prima e la quarta uguaglianza seguono dal secondo assioma nella Definizione 2.19 e la terza identità dal terzo assioma in versione finita. Da questa identità ricaviamo $P(\emptyset) = 0$. □

PROPOSIZIONE 2.26. *Se $E \in \mathcal{F}$, la probabilità del suo complementare E^c è $P(E^c) = 1 - P(E)$.*

Dimostrazione. Come prima cosa, sappiamo che P è definita in E^c, poiché $\mathcal{F}$ è una tribù ed è chiusa rispetto all'operazione di complementare. Inoltre, possiamo osservare che $E \cup E^c = \Omega$ e che $E \cap E^c = \emptyset$, quindi

$$1 = P(\Omega) = P(E \cup E^c) = P(E) + P(E^c),$$

in cui l'ultima uguaglianza segue dal terzo assioma (in versione finita) della Definizione 2.19. □

Questa è forse la proprietà della probabilità che sfrutteremo più di tutte nello svolgere esercizi e problemi: molte volte infatti ci verranno forniti dati incompleti, che potremo ricostruire in questo modo. Capiterà spesso che il calcolo diretto della probabilità di un evento sia molto complicato (ad esempio perché ci sono parecchi casi possibili), mentre passando al complementare i conti si semplificano notevolmente.

Esempio 2.27. In un ``Gratta e vinci''[2.8] ci sono premi di prima e seconda fascia. La probabilità di vincere un premio di prima fascia è $\frac{1}{1000000}$, quella di vincere un premio di seconda fascia è $\frac{1}{100}$. Con che probabilità, giocando, non si vince nulla?

Le due fasce cui appartengono i premi sono distinte tra loro, quindi la probabilità di vincere qualcosa è la somma delle due probabilità assegnate, cioè $\frac{1}{1000000} + \frac{1}{100} = \frac{10001}{1000000}$. Allo stesso tempo, non vincere nulla è l'evento complementare al vincere qualcosa, quindi la sua probabilità è

$$1 - \frac{10001}{1000000} = \frac{989999}{1000000} \approx 99\%.$$

Vediamo un'altra proprietà, che prende il nome di *monotonia* della probabilità.

PROPOSIZIONE 2.28. *Siano E, F due eventi in $\mathcal{F}$ tali che $E \subseteq F$. Allora vale la disuguaglianza $P(E) \leqslant P(F)$.*

Dimostrazione. Possiamo riscrivere F come

$$F = (E \cap F) \cup (E^c \cap F) = E \cup (E^c \cap F),$$

che è un'unione disgiunta. A questo punto

$$P(F) = P(E) + P(E^c \cap F) \geqslant P(E),$$

dove per l'uguaglianza sfruttiamo il terzo assioma (in versione finita) della Definizione 2.19, per la disuguaglianza la non negatività del primo assioma. □

Questo risultato dice formalmente quanto avevamo visto nell'Esempio 1, ossia che un evento che è un caso particolare di un altro ha necessariamente probabilità minore o uguale. Tornando alle proprietà, una conseguenza della monotonia della probabilità è la seguente.

COROLLARIO 2.29. *L'immagine della funzione di probabilità è contenuta nell'intervallo unitario* $[0,1]$.

2.8. Le probabilità usate in questo esempio non sono quelle vere, principalmente perché ``Gratta e vinci'' comprende un'ampia famiglia di lotterie istantanee, che cambia spesso e con premi in numero e taglia variabile. Sono comunque probabilità di un ordine di grandezza non dissimile da quello vero. Per chi volesse approfondire, le coordinate di riferimento sono quelle del sito dell'agenzia Dogane e Monopoli, dove per legge sono mostrate le probabilità dei vari premi nelle varie lotterie: `https://www.adm.gov.it/portale/monopoli/giochi/lotterie/lotterie_istantanee/lot_ist_note`. Sempre su questo tema e in generale su quello dei giochi d'azzardo e della probabilità a essi collegata, una lettura divertente e interessante è *Fate il nostro gioco*.

Dimostrazione. Segue immediatamente dal fatto che, per ogni evento $E \in \mathcal{F}$, $\emptyset \subseteq E \subseteq \Omega$ e dalla monotonia. □

Vediamo ora qualcosa che apparentemente abbiamo già incontrato: la probabilità dell'unione di due eventi.

PROPOSIZIONE 2.30. *Siano E, F due eventi in $\mathcal{F}$, allora la probabilità della loro unione è*

$$P(E \cup F) = P(E) + P(F) - P(E \cap F).$$

Dimostrazione. Come nelle dimostrazioni precedenti vogliamo andare a riscrivere questo insieme come unione disgiunta. Per farlo, osserviamo che $E \cup F = E \setminus F \cup F$ e che $E \setminus F \cap F = \emptyset$. Allora

$$P(E \cup F) = P(E \setminus F) + P(F).$$

Tuttavia, non sappiamo quale sia il valore[2.9] di $P(E \setminus F)$. Possiamo però riscrivere E come $E = (E \cap F) \cup (E \setminus F)$, notando che si tratta di un'unione disgiunta, quindi $P(E) = P(E \cap F) + P(E \setminus F)$. A questo punto dobbiamo solo andare a sostituire per ottenere la tesi. □

Confrontiamo quanto visto ora e l'enunciato in versione finita del terzo assioma: in quest'ultimo la probabilità dell'unione era la probabilità che accadesse esattamente uno dei due eventi (poiché erano mutualmente esclusivi). Qui invece abbiamo una vera unione: stiamo chiedendo che almeno uno degli eventi si sia verificato e contempliamo anche la possibilità che si siano verificati entrambi. In analogia con il principio di inclusione ed esclusione, togliamo la probabilità dell'evento intersezione, cioè "sono avvenuti entrambi", dal totale, per non contarla due volte.

Esempio 2.31. In un videogioco, la probabilità di trovare un oggetto raro in uno dei contenitori posti in giro è del 4%, mentre quella di trovare un oggetto magico è del 12%. La probabilità di trovare un oggetto raro che sia anche magico è dell'1%. Qual è la probabilità di trovare un oggetto che sia magico o raro?

Dobbiamo sommare la probabilità di avere un oggetto magico e quella di avere un oggetto raro, per un totale del 16%. Tuttavia, abbiamo contato due volte la probabilità di avere un oggetto che sia contemporaneamente magico e raro, uguale all'1%. Dobbiamo quindi sottrarre, ottenendo 15%.

Esempio 2.32. In una scuola, la probabilità che una studentessa o uno studente abbia in pagella un'insufficienza in matematica è $\frac{17}{24}$, che ne abbia una in inglese è $\frac{5}{6}$. Quanto vale, come minimo, la probabilità di avere un'insufficienza in entrambe le materie?

Questo problema sembra diverso da quello precedente, ma in realtà possiamo risolverlo in modo molto simile. Cominciamo sommando le due probabilità che conosciamo: $\frac{17}{24} + \frac{5}{6} = \frac{37}{24}$. Osserviamo che questa quantità è maggiore di 1.

Non abbiamo ancora scritto nulla che coinvolga la probabilità dell'intersezione, che chiamiamo p ed è la quantità che vogliamo calcolare. Sappiamo però che la probabilità di avere un'insufficienza in almeno una materia è $\frac{17}{24} + \frac{5}{6} - p = \frac{37}{24} - p$. Affinché sia una probabilità, questa quantità deve essere minore o uguale a 1, cioè $\frac{37}{24} - p \leqslant 1$, da cui $\frac{37}{24} - 1 \leqslant p$, quindi $p \geqslant \frac{13}{24}$.

Abbiamo una conseguenza immediata della Proposizione 2.30.

COROLLARIO 2.33. *Possiamo maggiorare la probabilità dell'unione di due eventi con la somma delle probabilità dei due eventi:*

$$P(E \cup F) \leqslant P(E) + P(F).$$

2.9. Anche se sappiamo che è definita, poiché $E \setminus F = E \cap F^c \in \mathcal{F}$.

Questa proprietà prende il nome di sub-additività.

Riguardiamo la Proposizione 2.30 e il parallelo fatto col principio di inclusione-esclusione. Non ci sorprende, a questo punto, che come il principio combinatorio vale per un qualunque numero finito di insiemi, la Proposizione 2.30 possa essere estesa a un generico numero n di eventi.

PROPOSIZIONE 2.34. *Sia $(E_i)_{i=1}^n$ una famiglia finita di eventi. Allora*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right). \tag{2.1}$$

Possiamo generalizzare a questo caso il Corollario 2.33.

COROLLARIO 2.35. *È possibile maggiorare la probabilità di un'unione finita di eventi con la somma delle probabilità:*

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leqslant \sum_{i=1}^{n} P(E_i).$$

Possiamo in realtà dare un risultato più raffinato, ripensando ancora una volta a quanto detto per il principio di inclusione-esclusione.

PROPOSIZIONE 2.36. *La probabilità dell'unione di un numero finito di eventi può essere stimata dall'alto troncando il secondo membro della (2.1) in modo che il primo termine che tralasciamo sia di segno negativo, oppure dal basso, se il primo termine che ignoriamo è di segno positivo. In particolare*

$$\sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) \leqslant P\left(\bigcup_{i=1}^{n} E_i\right) \leqslant \sum_{i=1}^{n} P(E_i).$$

Queste disuguaglianze prendono il nome di disuguaglianze di Bonferroni[2.10].

A differenza di quanto visto per la combinatoria, però, per la probabilità abbiamo anche il caso delle unioni numerabili: abbiamo stabilito nella definizione di tribù che tali unioni di eventi fossero esse stesse eventi. Un risultato elementare in questo contesto è la generalizzazione del Corollario 2.35 al caso numerabile, detta anche disuguaglianza di Boole[2.11].

PROPOSIZIONE 2.37. *Data una famiglia numerabile di eventi $(E_i)_{i=1}^{+\infty}$, possiamo stimare dall'alto la probabilità della sua unione con la somma delle probabilità dei singoli eventi:*

$$P\left(\bigcup_{i=1}^{+\infty} E_i\right) \leqslant \sum_{i=1}^{+\infty} P(E_i).$$

Questo significa che la probabilità è σ-sub-additiva.

Dimostrazione. Per l'unione numerabile al momento abbiamo solo l'assioma 3, quindi dobbiamo trovare un modo di riscrivere il problema in termini di unione di eventi disgiunti. Possiamo farlo nel modo seguente:

$$\begin{cases} F_1 = E_1 \\ F_k = E_k \setminus \bigcup_{i=1}^{k-1} F_i, \quad k \geqslant 2. \end{cases}$$

2.10. Carlo Emilio Bonferroni (1892 – 1960).

2.11. George Boole (1815 – 1864).

In questo modo gli eventi F_i sono a due a due disgiunti e la loro unione coincide con l'unione degli E_i. In più, per ogni $k \in \mathbb{N}$, $F_k \subseteq E_k$, quindi possiamo sfruttare la monotonia:

$$P\left(\bigcup_{i=1}^{+\infty} E_i\right) = P\left(\bigcup_{i=1}^{+\infty} F_i\right) = \sum_{i=1}^{+\infty} P(F_i) \leqslant \sum_{i=1}^{+\infty} P(E_i),$$

in cui abbiamo usato per seconda uguaglianza il terzo assioma nella Definizione 2.19 e per la disuguaglianza la Proposizione 2.28. □

2.4. PROBLEMI

Problema 2.2. Lanciando un dado a 12 facce in cui ogni faccia pari esce con probabilità $\frac{1}{18}$ e ogni faccia dispari con probabilità $\frac{1}{9}$, con che probabilità esce un multiplo di 3 o di 7?

Soluzione. I multipli di 3 che compaiono sul dado sono 3, 6, 9 e 12, mentre di multipli di 7 c'è solo il 7. Non ci sono numeri che siano multipli di 3 e 7, quindi non ne contiamo alcuno due volte. Di questi numeri, 3 sono dispari e 2 sono pari, quindi la probabilità cercata è

$$3 \cdot \frac{1}{9} + 2 \cdot \frac{1}{18} = \frac{4}{9}.$$

Problema 2.3. In una particolare estrazione del Lotto matematico su tutti i numeri naturali, gli infiniti numeri non escono tutti con la medesima probabilità. Sui numeri dispari abbiamo un po' di informazioni: 1 esce con probabilità $\frac{1}{3}$, 3 con probabilità $\frac{1}{9}$, 5 con probabilità $\frac{1}{27}$ e così via. In generale il k-esimo numero dispari esce con probabilità $\frac{1}{3^k}$. La probabilità che esca un numero pari, poi, è doppia rispetto alla probabilità che esca un numero pari positivo. Con che probabilità esce 0?

Soluzione. Abbiamo una bella collezione di eventi disgiunti, in numero infinito, ma numerabile: tutti i singoletti dei numeri dispari e i numeri pari. La probabilità che esca un numero dispari è la somma della serie geometrica, $\sum_{k=1}^{+\infty} 3^{-k} = \frac{1}{2}$. La probabilità che esca un numero pari è allora $1 - \frac{1}{2} = \frac{1}{2}$. Chiamiamo E l'insieme dei numeri pari positivi, allora $\frac{1}{2} = P(\{0\}) + P(E) = 2 \cdot P(E)$, da cui $P(\{0\}) = P(E) = \frac{1}{4}$.

Problema 2.4. Un'urna contiene 16 biglie bianche e 11 nere. Pescandone 4 assieme, qual è la probabilità che non siano tutte del medesimo colore?

Soluzione. Possiamo risolvere questo esercizio andando a considerare tutti i casi favorevoli, ossia 3 biglie bianche e 1 nera, 2 bianche e 2 nere, oppure 3 nere e 1 bianca. Tuttavia dovremmo tenere conto dei diversi modi di ottenere le varie combinazioni, oltre alle rispettive probabilità. Se invece calcoliamo la probabilità che le cose non vadano come vogliamo, abbiamo meno casi (e più semplici) da considerare.

Le biglie possono essere tutte del medesimo colore se sono tutte bianche, oppure tutte nere. La probabilità che siano tutte bianche è $\frac{16}{27} \cdot \frac{15}{26} \cdot \frac{14}{25} \cdot \frac{13}{24}$, che siano tutte nere è $\frac{11}{27} \cdot \frac{10}{26} \cdot \frac{9}{25} \cdot \frac{8}{24}$. Sommando le probabilità di questi due casi disgiunti abbiamo $\frac{51600}{421200} = \frac{43}{351}$. Dal momento che a noi interessa la probabilità dell'evento complementare, per avere la risposta non ci resta che sottrarre questo numero da 1:

$$1 - \frac{43}{351} = \frac{308}{351}.$$

Problema 2.5. A una festa ciascuna delle n invitate porta un regalo, chiuso in un pacchetto e incartato. Questi pacchetti, tutti della stessa dimensione e con la stessa carta, vengono messi su un tavolo e, nel momento clou della festa, ridistribuiti tra le partecipanti. Qual è la probabilità che almeno un'invitata riceva il regalo che ha portato?

Soluzione. Numeriamo le invitate da 1 a n e mettiamoci nei panni dell'invitata i. La probabilità che costei riceva il proprio pacchetto è $\frac{1}{n}$. Possiamo chiamare R_i l'evento "invitata i-esima riceve il proprio regalo". Quello che vogliamo calcolare è allora la probabilità dell'unione di questi R_i al variare di i tra 1 e n: $P(\bigcup_{i=1}^{n} R_i)$. Si tratta però di eventi non disgiunti. Infatti più invitate potrebbero ricevere il proprio pacchetto. Dobbiamo quindi calcolare le probabilità delle intersezioni e usare il principio di inclusione–esclusione.

Se consideriamo due invitate i e j, la probabilità che entrambe ricevano il proprio regalo è $\frac{1}{n} \cdot \frac{1}{n-1}$, corrispondente all'evento $R_i \cap R_j$. Osserviamo che questa probabilità dipende, come già $P(R_i)$, solamente da quante invitate sono coinvolte e non dalle loro identità. In generale, se guardiamo l'evento in cui k invitate ricevono il proprio regalo, esso avrà probabilità $\frac{1}{n} \cdot \frac{1}{n-1} \cdot \cdots \cdot \frac{1}{n-(k-1)} = \frac{(n-k)!}{n!}$. Per ogni k, però, abbiamo $\binom{n}{k}$ modi di scegliere k invitate tra le partecipanti. Allora la probabilità cercata, ossia la probabilità dell'unione, sarà data da

$$P\left(\bigcup_{i=1}^{n} R_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \binom{n}{k} \frac{(n-k)!}{n!} = \sum_{k=1}^{n} (-1)^{k+1} \frac{1}{k!}.$$

Il problema può dirsi concluso: per valori assegnati di n basta andare a calcolarsi questa somma (finita) a segni alterni. Tuttavia, chi avesse già incontrato la funzione esponenziale espressa come serie potrebbe riconoscere l'ultimo membro come l'approssimazione al grado n-esimo di $1-e^{-1}$ (e saper dare quindi più facilmente un'approssimazione del valore cercato).

Abbiamo anche la risposta al problema complementare: "Qual è la probabilità che nessuna abbia davanti a sé il proprio pacchetto?". Al crescere del numero delle invitate n, questa quantità tende a $e^{-1} \approx 37\%$.

Problema 2.6. Una coppia di rapinatori ha svaligiato una banca ed è in fuga a piedi verso il proprio covo. La città ha una struttura tipicamente romana, come si vede in Figura 2.2. Quanti sono i percorsi di lunghezza minima che i ladri hanno a disposizione?

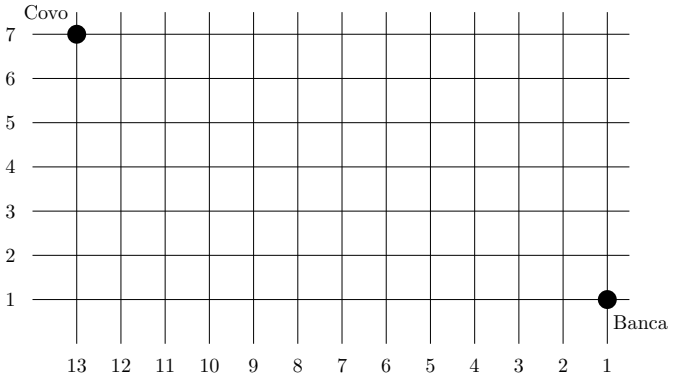


Figura 2.2. La fuga dei malfattori

Soluzione. I rapinatori si muovono solamente lungo le strade e devono necessariamente fare 6 isolati verso Nord e 12 isolati verso Ovest. Vogliono anche fare meno strada possibile, quindi non si muoveranno mai verso Est o verso Sud.

Vogliamo allora contare i percorsi costituiti esattamente da 6 spostamenti verso Nord e da 12 verso Ovest. Possiamo vederli come parole di $6+12=18$ lettere, di cui 6 "N" e 12 "O".

Ci siamo ricondotti quindi al conteggio degli anagrammi di

``NNNNNNOOOOOOOOOOOO''.

In tutto i percorsi di lunghezza minima sono $\binom{18}{6} = 18564$.

Problema 2.7. Nelle stesse ipotesi del problema precedente, la polizia ha avuto una soffiata e ha piazzato due posti di blocco, come indicato in Figura 2.3: se i rapinatori passano di lì, vengono arrestati. Se i rapinatori scelgono uniformemente a caso tra tutti i percorsi di lunghezza minima, con che probabilità verranno catturati?

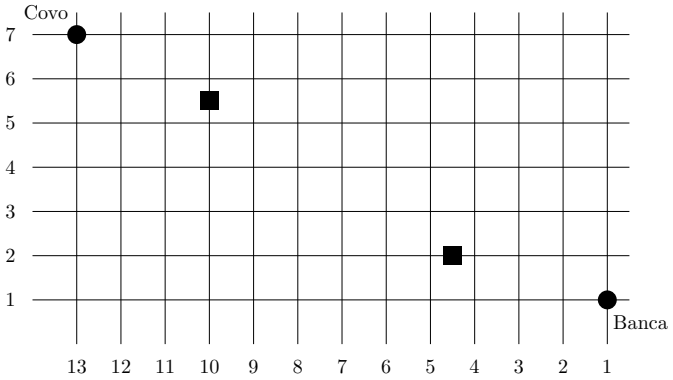


Figura 2.3. Malfattori in fuga con posti di blocco

Soluzione. I casi totali sono tanti quanti i percorsi di lunghezza minima che vanno al covo, già contati nel Problema 2.6: $\binom{18}{6}$. Quanti sono quelli che portano alla cattura? Tutti quelli che passano dal primo posto di blocco, più tutti quelli che passano dal secondo posto di blocco meno quelli che passano da entrambi[2.12], usando il principio di inclusione-esclusione. Possiamo contare il numero dei possibili percorsi esattamente come prima, eventualmente scomponendo il percorso in sotto-percorsi.

2.12. Osserviamo che questo dipende dalla particolare disposizione scelta dei posti di blocco: ci sono posizionamenti della polizia tali per cui non esistono percorsi di lunghezza minima che passano da entrambi.

Contiamo allora quanti sono i percorsi che passano dal primo posto di blocco: $\binom{4}{1}\cdot 1\cdot\binom{13}{5}$ perché dobbiamo raggiungere il nodo $(4,2)$, passare dal posto di blocco e poi andare dal nodo $(5,2)$ al covo in $(13,7)$. In modo simile contiamo quanti passano dal secondo, $\binom{13}{4}\cdot 1\cdot\binom{4}{1}$, e quanti da entrambi, $\binom{4}{1}\cdot 1\cdot\binom{8}{3}\cdot 1\cdot\binom{4}{1}$. In tutto i percorsi di lunghezza minima che portano alla cattura sono

$$4\cdot\left[\binom{13}{4}+\binom{13}{5}\right]-4\cdot\left[\binom{8}{3}\cdot 4\right]=4\cdot\left[\binom{14}{5}-4\cdot\binom{8}{3}\right]=7112.$$

La probabilità che vengano catturati è allora $\frac{7112}{18564}=\frac{254}{663}\approx 38.3\%$, facendo il rapporto con il numero di percorsi totali di lunghezza minima.

CAPITOLO 3

PROBABILITÀ CONDIZIONATA

In questo modo la probabilità diventa in un certo senso una misura di informazione, ed è quindi naturale pensare che, se accumuliamo nuovi dati riguardo a un evento, possiamo e dobbiamo aggiornare la probabilità che gli assegniamo. Questo punto di vista è molto vicino al metodo scientifico: non possiamo mai avere certezza di nulla, ma una serie di risultati sperimentali favorevoli a una nostra teoria farà aumentare la nostra confidenza nel fatto che tale teoria possa dare una buona spiegazione.

Supponiamo, in un certo spazio di probabilità, di venire a sapere che un certo evento F si è verificato. Se a questo punto vogliamo valutare di nuovo la probabilità di un altro evento E, vorremo tener conto delle informazioni in più che abbiamo, ossia che è successo F. Parliamo in questo caso di probabilità condizionata.

DEFINIZIONE 3.1. *Dato uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$ *e due eventi* E *ed* F *in* $\mathcal{F}$*, con* $P(F) \neq 0$*, definiamo la* probabilità di E condizionata a F *come*

$$P(E|F) = \frac{P(E \cap F)}{P(F)}.$$

Se guardiamo il numeratore della definizione, stiamo considerando solo gli esiti in E che possono verificarsi in un mondo nel quale sappiamo che F non è più solo una possibilità, ma un dato di fatto. Per quanto riguarda il denominatore, dividiamo per $P(F)$ perché il nostro universo è ora il solo F e, dal momento che vogliamo avere di nuovo una probabilità, dobbiamo rinormalizzare opportunamente. Nella Figura 3.1 possiamo vedere un'illustrazione in termini di insiemi della definizione.

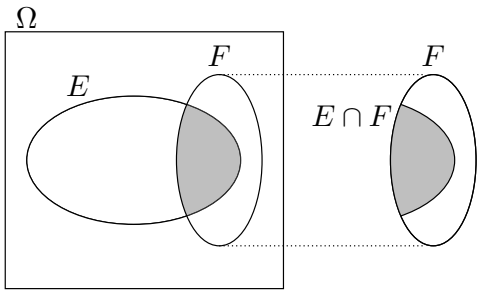


Figura 3.1. Nel condizionamento F è il ``nuovo'' universo

Esempio 3.2. Rosalia lancia un normale dado a 6 facce. Come spesso accade, il dado cade a terra e Rosalia non vede cos'è uscito. Stefano, che vede il risultato del dado, le dice che è uscito un numero dispari. Qual è la probabilità che Rosalia abbia fatto 3? E qual è la probabilità che non abbia fatto 1?

Stiamo considerando il lancio di un dado a 6 facce. Abbiamo quindi come possibile scelta dello spazio degli esiti $\Omega = \{1,2,3,4,5,6\}$, per l'algebra $\mathcal{F} = \mathcal{P}(\Omega)$ e per funzione di probabilità quella che dà peso $\frac{1}{6}$ a ogni singoletto. L'informazione fornita da Stefano è che si è verificato l'evento $F = \{1,3,5\}$. A questo punto possiamo usare la definizione per calcolare le quantità richieste.

La probabilità che sia uscito 3 è

$$P(\{3\}|F) = \frac{P(\{3\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{P(\{3\})}{P(\{1,3,5\})} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}.$$

La probabilità che non sia uscito 1 è

$$P(\{1\}^c|F) = \frac{P(\{2,3,4,5,6\} \cap \{1,3,5\})}{P(\{1,3,5\})} = \frac{P(\{3,5\})}{P(\{1,3,5\})} = \frac{\frac{2}{6}}{\frac{1}{2}} = \frac{2}{3}.$$

Osservazione 3.3. Se fissiamo un evento F di probabilità non nulla, allora la funzione $P_F\colon \mathcal{F} \to \mathbb{R}$ definita per ogni $E \in \mathcal{F}$ da $P_F(E) = P(E|F)$ è una funzione di probabilità, poiché soddisfa tutti gli assiomi visti. Abbiamo, infatti, che $P_F(E) \geqslant 0$ per qualunque evento E in $\mathcal{F}$, dal momento che stiamo facendo il rapporto tra una quantità non negativa e una positiva. Allo stesso tempo, $P_F(\Omega) = \frac{P(\Omega \cap F)}{P(F)} = 1$. Non resta che verificare che anche il terzo assioma sia soddisfatto: prendiamo una famiglia numerabile $(E_i)_{i=1}^{+\infty}$ di eventi a due a due disgiunti e scriviamone la probabilità condizionata a F dell'unione,

$$\begin{aligned} P_F\left(\bigcup_{i=1}^{+\infty} E_i\right) &= \frac{P[(\bigcup_{i=1}^{+\infty} E_i) \cap F]}{P(F)} = \frac{P[\bigcup_{i=1}^{+\infty} (E_i \cap F)]}{P(F)} \\ &= \frac{\sum_{i=1}^{+\infty} P(E_i \cap F)}{P(F)} = \sum_{i=1}^{+\infty} P_F(E_i), \end{aligned}$$

in cui abbiamo usato la distributività dell'intersezione rispetto all'unione.

Allora anche per P_F valgono le proprietà viste per una qualunque misura di probabilità P. Tutto questo non ci sorprende: abbiamo dato la definizione di probabilità condizionata proprio con l'idea di avere alla fine una misura di probabilità.

Esempio 3.4. Edoardo ama molto correre in montagna quindi, se le previsioni sono buone, la probabilità che passi la domenica sui monti è del 70%. Se le previsioni sono buone, con che probabilità rimane a casa?

Siano M l'evento ``correre in montagna'' e $S$ l'evento ``buone previsioni'':

$$P(M|S) + P(M^c|S) = P(M \cup M^c|S) = P(\Omega|S) = 1,$$

quindi $P(M^c|S) = 1 - P(M|S) = 30\%$.

Nel definire la probabilità condizionata, il nostro scopo era quantificare l'effetto di un evento su un altro, in termini di probabilità. Tuttavia, il bello delle identità è che possiamo rigirarle un po' per mettere in evidenza altri aspetti. In particolare, dalla definizione di probabilità condizionata possiamo ricavare un modo (anzi, due) per scrivere la probabilità dell'intersezione tra due eventi:

$$P(E \cap F) = P(E|F) \cdot P(F) = P(F|E) \cdot P(E). \tag{3.1}$$

La doppia identità (3.1) prende anche il nome di *teorema (o regola) del prodotto*. Osserviamo che nella (3.1) siamo stati un po' imprecisi: non abbiamo specificato che $P(E) \neq 0 \neq P(F)$. Tuttavia, se anche $P(E)$ o $P(F)$ fossero nulli, avremmo che $P(E \cap F) = 0$, perché l'intersezione $E \cap F$ è un evento contenuto in un evento di probabilità nulla (E o F): qualunque valore (finito) assegniamo a $P(E|F)$ (o $P(F|E)$), lo annulleremo moltiplicandolo per 0.

Potrebbe essere interessante caratterizzare quegli eventi che non interagiscono tra loro, quelli che intuitivamente chiameremmo eventi indipendenti. Come prossimo passo vogliamo quindi dare una definizione matematica di indipendenza tra eventi, per poi vedere come essa si sposi con l'idea intuitiva di eventi indipendenti.

DEFINIZIONE 3.5. *In uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, due eventi E ed F in $\mathcal{F}$ si dicono* indipendenti *se vale l'uguaglianza $P(E \cap F) = P(E) \cdot P(F)$.*

Questa definizione, a un primo sguardo, ci sorprende un po': com'è che parliamo di indipendenza tra eventi e ci ritroviamo con una "formula" per la probabilità dell'intersezione? In realtà grazie al legame tra probabilità dell'intersezione e probabilità condizionata possiamo dare un altro punto di vista sull'indipendenza appena definita. Infatti, se due eventi E ed F sono indipendenti, abbiamo

$$P(E)\cdot P(F) = P(E\cap F) = P(E|F)\cdot P(F) \quad \text{e} \quad P(E)\cdot P(F) = P(F|E)\cdot P(E),$$

cioè, supponendo $P(E)\neq 0$ e $P(F)\neq 0$,

$$P(E|F) = P(E) \quad \text{e} \quad P(F|E) = P(F).$$

Quindi sapere che è accaduto F non cambia quello che sappiamo della probabilità di E e, viceversa. Inoltre, avendo due catene di uguaglianze possiamo invertire il ragionamento: sapere che E non ci dà informazioni su F ed F non ci dà informazioni su E implica che E ed F sono indipendenti, per la definizione di indipendenza data sopra. La definizione data caratterizza proprio quello che ci aspettavamo e il termine usato è giustificato.

Se siamo invece interessati alla probabilità dell'intersezione di due eventi, sappiamo che essa è uguale al prodotto delle probabilità dei due eventi se questi ultimi sono tra loro indipendenti. Se non abbiamo questa informazione, dobbiamo usare la regola del prodotto (3.1) vista sopra, oppure l'identità incontrata in precedenza:

$$P(E\cap F) = P(E) + P(F) - P(E\cup F).$$

Esempio 3.6. Gaia sa che le sue professoresse di Storia e di Arte interrogano in ognuna delle due materie a sorteggio tra coloro che ancora non hanno un voto. Sapendo che nella classe, formata da 25 tra studentesse e studenti, nessuno è ancora stato interrogato in Storia e 6 persone (ma non Gaia) hanno un voto in Arte, con che probabilità Gaia verrà interrogata domani?

Gaia ha una probabilità di essere interrogata in Storia uguale a $P(S) = \frac{1}{25}$, come tutte le sue compagne e i suoi compagni di classe, e $P(A) = \frac{1}{19}$ di essere interrogata in Arte. Le due estrazioni vengono fatte da liste (o contenitori) diversi, quindi non si influenzano l'una con l'altra: possiamo allora considerare le due interrogazioni come indipendenti. La probabilità di interrogazione di Gaia è allora

$$P(S\cup A) = P(S) + P(A) - P(S\cap A) = \frac{1}{25} + \frac{1}{19} - \frac{1}{25\cdot 19} = \frac{43}{475},$$

in cui abbiamo usato l'identità per la probabilità dell'unione vista nella Proposizione 2.30 e, per valutare $P(S\cap A)$, l'indipendenza.

Esempio 3.7. Nicolò possiede un'auto sportiva gialla. Un giorno, in un parcheggio, vede che l'auto in sosta accanto alla sua è anch'essa una sportiva gialla. Mentre rientra verso casa, si chiede quanto sia probabile che un'auto sia una sportiva gialla. Da una rapida ricerca online scopre che solamente 1 auto ogni 100 è un'auto sportiva e che solo 1 auto su 200 è gialla. Ne conclude quindi che la probabilità che un'auto sia una sportiva gialla è $\frac{1}{20000}$.

In realtà questo ragionamento non è corretto, perché nulla garantisce che i due eventi ``auto sportiva'' e ``auto gialla'' siano indipendenti. In effetti, con un po' di attenzione, Nicolò scopre poco dopo che tra le auto gialle, 1 su 3 è un'auto sportiva, quindi la probabilità che un'auto a caso sia gialla e sportiva è

$$P(S\cap G) = P(S|G)\cdot P(G) = \frac{1}{3}\cdot\frac{1}{200} = \frac{1}{600} \neq \frac{1}{20000} = P(S)\cdot P(G).$$

Qui l'errore non ha gravi conseguenze, ma una svista simile ha contribuito alla condanna, poi annullata, di Malcolm Ricardo Collins[3.1].

3.1. È un caso giudiziario realmente accaduto negli anni Sessanta in California. Una discussione più dettagliata di questo e di altri esempi reali di errori matematici in ambito processuale si trova nel libro *Math on Trial*.

Dalla riscrittura in termini di probabilità condizionata dell'indipendenza, abbiamo che, se due eventi E ed F sono indipendenti, allora $P(F|E) = P(F)$, ma anche $P(F^c|E) = P(F^c)$. Ma che succede se abbiamo il complementare dall'altro lato del condizionamento?

Esempio 3.8. Prendiamo, in uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, due eventi E ed F tali che $0 < P(F) < 1$ e $P(E|F) = P(E|F^c)$. Possiamo dire che gli eventi E ed F sono indipendenti?

Potremmo sospettare un trabocchetto, quindi andiamo a scriverci con attenzione le quantità:

$$\frac{P(E \cap F)}{P(F)} = P(E|F) = P(E|F^c) = \frac{P(E \cap F^c)}{P(F^c)} = \frac{P(E \cap F^c)}{1 - P(F)}.$$

Ora prendiamo il primo e l'ultimo termine e moltiplichiamoli per $P(F)\,(1 - P(F))$

$$P(E \cap F)\,(1 - P(F)) = P(E \cap F^c)\,P(F)$$

e continuiamo raccogliendo i termini moltiplicati per $P(F)$ a secondo membro,

$$P(E \cap F) = (P(E \cap F) + P(E \cap F^c))\,P(F).$$

A questo punto possiamo osservare che i due eventi $E \cap F$ ed $E \cap F^c$ sono disgiunti e la loro unione è E, quindi, siccome la probabilità dell'unione disgiunta è la somma delle probabilità, abbiamo $P(E \cap F) = P(E)\,P(F)$, ossia l'indipendenza.

Torniamo allora a guardare il testo iniziale e proviamo a rileggere quello che c'è scritto. La condizione $P(E|F) = P(E|F^c)$ ci dice che sapere che F sia accaduto o no non dà alcuna informazione su E; infatti non modifica la sua probabilità.

Lezione 4

Nell'Esempio 3.8 abbiamo incontrato un'idea interessante: abbiamo scritto un evento dividendolo in due pezzi disgiunti, che però esaurissero tutte le possibilità. In realtà non c'è nulla di speciale nel fatto che siano due eventi complementari: le caratteristiche fondamentali sono che gli eventi siano tutti disgiunti, ma che allo stesso tempo coprano tutto lo spazio, cioè ne siano una partizione. Andare a riscrivere la probabilità di un evento in termini delle sue probabilità condizionate a una partizione di eventi è una tecnica molto importante che prende il nome di formula di fattorizzazione (o legge delle probabilità totali). La sua validità è garantita dal seguente teorema.

TEOREMA 3.9. *Dato uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, consideriamo una famiglia al più numerabile di eventi disgiunti $(E_i)_{i \in I}$ che sia anche una partizione di Ω. Supponiamo che ogni evento nella partizione abbia probabilità non nulla. Allora per ogni evento $F \in \mathcal{F}$,*

$$P(F) = \sum_{i \in I} P(F \cap E_i) = \sum_{i \in I} P(F|E_i) \cdot P(E_i).$$

Dimostrazione. Osserviamo che la seconda uguaglianza deriva, addendo per addendo, dalla definizione di probabilità condizionata. Per quanto riguarda la prima, basta osservare che

$$P(F) = P(F \cap \Omega) = P\left(F \cap \left(\bigcup_{i \in I} E_i\right)\right) = P\left(\bigcup_{i \in I} (F \cap E_i)\right)$$

e che l'unione è necessariamente disgiunta, dal momento che per ogni i risulta $F \cap E_i \subset E_i$. □

Osserviamo che in realtà la richiesta che gli eventi nella partizione non abbiano misura nulla non è cruciale: se è vero che non sappiamo determinare il valore di $P(F|E_i)$ per tali eventi, sappiamo che comunque è una probabilità, quindi un numero compreso tra 0 e 1. Questo numero compare moltiplicato per $P(E_i)$, cioè per 0, e ciò risolve i nostri problemi.

La formula di fattorizzazione va a estendere al mondo della probabilità quello che è il principio della somma nella combinatoria: ci permette di dividere il problema in sotto-problemi (auspicabilmente) più facili, dandoci un modo per combinarli alla fine. Di nuovo la strategia del *divide et impera*. Come accennato in precedenza, questo risultato è di grandissima utilità pratica, perché ci permette di spezzare il calcolo della probabilità in più sotto-casi, scelti opportunamente, spesso semplificando enormemente i conti.

Esempio 3.10. A un gruppo di studio per preparare l'esame di probabilità e statistica partecipano solo tre studenti: Carlo, Anita e Francesca. Carlo è un esperto di problemi di combinatoria e ne risolve sei su sette, le altre due preferiscono entrambe la statistica e risolvono gli esercizi di combinatoria solo una volta su quattro. Oggi lavorano indipendentemente su tre problemi, assegnati a caso, di cui solo uno di combinatoria. Qual è la probabilità che alla fine dell'incontro il gruppo abbia una soluzione per il problema di combinatoria?

Cosa sappiamo? Chiamiamo C l'evento ``il problema di combinatoria viene assegnato a Carlo'' e $R$ l'evento ``il problema di combinatoria viene risolto''. Allora il testo ci dice che

$$P(R|C) = \frac{6}{7}, \qquad P(R|C^c) = \frac{1}{4}, \qquad P(C) = \frac{1}{3}.$$

Grazie alla formula di fattorizzazione possiamo riscrivere la probabilità cercata come

$$P(R) = P(R|C) \cdot P(C) + P(R|C^c) \cdot P(C^c) = \frac{6}{7} \cdot \frac{1}{3} + \frac{1}{4} \cdot \frac{2}{3} = \frac{19}{42}.$$

Esempio 3.11. Da un recente sondaggio svolto nell'Arcipelago delle Tre Isole è emerso che nell'isola di Idilos 2 abitanti su 15 sono matematici, nell'isola di Iremun è matematico 1 abitante su 5, mentre sulla terza isola, Erettel, sono 3 su 25. Qual è la probabilità che un qualunque abitante dell'arcipelago sia un matematico, se il 30% vive su Idilos, il 45% su Iremun e il 25% su Erettel?

Indicando con M l'essere un matematico e con S, N ed E l'essere abitante dell'isola di Idilos, Iremun ed Erettel rispettivamente, abbiamo

$$\begin{aligned} P(M) &= P(M|S) \cdot P(S) + P(M|N) \cdot P(N) + P(M|E) \cdot P(E) \\ &= \frac{2}{15} \cdot \frac{30}{100} + \frac{1}{5} \cdot \frac{45}{100} + \frac{3}{25} \cdot \frac{25}{200} \\ &= 16\% \end{aligned}$$

In pratica quello che stiamo facendo è prendere la media delle probabilità dell'evento che ci interessa (essere matematici) condizionata ai casi disgiunti (vivere in una specifica isola), pesando questa media con le probabilità dei casi stessi.

Facciamo un passo indietro e torniamo all'indipendenza: l'abbiamo definita a partire dalla probabilità dell'intersezione e siamo poi passati al legame con la probabilità condizionata, che ci ha dato una caratterizzazione molto più intuitiva dell'indipendenza stessa. Perché allora non abbiamo usato direttamente la probabilità condizionata per dare la definizione?

Torniamo per un momento alla questione, lasciata in sospeso, del caso in cui abbiamo un evento di probabilità nulla. Cosa succede? Supponiamo che sia $P(E) = 0$. Allora, per monotonia, $P(E \cap F) \leqslant P(E) = 0$, cioè $P(E \cap F) = 0 = P(E) \cdot P(F)$, ossia un evento di probabilità nulla è indipendente rispetto a ogni evento, usando la definizione data. Cosa succede se andiamo a considerare le probabilità condizionate?

Quella che vogliamo guardare è $P(F|E)$, che però non è definita: questo ci obbligherebbe a dare una definizione più macchinosa di indipendenza, specificando a parte il caso in cui un evento ha probabilità nulla. Osserviamo che questo non è davvero influente, perché $P(F|E)$ compare moltiplicato per $P(E)$: $P(F|E)$ è una probabilità e ha un valore compreso tra 0 e 1, quindi anche se non ne conosciamo il valore, sappiamo che il prodotto varrà zero.

A questo punto abbiamo la curiosità di capire quanto possa valere $P(F|E)$ se $P(E)=0$. Quando andiamo ad analizzare i dettagli, però, ci accorgiamo che non ha un valore univoco. Se $E \subset F$, allora sapere che è avvenuto E ci dice automaticamente che è avvenuto F, con probabilità 1. Matematicamente questo torna (con qualche equilibrismo), perché $E \cap F = E$ e quindi abbiamo che i due termini uguali "si semplificano". Se invece $E \cap F = \emptyset$, cioè $E \subset F^c$, sapere che è avvenuto E assegna automaticamente probabilità 0 a F, cosa che possiamo immaginare vedendo la probabilità dell'evento nullo "più nulla" di tutte le altre.

Fin qui sembra andare tutto bene, a parte la seccatura di dover distinguere queste due possibilità. Purtroppo però questi non sono i soli casi possibili: infatti un evento di probabilità nulla può avere intersezione non vuota e differenza non vuota con un altro evento e, in questo caso, non sappiamo assegnare un valore sensato alla probabilità condizionata.

Proseguiamo ora con altre proprietà interessanti del condizionamento.

Esempio 3.12. Lanciamo per l'n-esima volta un dado a 6 facce. Lo spazio probabilizzabile che consideriamo è dunque $\Omega = \{1,2,3,4,5,6\}$ e $\mathcal{F} = \mathcal{P}(\Omega)$. Prendiamo i due eventi $E = \{2,4,6\}$ ed $F = \{3,6\}$.

Supponiamo che il dado sia bilanciato, quindi ogni faccia del dado (ogni singoletto) ha probabilità $P(\{i\}) = \frac{1}{6}$, per ogni $i = 1,\dots,6$. Allora

$$P(E) = \frac{1}{2}, \quad P(F) = \frac{1}{3} \quad \text{e} \quad P(E \cap F) = P(\{6\}) = \frac{1}{6} = P(E) \cdot P(F),$$

cioè i due eventi sono indipendenti.

Supponiamo invece che il dado sia truccato: allora abbiamo una nuova probabilità $\tilde{P}$ tale che

$$\tilde{P}(\{1\}) = \tilde{P}(\{2\}) = \tilde{P}(\{3\}) = \tilde{P}(\{4\}) = \frac{1}{12}, \quad \tilde{P}(\{5\}) = \tilde{P}(\{6\}) = \frac{1}{3}.$$

Dopo aver verificato che si tratta effettivamente di una probabilità, andiamo a calcolare

$$\tilde{P}(E) = \frac{1}{2}, \quad \tilde{P}(F) = \frac{5}{12} \quad \text{e} \quad \tilde{P}(E \cap F) = \tilde{P}(\{6\}) = \frac{1}{3} \neq \frac{5}{24} = \tilde{P}(E) \cdot \tilde{P}(F),$$

ossia sotto questa probabilità i due eventi non sono indipendenti.

Grazie a quest'ultimo esempio, notiamo che l'indipendenza tra due eventi non è una proprietà intrinseca degli eventi stessi, ma dipende dall'intero spazio di probabilità scelto e, in particolare, dalla misura di probabilità. Se consideriamo sullo stesso spazio probabilizzabile due probabilità distinte, può succedere che con una di esse due eventi siano indipendenti e con l'altra no.

C'è un caso particolare che ci interessa, arrivati a questo punto: mettiamo assieme il concetto di indipendenza e una particolare misura di probabilità, la probabilità condizionata. Se fissiamo nel nostro spazio di probabilità $(\Omega, \mathcal{F}, P)$ un evento $F \in \mathcal{F}$ di probabilità non nulla, abbiamo visto che la funzione $P_F \colon \mathcal{F} \to \mathbb{R}$ definita per ogni $E \in \mathcal{F}$ da $P_F(E) = P(E|F)$ è una probabilità e possiamo quindi considerare l'indipendenza tra eventi rispetto a essa. Ne nasce la seguente definizione.

DEFINIZIONE 3.13. *In uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, fissato un evento F tale che $P(F) \neq 0$, due eventi E_1 ed E_2 si dicono* indipendenti condizionalmente a F *se*

$$P(E_1 \cap E_2 | F) = P(E_1|F) \cdot P(E_2|F).$$

L'indipendenza condizionale è distinta dall'indipendenza, come vediamo nei due esempi successivi.

Esempio 3.14. Marko ha due cassetti nel suo armadio, nei quali tiene i suoi calzini. In uno ci sono solo calzini invernali lunghi, nell'altro ci sono calzini estivi sia lunghi sia corti (metà e metà). Marko pesca contemporaneamente due calzini da uno dei due cassetti.

Se chiamiamo S l'evento ``Marko pesca dal secondo cassetto'', gli eventi $L_1$: ``il primo calzino pescato è lungo'' e C_2: ``il secondo calzino pescato è corto'' sono indipendenti condizionalmente a S. Infatti $P(L_1|S) = 0.5 = P(C_2|S)$, quindi $P(L_1 \cap C_2|S) = 0.25 = P(L_1|S)\,P(C_2|S)$.

In generale, tuttavia, se supponiamo che Marko scelga con uguale probabilità dai due cassetti, i due eventi non sono indipendenti. Infatti abbiamo

$$\begin{aligned} P(L_1) &= P(L_1|S)\,P(S) + P(L_1|S^c)\,(1-P(S)) = \frac{1}{4} + \frac{1}{2} = \frac{3}{4} \\ P(C_2) &= P(C_2|S)\,P(S) + P(C_2|S^c)\,(1-P(S)) = \frac{1}{4} + 0 = \frac{1}{4} \\ P(L_1 \cap C_2) &= P(L_1 \cap C_2|S)\,P(S) + P(L_1 \cap C_2|S^c)\,(1-P(S)) = \frac{1}{8} \\ P(L_1) \cdot P(C_2) &= \frac{3}{16} \neq \frac{1}{8}. \end{aligned}$$

Esempio 3.15. Consideriamo ancora una volta il lancio di due dadi a 6 facce. I due eventi D_2: ``il primo dado ha come risultato 2'' ed $E_5$: ``il secondo dado ha come risultato 5'' sono tra loro indipendenti. Tuttavia se condizioniamo rispetto all'evento S_8: ``la somma dei dadi è 8'', vediamo che D_2 ed E_5 non sono indipendenti condizionalmente a S_8. Infatti

$$P(D_2 \cap E_5|S_8) = 0 \neq \frac{1}{25} = P(D_2|S_8) \cdot P(E_5|S_8),$$

poiché per avere la somma dei due dadi uguale a 8, ciascuno dei due può prendere uno dei 5 valori tra 2 e 6, quindi entrambi i fattori a ultimo membro sono $\frac{1}{5}$.

Possiamo prendere una variante dell'esempio precedente e osservare un altro aspetto.

Esempio 3.16. Siano D_2, E_5 come nell'esempio precedente e S_7: ``la somma dei dadi è 7''. Osserviamo che non solo D_2 ed E_5 sono indipendenti tra loro, ma ciascuno di loro è anche indipendente da S_7:

$$\begin{aligned} P(D_2|S_7) &= \frac{1}{6} = P(D_2) \quad P(S_7|D_2) = \frac{1}{6} = P(S_7) \\ P(E_5|S_7) &= \frac{1}{6} = P(E_5) \quad P(S_7|E_5) = \frac{1}{6} = P(S_7). \end{aligned}$$

Questo però non ci basta per dire che sono tutti e tre indipendenti tra loro, infatti

$$P(D_2 \cap E_5 \cap S_7) = \frac{1}{36} \neq \frac{1}{216} = P(D_2) \cdot P(E_5) \cdot P(S_7).$$

Concludiamo queste divagazioni sull'indipendenza con un ultimo esempio, in cui abbiamo indipendenza condizionale rispetto a una partizione.

Esempio 3.17. Prendiamo ora tre eventi D, E ed F sul nostro spazio di probabilità $(\Omega, \mathcal{F}, P)$, con $0 < P(F) < 1$ e supponiamo che D ed E siano indipendenti tra loro condizionalmente a F, ma anche a F^c. Possiamo dire che D ed E sono necessariamente indipendenti tra loro in senso stretto?

Da un lato avremmo la tentazione di rispondere affermativamente: sono indipendenti in ciascuna delle due possibilità determinate da F (sia con F vero, sia con F falso), quindi lo saranno anche globalmente. Allo stesso tempo, però, gli esempi precedenti ci hanno insegnato un po' di prudenza.

Proviamo allora a vedere se ci sono condizioni da soddisfare affinché questa indipendenza sia vera e, allo stesso tempo, se possiamo costruire un controesempio.

Dalla formula di fattorizzazione abbiamo le seguenti identità:

$$\begin{aligned} P(D) &= P(D|F) \cdot P(F) + P(D|F^c) \cdot (1-P(F)) \\ &= (P(D|F) - P(D|F^c)) \cdot P(F) + P(D|F^c) \end{aligned}$$

$$\begin{aligned} P(E) &= P(E|F) \cdot P(F) + P(E|F^c) \cdot (1-P(F)) \\ &= (P(E|F) - P(E|F^c)) \cdot P(F) + P(E|F^c), \end{aligned}$$

cioè, chiamando per semplicità $d = P(D|F)$, $d' = P(D|F^c)$, $e = P(E|F)$, $e' = P(E|F^c)$ e anche $a = P(D)$, $b = P(E)$, $c = P(F)$,

$$\begin{aligned} a &= dc + d'(1-c) = (d-d')c + d' \\ b &= ec + e'(1-c) = (e-e')c + e'. \end{aligned}$$

Allo stesso tempo abbiamo anche, grazie all'indipendenza di D ed E condizionalmente a F ed F^c,

$$\begin{aligned} P(D \cap E) &= P(D \cap E|F) \cdot P(F) + P(D \cap E|F^c) \cdot P(F^c) \\ &= P(D|F) \cdot P(E|F) \cdot P(F) + P(D|F^c) \cdot P(E|F^c) \cdot P(F^c) \\ &= dec + d'e'(1-c). \end{aligned}$$

Avremmo l'indipendenza se valesse $P(D \cap E) = P(D) \cdot P(E)$, cioè, con la nuova notazione, $dec + d'e'(1-c) = ab$. Studiamo allora questa identità.

$$\begin{aligned} dec + d'e' - d'e'c &= ab \\ &= (dc + d' - d'c)(ec + e' - e'c) \\ &= dec^2 + de'c - de'c^2 + d'ec \\ &\quad + d'e' - d'e'c - d'ec^2 - d'e'c + d'e'c^2. \end{aligned}$$

Possiamo semplificare un po' di termini, arrivando a

$$dec^2 + de'c - de'c^2 + d'ec - d'ec^2 - d'e'c + d'e'c^2 - dec = 0$$

che possiamo riscrivere, raccogliendo più volte i fattori in comune, come

$$c(c-1)(d-d')(e-e') = 0,$$

o, tornando esplicitamente alle probabilità,

$$P(F) \cdot P(F^c) \cdot (P(D|F) - P(D|F^c)) \cdot (P(E|F) - P(E|F^c)) = 0.$$

I primi due fattori per ipotesi non possono essere 0 (altrimenti non potremmo parlare di probabilità condizionali), quindi ci sono due possibilità: o la probabilità di D non cambia nelle due parti F ed F^c, o quella di E non cambia.

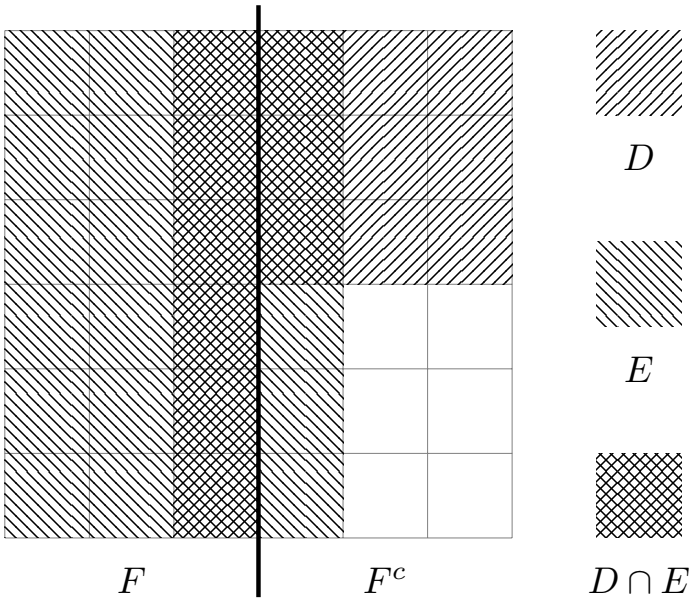


Figura 3.2. Un controesempio

Abbiamo allora tutti gli ingredienti per costruire un controesempio, rappresentato in Figura 3.2. In questo esempio abbiamo $P(F) = P(F^c) = \frac{1}{2}$. La probabilità di ciascun evento è data dalla sua area in quadratini divisa per l'area totale (sempre in quadratini). In F abbiamo $P(D|F) = \frac{1}{3}$, $P(E|F) = 1$ e $P(D \cap E|F) = \frac{1}{3}$, mentre in F^c valgono $P(D|F^c) = \frac{1}{2}$, $P(E|F^c) = \frac{1}{3}$ e $P(D \cap E|F^c) = \frac{1}{6}$. Allora, condizionalmente a F e F^c, D ed E sono indipendenti.

Se guardiamo però le probabilità di D ed E, vediamo $P(D) = \frac{5}{12}$ e $P(E) = \frac{2}{3}$, dunque $P(D) \cdot P(E) = \frac{5}{18}$, mentre $P(D \cap E) = \frac{1}{4}$, quindi D ed E non sono indipendenti.

3.1. TEOREMA DI BAYES

La probabilità condizionata non è simmetrica: in generale $P(E|F) \neq P(F|E)$. Da un punto di vista matematico la cosa è immediata: basta guardare la definizione e osservare che non è simmetrica nei due insiemi considerati. Tuttavia, se andiamo a considerare l'uso della probabilità nella vita di tutti i giorni, ci accorgiamo che questo è uno degli errori (o fallacie) più frequenti.

Esempio 3.18. Da una recente indagine[3.2] sui vaccini per l'influenza stagionale, in Italia la copertura vaccinale per le persone di età maggiore o uguale a 65 anni è del 53.1%. Nella popolazione generale la copertura si riduce al 15.8%. Questo non significa che, scegliendo un vaccinato a caso, la probabilità che abbia almeno 65 anni sia il 53.1%. Infatti gli italiani con almeno 65 anni sono circa 14 milioni, di cui circa 7.5 milioni sono vaccinati. Al tempo stesso la popolazione italiana è costituita da 60 milioni di persone circa, di cui 9.5 milioni vaccinati. Tra i vaccinati, gli over 65 sono quasi il 79%. In termini di probabilità condizionate abbiamo

$$P(\text{vaccinato} \mid \text{over 65}) = 53.1\% \neq P(\text{over 65} \mid \text{vaccinato}) = 78.9\%.$$

Purtroppo nel momento in cui ci si allontana dal contesto esplicitamente matematico, capita spesso che le due probabilità condizionate vengano confuse. Vediamo alcuni tipici esempi.

- "Se la maggior parte dei criminali appartiene a un certo gruppo, allora è altamente probabile che un generico membro del gruppo sia un criminale." Falso: tra i condannati per omicidio in Italia, oltre il 95% sono di sesso maschile, ma non ci verrebbe mai in mente di pensare che quasi tutti i maschi italiani siano assassini.
- "Se la probabilità che un imputato abbia indizi contro di lui pur essendo innocente è molto bassa, allora deve essere molto bassa anche la probabilità che sia innocente se ci sono indizi contro di lui." Falso: questo argomento prende il nome di *fallacia del procuratore* ed è stato ingrediente di molti casi di cattiva giustizia, con condanne annullate in fase di revisione dei processi, ad esempio il già citato caso Collins, ma anche con assoluzioni forse non meritate, come nel caso O. J. Simpson.
- "Se la maggior parte dei recenti attacchi terroristici in Europa è stata portata a termine da musulmani, allora la proporzione di musulmani che sono terroristi è molto alta." Falso anche questa volta: in realtà la probabilità che un musulmano europeo sia un terrorista è dell'ordine di $4 \cdot 10^{-6}$, cento volte più piccola della probabilità di essere colpiti da un fulmine nel corso della propria vita.

Pur non essendoci simmetria, le due probabilità condizionate $P(E|F)$ e $P(F|E)$ non sono completamente scollegate tra loro, come vediamo nel prossimo esempio.

Esempio 3.19. Tra i concorrenti delle Olimpiadi della Matematica[3.3], il 43% è del biennio, il rimanente 57% del triennio. Tra i concorrenti del biennio, il 51% sono ragazze, tra quelli del triennio tale percentuale scende al 23%. Se Giulietta è una concorrente, qual è la probabilità che sia una studentessa del biennio?

Indichiamo con B l'evento ``concorrente del biennio'' e con ♀ l'evento ``concorrente è una ragazza''. Allora dai dati del problema abbiamo:

$$P(B) = 0.43, \quad P(B^c) = 0.57, \quad P(♀|B) = 0.51, \quad P(♀|B^c) = 0.23.$$

3.2. Fonte: Ministero della Salute-ISS per la stagione 2018/19.

3.3. Un sottoinsieme ben determinato degli studenti di scuola secondaria di secondo grado.

Noi però vorremmo calcolare $P(B|\text{♀})$, poiché Giulietta è una ragazza. Cominciamo a calcolare qualcosa di diverso: $P(B \cap \text{♀})$, cioè la probabilità che la persona presa sia del biennio e sia una ragazza. Lo facciamo perché per definizione $P(B|\text{♀}) = \frac{P(B\cap\text{♀})}{P(\text{♀})}$ e stiamo in questo modo calcolando il numeratore. Dalla definizione di probabilità condizionata otteniamo che

$$P(B \cap \text{♀}) = P(\text{♀}|B) \cdot P(B)$$

dove le due quantità a secondo membro sono note. Possiamo allora calcolare esplicitamente $P(B \cap \text{♀}) = 0.51 \cdot 0.43 = 0.2193$.

Per calcolare $P(B|\text{♀})$, la quantità che cerchiamo, non resta che calcolare $P(\text{♀})$, cosa che possiamo fare aiutandoci con la formula di fattorizzazione,

$$P(\text{♀}) = P(\text{♀}|B) \cdot P(B) + P(\text{♀}|B^c) \cdot P(B^c).$$

Anche in questo caso tutte le quantità sono note (e addirittura abbiamo già calcolato il primo prodotto), quindi abbiamo

$$P(\text{♀}) = 0.2193 + 0.23 \cdot 0.57 = 0.3504.$$

Mettendo assieme il tutto, abbiamo che quanto cerchiamo, cioè la probabilità che Giulietta sia del biennio, è

$$P(B|\text{♀}) = \frac{P(B \cap \text{♀})}{P(\text{♀})} = \frac{P(\text{♀}|B) \cdot P(B)}{P(\text{♀})} = \frac{0.2193}{0.3504} \approx 0.6259.$$

Nell'esempio precedente abbiamo fatto qualcosa di interessante, che va oltre la risoluzione del problema assegnato: abbiamo calcolato una probabilità condizionata in funzione della sua speculare, ossia $P(E|F)$ a partire da $P(F|E)$. Possiamo fare la stessa cosa in generale, come mostrato dal seguente risultato.

TEOREMA 3.20. (BAYES) *Sia $(\Omega, \mathcal{F}, P)$ uno spazio di probabilità e siano E, F due eventi, entrambi di probabilità non nulla. Allora*

$$P(E|F) = \frac{P(F|E)}{P(F)} \cdot P(E).$$

Dimostrazione. Dalla definizione di probabilità condizionata abbiamo la seguente catena di uguaglianze:

$$P(E|F) = \frac{P(E \cap F)}{P(F)} = \frac{P(E \cap F)}{P(E)} \cdot \frac{P(E)}{P(F)} = \frac{P(F|E) \cdot P(E)}{P(F)}. \qquad \square$$

Possiamo poi combinare il Teorema di Bayes con il teorema delle probabilità totali, ricavando il seguente risultato.

COROLLARIO 3.21. *Data una partizione di Ω in eventi disgiunti di probabilità non nulla, se F è un evento in $\mathcal{F}$, allora per ogni evento E*

$$P(E|F) = \frac{P(F|E) \cdot P(E)}{\sum_{i \in I} P(F|E_i) \cdot P(E_i)}. \tag{3.2}$$

Un trucco di pigrizia: se scegliamo la partizione in modo che E ne faccia parte, il prodotto al numeratore sulla destra compare anche nella somma a denominatore, quindi dobbiamo calcolare il valore di un addendo in meno.

Esempio 3.22. Un laboratorio propone un nuovo test per determinare la positività (o negatività) al virus SARS-CoV-2. La proporzione di infetti che risultano positivi al test (detta anche sensibilità) è il 99.9%, mentre la proporzione di sani che sono negativi al test (detta anche specificità) è il 99.7%. In Italia il virus contagia 5 persone su 1000. Jacopo si sottopone a questo test. Se il test è positivo, con che probabilità Jacopo è davvero infetto?

La prima tentazione è di rispondere 99.9%. Tuttavia, avendo visto che il condizionamento non è simmetrico, sappiamo distinguere tra $P(M|+)$ e $P(+|M)$, dove M è l'evento ``Jacopo è malato'' e $+$ l'evento ``Jacopo è positivo''. Il dato del problema sulla sensibilità è $P(+|M)$, mentre il problema ci chiede $P(M|+)$. Il Teorema di Bayes, però, ci suggerisce la strada da prendere:

$$\begin{aligned} P(M|+) &= \frac{P(+|M)\cdot P(M)}{P(+)} \\ &= \frac{P(+|M)\cdot P(M)}{P(+|M)\cdot P(M) + P(+|M^c)\cdot P(M^c)}, \end{aligned} \tag{3.3}$$

dove abbiamo usato anche l'identità (3.2) e la formula di fattorizzazione. Sostituiamo i valori disponibili, che abbiamo già come dati o che ricaviamo facilmente:

$$P(M) = 0.005, \quad P(M^c) = 1 - P(M) = 0.995, \quad P(+|M) = 0.999$$

e anche

$$P(+|M^c) = 1 - P(-|M^c) = 0.003.$$

Tornando alla (3.3), abbiamo allora

$$\begin{aligned} P(M|+) &= \frac{0.999\cdot 0.005}{0.999\cdot 0.005 + 0.003\cdot 0.995} \\ &= \frac{0.004995}{0.00798} \approx 63\%. \end{aligned}$$

Questa probabilità, per quanto non trascurabile, è comunque inferiore rispetto a quella che ci aveva tentato inizialmente.

Spendiamo due parole per spiegare, per quanto in modo non approfondito, il motivo di questa discrepanza. Concentriamoci su quello che sappiamo: Jacopo è positivo al test. Quando succede questo? Se una persona è veramente malata, nel 99.9% dei casi il test sarà positivo, tuttavia l'incidenza della malattia, ossia la proporzione di persone effettivamente malate, è molto piccola. Allo stesso tempo, raramente (nello 0.3% dei casi) il test segnalerà come positivo qualcuno che è sano. Tuttavia la proporzione di persone sane è molto alta, quindi tra i positivi al test i falsi positivi sono una parte non trascurabile: più di un terzo.

L'esempio precedente, oltre a essere un buon esercizio, ci mostra anche quanto sia importante il Teorema di Bayes nella vita reale. Il cervello umano non è portato intuitivamente al ragionamento probabilistico ed è quindi facile incappare in errori. Il Teorema di Bayes è uno degli strumenti che ci permettono di aggirare ed evitare questi errori. Una delle sue applicazioni, in sintonia con il metodo scientifico, consiste nello spingerci ad aggiornare le nostre convinzioni.

Cosa vogliamo dire con questo? Ci aspettiamo di fare ipotesi e metterle alla prova con opportuni esperimenti. Facciamo entrare in gioco anche la probabilità, usandola come misura del livello di convinzione nella nostra ipotesi.

Ad esempio, Maestra Rita potrebbe supporre che la probabilità che Pierino non abbia studiato la lezione sia del 70%. In questo caso il fenomeno d'interesse è “lo studio da parte degli scolari” (in particolare da parte di Pierino) e abbiamo come ipotesi “Pierino non ha studiato”. Maestra Rita non è sicura di questa ipotesi: Pierino potrebbe finalmente aver capito ed essersi messo sui libri, ma se la maestra dovesse scommettere darebbe fiducia a Pierino solo al 30%. Maestra Rita però può mettere alla prova la sua ipotesi con un esperimento: interrogando Pierino ha modo di verificare se abbia studiato o no.

Scriviamo queste cose con la notazione della probabilità: H è la nostra ipotesi, (Pierino non ha studiato) che supponiamo vera con probabilità $P(H)$ (70% nell'esempio). Con E indichiamo il risultato di un esperimento (Pierino non sa rispondere alla domanda).

Prima di effettuare l'esperimento, possiamo assegnare le probabilità relative all'esperimento: $P(E|H)$ nel caso in cui H sia vera e $P(E|H^c)$ nel caso in cui H sia falsa. Nel caso di Pierino, Maestra Rita stima che $P(E|H) = 90\%$: se Pierino non ha studiato è probabile che non sappia rispondere, ma potrebbe avere fortuna e azzeccare la risposta. Viceversa, valuta $P(E|H^c) = 5\%$: se Pierino ha studiato, potrebbe comunque non rispondere correttamente, per qualche motivo, anche se è poco probabile.

Assegniamo queste probabilità condizionate prima di vedere l'effettivo risultato dell'esperimento. Quando però sappiamo cosa è successo, possiamo usare gli ingredienti che abbiamo preparato per vedere come cambia la nostra confidenza nell'ipotesi dopo aver visto il verificarsi di E. In altre parole siamo interessati a $P(H|E)$: quanto è convinta Maestra Rita che Pierino non abbia studiato se non ha saputo rispondere alla domanda che gli ha fatto?

Per il Teorema di Bayes,

$$P(H|E) = \frac{P(E|H)}{P(E|H)\,P(H) + P(E|H^c)\,P(H^c)} \cdot P(H)$$

e Maestra Rita, che prima pensava che ci fosse un 30% di possibilità che Pierino per una volta avesse studiato la lezione, dopo la scena muta aggiorna questa sua convinzione,

$$\frac{90}{100} \cdot \frac{10000}{90 \cdot 70 + 5 \cdot 30} \cdot \frac{70}{100} \approx 97.7\%$$

e ha quasi la certezza che Pierino non si sia preparato.

Tornando al caso generale, $P(H)$ è la probabilità che diamo alla verità di H prima di effettuare l'esperimento e prende quindi il nome di *probabilità a priori* (o *prior*). D'altra parte, $P(H|E)$ è la probabilità di H aggiornata dopo aver visto il risultato E dell'esperimento: prende il nome di *probabilità a posteriori* o *posterior*.

È allora più chiaro il parallelo col ragionamento scientifico. Nello studiare un fenomeno, facciamo un'ipotesi H di cui siamo convinti a un livello $P(H)$, per precedenti osservazioni o per altri motivi. Pianifichiamo un esperimento e, prima di effettuarlo, valutiamo con cura quali sono i possibili risultati e quanto li riteniamo plausibili in un mondo in cui H è vera e in uno in cui H è falsa, dando dei valori a $P(E|H)$ e $P(E|H^c)$, rispettivamente, per ogni possibile esito E dell'esperimento. A questo punto facciamo l'esperimento e ne osserviamo il risultato E. Possiamo poi aggiornare la nostra convinzione che H sia vera, con l'informazione in più raccolta con l'esperimento, calcolando $P(H|E)$ col Teorema di Bayes.

Nell'Esempio 3.22, prima di sottoporsi al test, Jacopo poteva stimare la probabilità di essere malato allo 0.5%; dopo il risultato positivo del test, rivaluta questa probabilità al 63%. Il modo in cui ha scelto la sua probabilità a priori di essere malato è di considerarsi un individuo qualunque della popolazione, all'interno della quale l'incidenza è 5 su 1000. Chiaramente altri fattori sarebbero potuti entrare in gioco: ad esempio se avesse avuto sintomi, magari avrebbe valutato diversamente la probabilità a priori.

Ci sono pochi vincoli sulla prior: deve essere una probabilità, quindi soddisfare le proprietà che ormai conosciamo (in particolare quella di monotonia). In più, se vogliamo poter usare il Teorema di Bayes in modo fruttuoso, non possiamo assegnare mai le probabilità 0 e 1.

Infatti se assegniamo a un evento probabilità 1, diciamo $P(H) = 1$, per quanti esperimenti contrari facciamo non potremo mai discostarci da quel valore:

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E|H)\,P(H) + P(E|H^c)\,P(H^c)} = \frac{P(E|H)}{P(E|H)} = 1$$

e analogamente per il caso $P(H) = 0$.

Questo ha senso, da un punto di vista astratto: se siamo certi di qualcosa nulla ci farà cambiare idea. In generale, però, quando dichiariamo di essere certi di qualcosa in un contesto sperimentale, questo significa che per cambiare idea avremo bisogno di una notevole quantità di evidenza contraria alla nostra convinzione precedente. Questo almeno finché vogliamo agire in modo razionale. I valori 0 e 1 sono quindi da evitare.

Quanto appena detto vale anche per i risultati degli esperimenti: anche se possono sembrare controesempi alla nostra ipotesi, dobbiamo tenerci un po' di margine (da valutare) che tenga conto di possibili errori nell'esperimento, ad esempio una lettura sbagliata da parte dello strumento. Quindi non raggiungeremo mai certezze: per la gioia degli scienziati sperimentali possiamo continuare a fare esperimenti all'infinito!

3.1.1. Esperimenti ripetuti (divagazione)

Se continuiamo a fare esperimenti, vorremo combinare i risultati osservati in ciascuno di essi per aggiornare la nostra $P(H)$. Come primo passo, vediamo il caso di due esperimenti: abbiamo due esiti E_1 ed E_2 e vogliamo capire quanto vale $P(H|E_1 \cap E_2)$, la probabilità a posteriori della nostra ipotesi dopo entrambi gli esperimenti. Facciamo un esempio.

Esempio 3.23. Torniamo al caso di Jacopo, incontrato all'Esempio 3.22. Se Jacopo si sottoponesse di nuovo al test e questo risultasse nuovamente positivo, quale sarebbe la probabilità che sia effettivamente malato?

L'impostazione del problema è simile a quella dell'Esempio 3.22, solo che ora abbiamo due eventi rispetto ai quali condizioniamo. Allora

$$\begin{aligned} P(M|+_2 \cap +_1) &= \frac{P(+_2|M \cap +_1)P(M|+_1)}{P(+_2|M \cap +_1)P(M|+_1) + P(+_2|M^c \cap +_1)P(M^c|+_1)} \\ &= \frac{P(+_2|M \cap +_1)}{P(+_2|M \cap +_1)P(M|+_1) + P(+_2|M^c \cap +_1)P(M^c|+_1)} \cdot \\ &\quad \cdot \frac{P(+_1|M)P(M)}{P(+_1|M)P(M) + P(+_1|M^c)P(M^c)}, \end{aligned} \tag{3.4}$$

in cui abbiamo messo in evidenza una specie di iterazione. Tuttavia non è facile semplificare ulteriormente questa espressione, a meno di non fare alcune ipotesi di indipendenza tra i due test. In particolare, prendiamo come ipotesi il fatto che i due test, ossia gli eventi $+_1$ e $+_2$, siano indipendenti *condizionatamente a* M ed M^c. In altre parole, quando sappiamo se Jacopo è malato o no, la positività dei due test è indipendente[3.4].

Se ora torniamo alla (3.4) abbiamo, sfruttando l'indipendenza condizionata,

$$P(M|+_1 \cap +_2) = \frac{P(+_2|M)}{P(+_2|M)P(M|+_1) + P(+_2|M^c)P(M^c|+_1)} \cdot \frac{P(+_1|M)P(M)}{P(+_1|M)P(M) + P(+_1|M^c)P(M^c)}$$

in cui il secondo fattore nel membro di destra è esattamente $P(M|+_1)$. In sostanza stiamo facendo esattamente la medesima cosa vista all'Esempio 3.22, solo che al secondo passaggio è cambiata la probabilità di partenza: non è più $P(M)$, bensì $P(M|+_1)$, perché dobbiamo tenere conto del primo test fatto.

Possiamo a questo punto sostituire nell'espressione i valori che conosciamo e ricavare la probabilità cercata: $P(M|+_1 \cap +_2) \approx 99.8\%$. Avere un secondo test positivo ci ha portati (quasi) alla certezza, nonostante quanto osservato prima (Esempio 3.22) sul fatto che in prima battuta i falsi positivi non sono trascurabili.

Può essere interessante notare, a margine di questo esempio, cosa succederebbe qualora il secondo test fosse negativo. La risposta di pancia potrebbe essere che il test positivo e quello negativo si ``annullano'' a vicenda, quindi che la probabilità che Jacopo sia malato ritorni a essere il valore di base 0.005. Le cose però non stanno proprio così: abbiamo

$$P(M|+_1 \cap -_2) = \frac{P(-_2|M)}{P(-_2|M)P(M|+_1) + P(-_2|M^c)P(M^c|+_1)} \cdot \\ \cdot \frac{P(+_1|M)P(M)}{P(+_1|M)P(M) + P(+_1|M^c)P(M^c)}$$

e, sostituendo i valori che conosciamo, otteniamo $P(M|+_1 \cap -_2) \approx 0.002$. Come mai? Il motivo è questo: se da un lato i falsi positivi non sono infrequenti, i falsi negativi lo sono molto meno, dato che complessivamente gli infetti sono una piccola parte della popolazione.

Ispirati dall'Esempio 3.23, possiamo fare alcune osservazioni generali.

3.4. Questo non significa che i due test siano indipendenti tra loro, nonostante stiamo considerandoli indipendenti se condizionati a un evento e al suo complementare, come abbiamo visto nell'Esempio 3.17.

Osservazione 3.24. Ripetere un esperimento non è inutile, nemmeno se dà nuovamente il medesimo risultato: la nostra valutazione della probabilità cambierà ulteriormente dopo la seconda osservazione. Lo stesso vale per due esperimenti con risultato opposto: in generale vederne i risultati non ci riporta al punto di partenza, ma ci lascia comunque delle informazioni aggiuntive, codificate dentro la probabilità.

Osservazione 3.25. La grandezza dell'effetto delle due osservazioni sulla probabilità non è la stessa. Nell'esempio, il risultato del primo test porta la probabilità di malattia da 0.005 a 0.63, con un aumento di 0.625. Il secondo esperimento la fa crescere ``solo'' di 0.368. Questo potrebbe sorprenderci: la seconda osservazione non è in sé diversa dalla prima. Il fenomeno, che prende il nome di *diminuzione dei ritorni marginali*, è del tutto naturale: ogni successivo esperimento con il medesimo risultato ha un impatto sempre minore sulla probabilità. Può sembrare contro-intuitivo, ma ciò accade solo perché le informazioni che raccogliamo interagiscono con la probabilità attraverso una moltiplicazione e non una somma, come il nostro cervello preferirebbe. Possiamo vedere una traccia di questo comportamento moltiplicativo nell'identità (3.4).

CAPITOLO 4

COSTRUIRE PROBABILITÀ

4.1. SPAZI FINITI O NUMERABILI

Cominciamo esaminando un caso semplice. Supponiamo di aver individuato lo spazio degli esiti Ω e di aver visto che esso è un insieme finito o numerabile. Come prima cosa prendiamo $\mathcal{F} = \mathcal{P}(\Omega)$, cioè l'insieme delle parti di Ω. In altre parole vogliamo che tutti i possibili sottoinsiemi di Ω siano eventi. Se siamo alla ricerca di un algoritmo generale, questa è una buona idea, perché qualunque insieme ci capiti di avere in Ω, esso potrà avere una probabilità.

Come abbiamo visto, la cardinalità di $\mathcal{F}$ è $2^{\#\Omega}$. Nel caso finito non è un grave problema, ma nel caso numerabile dovremo andare ad assegnare una probabilità a tanti eventi quanti sono i numeri reali (non solo infiniti, ma più che numerabili). Questo può sembrare un problema, dal momento che non lo possiamo fare ricorsivamente, a differenza di quanto accade nel caso di una quantità numerabile di oggetti.

Ma proprio qui sta il trucco: andiamo ad assegnare una probabilità a ciascun singoletto in Ω, in modo che per ogni $\omega \in \Omega$, $P(\{\omega\}) \geqslant 0$ e $\sum_{\omega\in\Omega} P(\{\omega\}) = 1$.

In pratica quello che facciamo è scegliere una funzione che soddisfi queste due proprietà, più eventuali altre condizioni imposte dal problema specifico, e a questo punto siamo a posto. Infatti per ogni $E \in \mathcal{F}$

$$P(E) := \sum_{\omega\in E} P(\{\omega\}),$$

dove abbiamo usato una proprietà che volevamo soddisfare, ossia che la probabilità di un'unione disgiunta sia la somma delle probabilità. Inoltre possiamo osservare che la somma si svolge su una quantità di indici al più numerabile, quindi non stiamo commettendo alcun abuso di notazione[4.1].

La difficoltà più grande in questo caso è individuare una funzione P definita su Ω che soddisfi le due proprietà enunciate sopra, cioè la non negatività e la somma a 1, e che al contempo catturi le proprietà del particolare problema che stiamo considerando.

Esempio 4.1. Tre amici si sfidano abitualmente nella corsa, sempre sullo stesso percorso. Prisca arriva per prima il doppio delle volte di Carlo, Daniele arriva primo la metà delle volte di Carlo. Qual è la probabilità che, in un giorno qualunque, Carlo sia il più veloce?

Indichiamo con d la frequenza con cui Daniele vince. Dai dati del problema sappiamo che Carlo vince con frequenza $2d$ e Prisca con frequenza $2 \cdot 2d = 4d$. Sappiamo anche che, dal momento che i concorrenti sono solo loro tre, $1 = 4d + 2d + d = 7d$, cioè Carlo arriva primo con probabilità $\frac{2}{7}$.

Non sempre, però, abbiamo le informazioni per dare una probabilità esplicita a ogni esito, come vedremo nel prossimo esempio. In questo caso abbiamo due possibilità: accontentarci di assegnare la probabilità solo su una tribù di eventi, oppure cambiare l'insieme Ω in modo che gli "eventi indivisibili" diventino esiti nella nuova rappresentazione.

4.1. Un vero abuso che si vede spesso è il seguente: si lasciano cadere le parentesi graffe e si identifica il singoletto di ω, un evento, con ω stesso, un esito. Anche se il desiderio di alleggerire la notazione è condivisibile, si tratta di una scelta pericolosa, perché genera ambiguità.

Esempio 4.2. Sull'isola dei matematici applicati c'è una particolare lotteria, in cui viene estratto un numero naturale a caso. Tuttavia, non tutti i numeri hanno la medesima probabilità di uscire: ciascun numero pari ha la stessa probabilità di uscire, il 7 esce con probabilità $\frac{1}{2}$, l'evento $\{1,2,3,5\}$ ha probabilità $\frac{1}{3}$, mentre gli eventi $\{9\}$, $\{9,11\}$ e $\{n : n \geqslant 9\}$ hanno la stessa probabilità.

Possiamo iniziare osservando che i numeri pari possono avere solamente probabilità 0: se così non fosse, avremmo una probabilità totale maggiore di 1, dal momento che i numeri naturali soddisfano la proprietà archimedea. Questo ci dice anche che $P(\{1,2,3,5\}) = P(\{1,3,5\}) = \frac{1}{3}$. Con le stesse idee possiamo anche mostrare che ogni numero naturale strettamente maggiore di 9 ha probabilità 0. A questo punto sappiamo che

$$1 = P(\Omega) = P(\{1,3,5\}) + P(7) + P(9) + P(\{0,2,4,6,8\}) + P(\{n : n > 9\}),$$

quindi $P(\{9\}) = \frac{1}{6}$. Osserviamo che, con i dati forniti, non siamo in grado di dire quali siano le probabilità degli eventi $\{1\}$, $\{3\}$, $\{5\}$, $\{1,3\}$, $\{1,5\}$, $\{1,7\}$... Possiamo considerare solo eventi in cui $\{1,3,5\}$ sia un blocco unico.

Un modo per ricondursi a quanto visto prima è scegliere un Ω diverso. In questo caso prendiamo, per esempio, Ω che ha per elementi l'insieme dei naturali pari, l'insieme dei naturali dispari maggiori di 5 e l'insieme $\{1,3,5\}$.

Nel caso numerabile, dato che ci sono infiniti singoletti, potremmo aspettarci che un numero infinito di essi dovrà necessariamente avere probabilità zero. Questo è falso, come possiamo vedere nel seguente esempio.

Esempio 4.3. Anche sull'isola dei matematici puri c'è una lotteria infinita, su tutti i numeri naturali, in cui ogni numero ha il doppio della probabilità di essere estratto rispetto al suo successore.

In questo caso abbiamo bisogno di sfruttare la serie geometrica. Sappiamo infatti che, posta z la probabilità di estrarre 0, la probabilità di estrarre n è $2^{-n} \cdot z$, ma anche che

$$1 = \sum_{n=0}^{+\infty} 2^{-n} \cdot z = z \cdot \sum_{n=0}^{+\infty} 2^{-n} = z \cdot 2,$$

da cui abbiamo che lo zero esce con probabilità $\frac{1}{2}$ e che in generale un numero naturale n esce con probabilità $2^{-(n+1)}$. In particolare, nessun numero naturale ha probabilità 0 di uscire.

4.2. LO SPAZIO DEI NUMERI REALI

Consideriamo ora il caso in cui Ω è l'intervallo di numeri reali $[0,1]$. Dobbiamo scegliere la tribù e valutare come definire una misura di probabilità. Cominciamo dalla tribù.

Come già accennato in precedenza, potremmo prendere come tribù l'insieme delle parti di $[0,1]$, ma questo ha cardinalità pari all'insieme potenza di $\mathbb{R}$, cioè $2^{(2^{\aleph_0})}$, che è un po' grande per i nostri gusti, visto che poi a ogni elemento della tribù andrà assegnata una probabilità[4.2]. Consideriamo insiemi di numeri reali. Quelli che ci possono venire in mente di solito[4.3] sono punti singoli, segmenti, semirette e loro combinazioni (unioni finite o numerabili, differenze e così via). Dato che per il momento ci stiamo interessando solamente all'intervallo $[0,1]$, intersecheremo quest'ultimo con gli insiemi visti sopra. Dentro alla nostra tribù dovranno esserci insiemi di questo tipo, perché è di questi che vogliamo calcolare la probabilità.

4.2. Ci sono anche altri motivi per non scegliere l'insieme delle parti: non possiamo farlo, se vogliamo definire una probabilità che soddisfi alcune ragionevoli condizioni. Discutere di questo, però, ci porterebbe un po' troppo fuori strada, verso la teoria della misura.

4.3. Qualcuno potrebbe pensare immediatamente a casi patologici come l'insieme di Vitali (Giuseppe Vitali, 1875 – 1932).

In altre parole, vogliamo la tribù generata da punti isolati, intervalli (aperti, chiusi, semiaperti a destra e a sinistra) e loro unioni numerabili, cioè la più piccola tribù che contiene tutti questi insiemi. Con un po' di teoria degli insiemi possiamo osservare che a partire dai soli intervalli chiusi in $[0,1]$ possiamo ottenere, attraverso il passaggio al complementare e all'unione numerabile:

- gli intervalli aperti (a,b), definendo per ogni $n \in \mathbb{N} \setminus \{0\}$, $I_n = \left[a + \frac{1}{n}, b - \frac{1}{n}\right]$, intervallo chiuso, e prendendone l'unione $\bigcup_{n\in\mathbb{N}} I_n = (a,b)$;
- gli intervalli semiaperti della forma $[a,b)$ e $(a,b]$, in maniera analoga;
- i singoletti;
- le intersezioni...

Quindi se vogliamo avere una tribù che contenga tutti questi insiemi, possiamo generarla a partire dai soli intervalli chiusi, dal momento che unioni numerabili e complementari di elementi di una tribù sono essi stessi nella tribù.

In realtà è possibile usare come generatori gli intervalli semiaperti a sinistra, ossia della forma $(a,b]$. Come vedremo tra poco, questo modo di procedere è anche più comodo. In modo analogo a quanto fatto sopra, a partire dagli intervalli semiaperti a sinistra possiamo ottenere (con unioni numerabili e passaggi al complementare) gli intervalli chiusi e, di conseguenza, tutti gli altri insiemi che ci interessano.

Insomma, sia usando gli intervalli chiusi, sia usando gli intervalli semichiusi a destra (cioè semiaperti a sinistra), generiamo una tribù che soddisfa le nostre richieste, poiché contiene gli insiemi che consideriamo interessanti. Essa prende il nome di *tribù dei Boreliani*[4.4] (su $[0,1]$) e viene indicata con $\mathcal{B}([0,1])$. La sua cardinalità è quella del continuo, cosa che non dimostreremo qui (si fa per induzione transfinita).

Ora che abbiamo Ω e $\mathcal{F}$, dobbiamo solo scegliere una misura di probabilità. Anche in questo caso, come in quello degli spazi finiti o numerabili, non esiste un'unica scelta: il modo in cui definiamo la probabilità dipende dal problema che stiamo considerando. Tuttavia possiamo stabilire una procedura per definire misure di probabilità valide: dal momento che dobbiamo assegnare una probabilità a ciascun evento, cioè a ciascun elemento della tribù dei Boreliani, cominciamo assegnando una probabilità a ciascun intervallo utilizzato per generare la tribù[4.5]. Vogliamo farlo in modo che la probabilità dipenda solo dai due estremi dell'intervallo, senza dimenticare che anche le altre proprietà devono essere soddisfatte.

Esempio 4.4. Una possibile scelta di misura di probabilità sull'intervallo unitario $[0,1]$ è la seguente: interpretiamo ogni intervallo $[a,b]$ contenuto in $[0,1]$ come un segmento e gli assegniamo come probabilità la sua lunghezza. Abbiamo allora $P([a,b]) = b-a$.

A partire da questo, possiamo calcolare le probabilità degli altri elementi della tribù, anche di forma diversa da $[a,b]$, come ad esempio $[a,b)$, sfruttando gli assiomi di Kolmogorov. Infatti, preso c tale che $b \leqslant c \leqslant 1$, abbiamo $[a,c] = [a,b) \cup [b,c]$, in cui l'unione è disgiunta. Allora

$$c - a = P([a,c]) = P([a,b)) + P([b,c]) = P([a,b)) + (c-b),$$

da cui $P([a,b)) = c-a-(c-b) = b-a$. In modo analogo possiamo calcolare la probabilità degli altri elementi della tribù, ad esempio quella dei singoletti.

Questa è *una* possibile scelta di probabilità sull'intervallo $[0,1]$, che prende anche il nome di *probabilità uniforme* o *misura di Lebesgue*[4.6], ma non è l'unica.

4.4. Émile Borel (1871 – 1956).

4.5. Il fatto che questo sia sufficiente a definire una probabilità su tutta la tribù dei Boreliani, anche se è intuitivo, non è un fatto banale. Esiste però un risultato, il Teorema di Carathéodory (Constantin Carathéodory, 1873 – 1950), che ce lo garantisce, ne vediamo l'enunciato poco oltre.

4.6. Henri Léon Lebesgue (1875 – 1941).

Se vogliamo che la probabilità di un intervallo dipenda solo dai suoi estremi, possiamo considerare una funzione $F\colon[0,1]\to\mathbb{R}$, tale che $P((a,b])=F(b)-F(a)$. Da questo punto di vista la probabilità nell'Esempio 4.4 è stata ottenuta scegliendo $F(x)=x$ (la funzione identità) nell'intervallo $[0,1]$. Chiaramente ci sono altre scelte possibili, come vedremo ora.

Esempio 4.5. Prendiamo la funzione $F\colon[0,1]\to\mathbb{R}$ definita da

$$F(x)=\begin{cases} x & 0\leqslant x<\dfrac{1}{2} \\ \dfrac{1}{2}(x+1) & \dfrac{1}{2}\leqslant x\leqslant 1 \end{cases}.$$

La probabilità definita sulla tribù dei Boreliani a partire da $P((a,b])=F(b)-F(a)$ non è più quella uniforme vista nell'Esempio 4.4. Per certi intervalli (e quindi per certi eventi) le due probabilità coincidono, ma possiamo vedere facilmente che su alcuni intervalli, come ad esempio $\left(\frac{1}{4},\frac{3}{4}\right]$, esse assumono valori diversi. Inoltre, in questo caso, non è sempre vero che $P((a,b])=P((a,b))$. Infatti $P\left(\left(\frac{1}{4},\frac{1}{2}\right]\right)=\frac{3}{4}-\frac{1}{4}=\frac{1}{2}$, mentre

$$\begin{aligned} P\left(\left(\tfrac{1}{4},\tfrac{1}{2}\right)\right) &= P\left(\bigcup_{n\in\mathbb{N}}\left(\tfrac{1}{4},\tfrac{1}{2}-\tfrac{1}{n}\right]\right) \\ &= \lim_{n\to+\infty} P\left(\left(\tfrac{1}{4},\tfrac{1}{2}-\tfrac{1}{n}\right]\right) \\ &= \lim_{n\to+\infty} F\left(\tfrac{1}{2}-\tfrac{1}{n}\right)-F\left(\tfrac{1}{4}\right) \\ &= \tfrac{1}{2}-\tfrac{1}{4}=\tfrac{1}{4}, \end{aligned}$$

in cui abbiamo dovuto scomodare il passaggio al limite per n che tende a $+\infty$.

Di conseguenza abbiamo anche che $P\left(\left\{\frac{1}{2}\right\}\right)=P\left(\left(\frac{1}{4},\frac{1}{2}\right]\right)-P\left(\left(\frac{1}{4},\frac{1}{2}\right)\right)=\frac{1}{2}-\frac{1}{4}=\frac{1}{4}$, mentre si può verificare (con l'ausilio dei limiti) che per ogni $x\neq\frac{1}{2}$ in $[0,1]$, $P(\{x\})=0$.

Non tutte le funzioni F vanno bene, però: non dobbiamo dimenticare che stiamo cercando delle probabilità, quindi gli assiomi dovranno essere soddisfatti. Abbiamo visto, nell'Esempio 4.5, che F non deve necessariamente essere continua. Tuttavia deve essere monotona debolmente crescente, poiché per $0\leqslant a<b<c\leqslant 1$,

$$F(b)-F(a)=P((a,b])\leqslant P((a,c])=F(c)-F(a),$$

per la Proposizione 2.28, quindi $F(c)\geqslant F(b)$. Questo ancora non basta: per vedere le altre proprietà di queste funzioni, conviene però passare al caso in cui Ω è l'intera retta reale e considerare $[0,1]$ come un caso speciale.

Se vogliamo lavorare sull'intera retta dei numeri reali $\mathbb{R}$, dobbiamo come prima cosa definire nuovamente la tribù che consideriamo. I Boreliani su $[0,1]$ non sono più sufficienti, ma basterà modificarli un po' per estenderli a tutto $\mathbb{R}$.

Quali sono queste modifiche? Per comodità, al posto dei segmenti prenderemo le semirette come mattoni base della nostra costruzione. In particolare, sostituiamo gli intervalli semiaperti $(a,b]$ con le semirette sinistre chiuse, cioè della forma $(-\infty,b]$, con $b\in\mathbb{R}$ (e non più limitato al solo intervallo $[0,1]$).

A partire da queste semirette possiamo generare, con le solite operazioni di unione numerabile e passaggio al complementare, gli intervalli (aperti, chiusi e semiaperti), i singoletti, le semirette sinistre aperte e le semirette destre aperte e chiuse, nonché tutte le loro unioni: abbiamo quindi dei buoni generatori. La tribù generata dalle semirette sinistre chiuse prende il nome di *tribù dei Boreliani* (su $\mathbb{R}$) e viene indicata con $\mathcal{B}(\mathbb{R})$ (o brevemente con $\mathcal{B}$)[4.7]. La sua cardinalità è anche in questo caso quella del continuo.

4.7. Si può ottenere la stessa tribù anche usando altri generatori, ma come vedremo questa scelta è particolarmente comoda per definire le probabilità.

Per definire una probabilità sullo spazio probabilizzabile $(\mathbb{R}, \mathcal{B})$, sfruttiamo la medesima idea vista per l'intervallo unitario: la faremo dipendere solamente dagli estremi. In questo caso però abbiamo un solo estremo "agibile": il secondo. Allora definiamo la probabilità della semiretta $(-\infty, b]$ come funzione del solo estremo b, mediante un'opportuna funzione F definita su tutti i reali: $P((-\infty, b]) = F(b)$. Questo è del tutto compatibile con quanto visto prima: per differenza di insiemi abbiamo infatti che $P((a,b]) = P((-\infty, b]) - P((-\infty, a]) = F(b) - F(a)$.

Non tutte le funzioni F vanno bene, tuttavia. Abbiamo già visto che F deve essere monotona non decrescente, ma ora non possiamo più avere come probabilità la lunghezza dei segmenti, ossia F uguale all'identità: dal momento che le semirette hanno lunghezza infinita, non potremmo più rispettare gli assiomi di Kolmogorov[4.8].

Abbiamo però una buona caratterizzazione delle funzioni ammesse: sono quelle funzioni $F\colon \mathbb{R} \to \mathbb{R}$ tali che

- F è non decrescente (o debolmente crescente);
- esiste il limite di $F(x)$ per x che tende a $+\infty$ e vale $\lim_{x \to +\infty} F(x) = 1$;
- esiste il limite di $F(x)$ per x che tende a $-\infty$ e vale $\lim_{x \to -\infty} F(x) = 0$;
- in ogni punto x_0 la funzione F è continua a destra, cioè $\lim_{x \to x_0^+} F(x) = F(x_0)$ e limitata a sinistra, ossia $\lim_{x \to x_0^-} F(x) \leqslant F(x_0)$.

La prima di queste proprietà ci è familiare e segue dalla monotonia della probabilità. Le due successive seguono dal fatto che $P(\Omega) = P(\mathbb{R}) = 1$. L'ultima proprietà (o meglio, le due proprietà all'ultimo punto) possono apparire più sorprendenti. In realtà servono per darci la possibilità di assegnare a un punto una probabilità diversa da 0:

$$\begin{aligned} P(\{x_0\}) &= P((-\infty, x_0]) - P\left(\bigcup_{n \in \mathbb{N}} \left(-\infty, x_0 - \tfrac{1}{n}\right] \right) \\ &= F(x_0) - \lim_{n \to +\infty} F\left(x_0 - \frac{1}{n}\right) \\ &= F(x_0) - \lim_{x \to x_0^-} F(x). \end{aligned}$$

Allo stesso tempo ci garantiscono che la probabilità si comporta bene anche in tali punti e, in particolare,

$$\lim_{n \to +\infty} F\left(b + \frac{1}{n}\right) - F(a) = \lim_{n \to +\infty} P\left(\left(a, b + \frac{1}{n}\right]\right) = P((a,b]) = F(b) - F(a).$$

Insomma, ci basta definire una funzione F di questo tipo per avere una probabilità sulla retta reale[4.9].

Esempio 4.6. Consideriamo la funzione $F\colon \mathbb{R} \to \mathbb{R}$ definita da

$$F(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x \geqslant 0 \end{cases}.$$

Questa funzione soddisfa le proprietà viste sopra (è addirittura continua in ogni punto), quindi definisce una probabilità. Possiamo in particolare vedere che ogni intervallo non vuoto nei reali positivi ha una probabilità strettamente positiva:

$$P((a,b]) = F(b) - F(a) = 1 - e^{-b} - 1 + e^{-a} = e^{-a} - e^{-b},$$

mentre ogni singoletto ha probabilità 0 (conseguenza del fatto che F è continua).

4.8. Non possiamo nemmeno prendere una funzione che sia proporzionale alla lunghezza dei segmenti, perché avremmo il medesimo problema.

4.9. Stiamo ancora imbrogliando, perché stiamo sfruttando in silenzio il Teorema di Carathéodory già nominato in precedenza.

Non possiamo davvero apprezzarlo qui, ma imparare a costruire una misura di probabilità (e quindi uno spazio di probabilità) sui reali è un ottimo investimento, se non addirittura il migliore che possiamo fare. È infatti possibile trasformare ogni esperimento aleatorio in uno equivalente in cui lo spazio probabilizzabile sia $(\mathbb{R}, \mathcal{B})$ e tutte le caratteristiche peculiari del problema siano codificate dalla probabilità P (cioè dalla funzione F, che come vedremo prende il nome di *funzione di ripartizione*). Questo è reso possibile dalla nozione di variabile aleatoria o casuale[4.10].

4.2.1. Il teorema di Carathéodory

Il Teorema di Carathéodory è lo strumento che permette di definire la probabilità su Ω dandone il valore su una piccola parte degli eventi e non su tutti quelli che stanno nella tribù. Per poter fare ciò, però, non possiamo prendere una famiglia qualunque di eventi, ma dobbiamo prenderne una abbastanza ricca. Una possibilità è quella di prendere un'algebra: se abbiamo una funzione su quest'algebra che si comporta come una probabilità, allora la possiamo estendere a una probabilità vera e propria definita sulla tribù generata da $\mathcal{A}$.

TEOREMA 4.7. (CARATHÉODORY) *Dati un insieme Ω, un'algebra $\mathcal{A}$ di sottoinsiemi di Ω e una funzione $P_0\colon \mathcal{A} \to [0,1]$ che ha le proprietà di una probabilità, allora esiste un'unica probabilità P su $(\Omega, \sigma(\mathcal{A}))$ che coincide con P_0 su $\mathcal{A}$.*

4.3. SPAZI PRODOTTO

In probabilità succede spesso che qualcosa possa essere visto come una combinazione di più fenomeni aleatori. Quando questi sono distinti e non si influenzano a vicenda (sono indipendenti, come visto nel Capitolo 3), possiamo descriverli tutti assieme come spazio prodotto, portandoci dietro quello che sappiamo sulle varie componenti. Per farci un'idea, vediamo qualche esempio.

Esempio 4.8. Se lanciamo un dado a 4 facce e una moneta, possiamo scrivere gli esiti come coppie ordinate in cui la prima componente è l'esito del lancio del dado e la seconda l'esito del lancio della moneta. In altre parole, $\Omega = \{(1,T), (2,T), (3,T), (4,T), (1,C), (2,C), (3,C), (4,C)\}$. Come insieme, questo è il prodotto cartesiano dei due insiemi universo $\Omega_1 = \{1,2,3,4\}$ e $\Omega_2 = \{T,C\}$, cioè $\Omega = \Omega_1 \times \Omega_2$. Se assumiamo che il dado e la moneta non si influenzino, possiamo definire una probabilità su questo spazio a partire dalle probabilità del dado e della moneta, ovviamente su un'opportuna tribù.

Esempio 4.9. Prendiamo ora n monete tutte uguali tra loro e lanciamole (o in alternativa prendiamo una sola moneta e lanciamola n volte). In questo caso uno spazio naturale per descrivere il fenomeno è quello delle n-uple ordinate di elementi di $\Omega_1 = \{T,C\}$, cioè $\Omega = (\Omega_1)^n$. E se pensassimo di lanciare la moneta infinite volte? Avremmo che un esito è una successione di elementi di Ω_1, cioè avremmo $\Omega = (\Omega_1)^{\mathbb{N}}$. In entrambi i casi, però, per poter calcolare probabilità di eventi abbiamo bisogno di definire una tribù $\mathcal{F}$ e una funzione di probabilità P. Nel secondo caso possiamo identificare Ω con $\{0,1\}^{\mathbb{N}}$, che a sua volta possiamo identificare coi numeri reali in $[0,1]$: potremmo allora usare quanto visto nella Sezione 4.2. Questo maschererebbe però la struttura di ``esperimento ripetuto'' che invece è più facilmente riconoscibile nella rappresentazione come prodotto (infinito).

Come primo caso, consideriamo il prodotto di due esperimenti aleatori descritti rispettivamente dagli spazi di probabilità $(\Omega_1, \mathcal{F}_1, P_1)$ e $(\Omega_2, \mathcal{F}_2, P_2)$. Vogliamo costruire uno spazio di probabilità $(\Omega, \mathcal{F}, P)$ che descriva la coppia di esperimenti. Iniziamo dallo spazio degli esiti: come abbiamo già detto nell'Esempio 4.8, è ragionevole prendere il prodotto cartesiano $\Omega = \Omega_1 \times \Omega_2$.

4.10. Incontreremo di nuovo le variabili aleatorie nel Capitolo 5.

Passiamo allora alla tribù: $\mathcal{F}$ sarà generata dai prodotti di elementi delle due tribù $\mathcal{F}_1$ e $\mathcal{F}_2$, quindi

$$\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\{E_1 \times E_2 : E_1 \in \mathcal{F}_1, E_2 \in \mathcal{F}_2\}),$$

cioè $\mathcal{F}$ è la tribù generata dai rettangoli in cui la prima coordinata è data dal primo esperimento e la seconda coordinata dal secondo. Non ci fermiamo alla famiglia dei rettangoli, ma ne prendiamo la tribù generata perché vogliamo essere sicuri di avere una famiglia di insiemi che sia una tribù. Usando l'analogia geometrica, vogliamo che nella tribù ci siano anche altre figure (triangoli, cerchi...), che costruiamo come unione numerabile di rettangoli (o complementari).

Come ultimo passo, dobbiamo parlare della probabilità P. Ancora una volta vogliamo mettere in evidenza che si tratta di una combinazione di esperimenti, quindi vorremmo che la proiezione su ogni coordinata fosse la probabilità del corrispondente esperimento singolo, cioè che la probabilità di ogni esperimento di Ω_1 fosse inalterata nel prodotto con Ω_2 e viceversa.

Possiamo ottenere una probabilità P con le proprietà richieste se la definiamo, per ogni rettangolo $E_1 \times E_2$ in $\mathcal{F} = \mathcal{F}_1 \otimes \mathcal{F}_2$, come

$$P(E_1 \times E_2) = P_1(E_1) \cdot P_2(E_2). \tag{4.1}$$

Questo giustifica anche la notazione $P = P_1 \otimes P_2$. Si può obiettare che la (4.1) non definisce da sola una probabilità per ogni elemento di $\mathcal{F}$ se, come abbiamo detto, non ogni elemento di $\mathcal{F}$ è un rettangolo. Tuttavia, potendo scrivere ogni elemento di $\mathcal{F}$ a partire da rettangoli, mediante unione e complementare, e sapendo come si comporta la probabilità rispetto all'unione (disgiunta) e al complementare, possiamo limitarci a definirla sui rettangoli e l'estensione sarà unica[4.11].

Esempio 4.10. Tornando all'Esempio 4.8, osserviamo che nella tribù $\mathcal{F}$ non ci sono solo i prodotti di elementi delle due tribù $\mathcal{F}_1$ e $\mathcal{F}_2$: infatti il complementare di $\{1\} \times \{C\}$ non può essere scritto come prodotto (in particolare non è $\{2,3,4\} \times \{T\}$, che non contiene la coppia $(1,T)$, che appartiene al complementare di $\{1\} \times \{C\}$). Dobbiamo prendere la tribù generata, che contiene anche tutti i complementari e le unioni di rettangoli (cioè di prodotti di elementi di $\mathcal{F}_1$ e $\mathcal{F}_2$).

La probabilità dell'evento $\{(1,T)\}$, supponendo il dado equilibrato e la moneta non truccata, sarà $P(\{(1,T)\}) = P_1(\{1\}) \cdot P_2(\{T\}) = \frac{1}{8}$, mentre quella dell'evento $(\{1,2\} \times \{C\})^c$ sarà

$$P((\{1,2\} \times \{C\})^c) = 1 - P(\{1,2\} \times \{C\}) = 1 - P_1(\{1,2\}) \cdot P_2(\{C\}) = 1 - \frac{1}{4} = \frac{3}{4}.$$

In modo analogo possiamo fare per un numero finito di esperimenti distinti quello che abbiamo mostrato per due esperimenti. Possiamo anche passare a una quantità numerabile, ma vedremo i dettagli solamente in un caso speciale: quello degli esperimenti ripetuti.

Parliamo di esperimenti ripetuti quando tutti gli esperimenti sono copie identiche del medesimo esperimento, cioè possono essere tutti descritti con lo stesso spazio di probabilità $(\Omega_S, \mathcal{F}_S, P_S)$. Ne abbiamo visti due nell'Esempio 4.9: il lancio di n monete uguali o quello di infinite (numerabili) monete uguali.

Nel caso di un numero finito di ripetizioni, abbiamo una versione semplificata di quanto visto per il prodotto di esperimenti qualunque: abbiamo infatti (considerando ad esempio due sole ripetizioni) che $\Omega = \Omega_1 \times \Omega_2 = \Omega_S^2$, perché i due spazi dei singoli elementi coincidono; abbiamo inoltre che la tribù

$$\mathcal{F} = \mathcal{F}_S \otimes \mathcal{F}_S = \mathcal{F}_S^2 = \sigma(\{E_1 \times E_2 : E_1 \in \mathcal{F}_S, E_2 \in \mathcal{F}_S\})$$

e che la probabilità $P = P_S^2$.

Esempio 4.11. Francesco lancia 6 volte una moneta truccata (o una volta sei monete truccate identiche tra loro), che dà testa con probabilità p e croce con probabilità $1-p$. Con che probabilità i primi due lanci sono entrambi testa? Con che probabilità i primi tre lanci non sono tutti uguali tra loro?

4.11. Ancora una volta stiamo facendo le cose più facili di quanto non siano in realtà: anche qui viene in aiuto il Teorema di Carathéodory, che garantisce che tale estensione è unica.

Come spazio Ω abbiamo $\{T,C\}^6$ o $\{0,1\}^6$. La tribù è quella generata dai rettangoli, mentre la probabilità su ciascuna componente vale 0 sull'insieme vuoto, p su $\{T\}$, $1-p$ su $\{C\}$ e 1 su $\Omega_S=\{T, C\}$. Il primo evento cui siamo interessati, ``i primi due lanci sono entrambi testa'', è $\{T\}\times\{T\}\times\Omega_S^4$, la cui probabilità è

$$P(\{T\}\times\{T\}\times\Omega_S^4)=P(\{T\})^2P(\Omega_S)^4=p^2 1^4=p^2.$$

Il secondo evento è un po' più complicato: lo possiamo scrivere come unione di rettangoli, oppure in modo più semplice come complementare di unione di rettangoli,

$$E=(\{T\}\times\{T\}\times\{T\}\times\Omega_S^3\cup\{C\}\times\{C\}\times\{C\}\times\Omega_S^3)^c.$$

Per quanto riguarda la probabilità abbiamo allora

$$\begin{aligned}P(E)&=1-P(\{T\}\times\{T\}\times\{T\}\times\Omega_S^3\cup\{C\}\times\{C\}\times\{C\}\times\Omega_S^3)\\&=1-P(\{T\}\times\{T\}\times\{T\}\times\Omega_S^3)-P(\{C\}\times\{C\}\times\{C\}\times\Omega_S^3)\\&=1-p^3-(1-p)^3\\&=3p-3p^2,\end{aligned}$$

dove nel secondo passaggio abbiamo usato che i due eventi $\{T\}\times\{T\}\times\{T\}\times\Omega_S^3$ e $\{C\}\times\{C\}\times\{C\}\times\Omega_S^3$ sono disgiunti, dal momento che le sestuple nei due insiemi hanno sicuramente le prime tre coordinate distinte e sono quindi diverse.

Passiamo al caso di infinite ripetizioni di uno stesso esperimento $(\Omega_S,\mathcal{F}_S,P_S)$: siamo alla ricerca di un unico spazio $(\Omega,\mathcal{F},P)$ che le descriva tutte assieme. Cominciamo come sempre dallo spazio Ω: esso sarà costituito da successioni di elementi di Ω_S, quindi $\Omega=\Omega_S^{\mathbb{N}}$. Fin qui nulla di difficile.

Ci dedichiamo ora alla tribù $\mathcal{F}$. Qui, almeno in apparenza, quando le ripetizioni sono infinite le cose si complicano: ci sono troppe componenti da controllare. Proviamo dunque a sfruttare le idee viste prima e a concentrarci solo sulla ricerca dei generatori della tribù. Non solo, cerchiamo anche di imparare da quanto visto nella Sezione 4.2 per $\mathbb{R}$.

Una cosa che possiamo fare è fissare un numero naturale n e mettere in un unico insieme tutti gli elementi $\omega\in\Omega$ che hanno in comune le prime n coordinate. Possiamo farlo per ogni numero naturale n, considerando per ciascun n tutte le possibili n-uple di elementi di Ω_S. Questi insiemi, al variare di n, prendono il nome di *n-cilindri*, perché come i cilindri geometrici sono caratterizzati dall'avere una sezione fissata (le prime n componenti).

Prendiamo allora la collezione $\mathcal{C}$ di tutti gli n-cilindri al variare di n: la chiamiamo *famiglia degli insiemi cilindrici*. Analogamente a quanto abbiamo visto per i rettangoli, la famiglia dei cilindri in generale non è una tribù. Possiamo però usarla per generarne una: $\mathcal{F}=\sigma(\mathcal{C})$, che è una tribù su $\Omega^{\mathbb{N}}$.

Avendo costruito spazio e tribù, non resta che l'ultimo passo, la probabilità. Per definirla usiamo la forma dei cilindri che generano la tribù $\mathcal{F}$ e il fatto che stiamo parlando di esperimenti ripetuti: su ogni cilindro definiamo la probabilità come il prodotto della probabilità P_S su ciascuna delle n componenti del cilindro e di fattori 1 per tutte le altre (in sostanza le stiamo ignorando). In questo modo abbiamo una probabilità che generalizza al caso infinito quanto già visto per il caso del prodotto finito: per una successione di eventi $E_i\in\mathcal{F}_S$ abbiamo che la probabilità dell'evento $\bigotimes_{i=1}^{+\infty}E_i\in\mathcal{F}$ è

$$P\left(\bigotimes_{i=1}^{+\infty}E_i\right)=\prod_{i=1}^{+\infty}P_S(E_i).$$

In realtà, stiamo tacendo molti dettagli: non abbiamo la pretesa di essere precisi e nemmeno lo spazio o i prerequisiti per poterlo fare, ma vogliamo solo farci un'idea. Per vedere a fondo tutti i dettagli, ancora una volta, è necessario prendere in mano un libro di testo avanzato o seguire un corso universitario di probabilità o di teoria della misura.

Un'ultima osservazione, prima di passare a qualche esempio: quello che abbiamo fatto per la ripetizione infinita di un esperimento può essere adattato al caso del prodotto infinito di esperimenti non necessariamente uguali tra loro. Anche in tal caso possiamo definire dei cilindri, in cui però le componenti devono essere "pescate" dagli spazi corrispondenti alla coordinata in questione. Questo appesantisce la notazione, ma non cambia la sostanza.

Esempio 4.12. Federico ha infinite monete identiche tra loro, ciascuna delle quali dà testa con probabilità p e croce con probabilità $1-p$. Come sempre possiamo pensare che in realtà ne abbia una sola e la lanci infinite volte. Con che probabilità Federico ottiene la prima testa al k-esimo lancio?

Osserviamo che in questo esempio non possiamo fissare a priori un numero massimo di lanci (o di monete), perché qualunque sia questo numero, potremmo avere croci in tutti questi lanci (improbabile, al crescere del numero dei lanci, ma mai con probabilità identicamente zero). Ha allora senso considerare una ripetizione infinita dell'esperimento ``lancio di una moneta''[4.12]. Qual è l'evento del quale vogliamo calcolare la probabilità? È un k-cilindro le cui prime $k-1$ componenti sono C e la cui k-esima componente è T. Delle successive non ci interessa. Sappiamo che i cilindri stanno nella tribù, dal momento che ne sono i generatori. La probabilità di questo cilindro è

$$P_S(\{\mathrm{C}\})^{k-1} P_S(\{\mathrm{T}\}) \prod_{i=k+1}^{+\infty} 1 = (1-p)^{k-1} p.$$

4.4. FARSI LE OSSA

Nelle sezioni precedenti abbiamo visto alcuni modi per costruire spazi di probabilità. Non è però garantito che siano i migliori per i particolari problemi che incontreremo, né che in ogni problema avremo tutte le informazioni necessarie per costruirli nei modi visti, pur avendo magari tutto quello che ci serve per arrivare a una soluzione.

Esempio 4.13. Supponiamo di avere un dado a 6 facce, di cui sappiamo che $P(1) = \frac{1}{6}$. Se volessimo procedere come visto nella Sezione 4.1 e ci concentrassimo su $\Omega = \{1,\dots,6\}$, dovremmo assegnare una probabilità a tutte le facce del dado. Però non possiamo farlo, perché non abbiamo alcuna informazione sulle altre facce. Potremmo assumere che il dado sia bilanciato e che quindi tutte le facce escano con la medesima probabilità, ma il risultato che otterremmo sarebbe vero solo se questa ipotesi fosse soddisfatta, cosa che non abbiamo la possibilità di controllare.

Alle volte la formulazione del problema ci suggerisce una costruzione diversa da quella standard. Come possiamo accorgercene? È un'attività creativa e non meccanica: dobbiamo fare apprendistato, il solo modo di allenare l'occhio e la mano è fare tanti esercizi. Dobbiamo cercare soluzioni diverse alle quali ispirarci in futuro per affrontare altri problemi, in particolare quelli in cui i metodi standard non funzioneranno. Per questo stesso motivo, uno dei metodi migliori per allenarsi a risolvere problemi è risolvere altri problemi e confrontare le proprie soluzioni con quelle altrui. Pólya[4.13] nel suo libro *Come risolvere i problemi di matematica* indica tra le diverse euristiche per trovare una soluzione a un problema quella di cercare un altro problema, analogo o simile, del quale ci sia nota una soluzione, per poi cercare di adattare quest'ultima al problema corrente. Parla però di euristiche, non di teoremi, perché questa somiglianza non è definita in modo formale, poiché esula dagli scopi del testo: sta a noi individuarla e sfruttarla.

4.12. Questo esperimento costituito da una ripetizione infinita del lancio di una moneta si chiama anche *processo (o schema) di Bernoulli* (Jakob Bernoulli 1654 – 1705). Il modello che descrive il primo istante di successo in un processo di Bernoulli si chiama, per alcuni, *geometrico*. Lo rivedremo più avanti, nel Capitolo 8.

4.13. György Pólya (1887 – 1985).

Nel caso dei problemi di probabilità, dobbiamo imparare a estrarre gli oggetti giusti dal testo che abbiamo: non solo la probabilità, ma prima di essa lo spazio degli esiti e la tribù. In particolare è un ottimo esercizio, soprattutto all'inizio, essere molto precisi (quasi noiosi) nello scrivere esplicitamente cosa scegliamo come spazio degli esiti e come tribù, perché quest'accortezza ci eviterà di prendere dei granchi, come ad esempio definire una probabilità su coppie ordinate, quando gli oggetti su cui stiamo lavorando magari sono coppie non ordinate.

Esempio 4.14. In una noiosa serata di lockdown, due amici si danno appuntamento su Zoom per passare assieme la serata. Per rendere più elettrizzante l'appuntamento, si mettono d'accordo nel modo seguente: ciascuno di loro si impegna a connettersi in un orario compreso tra le 22 e le 23 e a restare online in attesa per 5 minuti. Passati questi 5 minuti (o allo scoccare delle 23) si disconnetterà. Con che probabilità i due amici si incontreranno su Zoom?

La prima volta che si affronta un problema di questo tipo, la tentazione più forte è quella di discretizzare in minuti o secondi. In questo caso, però, il tempo va considerato una quantità continua. Concentriamoci allora su uno dei due amici. Con che probabilità arriverà nei primi dieci minuti dell'ora? Il segmento favorevole è lungo $\frac{1}{6}$ del segmento totale (10 minuti su 60), quindi la probabilità che arrivi in quei dieci minuti è proprio $\frac{1}{6}$. Analogamente, la probabilità che il secondo amico arrivi tra le 22.21 e le 22.41 è $\frac{1}{3}$, poiché c'è un intervallo lungo 20 (minuti) favorevole su un intervallo totale lungo 60 (minuti).

In questo ragionamento, però, stiamo considerando i due amici separatamente e stiamo trascurando il fatto che sono disposti ad aspettare. Se sapessimo che il primo amico arriva alle 22.13, allora la probabilità che i due si incontrino sarebbe uguale alla probabilità che il secondo arrivi nei 5 minuti precedenti alle 22.13 o nei 5 minuti successivi, cioè $\frac{1}{6}$.

È arrivato il momento di passare dai segmenti ai quadrati. Mettiamo sull'asse delle ascisse l'orario di arrivo del primo amico e su quello delle ordinate quello del secondo. Le coordinate interne al quadrato rappresentano le combinazioni di arrivi dei due amici. I due si incontrano se le due coordinate non differiscono più di 5. Ma geometricamente questo cosa significa?

Se arrivano insieme, si incontrano. Questi punti sono la diagonale del quadrato. Ma non sono, come detto, la loro unica possibilità di incontrarsi[4.14]. Possiamo spostarci orizzontalmente o verticalmente di 5 minuti, rispetto alla diagonale, cioè considerare la diagonale ingrassata, colorata in grigio chiaro nella Figura 4.1. Questa superficie rappresenta tutte le coppie di orario d'arrivo per cui i due amici si incontrano. Per calcolare la probabilità richiesta dobbiamo considerare il rapporto tra quest'area e quella totale, che rappresenta tutte le possibili coppie di tempi d'arrivo dei due amici. Questo rapporto è $\frac{11}{36}$.

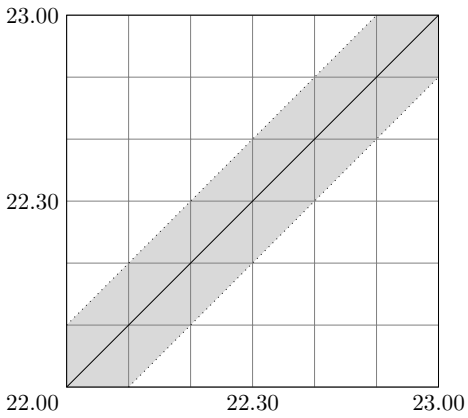


Figura 4.1. Incontro sotto il Grattacielo

4.14. In realtà la probabilità che arrivino insieme è 0, come possiamo vedere calcolando il rapporto tra l'area della diagonale (nulla) e quella del quadrato.

Possiamo esaminare questo stesso esercizio sotto la lente più formale introdotta nella prima parte di questo capitolo. Quello che cambia è solamente il linguaggio, non l'idea sottostante, né tanto meno il risultato. Giusto per dare uno spunto: abbiamo considerato per ciascuno dei due amici uno spazio di probabilità in cui $\Omega = [0,60]$ e $\mathcal{F} = \mathcal{B}([0,60])$ (cioè la tribù dei Boreliani generata dagli intervalli semiaperti in $[0,60]$, il che equivale a prendere la tribù dei Boreliani su $\mathbb{R}$ intersecata con l'intervallo che ci interessa). Per quanto riguarda P stiamo prendendo la lunghezza dei segmenti riscalata (in modo che $P([0,60]) = 1$), cioè $P([a,b]) = \frac{b-a}{60}$. Possiamo vedere la stessa probabilità come generata dalla funzione

$$F(x) = \begin{cases} 0 & x < 0 \\ \dfrac{x}{60} & 0 \leqslant x \leqslant 60 \\ 1 & x > 60 \end{cases}.$$

Quando poi passiamo a considerare insieme i due amici, siamo in uno spazio prodotto (in realtà il quadrato dello stesso spazio), con la misura prodotto, che è l'area delle porzioni del quadrato (il nostro Ω^2), rinormalizzata dividendo per $60 \cdot 60 = 3600$, in modo da avere una probabilità.

Osservazione 4.15. Un dettaglio interessante, anche se non necessario per il problema appena esaminato, è il seguente: possiamo calcolare la probabilità anche di eventi che nello spazio bidimensionale non sono rettangoli, ma che si ottengono come unione (eventualmente numerabile) di rettangoli, come ad esempio triangoli, poligoni, cerchi o altre figure convesse.

Esempio 4.16. Prendendo a caso due punti su un segmento, lo si divide in tre parti. Con che probabilità questi tre segmenti possono formare un triangolo?

Come prima cosa, fissiamo uguale a 1 la lunghezza del segmento iniziale[4.15]. Consideriamo questo segmento unitario in un sistema cartesiano dove l'origine coincide con il primo estremo e chiamiamo P e Q i due punti presi a caso su di esso. Ciascuno di essi è univocamente identificato dalla sua distanza dall'origine, che indichiamo rispettivamente con p e q.

Osserviamo che l'evento in cui P e Q coincidono (e dunque $p = q$) è un punto e ha quindi probabilità nulla[4.16] di accadere, per le proprietà della probabilità uniforme sul segmento $[0,1]$. Possiamo quindi trascurarlo e abbiamo così due casi possibili: $p < q$ e $p > q$. Data la simmetria del problema, possiamo studiare solamente uno dei due casi (a patto di ricordarcene nel momento in cui consideriamo Ω o di moltiplicare per 2 il risultato, se Ω non tiene conto della simmetria).

Se $p < q$, allora i tre segmenti hanno lunghezza p, $q-p$ e $1-q$ (quello che resta a destra del secondo punto). Affinché possano essere le lunghezze dei lati di un triangolo, devono soddisfare le disuguaglianze triangolari, cioè ogni lunghezza deve essere minore della somma delle altre due:

$$\begin{cases} p < q - p + 1 - q = 1 - p \\ q - p < p + 1 - q \\ 1 - q < p + q - p = q. \end{cases}$$

In ciascuna di queste disuguaglianze sommiamo a entrambi i membri quanto compare a primo membro, ottenendo

$$\begin{cases} 2p < 1 \\ 2(q-p) < 1 \\ 2(1-q) < 1, \end{cases}$$

da cui risulta che le tre lunghezze p, $q-p$ e $1-q$ devono tutte essere minori di $\frac{1}{2}$.

4.15. In sostanza stiamo assumendo come unità di misura ``la lunghezza di questo segmento''.

4.16. Abbiamo già visto un fenomeno simile nell'Esempio 4.14: in entrambi i casi abbiamo che, con una distribuzione uniforme di probabilità, oggetti geometrici di dimensione più bassa (come i punti in un segmento o i segmenti in una superficie) hanno misura nulla.

Anche in questo problema, come nel precedente, abbiamo però un continuo di valori possibili per p e q. Rappresentiamoli anche in questo caso in due dimensioni: siccome abbiamo assunto $p<q$, il nostro Ω sarà il solo triangolo sopra la diagonale, colorato in grigio chiaro. Quali sono in questo triangolo le coppie (p,q) che vanno bene? Cominciamo scartando tutti i punti in cui $p>\frac{1}{2}$ o $q<\frac{1}{2}$. Rimane solamente un vincolo da considerare: $q-p<\frac{1}{2}$, cioè scartiamo i punti che distano più di $\frac{1}{2}$ dalla diagonale. Rimane il triangolo colorato in grigio scuro nella Figura 4.2, che è rettangolo isoscele di cateto $\frac{1}{2}$ e area $\frac{1}{8}$. La probabilità cercata è il rapporto tra quest'area e l'area dell'intero triangolo sopra la diagonale, che vale $\frac{1}{2}$. La probabilità cercata è allora $\frac{1}{4}$.

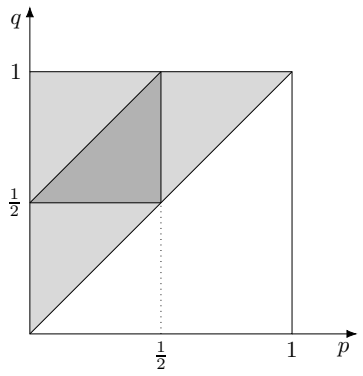


Figura 4.2. Spezzare un segmento per avere triangoli, prima soluzione

Per evitare di calcolare le aree, possiamo osservare che il triangolo più grande è diviso in quattro triangolini equivalenti dalle linee che abbiamo tracciato, di cui solo uno, quello più scuro, è costituito da punti che rappresentano casi favorevoli.

Se non siamo convinti che sia sufficiente considerare solo il caso $p<q$, possiamo guardare i casi favorevoli all'interno dell'intero quadrato, considerando anche il caso simmetrico in cui $p>q$.

Vediamo ora una seconda soluzione. Mettiamoci in un sistema di riferimento cartesiano tridimensionale, con i tre assi che rappresentano le lunghezze dei tre segmenti. Le terne possibili (cioè quelle nel primo ottante in cui la somma delle tre coordinate è uguale a 1) giacciono tutte su un piano e, in particolare, sulla superficie di un triangolo (all'interno del cubo unitario) di vertici $(1,0,0),(0,1,0),(0,0,1)$, colorato in grigio nella Figura 4.3 a sinistra.

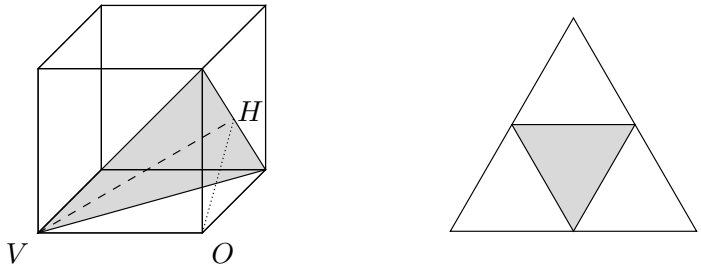


Figura 4.3. Spezzare un segmento per avere triangoli, seconda soluzione

Possiamo proiettare i tre assi cartesiani sul triangolo, in modo che siano le altezze rispetto ai tre lati (ad esempio, nella Figura 4.3 a sinistra proiettiamo l'asse OV sull'altezza VH): in altre parole, possiamo scegliere di rappresentare le lunghezze dei tre segmenti come le tre altezze del triangolo, che consideriamo dunque tutte di lunghezza 1. Il Teorema di Viviani[4.17] ci dice che i punti del triangolo sono precisamente i punti in cui la somma delle distanze dai lati è uguale alla lunghezza di una delle altezze e quindi a 1. I punti interni al triangolo sono quindi tutti i modi possibili in cui un segmento unitario può essere spezzato in tre parti, il nostro Ω.

4.17. Vincenzo Viviani (1622 -- 1703).

Ora dobbiamo rappresentare le altre condizioni, già viste nella prima soluzione: nessuno dei tre segmenti può essere più lungo di $\frac{1}{2}$. L'intersezione di queste tre condizioni è il triangolino centrale in Figura 4.3 a destra, la cui superficie è $\frac{1}{4}$ della superficie totale.

Qualche commento prima di proseguire. Nella seconda soluzione dell'Esempio 4.16 abbiamo trasformato il nostro problema di probabilità in un problema di geometria. E in un certo senso avevamo fatto la stessa cosa anche nella prima soluzione e in quella dell'Esempio 4.14. Ciò accade perché, se proviamo a suddividere la matematica in compartimenti stagni risolvendo problemi di probabilità o di geometria o di teoria dei numeri, la matematica si mostrerà comunque come un tutt'uno, in cui lo stesso problema può (e alle volte deve) essere affrontato con tecniche diverse che arrivano da ambiti apparentemente distinti.

Inoltre sembra che ci siano spesso più strade che portano al medesimo risultato. Anche questa è una verità più generale della matematica e non solo dei problemi di probabilità. Questa osservazione, però, ci suggerisce un buon esercizio: cercare nuove soluzioni a un problema già visto, che magari potranno tornare utili per altri problemi che incontreremo in futuro.

È importante che le diverse soluzioni portino davvero al medesimo risultato: guadagneremo così un po' di confidenza nella correttezza di quanto abbiamo ottenuto. Se i risultati invece saranno tutti diversi, avremo la certezza che almeno uno di quelli trovati sia sbagliato. Infatti in probabilità può capitare, se non si fa attenzione a scrivere tutto nei dettagli, di trovarsi in situazioni paradossali.

CAPITOLO 5

VARIABILI ALEATORIE

Lezione 7 Cominciamo con un esempio già visto.

Esempio 5.1. Lanciamo due dadi bilanciati a 6 facce e ne consideriamo la somma. Quali sono le probabilità dei vari risultati possibili della somma?

Abbiamo già visto questo esempio: avevamo scelto come spazio degli esiti l'insieme delle coppie ordinate in cui ciascuno degli elementi è un numero naturale compreso tra 1 e 6. Per i vari risultati della somma, che indichiamo con S abbiamo le seguenti probabilità:

S	elementi	P
$S=0$	$\emptyset$	0
$S=1$	$\emptyset$	0
$S=2$	$\{(1,1)\}$	$\frac{1}{36}$
$S=3$	$\{(1,2),(2,1)\}$	$\frac{2}{36}$
$S=4$	$\{(1,3),(2,2),(3,1)\}$	$\frac{3}{36}$
$S=5$	$\{(1,4),(2,3),(3,2),(4,1)\}$	$\frac{4}{36}$
$S=6$	$\{(1,5),(2,4),(3,3),(4,2),(5,1)\}$	$\frac{5}{36}$
$S=7$	$\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$	$\frac{6}{36}$
$S=8$	$\{(2,6),(3,5),(4,4),(5,3),(6,2)\}$	$\frac{5}{36}$
$S=9$	$\{(3,6),(4,5),(5,4),(6,3)\}$	$\frac{4}{36}$
$S=10$	$\{(4,6),(5,5),(6,4)\}$	$\frac{3}{36}$
$S=11$	$\{(5,6),(6,5)\}$	$\frac{2}{36}$
$S=12$	$\{(6,6)\}$	$\frac{1}{36}$
$S\geqslant 13$	$\emptyset$	0

Tabella 5.1. Somma di due dadi.

Ciascuna riga della tabella corrisponde a un evento: con $\{S=11\}$, $\{S=6\}$ stiamo indicando degli eventi. La scrittura $\{S=x\}$ un modo compatto per indicare quello che sta nella seconda colonna della tabella, che sarebbe l'evento vero e proprio, ossia il sottoinsieme di Ω. Però sottolinea un dettaglio importante: in questo caso non ci interessa il risultato dell'esperimento, ma una sua funzione. Ci è indifferente che la somma uguale a 4 sia stata ottenuta come $(1,3)$ o $(2,2)$. Se pensiamo al lavoro fatto per completare la tabella, molto di quel lavoro è stato inutile, abbiamo esplicitato molte informazioni che, per risolvere il problema in questione, non ci occorrono.

Continuiamo con un secondo esempio.

Esempio 5.2. Un'azienda produce calcolatori. Il costo di produzione di un singolo calcolatore è pari a 1000 €, mentre il prezzo di vendita è 2000 €. La probabilità che ci siano guasti irreparabili durante la produzione di un calcolatore è del 10%. I calcolatori con guasti non vengono venduti.

1. Con un ordine di un calcolatore, qual è la probabilità per l'azienda di avere un guadagno?
2. Con un ordine di tre calcolatori, qual è la probabilità per l'azienda di avere un guadagno?

Per avere un calcolatore funzionante, l'azienda deve continuare a produrne finché non ne esce uno senza guasti (e quindi vendibile). Con che probabilità accade questo?

Può succedere al primo tentativo: viene prodotto un solo calcolatore e questo non ha difetti. La probabilità che questo accada è $1-p=0.9$, dove con p indichiamo la probabilità che il calcolatore prodotto sia guasto.

Può succedere al secondo tentativo: viene prodotto un primo calcolatore, ma è guasto, dopodiché ne viene prodotto un secondo, funzionante. La probabilità che questo accada è $p\cdot(1-p)=0.09$. È importante l'ordine: se il primo fosse funzionante, non ci sarebbe bisogno di produrre il secondo. (Vale la pena osservare che stiamo assumendo l'indipendenza della presenza di errori tra calcolatori diversi.)

Può succedere al terzo tentativo, al quarto e così via. In generale la probabilità che il primo calcolatore vendibile sia prodotto all'n-simo tentativo è $p^{n-1}\cdot(1-p)$.

La domanda del problema, però, è molto specifica: richiede la probabilità che l'azienda abbia un guadagno, cosa che avviene se vengono prodotti n calcolatori con $2000-1000\,n>0$, ossia l'azienda guadagna solo se $n=1$, nel caso $n=2$ va in pari e per $n\geqslant 3$ va in perdita. Quindi la probabilità di guadagno è 0.9, quella di non perderci è 0.99 e quella di andare in rosso è $0.00\overline{9}=0.01$.

Non abbiamo scritto esplicitamente chi fosse Ω, ma possiamo usare $\Omega=\mathbb{N}\setminus\{0\}$ pensando come esiti il numero di calcolatori prodotti fino al primo calcolatore funzionante.

Passiamo ora alla seconda domanda: l'ordine è ora di 3 calcolatori. Abbiamo un successo quando sono stati prodotti 3 calcolatori funzionanti, quindi $\Omega=\mathbb{N}\setminus\{0,1,2\}$, perché l'azienda ne dovrà produrre almeno 3.

L'ordine può essere evaso non appena vengono prodotti 3 calcolatori, se tutti e tre sono funzionanti, cosa che avviene con probabilità $(1-p)^3=0.729$.

Alternativamente l'ordine può essere evaso con la produzione di 4 calcolatori, purché ce ne sia esattamente uno tra i primi tre che è guasto. L'ultimo non può essere guasto, altrimenti non sarebbe nemmeno stato prodotto. La probabilità che questo avvenga è $\binom{3}{1}p\,(1-p)^3=0.2187$. Il ruolo del coefficiente binomiale è quello di contare tutti i casi in cui possiamo avere un calcolatore guasto tra i primi 3.

Similmente potrebbero occorrere 5 calcolatori prodotti, se 2 dei primi 4 sono guasti. La probabilità di questo evento è $\binom{4}{2}p^2\,(1-p)^3=0.04374$.

In generale occorreranno n calcolatori prodotti per averne 3 funzionanti (di cui l'ultimo prodotto) con probabilità $\binom{n-1}{n-3}p^{n-3}\,(1-p)^3$.

Anche in questo caso dobbiamo calcolare la probabilità che l'azienda ci guadagni. Questo avviene per quegli n tali che $3\cdot 2000-n\cdot 1000>0$, ossia $n<6$ (anche in questo caso per $n=6$ c'è pareggio di bilancio). La probabilità che l'azienda guadagni, quindi, è la somma delle probabilità che debba produrre 3, 4 o 5 calcolatori:

$$P=\sum_{n=3}^{5}\binom{n-1}{n-3}p^{n-3}\,(1-p)^3=0.729+0.2187+0.04374=0.99144.$$

Possiamo osservare che abbiamo prestato molta poca attenzione a $(\Omega,\mathcal{F},P)$. Inoltre in Ω non c'era nulla che descrivesse il guadagno: abbiamo solo contato *quanti* calcolatori era necessario costruire per averne 1 (o 3) non guasti. Possiamo però scrivere un Ω diverso per i guadagni, ora:

$$\Omega_G^1=\{1000,0,-1000,-2000,\dots\},\qquad \Omega_G^3=\{3000,2000,1000,0,-1000,\dots\}$$

dove consideriamo rispettivamente la vendita di un calcolatore e di tre. Possiamo anche definire una probabilità direttamente su questi Ω (che hanno cardinalità sempre numerabile) "sfruttando" quanto calcolato nello spazio di probabilità precedente:

$$P_G^1(1000)=0.9, P_G^1(0)=0.99,\dots$$

e (nel caso dei 3 calcolatori)

$$P_G^3(3000)=0.729, P_G^3(2000)=0.2187, P_G^3(1000)=0.04374,\dots$$

Anche in questo esempio, come nel precedente abbiamo considerato una funzione del risultato dell'esperimento aleatorio di partenza,.

DEFINIZIONE 5.3. *Dato uno spazio probabilizzabile* $(\Omega, \mathcal{F})$, *si dice* variabile aleatoria *o* variabile casuale *ogni funzione* $X : \Omega \to \mathbb{R}$ *tale che per ogni* $x \in \mathbb{R}$, *l'insieme* $\{\omega \in \Omega : X(\omega) \leqslant x\} \in \mathcal{F}$.

Esempio 5.4. La funzione S dell'Esempio 5.1 che calcola la somma dei risultati dei due dadi (cioè delle due componenti di ogni elemento ω) è una variabile aleatoria.

Osservazione 5.5. Proviamo a leggere più a fondo questa definizione.

1. Chiamiamo queste *funzioni* "variabili aleatorie" perché il valore della funzione dipende dal risultato ω di un esperimento casuale.
2. Siamo partiti da uno spazio *probabilizzabile*, non di probabilità. Può sembrare strano, visto che siamo partiti dall'idea di assegnare una probabilità a funzioni di esiti, ma la definizione che abbiamo dato non dipende da una particolare probabilità.
3. Chiediamo che $\{\omega \in \Omega : X(\omega) \leqslant x\} \in \mathcal{F}$ perché vorremmo assegnare una probabilità a questi insiemi. Qualunque probabilità abbiamo sullo spazio $(\Omega, \mathcal{F})$, la possiamo "esportare" a questi insiemi.
4. Come mai consideriamo proprio gli insiemi di questa forma? Cosa ci ricorda la condizione $X(\omega) \leqslant x$? Lo spazio di arrivo è lo spazio $\mathbb{R}$ dei numeri reali, se vogliamo avere una probabilità ci serve come prima cosa una tribù e, per $\mathbb{R}$, abbiamo visto la tribù $\mathcal{B}$ dei Boreliani, che ha come possibili generatori le semirette $(-\infty, x]$.
5. Cominciamo a vedere dove vogliamo arrivare. Resta però una domanda: come mai stiamo procedendo "al contrario"? Perché "portiamo indietro" gli insiemi misurabili da $\mathcal{B}$ a $\mathcal{F}$ e non viceversa?

Esempio 5.6. Siano $\Omega = \{1,2,3\}$, $\mathcal{F} = \{\emptyset, \{1\}, \{2,3\}, \{1,2,3\}\}$ e sia $\tilde{\Omega} = \{1,2\}$. Se prendiamo la funzione $f : \Omega \to \tilde{\Omega}$ tale che $f(1) = f(2) = 1$ e $f(3) = 2$, la famiglia di insiemi $\tilde{\mathcal{F}} = \{f(E) : E \in \mathcal{F}\}$ non è una tribù, infatti $\tilde{\mathcal{F}} = \{f(\emptyset), f(\{1\}), f(\{2,3\}), f(\{1,2,3\})\} = \{\emptyset, \{1\}, \{1,2\}\}$, cui manca $\{2\}$ per essere una tribù.

Nota 5.7. Nell'enunciato precedente E è un insieme e con $f(E)$ ne stiamo prendendo l'immagine, ossia l'insieme $f(E) = \{\tilde{\omega} \in \tilde{\Omega} : \exists \omega \in E \subseteq \Omega : \tilde{\omega} = f(\omega)\}$. Osserviamo che $f(\Omega) \subseteq \tilde{\Omega}$, ma non necessariamente vale l'uguaglianza: per averla f deve essere suriettiva.

In modo analogo possiamo definire la *preimmagine* di un insieme di $\tilde{\Omega}$ mediante f: essa è l'insieme

$$f^{-1}(\tilde{E}) = \{\omega \in \Omega : f(\omega) \in \tilde{E}\},$$

definito per ogni sottoinsieme $\tilde{E}$ di $\tilde{\Omega}$. In questo caso f^{-1} *non* è la funzione inversa, che potrebbe anche non esistere, dal momento che non abbiamo fatto ipotesi sull'invertibilità di f. Non la stiamo vedendo come funzione degli elementi di $\tilde{\Omega}$, bensì come mappa di insiemi. In particolare, siccome f è una funzione, $f^{-1}(\tilde{\Omega}) = \Omega$, non occorre che f sia iniettiva né suriettiva.

D'altra parte, la scelta di tornare indietro funziona, come ci garantisce il seguente risultato.

TEOREMA 5.8. *Sia* $(\tilde{\Omega}, \tilde{\mathcal{F}})$ *uno spazio probabilizzabile. Siano inoltre* Ω *un insieme e* $X : \Omega \to \tilde{\Omega}$ *una funzione. Allora* $\mathcal{F} = \{X^{-1}(\tilde{E}) : \tilde{E} \in \tilde{\mathcal{F}}\}$ *è una tribù su* Ω.

Dimostrazione. Controlliamo che siano soddisfatte le tre proprietà che caratterizzano una tribù:

i. $X^{-1}(\tilde{\Omega}) = \Omega$, quindi $\Omega \in \mathcal{F}$;

ii. $X^{-1}(\tilde{E}^c) = X^{-1}(\tilde{\Omega} \setminus \tilde{E}) = \Omega \setminus X^{-1}(\tilde{E}) = (X^{-1}(\tilde{E}))^c$, quindi $(X^{-1}(\tilde{E}))^c \in \mathcal{F}$;

iii. $X^{-1}(\bigcup_{i=1}^{\infty} \tilde{E}_i) = \bigcup_{i=1}^{\infty} X^{-1}(\tilde{E}_i)$, quindi $\mathcal{F}$ è chiusa rispetto all'unione numerabile. □

Non solo, possiamo anche usare questa stessa idea per "portare avanti" una tribù.

TEOREMA 5.9. *Sia $(\Omega, \mathcal{F})$ uno spazio probabilizzabile. Siano inoltre $\tilde{\Omega}$ un insieme e $X: \Omega \to \tilde{\Omega}$ una funzione. Allora $\tilde{\mathcal{F}} = \{\tilde{E} \subseteq \tilde{\Omega} : X^{-1}(\tilde{E}) \in \mathcal{F}\}$ è una tribù.*

Dimostrazione. Controlliamo che siano soddisfatte le tre proprietà che caratterizzano una tribù:

i. $X^{-1}(\tilde{\Omega}) = \Omega \in \mathcal{F}$, quindi $\tilde{\Omega} \in \tilde{\mathcal{F}}$;

ii. se $\tilde{E} \in \tilde{\mathcal{F}}$, allora $X^{-1}(\tilde{E}) \in \mathcal{F}$, quindi $(X^{-1}(\tilde{E}^c)) = \Omega \setminus X^{-1}(\tilde{E}) = (X^{-1}(\tilde{E}))^c \in \mathcal{F}$, dunque $\tilde{E}^c \in \tilde{\mathcal{F}}$;

iii. se abbiamo una successione $(\tilde{E}_i)_i \subseteq \tilde{\mathcal{F}}$, allora la successione $(X^{-1}(\tilde{E}_i))_i \subseteq \mathcal{F}$, di conseguenza $\mathcal{F} \ni \bigcup_{i=1}^{\infty} X^{-1}(\tilde{E}_i) = X^{-1}(\bigcup_{i=1}^{\infty} \tilde{E}_i)$ e $\bigcup_{i=1}^{\infty} \tilde{E}_i \in \tilde{\mathcal{F}}$. □

Osservazione 5.10. Noi siamo interessati a un caso speciale, in cui $\tilde{\Omega} = \mathbb{R}$ e $\tilde{\mathcal{F}} = \mathcal{B}$ e $X: \Omega \to \mathbb{R}$ è una variabile aleatoria. In questo contesto la famiglia di insiemi

$$\sigma(X) := \{X^{-1}(B) : B \in \mathcal{B}\}$$

è una tribù, detta *tribù generata da X*. Inoltre $\sigma(X) \subseteq \mathcal{F}$, con l'inclusione invece dell'uguaglianza, perché non è detto che tutti gli elementi di $\mathcal{F}$ siano controimmagine di qualche Boreliano B.

Gli eventi in $\mathcal{F}$ sono quelli per cui abbiamo un valore della funzione probabilità, nel momento in cui ne definiamo una su $(\Omega, \mathcal{F})$. Gli eventi in $\sigma(X)$ sono tutti gli eventi che "hanno a che fare" con X. Dal momento che $\sigma(X)$ è un sottoinsieme di $\mathcal{F}$, abbiamo automaticamente una probabilità anche per tutti gli eventi in $\sigma(X)$.

Esempio 5.11. Nel momento in cui abbiamo una probabilità sui risultati dei due dadi nell'Esempio 5.1 (quella dei dadi bilanciati, ad esempio, ma anche qualche probabilità diversa, che descriva dadi truccati), attraverso S abbiamo immediatamente una probabilità sui numeri reali che descrive la probabilità della somma dei due dadi.

Osservazione 5.12. Non sempre è facile capire se una funzione sia o meno una variabile aleatoria: c'è una branca della matematica che si occupa (anche) di questo, chiamata *Teoria della misura*. Tuttavia ci viene in aiuto, nel caso $\Omega = \mathbb{R}$ e $\mathcal{F} = \mathcal{B}$, il seguente teorema.

TEOREMA 5.13. *Sia $X: (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ una funzione continua o monotona crescente o monotona decrescente. Allora X è una variabile aleatoria.*

Osservazione 5.14. La notazione $X: (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ nell'enunciato precedente sottolinea il fatto che siamo interessati non solo agli insiemi di partenza e arrivo della funzione X, ma che essi ci interessano come spazi probabilizzabili, in particolare entrambi con la tribù dei Boreliani.

Può sembrare che stiamo introducendo molta notazione e che, sotto sotto, ci stiamo complicando la vita: in fondo che male c'è ad avere tanti spazi probabilizzabili diversi per descrivere esperimenti aleatori diversi? In realtà sapere che possiamo riscrivere un esperimento aleatorio sullo spazio $(\mathbb{R}, \mathcal{B})$ ci dice che possiamo concentrarci a definire probabilità su quello spazio e abbiamo visto che per farlo abbiamo bisogno di funzioni F sui reali (con un po' di caratteristiche, che abbiamo visto nella Definizione XXX).

Ora mettiamo alla prova il nostro strumento, le variabili aleatorie, per descrivere un particolare esperimento aleatorio.

Esempio 5.15. Lanciamo una moneta, ancora una volta. Questa volta siamo interessati a una successione di lanci di una moneta bilanciata. Vogliamo calcolare la probabilità di ottenere testa per la prima volta in un lancio dispari (ad esempio perché stiamo giocando in due, lanciando alternatamente, col primo a ottenere testa che vince).

Lo spazio di probabilità che descrive questo esperimento è $(\Omega, \mathcal{F}, P)$, con $\Omega = \{T, C\}^{\mathbb{N} \setminus \{0\}}$ lo spazio prodotto delle successioni di teste e croci, $\mathcal{F}$ la tribù generata dai cilindri, cioè quegli eventi in cui fissiamo un numero finito di indici, e P è la probabilità prodotto.

Siamo interessati a calcolare la probabilità che testa esca la prima volta a un lancio dispari. Consideriamo allora come variabile aleatoria X la funzione che ci dice qual è il primo lancio in cui esce testa. In questo modo quello che vogliamo calcolare è $P(X \in \{2k+1, k \in \mathbb{N}\})$. Com'è fatta questa funzione? Possiamo scriverla come $X(\omega) = \inf\{i \geqslant 1 : \omega_i = T\}$.

Ora che abbiamo X, possiamo ricavarci la tribù generata da X, $\sigma(X)$. Cominciamo a vedere come sono fatte le controimmagini dei singoletti di numeri naturali positivi (anche perché ci aspettiamo che siano gli eventi in cui la probabilità sarà non nulla). Abbiamo

$$X^{-1}(\{4\}) = \{\omega \in \Omega : \omega_1 = \omega_2 = \omega_3 = C, \omega_4 = T\}$$

cioè il cilindro le cui prime 3 componenti sono C e la quarta è T. Più in generale, $\sigma(X)$ è formata da unioni finite o numerabili di cilindri della forma

$$H_k := \{\omega \in \Omega : \omega_i = C, i < k, \omega_k = T\}.$$

Quindi

$$\begin{aligned} P(X \text{ è dispari}) &= \sum_{i=0}^{\infty} P(\omega \in H_{2i+1}) \\ &= \sum_{i=0}^{\infty} \frac{1}{2^{2i+1}} = \frac{1}{2} \sum_{i=0}^{\infty} 4^{-i} = \frac{1}{2} \cdot \frac{1}{1 - \frac{1}{4}} = \frac{2}{3}. \end{aligned}$$

Potevamo arrivare allo stesso risultato osservando che $P(X \text{ pari}) + P(X \text{ dispari}) = 1$ e che

$$P(X \text{ pari}) = \frac{1}{2} P(X \text{ dispari}),$$

perché è la probabilità che il primo lancio sia C e che poi contiamo i lanci a partire dal primo del secondo giocatore. Quindi, ponendo $x = P(X \text{ dispari})$, $x + \frac{1}{2}x = 1$ e $x = \frac{2}{3}$.

Lezione 8

Abbiamo introdotto le *variabili aleatorie*, funzioni dall'insieme degli esiti di un esperimento aleatorio all'insieme dei numeri reali, già dotato della tribù dei Boreliani. Pensarle come funzioni può aiutare a capirle meglio: non pensiamo tanto al *valore* della funzione in un punto, cioè al valore della funzione per un particolare esito ω, ma *alla funzione in sé*, in senso globale. Ci interessano di più i valori che può assumere e con quale probabilità li può assumere.

Vediamo alcuni esempi di variabili aleatorie molto semplici. Fissiamo, per tutti gli esempi seguenti, uno spazio di probabilità $(\Omega, \mathcal{F}, P)$.

Esempio 5.16. (Variabili aleatorie degeneri) Per ogni $c \in \mathbb{R}$ la funzione costante $X(\omega) \equiv c$, per ogni $\omega \in \Omega$, è una variabile aleatoria, detta *variabile aleatoria degenere*. Una volta fissato c (e quindi una particolare funzione costante, cioè una particolare variabile aleatoria X), possiamo chiederci quale sia la probabilità che la funzione X assuma un certo valore a. Ci aspettiamo che questa probabilità sia 1 se $a = c$ e 0 altrimenti e così è:

$$P(X = a) = \begin{cases} 1 & a = c \\ 0 & a \neq c \end{cases}$$

Più in generale vorremo calcolare la probabilità di un evento, ossia di un insieme, cioè con la terminologia delle variabili aleatorie, $P(X \in A)$. Questo possiamo farlo se $A \in \mathcal{B}$, grazie alle proprietà delle tribù, in particolare dei Boreliani, e della preimmagine nella definizione di variabile aleatoria:

$$P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}) = P(X^{-1}(A)),$$

in cui $X^{-1}(A) \in \mathcal{F}$ e quindi l'ultima probabilità è ben definita.

Nel caso particolare della variabile aleatoria degenere $X \equiv c$, se $A \in \mathcal{B}$,

$$P(X \in A) = \begin{cases} 1 & \text{se } c \in A \\ 0 & \text{altrimenti.} \end{cases}$$

Qual è la tribù generata da X? Abbiamo in questo caso $\sigma(X) = \{\emptyset, \Omega\}$: la preimmagine di un insieme A in $\mathcal{B}$ è tutto Ω se $c \in A$ ed è l'insieme vuoto altrimenti.

Esempio 5.17. (Variabili aleatorie indicatrici) In questo caso partiamo con un evento nello spazio di partenza $(\Omega, \mathcal{F}, P)$: $E \in \mathcal{F}$. La *variabile aleatoria indicatrice di E* è definita come

$$I_E(\omega) = \mathbb{1}_E(\omega) = \begin{cases} 1 & \text{se } \omega \in E \\ 0 & \text{se } \omega \in E^c. \end{cases}$$

Quindi la variabile aleatoria I_E può assumere due valori, 0 oppure 1. Com'è fatta allora $\sigma(I_E)$? Dobbiamo vedere chi sono gli insieme pre-immagine dei Boreliani. Abbiamo sicuramente $\emptyset$ e Ω, ma anche $E = I_E^{-1}(\{1\})$ ed $E^c = I_E^{-1}(\{0\})$. Inoltre ogni insieme $A \in \mathcal{B}$ che contiene 1 ma non 0 ha come preimmagine E, ogni insieme $B \in \mathcal{B}$ che contiene 0 ma non 1 ha come preimmagine E^c. Se un insieme di $\mathcal{B}$ non contiene alcuno tra 0 e 1, la sua preimmagine è $\emptyset$, mentre se li contiene entrambi la sua preimmagine è Ω.

Per quanto riguarda la probabilità, abbiamo, per $A \in \mathcal{B}$

$$P(I_E \in A) = \begin{cases} P(E) & \text{se } 1 \in A \text{ e } 0 \notin A \\ P(E^c) & \text{se } 1 \notin A \text{ e } 0 \in A \\ 1 & \text{se } 1 \in A \text{ e } 0 \in A \\ 0 & \text{se } 1 \notin A \text{ e } 0 \notin A. \end{cases}$$

Questa funzione è proprio la funzione indicatrice dell'insieme E, ma la sua probabilità non è una funzione indicatrice.

Esempio 5.18. (Variabili aleatorie semplici) Una volta che abbiamo le variabili aleatorie indicatrici, ne possiamo considerare delle combinazioni lineari[5.1]. Per esempio, dati $E, F \in \mathcal{F}$, possiamo prendere $X = I_E - 3I_F$.

In analogia all'esempio delle variabili aleatorie indicatrici, come prima cosa ci chiediamo quali siano i possibili valori di X e quali siano, di conseguenza, gli elementi di $\sigma(X)$. Abbiamo

$$X(\omega) = \begin{cases} 0 & \omega \notin E \cup F \\ 1 & \omega \in E \setminus F \\ -3 & \omega \in F \setminus E \\ -2 & \omega \in E \cap F. \end{cases}$$

Quindi abbiamo che la tribù generata da X è la tribù generata da E ed F:

$$\begin{aligned} \sigma(X) &= \sigma(E, F) \\ &= \{\emptyset, E, F, E^c, F^c, E \cap F, (E \cap F)^c = E^c \cup F^c, E \cup F, (E \cup F)^c = E^c \cap F^c, E \setminus F = E \cap F^c, \\ &\quad (E \setminus F)^c = E^c \cup F, F \setminus E = F \cap E^c, (F \setminus E)^c = F^c \cup E, E \Delta F = (E \cap F^c) \cup (E^c \cap F), \\ &\quad (E \Delta F)^c = (E^c \cup F) \cap (F^c \cup E) = (E^c \cap F^c) \cup (E \cap F), \Omega\}. \end{aligned}$$

A questo punto, in modo del tutto analogo a quanto visto prima, possiamo assegnare, per $A \in \mathcal{B}$, dei valori di probabilità $P(X \in A)$. Quanti sono questi valori? Quali sono? Quanti ne dobbiamo calcolare?

◦ Soluzione:

Negli esempi precedenti abbiamo parlato di probabilità della variabile aleatoria, ma prima di continuare andiamo a formalizzare quanto già fatto. Sia $(\Omega, \mathcal{F}, P)$ uno spazio di probabilità, $X : \Omega \to \mathbb{R}$ una variabile aleatoria e $A \in \mathcal{B}$. Allora

$$P(X \in A) = P(\{\omega \in \Omega : X(\omega) \in A\}) = P(X^{-1}(A))$$

e questa quantità è ben definita, infatti A è un Boreliano, quindi può essere scritto mediante unione (numerabile) e complementare di semirette della forma $(-\infty, a]$ e, per definizione di variabile aleatoria, ogni preimmagine di semiretta (e quindi ogni preimmagine di Boreliani) è in $\mathcal{F}$ e, per concludere, per ogni elemento di $\mathcal{F}$ la probabilità P è ben definita.

In questo senso, quindi, la funzione $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B})$ "trasporta" su $(\mathbb{R}, \mathcal{B})$ una qualunque probabilità P definita sullo spazio probabilizzabile $(\Omega, \mathcal{F})$. Possiamo chiamare P_X questa probabilità su $(\mathbb{R}, \mathcal{B})$.

5.1. Questo dovrebbe richiamare memorie del corso di Analisi.

DEFINIZIONE 5.19. *Dati uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$ *e una variabile aleatoria* $X\colon(\Omega,\mathcal{F})\to(\mathbb{R},\mathcal{B})$, *si dice* legge *o* distribuzione *di* X *la funzione di probabilità* P_X *definita su* $(\mathbb{R},\mathcal{B})$ *per ogni* $A\in\mathcal{B}$ *da*

$$P_X(A) := P(X\in A) = P(X^{-1}(A)).$$

Esempio 5.20. Tornando alla situazione descritta nell'Esempio 5.1, cioè la somma delle facce di due dadi bilanciati indipendenti, possiamo ora scrivere $P_S(\{7\})$ per indicare $P(S\in\{7\})$.

DEFINIZIONE 5.21. *Siano* X *e* Y *due variabili aleatorie definite su due spazi di probabilità,* $(\Omega,\mathcal{F},P)$ *e* $(\tilde\Omega,\tilde{\mathcal{F}},\tilde P)$ *rispettivamente:*

$$X\colon(\Omega,\mathcal{F},P)\to(\mathbb{R},\mathcal{B})$$
$$Y\colon(\tilde\Omega,\tilde{\mathcal{F}},\tilde P)\to(\mathbb{R},\mathcal{B}).$$

Chiamiamo P_X *e* P_Y *le loro leggi, cioè funzioni di probabilità definite su* $(\mathbb{R},\mathcal{B})$ *come*

$$P_X(\cdot) = P(X^{-1}(\cdot)); \qquad P_Y(\cdot) = \tilde P(Y^{-1}(\cdot)).$$

Se le due funzioni di probabilità P_X *e* P_Y *sono uguali, cioè se assegnano la medesima probabilità ad ogni elemento di* $\mathcal{B}$*, diciamo che le variabili aleatorie* X *e* Y *sono* identicamente distribuite *e scriviamo* $X\sim Y$.

Se leggiamo meglio la definizione appena data, dire che $X\sim Y$ equivale ad affermare che sono due copie dello stesso esperimento, almeno dal punto di vista della probabilità. Notiamo che non è necessario che X e Y siano definite sullo stesso spazio di probabilità. In altre parole stiamo dicendo che possiamo rappresentare esperimenti aleatori equivalenti in spazi diversi senza che questo abbia effetti sulla probabilità. Allo stesso tempo è anche possibile che esperimenti aleatori a priori diversi tra loro siano descritti da variabili aleatorie identicamente distribuite e che quindi, dal punto di vista della probabilità, siano essenzialmente lo stesso esperimento.

Esempio 5.22. Non importa se per descrivere il lancio di un dado a 6 facce scegliamo $\Omega=\{1,2,3,4,5,6\}$, $\Omega=\{\text{uno, due, tre, quattro, cinque, sei}\}$, $\Omega=\{U,D,T,Q,C,S\}$: anche se formalmente sono spazi degli esiti distinti (e quindi ne derivano spazi di probabilità diversi), possiamo in tutti i casi scrivere una codifica dell'esperimento in $\mathbb{R}$ attraverso un'opportuna variabile aleatoria (il risultato del lancio), in modo che ciascuna di esse sia identicamente distribuita rispetto alle altre e quindi equivalente dal punto di vista della probabilità.

Esempio 5.23. Consideriamo un'urna con 50 biglie bianche e 50 biglie nere, da cui estraiamo una biglia. Sia X la variabile aleatoria indicatrice dell'evento "è uscita una biglia bianca". Allora, per $A\in\mathcal{B}$,

$$P_X(A) = \begin{cases} 1 & \text{se } \{0,1\}\subseteq A \\ \frac{1}{2} & \text{se } \#\{\{0,1\}\cap A\}=1 \\ 0 & \text{se } \{0,1\}\cap A=\emptyset. \end{cases}$$

Prendiamo ora una moneta bilanciata, che lanciamo una sola volta, e definiamo Y la variabile aleatoria indicatrice dell'evento "è uscita croce". Per $B\in\mathcal{B}$ abbiamo

$$P_Y(B) = \begin{cases} 1 & \text{se } \{0,1\}\subseteq B \\ \frac{1}{2} & \text{se } \#\{\{0,1\}\cap B\}=1 \\ 0 & \text{se } \{0,1\}\cap B=\emptyset. \end{cases}$$

Allo stesso modo, consideriamo una sfida tra due giocatori, Cassandra e Daniele, in cui ciascuno abbia la medesima probabilità di vincere (e non sia possibile pareggiare) e chiamiamo Z la variabile indicatrice dell'evento "vince Cassandra". Per $C\in\mathcal{B}$ vale ancora una volta

$$P_Z(C) = \begin{cases} 1 & \text{se } \{0,1\}\subseteq C \\ \frac{1}{2} & \text{se } \#\{\{0,1\}\cap C\}=1 \\ 0 & \text{se } \{0,1\}\cap C=\emptyset. \end{cases}$$

Astraendo i tre esperimenti alle sole proprietà o caratteristiche che riguardano la probabilità, possiamo osservare che essi sono lo stesso esperimento, fatto codificato dalla notazione $X \sim Y \sim Z$.

Questo è uno dei motivi per cui monete, dadi e urne sono così comuni negli esempi di probabilità: permettono di rappresentare in termini molto semplici e vicini all'esperienza comune esperimenti aleatori magari molto complicati ma che, dal punto di vista della probabilità, non aggiungono nulla. Un esempio classico è quello della descrizione della diffusione di un'infezione mediante urne.

Osservazione 5.24. Nel momento in cui assegniamo una legge a una variabile aleatoria, non è più necessario specificare lo spazio di probabilità sottostante. Le variabili aleatorie ci permettono di riprodurre nello spazio probabilizzabile $(\mathbb{R}, \mathcal{B})$ gli esperimenti aleatori, mediante la scelta di un'opportuna probabilità, la legge della variabile aleatoria. In questo modo abbiamo semplificato la portata della teoria che dobbiamo sviluppare: non occorre farlo per tutti i possibili spazi probabilizzabili, ma per il solo $(\mathbb{R}, \mathcal{B})$.

Insomma, ci basta parlare di funzioni di probabilità sullo spazio $(\mathbb{R}, \mathcal{B})$. Abbiamo già visto come definirle: ci basta assegnarle su una particolare famiglia di generatori della tribù $\mathcal{B}$ dei Boreliani, le semirette di forma $(-\infty, a]$, al variare di $a \in \mathbb{R}$. Quindi, tornando al contesto delle leggi delle variabili aleatorie, è sufficiente specificare il valore di una legge P_X sulle semirette per averne una definizione univoca su tutta la tribù $\mathcal{B}$. Questo ci permette di lavorare con una funzione su $\mathbb{R}$, dal momento che le semirette sono in relazione biunivoca con i numeri reali, invece che con una funzione su $\mathcal{B}$, cioè con una funzione su numeri invece che su insiemi di numeri, qualcosa cui siamo più abituati.

DEFINIZIONE 5.25. *Data una variabile aleatoria X sullo spazio di probabilità $(\Omega, \mathcal{F}, P)$, la* funzione di ripartizione *o* funzione cumulativa[5.2] *di X è la funzione $F_X : \mathbb{R} \to \mathbb{R}$ definita per ogni $y \in \mathbb{R}$ da*

$$\begin{aligned} F_X(y) &:= P_X((-\infty, y]) \\ &= P(X \in (-\infty, y]) \\ &= P(\{\omega \in \Omega : X(\omega) \leqslant y\}) \\ &= P(X \leqslant y). \end{aligned}$$

Con un leggero abuso di notazione si scrive $X \sim F_X$ per dire che la variabile aleatoria X ha funzione di ripartizione F_X.

Esempio 5.26. Consideriamo una variabile aleatoria degenere $X \equiv c$. La sua funzione di ripartizione è

$$F_X(y) = P(X \leqslant y) = \begin{cases} 1 & \text{se } y \geqslant c \\ 0 & \text{se } y < c \end{cases}$$

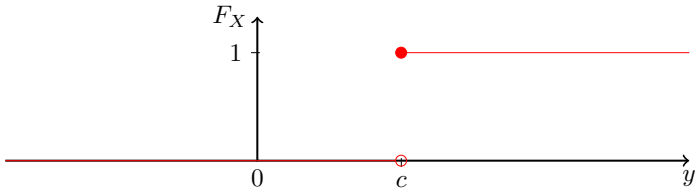


Figura 5.1. Funzione di ripartizione della v.a. degenere $X \equiv c$

Esempio 5.27. Consideriamo ora una variabile aleatoria indicatrice. Per far ciò fissiamo un evento $E \in \mathcal{F}$: la variabile aleatoria indicatrice associata a E, indicata con I_E o $\mathbb{1}_E$ è la funzione

$$I_E(\omega) = \mathbb{1}_E(\omega) = \begin{cases} 1 & \text{se } \omega \in E \\ 0 & \text{se } \omega \in E^c. \end{cases}$$

5.2. Ci sono anche altre varianti, che nascono dal termine inglese, *cumulative distribution function*, da cui viene anche l'abbreviazione *cdf*.

La funzione di ripartizione di questa variabile aleatoria è

$$F_{I_E}(y) = P(I_E \leqslant y) = \begin{cases} 0 & \text{se } y < 0 \\ P(E^c) & \text{se } 0 \leqslant y < 1 \\ 1 & \text{se } y \geqslant 1 \end{cases}$$

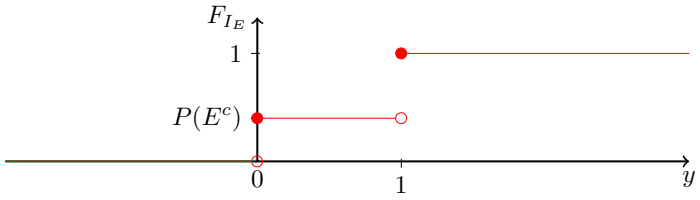


Figura 5.2. Funzione di ripartizione della v.a. indicatrice I_E

Esempio 5.28. Prendiamo ora il lancio di un dado bilanciato a 4 facce: lo rappresentiamo con la variabile aleatoria D_4, la cui funzione di ripartizione è

$$F_{D_4}(y) = P(D_4 \leqslant y) = \begin{cases} 0 & \text{se } y < 1 \\ 1/4 & \text{se } 1 \leqslant y < 2 \\ 2/4 & \text{se } 2 \leqslant y < 3 \\ 3/4 & \text{se } 3 \leqslant y < 4 \\ 1 & \text{se } y \geqslant 4 \end{cases}$$

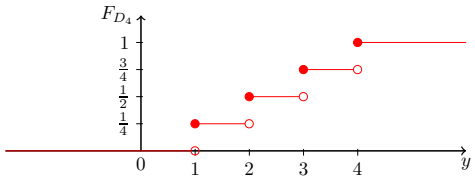


Figura 5.3. Funzione di ripartizione della v.a. di un dado a 4 facce D_4

Esempio 5.29. Consideriamo ora lo spazio di probabilità $(\Omega, \mathcal{F}, P)$ in cui $\Omega = [0,1]$ è l'intervallo unitario dei numeri reali, $\mathcal{F} = \mathcal{B}([0,1])$ è la tribù dei Boreliani ristretta all'intervallo $[0,1]$ e come probabilità abbiamo $P([a,b]) = b - a$, la lunghezza dei segmenti (detta anche misura di Lebesgue). Prendiamo la variabile aleatoria $X = \mathrm{Id} : [0,1] \to \mathbb{R}$. Vogliamo determinarne la funzione di ripartizione:

$$\begin{aligned} F_X(y) &= P(X \leqslant y) \\ &= P(\{\omega \in \Omega : X(\omega) \leqslant y\}) \\ &= \begin{cases} P(\emptyset) = 0 & \text{se } y < 0 \\ P([0,y]) = y & \text{se } 0 \leqslant y < 1 \\ P([0,1]) = 1 & \text{se } y \geqslant 1 \end{cases} \end{aligned}$$

In questo caso possiamo osservare che la funzione di ripartizione di questa variabile aleatoria, diversamente da quanto visto negli esempi precedenti, è una funzione continua. Questa particolare variabile aleatoria è molto importante e prende il nome di *variabile aleatoria uniforme su* $[0,1]$.

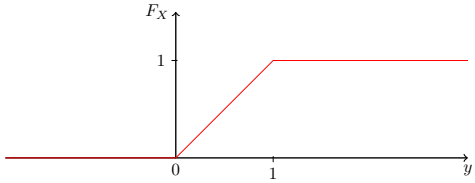


Figura 5.4. Funzione di ripartizione della v.a. uniforme su $[0,1]$

Ripartiamo dagli ultimi esempi visti e richiamiamo alcune proprietà già viste. Abbiamo visto che, per assegnare una legge a una variabile aleatoria, è sufficiente assegnare una funzione di probabilità su $\mathbb{R}$, cosa che possiamo fare in particolare attraverso una funzione di ripartizione.

Osservazione 5.30. Se abbiamo assegnato una funzione di ripartizione F_X su $\mathbb{R}$, possiamo calcolare non solo la probabilità P_X di una semiretta $(-\infty, y]$, ma anche di tutti gli altri insiemi Boreliani su $\mathbb{R}$. In particolare, la probabilità dell'intervallo $(a,b]$, con $a<b$, è

$$P_X((a,b]) = P((-\infty,b]) - P((-\infty,a]) = F_X(b) - F_X(a),$$

grazie alle proprietà delle funzioni di probabilità e alle definizioni date. Possiamo allora recuperare dalla discussione fatta precedentemente le proprietà della funzione di ripartizione F_X.

PROPOSIZIONE 5.31. *Data una variabile aleatoria X, la sua funzione di ripartizione F_X soddisfa le seguenti proprietà:*

i. è non decrescente

ii. $\lim_{x\to-\infty} F_X(x) = 0$ *e* $\lim_{x\to+\infty} F_X(x) = 1$

iii. è cadlag, *ossia continua a destra (*$\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$*) e limitata a sinistra (*$\lim_{x\to x_0^-} F_X(x) = F_X(x_0) - P(X = x_0)$*).*

Dimostrazione. Mostriamo, in ordine, le varie proprietà. Per $s \leqslant t$ abbiamo

$$\begin{aligned} F_X(s) &= P(X \leqslant s) = P(\{\omega \in \Omega : X(\omega) \leqslant s\}) \\ &\leqslant P(\{\omega \in \Omega : X(\omega) \leqslant t\}) = P(X \leqslant t) = F_X(t) \end{aligned}$$

in cui, per la disuguaglianza abbiamo usato la monotonia delle funzioni di probabilità e l'inclusione

$$\{\omega \in \Omega : X(\omega) \leqslant s\} \subseteq \{\omega \in \Omega : X(\omega) \leqslant t\}.$$

Per quanto riguarda i limiti agli estremi,

$$\begin{aligned} \lim_{x\to-\infty} F_X(x) &= \lim_{x\to-\infty} P(X \leqslant x) \\ &= \lim_{x\to-\infty} P(\{\omega \in \Omega : X(\omega) \leqslant x\}) \\ &= P\left(\bigcap_{n\to+\infty} \{\omega \in \Omega : X(\omega) \leqslant -n\} \right) \\ &= P(\emptyset) = 0 \end{aligned}$$

e, in modo del tutto analogo,

$$\begin{aligned} \lim_{x\to+\infty} F_X(x) &= \lim_{x\to+\infty} P(X \leqslant x) \\ &= \lim_{x\to+\infty} P(\{\omega \in \Omega : X(\omega) \leqslant x\}) \\ &= P\left(\bigcup_{n\to+\infty} \{\omega \in \Omega : X(\omega) \leqslant n\} \right) \\ &= P(X^{-1}(\mathbb{R})) = P(\Omega) = 1. \end{aligned}$$

In un generico punto interno $x_0 \in \mathbb{R}$ abbiamo, per il limite da destra,

$$\begin{aligned} \lim_{x\to x_0^+} F_X(x) &= \lim_{x\to x_0^+} P(X \leqslant x) \\ &= P\left(\bigcap_{n\to+\infty} \left\{\omega \in \Omega : X(\omega) \leqslant x_0 + \frac{1}{n}\right\} \right) \\ &= P(X \leqslant x_0) = F_X(x_0) \end{aligned}$$

e per quello da sinistra

$$\begin{aligned}\lim_{x\to x_0^-} F_X(x) &= \lim_{x\to x_0^-} P(X\leqslant x)\\ &= P\left(\bigcup_{n\to+\infty}\left\{\omega\in\Omega: X(\omega)\leqslant x_0-\frac{1}{n}\right\}\right)\\ &= P(X<x_0)\\ &= P(X\leqslant x_0)-P(X=x_0)\\ &= F_X(x_0)-P(X=x_0)\end{aligned}$$

in cui abbiamo usato la seguente osservazione: se $\omega\in\bigcup_{n\to+\infty}\left\{X(\omega)\leqslant x_0-\frac{1}{n}\right\}$, allora $X(\omega)\leqslant x$ per qualche $x<x_0$ e dunque $X(\omega)<x_0$. □

Osservazione 5.32. Fissato un qualunque punto $x_0\in\mathbb{R}$, chiedere che la funzione di ripartizione F_X sia continua in x_0, cioè chiedere che $\lim_{x\to x_0^+} F_X(x)=\lim_{x\to x_0^-} F_X(x)$ è equivalente a chiedere che la probabilità che X assuma il valore x_0 sia nulla, cioè $P(X=x_0)=0$.

Osservazione 5.33. Come già osservato in precedenza, a partire da una funzione di ripartizione F_X possiamo calcolare la probabilità in un qualunque Boreliano. Ad esempio:

- $P(X\in(a,b])=F_X(b)-F_X(a)$
- $P(X\in[a,b])=F_X(b)-F_X(a)+P(X=a)=F_X(b)-\lim_{x\to a^-}F_X(x)$
- $P(X<a)=\lim_{x\to a^-}F_X(x)$
- $P(X>b)=1-F_X(b)$

e così via.

5.1. VARIABILI ALEATORIE DISCRETE E CONTINUE

Le variabili aleatorie si possono dividere in tre classi:

- variabili aleatorie discrete
- variabili aleatorie (assolutamente) continue
- variabili aleatorie miste.

DEFINIZIONE 5.34. *Una variabile aleatoria che può assumere al più un numero finito o numerabile di valori si dice* variabile aleatoria discreta.

Osservazione 5.35. Una caratterizzazione equivalente di variabile aleatoria discreta può essere data in termini della funzione di ripartizione. Una variabile aleatoria è discreta se e solo se la sua funzione di ripartizione è discontinua e costante a tratti, con un numero finito o numerabile di discontinuità. I punti di discontinuità sono i valori che la variabile aleatoria può assumere.

DEFINIZIONE 5.36. *Una variabile aleatoria X si dice* continua *se la sua funzione di ripartizione F_X è continua. Se, inoltre, esiste una funzione non negativa $f_X:\mathbb{R}\to\mathbb{R}$ tale che, per ogni $x\in\mathbb{R}$,*

$$F_X(x)=\int_{-\infty}^{x} f_X(y)\,\mathrm{d}y$$

allora X si dice assolutamente continua.

Osservazione 5.37. Al momento il motivo di questa richiesta addizionale per la funzione di ripartizione non è chiarissima, ma ne capiremo il motivo tra qualche pagina.

DEFINIZIONE 5.38. *Una variabile aleatoria che non sia né discreta né (assolutamente) continua*[5.3] *di dice* variabile aleatoria mista.

5.3. Il fatto che si escludano tutte le continue o solo le assolutamente continue dipende dai casi. Noi escluderemo le assolutamente continue.

Osservazione 5.39. La famiglia delle variabili aleatorie miste è la classe più grande, ma anche quella di cui possiamo dire di meno, in particolare in un corso introduttivo come questo. Nel seguito non le tratteremo quasi mai.

Lezione 9

Esempio 5.40. Vediamo come costruire un'interessante variabile aleatoria mista. Per farlo, partiamo da una variabile aleatoria uniforme sull'intervallo $[0,1]$, scrivendone i possibili valori in binario. A questo punto, sostituiamo, nella rappresentazione come allineamento dei valori, ogni 1 con un 2 e leggiamo i valori risultanti come se fossero in base 3.

La funzione di ripartizione di questa variabile aleatoria è continua e costante a tratti, ma non ammette primitiva.

- Codice R per generare la figura

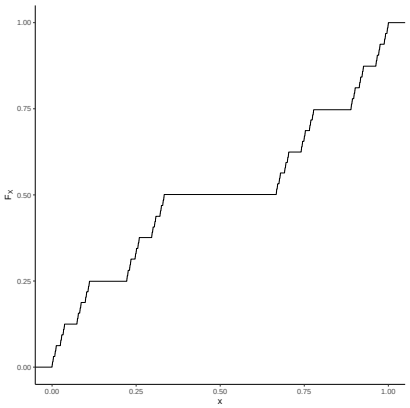


Figura 5.5. La funzione di ripartizione della variabile aleatoria di Cantor

Esempio 5.41. Un altro esempio, meno drammatico, di variabile aleatoria mista è il seguente:

$$F_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ x & \text{se } 0 \leqslant x < 1/2 \\ 1 & \text{se } 1/2 \leqslant x \end{cases}$$

Questa variabile aleatoria non può essere continua, perché ha una discontinuità in $\frac{1}{2}$, ma allo stesso tempo non è discreta, poiché, pur avendo un numero finito di discontinuità, non è costante a tratti.

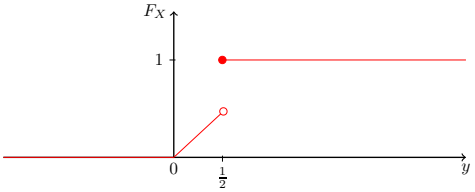


Figura 5.6. Funzione di ripartizione di una variabile aleatoria mista

5.1.1. Variabili aleatorie discrete

Ci concentriamo ora sulle variabili aleatorie discrete.

DEFINIZIONE 5.42. *Sia X una variabile aleatoria discreta. Chiamiamo* densità discreta *o* funzione di massa di probabilità *(a volte abbreviata con* pmf*) la funzione $\varphi_X : \mathbb{R} \to [0,1]$, $\varphi_X(x) := P(X = x)$. L'insieme $\mathcal{R}_X = \{x_i\}_{i \in I}$, al più numerabile, dei possibili valori assunti da X (e quindi punti in cui φ_X è non nulla) prende il nome di* supporto di X *o di φ_X.*

TEOREMA 5.43. (PROPRIETÀ DELLA PMF) *Sia X una variabile aleatoria discreta e sia φ_X la sua densità discreta. Allora:*

i. per ogni $x \in \mathbb{R}$, $\varphi_X(x) \geqslant 0$

ii. per ogni $x \in \mathcal{R}_X^c$, $\varphi_X(x) = 0$
iii. $\sum_{x \in \mathcal{R}_X} \varphi_X(x) = 1$
iv. se $E \in \mathcal{B}$, *allora* $P_X(E) = \sum_{x \in \mathcal{R}_X \cap E} \varphi_X(x) = \sum_{x \in \mathcal{R}_X} \mathbb{1}_E(x)\, \varphi_X(x)$.

Dimostrazione. Come prima cosa osserviamo che le proprietà i e ii sono immediate dalla Definizione 5.42. Inoltre, le somme in iii e iv sono ben definite, perché $\mathcal{R}_X$ è al più numerabile. Mostriamo la iv e la iii seguirà come caso particolare. Abbiamo

$$\begin{aligned} P_X(E) &= P(X \in E) = P(\{\omega \in \Omega : X(\omega) \in E\} \cap \Omega) \\ &= P\left(\{\omega \in \Omega : X(\omega) \in E\} \cap \bigcup_{x \in \mathcal{R}_X} \{\omega \in \Omega : X(\omega) = x\}\right) \\ &= P\left(\bigcup_{x \in \mathcal{R}_X} (\{\omega \in \Omega : X(\omega) \in E\} \cap \{\omega \in \Omega : X(\omega) = x\})\right) \\ &= \sum_{x \in \mathcal{R}_X} P((X \in E) \cap (X = x)) \\ &= \sum_{x \in \mathcal{R}_X} \mathbb{1}_E(x)\, \varphi_X(x). \end{aligned}$$

Come detto, la iii si ottiene nel caso particolare $E = \Omega$. □

Osservazione 5.44. Se X è una variabile aleatoria discreta di densità discreta φ_X, allora la sua funzione di ripartizione F_X è tale che

$$F_X(y) = \sum_{x \in \mathcal{R}_X} \mathbb{1}_{(-\infty, y]}(x)\, \varphi_X(x).$$

In particolare questo ripete quanto osservato in precedenza: la funzione di ripartizione è costante a tratti con salti nei punti in $\mathcal{R}_X$. L'ampiezza dei salti, inoltre, è proprio la probabilità che la variabile aleatoria assuma quel valore.

5.1.2. Variabili aleatorie assolutamente continue

Nel caso delle variabili aleatorie continue, e a maggior ragione nel caso di quelle assolutamente continue, non possiamo aspettarci una funzione come la densità discreta. Infatti, poiché per definizione F_X è continua, in ogni punto $P(X = x) = 0$ e quindi φ_X sarebbe identicamente nulla.

DEFINIZIONE 5.45. *Sia X una variabile aleatoria assolutamente continua. La funzione non negativa $f_X : \mathbb{R} \to \mathbb{R}$ tale che $F_X(x) = \int_{-\infty}^{x} f_X(y)\, \mathrm{d}y$ prende il nome di* funzione di densità di probabilità *(o più semplicemente* densità*) di X, a volte abbreviata con* pdf. *Per una variabile aleatoria assolutamente continua, $\mathcal{R}_X = \{x \in \mathbb{R} : f_X(x) \neq 0\}$.*

TEOREMA 5.46. (PROPRIETÀ DELLA PDF) *Sia X una variabile aleatoria assolutamente continua e sia f_X la sua densità. Allora:*

i. $\int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = 1$
ii. $\int_a^b f_X(x)\, \mathrm{d}x = F_X(b) - F_X(a)$.

Dimostrazione. Entrambe le proprietà seguono direttamente dalla definizione. Infatti

$$\int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = \lim_{x \to +\infty} F_X(x) = 1$$

per le proprietà della funzione di ripartizione e

$$\int_a^b f_X(x)\, \mathrm{d}x = \int_{-\infty}^{b} f_X(x)\, \mathrm{d}x - \int_{-\infty}^{a} f_X(x)\, \mathrm{d}x = F_X(b) - F_X(a)$$

riconducendoci alla definizione di funzione di ripartizione. □

Osservazione 5.47. A differenza della funzione di densità discreta, la funzione di densità non è necessariamente limitata all'intervallo $[0,1]$. È non negativa, per definizione, ma può assumere valori maggiori di 1, purché l'integrale su $\mathbb{R}$ sia uguale a 1.

Osservazione 5.48. Nei punti in cui la funzione di ripartizione F_X è differenziabile, il teorema fondamentale del calcolo ci dice che $F_X'(x) = f_X(x)$, cioè la densità è la derivata della funzione di ripartizione, ossia la "velocità" con cui sta cambiando la probabilità in quel punto.

Possiamo vedere la stessa cosa anche nel modo seguente, sfruttando il teorema del valor medio:

$$P(x-\varepsilon \leqslant X \leqslant x+\varepsilon) = \int_{x-\varepsilon}^{x+\varepsilon} f_X(y)\,\mathrm{d}y \approx 2\,\varepsilon \cdot f_X(x),$$

in cui $2\,\varepsilon$ è l'ampiezza dell'intervallo. Al tendere di ε a 0 abbiamo così la probabilità di un ε-intorno di x, cioè una palla di raggio ε centrata in x.

Ci possono essere punti in cui F_X non è differenziabile e, quindi, non possiamo ricavare in modo univoco f_X da F_X. Questo non è un problema se questi punti sono in numero al più numerabile, perché l'integrale ignora questi punti e quindi non hanno influsso sulla probabilità.

Esempio 5.49. (Variabile aleatoria uniforme in [0,1]) Abbiamo già visto la funzione di ripartizione della variabile aleatoria uniforme sull'intervallo $[0,1]$:

$$F_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ x & \text{se } 0 \leqslant x < 1 \\ 1 & \text{se } x \geqslant 1. \end{cases}$$

Questa funzione è derivabile in ogni punto, tranne 0 e 1, quindi potremo definire f_X su $\mathbb{R} \setminus \{0,1\}$:

$$f_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ 1 & \text{se } 0 < x < 1 \\ 0 & \text{se } x > 1 \end{cases}$$

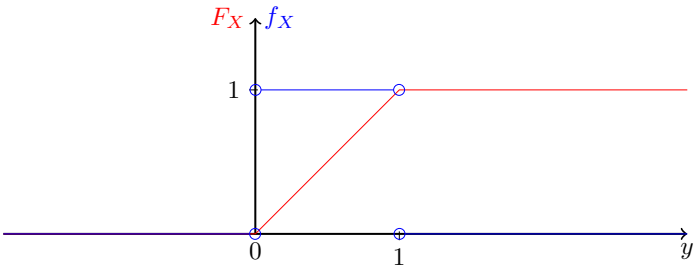


Figura 5.7. Funzione di ripartizione e densità della v.a. uniforme su $[0,1]$

Osservazione 5.50. Come la funzione di massa di probabilità φ_X non ha senso per una variabile aleatoria assolutamente continua, così la densità f_X non ha senso per le variabili aleatorie discrete.

CAPITOLO 6

TRASFORMAZIONI DI VARIABILI ALEATORIE

Abbiamo caratterizzato le variabili aleatorie come funzioni, quindi è naturale chiedersi quale sia il loro comportamento quando le trasformiamo. Ci interessa in particolare il modo in cui una trasformazione influenza la legge della variabile aleatoria.

6.1. TRASFORMAZIONI LINEARI

Cominciamo dal caso più semplice: data una variabile aleatoria X, prendiamone una trasformazione lineare, ad esempio $2X+3$. Com'è fatta questa nuova variabile aleatoria? Vediamolo in un esempio.

Esempio 6.1. Consideriamo il lancio di un dado a 4 facce, rappresentato dalla variabile aleatoria D_4. La funzione di ripartizione, come abbiamo già visto, è

$$F_{D_4}(x) = P(D_4 \leqslant x) = \begin{cases} 0 & \text{se } x < 1 \\ 1/4 & \text{se } 1 \leqslant x < 2 \\ 2/4 & \text{se } 2 \leqslant x < 3 \\ 3/4 & \text{se } 3 \leqslant x < 4 \\ 1 & \text{se } x \geqslant 4 \end{cases}$$

e la sua funzione di densità discreta φ_{D_4} è

$$\varphi_{D_4}(x) = \begin{cases} \frac{1}{4} & \text{se } x \in \{1,2,3,4\} \\ 0 & \text{altrimenti.} \end{cases}$$

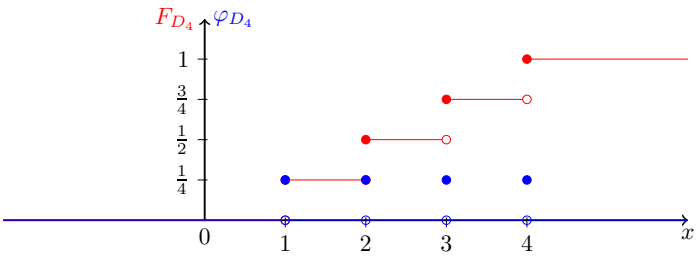


Figura 6.1. Cdf e pmf della v.a. di un dado a 4 facce D_4

Sia ora $Y = 2D_4 + 3$, quali sono i valori che può assumere[6.1]?

D_4	1	2	3	4
Y	5	7	9	11

e sia la densità sia la funzione di ripartizione sono traslate e dilatate in ascissa ma non in ordinata.

6.1. Attenzione in questo caso $2D_4$ non è da intendersi uguale al lancio di due dadi a 4 facce (cosa che sarebbe $D_4 + \widetilde{D_4}$), è un solo dado il cui risultato viene raddoppiato.

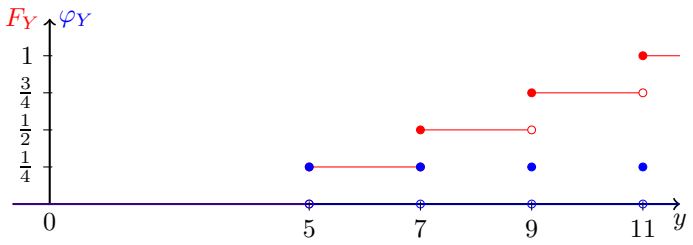


Figura 6.2. Cdf e pmf della v.a. $Y = 2D_4 + 3$

Possiamo infatti osservare che per la funzione di densità discreta

$$\varphi_Y(y) = P(Y = y) = P(2D_4 + 3 = y) = P\left(D_4 = \frac{y-3}{2}\right) = \varphi_{D_4}\left(\frac{y-3}{2}\right)$$

e, per la funzione di ripartizione

$$F_Y(y) = P(Y \leqslant y) = P(2D_4 + 3 \leqslant y) = P\left(D_4 \leqslant \frac{y-3}{2}\right) = F_{D_4}\left(\frac{y-3}{2}\right).$$

Continuiamo con un secondo esempio, questa volta usando una variabile aleatoria assolutamente continua (la sola che abbiamo incontrato finora).

Esempio 6.2. Sia X la variabile aleatoria uniforme sull'intervallo $[0,1]$. Ne conosciamo già sia la funzione di ripartizione F_X, sia la densità f_X.

Sia $Y = 2X + 3$. Non possiamo andarci a calcolare come prima i valori possibili, prendendo gli elementi (in numero finito) di $\mathcal{R}_X$ e calcolandone il valore trasformato, perché questa volta la cardinalità di $\mathcal{R}_X$ è quella del continuo. Tuttavia possiamo declinare la stessa idea: cerchiamo il supporto della densità f_Y a partire dal supporto della densità f_X. Abbiamo visto che $\mathcal{R}_X = (0,1)$. Inoltre, per definizione, il supporto $\mathcal{R}_Y$ di Y è l'insieme dei numeri reali y tali che $f_Y(y) \neq 0$ o, equivalentemente, tali che $f_{2X+3}(y) \neq 0$, o anche l'insieme dei numeri reali x tali che $f_Y(2x+3) \neq 0$.

Possiamo allora convincerci (vedremo i dettagli per assicurarcene in seguito) che il supporto di Y sia il supporto di X dilatato e traslato: $\mathcal{R}_Y = 2\,\mathcal{R}_X + 3$, ogni numero x in $(0,1)$ viene trasformato in un numero $y \in (3,5)$.

Come cambia la funzione di densità? Ci aspettiamo, essendo una trasformazione lineare, che sia qualcosa della stessa forma, quindi una costante c non nulla in $(3,5)$ e costantemente nulla altrove. La tentazione di dire $c = 1$ è forte: per le variabili aleatorie discrete abbiamo visto che la densità discreta era trasformata in ascissa ma non in ordinata. Tuttavia questo non può essere il caso, infatti deve essere, per ogni densità,

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = \int_{\mathcal{R}_X} f_X(x)\,\mathrm{d}x = 1$$

e se fosse $c = 1$ avremmo

$$\int_{-\infty}^{+\infty} f_Y(y)\,\mathrm{d}y = \int_3^5 1\,\mathrm{d}y = 2.$$

Proviamo allora a passare dalla funzione di ripartizione: come vedremo è quella la via maestra. Abbiamo infatti

$$F_Y(y) = P(Y \leqslant y) = P(2X + 3 \leqslant y) = P\left(X \leqslant \frac{y-3}{2}\right) = F_X\left(\frac{y-3}{2}\right)$$

che nel caso specifico diventa

$$F_Y(y) = \begin{cases} 0 & \text{se } \frac{y-3}{2} < 0, \text{ cioè } y < 3 \\ \frac{y-3}{2} & \text{se } 0 \leqslant \frac{y-3}{2} < 1, \text{ cioè } 3 \leqslant y < 5 \\ 1 & \text{se } \frac{y-3}{2} \geqslant 1, \text{ cioè } y \geqslant 5. \end{cases}$$

Possiamo ora ricavarci per derivazione la densità f_Y, ottenendo

$$f_Y(y) = \begin{cases} \frac{1}{2} & y \in (3,5) \\ 0 & y \in [3,5]^c \end{cases}$$

confermando quindi che il supporto di Y è la trasformazione del supporto di X, come ipotizzato. Inoltre,

$$f_Y(y) = \frac{1}{2} f_X\left(\frac{y-3}{2}\right),$$

rispetto al caso discreto compare un coefficiente.

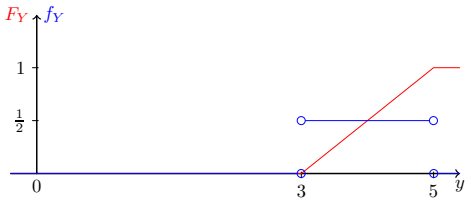


Figura 6.3. Cdf e pdf di una trasformazione lineare della v.a. uniforme su $[0,1]$

In generale vale il seguente risultato.

PROPOSIZIONE 6.3. *Sia X una variabile aleatoria e sia $Y = aX + b$ con $a \neq 0, b \in \mathbb{R}$ una sua trasformazione lineare. Allora se $a > 0$, $F_Y(y) = F_X\left(\frac{y-b}{a}\right)$, mentre se $a<0$*

$$F_Y(y) = \begin{cases} 1 - F_X\left(\frac{y-b}{a}\right) & \text{se } X \text{ è ass. continua} \\ 1 - F_X\left(\frac{y-b}{a}\right) + \varphi_X\left(\frac{y-b}{a}\right) & \text{se } X \text{ è discreta.} \end{cases}$$

Inoltre, se X è continua

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right),$$

mentre se è discreta $\varphi_Y = \varphi_X\left(\frac{y-b}{a}\right)$.

Dimostrazione. Cominciamo dalla funzione di ripartizione. Dalla definizione abbiamo

$$F_Y(y) = P(Y \leqslant y) = P(aX + b \leqslant y) = P(aX \leqslant y - b).$$

Ora dobbiamo scindere in due casi in base al segno di a. Se è positivo

$$F_Y(y) = P(aX \leqslant y - b) = P\left(X \leqslant \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right).$$

Se invece a è negativo,

$$F_Y(y) = P(aX \leqslant y - b) = P\left(X \geqslant \frac{y-b}{a}\right) = 1 - P\left(X < \frac{y-b}{a}\right) = 1 - \lim_{x \to \left(\frac{y-b}{a}\right)^-} F_X(x)$$

che per X (e dunque F_X) continua dà $F_Y(y) = 1 - F_X\left(\frac{y-b}{a}\right)$, mentre per X discreta

$$F_Y(y) = 1 - F_X\left(\frac{y-b}{a}\right) + \varphi_X\left(\frac{y-b}{a}\right).$$

Per la densità f_Y, nel caso assolutamente continuo, è sufficiente usare la regola della catena, facendo attenzione ai segni: se $a > 0$,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X\left(\frac{y-b}{a}\right) = \frac{1}{a} F_X'\left(\frac{y-b}{a}\right) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

Se invece $a<0$,

$$f_Y(y)=\frac{\mathrm{d}}{\mathrm{d}y}F_Y(y)=\frac{\mathrm{d}}{\mathrm{d}y}\left(1-F_X\left(\frac{y-b}{a}\right)\right)=-\frac{\mathrm{d}}{\mathrm{d}y}F_X\left(\frac{y-b}{a}\right)=-\frac{1}{a}f_X\left(\frac{y-b}{a}\right),$$

da cui $f_Y(y)=\frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$.

Per la densità discreta, infine,

$$\varphi_Y(y)=P(Y=y)=P(aX+b=y)=P\left(X=\frac{y-b}{a}\right)=\varphi_X\left(\frac{y-b}{a}\right)$$

in cui il fatto che ci sia l'uguaglianza rende insignificante il segno di a. □

6.1.1. La costante di rinormalizzazione

Ispirandoci a quanto abbiamo appena visto per la funzione di densità trasformata, consideriamo un problema più generale. Supponiamo di avere una funzione, quand'è che essa è la densità di una variabile aleatoria? Innanzitutto dobbiamo controllare che sia non negativa, dopodiché passiamo alla condizione sull'integrale.

Se abbiamo una funzione $f\geqslant 0$ il cui integrale su $\mathbb{R}$ è finito e positivo ma diverso da 1, possiamo ricavare da f una funzione di densità prendendo la funzione $c\cdot f$, per un'opportuna costante (positiva) c. Come facciamo a determinare questa costante? Deve essere

$$1=\int_{-\infty}^{+\infty}cf(x)\,\mathrm{d}x=c\int_{-\infty}^{+\infty}f(x)\,\mathrm{d}x,$$

quindi la scelta di c è obbligata:

$$c=\left(\int_{-\infty}^{+\infty}f(x)\,\mathrm{d}x\right)^{-1}.$$

Il nome *costante di rinormalizzazione* viene dal fatto che stiamo riscalando la funzione f in modo che il suo integrale su $\mathbb{R}$ sia 1.

Esempio 6.4. Consideriamo la funzione $f(x)=e^{-x}$ per $x\in(0,1)$ e costantemente nulla sul resto di $\mathbb{R}$. Possiamo trasformarla nella densità di una variabile aleatoria moltiplicandola per un'opportuna costante, visto che è una funzione non negativa, purché il suo integrale sia positivo.

Cominciamo allora con il calcolo di

$$\int_{-\infty}^{+\infty}f(x)\,\mathrm{d}x=\int_0^1 e^{-x}\,\mathrm{d}x=1-e^{-1}<1.$$

Allora la funzione

$$f_X(x)=\begin{cases}(1-e^{-1})^{-1}e^{-x} & x\in(0,1)\\ 0 & \text{altrimenti}\end{cases}$$

è una densità di probabilità.

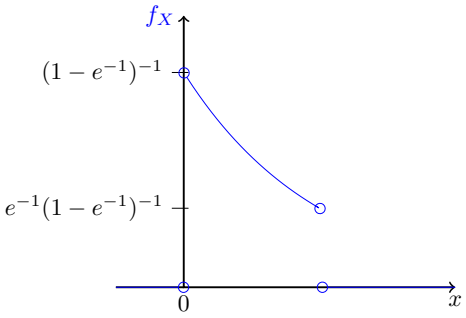


Figura 6.4. Densità di una trasformazione lineare della v.a. uniforme su $[0,1]$

Esempio 6.5. Sia $f(x)=c$ nell'intervallo $(0,\pi)$ e identicamente nulla altrimenti. Esistono (e se sì, quali sono) valori di c tali che f sia una densità di probabilità?

La funzione f è non negativa a patto che $c \geqslant 0$, restringendo quindi i potenziali valori di c. Inoltre deve essere

$$1 = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_0^{\pi} c\,\mathrm{d}x = c \cdot \pi,$$

da cui abbiamo $c = 1/\pi$. Questa è la variabile aleatoria uniforme sull'intervallo $[0, \pi]$, la cui funzione di ripartizione (che possiamo ricavare integrando f) è

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 0 & x < 0 \\ \frac{1}{\pi} x & 0 \leqslant x < \pi \\ 1 & x > \pi. \end{cases}$$

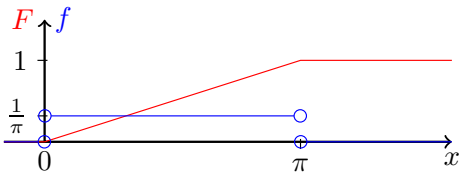


Figura 6.5. Funzione di ripartizione e densità della v.a. uniforme su $[0, \pi]$

Esempio 6.6. Sia ora $f(x) = c\,e^{-|x|}$ una funzione definita su $\mathbb{R}$. Per quali valori (eventualmente anche nessuno) di c è la densità di una variabile aleatoria?

Anche in questo caso osserviamo che necessariamente $c \geqslant 0$ per garantire la non negatività di f. Passiamo poi alla condizione sull'integrale,

$$1 = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} c\,e^{-|x|}\,\mathrm{d}x = c\left(\int_{-\infty}^{0} e^{x}\,\mathrm{d}x + \int_0^{+\infty} e^{-x}\,\mathrm{d}x\right) = c\,[e^{x}]_{-\infty}^{0} + c\,[-e^{-x}]_0^{+\infty} = 2c,$$

da cui $c = 1/2$, quindi $f(x) = \frac{1}{2} e^{-|x|}$ è una densità di probabilità. La corrispondente funzione di ripartizione è

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \frac{1}{2}\int_{-\infty}^{x} e^{-|t|}\mathrm{d}t = \begin{cases} \frac{1}{2}\int_{-\infty}^{x} e^{-t}\,\mathrm{d}t = \frac{1}{2}e^{x} & x < 0 \\ \frac{1}{2}\int_{-\infty}^{0} e^{t}\,\mathrm{d}t + \frac{1}{2}\int_0^{x} e^{-t}\,\mathrm{d}t = 1 - \frac{1}{2}e^{-x} & x > 0. \end{cases}$$

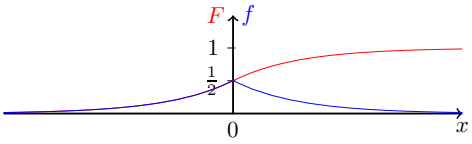


Figura 6.6. Densità di una trasformazione lineare della v.a. uniforme su $[0, 1]$

Lezione 10

6.2. TRASFORMAZIONI NON LINEARI

Anche in questo caso abbiamo una variabile aleatoria X, di cui conosciamo la legge, ossia la funzione di ripartizione F_X. Questo è essenzialmente equivalente al conoscerne la densità f_X, nel caso di variabili aleatorie assolutamente continue, o la densità discreta φ_X, nel caso di variabili aleatorie discrete.

Ora, però, invece di una trasformazione lineare abbiamo una funzione $g : \mathbb{R} \to \mathbb{R}$, che supponiamo nonlineare (se fosse lineare ricadremmo nel caso precedente), ad esempio $g(x) = \sqrt{|x|}$ o

$$g(x) = \begin{cases} 3x^2 + \log(x) & x > 0 \\ 0 & x \leqslant 0. \end{cases}$$

L'obiettivo è il medesimo di prima: determinare la legge della variabile aleatoria $Y = g(X)$.

Per le variabili aleatorie discrete le cose sono anche in questo caso molto semplici: dobbiamo solamente fare attenzione al fatto che g non è necessariamente iniettiva o suriettiva, quindi ogni valore di y può avere nessuna, una o più di una preimmagine rispetto a g. Ogni y che è immagine di almeno un punto $x \in \mathcal{R}_X$ eredita da ogni sua preimmagine la corrispondente probabilità:

$$\varphi_Y(y) = \sum_{x \in g^{-1}(\{y\})} \varphi_X(x).$$

Il supporto di Y è l'immagine mediante g del supporto di X, $\mathcal{R}_Y = g(\mathcal{R}_X)$ e la funzione di ripartizione si ricava dalla densità discreta.

Se invece X è assolutamente continua abbiamo in questo caso più generale rispetto a quello lineare almeno due modi di farlo, ciascuno coi suoi pro e i suoi contro[6.2]:

1. Possiamo ricavare la legge di Y sfruttando la forma della variabile aleatoria X (in particolare la forma della sua funzione di ripartizione F_X) e della funzione g. Questa è una strategia che richiede di adattarsi alla specifica coppia (X, g) che consideriamo. Spesso è facile, ma è anche facile sbagliare.
2. Possiamo usare un teorema generale. Purtroppo le ipotesi del teorema non sono sempre soddisfatte e, anche quando lo sono, l'applicazione del teorema può essere difficile.

Vediamo la prima strategia, necessariamente, con due esempi.

Esempio 6.7. Sia X una variabile aleatoria (assolutamente continua) di densità

$$f_X(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

e sia $g : \mathbb{R} \to \mathbb{R}$ la funzione $g(x) = e^{-x}$. Vogliamo determinare la legge della variabile aleatoria $Y = e^{-X}$.

Cominciamo dalla definizione:

$$F_Y(y) = P(Y \leqslant y) = P(e^{-X} \leqslant y) = \begin{cases} P(-X \leqslant \log y) = P(X \geqslant -\log y) & y > 0 \\ 0 & y \leqslant 0 \end{cases}$$

dove abbiamo usato, oltre alle definizioni, la monotonia crescente del logaritmo.

Ora per proseguire ci occorre la funzione di ripartizione di X, che possiamo ricavare da f_X:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_0^x e^{-t}\,\mathrm{d}t = 1 - e^{-x}, \qquad x > 0$$

e identicamente nulla altrimenti. Allora

$$F_Y(y) = \begin{cases} 0 & y \leqslant 0 \\ 1 - (1 - e^{-(-\log y)}) = y & 0 < y < 1 \\ 1 & y \geqslant 1 \end{cases}$$

in cui l'ultimo caso si verifica quando $y > 0$ e $-\log y \leqslant 0$. Quindi in questo caso Y è la variabile aleatoria uniforme su $[0, 1]$.

Se volessimo avere anche f_Y, potremmo farlo derivando F_Y, oppure, se non avessimo calcolato F_Y derivandola in astratto. In questo secondo caso abbiamo (per $y > 0$)

$$\begin{aligned} f_Y(y) &= F_Y'(y) = \frac{\mathrm{d}}{\mathrm{d}y}\,(1 - F_X(-\log y)) \\ &= -F_X'(-\log y)\,\frac{\mathrm{d}}{\mathrm{d}y}\,(-\log y) = -f_X(-\log y)\left(-\frac{1}{y}\right) \\ &= \frac{1}{y} f_X(-\log y) \\ \text{per } -\log y > 0 \Leftrightarrow y < 1 &= \frac{1}{y} e^{-(-\log y)} = 1. \end{aligned}$$

6.2. Contrariamente a quanto si pensa, raramente in matematica esiste una sola ricetta per risolvere problemi.

Riassumendo abbiamo

$$f_Y(y) = \begin{cases} 0 & y<0 \\ 1 & 0<y<1 \\ 0 & y>1. \end{cases}$$

Esempio 6.8. Sia X una variabile aleatoria (assolutamente continua) di densità

$$f_X(x) = \begin{cases} e^{-x} & x>0 \\ 0 & x\leqslant 0 \end{cases}$$

e sia $Z=(1-X)^2$, cioè $g(x)=(1-x)^2$. Vogliamo determinare la legge di Z.

Come prima cosa, consideriamo la funzione di ripartizione F_Z,

$$F_Z(z) = P(Z\leqslant z) = P((1-X)^2\leqslant z) = \begin{cases} P(|1-X|\leqslant\sqrt{z}) & z>0 \\ 0 & z\leqslant 0 \end{cases}$$

in cui $z=0$ può stare equivalentemente sopra o sotto, tanto la probabilità di un punto è nulla, siccome la variabile è continua. Proseguendo, nel caso $z>0$, abbiamo

$$\begin{aligned} P(|1-X|\leqslant\sqrt{z}) &= P(-\sqrt{z}\leqslant 1-X\leqslant\sqrt{z}) \\ &= P(1+\sqrt{z}\geqslant X\geqslant 1-\sqrt{z}) \\ &= F_X(1+\sqrt{z}) - F_X(1-\sqrt{z}) \\ &= (1-e^{-(1+\sqrt{z})})\,\mathbb{1}_{\{1+\sqrt{z}>0\}} - (1-e^{-(1-\sqrt{z})})\,\mathbb{1}_{\{1-\sqrt{z}>0\}} \end{aligned}$$

e, riassumendo,

$$F_Z(z) = \begin{cases} 0 & z\leqslant 0 \\ e^{-(1-\sqrt{z})} - e^{-(1+\sqrt{z})} & 0<z<1 \\ 1-e^{-(1+\sqrt{z})} & z\geqslant 1. \end{cases}$$

Passiamo alla densità f_Z. Avendo la forma esplicita di F_Z possiamo ricavarla derivando direttamente quest'ultima, ma se non l'avessimo già calcolata potremmo ricondurci a f_X nel modo seguente, per $z>0$,

$$\begin{aligned} f_Z(z) &= f_X(1+\sqrt{z})\frac{\mathrm{d}}{\mathrm{d}z}(1+\sqrt{z}) - f_X(1-\sqrt{z})\frac{\mathrm{d}}{\mathrm{d}z}(1-\sqrt{z}) \\ &= \frac{1}{2\sqrt{z}}\left(f_X(1+\sqrt{z}) + f_X(1-\sqrt{z})\right). \end{aligned}$$

Ora possiamo andare a inserire la forma esplicita di f_X, facendo attenzione al suo dominio, in particolare nel secondo addendo. Abbiamo

$$f_Z(z) = \begin{cases} 0 & z<0 \\ \frac{1}{2\sqrt{z}}\left(e^{-(1+\sqrt{z})} + e^{-(1-\sqrt{z})}\right) & 0<z<1 \\ \frac{1}{2\sqrt{z}}e^{-(1+\sqrt{z})} & z>1. \end{cases}$$

Anche in questo caso il valore della densità in 0 e 1 non è rilevante. È invece un esercizio di Analisi verificare che la funzione f_Z appena scritta sia una densità di probabilità, cioè che sia non negativa e abbia integrale uguale a 1.

Passiamo ora alla seconda strategia. Essa si basa sul seguente risultato.

TEOREMA 6.9. (CAMBIO DI VARIABILE) *Sia X una variabile aleatoria assolutamente continua, di densità f_X. Sia inoltre $Y=g(X)$, con $g:\mathbb{R}\to\mathbb{R}$ funzione C^1 a tratti e tale che $P(g'(X)=0)=0$. Allora*

$$f_Y(y) = \sum_{\{x\in g^{-1}(\{y\})\}} \frac{f_X(x)}{|g'(x)|}.$$

Osservazione 6.10. Cerchiamo di capire cosa significa la condizione $P(g'(X)=0)=0$ nel Teorema 6.9. La scrittura $g'(X)=0$ rappresenta un insieme, in particolare

$$\{g'(X)=0\}=\{\omega\in\Omega: g'(X(\omega))=0\}=\bigcup_{x:g'(x)=0}\{\omega\in\Omega:\omega\in X^{-1}(\{x\})\}$$

sono quindi tutti quegli esiti che finiscono, attraverso X, nei punti in cui si annulla la derivata di g. Stiamo quindi chiedendo che g' si annulli su insiemi di $\mathbb{R}$ in cui X ha valore con probabilità 0. Possono essere punti, dunque, anche in quantità numerabile, ma in genere non intervallini. In particolare, nella somma possiamo trascurare eventuali punti x in cui il denominatore si annulla.

Inoltre possiamo osservare che l'insieme $\{x\in g^{-1}(\{y\})\}=\{x: g(x)=y\}$, grazie a questa ipotesi sugli zeri di g', ha un numero finito di elementi, quindi la somma è ben definita.

Vediamo come usare il Teorema 6.9 in un esempio.

Esempio 6.11. Rimettiamoci nello stesso caso dell'Esempio 6.8. La funzione $g:\mathbb{R}\to\mathbb{R}$ è una funzione C^1 e la sua derivata $g'(x)=2x-2$ si annulla solamente in $x=1$, ma per la forma della variabile aleatoria X la probabilità che $X=1$ è nulla.

Siamo allora nelle ipotesi del Teorema 6.9. Per usarne il risultato, come prima cosa andiamo a studiare come sono fatti gli insiemi $g^{-1}(\{z\})$ al variare di $z\in\mathbb{R}$. Abbiamo

$$g^{-1}(\{z\})=\begin{cases}\emptyset & z<0\\ \{1\} & z=0\\ \{1-\sqrt{z},1+\sqrt{z}\} & z>0.\end{cases}$$

Se ora passiamo a f_Z abbiamo, dal Teorema 6.9

$$f_Z(z)=\begin{cases}0 & z<0\\ \frac{f_X(1)}{|g'(1)|}=+\infty & z=0\\ \frac{f_X(1-\sqrt{z})}{|g'(1-\sqrt{z})|}+\frac{f_X(1+\sqrt{z})}{|g'(1+\sqrt{z})|} & z>0.\end{cases}$$

La prima parte, per $z<0$, è a posto. Per $z=0$ abbiamo un momento di fastidio, ma poi pensiamo al fatto che non ci interessa il valore in un singolo punto, per f_Z: possiamo non definirla in $z=0$. Resta da sistemare l'ultimo caso, $z>0$. In tal caso

$$f_Z(z)=\frac{f_X(1-\sqrt{z})}{|g'(1-\sqrt{z})|}+\frac{f_X(1+\sqrt{z})}{|g'(1+\sqrt{z})|}=\begin{cases}\frac{e^{-(1-\sqrt{z})}}{2|1-\sqrt{z}-1|}+\frac{e^{-(1+\sqrt{z})}}{2|1+\sqrt{z}-1|} & 1-\sqrt{z}>0\\ \frac{e^{-(1+\sqrt{z})}}{2|1+\sqrt{z}-1|} & 1-\sqrt{z}<0\end{cases}$$

e, mettendo assieme tutti i pezzi, otteniamo che

$$f_Z(z)=\begin{cases}0 & z<0\\ \frac{e^{-(1-\sqrt{z})}}{2\sqrt{z}}+\frac{e^{-(1+\sqrt{z})}}{2\sqrt{z}} & 0<z<1\\ \frac{e^{-(1+\sqrt{z})}}{2\sqrt{z}} & z>1.\end{cases}$$

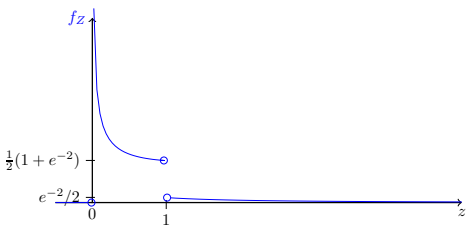


Figura 6.7. Densità della variabile aleatoria Z

CAPITOLO 7

VETTORI ALEATORI

Finora abbiamo considerato le variabili aleatorie una per volta. Tuttavia potremmo avere, sullo spazio di probabilità $(\Omega, \mathcal{F}, P)$ due variabili aleatorie X, Y che vogliamo trattare assieme, ad esempio perché ci interessa conoscere la probabilità che $X < Y$, oppure che $|X+Y| > 1$. Se ci pensiamo questo non è molto diverso dal cercare la probabilità di $g(X) < \alpha$, per qualche funzione g e qualche valore α.

Ad esempio, possiamo prendere $g(x) = |x|$ e $\alpha = 1$ e riscrivere $P(g(X) \leqslant \alpha) = P(|X| \leqslant 1)$ come

$$P(|X| \leqslant 1) = P(-1 \leqslant X \leqslant 1) = F_X(1) - F_X(-1) + P(X = -1).$$

Abbiamo calcolato la probabilità di tutti gli ω in Ω tali che $X(\omega)$ sia nell'intervallo chiuso $[-1,1]$. In maniera del tutto analoga, calcolare $P(X < Y)$ significherà trovare la probabilità di tutti quegli ω tali che $X(\omega) < Y(\omega)$. Gli esiti ω però devono essere gli stessi in contemporanea in X e Y, quindi a priori non possiamo usare separatamente le leggi di X e Y. Procediamo per passi.

DEFINIZIONE 7.1. *Dati uno spazio di probabilità $(\Omega, \mathcal{F}, P)$ e due variabili aleatorie X e Y su di esso, si chiama* coppia di variabili aleatorie *o* variabile aleatoria doppia *o* vettore aleatorio di dimensione 2 *la funzione $V : \Omega \to \mathbb{R}^2$ definita da $V(\omega) = (X(\omega), Y(\omega))$. Il vettore aleatorio V ha supporto*

$$\mathcal{R}_V = \mathcal{R}_{X,Y} = \mathcal{R}_X \times \mathcal{R}_Y = \{(x,y) \in \mathbb{R}^2 : x \in \mathcal{R}_X, y \in \mathcal{R}_Y\}.$$

Osservazione 7.2. Possiamo pensare a una vettore aleatorio di dimensione 2 come a una variabile aleatoria a valori sul piano $\mathbb{R}^2$ invece che sulla retta $\mathbb{R}$. Un singolo esito ω viene mandato dal vettore in un punto del piano.

In modo del tutto analogo possiamo definire e studiare vettori aleatori di dimensione $d \geqslant 1$.

DEFINIZIONE 7.3. *Data una variabile aleatoria doppia (X, Y), la sua* funzione di ripartizione *è*

$$F_{X,Y}((x,y)) = F_{X,Y}(x,y) = P(X \leqslant x, Y \leqslant y).$$

Tale funzione $F_{X,Y}$ si dice anche funzione di ripartizione congiunta di X e Y.

In maniera del tutto analoga possiamo definire la funzione di ripartizione congiunta per vettori aleatori d-dimensionali.

Osservazione 7.4. In generale non è sufficiente conoscere le funzioni di ripartizione F_X ed F_Y per conoscere la funzione di ripartizione congiunta $F_{X,Y}$. Viceversa, nota $F_{X,Y}$ possiamo ricavare da essa F_X ed F_Y, che in questo caso prendono il nome di *funzioni di ripartizione marginali*. Infatti

$$\begin{aligned} F_X(x) &= P(X \leqslant x, \forall Y) \\ &= P(X \leqslant x, Y < +\infty) \\ &= \lim_{y \to +\infty} F_{X,Y}(x,y) \end{aligned}$$

e analogamente $F_Y(y) = \lim_{x \to +\infty} F_{X,Y}(x,y)$.

DEFINIZIONE 7.5. *Data una variabile aleatoria doppia (X, Y) si dice* funzione di ripartizione di X condizionata a Y *la funzione $F_{X|Y}(x \mid y) := \frac{F_{X,Y}(x,y)}{F_Y(y)}$.*

Questa funzione è la probabilità dell'evento che immaginiamo: $F_{X|Y}(x \mid y) = P(X \leqslant x \mid Y \leqslant y)$. Non ci sorprende dunque la prossima definizione.

DEFINIZIONE 7.6. *Dati uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$ *e due sottotribù* $\mathcal{F}_1$ *ed* $\mathcal{F}_2$ *di* $\mathcal{F}$*, diciamo che* $\mathcal{F}_1$ *ed* $\mathcal{F}_2$ *sono* indipendenti *se ogni evento in* $\mathcal{F}_1$ *è indipendente da ogni evento di* $\mathcal{F}_2$*, ossia se per ogni* $E_1 \in \mathcal{F}_1$ *ed* $E_2 \in \mathcal{F}_2$*,* $P(E_1 \cap E_2) = P(E_1) \cdot P(E_2)$*.*

DEFINIZIONE 7.7. *Dati uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$ *e due variabili aleatorie* X *e* Y *su di esso, tali variabili aleatorie sono* indipendenti *se lo sono le tribù* $\sigma(X)$ *e* $\sigma(Y)$ *da esse generate.*

La definizione di indipendenza tra variabili aleatorie ha lo svantaggio di non essere molto pratica nelle applicazioni, perché per essere verificata richiede di controllare che tutte le coppie di eventi nel prodotto delle tribù generate siano indipendenti. Esistono però delle condizioni equivalenti, di più facile uso.

PROPOSIZIONE 7.8. *Due variabili aleatorie* X *e* Y *sullo stesso spazio di probabilità sono indipendenti se e solo se per ogni* $(x, y) \in \mathbb{R}^2$*,* $F_{X,Y}(x, y) = F_X(x) \cdot F_Y(y)$*.*

Dimostrazione. Mostriamo l'implicazione diretta $\Rightarrow$. Abbiamo

$$F_{X,Y}(x, y) = P(\{X \leqslant x\} \cap \{Y \leqslant y\}) = P(X \leqslant x)\, P(Y \leqslant y) = F_X(x)\, F_Y(y),$$

in cui nella seconda uguaglianza abbiamo usato che $\{X \leqslant x\} \in \sigma(X)$, $\{Y \leqslant y\} \in \sigma(Y)$ e che, per definizione di indipendenza, le tribù generate sono equivalenti.

L'implicazione inversa $\Leftarrow$ è una conseguenza non banale del teorema di Carathéodory e non viene affrontata in questo corso. □

PROPOSIZIONE 7.9. *Due variabili aleatorie* X *e* Y *sullo stesso spazio di probabilità sono indipendenti se e solo se per ogni* $(x, y) \in \mathbb{R}^2$*,* $F_X(x) = F_{X|Y}(x \,|\, y)$ *e* $F_Y(y) = F_{Y|X}(y \,|\, x)$*.*

Osservazione 7.10. Se abbiamo più variabili aleatorie, cioè se abbiamo un vettore aleatorio di dimensione $d \geqslant 3$, per avere l'indipendenza dobbiamo considerare tutti i raggruppamenti possibili (a 2 a 2, a 3 a 3 e così via) come già visto per gli eventi.

7.1. VETTORI ALEATORI DISCRETI

Quanto detto finora sulle coppie di variabili aleatorie riguardava solamente le funzioni di ripartizione, quindi vale tanto per le variabili aleatorie discrete quanto per quelle assolutamente continue (e anche per quelle miste). Nel caso in cui entrambe le variabili aleatorie X e Y siano discrete, possiamo indagare più a fondo, facendo entrare in gioco le densità discrete.

DEFINIZIONE 7.11. *Siano* X *e* Y *due variabili aleatorie discrete sullo stesso spazio di probabilità* $(\Omega, \mathcal{F}, P)$*. Si dice* densità discreta congiunta di X e Y *la funzione* $\varphi_{X,Y} \colon \mathbb{R}^2 \to [0, 1]$ *definita da*

$$\varphi_{X,Y}(x, y) = P(X = x, Y = y).$$

Si dice inoltre densità condizionale di X data Y *la funzione*

$$\varphi_{X|Y}(x \,|\, y) = \begin{cases} P(X = x \,|\, Y = y) & y \in \mathcal{R}_Y \\ 0 & y \in \mathcal{R}_Y^c. \end{cases}$$

Osservazione 7.12. Dalla definizione di densità discreta congiunta ricaviamo immediatamente le seguenti proprietà:

- per ogni (x, y) in $\mathbb{R}^2$, $0 \leqslant \varphi_{X,Y}(x, y) \leqslant 1$
- $\varphi_{X,Y}(x, y) = 0$ sui valori impossibili, cioè se $x \in \mathcal{R}_X^c$ o $y \in \mathcal{R}_Y^c$
- vale l'identità

$$\sum_{(x,y) \in \mathbb{R}^2} \varphi_{X,Y}(x, y) = 1.$$

Forse la sola cosa che ci può sorprendere è il fatto che abbiamo scritto una somma su tutte le coppie in $\mathbb{R}^2$, nell'ultima identità. Ma questo abuso di notazione è innocuo, dal momento che tranne che per un numero finito o numerabile di valori di x e di y e dunque di coppie (x,y), la funzione $\varphi_{X,Y}$ è identicamente nulla. In effetti possiamo scrivere

$$\sum_{(x,y)\in\mathbb{R}^2} \varphi_{X,Y}(x,y) = \sum_{x\in\mathcal{R}_X, y\in\mathcal{R}_Y} \varphi_{X,Y}(x,y) = 1.$$

Vediamo ora alcune proprietà delle coppie di variabili aleatorie discrete.

PROPOSIZIONE 7.13. *Sia (X,Y) una coppia di variabili aleatorie. Valgono le seguenti uguaglianze:*

i. per ogni $(x,y)\in\mathbb{R}^2$,

$$F_{X,Y}(x,y) = \sum_{(\xi,\eta)\in\mathcal{R}_{X,Y}} \mathbb{1}_{\{\xi\leqslant x\}}\,\mathbb{1}_{\{\eta\leqslant y\}}\,\varphi_{X,Y}(\xi,\eta)$$

ii. per ogni $(x,y)\in\mathbb{R}^2$, $\varphi_{X,Y}(x,y) = \varphi_{X|Y}(x\,|\,y)\,\varphi_Y(y)$

iii. per ogni $x\in\mathbb{R}$, $\varphi_X(x) = \sum_{y\in\mathcal{R}_Y}\varphi_{X,Y}(x,y)$

iv. le variabili aleatorie X e Y sono indipendenti se e solo se, per ogni $(x,y)\in\mathcal{R}_{X,Y}$, $\varphi_{X,Y}(x,y) = \varphi_X(x)\,\varphi_Y(y)$

v. le variabili aleatorie X e Y sono indipendenti se e solo se, per ogni $(x,y)\in\mathcal{R}_{X,Y}$, $\varphi_X(x) = \varphi_{X|Y}(x\,|\,y)$ e $\varphi_Y(y) = \varphi_{Y|X}(y\,|\,x)$.

Lezione 11

Dimostrazione. Procediamo in ordine.

i. Segue immediatamente dalle definizioni.

ii. Se $y\in\mathcal{R}_Y^c$ l'identità è immediata, se $y\in\mathcal{R}_Y$ abbiamo

$$\begin{aligned}\varphi_{X|Y}(x\,|\,y)\,\varphi_Y(y) &= P(X=x\,|\,Y=y)\,P(Y=y)\\ &= \frac{P(X=x,Y=y)}{P(Y=y)}\,P(Y=y)\\ &= \varphi_{X,Y}(x,y).\end{aligned}$$

iii. Iniziamo riscrivendo il secondo membro, sfruttando l'identità appena mostrata:

$$\begin{aligned}\sum_{y\in\mathcal{R}_Y}\varphi_{X,Y}(x,y) &= \sum_{y\in\mathcal{R}_Y}\varphi_{X|Y}(x\,|\,y)\,\varphi_X(x)\\ &= \varphi_X(x)\sum_{y\in\mathcal{R}_Y}\varphi_{X|Y}(x\,|\,y)\\ &= \varphi_X(x)\sum_{y\in\mathcal{R}_Y}P(Y=y\,|\,X=x)\\ &= \varphi_X(x),\end{aligned}$$

poiché $P(\cdot\,|\,X=x)$ è una probabilità e stiamo sommando su tutti i possibili eventi disgiunti.

iv. L'implicazione $\Rightarrow$ è immediata dalle definizioni. Viceversa, l'implicazione $\Leftarrow$ si ottiene usando la prima proprietà e l'analogo risultato visto per le funzioni di ripartizione.

v. Segue dalla iv. e dalla ii. □

Osservazione 7.14. Vale la pena osservare che, se è nota $\varphi_{X,Y}$, la proprietà iv. è molto pratica per verificare (o confutare) l'indipendenza di X e Y.

Esempio 7.15. Abbiamo due variabili aleatorie X e Y, entrambe discrete. La variabile X descrive il lancio di una moneta bilanciata, mentre Y è il lancio di un dado a 6 facce se $X=0$ e il lancio di un dado a 8 facce se $X=1$. Vogliamo ottenere la legge di Y.

Come prima cosa vogliamo scrivere in modo preciso i dati del problema:

$$\varphi_{Y|X}(y\,|\,0) = \begin{cases}1/6 & y\in\{1,\dots,6\}\\ 0 & \text{altrimenti}\end{cases} \qquad \varphi_{Y|X}(y\,|\,1) = \begin{cases}1/8 & y\in\{1,\dots,8\}\\ 0 & \text{altrimenti,}\end{cases}$$

che potremmo anche scrivere in modo più compatto come

$$\varphi_{Y|X}(y\,|\,x)=\begin{cases}1/6 & x=0, y\in\{1,\dots,6\}\\ 1/8 & x=1, y\in\{1,\dots,8\}\\ 0 & \text{altrimenti.}\end{cases}$$

Poi andiamo a ricavarci la densità discreta congiunta, ricordando che $\varphi_{X,Y}(x,y)=\varphi_{Y|X}(y\,|\,x)\,\varphi_X(x)$:

$$\varphi_{X,Y}(x,y)=\begin{cases}1/12 & x=0, y\in\{1,\dots,6\}\\ 1/16 & x=1, y\in\{1,\dots,8\}\\ 0 & \text{altrimenti.}\end{cases}$$

A questo punto abbiamo tutti gli ingredienti necessari per calcolare la densità discreta di Y, sommando su tutti i possibili valori di X:

$$\varphi_Y(y)=\sum_{x\in\mathcal{R}_X}\varphi_{X,Y}(x,y)=\begin{cases}7/48 & y\in\{1,\dots,6\}\\ 1/16 & y\in\{7,8\}\\ 0 & \text{altrimenti.}\end{cases}$$

A questo punto possiamo anche chiederci se le variabili aleatorie X e Y siano o meno indipendenti. Dal momento che conosciamo la densità discreta congiunta ed entrambe le densità marginali, è sufficiente verificare se $\varphi_{X,Y}(x,y)=\varphi_X(x)\,\varphi_Y(y)$, ma

$$\varphi_X(x)\,\varphi_Y(y)=\begin{cases}7/96 & x=0, y\in\{1,\dots,6\}\\ 7/96 & x=1, y\in\{1,\dots,6\}\\ 1/32 & x=1, y\in\{7,8\}\\ 0 & \text{altrimenti}\end{cases}\neq\varphi_{X,Y}(x,y),$$

quindi (come potevamo aspettarci, vista la definizione di Y) X e Y non sono indipendenti tra loro.

Data una coppia di variabili aleatorie, in molti casi siamo interessati a qualche loro funzione. Un esempio, semplice ma molto utile, è la somma di due variabili aleatorie, che vediamo, nel caso discreto, nel seguente risultato.

PROPOSIZIONE 7.16. (SOMMA DI VARIABILI ALEATORIE DISCRETE) *Siano X e Y due variabili aleatorie sullo stesso spazio di probabilità $(\Omega,\mathcal{F},P)$ con densità congiunta $\varphi_{X,Y}$. La loro somma ha densità discreta*

$$\varphi_{X+Y}(z)=\sum_{x\in\mathcal{R}_X}\varphi_{X,Y}(x,z-x).$$

Dimostrazione. Dalle definizioni abbiamo

$$\begin{aligned}\varphi_{X+Y}(z) &= P(X+Y=z)\\ &= P\left(\bigcup_{x\in\mathcal{R}_X}\{X=x, X+Y=z\}\right)\\ &= P\left(\bigcup_{x\in\mathcal{R}_X}\{X=x, Y=z-x\}\right)\\ &= \sum_{x\in\mathcal{R}_X}P(X=x, Y=z-x)\\ &= \sum_{x\in\mathcal{R}_X}\varphi_{X,Y}(x,z-x)\end{aligned}$$

e abbiamo così la densità discreta della variabile aleatoria $Z=X+Y$. □

Osservazione 7.17. Se le variabili aleatorie X e Y sono indipendenti, allora

$$\varphi_{X+Y}(z)=\sum_{x\in\mathcal{R}_X}\varphi_X(x)\,\varphi_Y(z-x).$$

Esempio 7.18. Siano X e Y due variabili aleatorie (indipendenti) che descrivono ciascuna il lancio di un dado a 10 facce. Indichiamo con $S = X + Y$ la loro variabile aleatoria somma. Vogliamo scrivere la densità discreta congiunta di S e X, $\varphi_{S,X}(s,x)$ e la densità discreta condizionata di S data X, $\varphi_{S|X}(s\,|\,x)$.

Partendo dalla definizione,

$$\begin{aligned}\varphi_{S,X}(s,x) &= P(S=s, X=x) = P(X+Y=s, X=x)\\ &= P(Y=s-x, X=x) = \varphi_{X,Y}(x, s-x) = \varphi_X(x)\,\varphi_Y(s-x),\end{aligned}$$

dove nell'ultimo passaggio abbiamo sfruttato il fatto che X e Y siano indipendenti.

Passiamo alla densità discreta condizionata,

$$\varphi_{S|X}(s\,|\,x) = \frac{\varphi_{S,X}(s,x)}{\varphi_X(x)} = \varphi_Y(s-x).$$

Finora non abbiamo usato la particolare forma delle densità discrete di X e Y:

$$\varphi_X(x) = \varphi_Y(x) = \begin{cases} 1/10 & x \in \{1,\dots,10\} \\ 0 & \text{altrimenti,}\end{cases}$$

se andiamo a scriverle nelle identità ottenute sopra, abbiamo

$$\varphi_{S,X}(s,x) = \begin{cases} 1/100 & s \in \{2,\dots,20\}, x \in \{1,\dots,s-1\} \\ 0 & \text{altrimenti}\end{cases}$$

in cui la prima riga cattura tutti i possibili risultati della somma in cui x sia un addendo ammissibile. Per la densità discreta condizionata di S data X,

$$\varphi_{S|X}(s\,|\,x) = \begin{cases} 1/10 & x \in \{1,\dots,10\}, s \in \{x+1,\dots,x+10\} \\ 0 & \text{altrimenti.}\end{cases}$$

Per quanto riguarda la densità discreta di S abbiamo

$$\varphi_S(s) = \sum_{x=1}^{10} \varphi_{S,X}(s,x) = \begin{cases} 1/100 & s=2 \\ 2/100 & s=3 \\ \quad\vdots & \quad\vdots \\ 10/100 & s=11 \\ \quad\vdots & \quad\vdots \\ 1/100 & s=20 \\ 0 & s \in \{2,\dots,20\}^c.\end{cases}$$

7.2. VETTORI ALEATORI ASSOLUTAMENTE CONTINUI

Dopo aver visto il caso particolare di coppie di variabili aleatorie discrete, consideriamo ora le coppie di variabili aleatorie assolutamente continue. Come osservato, per le funzioni di ripartizione la teoria non dipende dal particolare tipo di variabile aleatoria considerato, quindi quello che vedremo, specifico per le assolutamente continue, sarà legato alle densità di probabilità.

DEFINIZIONE 7.19. *Siano X e Y due variabili aleatorie assolutamente continue sullo stesso spazio di probabilità $(\Omega, \mathcal{F}, P)$. Chiamiamo* densità congiunta di X e Y *la funzione $f_{X,Y}\colon \mathbb{R}^2 \to \mathbb{R}$ tale che per ogni evento E nella tribù prodotto $\mathcal{B} \otimes \mathcal{B}$*

$$P((X,Y) \in E) = \iint_E f_{X,Y}(s,t)\,\mathrm{d}s\,\mathrm{d}t.$$

Osservazione 7.20. C'è una sola probabilità e non una probabilità prodotto perché non stiamo considerando coppie di esiti, ma un solo esito, sul quale costruiamo la coppia $(X(\omega), Y(\omega))$:

$$P((X,Y) \in E) = P(\{\omega \in \Omega\colon (X(\omega), Y(\omega)) \in E\}).$$

In effetti la tribù prodotto è in $\mathbb{R}^2$, non in $\Omega \times \Omega$.

PROPOSIZIONE 7.21. *Siano X e Y due variabili aleatorie assolutamente continue sullo stesso spazio di probabilità $(\Omega, \mathcal{F}, P)$. Allora:*

i. per ogni $(x,y) \in \mathbb{R}^2$,

$$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\,\mathrm{d}t\,\mathrm{d}s$$

ii. per ogni $x, y \in \mathbb{R}$ possiamo scrivere le densità marginali di X e Y come

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,t)\,\mathrm{d}t \qquad\qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(s,y)\,\mathrm{d}s$$

iii. X e Y sono indipendenti se e solo se per ogni $(x,y) \in \mathbb{R}^2$, $f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$.

Dimostrazione. La prima uguaglianza, che possiamo anche scrivere in forma differenziale come

$$f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x\,\partial y}(x,y)$$

segue immediatamente dalla definizione di densità congiunta e di funzione di ripartizione. La seconda coppia di uguaglianze è un'applicazione del teorema del calcolo integrale. Per quanto riguarda la terza proprietà, l'implicazione diretta $\Rightarrow$ segue dalla definizione, mentre quella inversa $\Leftarrow$ si mostra passando dalle corrispondenti proprietà della funzione di ripartizione. □

Osservazione 7.22. Anche nel caso assolutamente continuo abbiamo l'analogo dell'Osservazione 7.12, cioè alcune proprietà immediate della funzione di densità congiunta. Abbiamo infatti, per ogni $(x,y) \in \mathbb{R}^2$, che $f_{X,Y}(x,y) \geqslant 0$. Osserviamo però che, a differenza di $\varphi_{X,Y}$, non abbiamo un limite dall'alto del valore della densità congiunta, in analogia a quanto visto per la densità di una variabile aleatoria assolutamente continua (anch'essa non negativa) e la densità discreta di una discreta (la cui immagine è contenuta in $[0,1]$).

Inoltre, vale l'identità

$$\iint_{\mathbb{R}^2} f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = 1,$$

da cui possiamo sviluppare un discorso sulle costanti di rinormalizzazione analogo a quello fatto in precedenza.

Esempio 7.23. Sia $f_{X,Y}(x,y) = e^{-x}$ per $0 \leqslant y \leqslant x$ e nulla altrimenti. Vogliamo la densità marginale f_X di X.

Per ottenerla ci basta integrare la densità congiunta su tutti i possibili valori di y,

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^x e^{-x}\,\mathrm{d}y = x\,e^{-x}.$$

Se vogliamo anche la funzione di ripartizione di X, possiamo ottenerla come

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = 1 - (x+1)\,e^{-x}$$

per $x > 0$ (e 0 altrimenti), oppure direttamente dalla densità congiunta,

$$F_X(x) = \lim_{y\to+\infty} F_{X,Y}(x,y) = \lim_{y\to+\infty} \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s,t)\,\mathrm{d}t\,\mathrm{d}s = \int_{-\infty}^{x} \int_{-\infty}^{+\infty} f_{X,Y}(s,t)\,\mathrm{d}t\,\mathrm{d}s.$$

DEFINIZIONE 7.24. *Siano X e Y due variabili aleatorie assolutamente continue sullo stesso spazio probabilità $(\Omega, \mathcal{F}, P)$. Chiamiamo* densità condizionale di X rispetto a Y *la funzione $f_{X|Y}$ definita come*

$$f_{X|Y}(x\,|\,y) = \frac{f_{X,Y}(x,y)}{f_Y(y)},$$

per $y \in \mathcal{R}_Y$ e identicamente nulla altrimenti.

Osservazione 7.25. Anche in questo caso possiamo ricavare dalla densità condizionale e dalla densità marginale di Y la densità congiunta:

$$f_{X,Y}(x,y) = f_{X|Y}(x\,|\,y)\,f_Y(y).$$

Osservazione 7.26. Se guardiamo $f_{X|Y}$ come funzione della sola x, per y fissato, abbiamo che $f_{X|Y}(x|y)$ è la densità di X "condizionata" all'evento $\{Y=y\}$. Le virgolette sono necessarie perché, avendo preso Y assolutamente continua, l'evento $\{Y=y\}$ ha probabilità 0 e non può essere usato in un condizionamento.

Esempio 7.27. Due variabili aleatorie X e Y, assolutamente continue, hanno densità congiunta

$$f_{X,Y}(x,y) = 6e^{-2x}e^{-3y}$$

per $x>0$ e $y>0$ e nulla altrimenti. Vogliamo determinare se X e Y sono indipendenti.

Come prima cosa ci ricaviamo, dalla densità congiunta, le densità marginali. Per X,

$$f_X(x) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,\mathrm{d}y = \begin{cases} \int_0^{+\infty} 6e^{-2x}e^{-3y}\,\mathrm{d}y = 2e^{-2x} & x>0 \\ 0 & x\leqslant 0 \end{cases}$$

e per Y

$$f_Y(y) = \int_{-\infty}^{+\infty} f_{X,Y}(x,y)\,\mathrm{d}x = \begin{cases} \int_0^{+\infty} 6e^{-2x}e^{-3y}\,\mathrm{d}x = 3e^{-3y} & y>0 \\ 0 & y\leqslant 0. \end{cases}$$

A questo punto non ci resta che verificare l'indipendenza confrontando il prodotto delle densità marginali con la densità congiunta:

$$f_X(x)\,f_Y(y) = 2e^{-2x}3e^{-3y} = 6e^{-2x}e^{-3y} = f_{X,Y}(x,y).$$

Le due variabili aleatorie sono allora indipendenti tra loro.

Come per le variabili aleatorie discrete, ci interessiamo a un problema particolare: la somma di una coppia di variabili aleatorie assolutamente continue.

PROPOSIZIONE 7.28. *Siano X e Y due variabili aleatorie assolutamente continue sullo stesso spazio probabilità $(\Omega,\mathcal{F},P)$, con densità congiunta $f_{X,Y}$. La densità della loro somma è*

$$f_{X+Y}(z) = \int_{-\infty}^{+\infty} f_{X,Y}(x,z-x)\,\mathrm{d}x.$$

Dimostrazione. Cominciamo considerando la funzione di ripartizione, $F_{X+Y}(z) = P(X+Y\leqslant z)$. Possiamo vedere questa probabilità come $P((X,Y)\in E)$ per qualche $E\in\mathcal{B}\otimes\mathcal{B}$: infatti

$$X+Y\leqslant z \quad\Longleftrightarrow\quad Y\leqslant z-X.$$

Se lo vediamo nel piano cartesiano $\mathbb{R}^2$, sono i punti al di sotto della retta $y=-x+z$, quindi

$$P((X,Y)\in E) = \iint_E f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x = \int_{-\infty}^{+\infty}\int_{-\infty}^{z-x} f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

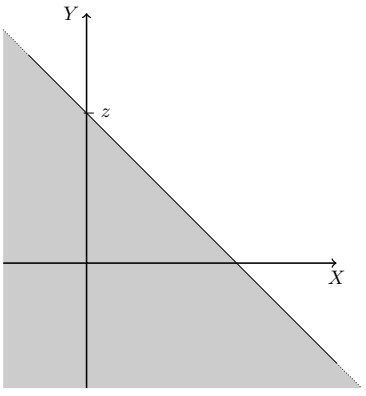


Figura 7.1. L'evento E è quello in grigio in figura

Infine, visto che siamo interessati alla densità, deriviamo in z e abbiamo concluso. □

Esempio 7.29. Siano X e Y variabili aleatorie assolutamente continue tali che $f_{X,Y}(x,y) = e^{-x}$ per $0 \leqslant y \leqslant x$ e nulla altrimenti. Qual è la legge della somma $X+Y$?

Come prima cosa determiniamo in quali punti del piano $\mathbb{R}^2$ è supportata (cioè è diversa da 0) la funzione $f_{X,Y}$: dalla definizione abbiamo che $f_{X,Y} = 0$ se $y \leqslant 0$ o se $y > x$.

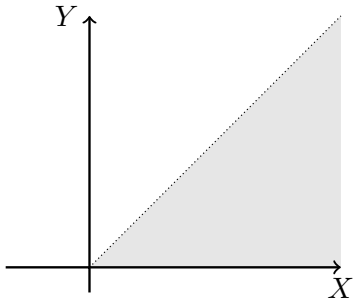


Figura 7.2. In grigio il supporto di $f_{X,Y}$

Sappiamo, dalla Proposizione 7.28, che

$$F_{X+Y}(z) = \iint_E f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x$$

con $E = \{(x,y) : x+y \leqslant z\}$ (rappresentato in Figura 7.1). Ma possiamo mettere assieme questa informazione col supporto di $f_{X,Y}$, perché al di fuori di quest'ultimo l'integrale è identicamente nullo. Quindi possiamo integrare sul dominio E', dato dall'intersezione di E col supporto di $f_{X,Y}$ e rappresentato in Figura 7.3.

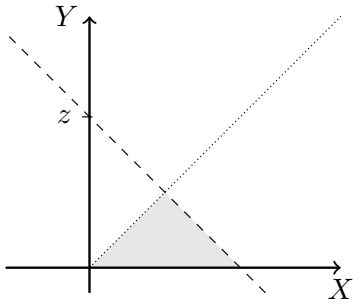


Figura 7.3. In grigio il dominio di integrazione di $f_{X,Y}$

Ora non ci resta che calcolare l'integrale, ma come possiamo farlo? L'integrale ha forma

$$\int_{\square}^{\square}\int_{\square}^{\square} e^{-x}\,\mathrm{d}x\,\mathrm{d}y$$

in cui dobbiamo però determinare gli estremi di integrazione. Possiamo osservare che y varia tra 0 e $z/2$ e che, per y fissato, x varia tra y e $z-y$. Allora

$$\begin{aligned} F_{X+Y}(z) &= \int_0^{\frac{z}{2}}\int_y^{z-y} e^{-x}\,\mathrm{d}x\,\mathrm{d}y = \int_0^{\frac{z}{2}} [-e^{-x}]_y^{z-y}\,\mathrm{d}y \\ &= \int_0^{\frac{z}{2}} e^{-y} - e^{-(z-y)}\,\mathrm{d}y = [-e^{-y}]_0^{z/2} - e^{-z}\,[e^{y}]_0^{z/2} \\ &= 1 - e^{-\frac{z}{2}} - e^{-\frac{z}{2}} + e^{-z} = \left(1 - e^{-\frac{z}{2}}\right)^2 \end{aligned}$$

per ogni $z \geqslant 0$. A questo punto, per avere f_{X+Y} possiamo derivare in z.

Esempio 7.30. Siano X, Y due variabili aleatorie indipendenti e identicamente distribuite con densità $f(t) = e^{-t}$ per $t > 0$ e 0 altrimenti. Sia $S = X + Y$ la loro somma. Qual è la densità di X condizionata a S?

Determiniamo come prima cosa la densità congiunta. Grazie all'ipotesi di indipendenza tra le variabili aleatorie abbiamo

$$f_{X,Y}(x,y) = e^{-x}e^{-y}$$

per $x, y > 0$, nel primo quadrante, e 0 altrove.

Ora vogliamo determinare la legge congiunta di X e S. Per farlo, passiamo dalla funzione di ripartizione $F_{X,S}$:

$$\begin{aligned} F_{X,S}(x,z) &= P(X \leqslant x, S \leqslant z) = P(X \leqslant x, Y \leqslant z - X) \\ &= P((X,Y) \in E) = \iint_E f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \end{aligned}$$

dove il dominio E cambia a seconda che $x < z$ o $x \geqslant z$, come illustrato nella Figura 7.4.

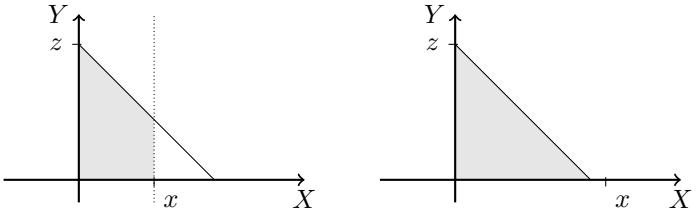


Figura 7.4. Il dominio di integrazione E in grigio, a sinistra se $0 < x < z$, a destra se $x \geqslant z$

Allora, se $x \geqslant z$,

$$F_{X,S}(x,z) = \int_0^z e^{-t} \int_0^{z-t} e^{-u}\,\mathrm{d}u\,\mathrm{d}t = \int_0^z e^{-t}(1 - e^{-z}e^{t})\,\mathrm{d}t = 1 - e^{-z} - ze^{-z}.$$

Se invece $0 < x < z$,

$$F_{X,S}(x,z) = \int_0^x e^{-t} \int_0^{z-t} e^{-u}\,\mathrm{d}u\,\mathrm{d}t = \int_0^x e^{-t}(1 - e^{-z}e^{t})\,\mathrm{d}t = 1 - e^{-x} - xe^{-z}.$$

Ora possiamo ricavare la densità congiunta derivando in x e z:

$$f_{X,S}(x,z) = \frac{\partial^2 F_{X,S}}{\partial x\,\partial z} = \begin{cases} e^{-z} & 0 < x < z \\ 0 & x \geqslant z. \end{cases}$$

Vogliamo determinare $f_{X|S}(x\,|\,z) = \frac{f_{X,S}(x,z)}{f_S(z)}$, ma ci occorre ancora f_S,

$$f_S(z) = \int_{\mathbb{R}} f_{X,S}(x,z)\,\mathrm{d}x = \int_0^z e^{-z}\,\mathrm{d}x = ze^{-z}.$$

Quindi abbiamo, per $0 < x < z$,

$$f_{X|S}(x\,|\,z) = \frac{e^{-z}}{ze^{-z}} = \frac{1}{z}.$$

Lezione 12

7.3. VETTORI ALEATORI MISTI

Nelle sezioni precedenti abbiamo considerato coppie aleatorie omogenee, in cui entrambe le variabili aleatorie sono dello stesso tipo, o discrete o assolutamente continue. Vediamo ora, in un esempio, cosa succede se le due variabili aleatorie in una coppia sono una discreta e una assolutamente continua.

Esempio 7.31. Tra gli studenti dell'Università di Otnert, il 52% studiano materie umanistiche e il 48% studiano materie scientifiche. Il tempo di studio al giorno per gli studenti delle materie scientifiche è distribuito in modo uniforme tra 155 e 180 minuti, mentre per gli studenti delle materie umanistiche è distribuito in modo uniforme tra 143 e 166 minuti. Ci chiediamo:

1. Qual è, se esiste, la legge congiunta delle variabili aleatorie X (indirizzo di studio) e Y (tempo di studio quotidiano).
2. Qual è la probabilità che un generico studente dell'università studi al più 160 minuti.
3. Come sono suddivisi tra i due indirizzi gli studenti che passano sui libri meno di 160 minuti.
4. Come sono suddivisi tra i due indirizzi gli studenti che passano sui libri esattamente 160 minuti.

Cominciamo con lo scrivere formalmente i dati del nostro problema. La variabile aleatoria X è discreta e, in particolare, identicamente distribuita a una moneta sbilanciata: se codifichiamo con 0 l'indirizzo scientifico e con 1 l'indirizzo umanistico abbiamo

$$\varphi_X(x) = \begin{cases} 0.48 & x = 0 \\ 0.52 & x = 1 \\ 0 & x \in \{0,1\}^c. \end{cases}$$

Abbiamo poi per Y le seguenti densità condizionate:

$$f_{Y|X}(y\,|\,0) = \begin{cases} c_S & y \in [155,180] \\ 0 & \text{altrimenti} \end{cases} \qquad f_{Y|X}(y\,|\,1) = \begin{cases} c_U & y \in [143,166] \\ 0 & \text{altrimenti} \end{cases}$$

dove c_S e c_U sono due costanti positive che dobbiamo determinare, in modo che le densità condizionate siano effettivamente delle densità, cioè abbiano integrale 1,

$$c_S\,(180-155) = 1 \quad \Rightarrow \quad c_S = \frac{1}{25} \qquad\qquad c_U\,(166-143) = 1 \quad \Rightarrow \quad c_S = \frac{1}{23}.$$

In realtà dovremmo scrivere la legge condizionata per ogni $(x,y) \in \mathbb{R}^2$,

$$f_{Y|X}(y\,|\,x) = \begin{cases} 1/25 & y \in [155,180], x = 0 \\ 1/23 & y \in [143,166], x = 1 \\ 0 & \text{altrimenti.} \end{cases}$$

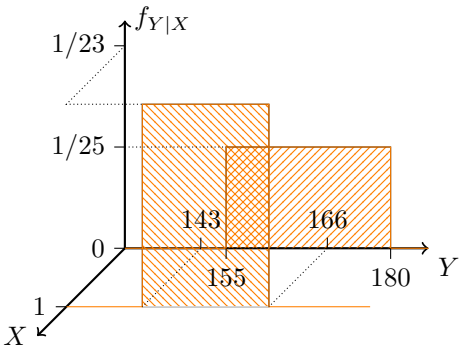


Figura 7.5. Densità condizionale di Y data X. Entrambi i rettangolini hanno area 1

Per avere la legge congiunta, dobbiamo passare attraverso le funzioni di ripartizione, per vedere che succede. Quello che otteniamo è $F_{X,Y}(x,y) = F_{Y|X}(y\,|\,x)\,F_X(x)$, che ci suggerisce la forma seguente per la "densità",

$$f_{X,Y}(x,y) = f_{Y|X}(y\,|\,x)\,\varphi_X(x),$$

un ibrido tra una densità congiunta e una densità discreta congiunta,

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{25}\cdot 0.48 = 0.0192 & y \in [155,180], x = 0 \\ \frac{1}{23}\cdot 0.52 = 0.0226087 & y \in [143,166], x = 1 \\ 0 & \text{altrimenti.} \end{cases}$$

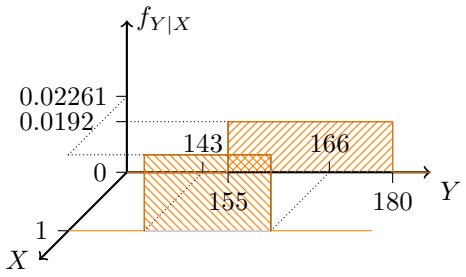


Figura 7.6. Densità congiunta di X e Y. La somma delle aree dei rettangolini è 1

Per trovare la legge di Y dobbiamo marginalizzare la legge congiunta, sommando su tutti i valori possibili di X, che sono solo due:

$$f_Y(y) = f_{X,Y}(0,y) + f_{X,Y}(1,y) = \begin{cases} 0.0226087 & 143 < y < 155 \\ 0.0418087 & 155 < y < 166 \\ 0.0192 & 166 < y < 180 \\ 0 & \text{altrimenti.} \end{cases}$$

Siccome vogliamo sapere la probabilità che uno studente passi al più 160 minuti sui libri, dobbiamo ricavare $F_Y(y)$, integrando f_Y,

$$F_Y(y) = \begin{cases} 0 & y < 143 \\ 0.0226087\,(y-143) & 143 \leqslant y < 155 \\ 0.2713044 + 0.0418087\,(y-155) & 155 \leqslant y < 166 \\ 0.7312001 + 0.0192\,(y-166) & 166 \leqslant y < 180 \\ 1 & y \geqslant 180, \end{cases}$$

quindi la probabilità cercata è $F_Y(160) = 0.4803479$.

Chiedere come sono distribuiti tra i due indirizzi gli studenti che passano meno di 160 minuti sui libri equivale a calcolare, per $x = 0, 1$, le probabilità

$$P(X = x \mid Y < 160) = \frac{P(X = x, Y < 160)}{P(Y < 160)} = \frac{\int_{143}^{160} f_{X,Y}(x,y)\,\mathrm{d}y}{F_Y(160)} \approx \begin{cases} 0.2 & x = 0 \\ 0.8 & x = 1. \end{cases}$$

Se invece siamo interessati alla probabilità di appartenenza ai due indirizzi di uno studente che studia esattamente 160 minuti, non possiamo fare allo stesso modo, perché l'evento $Y = 160$ ha probabilità nulla[7.1]. Tuttavia possiamo usare la densità condizionata ibrida

$$f_{X|Y}(x \mid 160) = \frac{f_{X,Y}(x,160)}{f_Y(160)} = \frac{f_{X,Y}(x,160)}{0.0418087} \approx \begin{cases} 0.46 & x = 0 \\ 0.54 & x = 1. \end{cases}$$

Quello che otteniamo è una densità discreta, ossia la probabilità che uno studente appartenga a uno dei due indirizzi. Questo non ci dovrebbe sorprendere, perché nel momento in cui ci restringiamo a un valore specifico di Y, geometricamente stiamo sezionando la legge congiunta, restringendola al piano $y = 160$. Su questo piano abbiamo una funzione (in x) costantemente uguale a 0, tranne in 0 e 1. A questo punto (modulo un riscalamento per $f_Y(160)$) abbiamo una funzione che soddisfa tutte le proprietà di una densità discreta.

7.1. Se un evento è impossibile, allora ha probabilità nulla, ma non è vero il viceversa: se abbiamo una variabile aleatoria continua, ciascun suo valore a priori ha probabilità 0 di uscire, eppure uno di essi esce, quindi, a posteriori, non possiamo dire che fosse impossibile.

CAPITOLO 8

MODELLI DI VARIABILI ALEATORIE DISCRETE

Come abbiamo visto in molti degli esempi, ci sono alcune variabili aleatorie che ricorrono abbastanza spesso. In parte questo è dovuto al fatto che finora non abbiamo introdotto moltissimi esempi di variabili aleatori, ma in parte è anche legato all'esistenza di un certo numero di variabili aleatorie ben conosciute che vengono usate per descrivere esperimenti aleatori con certe caratteristiche.

Parleremo indifferentemente di variabili aleatorie o di distribuzioni (ad esempio, una variabile aleatoria Bernoulliana o a distribuzione Bernoulliana o distribuita come una Bernoulliana) perché siamo interessati alla legge, alla distribuzione, appunto, perché è questa a caratterizzare il comportamento probabilistico della variabile aleatoria, indipendentemente dallo spazio di partenza.

8.1. BERNOULLIANE

Cominciamo dall'esperimento più semplice (ma non banale) che possiamo immaginare: qualcosa il cui esito è binario, può essere "sì" o "no", positivo o negativo e così via, con una certa probabilità. Dovrebbe ricordarci qualcosa…

DEFINIZIONE 8.1. *Una variabile aleatoria discreta X si dice* Bernoulliana *(o* variabile aleatoria di Bernoulli[8.1]*)* di parametro p*, con* $p \in [0,1]$ *se ha densità discreta*

$$\varphi_X(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \\ 0 & \textit{altrimenti,} \end{cases}$$

o equivalentemente se ha funzione di ripartizione

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1-p & 0 \leqslant x < 1 \\ 1 & x \geqslant 1. \end{cases}$$

Se X è una Bernoulliana, scriviamo $X \sim \mathrm{bin}(1,p)$.

Se guardiamo la funzione di ripartizione (o la densità discreta) possiamo riconoscere in questa una variabile aleatoria che abbiamo già incontrato in precedenza: è la variabile aleatoria indicatrice, vista negli Esempi 5.17 e 5.27, in questo caso indicatrice dell'evento "successo", che ha probabilità p. Nello scrivere l'esperimento aleatorio come variabile casuale abbiamo codificato il successo con 1 e l'insuccesso con 0. In realtà l'abbiamo incontrata altre volte, spesso con $p = 0.5$: è il lancio di una moneta. Se $p = 0.5$, la moneta è equa, altrimenti è non bilanciata.

A questo punto sappiamo facilmente proporre un candidato per lo spazio di probabilità su cui è definita: $\Omega = \{0,1\}$, $\mathcal{F} = \mathcal{P}(\Omega) = \{\emptyset, \{0\}, \{1\}, \Omega\}$ e P definita sui singoletti come $P(\{0\}) = 1-p$ e $P(\{1\}) = p$.

8.1. Jakob Bernoulli (1655 – 1705) uno dei molti matematici della famiglia. È legato alla probabilità dalla sua opera *Ars Conjectandi*, pubblicata postuma nel 1713.

8.2. BINOMIALI

Consideriamo ora il caso in cui abbiamo n variabili aleatorie Bernoulliane, indipendenti e identicamente distribuite (i.i.d.); ne prendiamo la somma S. Se riguardiamo alla caratterizzazione che abbiamo dato poco sopra delle Bernoulliane, S è la variabile aleatoria che conta il numero di successi in n lanci di una moneta (cioè in n ripetizioni di un esperimento Bernoulliano).

DEFINIZIONE 8.2. *Diciamo che una variabile aleatoria discreta X è una* binomiale di parametri n e p, *con $n \in \mathbb{N} \setminus \{0\}$ e $p \in [0,1]$, se è la somma di n variabili aleatorie di Bernoulli indipendenti e identicamente distribuite di parametro p. In questo caso scriviamo $X \sim \text{bin}(n,p)$.*

Osservazione 8.3. Come suggerito dalla notazione, una Bernoulliana è una binomiale di parametri $n = 1$ e p, cioè la somma di una sola Bernoulliana di parametro p.

Mentre nella definizione delle variabili aleatorie Bernoulliane avevamo identificato queste ultime mediante la loro legge, ossia la loro densità discreta o funzione di ripartizione, nel caso delle binomiali le abbiamo caratterizzate a partire da altre variabili aleatorie. Come prima cosa, quindi, andiamo a ricavare la legge di una binomiale di parametri n e p.

PROPOSIZIONE 8.4. *Se $X \sim \text{bin}(n,p)$ con $n \in \mathbb{N} \setminus \{0\}$ e $p \in [0,1]$, allora*

$$\varphi_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k \in \{0,\dots,n\} \\ 0 & \textit{altrimenti} \end{cases} \qquad F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}.$$

Dimostrazione. Se $X \sim \text{bin}(n,p)$, allora il suo supporto, ossia l'insieme dei valori che può assumere, è l'insieme $\mathcal{R}_X = \{0,1,\dots,n\}$, dal momento che tra le n Bernoulliane considerate possiamo non avere alcun successo, averne uno solo e così via fino a n successi.

Prendiamo ora $k \in \{0,\dots,n\}$, vogliamo calcolare $\varphi_X(k)$. Ricordiamo che, dalle definizioni,

$$\varphi_X(k) = P(X=k) = P\left(\sum_{i=1}^{n} X_i = k\right)$$

con $X_i \sim \text{bin}(1,p)$ per $i \in \{1,\dots,n\}$. Chiamiamo $N = \{1,\dots,n\}$ l'insieme degli indici delle Bernoulliane e indichiamo con $\mathcal{I}_k \subseteq \mathcal{P}(N)$ la famiglia degli insiemi I_k tali che $I_k \subseteq N$ e $\#I_k = k$. Per ciascun insieme (di indici) I_k definiamo l'evento

$$E_{I_k} = \bigcap_{i \in I_k} \{X_i = 1\} \cap \bigcap_{i \in N \setminus I_k} \{X_i = 0\}$$

che contiene gli esiti per cui tutti e soli i successi sono nelle Bernoulliane i cui indici sono in I_k.

Possiamo ora scrivere l'evento $\{X = k\}$ di tutti gli esiti che danno luogo a esattamente k successi come

$$\{X = k\} = \bigcup_{I_k \in \mathcal{I}_k} E_{I_k}.$$

Per come sono costruiti questi sono tutti e soli i modi di avere esattamente k successi negli n tentativi. Non solo, per come abbiamo definito gli E_k, essi sono eventi tra loro disgiunti, quindi

$$\varphi_X(k) = P(X=k) = P\left(\bigcup_{I_k \in \mathcal{I}_k} E_{I_k}\right) = \sum_{I_k \in \mathcal{I}_k} P(E_{I_k}).$$

Il passo successivo è quindi calcolare $P(E_{I_k})$ per ogni $I_k \in \mathcal{I}_k$:

$$\begin{aligned} P(E_{I_k}) &= P\left(\bigcap_{i \in I_k} \{X_i = 1\} \cap \bigcap_{i \in N \setminus I_k} \{X_i = 0\}\right) \\ &= \prod_{i \in I_k} P(X_i = 1) \prod_{i \in N \setminus I_k} P(X_i = 0) \\ &= p^{\#I_k} (1-p)^{\#(N \setminus I_k)} = p^k (1-p)^{n-k}, \end{aligned}$$

in cui abbiamo sfruttato l'indipendenza delle Bernoulliane nella seconda identità e il fatto che siano identicamente distribuite nella terza.

Osserviamo che la probabilità $P(E_{I_k})$ così trovata è uguale per tutti gli insiemi $I_k \in \mathcal{I}_k$, dal momento che dipende solamente dalla cardinalità di I_k. Allora

$$\varphi_X(k) = \sum_{I_k \in \mathcal{I}_k} P(E_{I_k}) = p^k (1-p)^{n-k} \sum_{I_k \in \mathcal{I}_k} 1$$

e, per concludere, non ci resta che contare quanti elementi ha la famiglia di insiemi $\mathcal{I}_k$. Ci siamo ricondotti al problema, già visto, di contare in quanti modi diversi possiamo scegliere k indici tra n, cioè $\binom{n}{k}$. In conclusione, per $k \in \{0, \dots, n\}$, $\varphi_X(k) = \binom{n}{k} p^k (1-p)^{n-k}$.

Per ricavare la funzione di ripartizione $F_X(x)$ dobbiamo sommare, per tutti gli interi non negativi minori di x, la probabilità di assumere tali valori, ossia la densità discreta:

$$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \varphi_X(k) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k},$$

concludendo così la dimostrazione. □

Osservazione 8.5. Nella dimostrazione abbiamo osservato che per una variabile aleatoria binomiale X, $\varphi_X(k) = P(\sum_{i=1}^n X_i = k)$, con le $X_i \sim \text{bin}(1, p)$. Queste ultime sono tutte variabili aleatorie discrete, quindi potremmo farne ricorsivamente la somma, come visto nel Capitolo 7. Tuttavia questa strategia si presta più facilmente a errori e, nella sostanza, è analoga a quanto visto nella dimostrazione.

Osservazione 8.6. Possiamo controllare immediatamente che, per una variabile aleatoria binomiale $X \sim \text{bin}(n, p)$, $F_X(x) = 1$ per ogni $x \geqslant n$: basta leggere la somma come binomio di Newton di esponente n,

$$F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \varphi_X(k) = \sum_{k=0}^{n} \varphi_X(k) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1.$$

Esempio 8.7. (Ross 5.1.1) Un'azienda produce pennette USB che sono difettose, indipendentemente l'una dall'altra, con probabilità $p = 0.02$. Vende questi oggetti in confezioni da 15 e rimborsa i propri clienti se c'è più di una pennetta difettosa nella confezione. Quale percentuale di confezioni viene rimborsata? Comprando 4 confezioni, con che probabilità esattamente una di queste sarà rimborsabile?

Cominciamo a descrivere la situazione in questo problema in termini di variabili aleatorie. Chiamiamo O la variabile aleatoria che descrive, per una singola pennetta, il suo essere o meno difettosa e con N il numero di pennette difettose in una confezione. La variabile aleatoria O è una Bernoulliana di parametro $p = 0.02$, cioè $O \sim \text{bin}(1, 0.02)$. La variabile aleatoria N è la somma di 15 variabili aleatorie indipendenti e distribuite come O, quindi è una binomiale di parametri $n = 15$ e $p = 0.02$, $N \sim \text{bin}(15, 0.02)$.

Possiamo a questo punto riformulare la prima domanda come la probabilità che $N > 1$,

$$\begin{aligned} P(N>1) &= \sum_{k=2}^{15} P(N=k) = 1 - P(N=0) - P(N=1) \\ &= 1 - \varphi_N(0) - \varphi_N(1) \\ &= 1 - \binom{15}{0} (1-0.02)^{15} - \binom{15}{1} 0.02\,(1-0.02)^{14} \approx 3.5\% \end{aligned}$$

Osserviamo che avremmo potuto equivalentemente scrivere $P(N > 1) = 1 - F_N(1)$.

Per rispondere alla seconda domanda introduciamo una nuova variabile aleatoria S che dice se una scatola è rimborsabile o meno: $S \sim \text{bin}(1, 0.035)$. A noi però interessa il totale di scatole da rimborsare tra le 4 comprate: questa è una variabile aleatoria $R \sim \text{bin}(4, 0.035)$. La risposta alla seconda domanda, ossia la probabilità di farsi rimborsare esattamente una scatola tra 4 acquistate, è

$$\varphi_R(1) = \binom{4}{1} 0.035\,(1-0.035)^3 \approx 12.7\%.$$

8.2.1. Bernoulliane e binomiali in R

Uno dei motivi per cui usiamo R a supporto degli esercizi è il suo essere orientato alla probabilità e alla statistica. In particolare contiene già le leggi delle principali variabili aleatorie. Cominciamo a vedere cosa offre per la binomiale (e per la Bernoulliana).

Densità discreta La funzione di densità discreta per una binomiale è la funzione `dbinom(x, size, prob)` che ha come parametri il punto `x` in cui vogliamo calcolare la densità φ, il numero `size` di tentativi (quello che nella Definizione 8.2 abbiamo indicato con n) e la probabilità `prob` di successo di ogni tentativo (quella che nella Definizione 8.2 abbiamo indicato con p).

Per calcolare la densità discreta $\varphi_X(11)$ di una variabile aleatoria $X \sim \text{bin}(44, 0.2)$ in R useremo il comando `dbinom(x = 11, size = 44, prob = 0.2)`. Se nominiamo i parametri, possiamo anche passarli in ordine diverso (ad esempio `dbinom(size = 44, prob = 0.2, x = 11)`), in alternativa possiamo passarli anche senza nominarli, ma in questo caso devono essere nell'ordine predefinito: `dbinom(11,44,0.2)`.

Funzione di ripartizione La funzione di ripartizione per una binomiale è la funzione `pbinom(q, size, prob, lower.tail = TRUE)`, i cui parametri `size` e `prob` sono esattamente come sopra, mentre `q` è il punto[8.2] in cui vogliamo calcolare la funzione di ripartizione F e `lower.tail` è un parametro logico (posto vero di default) che determina se calcoliamo la funzione di ripartizione F_X nel punto q (ossia $P(X \leqslant q)$, la coda inferiore), in corrispondenza del valore `TRUE` o il suo complementare $1 - F_X(q)$ (cioè $P(X > q)$, la coda superiore), in corrispondenza del valore `FALSE`.

Per calcolare la funzione di ripartizione $F_X(12.5)$ di una variabile aleatoria $X \sim \text{bin}(23, 0.5)$ in R useremo quindi il comando `pbinom(q = 12.5, size = 23, prob = 0.5)`. Non abbiamo bisogno di specificare il valore `lower.tail = TRUE`, perché stiamo prendendo il valore di default. Se invece fossimo interessati alla probabilità $P(X > 10)$, potremmo scrivere `pbinom(10, 23, 0.5, lower.tail = FALSE)`.

Altre funzioni Ci sono altre due funzioni nella famiglia binomiale di R: `rbinom` e `qbinom`. La prima è un generatore casuale di risultati distribuiti come una binomiale dei parametri assegnati. In sostanza ci genera dei *valori* $X(\omega) \in \mathbb{R}$, con $X \sim \text{bin}(n, p)$. La sintassi di questa funzione è la seguente: `rbinom(n, size, prob)`, in cui `size` e `prob` sono gli stessi parametri già incontrati sopra, mentre `n` indica il numero di realizzazioni da generare.

Se vogliamo un campione di 100 realizzazioni di una binomiale di parametri $n = 1$ e $p = 0.5$ (cioè 100 lanci di una moneta bilanciata), possiamo scrivere `rbinom(n=100,size=1,prob=0.5)`.

La funzione `qbinom` è la funzione quantile, che incontreremo più avanti.

Esempio 8.8. Avremmo potuto rispondere alle domande nell'Esempio 8.7 con il seguente codice:

```
p <- pbinom(q = 1, size = 15, prob = 0.02, lower.tail = FALSE)
p #per visualizzare la prima probabilità
dbinom(x = 1, size = 4, prob = p)
```

Lezione 13

8.3. LO SCHEMA DI BERNOULLI

Nella sezione precedente abbiamo introdotto le binomiali come somma finita di variabili aleatorie Bernoulliane. Possiamo vedere l'esperimento sottostante come una ripetizione finita di esperimenti Bernoulliani. Il passo successivo, però, è considerare una ripetizione infinita, almeno in potenza (nel senso che ci aspettiamo che a un certo punto finisca, ma non sappiamo dare a priori un limite superiore al numero di ripetizioni).

8.2. La lettera `q` viene dal termine *quantile* che essa rappresenta, che definiremo in seguito.

Se ci pensiamo bene, abbiamo già visto un esperimento di questo tipo nell'Esempio 5.15. Possiamo generalizzarlo, considerando una successione infinita di prove, tra loro indipendenti, che abbiano successo con probabilità comune p (e insuccesso con probabilità $1-p$). Avendo introdotto le variabili aleatorie di Bernoulli, possiamo dire che si tratta di una successione infinita di variabili aleatorie Bernoulliane indipendenti e identicamente distribuite, di parametro p. L'intera successione di prove prende anche il nome di *processo di Bernoulli*.

Vogliamo rappresentare il processo di Bernoulli in linguaggio matematico, come spazio di probabilità $(\Omega, \mathcal{F}, P)$, come avevamo già accennato nell'Esempio 5.15. Lo spazio degli esiti Ω è l'insieme delle successioni a valori in $\{0,1\}$, quindi $\Omega = \{0,1\}^{\mathbb{N}\setminus\{0\}}$.

Stiamo considerando uno spazio prodotto infinito, quindi vogliamo prendere per $\mathcal{F}$ la tribù generata dai cilindri, ossia i sottoinsiemi di Ω ottenuti fissando un numero finito degli indici iniziali: un insieme $C \subseteq \Omega$ è un cilindro se esistono un numero naturale n e un vettore $v \in \{0,1\}^n$ tali che le prime n componenti di ogni elemento $\omega \in C$ coincidono col vettore v, ossia

$$C = \{\omega \in \Omega : \omega_i = v_i, 1 \leqslant i \leqslant n\}.$$

La probabilità sullo spazio prodotto è il prodotto delle probabilità sulle varie componenti, uguale a p o $(1-p)$.

Esempio 8.9. Vediamo alcuni esempi di probabilità di eventi (cilindrici) in un processo di Bernoulli di parametro p.

- La probabilità di avere un successo seguito da due insuccessi è $P(100*) = p\,(1-p)^2$, in cui abbiamo rappresentato con $*$ una qualunque successione di 0 e 1. Possiamo calcolare la probabilità di $100*$ perché è un cilindro, le cui prime tre componenti sono $(1,0,0)$.
- La probabilità che il primo successo sia alla k-sima prova è $P(0\ldots01*) = (1-p)^k p$. In questo caso il cilindro è determinato dal vettore di lunghezza k le cui prime $k-1$ componenti sono 0 e la k-sima è un 1.
- La probabilità che il terzo lancio sia un successo. L'evento che ci interessa è $\{\cdot\cdot 1*\}$, cioè due componenti qualunque (a scelta tra 0 e 1), seguite da un 1, seguito a sua volta da qualunque cosa. Scritto così, $\{\cdot\cdot 1*\}$ non è un cilindro, ma possiamo generarlo con cilindri, ossia scriverlo come unione numerabile di cilindri e loro complementari:

  $$\{\cdot\cdot 1*\} = \{001*\} \cup \{011*\} \cup \{101*\} \cup \{111*\}.$$

  Questa unione è disgiunta, quindi possiamo calcolare la probabilità cercata sommando le probabilità dei quattro cilindri:

  $$\begin{aligned} P(\cdot\cdot 1*) &= P(001*) + P(011*) + P(101*) + P(111*) \\ &= (1-p)^2 p + (1-p)\,p^2 + p\,(1-p)\,p + p^3 \\ &= p((1-p)^2 + 2p\,(1-p) + p^2) \\ &= p\,(p + (1-p))^2 = p. \end{aligned}$$

- La probabilità che il primo successo sia in un lancio dispari.

8.4. GEOMETRICHE

Consideriamo uno schema di Bernoulli di parametro p: siamo interessati al numero di insuccessi prima di ottenere un successo. Chiamiamo T_1 l'istante di primo successo in uno schema di Bernoulli. Gli esiti sono della forma $\omega = (\omega_1, \omega_2, \ldots)$, quindi possiamo scrivere

$$T_1 = \inf\{i \geqslant 1 : \omega_i = 1\}.$$

Allora T_1 è una variabile aleatoria[8.3]: è in particolare la variabile aleatoria indicatrice del primo successo. Tuttavia non descrive proprio quello che cercavamo: noi siamo interessati agli insuccessi che precedono il primo successo, ossia $T_1 - 1$. Qual è la loro distribuzione?

DEFINIZIONE 8.10. *Diciamo che una variabile aleatoria X è* geometrica di parametro *p se è l'istante precedente al primo successo di uno schema di Bernoulli di parametro p. In questo caso scriviamo $X \sim \mathrm{geom}(p)$.*

Dalla definizione possiamo ricavare immediatamente la densità discreta di $X \sim \mathrm{geom}(p)$:

$$\varphi_X(k) = (1-p)^k p$$

per $k \in \mathbb{N}$ (e 0 altrimenti). Infatti, dire che $X = k$ significa che i primi k tentativi sono insuccessi e il $(k+1)$-simo è un successo, da cui, rispettivamente, i due fattori $(1-p)^k$ e p.

Osservazione 8.11. Nella densità discreta non compare un coefficiente binomiale perché non stiamo chiedendo "un successo nei primi $k+1$ lanci", ma stiamo imponendo un ordine: i primi k sono insuccessi, il $(k+1)$-simo è un successo.

Dobbiamo verificare che la definizione data di densità discreta sia ben posta, in particolare che la somma su tutti i possibili valori sia uguale a 1. Se $p = 1$, allora $X \equiv 0$, quindi $\varphi_X(k) = \mathbb{1}_{\{k=0\}}$. Se $p < 1$, allora

$$\sum_{k=0}^{+\infty} \varphi_X(k) = \sum_{k=0}^{+\infty} (1-p)^k p = p \sum_{k=0}^{+\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1,$$

in cui abbiamo usato il fatto che la serie $\sum_{k=0}^{+\infty} (1-p)^k$ è geometrica[8.4] di ragione $1-p$ compresa tra 0 e 1.

Osservazione 8.12. La definizione di variabile geometrica non è unica: molti definiscono come variabile aleatoria geometrica il primo istante di successo (quella che abbiamo chiamato T_1). In questo caso la densità discreta è leggermente diversa:

$$\varphi_{T_1}(k) = (1-p)^{k-1} p$$

per $k \in \mathbb{N} \setminus \{0\}$. La sostanza non cambia molto, ma i numeri sì, quindi è opportuno prestare attenzione. Di solito si distingue tra le due specificando il dominio: nella Definizione 8.10 il dominio era $\mathbb{N}$, per il primo successo il dominio è $\mathbb{N} \setminus \{0\}$.

Scegliere l'una o l'altra dipende dal gusto estetico o dalla comodità. Nel caso di questo corso ci siamo allineati, per semplicità, alla scelta fatta in R.

Abbiamo ricavato la densità discreta, adesso vediamo com'è fatta la funzione di ripartizione di una variabile aleatoria X geometrica di parametro p:

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \sum_{k=0}^{\lfloor x \rfloor} \varphi_X(k) = p \sum_{k=0}^{\lfloor x \rfloor} (1-p)^k = 1 - (1-p)^{\lfloor x \rfloor + 1} & x \geqslant 0. \end{cases}$$

Osserviamo anche che possiamo ottenere lo stesso risultato in maniera più diretta passando dal complementare: se $n \in \mathbb{N}$, $P(X > n) = 1 - F_X(n)$ è la probabilità che nei primi $n+1$ lanci abbiamo avuto solamente insuccessi, quindi

$$P(X > n) = 1 - F_X(n) = (1-p)^{n+1}. \tag{8.1}$$

8.3. Appartiene a una famiglia di variabili aleatorie i cui elementi sono detti *tempi aleatori* o *tempi casuali* e, in particolare, si tratta di un *tempo d'arresto*, ossia il primo istante in cui una condizione viene soddisfatta.

8.4. Qualche informazione in più sulla serie geometrica è disponibile in Appendice A.2

PROPOSIZIONE 8.13. *Una variabile aleatoria geometrica X gode della proprietà di assenza di memoria, cioè per ogni $n,k\in\mathbb{N}$*

$$P(X\geqslant n+k\,|\,X\geqslant n)=P(X\geqslant k). \tag{8.2}$$

Dimostrazione. Partiamo dalla definizione di probabilità condizionata,

$$\begin{aligned} P(X\geqslant n+k\,|\,X\geqslant n) &= \frac{P((X\geqslant n+k)\cap(X\geqslant n))}{P(X\geqslant n)} \\ &= \frac{P(X\geqslant n+k)}{P(X>n-1)} \\ &= \frac{(1-p)^{n+k}}{(1-p)^n} \\ &= (1-p)^k \\ &= P(X\geqslant k), \end{aligned}$$

in cui abbiamo usato ripetutamente la (8.1). □

Osservazione 8.14. Da dove viene il nome *assenza di memoria* di questa proprietà? Possiamo leggere la (8.2) in questo modo: se dopo n lanci non abbiamo ancora visto un successo ($X\geqslant n$), la probabilità di avere ancora almeno k insuccessi ($X\geqslant n+k$) è uguale alla probabilità che, iniziando ora uno schema di Bernoulli di uguale parametro, avremo almeno k insuccessi prima del primo successo ($X\geqslant k$). In altre parole, il processo non ha memoria di quanti insuccessi ha già avuto: sapere che ci sono stati un certo numero di insuccessi non ci dà alcuna informazione aggiuntiva sull'istante di primo successo (e quindi sull'ultimo istante prima del primo successo).

Esempio 8.15. Se nel Superenalotto il 55 non esce da 60 estrazioni[8.5], quanto è probabile che esca alla prossima estrazione? E che esca per la prima volta tra almeno altre 30?

A ogni estrazione vengono scelti 6 numeri tra 90, la probabilità che esca un particolare numero (ad esempio il 55) è

$$1-\frac{\binom{89}{6}}{\binom{90}{6}}=\frac{\binom{89}{5}}{\binom{90}{6}}=\frac{89!}{84!\,5!}\cdot\frac{84!\,6!}{90!}=\frac{6}{90}=\frac{1}{15}\approx 6.6\%.$$

A ogni estrazione la variabile aleatoria indicatrice dell'evento "esce il 55 al Superenalotto" è una Bernoulliana di parametro $\frac{1}{15}$. Consideriamo lo schema di Bernoulli corrispondente e, in particolare, la probabilità che 55 esca alla prossima estrazione sapendo che non è uscito nelle prime 60. Chiamiamo X la variabile aleatoria che descrive l'ultima estrazione in cui non esce 55: la nostra richiesta è allora

$$P(X=60\,|\,X\geqslant 60)=\frac{P(X=60)}{P(X>59)}=\frac{\left(1-\frac{1}{15}\right)^{60}\frac{1}{15}}{\left(1-\frac{1}{15}\right)^{59+1}}=\frac{1}{15},$$

cioè è identica alla probabilità che 55 esca al primo tentativo.

Analogamente, la probabilità che esca per la prima volta tra almeno altre 30 estrazioni è

$$P(X\geqslant 60+30\,|\,X\geqslant 60)=P(X\geqslant 30)=\left(1-\frac{1}{15}\right)^{30}\approx 12.6\%.$$

La morale di questo esempio è che in termini di probabilità è assolutamente irrilevante che il 55 sia "in ritardo" da 60 estrazioni: la probabilità che esca alla prossima estrazione è esattamente la stessa che esca alla prima estrazione. Non solo, la probabilità che si debba attendere un po' è spesso sottostimata: la probabilità che un numero esca alla prossima estrazione è approssimativamente uguale alla probabilità che non esca prima di 38 estrazioni (6.67% contro 6.78%).

8.5. Dati al 13 aprile 2021, non che sia rilevante, come vedremo.

8.4.1. Geometriche in R

Come abbiamo detto, la scelta della definizione di variabile aleatoria con distribuzione geometrica fatta è stata dettata dalla scelta degli sviluppatori di R (e prima di S).

Le funzioni per una variabile aleatoria geometrica sono `dgeom(x, prob)` per la densità discreta, dove `x` è il punto in cui vogliamo calcolare la densità φ e `prob` è la probabilità di successo di ogni tentativo (che abbiamo indicato con p).

La funzione di ripartizione per una geometrica è la funzione `pgeom(q, prob, lower.tail = TRUE)`, il cui il parametro `prob` è esattamente come sopra, mentre `q` e `lower.tail` sono come nelle corrispondenti funzioni della binomiale: il primo è il punto in cui calcoliamo la funzione di ripartizione F, mentre il secondo è un parametro logico che determina se calcoliamo $F_X(q)$ (il default) o il suo complementare $1-F_X(q)$.

In modo del tutto analogo alla binomiale abbiamo anche per la geometrica altre due funzioni in R: `rgeom` e `qgeom`, rispettivamente la generatrice di valori casuali distribuiti come una geometrica di parametri assegnati e la funzione quantile.

Esempio 8.16. Avremmo potuto rispondere alle domande nell'Esempio 8.15 con il seguente codice:

```
dgeom(x = 60, prob = 1/15)/pgeom(q = 59, prob = 1/15, lower.tail
= FALSE)
pgeom(89, 1/15, FALSE)/pgeom(59,1/15,FALSE)
```

8.5. BINOMIALI NEGATIVE

Se per le variabili aleatorie geometriche siamo partiti dall'istante di primo successo di uno schema di Bernoulli, ora proviamo a generalizzare, considerando i tempi d'attesa del n-simo successo in uno schema di Bernoulli: per $n \in \mathbb{N} \setminus \{0\}$

$$T_n = \inf\left\{ i \geqslant 1 : \sum_{k=1}^{i} \omega_k = n \right\},$$

variabili aleatorie che possiamo anche definire ricorsivamente,

$$\begin{cases} T_1 = \inf\{i \geqslant 1 : \omega_i = 1\} \\ T_{n+1} = \inf\{i > T_n : \omega_i = 1\} \quad n \geqslant 1. \end{cases}$$

Se però con le variabili aleatorie geometriche eravamo interessati al numero di insuccessi prima del primo successo, ora considereremo il numero di insuccessi prima dell'n-simo successo (ossia la variabile aleatoria $T_n - n$. Qual è la sua distribuzione?

DEFINIZIONE 8.17. *Diciamo che una variabile aleatoria X è* binomiale negativa (o di Pascal) di parametri n e p *se è il numero di insuccessi precedenti all'n-simo successo di uno schema di Bernoulli di parametro p. In questo caso scriviamo $X \sim \mathrm{NB}(n,p)$.*

Anche in questo caso iniziamo a ricavare, dalla definizione, la funzione di densità discreta di $X \sim \mathrm{NB}(n,p)$, per $k \geqslant 0$

$$\begin{aligned} \varphi_X(k) &= P(X=k) = P(T_n = k+n) \\ &= P\left(\omega_{k+n} = 1, \sum_{i=1}^{k+n-1} \omega_i = n-1 \right) \\ &= p \binom{k+n-1}{n-1} p^{n-1} (1-p)^k \\ &= \binom{k+n-1}{n-1} p^n (1-p)^k \end{aligned}$$

in cui abbiamo iniziato osservando che se abbiamo k insuccessi prima di avere l'n-simo successo, questo sarà al tentativo $k+n$, allora nei precedenti $k+n-1$ tentativi ci sono $n-1$ successi e, nella penultima riga, abbiamo resa esplicita l'associazione con le binomiali, visto che è il prodotto della probabilità di successo alla $(k+n)$-sima prova (la densità discreta di una Bernoulliana di parametro p calcolata in 1) e della probabilità di avere $n-1$ successi nelle $k+n-1$ prove precedenti (la densità discreta di una binomiale di parametri $k+n-1$ e p calcolata in $n-1$).

Osservazione 8.18. Anche per le binomiali negative vale la pena dare una parola di avvertimento. Come per le geometriche, anche qui alcuni scelgono di definire le binomiali negative come i tempi d'attesa, ossia il numero di tentativi totali prima del successo n-simo. C'è però anche un'ulteriore difficoltà, perché è possibile estendere la definizione al caso in cui $n \in \mathbb{R}^+$, perdendo però l'interpretazione come istante (precedente) al tempo d'arresto. Se $n \in \mathbb{R}^+ \setminus \mathbb{N}$ la binomiale negativa prende anche il nome di *distribuzione di Pólya*.

Esempio 8.19. In un gioco, un personaggio è rimasto intrappolato sul fondo di una buca profonda. Per uscire ha bisogno di ottenere 3 risultati maggiori di 15 lanciando un dado a 20 facce. Ogni lancio di dado corrisponde a 5 minuti di tentativi, nel gioco: con che probabilità il personaggio impiegherà al più mezz'ora per uscire dalla buca?

I lanci ripetuti del dado nascondono uno schema di Bernoulli, in cui la probabilità di successo è $p = \frac{5}{20} = \frac{1}{4}$. Per rispondere alla domanda, osserviamo come prima cosa che il tempo necessario è $5\,T_3$, perché chiediamo che ci siano almeno tre tentativi con successo e ciascun tentativo (di successo o meno) richiede 5 minuti. Chiedere che il personaggio impieghi al più mezz'ora per uscire dalla buca equivale allora a chiedere che faccia al più 6 tentativi totali per di ottenere 3 successi ossia che abbia al più 3 insuccessi. Avendo riformulato il problema in questo modo, possiamo descriverlo usando una binomiale negativa X, di parametri $n = 3$ e $p = \frac{1}{4}$. La probabilità cercata è allora

$$\sum_{k=0}^{3} \varphi_X(k) = \sum_{k=0}^{3} \binom{k+2}{2} \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^k \approx 17\%.$$

8.5.1. Binomiali negative in R

Le funzioni per una variabile aleatoria binomiale negativa sono `dnbinom(x, size, prob)` per la densità discreta, dove `x` è il punto in cui vogliamo calcolare la densità φ, `size` è il numero di successi da raggiungere (quello che abbiamo chiamato n) e `prob` è la probabilità di successo di ogni tentativo (che abbiamo indicato con p).

La funzione di ripartizione per una binomiale negativa è la funzione `pnbinom(q, size, prob, lower.tail = TRUE)`, il cui i parametri `size` e `prob` sono esattamente come sopra, mentre `q` e `lower.tail` sono come nelle corrispondenti funzioni della geometrica e della binomiale: il primo è il punto in cui calcoliamo la funzione di ripartizione F, mentre il secondo è un parametro logico che determina se calcoliamo $F_X(q)$ (il default) o il suo complementare $1 - F_X(q)$. Attenzione che in realtà `pnbinom` (così come le altre funzioni della famiglia binomiale negativa) ha un ulteriore parametro `mu`, che a noi non interessa, ma che rende necessario esplicitare sempre il nome del parametro `lower.tail`: dobbiamo scrivere `pnbinom(2, 4, 0.01, lower.tail = TRUE)`, perché `pnbinom(2, 4, 0.01, TRUE)` dà errore (il valore `TRUE` è dove la funzione si aspetta il valore per `mu`).

In modo del tutto analogo alle altre distribuzioni viste finora, abbiamo altre due funzioni in R: `rnbinom` e `qnbinom`, rispettivamente la generatrice di valori casuali distribuiti come una binomiale negativa di parametri assegnati e la funzione quantile.

Avremmo potuto rispondere alla domanda nell'Esempio 8.19 usando il seguente codice R: `pnbinom(q = 3, size = 3, prob = 0.25, lower.tail = TRUE)`.

Lezione 14

8.5.2. Riproducibilità

Le distribuzioni che abbiamo incontrato finora sono tutte parametrizzate, ossia dipendono da almeno un parametro. Potremmo allora più correttamente parlare di *famiglie* di distribuzioni o di leggi di probabilità. Perché farlo?

Abbiamo visto come trattare la somma di variabili aleatorie, come determinare la legge della variabile aleatoria risultante. Possiamo chiederci se, prese due variabili aleatorie (indipendenti) di uguale distribuzione la loro somma sia a sua volta una variabile aleatoria con medesima legge. Pensandoci un attimo, questo è probabilmente chiedere troppo. Ma forse, lasciando un po' di margine di manovra come ad esempio i parametri di una famiglia, ci possiamo riuscire: sommando due variabili aleatorie della stessa famiglia possiamo ottenere una variabile aleatoria della stessa famiglia, magari con parametri diversi.

DEFINIZIONE 8.20. *Diciamo che una famiglia di leggi di probabilità è* riproducibile *se sommando due variabili aleatorie indipendenti con leggi di quella famiglia, se ne ottiene un'altra della stessa famiglia.*

Andiamo allora a vedere se le distribuzioni discrete che abbiamo incontrato finora sono riproducibili o no.

Esempio 8.21. Siano X e Y due variabili aleatorie indipendenti e identicamente distribuite, di legge geometrica di parametro p. Cosa possiamo dire della loro somma S?

Da quanto visto nella Proposizione 7.16, abbiamo

$$\begin{aligned}\varphi_S(k) &= \sum_{j\in\mathcal{R}_X} \varphi_X(j)\,\varphi_Y(k-j)\\ &= \sum_{j=0}^{+\infty} (1-p)^j p\,\mathbb{1}_{\{k-j\geqslant 0\}}\,(1-p)^{k-j}p\\ &= \sum_{j=0}^{k} p^2(1-p)^k\\ &= (k+1)\,p^2(1-p)^k.\end{aligned}$$

Questa non è la densità discreta di una geometrica, ma ci dovrebbe ricordare qualcosa di altro: una binomiale negativa. In effetti possiamo osservare che $\mathrm{NB}(2,p)$ ha densità discreta

$$\binom{n+k-1}{n-1} p^n (1-p)^k = (k+1)\,p^2(1-p)^k,$$

ossia la somma di due variabili aleatorie indipendenti geometriche di parametro p è una binomiale negativa di parametri 2 e p.

Le geometriche non sono quindi riproducibili, ma se osserviamo che una geometrica di parametro p è una binomiale negativa di parametri 1 e p, abbiamo un suggerimento su quale possa essere una famiglia di distribuzioni riproducibili.

PROPOSIZIONE 8.22. *La famiglia delle distribuzioni binomiali negative a parametro p fissato è riproducibile. In particolare la somma di una binomiale negativa di parametri n e p e di una (indipendente dalla prima) di parametri m e p è distribuita come una binomiale negativa di parametri $n+m$ e p.*

Dimostrazione. Si può fare come conto, oppure usando il fatto che una binomiale negativa è somma di geometriche. [TBA] □

Osservazione 8.23. Il significato di questo risultato non è molto sorprendente, se lo scriviamo esplicitamente: date $X\sim\mathrm{NB}(n,p)$ e $Y\sim\mathrm{NB}(m,p)$ indipendenti, allora $X+Y\sim\mathrm{NB}(m+n,p)$. Se pensiamo all'interpretazione di X e Y, stiamo dicendo che la il numero di insuccessi prima di ottenere n successi più il numero di insuccessi prima di ottenere m successi ha la stessa distribuzione del numero di insuccessi prima di avere $n+m$ successi. In altre parole, stiamo "resettando" dopo aver raggiunto i primi n successi.

La proprietà che c'è sotto è l'assenza di memoria delle geometriche, perché alla fine possiamo scrivere ogni binomiale negativa come somma di geometriche.

Quanto abbiamo visto con le binomiali negative dovrebbe suggerirci qualcos'altro: se abbiamo definito le binomiali come somme di Bernoulliane, è lecito chiedersi se anche le binomiali siano riproducibili.

PROPOSIZIONE 8.24. *La famiglia delle distribuzioni binomiali a parametro p fissato è riproducibile. In particolare la somma di una binomiale di parametri n e p e di un'altra (indipendente dalla prima) di parametri m e p è distribuita come una binomiale di parametri $n+m$ e p.*

Dimostrazione. Consideriamo $X \sim \text{bin}(n,p)$ e $Y \sim \text{bin}(m,p)$, tra loro indipendenti. Dalla Proposizione 7.16 abbiamo

$$\varphi_{X+Y}(l) = \sum_{k=0}^{n} \varphi_X(k)\,\varphi_Y(l-k).$$

Ora possiamo osservare che gli addendi non si annullano solo se $0 \leqslant l-k \leqslant m$, cioè per $l-m \leqslant k \leqslant l$, con la condizione che $l-k$ non vada oltre a m. Continueremo a scrivere le somme per intero, con i corrispondenti coefficienti binomiali e potenze, da ignorare nel caso abbiano argomenti o esponenti negativi.

Abbiamo quindi

$$\begin{aligned}
\varphi_{X+Y}(l) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}\, \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{m}{l-k} p^{l-k} (1-p)^{m-(l-k)} \\
&= \sum_{k=0}^{n} p^l (1-p)^{n+m-l}\, \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{n}{k}\binom{m}{l-k} \\
&= p^l (1-p)^{n+m-l} \sum_{k=0}^{n} \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{n}{k}\binom{m}{l-k}
\end{aligned}$$

e per concludere ci basta mostrare che

$$\sum_{k=0}^{n} \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{n}{k}\binom{m}{l-k} = \binom{n+m}{l}.$$

Lo facciamo per induzione su n. Lasciamo cadere la funzione indicatrice e osserviamo che combinatoricamente possiamo mettere a 0 i coefficienti binomiali con la "parte sotto" fuori dai limiti. Per $n=0$,

$$\binom{0}{0}\binom{m}{l} = \binom{m}{l} = \binom{0+m}{l}.$$

Supponiamo ora che la proprietà valga per n e mostriamo che vale per $n+1$:

$$\begin{aligned}
\sum_{k=0}^{n+1} \binom{n+1}{k}\binom{m}{l-k} &= \sum_{k=0}^{n+1} \left(\binom{n}{k} + \binom{n}{k-1}\right)\binom{m}{l-k} \\
&= \sum_{k=0}^{n} \binom{n}{k}\binom{m}{l-k} + \sum_{k=1}^{n+1} \binom{n}{k-1}\binom{m}{l-k} \\
&= \sum_{k=0}^{n} \binom{n}{k}\binom{m}{l-k} + \sum_{h=0}^{n} \binom{n}{h}\binom{m}{l-h-1} \\
&= \binom{n+m}{l} + \binom{n+m}{l-1} \\
&= \binom{n+m+1}{l}
\end{aligned}$$

in cui abbiamo usato due volte l'identità 4. della Proposizione 1.26. □

8.6. IPERGEOMETRICHE

Un altro esperimento descritto da una variabile aleatoria Bernoulliana è l'estrazione di una biglia da un'urna di composizione nota. In un'urna ci sono m palline bianche e n palline nere, cioè la proporzione di biglie bianche sul totale è $p = \frac{m}{m+n}$, la variabile aleatoria "estrazione di una biglia bianca" ha legge Bernoulliana $\text{bin}(1,p)$. Possiamo allora vedere la variabile aleatoria che conta le biglie bianche tra k estratte *con reimmissione* dall'urna come una binomiale $\text{bin}(k,p)$.

Se siamo invece interessati alla variabile aleatoria che conta il numero di biglie bianche tra k estratte *senza reimmissione*, abbiamo bisogno di introdurre una nuova distribuzione.

DEFINIZIONE 8.25. *Data un'urna contenente m biglie bianche e n biglie nere, chiamiamo* ipergeometrica di parametri k, n e m *la variabile aleatoria X che conta il numero di palline bianche tra k estratte dall'urna senza reimmissione. Scriviamo in questo caso* $X \sim \text{hyp}(k,m,n)$.

Ricaviamo la densità discreta di una variabile aleatoria ipergeometrica di parametri k, m e n. Innanzitutto abbiamo dei vincoli su k, $0 \leqslant k \leqslant m+n$, perché non possiamo estrarre più biglie di quelle presenti nell'urna. In tutto abbiamo $\binom{n+m}{k}$ modi di estrarre k biglie tra $m+n$. Vogliamo contare il numero b di biglie bianche estratte, in altre parole b delle k estratte saranno bianche e le rimanenti $k-b$ saranno nere. Anche questo impone dei vincoli su b: da un lato $0 \leqslant b \leqslant m$, perché non possiamo pescare più biglie bianche di quelle che ci sono, dall'altro $0 \leqslant k-b \leqslant n$ (ossia $k-n \leqslant b \leqslant k$) perché non possiamo pescare più biglie nere di quelle che ci sono. Possiamo ora contare quanti sono le possibili estrazioni a noi favorevoli: dobbiamo scegliere b biglie bianche tra le m disponibili e $k-b$ biglie nere tra le n disponibili e lo possiamo fare in $\binom{m}{b}\binom{n}{k-b}$ modi. Allora

$$\varphi_X(b) = \begin{cases} \frac{\binom{m}{b}\binom{n}{k-b}}{\binom{n+m}{k}} & \max\{0,k-n\} \leqslant b \leqslant \min\{k,m\} \\ 0 & \text{altrimenti.} \end{cases}$$

Osservazione 8.26. Grazie a quanto visto nella dimostrazione della Proposizione 8.24, abbiamo immediatamente che

$$\sum_{b=0}^{k} \varphi_X(b) = \sum_{b=\max\{0,k-n\}}^{\min\{k,m\}} \varphi_X(b) = \frac{\sum_{b=\max\{0,k-n\}}^{\min\{k,m\}} \binom{m}{b}\binom{n}{k-b}}{\binom{n+m}{k}} = 1.$$

Esempio 8.27. Un'azienda produce 400 tastiere al giorno e di queste 10 sono difettose. Se ogni giorno l'azienda controlla 5 tastiere tra quelle prodotte, come sarà distribuito il numero di quelle difettose tra le tastiere testate?

Chiamiamo D la variabile aleatoria che conta il numero di tastiere difettose tra quelle testate in un dato giorno. Possiamo vedere la situazione in termini di un'urna contenente 400 palline (le tastiere prodotte), di cui 10 bianche (le tastiere difettose): estraiamo senza reimmissione 5 biglie (le tastiere da controllare) e ci chiediamo quante di queste siano bianche. Allora $D \sim \text{hyp}(5, 10, 400-10)$.

Massima verosimiglianza (divagazione) Nella realtà, però, in una situazione come quella dell'Esempio 8.27 l'azienda *non* sa quante siano le tastiere difettose, ma sa quante sono quelle difettose tra quelle testate. L'uso della probabilità per l'azienda sta nello *stimare* il valore più plausibile del numero di tastiere difettose prodotte, sapendo quante ne ha viste di difettose tra le testate. Cerchiamo di riscrivere in modo più esplicito questo problema. Cominciamo con gli ingredienti:

- M è il numero di tastiere difettose prodotte al giorno; è la quantità (incognita) che vogliamo stimare.
- t è il numero di tastiere prodotte al giorno; è noto.
- N è il numero di tastiere non difettose prodotte al giorno; non è noto, ma sappiamo che $N = t-M$.
- k è il numero (noto) di tastiere controllate ogni giorno.

– b è il numero *osservato* di tastiere controllate e difettose.

Inoltre, siccome sappiamo che il numero di tastiere difettose tra quelle testate (visto come variabile aleatoria, prima di osservare b) ha una distribuzione $D \sim \text{hyp}(k, M, t-M)$, possiamo scrivere che

$$P(D=b) = \frac{\binom{M}{b}\binom{t-M}{k-b}}{\binom{t}{k}},$$

cioè dati i valori noti t e k, se sapessimo M avremmo la probabilità di osservare proprio b tastiere difettose tra quelle controllate. Ma cambiamo il punto di vista: sapendo che abbiamo visto $D=b$, qual è il valore di M per cui era massima la probabilità di vedere proprio b? Qual è *a posteriori* il valore più verosimile per M?

Infatti, a ogni possibile valore m di M corrisponde una certa probabilità di avere $D=b$, come abbiamo visto nella prima parte dell'esempio, cosa che possiamo scrivere in forma di probabilità condizionata come

$$P(D=b \mid M=m) = \frac{\binom{m}{b}\binom{t-m}{k-b}}{\binom{t}{k}},$$

ma grazie al teorema di Bayes

$$P(M=m \mid D=b) = \frac{P(D=b \mid M=m)\, P(M=m)}{P(D=b)}, \tag{8.3}$$

che è la probabilità che vogliamo massimizzare, variando m, per trovare il valore m più verosimile, più compatibile con l'osservazione fatta che $D=b$.

Per massimizzare la (8.3) iniziamo con l'osservare che $P(D=b)$ è costante al variare di m (è uguale per tutti i valori m di M), quindi non gioca alcun ruolo nella massimizzazione, ce ne possiamo dimenticare. Passiamo allora al termine $P(M=m)$ che compare al numeratore. Questo contiene la nostra valutazione *a priori* della plausibilità dei valori di M. In assenza di altre informazioni (per esempio il primo giorno in cui vengono fatti i test) possiamo ipotizzare che M sia equidistribuita tra i valori possibili, cioè l'insieme $\{0,1,\ldots,t\}$, ossia che per ogni $m \in \{0,1,\ldots,t\}$, $P(M=m) = \frac{1}{t+1}$.

Il problema di massimizzare la (8.3) è diventato allora trovare il valore $\overline{m} \in \{0,1,\ldots,t\}$ che massimizza l'ipergeometrica:

$$\overline{m} = \text{argmax}_{m \in \{0,1,\ldots,t\}} \frac{\binom{m}{b}\binom{t-m}{k-b}}{\binom{t}{k}} = \text{argmax}_m \binom{m}{b}\binom{t-m}{k-b}.$$

In questo modo otteniamo il candidato più verosimile come numero di tastiere difettose, avendone testate k, di cui b erano guaste.

Proviamo con qualche numero: $t=400$, $k=5$, $b=2$, ossia delle 5 testate, 2 sono difettose.

```
m <- 0:400
a <- choose(m,2)*choose(400-m,3)
m[which.max(a)]
```

8.6.1. Ipergeometriche in R

Le funzioni per una variabile aleatoria ipergeometrica sono `dhyper(x, m, n, k)` per la densità discreta, dove `x` è il punto in cui vogliamo calcolare la densità φ, `m` è il numero di biglie bianche nell'urna, `n` il numero di biglie nere nell'urna e `k` il numero di biglie estratte dall'urna (i nomi dei parametri coincidono con quelli dati sopra nella Definizione 8.25).

La funzione di ripartizione per un'ipergeometrica è la funzione `phyper(q, m, n, k, lower.tail = TRUE)`, il cui i parametri `m`, `n` e `k` sono esattamente come sopra, mentre `q` e `lower.tail` sono come nelle corrispondenti funzioni già viste per le altre variabili aleatorie: il primo è il punto in cui calcoliamo la funzione di ripartizione F, mentre il secondo è un parametro logico che determina se calcoliamo $F_X(q)$ (il default) o il suo complementare $1-F_X(q)$.

In modo del tutto analogo alle altre distribuzioni viste finora, abbiamo altre due funzioni in R: `rhyper` e `qhyper`, rispettivamente la generatrice di valori casuali distribuiti come una binomiale negativa di parametri assegnati e la funzione quantile. Per `rhyper` dobbiamo solamente fare attenzione al fatto che il numero di realizzazioni (che per le altre variabili era `n`) è in questo caso `nn`: `rhyper(nn, m, n, k)`.

Esempio 8.28. Il Blackjack (o 21) è un gioco d'azzardo, molto diffuso negli Stati Uniti e reso famoso da un film (21, appunto). Nella sua versione base[8.6]a un solo giocatore contro il banco, si gioca con un normale mazzo da 52 carte, di cui 2 vengono date al giocatore. Le figure hanno un valore pari a 10, gli assi hanno un valore uguale a 1 o 11 e le altre carte hanno il loro valore nominale (un 7 vale 7). Il giocatore fa blackjack se le sue due carte sono una carta di valore uguale a 10 e un asso (per un totale di 21 punti). Qual è la probabilità di fare blackjack?

Cominciamo con il calcolare la probabilità che le due carte del giocatore siano entrambe o assi o carte di valore 10. Usiamo una distribuzione ipergeometrica di parametri $k=2$ (le carte date al giocatore), $m=20$ (3 figure, un asso e un 10 per ciascuno dei quattro semi) e $n=32$ (52 carte totali meno le 16 "buone"). Vogliamo $\varphi(2)\approx 14\%$ (possiamo calcolarla come `dhyper(x = 2, m = 20, n = 32, k = 2)` in R).

Ora calcoliamo la probabilità che entrambe le carte siano assi, usando un'ipergeometrica di parametri $k=2, m=4, n=48$: $\varphi(2)\approx 0.5\%$ (in R `dhyper(x = 2, m = 4, n = 48, k = 2)`).

Ancora, calcoliamo la probabilità che entrambe le carte abbiano valore 10, questa volta con un'ipergeometrica di parametri $k=2$, $m=12$, $n=40$: $\varphi(2)\approx 9\%$ (in R `dhyper(x = 2, m = 16, n = 36, k = 2)`).

A questo punto possiamo ricavare la probabilità di fare blackjack sottraendo le ultime due probabilità dalla prima: otteniamo all'incirca il 4.8%.

Esempio 8.29. Nella versione del poker nota come Texas hold'em, ogni giocatore ha una mano di 2 carte personali, da combinare con le carte comuni (fino a 5). Il mazzo è un normale mazzo a 52 carte, con 13 carte per ognuno dei 4 semi. Un giocatore ha in mano un 3 di cuori e un 7 di picche. Le prime due carte che escono sul tavolo sono il 3 di picche e la donna di picche. Con che probabilità le prossime tre carte gli faranno avere colore o un poker?

Cominciamo con il colore (ossia avere 5 carte dello stesso seme, non necessariamente ordinate, nel qual caso sarebbe scala colore). La probabilità che lo ottenga è pari alla probabilità che due delle prossime tre carte estratte dal mazzo siano di picche, di cui nel mazzo ne restano $13-3=10$: abbiamo quindi un'ipergeometrica di parametri $k=3$, $m=10$ e $n=38$ (perché nel mazzo restano 48 carte e di queste 10 sono quelle che vanno bene), di cui vogliamo calcolare la densità discreta in $b=2$ e $b=3$. Aiutandoci con R (`phyper(q = 1, m = 10, n = 38, k = 3, lower.tail = FALSE)`) otteniamo 10.6%. A questa dobbiamo però sottrarre la probabilità di ottenere una scala colore, che possiamo avere con le carte di picche da 3 a 7 (ossia estraendo 4, 5 e 6 di picche). Questa probabilità è minore di 10^{-4} (`dhyper(x = 3, m = 3, n = 45, k = 3)`) e la differenza è allora circa 10.6%.

Passiamo ora al poker: l'unica coppia che ha già in mano è quella di 3, per completarla dovrebbero uscire i due 3 rimanenti. Di nuovo abbiamo un'ipergeometrica di parametri $k=3$, $m=2$ e $n=46$, di cui calcoliamo la densità discreta in $b=2$. La probabilità è 0.27%. In alternativa, può ottenere un poker anche pescando le 3 carte rimanenti di un segno tra 7 e donna. Per ciascuno di questi casi abbiamo una ipergeometrica di parametri $k=3$, $m=3$ e $n=45$, di cui calcoliamo la densità discreta in $b=3$, ottenendo complessivamente (cioè sommando le due probabilità) 0.01%.

Non è possibile che abbiamo contemporaneamente due poker o un poker e colore, quindi possiamo sommare le probabilità dei vari eventi, mutualmente esclusivi. Complessivamente, dunque, la probabilità cercata è 10.9%.

8.6. Quella giocata nei casinò è leggermente diversa, con le modifiche che la rendono interessante per la storia (basata su fatti realmente accaduti) narrata nel film.

Osservazione 8.30. Avendo parlato di riproducibilità per altre famiglie di variabili aleatorie, è abbastanza naturale chiedersi se anche la famiglia delle ipergeometriche sia riproducibile. Un dubbio sulla possibilità che questo sia vero può venirci dal fatto che, a differenza dei due casi di variabili riproducibili visti finora, qui non abbiamo un parametro da tenere fisso. In realtà questa non è in sé una condizione necessaria, come vedremo.

Pensiamo a uno dei casi più semplici di variabile aleatoria ipergeometrica, $X \sim \mathrm{hyp}(1,m,n)$, ossia estraiamo, da un'urna con m biglie bianche e n biglie nere una sola biglia. Prendiamo ora $Y \sim X$, ma indipendente e consideriamo la somma $X+Y$. Se estraiamo una sola biglia, la differenza tra estrazione con e senza reimmissione si perde, e come abbiamo già osservato $X \sim \mathrm{bin}\left(1, \frac{m}{m+n}\right)$. Allora dalla riproducibilità delle binomiali, sappiamo che $X+Y \sim \mathrm{bin}\left(2, \frac{m}{m+n}\right)$, che però non è un'ipergeometrica. Infatti, nel momento in cui andiamo a estrarre due biglie, la differenza tra estrazione con o senza reimmissione diventa significativa: nel primo caso le estrazioni sono tra loro indipendenti (e abbiamo la binomiale), nel secondo non lo sono, hanno influenza le une sulle altre (e abbiamo l'ipergeometrica). In particolare la famiglia di variabili aleatorie ipergeometriche non è riproducibile. Un altro modo di convincersene è provare a semplificare i coefficienti binomiali che si ottengono scrivendo la densità discreta della somma di due ipergeometriche indipendenti.

Lezione 15

Vediamo ora un legame tra le variabili aleatorie con distribuzione ipergeometrica e quelle con distribuzione binomiale.

PROPOSIZIONE 8.31. *Siano $\{a_i\}_i$ e $\{b_i\}_i$ due successioni di numeri interi non negativi che tendono monotonamente a $+\infty$ e tali che $\lim_{i\to+\infty} \frac{a_i}{a_i+b_i} = \alpha$, per qualche $\alpha \in [0,1]$. Allora*

$$\frac{\binom{a_i}{k}\binom{b_i}{n-k}}{\binom{a_i+b_i}{n}} \xrightarrow[i\to+\infty]{} \binom{n}{k} \alpha^k (1-\alpha)^{n-k}.$$

Dimostrazione. Procediamo per passi.

i. Osserviamo che $\frac{b_i}{a_i+b_i} = 1 - \frac{a_i}{a_i+b_i} \xrightarrow[i\to+\infty]{} 1-\alpha$.

ii. Per ogni c, d costanti, $\frac{a_i-c}{a_i+b_i-d} = \frac{a_i}{a_i+b_i} \cdot \frac{1-\frac{c}{a_i}}{1-\frac{d}{a_i+b_i}} \xrightarrow[i\to+\infty]{} \alpha$.

iii. Combinando i primi due punti, per ogni c, d costanti, $\frac{b_i-c}{a_i+b_i-d} \xrightarrow[i\to+\infty]{} 1-\alpha$.

iv. A questo punto abbiamo tutto quello che ci occorre:

$$\begin{aligned}
\frac{\binom{a_i}{k}\binom{b_i}{n-k}}{\binom{a_i+b_i}{n}} &= \frac{(a_i)!\,(b_i)!\,(a_i+b_i-n)!\,n!}{k!\,(a_i-k)!\,(n-k)!\,(b_i-n+k)!\,(a_i+b_i)} \\
&= \binom{n}{k} \frac{(a_i)!}{(a_i-k)!} \frac{(b_i)!}{(b_i-n+k)!} \frac{(a_i+b_i-n)!}{(a_i+b_i)!} \\
&= \binom{n}{k} \frac{a_i\,(a_i-1)\cdots(a_i-(k-1))}{(a_i+b_i)\,(a_i+b_i-1)\cdots(a_i+b_i-k+1)} \frac{b_i\cdots(b_i-(n-k-1))}{(a_i+b_i-k)\cdots(a_i+b_i-n+1)} \\
&\xrightarrow[i\to+\infty]{} \binom{n}{k} \alpha^k (1-\alpha)^{n-k}
\end{aligned}$$

perché nella penultima riga abbiamo k termini della forma $\frac{a_i-c}{a_i+b_i-d}$ ed $n-k$ termini della forma $\frac{b_i-c}{a_i+b_i-d}$. □

Osservazione 8.32. In termini di variabili aleatorie stiamo dicendo che se una popolazione a_i+b_i cresce all'infinito, convergendo però a una proporzione determinata di "a" e "b" (rispettivamente α e $1-\alpha$), allora se abbiamo una successioni di variabili aleatorie ipergeometriche

$$X_i \sim \mathrm{hyp}(n, a_i, b_i)$$

e una variabile aleatoria $X \sim \text{bin}(n, \alpha)$, allora la successione delle densità discrete φ_{X_i} converge alla densità discreta φ_X, ossia in un qualche senso[8.7] le variabili ipergeometriche tendono a una binomiale o, meglio, le leggi delle ipergeometriche tendono alla legge binomiale.

8.7. POISSON

Abbiamo rotto il ghiaccio con le sequenze di variabili aleatorie. Vediamone ora altre che entrano in gioco nel seguente problema.

Esempio 8.33. In una partita di Premier League vengono segnati in media[8.8] 2.5 gol a partita[8.9]. Vorremmo sapere quale può essere una distribuzione di probabilità del numero di gol in una partita.

Possiamo pensare, in prima approssimazione, di descrivere questo fenomeno nel modo seguente: dividiamo la partita (da 90') in 5 periodi da 18', in ciascuno dei quali abbiamo una probabilità $\frac{1}{2}$ di vedere un gol. Volendo una variabile aleatoria che conta i gol, ci riconduciamo a un modello che già conosciamo: il segnare un gol è per ogni periodo da 18' una Bernoulliana di parametro $\frac{1}{2}$ e il numero di gol in una partita è quindi una binomiale di parametri $n = 5$ e $p = \frac{1}{2}$. Chiamiamo allora questa variabile aleatoria che conta i gol $X_1 \sim \text{bin}\left(5, \frac{1}{2}\right)$. Osserviamo che, anche se non abbiamo ancora definito cosa sia la media di una Bernoulliana o di una binomiale, da un punto di vista intuitivo, ogni 18' ci aspettiamo di vedere $\frac{1}{2}$ gol, per un totale di 2.5 gol in una partita.

Questa descrizione del fenomeno, però, non ci piace troppo: dalle proprietà delle binomiali, sappiamo che il supporto di X_1 è l'insieme $\{0, \dots, 5\}$, cioè non è possibile che ci siano più di 5 gol a partita, cosa che non accade troppo spesso, ma che non possiamo escludere del tutto. Più grave, da un punto di vista modellistico, è che non è possibile avere più di un gol in ogni periodo da 18'.

Possiamo allora pensare di passare ad una griglia più fine: 10 periodi da 9' ciascuno, in cui però la probabilità di vedere un gol si è anch'essa dimezzata, passando da $\frac{1}{2}$ a $\frac{1}{4}$. Abbiamo quindi una seconda variabile aleatoria candidata a contare il numero di gol: $X_2 \sim \text{bin}\left(10, \frac{1}{4}\right)$. È un miglioramento rispetto a prima, ora possiamo vedere fino a 10 gol in una partita e fino a 1 gol in ogni periodo da 9', ma possiamo continuare a raffinare la nostra griglia:

$$X_3 \sim \text{bin}\left(20, \frac{1}{8}\right)$$
$$X_4 \sim \text{bin}\left(40, \frac{1}{16}\right)$$

E in realtà nessuno ci obbliga a dimezzare la durata degli intervallini ogni volta, quello che importa è mantenere costante il prodotto $n \cdot p = 2.5$ (come vedremo, $n \cdot p$ è proprio la media o *valore atteso* di una variabile aleatoria binomiale di parametri n e p). Quindi continuiamo con

$$X_5 \sim \text{bin}\left(45, \frac{1}{18}\right),$$

in cui abbiamo 45 Bernoulliane che descrivono periodi di gioco da 2 minuti ciascuna, e ancora

$$X_6 \sim \text{bin}\left(90, \frac{1}{36}\right)$$
$$X_7 \sim \text{bin}\left(180, \frac{1}{72}\right)$$

in cui stiamo considerando finestre da 1 minuto o da 30″ ciascuna.

8.7. Vedremo meglio più avanti i concetti di convergenza per variabili aleatorie.

8.8. Non abbiamo ancora definito il concetto di media per una variabile aleatoria, anche se non manca molto, lo incontreremo nel Capitolo 9. Tuttavia in questo caso stiamo parlando della media *empirica*, ossia il numero totale di gol segnati in Premier League diviso per il numero delle partite.

8.9. Dati del 14.04.2021.

Cosa succede se continuiamo a sviluppare una successione di questo tipo? Converge a qualcosa? E se sì, a cosa converge?

DEFINIZIONE 8.34. *Diciamo che una variabile aleatoria discreta X è* di Poisson[8.10] (o Poissoniana) di parametro λ, *con λ numero reale positivo, se ha densità discreta*

$$\varphi_X(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & k \in \mathbb{N} \\ 0 & \textit{altrimenti}. \end{cases} \tag{8.4}$$

In questo caso scriviamo $X \sim \text{Pois}(\lambda)$.

Osservazione 8.35. La funzione φ_X definita in (8.4) soddisfa le proprietà di una densità discreta di probabilità, in particolare è non negativa e ha somma uguale a 1 sul proprio supporto. Per convincerci di questa seconda proprietà, osserviamo che

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \cdots = \sum_{k=0}^{+\infty} \frac{x^k}{k!},$$

quindi abbiamo

$$\sum_{k \in \mathcal{R}_X} \varphi_X(k) = \sum_{k \in \mathbb{N}} \left(\frac{\lambda^k}{k!} e^{-\lambda} \right) = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Osservazione 8.36. Come abbiamo già visto, una variabile aleatoria binomiale conta il numero di successi in n prove indipendenti, tutte con uguale probabilità p di successo. Tuttavia n o p possono non essere noti con precisione, oppure possono essere, rispettivamente, molto grande e molto piccolo. In questa situazione può venirci in aiuto la variabile aleatoria di Poisson, per cui abbiamo solamente bisogno di conoscere un parametro λ che gioca il ruolo di $n \cdot p$ (e che intuitivamente è il numero di successi che ci aspettiamo in media, come poi confermeremo rigorosamente più avanti, nel Capitolo 9).

Alcuni esempi tipici in cui possiamo usare una variabile aleatoria di Poisson per descrivere (o modellizzare) il fenomeno sono:

- il numero di email ricevute da un utente nel corso di una giornata (a priori non possiamo dare un limite superiore al numero di email che potrebbe ricevere);
- il numero di morti sul lavoro in Italia in un dato giorno (abbiamo molta incertezza sul numero n dei lavoratori attivi quel giorno, pur avendo un'idea del suo ordine di grandezza, così come sul valore più plausibile di p, ma abbiamo delle statistiche storiche che ci dicono che il numero medio di morti sul lavoro al giorno è stato 3.5 nel 2020[8.11]);
- il numero di domande di iscrizione a Informatica a Trento, anno dopo anno (in media non cambieranno troppo, ma non sappiamo quantificare con certezza il numero n di coloro che considerano Informatica come indirizzo di studi e, per ciascuno di essi, quale sia la probabilità p che alla fine si iscrivano).

Possiamo ora riprendere l'Esempio 8.33 e rendere matematicamente solido quanto detto prima sul comportamento limite della successione di binomiali.

PROPOSIZIONE 8.37. *Sia $\{p_n\}_n$ una successione di numeri in $[0,1]$ tali che $\lim_{n\to+\infty} p_n \cdot n = \lambda$, per qualche numero reale positivo λ. Allora, per ogni k naturale*

$$\lim_{n\to+\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

8.10. Siméon Denis Poisson (1781 – 1840).

8.11. Dati INAIL sul numero di denunce di infortuni con esito mortale. Non è detto che diano una rappresentazione completa del fenomeno, a causa degli infortuni (anche mortali) non denunciati.

Dimostrazione. Cominciamo con lo scrivere esplicitamente il primo membro:

$$\begin{aligned}
\lim_{n\to+\infty}\binom{n}{k}p_n^k(1-p_n)^{n-k} &= \lim_{n\to+\infty}\frac{n(n-1)\cdots(n-k+1)}{k!\,n^k}n^k p_n^k(1-p_n)^{n-k}\\
&= \frac{1}{k!}\lim_{n\to+\infty}\left[\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\right](n\cdot p_n)^k\left(1-\frac{n\cdot p_n}{n}\right)^{n-k}\\
&= \frac{1}{k!}\lambda^k\lim_{n\to+\infty}\left(1-\frac{\lambda}{n}\right)^{n-k}\\
&= \frac{\lambda^k}{k!}e^{-\lambda}\lim_{n\to+\infty}\left(1-\frac{\lambda}{n}\right)^{-k}\\
&= \frac{\lambda^k}{k!}e^{-\lambda}
\end{aligned}$$

in cui nella prima riga abbiamo moltiplicato e diviso per n^k, nella seconda riga, i termini in arancione convergono a 1 al limite e i termini in verde convergono a λ e nel passare dalla terza alla quarta riga abbiamo usato una caratterizzazione della funzione esponenziale. □

Osservazione 8.38. In termini di variabili aleatorie stiamo dicendo che se abbiamo una successione di variabili aleatorie binomiali

$$X_n \sim \mathrm{bin}(n,p_n)$$

e una variabile aleatoria $X\sim\mathrm{Pois}(\lambda)$, con $\lambda=\lim_{n\to+\infty}n\cdot p_n$, allora la successione delle densità discrete φ_{X_n} converge alla densità discreta φ_X, ossia la variabile aleatoria di Poisson è il "limite" delle variabili aleatorie binomiali.

PROPOSIZIONE 8.39. *Le variabili aleatorie Poissoniane sono riproducibili.*

Dimostrazione. Vogliamo mostrare che, date due variabili aleatorie indipendenti $X_1\sim\mathrm{Pois}(\lambda_1)$ e $X_2\sim\mathrm{Pois}(\lambda_2)$, anche la loro somma ha legge Poissoniana. Consideriamone la densità discreta:

$$\begin{aligned}
\varphi_{X_1+X_2}(n) &= \sum_{k=0}^{n}\varphi_{X_1}(k)\,\varphi_{X_2}(n-k)\\
&= \sum_{k=0}^{n}\frac{n!}{n!}\frac{\lambda_1^k}{k!}e^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}e^{-\lambda_2}\\
&= \left[\sum_{k=0}^{n}\binom{n}{k}\lambda_1^k\lambda_2^{n-k}\right]\frac{e^{-(\lambda_1+\lambda_2)}}{n!}\\
&= \frac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)}
\end{aligned}$$

in cui abbiamo messo in evidenza il binomio di Newton nel passare dalla penultima all'ultima riga. Osserviamo che in questo modo abbiamo mostrato che $X_1+X_2\sim\mathrm{Pois}(\lambda_1+\lambda_2)$. □

Esempio 8.40. Siano $X_1\sim\mathrm{Pois}(\lambda_1)$ e $X_2\sim\mathrm{Pois}(\lambda_2)$ due variabili aleatorie indipendenti e sia S la loro somma. Vogliamo determinare la legge di X_1 condizionata a S.

Partiamo dalla definizione di densità discreta condizionata:

$$\begin{aligned}
\varphi_{X_1|S}(k|n) &= P(X_1=k|S=n)\\
&= \frac{P(X_1=k,S=n)}{P(S=n)}\\
&= \frac{P(X_1=k,X_2=n-k)}{P(S=n)}\\
&= \frac{P(X_1=k)\,P(X_2=n-k)}{P(S=n)}\\
&= \frac{\lambda_1^k}{k!}e^{-\lambda_1}\frac{\lambda_2^{n-k}}{(n-k)!}e^{-\lambda_2}\frac{n!}{(\lambda_1+\lambda_2)^n}e^{\lambda_1+\lambda_2}\\
&= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}.
\end{aligned}$$

Quindi $\varphi_{X_1|S}(\cdot|n) \sim \text{bin}\left(n, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$.

8.7.1. Poissoniane in R

Per una variabile aleatoria di Poisson, le funzioni in R sono:

- la densità discreta `dpois(x, lambda)`, con `x` il punto in cui vogliamo calcolare la densità discreta φ e `lambda` il parametro della Poisson;
- la funzione di ripartizione `ppois(q, lambda, lower.tail = TRUE)`, in cui `lambda` è lo stesso della funzione di densità discreta, mentre `q` e `lower.tail` sono come nelle corrispondenti funzioni già viste per le altre variabili aleatorie;
- il generatore casuale a distribuzione Poissoniana è `rpois(n, lambda)`, con `n` il numero di realizzazioni da generare;
- la funzione quantile è `qpois(p, lambda, lower.tail = TRUE)`, ma la vedremo meglio più avanti.

Se torniamo all'Esempio 8.33, possiamo generare il numero di gol nelle 10 partite di una giornata con il comando `rpois(n = 10, lambda = 2.5)`, ottenendo (ad esempio) la seguente realizzazione `3 2 3 3 2 2 5 2 3 4`, oppure `4 3 4 2 1 1 1 1 2 1`.

CAPITOLO 9

SPERANZA, VARIANZA E ALTRI INDICATORI

Spesso la legge di una variabile aleatoria X non è nota. Un modo di avere qualche informazione su X è considerarne alcuni indicatori, quantità deterministiche che riassumono alcune delle caratteristiche della distribuzione di una variabile aleatoria. Il primo indicatore che consideriamo è la media, ossia un valore *deterministico* (cioè un numero) che ci dà in un certo senso[9.1] il centro della distribuzione. Conoscere la media è però avere molta meno informazione rispetto al conoscere la legge: quest'ultima è una funzione, mentre la media è un numero.

9.1. VARIABILI ALEATORIE DISCRETE

Cominciamo col vedere la definizione di valore atteso per le variabili aleatorie discrete.

DEFINIZIONE 9.1. *Chiamiamo* valore atteso, speranza matematica *o* media *di una variabile discreta X il baricentro della sua distribuzione, ossia*

$$\mathbb{E}[X] = E[X] = \sum_{k \in \mathcal{R}_X} k\, \varphi_X(k).$$

Possiamo notare che la speranza è una media pesata dei possibili valori k assunti da X, i cui pesi sono le corrispondenti probabilità $\varphi_X(k) = P(X = k)$.

Osservazione 9.2. Non è detto che la speranza di una variabile aleatoria sia finita (Esempio 9.3), che sia positiva o che sia definita (la serie che la definisce potrebbe non convergere[9.2]). In generale, nel seguito, considereremo solamente variabili aleatorie X la cui speranza è definita e finita, salvo diversamente specificato.

Esempio 9.3. (Paradosso di San Pietroburgo) A Nicholas viene proposto il seguente gioco: lancia una moneta equilibrata e, se la prima Testa esce al lancio n, vince 2^n monete. Quante monete vincerà in media?

Usiamo la definizione appena data, chiamando X la variabile aleatoria che rappresenta la vincita. Ci occorre solamente $\varphi_X(2^n) = P(T_1 = n)$, dove T_1 è l'istante di prima uscita di una testa nel corrispondente schema di Bernoulli. Dobbiamo allora ricordare quale sia la probabilità che la prima testa esca al lancio n-simo, cioè $\left(\frac{1}{2}\right)^{n-1}\frac{1}{2} = \frac{1}{2^n}$. Allora

$$E[X] = \sum_{x \in \mathcal{R}_X} x\, \varphi_X(x) = \sum_{n \in \mathbb{N} \setminus \{0\}} 2^n \cdot \frac{1}{2^n} = \sum_{n \in \mathbb{N} \setminus \{0\}} 1 = +\infty.$$

Vale la pena notare che pur avendo speranza infinita, X è una variabile aleatoria finita con probabilità 1, infatti

$$P(X = +\infty) = \lim_{n \to +\infty} 2^{-n} = 0,$$

perché chiedere che sia infinita significa che tutti i lanci devono essere Croce.

9.1. Vedremo più avanti che non è il solo.

9.2. Consideriamo separatamente i casi in cui la serie diverge a $\pm\infty$ rispetto a quelli in cui non c'è proprio convergenza, ossia non esiste limite, né finito né infinito.

La Definizione 9.1 richiede di pesare ogni possibile risultato con la sua probabilità. Un'immediata generalizzazione, allora, è quella in cui consideriamo la probabilità condizionata a un evento H,

$$E[X|H] = \sum_{x \in \mathcal{R}_X} x \cdot P(X = x|H)$$

detta *speranza di X condizionata ad H*, ma anche a eventi speciali, quale ad esempio il valore assunto da un'altra variabile aleatoria,

$$E[X|Y = y] = \sum_{x \in \mathcal{R}_X} x \cdot P(X = x|Y = y) = \sum_{x \in \mathcal{R}_X} x \cdot \varphi_{X|Y}(x|y),$$

la *speranza di X condizionata al fatto che Y assuma il valore y*. Si può andare ancora oltre e considerare la speranza di una variabile aleatoria condizionata a un'altra variabile aleatoria (e non al suo valore) o a una tribù. Si parla in questo caso di *speranza condizionata*, ma non la tratteremo in questo corso.

Esempio 9.4. Sia $X \sim \text{bin}(1,p)$, calcoliamone la speranza. Dalla definizione abbiamo

$$E[X] = \sum_{k=0}^{1} k \cdot \varphi_X(k) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Avendo la definizione, possiamo usarla per calcolare la media di altre distribuzioni note e, in generale, di una qualunque variabile aleatoria discreta. Tuttavia per farlo abbiamo bisogno di conoscere la densità discreta della variabile aleatoria. I prossimi risultati ci danno delle scorciatoie.

TEOREMA 9.5. *Siano X una variabile aleatoria di densità discreta φ_X e $Y = g(X)$. Allora*

$$E[Y] = \sum_{k \in \mathcal{R}_X} g(k)\, \varphi_X(k).$$

Dimostrazione. Partiamo dalla definizione di speranza e sfruttiamo un risultato sulle trasformazioni di variabili aleatorie discrete visto nel Capitolo 6,

$$\begin{aligned} E[Y] &= \sum_{y \in \mathcal{R}_Y} y \cdot \varphi_Y(y) \\ &= \sum_{y \in \mathcal{R}_Y} y \cdot \sum_{x \in g^{-1}(\{y\})} \varphi_X(x) \\ &= \sum_{y \in \mathcal{R}_Y} \sum_{x \in g^{-1}(\{y\})} g(x) \cdot \varphi_X(x) \\ &= \sum_{x \in \mathcal{R}_X} g(x) \cdot \varphi_X(x), \end{aligned}$$

in cui abbiamo approfittato del fatto che se $x \in g^{-1}(\{y\})$, allora $y = g(x)$ e che $\mathcal{R}_Y = g(\mathcal{R}_X)$. □

Notiamo che per calcolare $E[Y]$ mediante il teorema precedente non abbiamo bisogno di calcolare esplicitamente φ_Y.

TEOREMA 9.6. *Siano (X,Y) un vettore aleatorio di variabili aleatorie discrete con densità congiunta $\varphi_{X,Y}$ e sia $Z = g(X,Y)$, per qualche funzione $g : \mathbb{R}^2 \to \mathbb{R}$. Allora*

$$E[Z] = \sum_{j \in \mathcal{R}_X} \sum_{k \in \mathcal{R}_Y} g(j,k) \cdot \varphi_{X,Y}(j,k).$$

Dimostrazione. [TBA] □

Esempio 9.7. Siano X e Y due d20 indipendenti tra loro. Sia $Z = \min(X,Y)$. Qual è il valore atteso della variabile aleatoria Z[9.3]?

9.3. Detta anche "tiro con svantaggio".

Siamo nelle ipotesi del Teorema 9.6, quindi

$$\begin{aligned} E[Z] &= \sum_{j=1}^{20}\sum_{k=1}^{20} \min(j,k)\,\varphi_{X,Y}(j,k) = \frac{1}{400}\sum_{j=1}^{20}\sum_{k=1}^{20} \min(j,k) \\ &= \frac{1}{400}\sum_{j=1}^{20}\left(\sum_{k=1}^{j} k + \sum_{k=j+1}^{20} j\right) = \frac{1}{400}\sum_{j=1}^{20}\left(\frac{j(j+1)}{2} + (20-j)\,j\right) \\ &= \frac{1}{400}\sum_{j=1}^{20}\left(-\frac{j^2}{2} + \frac{41}{2}j\right) = \frac{1}{800}\sum_{j=1}^{20}(j(41-j)) \\ &= \frac{5740}{800} = \frac{287}{40} = 7.175, \end{aligned}$$

mentre un normale d20 ha media 10.5.

PROPOSIZIONE 9.8. *Il valore atteso (per variabili aleatorie discrete) gode delle seguenti proprietà.* Lezione 16

Linearità. *Date due variabili aleatorie discrete X e Y e due numeri reali a e b,*

$$E[aX+Y+b] = aE[X]+E[Y]+b.$$

Prodotto di variabili aleatorie indipendenti. *Siano X e Y due variabili aleatorie discrete tra loro indipendenti, allora*

$$E[X\cdot Y] = E[X]\cdot E[Y].$$

Monotonia. *Sia X una variabile aleatoria discreta. Se $X \geqslant 0$, allora $E[X] \geqslant 0$. Inoltre l'uguaglianza vale solamente se $X \equiv 0$.*

Dimostrazione. Dimostriamo separatamente le tre proprietà.

Linearità. Per questa dimostrazione sfruttiamo il Teorema 9.6, con $g(x,y) = ax+y+b$,

$$\begin{aligned} E[aX+Y+b] &= E[g(X,Y)] = \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} g(j,k)\,\varphi_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} (aj+k+b)\,\varphi_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X} aj \sum_{k\in\mathcal{R}_Y} \varphi_{X,Y}(j,k) + \sum_{k\in\mathcal{R}_Y} k \sum_{j\in\mathcal{R}_X} \varphi_{X,Y}(j,k) + b\sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} \varphi_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X} aj\,\varphi_X(j) + \sum_{k\in\mathcal{R}_Y} k\,\varphi_Y(k) + b \\ &= aE[X]+E[Y]+b, \end{aligned}$$

in cui, nella penultima uguaglianza, abbiamo marginalizzato la densità discreta congiunta.

Prodotto. Anche in questo caso il nostro riferimento è il Teorema 9.6, con $g(x,y) = x\cdot y$,

$$\begin{aligned} E[XY] &= E[g(X,Y)] = \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} g(j,k)\,\varphi_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} jk\,\varphi_X(j)\,\varphi_Y(k) \\ &= \sum_{j\in\mathcal{R}_X} j\,\varphi_X(j) \sum_{k\in\mathcal{R}_Y} k\,\varphi_Y(k) = E[X]\,E[Y], \end{aligned}$$

in cui vale la pena sottolineare la necessità dell'ipotesi di indipendenza, per riscrivere la densità discreta congiunta come prodotto delle densità discrete marginali.

Monotonia. In questo caso partiamo dalla definizione,

$$E[X] = \sum_{k\in\mathcal{R}_X} k\,\varphi_X(k) \geqslant 0$$

perché $\varphi_X \geqslant 0$ e, per ipotesi, X è non negativa, ossia ogni elemento nel supporto $\mathcal{R}_X$ è maggiore o uguale di zero. La somma può essere nulla solamente se tutti gli addendi sono nulli, ossia se X assume solamente il valore 0. □

COROLLARIO 9.9. *Se X e Y sono due variabili aleatorie discrete tali che $P(X \geqslant Y) = 1$ (ossia $X \geqslant Y$ quasi certamente), allora $E[X] \geqslant E[Y]$. Inoltre, se vale $E[X] = E[Y]$, allora $X = Y$.*

Dimostrazione. Definiamo la variabile aleatoria $Z = X - Y$. Grazie all'ipotesi $P(X \geqslant Y) = 1$, abbiamo $P(Z \geqslant 0) = 1$ e, per linearità e monotonia della speranza,

$$E[X] - E[Y] = E[Z] \geqslant 0,$$

da cui $E[X] \geqslant E[Y]$. □

Osservazione 9.10. In generale non è vero che, date due variabili aleatorie discrete X e Y, se i loro valori attesi sono uguali, $E[X] = E[Y]$, allora le due variabili sono uguali. È necessaria l'ipotesi $P(X \geqslant Y) = 1$. Lasciandola cadere possiamo costruire dei controesempi, ad esempio $X \equiv 0$ e

$$Y = -1 + 2 \cdot \text{bin}\left(1, \frac{1}{2}\right)$$

(cioè una variabile aleatoria che assume i valori -1 e 1 ciascuno con probabilità $\frac{1}{2}$) hanno entrambe media 0, ma non sono uguali.

Abbiamo enunciato e dimostrato queste proprietà solamente per le variabili aleatorie discrete, dal momento che, per ora, abbiamo definito il valore atteso solamente per queste variabili aleatorie. Tuttavia, come vedremo nella prossima sezione, queste proprietà valgono anche per la speranza di variabili aleatorie assolutamente continue.

9.1.1. Valore atteso di alcune variabili aleatorie note

Calcoliamo ora la speranza dei modelli di variabili aleatorie discrete che abbiamo definito nel Capitolo 8.

Bernoulliane Come abbiamo già visto nell'Esempio 9.4, se $X \sim \text{bin}(1, p)$, allora $E[X] = p$.

Binomiali Sia $X \sim \text{bin}(n, p)$. Per calcolarne la speranza, possiamo usare la definizione di valore atteso, oppure la definizione di binomiale e le proprietà della speranza. Seguiamo questa seconda strada. Abbiamo che $X = \sum_{i=1}^{n} Y_i$, con le Y_i indipendenti e identicamente distribuite, $Y_i \sim \text{bin}(1, p)$. Allora, per linearità del valore atteso,

$$E[X] = E\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} E[Y_i] = \sum_{i=1}^{n} p = np.$$

Osserviamo che questo giustifica quanto avevamo detto euristicamente nell'Esempio 8.33, introducendo le variabili aleatorie di Poisson come limite di binomiali.

Poissoniane Consideriamo $X \sim \text{Pois}(\lambda)$, allora la sua densità discreta è, come abbiamo visto,

$$\varphi_X(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Per ricavare la speranza di X, usiamo la definizione di valore atteso,

$$\begin{aligned} E[X] &= \sum_{k \in \mathcal{R}_X} k \cdot \varphi_X(k) = \sum_{k=0}^{+\infty} k \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k=1}^{+\infty} k \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=0}^{+\infty} \lambda \frac{\lambda^{k-1}}{(k-1)!} e^{-\lambda} = \lambda \sum_{h=0}^{+\infty} \frac{\lambda^h}{h!} e^{-\lambda} = \lambda \sum_{h=0}^{+\infty} \varphi_X(h) = \lambda, \end{aligned}$$

in cui abbiamo usato la proprietà delle densità discrete per cui la somma sul supporto è 1.

Ipergeometriche Sia $X \sim \text{hyp}(k, m, n)$. Ricordiamo cosa rappresenta X: è il numero di biglie bianche tra le k estratte da un'urna che ne contiene m bianche e n nere. La densità discreta è

$$\varphi_X(b) = \frac{\binom{m}{b}\binom{n}{k-b}}{\binom{n+m}{k}},$$

per $b \in \{\max\{0, k-n\}, \dots, \min\{k, m\}\}$. Potremmo usare la definizione di valore atteso per calcolare la speranza di X, ma proviamo a sfruttare la definizione di X e le proprietà della speranza.

Per fare questo, chiamiamo $(Y_i)_{i=1}^k$ le variabili indicatrici, per ciascuna estrazione, del fatto che la pallina sia bianca oppure no:

$$Y_i = \begin{cases} 1 & \text{se la } i\text{-sima pallina è bianca} \\ 0 & \text{se la } i\text{-sima pallina è nera.} \end{cases}$$

A questo punto possiamo vedere X come la somma di queste indicatrici: $X = \sum_{i=1}^k Y_i$. Le Y_i non sono tra loro indipendenti, ma questo non ci crea problemi, perché puntiamo a usare la proprietà di linearità della speranza, che non richiede indipendenza tra le variabili aleatorie che sommiamo. Abbiamo però bisogno di sapere la densità discreta delle Y_i.

Per $i = 1$, abbiamo $P(Y_1 = 1) = \frac{m}{m+n}$. Per $i \in \{2, \dots, k\}$, quanto vale $P(Y_i = 1)$? Se sapessimo che biglie abbiamo estratto in precedenza, potremmo "aggiornare" la composizione dell'urna, ma questa sarebbe la probabilità di $Y_i = 1$ condizionata ai valori delle indicatrici Y_j con $1 \leqslant j < i$. A noi, però interessa la probabilità $P(Y_i = 1)$, senza avere altre informazioni: essa è la stessa per ogni i,

$$\varphi_{Y_i}(x) = \begin{cases} \frac{m}{m+n} & x = 1 \\ \frac{n}{m+n} & x = 0 \end{cases}$$

e 0 altrimenti. In altre parole le Y_i sono identicamente distribuite, sono tutte Bernoulliane di parametro $\frac{m}{m+n}$.

A questo punto possiamo usare la linearità della speranza:

$$E[X] = E\left[\sum_{i=1}^k Y_i\right] = \sum_{i=1}^k E[Y_i] = \sum_{i=1}^k \frac{m}{m+n} = \frac{km}{m+n}.$$

Geometriche Consideriamo ora $X \sim \text{geom}(p)$. La densità discreta è $\varphi_X(k) = p\,(1-p)^k$, ma non avremo bisogno di usarla esplicitamente. Per la speranza abbiamo infatti

$$\begin{aligned} E[X] &= \sum_{k \in \mathcal{R}_X} k\,\varphi_X(k) = \sum_{k=0}^{+\infty} k P(X = k) \\ &= \sum_{k=1}^{+\infty} \sum_{i=0}^{k-1} P(X = k) = \sum_{i=0}^{+\infty} \sum_{k=i+1}^{+\infty} P(X = k) \\ &= \sum_{i=0}^{+\infty} P(X > i) = \sum_{i=0}^{+\infty} (1-p)^{i+1} \\ &= (1-p)\,\frac{1}{1-(1-p)} = \frac{1-p}{p}, \end{aligned}$$

[Inserire disegno triangolo con dominio di somma] in cui abbiamo usato la (8.1) e la somma di una serie geometrica di ragione $1-p$.

Binomiali negative Per calcolare il valore atteso di una variabile aleatoria binomiale, ne sfruttiamo la caratterizzazione come somma di variabili aleatorie geometriche, quindi se $X \sim \text{NB}(n, p)$, e $Y_i \sim \text{geom}(p)$ per $i = 1, \dots, n$ allora la sua speranza è

$$E[X] = E\left[\sum_{i=1}^n Y_i\right] = \frac{n\,(1-p)}{p}.$$

Osservazione 9.11. Attenzione che, a seconda della definizione data di geometrica e binomiale negative, cambia il valore del valore atteso. In particolare la binomiale negativa può essere scritta anche come numero di successi dato un numero massimo di fallimenti, nel qual caso $1-p$ e p si scambiano i ruoli, motivo per cui in alcune fonti la media è $\frac{np}{1-p}$.

9.2. VARIABILI ALEATORIE ASSOLUTAMENTE CONTINUE

In analogia a quanto fatto nella sezione precedente per le variabili aleatorie discrete, definiamo ora il valore atteso per le variabili aleatorie assolutamente continue. Non possiamo farlo nello stesso modo, dal momento che la densità discreta (o in generale la probabilità di un singolo punto) non è definita. Possiamo però ricordare che la probabilità che una variabile aleatoria assolutamente continua X abbia valori in un intervallo $[a,b]$ è uguale all'integrale della densità:

$$P(X \in [a,b]) = \int_a^b f_X(x)\,\mathrm{d}x.$$

Questo ci dà una giustificazione euristica per la prossima definizione.

DEFINIZIONE 9.12. *Chiamiamo* valore atteso, speranza matematica *o* media *di una variabile aleatoria assolutamente continua X la quantità*

$$E[X] = \int_{-\infty}^{+\infty} x f_X(x)\,\mathrm{d}x.$$

Anche nel caso assolutamente continuo, come già nel caso discreto, la speranza può non essere definita (se l'integrale non esiste), oppure valere $\pm\infty$.

Valgono nel caso di variabili aleatorie assolutamente continue, risultati analoghi ai Teoremi 9.5 e 9.6, che ci permettono di calcolare la speranza della trasformazione di una variabile aleatoria o di una funzione di un vettore aleatorio.

TEOREMA 9.13. *Siano X una variabile aleatoria assolutamente continua di densità f_X e $Y = g(X)$ una sua trasformazione. Allora*

$$E[Y] = \int_{\mathbb{R}} g(x)\, f_X(x)\,\mathrm{d}x.$$

Dimostrazione. [TBA] Del tutto analoga a quella vista per il caso discreto. □

TEOREMA 9.14. *Siano (X,Y) un vettore aleatorio assolutamente continuo di legge $f_{X,Y}$ e $Z = g(X,Y)$ per qualche funzione $g\colon \mathbb{R}^2 \to \mathbb{R}$. Allora*

$$E[Z] = \iint_{\mathbb{R}^2} g(x,y)\, f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Dimostrazione. [TBA] □

Possiamo estendere i Teoremi 9.6 e 9.14 al caso di vettori aleatori misti, visti nella Sezione 7.3.

TEOREMA 9.15. *Siano X una variabile aleatoria discreta e Y una variabile aleatoria assolutamente continua e che il vettore (X,Y) abbia densità mista*[9.4] *$f_{X,Y}$. Sia inoltre $Z = g(X,Y)$, per qualche funzione $g\colon \mathbb{R}^2 \to \mathbb{R}$. Allora*

$$E[Z] = \sum_{x \in \mathcal{R}_X} \int_{\mathbb{R}} g(x,y)\, f_{X,Y}(x,y)\,\mathrm{d}y.$$

Dimostrazione. [TBA] □

Osservazione 9.16. Visto che stiamo parlando di vettori aleatori e di speranza, possiamo chiederci cosa sia la speranza di un vettore aleatorio. Chiamiamo $V = (X,Y)$ il vettore aleatorio. Abbiamo detto che la speranza è il baricentro di una distribuzione e V è un 2-vettore a valori nel piano $\mathbb{R}^2$. Il suo baricentro sarà anch'esso un punto del piano e, in particolare, sarà quindi una coppia ordinata di numeri reali. Mostriamo ora che, come ci potevamo aspettare, è proprio il vettore le cui componenti sono le speranze di X e Y rispettivamente.

9.4. Abbiamo parlato della densità congiunta mista nella Sezione 7.3.

Per dimostrare quindi che $E[V] = (E[X], E[Y])$, usiamo (supponendo che sia X sia Y siano assolutamente continue) il Teorema 9.14: se per $i = 1, 2$ chiamiamo $g_i \colon \mathbb{R}^2 \to \mathbb{R}$ la proiezione sulla i-sima componente, allora

$$E[g_i(V)] = \iint_{\mathbb{R}^2} g_i(x,y)\, f_{X,Y}(x,y)\, \mathrm{d}x\, \mathrm{d}y$$

e siccome le proiezioni di V sulle due componenti sono proprio X e Y, abbiamo il risultato.

Il valore atteso per le variabili aleatorie assolutamente continue gode delle stesse proprietà viste nella Proposizione 9.8 per la speranza di variabili aleatorie discrete. Possiamo quindi enunciare il seguente risultato più generale.

PROPOSIZIONE 9.17. *Il valore atteso gode delle seguenti proprietà.*

Linearità. *Date due variabili aleatorie X e Y e due numeri reali a e b,*

$$E[aX + Y + b] = a\,E[X] + E[Y] + b.$$

Prodotto di variabili aleatorie indipendenti. *Siano X e Y due variabili aleatorie tra loro indipendenti, allora*

$$E[X \cdot Y] = E[X] \cdot E[Y].$$

Monotonia. *Sia X una variabile aleatoria. Se $X \geqslant 0$, allora $E[X] \geqslant 0$. Inoltre l'uguaglianza vale solamente se $X \equiv 0$.*

Dimostrazione. [TBA] Le idee sono le stesse del caso discreto. □

Esempio 9.18. Sia X una variabile aleatoria di densità $f_X(x) = e^{-x}$ per $x > 0$ (e nulla altrimenti). Quanto vale la speranza di X? E quella di $X^{1/2}$?

Per quanto riguarda $E[X]$, usiamo la definizione:

$$E[X] = \int_0^{+\infty} x e^{-x}\, \mathrm{d}x = -[x e^{-x}]_0^{+\infty} + \int_0^{+\infty} e^{-x}\, \mathrm{d}x = [-e^{-x}]_0^{+\infty} = 1.$$

Passiamo ora a $E[X^{1/2}]$ e usiamo il Teorema 9.13, con $g(t) = \sqrt{t}$,

$$\begin{aligned} E[X^{1/2}] &= \int_0^{+\infty} \sqrt{x}\, e^{-x}\, \mathrm{d}x \\ &= \int_0^{+\infty} \frac{\xi}{\sqrt{2}} e^{-\xi^2/2} \xi\, \mathrm{d}\xi = \frac{1}{\sqrt{2}} \int_0^{+\infty} \xi^2 e^{-\xi^2/2}\, \mathrm{d}\xi \\ &= \left[\frac{\xi}{\sqrt{2}} e^{-\frac{\xi^2}{2}} \right]_0^{+\infty} + \frac{1}{\sqrt{2}} \int_0^{+\infty} e^{-\frac{\xi^2}{2}}\, \mathrm{d}\xi = 0 + \frac{\sqrt{\pi}}{2}, \end{aligned}$$

in cui abbiamo fatto un cambio di variabili nell'integrale, $x = \xi^2/2$, da cui $\mathrm{d}x = \xi\, \mathrm{d}\xi$ e abbiamo integrato per parti. L'integrale $\int_0^{+\infty} e^{-x^2/2}\, \mathrm{d}x$ è particolarmente importante in probabilità, come vedremo più avanti. Qualche informazione in più su come calcolarlo è in Appendice A.3.

9.3. MOMENTI DI UNA VARIABILE ALEATORIA

Possiamo considerare altri indicatori di una variabile aleatoria, oltre alla sua media. Una immediata generalizzazione del valore atteso è data dai momenti. Anche in questo caso, come già per il valore atteso, è possibile che non tutti i momenti siano definiti o che siano finiti.

DEFINIZIONE 9.19. *Per ogni $n \in \mathbb{N} \setminus \{0\}$, chiamiamo* momento n-simo *di una variabile aleatoria X il numero reale $E[X^n]$.*

Chiamiamo inoltre momento centrato n-simo *di X il numero reale $E[(X - E[X])^n]$.*

Osservazione 9.20. Con questa definizione il valore atteso è il momento primo di una variabile aleatoria e inoltre il momento centrato primo è nullo per ogni variabile aleatoria: per linearità

$$E[X - E[X]] = E[X] - E[X] = 0.$$

DEFINIZIONE 9.21. *Data una variabile aleatoria X, il suo momento centrato secondo prende anche il nome di* varianza *e viene denotato con*

$$\mathrm{Var}[X] = E[(X - E[X])^2].$$

Possiamo dare un'interpretazione fisica di media e varianza: la prima rappresenta il *baricentro* di una distribuzione di probabilità, mentre la seconda ne è il *momento di inerzia*.

PROPOSIZIONE 9.22. *Per la varianza vale la seguente uguaglianza:* $\mathrm{Var}[X] = E[X^2] - (E[X])^2$.

Dimostrazione. Scriviamo, per comodità, $E[X] = \mu$. Allora

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

in cui abbiamo usato la linearità della speranza. □

Se vogliamo calcolare la varianza di una variabile aleatoria X, grazie alla Proposizione 9.22 possiamo farlo calcolando la media $E[X]$ di X e il valore atteso della variabile aleatoria X^2, usando il Teorema 9.13 (o il Teorema 9.5, se X è discreta).

Possiamo ora mostrare alcune proprietà della varianza.

PROPOSIZIONE 9.23. *Siamo X una variabile aleatoria e* $\mathrm{Var}[X]$ *la sua varianza. Allora*

i. $\mathrm{Var}[X] \geqslant 0$ *e l'uguaglianza vale solamente se X è costante;*

ii. siano $a, b \in \mathbb{R}$, $\mathrm{Var}[aX + b] = a^2 \mathrm{Var}[X]$.

Dimostrazione. La prima proprietà è una conseguenza immediata della monotonia della speranza, infatti $(X - E[X])^2 \geqslant 0$, dunque

$$\mathrm{Var}[X] = E[(X - E[X])^2] \geqslant 0.$$

Inoltre, sempre per la monotonia, $E[(X - E[X])^2] = 0$ se e solo se l'argomento della speranza è nullo, ossia se $X = E[X]$, cioè se la variabile aleatoria X è costante, visto che $E[X]$ è un numero.

Passiamo alla seconda proprietà,

$$\begin{aligned} \mathrm{Var}[aX + b] &= E[(aX + b - E[aX + b])^2] \\ &= E[(aX + b - aE[X] - b)^2] \\ &= a^2 E[(X - E[X])^2] = a^2 \mathrm{Var}[X], \end{aligned}$$

usando la linearità della speranza. □

Osservazione 9.24. La proposizione precedente ci mostra in particolare che la varianza di una variabile aleatoria *non* è lineare: nel calcolare $\mathrm{Var}[aX + b]$ il termine noto non gioca alcun ruolo, mentre il coefficiente a esce dall'operatore Var al quadrato.

La seconda proprietà nella Proposizione 9.23 non copre il caso della varianza della combinazione lineare di due variabili aleatorie, a differenza di quanto visto per la speranza. Per poter trattare in generale la varianza della somma di due variabili aleatorie dobbiamo aspettare ancora un po', fino alla Sezione 9.5. Nel prossimo risultato, però, affrontiamo almeno una situazione particolare.

PROPOSIZIONE 9.25. *Se X e Y sono due variabili aleatorie indipendenti per cui sia definita la varianza, allora* $\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$.

Dimostrazione. Iniziamo sfruttando la Proposizione 9.22,

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

in cui l'ultimo addendo si annulla perché per l'indipendenza di X e Y, $E[XY] = E[X]E[Y]$. □

Avendone viste alcune proprietà, vogliamo ora provare a dare un'interpretazione intuitiva di cosa sia la varianza di una variabile aleatoria. Consideriamo come esempio la variabile aleatoria X, Bernoulliana di parametro $\frac{1}{2}$, ossia il lancio di una moneta bilanciata. Questa variabile aleatoria ha media $\frac{1}{2}$, proviamo a calcolarne la varianza: Lezione 17

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = 1^2 \cdot \frac{1}{2} + 0^2 \cdot \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4},$$

che però non sembra comparire in modo evidente nella descrizione della variabile aleatoria. Se riguardiamo la definizione di varianza come momento centrato secondo, $\mathrm{Var}[X] = E[(X - E[X])^2]$, vediamo che la varianza è la media del quadrato della distanza tra la variabile aleatoria e la sua media. Possiamo quindi aspettarci che in qualche senso misuri la "larghezza" della variabile aleatoria.

Per valutare meglio questa idea, modifichiamo la variabile X e consideriamone la seguente trasformazione: $Y = 2X$. Allora la media di Y sarà $E[Y] = 1$ e la varianza $\mathrm{Var}[Y] = 4\,\mathrm{Var}[X] = 1$, cioè al raddoppiare della "larghezza" della variabile aleatoria, la varianza è quadruplicata.

Abbiamo quindi una conferma euristica del fatto che la varianza misuri la dispersione di una variabile aleatoria, ossia quanto sono distanti "in media" i valori della variabile aleatoria dalla media della variabile aleatoria stessa. Allo stesso tempo, questa misura non è proprio la "larghezza", dal momento che varia quadraticamente.

DEFINIZIONE 9.26. *Chiamiamo* deviazione standard *di una variabile aleatoria X la radice quadrata della sua varianza, $\sigma_X = \sqrt{\mathrm{Var}[X]}$. Possiamo quindi indicare* $\mathrm{Var}[X]$ *con* σ_X^2.

In questo modo abbiamo un indicatore (ossia un numero) che misura proprio la media della distanza di una variabile aleatoria X dalla sua media. Essendo la radice quadrata della varianza, essa ha la stessa unità di misura di X e della sua media $E[X]$: se $Y = aX$, allora

$$\sigma_Y = \sqrt{\mathrm{Var}[Y]} = \sqrt{a^2 \mathrm{Var}[X]} = a\,\sigma_X.$$

Esempio 9.27. Consideriamo la variabile aleatoria X uniforme sull'intervallo $[0,1]$. Essa ha media

$$E[X] = \int_0^1 x\,1\,\mathrm{d}x = \frac{1}{2}$$

e varianza

$$\begin{aligned} \mathrm{Var}[X] &= E[X^2] - E[X]^2 \\ &= \int_0^1 x^2 \mathrm{d}x - \frac{1}{4} \\ &= \frac{1}{3} - \frac{1}{4} = \frac{1}{12} \end{aligned}$$

e la sua deviazione standard è $\sigma_X = \frac{1}{\sqrt{12}} \approx 0.29$.

9.3.1. Varianza di alcune variabili aleatorie note

Calcoliamo ora la varianza (e quindi la deviazione standard) di alcuni[9.5] modelli di variabili aleatorie discrete che abbiamo definito nel Capitolo 8.

Bernoulliane Sia $X \sim \mathrm{bin}(1,p)$. Allora

$$\begin{aligned} \mathrm{Var}[X] &= E[X^2] - E[X]^2 \\ &= p - p^2 = p\,(1-p). \end{aligned}$$

9.5. Non ricaviamo qui la varianza delle ipergeometriche, perché per farlo abbiamo bisogno della covarianza, che introdurremo solamente nella Sezione 9.5.

Binomiali Sia $X \sim \text{bin}(n,p)$. La variabile aleatoria X è la somma di n Bernoulliane indipendenti e identicamente distribuite Y_i di legge $\text{bin}(1,p)$, quindi

$$\text{Var}[X] = \text{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \text{Var}[Y_i] = np\,(1-p),$$

in cui abbiamo usato la Proposizione 9.25 e la forma della varianza per le Bernoulliane.

Geometriche Sia ora $X \sim \text{geom}(p)$. Se volessimo ripetere quanto fatto per il valore atteso, avremmo una difficoltà: non possiamo liberarci altrettanto facilmente del termine k^2 in $E[X^2]$ attraverso una somma ausiliaria. Vediamo quindi una strategia alternativa, con cui calcoliamo sia la media, sia la varianza.

Per definizione X è la variabile aleatoria che conta il numero di insuccessi prima del primo successo in un processo di Bernoulli. Sia invece Y la variabile aleatoria che conta il numero di insuccessi prima del primo successo escludendo il risultato del primo tentativo. In altre parole

$$Y = \inf\{n \geqslant 2 : \omega_n = 1\} - 2$$

(troviamo l'istante di primo successo successivo a 1 e togliamo 1 per trascurare il primo tentativo e 1 perché vogliamo contare il numero di insuccessi). Ne andiamo a scrivere la densità discreta,

$$\varphi_Y(k) = P(Y=k) = 1 \cdot (1-p)^k \cdot p,$$

del momento che trascuriamo il primo lancio, abbiamo k insuccessi e infine un successo. Ma questa è la densità discreta di una geometrica di parametro p, $Y \sim X \sim \text{geom}(p)$.

Ora possiamo scrivere la speranza di X come

$$\begin{aligned}
E[X] &= \sum_{k=0}^{+\infty} k\,P(X=k) \\
&= \sum_{k=0}^{+\infty} \left(k\,P(X=k \mid \omega_1=0)\,P(\omega_1=0) + k\,P(X=k \mid \omega_1=1)\,P(\omega_1=1)\right) \\
&= \sum_{k=0}^{+\infty} k\,P(X=k \mid \omega_1=0)\,(1-p) + \sum_{k=0}^{+\infty} k\,P(X=k \mid \omega_1=1)\,p \\
&= E[X \mid \omega_1=0]\,(1-p) + E[X \mid \omega_1=1]\,p \\
&= \sum_{k=0}^{+\infty} k\,P(Y+1=k \mid \omega_1=0)\,(1-p) + 0 \cdot P(X=0 \mid \omega_1=1)\,p \\
&= E[Y+1 \mid \omega_1=0]\,(1-p) \\
&= E[Y+1]\,(1-p) \\
&= E[Y]\,(1-p) + 1 - p,
\end{aligned}$$

da cui, siccome $E[X] = E[Y]$, ricaviamo (supponendo che $E[X]$ sia finita) $E[X] = \frac{1-p}{p}$. Nella catena di uguaglianze abbiamo usato che $P(X=0 \mid \omega_1=1) = 1$, che se $\omega_1 = 0$ allora $X = Y+1$ e che Y è indipendente dall'evento $\omega_1 = 0$.

Allo stesso modo abbiamo, per $E[X^2]$,

$$\begin{aligned}
E[X^2] &= E[X^2 \mid \omega_1=0]\,(1-p) + E[X^2 \mid \omega_1=1]\,p \\
&= E[(Y+1)^2 \mid \omega_1=0]\,(1-p) + 0 \cdot p \\
&= (E[Y^2] + 2E[Y] + 1)\,(1-p) \\
&= E[Y^2]\,(1-p) + \frac{2(1-p)^2}{p} + (1-p),
\end{aligned}$$

da cui (assumendo che $E[X^2] < +\infty$), $E[X^2] = \frac{2\,(1-p)^2 + p\,(1-p)}{p^2}$. Allora, per la varianza,

$$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{2\,(1-p)^2 + p\,(1-p)}{p^2} - \frac{(1-p)^2}{p^2} = \frac{1-p}{p^2}.$$

Binomiali negative Una variabile aleatoria $X \sim \mathrm{NB}(n,p)$ è la somma di n variabili aleatorie geometriche indipendenti e identicamente distribuite $Y_i \sim \mathrm{geom}(p)$. Allora

$$\mathrm{Var}[X] = \frac{n(1-p)}{p^2}.$$

Poissoniane Sia ora $X \sim \mathrm{Pois}(\lambda)$. Sappiamo che $E[X] = \lambda$, quindi per ricavare $\mathrm{Var}[X]$ ci basta calcolare il momento secondo $E[X^2]$. Possiamo provare a farlo usando il Teorema 9.5, ma

$$E[X^2] = \sum_{k=0}^{+\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda}$$

non sembra semplicissima da trattare. Cerchiamo allora di arrivare al risultato usando un trucco:

$$\begin{aligned} E[X^2 - X] &= E[X(X-1)] = \sum_{k=0}^{+\infty} k(k-1) \frac{\lambda^k}{k!} e^{-\lambda} \\ &= \sum_{k=2}^{+\infty} k(k-1) \frac{\lambda^2 \cdot \lambda^{k-2}}{k(k-1)(k-2)!} \\ &= \lambda^2 \sum_{j=0}^{+\infty} \frac{\lambda^j}{j!} e^{-\lambda} = \lambda^2 \end{aligned}$$

dove abbiamo usato il fatto che i primi due addendi nella prima somma sono nulli e abbiamo messo in evidenza, con il cambio di variabile $j = k-2$, la somma delle densità discrete sul supporto di una Poissoniana, somma che sappiamo essere uguale a 1. A questo punto, per linearità,

$$E[X^2] = E[X^2 - X] + E[X] = \lambda^2 + \lambda$$

e per la varianza abbiamo

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

9.4. DISUGUAGLIANZE

Nell'introduzione di questo capitolo, abbiamo detto che vogliamo usare gli indicatori per avere alcune informazioni su una distribuzione di probabilità ignota. Vediamo ora alcuni risultati che ci permettono di dire qualcosa su una variabile aleatoria e la sua distribuzione a partire dalla sua speranza e dalla sua varianza.

PROPOSIZIONE 9.28. (DISUGUAGLIANZA DI MARKOV[9.6]) *Sia X una variabile aleatoria non negativa di media finita. Allora, per ogni $a > 0$*

$$P(X \geqslant a) \leqslant \frac{E[X]}{a}. \tag{9.1}$$

Dimostrazione. Se $P(X \geqslant a) = 0$, la tesi segue dal fatto che $E[X] \geqslant 0$ e che quindi il secondo membro è sicuramente non negativo. Supponiamo allora che $P(X \geqslant a) > 0$: abbiamo

$$\begin{aligned} E[X] &= E[X \mid X < a] P(X < a) + E[X \mid X \geqslant a] P(X \geqslant a) \\ &\geqslant E[X \mid X \geqslant a] P(X \geqslant a) \\ &\geqslant a P(X \geqslant a) \end{aligned}$$

dal momento che stiamo facendo la media per valori che sono almeno a. □

Osservazione 9.29. È possibile dare una dimostrazione alternativa più diretta di questo fatto, usando la definizione di speranza nei casi discreto e assolutamente continuo, assieme alle proprietà di somma e integrale. È però un processo più laborioso.

9.6. Andrej Andreevič Markov (1856 – 1922). In realtà pare che questo risultato sia dovuto a Pafnutij L'vovič Čebyšëv (1821 – 1894), spesso traslitterato come Chebychev, di cui Markov fu allievo.

Osservazione 9.30. Esistono altre varianti della disuguaglianza di Markov (9.1), anche più "forti", che spesso vengono chiamate con lo stesso nome. Ad esempio possiamo lasciar cadere l'ipotesi che X abbia media finita, nel qual caso la disuguaglianza è banalmente vera, senza essere però molto utile.

PROPOSIZIONE 9.31. (DISUGUAGLIANZA DI CHEBYCHEV) *Sia X una variabile aleatoria con varianza finita. Allora, per ogni $a>0$*

$$P(|X-E[X]|\geqslant a)\leqslant\frac{\mathrm{Var}[X]}{a^2} \tag{9.2}$$

o equivalentemente

$$P(|X-E[X]|\geqslant a\cdot\sqrt{\mathrm{Var}[X]})\leqslant\frac{1}{a^2}. \tag{9.3}$$

Dimostrazione. Osserviamo innanzitutto che sia a sia $|X-E[X]|$ sono non negativi. Allora

$$\begin{aligned} P(|X-E[X]|\geqslant a) &= P((X-E[X])^2\geqslant a^2)\\ &\leqslant \frac{E[(X-E[X])^2]}{a^2}\\ &= \frac{\mathrm{Var}[X]}{a^2}, \end{aligned}$$

in cui abbiamo usato la disuguaglianza di Markov (9.1), sfruttando il fatto che la variabile aleatoria $(X-E[X])^2$ è non negativa. La forma equivalente (9.3) segue dalla (9.2) sostituendo ad a il numero reale positivo[9.7] $a\sqrt{\mathrm{Var}[X]}$ □

Grazie alla disuguaglianza di Chebychev (9.3) possiamo formalizzare quanto detto prima sul significato della deviazione standard: $\sqrt{\mathrm{Var}[X]}$ misura quanto X sia larga o dispersa, infatti possiamo usarla per valutare la probabilità che X si allontani dalla propria media.

Esempio[TBA]

9.5. COVARIANZA E CORRELAZIONE

Gli indicatori che abbiamo visto finora riguardano una singola variabile aleatoria. In certe situazioni, tuttavia, ci farebbe comodo avere un indicatore che misuri quanto due variabili aleatorie sono legate tra loro. Infatti sappiamo determinare se sono indipendenti o meno, ma non sappiamo modulare questo secondo caso.

DEFINIZIONE 9.32. *Date due variabili aleatorie X e Y, chiamiamo* covarianza *di X e Y la quantità*

$$\mathrm{Cov}[X,Y]=E[(X-E[X])\,(Y-E[Y])]. \tag{9.4}$$

La covarianza generalizza la varianza: $\mathrm{Cov}[X,X]=\mathrm{Var}[X]$. Anche per la covarianza, come già visto per la varianza, abbiamo una seconda formulazione equivalente, spesso più pratica da usare

$$\mathrm{Cov}[X,Y]=E[XY]-E[X]\,E[Y].$$

PROPOSIZIONE 9.33. *Vediamo alcune proprietà della covarianza.*

i. La covarianza è simmetrica: per ogni coppia di variabili aleatorie X e Y, $\mathrm{Cov}[X,Y]=\mathrm{Cov}[Y,X]$.

ii. Se X e Y sono variabili aleatorie indipendenti, allora $\mathrm{Cov}[X,Y]=0$.

Dimostrazione. [TBA] □

DEFINIZIONE 9.34. *Se due variabili aleatorie X e Y hanno covarianza nulla (cioè* $\mathrm{Cov}[X,Y]=0$*), diciamo che sono* scorrelate.

9.7. In realtà c'è il caso in cui $\mathrm{Var}[X]=0$, ma allora $X=E[X]$ e la disuguaglianza non è molto interessante.

Osservazione 9.35. È importante ricordare che non vale il viceversa della seconda proprietà nella Proposizione 9.33: non è necessariamente vero che due variabili aleatorie scorrelate siano indipendenti, anche se due variabili indipendenti sono anche scorrelate.

Esempio 9.36. Siano X e Y due variabili aleatorie di legge congiunta

$$\varphi_{X,Y}(x,y)=\begin{cases}\frac{1}{4} & (x,y)\in\{(-1,-1),(1,-1)\}\\ \frac{1}{2} & (x,y)=(0,1)\\ 0 & \text{altrimenti.}\end{cases}$$

Possiamo ricavarci le leggi marginali di X e Y:

$$\varphi_X(x)=\begin{cases}\frac{1}{4} & x\in\{-1,1\}\\ \frac{1}{2} & x=0\\ 0 & \text{altrimenti}\end{cases}\qquad\qquad \varphi_Y(y)=\begin{cases}\frac{1}{2} & y\in\{-1,1\}\\ 0 & \text{altrimenti.}\end{cases}$$

A questo punto possiamo facilmente verificare che X e Y non sono indipendenti, infatti

$$\varphi_{X,Y}(x,y)\neq\varphi_X(x)\cdot\varphi_Y(y)=\begin{cases}\frac{1}{8} & (x,y)\in\{(-1,-1),(-1,1),(1,-1),(1,1)\}\\ \frac{1}{4} & (x,y)\in\{(0,-1),(0,1)\}\\ 0 & \text{altrimenti.}\end{cases}$$

Allo stesso tempo, le due variabili aleatorie sono scorrelate, infatti

$$\mathrm{Cov}[X,Y]=E[XY]-E[X]\,E[Y]=1\cdot\frac{1}{4}+0\cdot\frac{1}{2}+(-1)\cdot\frac{1}{4}-0=0.$$

Il precedente esempio ci suggerisce un'osservazione più generale sul calcolo della covarianza: a partire dalla (9.4), possiamo usare i teoremi noti sulla speranza della funzione di un vettore aleatorio, nel caso particolare in cui $g(x,y)=x\cdot y$,

$$\begin{aligned}\mathrm{Cov}[X,Y] &= E[XY]-E[X]\,E[Y]\\ &= E[g(X,Y)]-E[X]\,E[Y]\\ &= \iint_{\mathbb{R}^2} xy f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y-\int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x\int_{\mathbb{R}} y f_Y(y)\,\mathrm{d}y\end{aligned}$$

nel caso X e Y siano assolutamente continue, oppure

$$\mathrm{Cov}[X,Y]=\sum_{x\in\mathcal{R}_X}\sum_{y\in\mathcal{R}_Y} xy\,\varphi_{X,Y}(x,y)-\sum_{x\in\mathcal{R}_X} x\,\varphi_X(x)\sum_{y\in\mathcal{R}_Y} y\,\varphi_Y(y)$$

nel caso siano entrambe discrete o ancora

$$\mathrm{Cov}[X,Y]=\sum_{x\in\mathcal{R}_X}\int_{\mathbb{R}} xy f_{X,Y}(x,y)\,\mathrm{d}y-\sum_{x\in\mathcal{R}_X} x\,\varphi_X(x)\int_{\mathbb{R}} y f_Y(y)\,\mathrm{d}y$$

nel caso X sia discreta e Y sia assolutamente continua.

PROPOSIZIONE 9.37. *La covarianza di due variabili aleatorie X e Y soddisfa anche le seguenti proprietà.*

i. Ci permette di calcolare la varianza della loro somma anche nel caso in cui X e Y non siano indipendenti: $\mathrm{Var}[X+Y]=\mathrm{Var}[X]+\mathrm{Var}[Y]+2\mathrm{Cov}[X,Y]$.

ii. La covarianza è lineare separatamente in ciascun argomento:

$$\mathrm{Cov}[aX+bY,Z]=a\,\mathrm{Cov}[X,Z]+b\,\mathrm{Cov}[Y,Z].$$

iii. La covarianza è bilineare: se $(a_i)_{i=1}^n$ e $(b_j)_{j=1}^m$ sono due vettori di numeri reali e $(X_i)_{i=1}^n$ e $(Y_j)_{j=1}^m$ sono due vettori aleatori, allora

$$\mathrm{Cov}\left[\sum_{i=1}^n a_i X_i,\sum_{j=1}^m b_j Y_j\right]=\sum_{i=1}^n\sum_{j=1}^m a_i b_j\,\mathrm{Cov}[X_i,Y_j].$$

Dimostrazione. [TBA] □

Osservazione 9.38. La Proposizione 9.37 generalizza le proprietà viste per la varianza. Infatti da un lato ci permette di calcolare la varianza della somma di due variabili aleatorie qualunque, dall'altro possiamo anche passare alla varianza di una qualsiasi combinazione lineare di variabili aleatorie: se $(a_i)_{i=1}^n$ è un vettore di numeri reali e $(X_i)_{i=1}^n$ è un vettore aleatorio, allora

$$\mathrm{Var}\left[\sum_{i=1}^n a_i X_i\right] = \mathrm{Cov}\left[\sum_{i=1}^n a_i X_i, \sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}[X_i, X_j]. \tag{9.5}$$

DEFINIZIONE 9.39. *Se $(X_i)_{i=1}^n$ è un vettore aleatorio, chiamiamo* matrice di covarianza *la matrice $n \times n$ le cui componenti sono $\mathrm{Cov}[X_i, X_j]$. Questa matrice è spesso indicata con $\Sigma(X,Y)$ o, in breve, con Σ.*

Possiamo allora riscrivere la (9.5) in maniera più compatta come

$$\begin{aligned}\mathrm{Var}\left[\sum_{i=1}^n a_i X_i\right] &= \mathrm{Var}[\vec{a}\cdot\vec{X}] = \mathrm{Var}[\langle \vec{a}, \vec{X}\rangle] \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}[X_i, X_j] \\ &= \vec{a}\cdot\Sigma\vec{a} = \vec{a}^t\Sigma\vec{a}\end{aligned}$$

mettendo in evidenza che si tratta di prodotti interni di vettori (prodotto scalare o prodotto componente per componente). La notazione vettoriale o matriciale è particolarmente comoda in R per evitare i cicli `for` tutte le volte che possiamo scrivere il problema in modo equivalente come operazioni su matrici: il costo computazionale si riduce notevolmente e il codice è molto più leggibile.

PROPOSIZIONE 9.40. *Valgono le seguenti disuguaglianze,*

$$-\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]} \leqslant \mathrm{Cov}[X,Y] \leqslant \sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}. \tag{9.6}$$

Dimostrazione. [TBA] □

Se a valori grandi di X corrispondono in genere valori grandi di Y e a valori piccoli della prima corrispondono valori piccoli della seconda, allora $\mathrm{Cov}[X,Y] > 0$ e diciamo che le due variabili aleatorie sono *positivamente correlate*. Se invece a valori grandi di X corrispondono in genere valori piccoli di Y e, viceversa, a valori piccoli di X corrispondono valori grandi di Y, $\mathrm{Cov}[X,Y] < 0$ e diciamo che le due variabili aleatorie sono *negativamente correlate*.

Esempio 9.41. Sono esempi di variabili aleatorie correlate il livello degli studi completati e il reddito, mentre il numero di core di un calcolatore e il tempo di calcolo sono variabili aleatorie negativamente correlate.

DEFINIZIONE 9.42. *Date due variabili aleatorie X e Y, chiamiamo* correlazione *o* coefficiente di correlazione lineare *il numero reale*

$$\rho(X,Y) = \mathrm{corr}[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}.$$

La correlazione $\rho(X,Y)$ tra due variabili aleatorie è, per la (9.6), un numero in $[-1,1]$ ed è una versione normalizzata della covarianza. In particolare $\rho \approx 1$ indica un'alta correlazione positiva tra le due variabili, $\rho \approx -1$ indica un'alta correlazione negativa e $\rho \approx 0$ indica correlazione bassa o assente.

Esempio 9.43. Avendo definito la covarianza, possiamo ora calcolare la varianza delle ipergeometriche. Sia $X \sim \mathrm{hyp}(k,m,n)$.

Come abbiamo visto nella Sotto-sezione 9.1.1 nel calcolare la speranza di una ipergeometrica, possiamo scrivere $X = \sum_{i=1}^{k} Y_i$, dove ogni Y_i è una variabile aleatoria che indica se la i-sima biglia estratta è bianca oppure no. Le Y_i non sono tra loro indipendenti, ma sono identicamente distribuite come Bernoulliane di parametro $\frac{m}{m+n}$.

Partendo dalla (9.5) possiamo scrivere

$$\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{i=1}^{k} Y_i\right] = \sum_{i=1}^{k} \mathrm{Var}[Y_i] + 2 \sum_{1 \leqslant i < j \leqslant k} \mathrm{Cov}[Y_i, Y_j].$$

Conosciamo $\mathrm{Var}[Y_i] = \frac{nm}{(n+m)^2}$ (indipendente da i), quindi dobbiamo calcolare

$$\mathrm{Cov}[Y_i, Y_j] = E[Y_i Y_j] - E[Y_i]\, E[Y_j].$$

Sappiamo già le medie $E[Y_i] = E[Y_j] = \frac{n}{m+n}$, non ci resta che ricavare $E[Y_i Y_j]$. Per farlo, osserviamo che $Y_i Y_j$ assume solamente i valori 0 o 1: sono binomiali di parametro $p = P(Y_i Y_j = 1)$ (che è anche la media) ed è quindi l'ultimo ingrediente che ci occorre,

$$\begin{aligned} E[Y_i Y_j] &= P(Y_i Y_j = 1) = P(Y_i = 1, Y_j = 1) = P(Y_i = 1)\, P(Y_j = 1 \mid Y_i = 1) \\ &= \frac{n}{n+m} \cdot \frac{n-1}{n+m-1} = \frac{n^2 - n}{(n+m)\,(n+m-1)}. \end{aligned}$$

La covarianza è quindi

$$\mathrm{Cov}[Y_i, Y_j] = \frac{n^2 - n}{(n+m)\,(n+m-1)} - \frac{n^2}{(n+m)^2} = \frac{\cancel{n^3} + n^2 m - \cancel{n^2} - nm - \cancel{n^3} - n^2 m + \cancel{n^2}}{(n+m)^2\,(n+m-1)}.$$

Ora possiamo mettere assieme il tutto,

$$\begin{aligned} \mathrm{Var}[X] &= k \cdot \frac{nm}{(n+m)^2} - k\,(k-1)\,\frac{nm}{(n+m)^2\,(n+m-1)} \\ &= \frac{knm}{(n+m)^2}\left(1 - \frac{k-1}{n+m-1}\right). \end{aligned}$$

Un'ultima osservazione: nel caso con reimmissione in un'urna di uguale composizione, la varianza sarebbe $\frac{knm}{(n+m)^2}$, quindi la "penalità" dovuta alla mancata reimmissione è un fattore $\frac{k-1}{n+m-1}$, che per un'urna molto grande (ossia per $n+m \to +\infty$) diventa trascurabile.

9.6. ALTRI INDICATORI DI UNA DISTRIBUZIONE

Lezione 18

Se torniamo agli indicatori di una sola variabile aleatoria, speranza e varianza non esauriscono le opzioni disponibili. Vediamo alcuni indicatori particolarmente importanti.

DEFINIZIONE 9.44. *Chiamiamo* mediana *di una variabile aleatoria X un numero m_X tale che*

$$P(X \leqslant m_X) = P(X \geqslant m_X). \tag{9.7}$$

Osservazione 9.45. Cerchiamo un valore m_X che sia al centro della distribuzione nel senso seguente: la probabilità che una realizzazione di X sia minore o uguale di m_X è uguale alla probabilità che sia maggiore o uguale di m_X, cioè sono entrambe uguali a $\frac{1}{2}$. Questo m_X è al centro della distribuzione, ma in un senso diverso da quello della media.

Possiamo riscrivere la caratterizzazione della mediana (9.7) in termini della funzione di ripartizione di X: m_X è tale che $F_X(m_X) = 1 - F_X(m_X)$, cioè $F_X(m_X) = \frac{1}{2}$, quindi ci verrebbe da dire che $m_X = F^{-1}\left(\frac{1}{2}\right)$, ma non sappiamo se F_X sia invertibile, quindi possiamo solamente affermare che $m_X \in F^{-1}\left(\left\{\frac{1}{2}\right\}\right)$, cioè appartiene alla preimmagine di $\frac{1}{2}$. Consideriamo ora separatamente i casi in cui X sia discreta o assolutamente continua. Iniziamo da quest'ultimo.

Se X è assolutamente continua, F_X è una funzione continua e monotona crescente tra 0 e 1, quindi l'insieme $F^{-1}\left(\left\{\frac{1}{2}\right\}\right)$ è non vuoto, ma non è detto che sia un singoletto e quindi non è detto che esista *la* mediana. Un controesempio all'unicità della mediana è rappresentato in Figura 9.1.

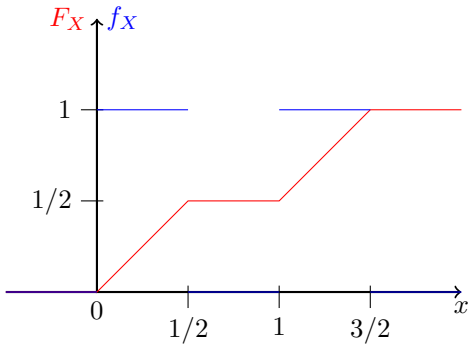


Figura 9.1. Esempio di non unicità della mediana: tutti i punti dell'intervallo $\left[\frac{1}{2}, 1\right]$ sono mediane.

Nel caso in cui X sia una variabile aleatoria discreta, le cose possono andare anche peggio: non solo la mediana può non essere unica, ma può addirittura non esistere, infatti in questo caso la funzione di ripartizione F_X non è più continua, quindi $F_X^{-1}\left(\left\{\frac{1}{2}\right\}\right)$ può essere vuoto.

Esempio 9.46. Sia $X \sim \text{bin}\left(1, \frac{1}{2}\right)$: allora per ogni $x \in (0,1)$ abbiamo $P(X \leqslant x) = P(X = 0) = \frac{1}{2}$, ma anche $P(X \geqslant x) = P(X = 1) = \frac{1}{2}$, quindi ogni x nell'intervallo aperto[9.8] $(0,1)$ è una mediana di X.

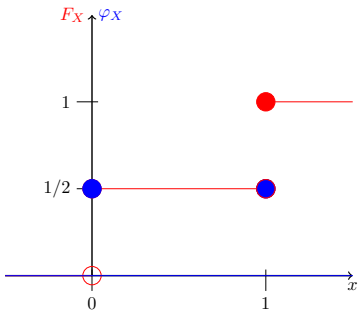


Figura 9.2. Tutti i punti in $(0,1)$ sono mediane.

Esempio 9.47. Definiamo ora la variabile aleatoria discreta X nel modo seguente:

$$X = \begin{cases} 0 \text{ con probabilità } \frac{1}{6} \\ 1 \text{ con probabilità } \frac{1}{2} \\ 2 \text{ con probabilità } \frac{1}{3}. \end{cases}$$

In questo caso, come vediamo nella Figura 9.3 $F_X^{-1}\left(\left\{\frac{1}{2}\right\}\right) = \emptyset$.

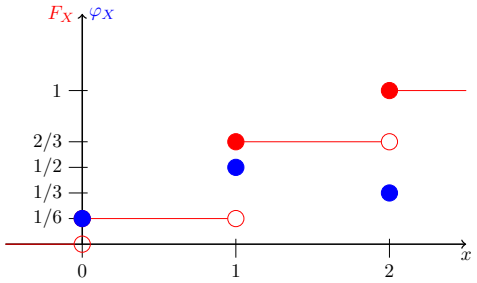


Figura 9.3. La variabile aleatoria X non ammette mediana.

9.8. Gli estremi non sono inclusi, come mai?

Per ogni $x \in (-\infty, 1)$ abbiamo

$$P(X \leqslant x) \leqslant P(X=0) = \frac{1}{6} \qquad P(X \geqslant x) \geqslant P((X=1) \cup (X=2)) = \frac{5}{6}$$

cioè tutti questi x sono troppo sbilanciati verso sinistra rispetto a quella che vorremmo come mediana[9.9]. Allo stesso tempo per ogni $x \in (1, +\infty)$

$$P(X \leqslant x) \geqslant P((X=0) \cup (X=1)) = \frac{2}{3} \qquad P(X \geqslant x) \leqslant P(X=2) = \frac{1}{3}$$

quindi questi valori di x sono "troppo a destra" per essere delle mediane. Non ci resta che sperare in $x=1$, ma

$$P(X \leqslant 1) = \frac{2}{3} \neq \frac{5}{6} = P(X \geqslant 1).$$

Questa variabile aleatoria, allora, non ammette alcuna mediana secondo la Definizione 9.44.

Dal momento che può essere utile avere un concetto di mediana definito per ogni variabile aleatoria, possiamo darne una definizione indebolita.

DEFINIZIONE 9.48. *Chiamiamo* mediana impropria *di una variabile aleatoria X un numero reale $\tilde{m}_X$ tale che $P(X \leqslant \tilde{m}_X) \geqslant \frac{1}{2}$ e $P(X \geqslant \tilde{m}_X) \geqslant \frac{1}{2}$.*

Osservazione 9.49. Con questa definizione, la mediana impropria è il valore soglia tra quelli "troppo a sinistra" e quelli "troppo a destra", anche se non soddisfa l'uguaglianza (9.7). La variabile aleatoria X nell'Esempio 9.47 ammette come mediana impropria $\tilde{m}_X = 1$.

Possiamo ora pensare di generalizzare quanto visto per la mediana, cercando i punti in cui "tagliare" una distribuzione in modo che una realizzazione della variabile aleatoria corrispondente abbia una probabilità predeterminata di essere minore o uguale al taglio. In altre parole, fissiamo $p \in [0,1]$ e cerchiamo i numeri reali x per cui $P(X \leqslant x) = F_X(x) = p$.

DEFINIZIONE 9.50. *Dati una variabile aleatoria X di legge F_X e $p \in (0,1)$, chiamiamo* quantile p *(o* p-quantile*) il numero reale $Q_X(p)$ tale che*

$$Q_X(p) = \inf\{x \in \mathbb{R} : F_X(x) \geqslant p\}. \tag{9.8}$$

Osservazione 9.51. Per $p = \frac{1}{2}$ abbiamo qualcosa di molto simile alla mediana, ma in questo caso viene sempre scelto un solo valore[9.10]. Non solo, dal momento che la funzione di ripartizione F_X è sempre continua a destra, l'inf nella (9.8) è in realtà un minimo.

Se la funzione di ripartizione F_X è continua e strettamente crescente, allora per ogni $p \in (0,1)$ abbiamo che $Q_X(p)$ è proprio quel valore che soddisfa $F_X(Q_X(p)) = p$. Più in generale se F_X è invertibile in quel punto, allora $Q_X(p) = F_X^{-1}(p)$.

DEFINIZIONE 9.52. *Chiamiamo* funzione quantile della variabile aleatoria X *la funzione*

$$Q : p \mapsto Q_X(p)$$

che associa ad ogni p il quantile corrispondente.

Abbiamo menzionato più volte prima d'ora le funzioni quantile associate alle varie distribuzioni. Possiamo usarle per calcolare quale sia il punto x che si lascia a sinistra al più probabilità p. Per esempio, se X è una variabile aleatoria di Poisson di parametro $\lambda = 2$, la funzione `qpois(p = 1/3, lambda = 2)` ci restituisce 1, infatti

9.9. Cosa intendiamo con "troppo a sinistra" o "troppo a destra"? La mediana è (almeno moralmente) il punto in cui la funzione di ripartizione F_X e il suo complemento a 1, $1 - F_X$ si bilanciano. Sappiamo però che al crescere di x la funzione $F_X(x)$ è crescente e, conseguentemente, la funzione $1 - F_X(x)$ è decrescente. Quindi se $F_X(x) < \frac{1}{2}$ siamo "troppo a sinistra", mentre se $1 - F_X(x) < \frac{1}{2}$ siamo "troppo a destra".

9.10. Quale?

$F_X(1) = P(X \leqslant 1) \approx 40\%$ e, per $x \in (0,1)$, $F_X(x) = F_X(0) = P(X \leqslant 0) \approx 14\%$.

Possiamo anche per la funzione quantile, come per la funzione di ripartizione, specificare la "coda" della distribuzione cui siamo interessati: di default è (come per la funzione di ripartizione) `lower.tail = TRUE`, ossia guardiamo la coda sinistra. Se invece ci interessa la coda destra, possiamo passare il valore `lower.tail = FALSE`. In questo caso la funzione ci restituirà il più piccolo valore di x per cui $P(X > x) = 1 - F_X(x) \leqslant p$. In pratica

```
qpois(p,lambda,FALSE)=qpois(1-p,lambda,TRUE)
```

e questo vale anche per le altre distribuzioni.

Osservazione 9.53. Per alcune scelte di p i quantili hanno nomi particolari:

- per $p = \frac{k}{4}$, con $k \in \{1,2,3\}$ parliamo di *quartili* (primo, secondo e terzo);
- per $p = \frac{k}{10}$, con $k \in \{1,\dots,9\}$ parliamo di *decili*;
- per $p = \frac{k}{100}$, con $k \in \{1,\dots,99\}$ parliamo di *percentili*.

DEFINIZIONE 9.54. *Chiamiamo* moda *di una variabile aleatoria X un numero $x \in \mathcal{R}_X$ tale che*

- *se X è discreta, φ_X è massima in x, cioè $x \in \operatorname{argmax}_y \varphi_X(y)$*
- *se X è assolutamente continua, f_X è massima in x, cioè $x \in \operatorname{argmax}_y f_X(y)$.*

Il caso discreto ci suggerisce quale sia il significato intuitivo della moda: è il valore più probabile (o meglio, uno dei valori più probabili). Questo non è del tutto corretto nel caso in cui X sia assolutamente continua, visto che la probabilità nei punti è sempre nulla.

Come accennato, non è detto che la moda di una distribuzione sia unica. Se lo è diciamo che è una legge *unimodale*, se ha due mode diciamo che è *bimodale* e in generale se non è unimodale allora è *multimodale*.

Esempio 9.55. Vediamo alcuni esempi di variabili aleatorie e delle loro mode.

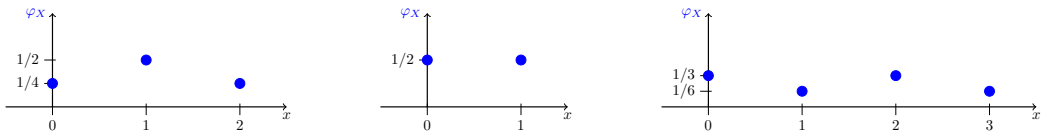


Figura 9.4. Tre variabili discrete. Nella prima la moda è 1, nella seconda sia 0 sia 1 sono mode, nella terza sia 0 sia 2 sono mode.

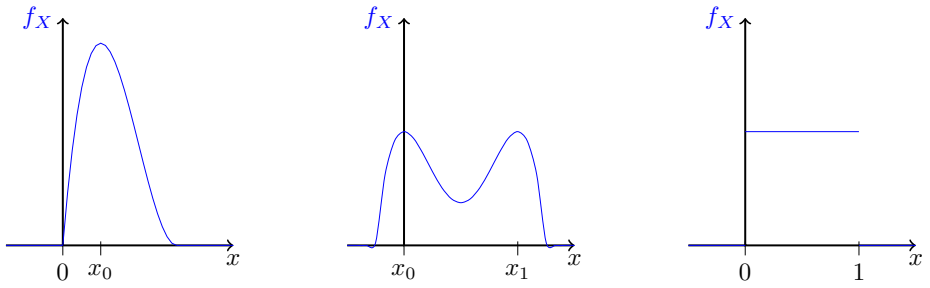


Figura 9.5. Tre variabili assolutamente continue. Nella prima la moda è x_0, nella seconda sono mode x_0 e x_1, nella terza tutti i punti tra 0 e 1 sono mode.

Osservazione 9.56. Nel caso continuo potrebbe venirci la tentazione di prendere la derivata della funzione densità e cercare i punti in cui si annulla. Purtroppo non sempre funziona, come possiamo vedere dagli esempi in Figura 9.6.

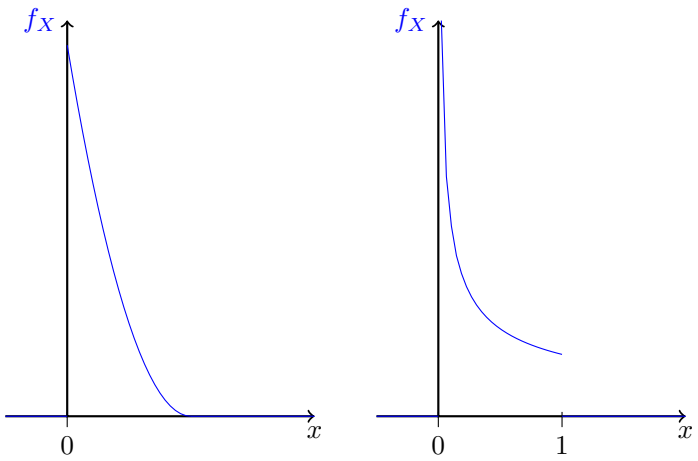


Figura 9.6. Due variabili assolutamente continue. Nella prima la moda è 0, ma la derivata è nulla per $x<0$ e per $x>1$. Nella seconda la moda è in corrispondenza dell'asintoto verticale in 0 (la densità in corrispondenza della moda è infinita).

Osservazione 9.57. Media, mediana e moda sono tre modi diversi per predire il valore di una variabile aleatoria, ottimizzando su criteri diversi.

La *media* è il valore che minimizza lo scarto (o errore) quadratico medio: per ogni $c\in\mathbb{R}$

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

dal momento che $(E[X]-c)^2\geqslant 0$.

La *mediana*, invece, minimizza la media dell'errore assoluto: per ogni $c\in\mathbb{R}$

$$E[|X-c|]\geqslant E[|X-m_X|].$$

Vediamolo nel caso in cui X sia assolutamente continua:

$$\begin{aligned} E[|X-c|] &= \int_{-\infty}^{+\infty}|x-c|\,f_X(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{c}-(x-c)\,f_X(x)\,\mathrm{d}x+\int_{c}^{+\infty}(x-c)\,f_X(x)\,\mathrm{d}x. \end{aligned}$$

Volendo minimizzare questa quantità al variare di c ne possiamo prendere la derivata in c e porla uguale a 0:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}c}E[|X-c|] &= -\frac{\mathrm{d}}{\mathrm{d}c}\int_{-\infty}^{c}(x-c)\,f_X(x)\,\mathrm{d}x+\frac{\mathrm{d}}{\mathrm{d}c}\int_{c}^{+\infty}(x-c)\,f_X(x)\,\mathrm{d}x \\ &= -\int_{-\infty}^{c}f_X(x)\,\mathrm{d}x+\int_{c}^{+\infty}f_X(x)\,\mathrm{d}x \end{aligned}$$

da cui segue che il minimo di $E[|X-c|]$ è in corrispondenza di $\tilde{c}$ tale che

$$\int_{-\infty}^{\tilde{c}}f_X(x)\,\mathrm{d}x=\int_{\tilde{c}}^{+\infty}f_X(x)\,\mathrm{d}x$$

ossia di $\tilde{c}$ per cui

$$F_X(\tilde{c})=1-F_X(\tilde{c}),$$

ma questa è proprio la caratterizzazione (9.7) della mediana nel caso X sia assolutamente continua.

La *moda*, infine, è il valore che massimizza la probabilità.

In generale questi tre numeri non coincidono. Quello "giusto" da usare dipende dal contesto.

Esempio 9.58. Supponiamo di avere un d6 sbilanciato in cui 1 esce con probabilità $\frac{1}{6}+5\varepsilon$ e le altre facce ciascuna con probabilità $\frac{1}{6}-\varepsilon$, per $\varepsilon=10^{-3}$. In questo caso la media (o valore atteso) è 3.485, la mediana non è definita e la mediana impropria è 3. La moda è 1.

Se vogliamo scommettere su un numero ci conviene scegliere la moda, se vogliamo minimizzare l'errore assoluto tra il numero che scegliamo e il numero che esce, scegliamo la mediana (impropria) e se vogliamo minimizzare l'errore quadratico medio scegliamo il valore atteso.

Per finire questa carrellata sugli indicatori, vediamone alcuni associati ai momenti centrati di ordine superiore a 2.

DEFINIZIONE 9.59. *Chiamiamo* skewness *di una variabile aleatoria X il suo momento terzo centrato e standardizzato, cioè*

$$\mathrm{sk}[X] = E\left[\left(\frac{X - E[X]}{\sqrt{\mathrm{Var}[X]}}\right)^3\right] = \frac{E[(X-E[X])^3]}{(\sqrt{\mathrm{Var}[X]})^3}.$$

La skewness è un indicatore della simmetria di X: se X è simmetrica, allora $\mathrm{sk}[X] = 0$ (ma non è necessariamente vero il viceversa), se $\mathrm{sk}[X] > 0$ la variabile aleatoria X ha una "scentratura" verso sinistra rispetto alla media, se $\mathrm{sk}[X] < 0$ ha una "scentratura" verso destra.

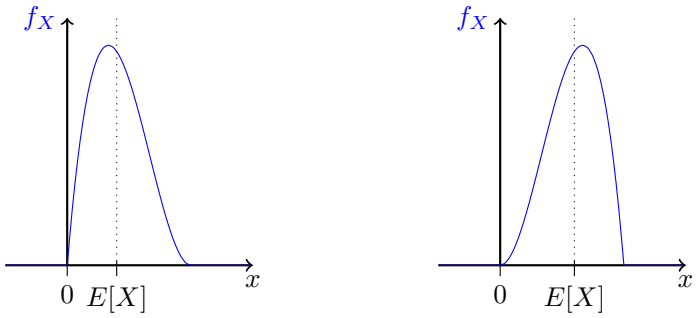


Figura 9.7. Due variabili aleatorie continue. Quella di sinistra ha skewness positiva, quella di destra ha skewness negativa.

DEFINIZIONE 9.60. *Chiamiamo* curtosi *o* kurtosis *di una variabile aleatoria X il suo momento quarto centrato e standardizzato, cioè*

$$\mathrm{kr}[X] = E\left[\left(\frac{X - E[X]}{\sqrt{\mathrm{Var}[X]}}\right)^4\right] = \frac{E[(X-E[X])^4]}{(\mathrm{Var}[X])^2}.$$

La kurtosis misura la concentrazione di una distribuzione e in questo caso il valore soglia è 3: una variabile con kurtosis maggiore di 3 ha un picco molto alto della densità attorno alla media e delle code pesanti. Viceversa una variabile con kurtosis minore di 3 ha un "plateau" attorno alla sua media e code leggere. Il metro di paragone (con kurtosis uguale a 3) è la variabile normale standard.

9.7. SPERANZA E VARIANZA CONDIZIONATE

Come abbiamo già accennato nell'introdurre il concetto di speranza, ma anche nel calcolare la speranza delle geometriche, ha senso parlare di speranza condizionata e per definirla non dobbiamo fare altro che considerare al posto di P la probabilità condizionata a un evento, $P(\cdot|E)$ e dunque la densità (discreta) della variabile aleatoria di interesse condizionata a tale evento.

PROPOSIZIONE 9.61. *Date una partizione $(E_i)_i$ di Ω in eventi disgiunti e una variabile aleatoria X vale*

$$E_X[X] = \sum_i E_X[X|E_i]\,P(E_i),$$

identità che prende il nome di fattorizzazione della speranza[9.11].

COROLLARIO 9.62. *Siano Y una variabile aleatoria discreta e X una variabile aleatoria qualsiasi. Allora*

$$E_X[X] = \sum_{y \in \mathcal{R}_X} E_X[X|Y=y]\,P(Y=y) = E_Y[E_X[X|Y]].$$

Dimostrazione. Supponiamo che anche X sia una variabile aleatoria discreta. Allora

$$\begin{aligned} E_Y[E_X[X|Y]] &= \sum_{y \in \mathcal{R}_Y} E_X[X|Y=y]\,P(Y=y) \\ &= \sum_{y \in \mathcal{R}_Y} \sum_{x \in \mathcal{R}_X} x \cdot \frac{P(X=x, Y=y)}{P(Y=y)} P(Y=y) \\ &= \sum_{x \in \mathcal{R}_X} x \sum_{y \in \mathcal{R}_Y} P(X=x, Y=y) \\ &= \sum_{x \in \mathcal{R}_X} x\,P(X=x) \\ &= E_X[X] \end{aligned}$$

in cui abbiamo marginalizzato in X la densità discreta congiunta. □

Osservazione 9.63. Risultati analoghi valgono anche per variabili aleatorie assolutamente continue, con l'accortezza di usare le densità al posto delle densità discrete.

Osservazione 9.64. Possiamo notare che $E[X|Y]$ è essa stessa una variabile aleatoria. La sua parte "casuale" è ereditata da Y. Ad esempio, se $Y \sim \text{bin}(1,p)$ allora

$$E[X|Y] = \begin{cases} E[X|Y=0] & \text{con probabilità } 1-p \\ E[X|Y=1] & \text{con probabilità } p. \end{cases}$$

Come abbiamo introdotto la speranza condizionata, possiamo definire anche la varianza condizionata: sia F un evento, allora

$$\text{Var}[X|F] = E[(X - E[X|F])^2|F] = \begin{cases} \sum_x (x - E[X|F])^2 \varphi_{X|F}(x|F) \\ \int (x - E[X|F])^2 f_{X|F}(x|F)\,\mathrm{d}x. \end{cases}$$

Nel caso particolare in cui F è determinato da una variabile aleatoria Y (che supponiamo discreta)

$$\text{Var}[X|Y=y] = \begin{cases} \sum_x (x - E[X|Y=y])^2 \varphi_{X|Y}(x|y) \\ \int (x - E[X|Y=y])^2 f_{X|Y}(x|y)\,\mathrm{d}x, \end{cases}$$

in cui l'ultima densità congiunta è mista. Anche $\text{Var}[X|Y]$ può essere vista come una variabile aleatoria che eredita la casualità da Y.

PROPOSIZIONE 9.65. *Date due variabili aleatorie X e Y vale la seguente identità*

$$\text{Var}_X[X] = E_Y[\text{Var}_X[X|Y]] + \text{Var}_Y[E_X[X|Y]],$$

detta scomposizione (o fattorizzazione) della varianza.

Dimostrazione. Riscriviamo il secondo membro dell'uguaglianza

$$\begin{aligned} E_Y[\text{Var}_X[X|Y]] + \text{Var}_Y[E_X[X|Y]] &= E_Y[E_X[X^2|Y] - E_X[X|Y]^2] + E_Y[E_X[X|Y]^2] - (E_Y[E_X[X|Y]])^2 \\ &= E_Y[E_X[X^2|Y]] - E_X[X]^2 \\ &= E_X[X^2] - E_X[X]^2 = \text{Var}_X[X] \end{aligned}$$

in cui abbiamo usato la linearità della speranza e, due volte, il Corollario 9.62. □

9.11. La notazione E_X serve per mettere in evidenza che si tratta della speranza associata alla variabile aleatoria X. In realtà non è necessario indicarlo e di solito non lo si fa.

CAPITOLO 10

MODELLI ASSOLUTAMENTE CONTINUI

In analogia a quanto fatto nel Capitolo 8 per le variabili aleatorie discrete, vediamo ora alcuni modelli di variabili aleatorie assolutamente continue.

10.1. UNIFORMI

Le abbiamo già incontrate più volte, ma diamone comunque una definizione.

DEFINIZIONE 10.1. *Dati due numeri reali $a<b$, chiamiamo* uniforme su $[a,b]$ *una variabile aleatoria assolutamente continua X la cui densità f_X è costante in $[a,b]$ e nulla altrove. Scriviamo in questo caso $X\sim\text{unif}[a,b]$ o $X\sim\text{unif}(a,b)$.*

Come abbiamo già visto, il valore costante non nullo c di f_X in $[a,b]$ è determinato da a e b:

$$1=\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x=\int_a^b c\,\mathrm{d}x=c\,(b-a),$$

da cui ricaviamo

$$f_X(x)=\begin{cases}\frac{1}{b-a} & x\in(a,b)\\ 0 & x\in[a,b]^c.\end{cases}$$

Per definizione la funzione di ripartizione è l'integrale della densità, quindi per $X\sim\text{unif}[a,b]$,

$$F_X(x)=\begin{cases}0 & x\leqslant a\\ \frac{x-a}{b-a} & a\leqslant x\leqslant b\\ 1 & x\geqslant b.\end{cases}$$

Esempio 10.2. La variabile aleatoria uniforme su $[1,3]$ è rappresentata in Figura 10.1.

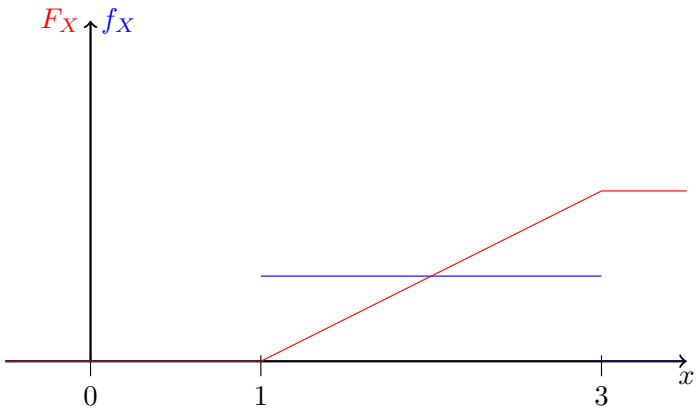


Figura 10.1. Funzione di ripartizione e di densità di $X\sim\text{unif}(1,3)$.

10.1.1. Uniformi in R

La funzione densità per un'uniforme è la funzione `dunif(x, min = 0, max = 1)`. Con `x` indichiamo il punto in cui la vogliamo calcolare, mentre `min` è il primo estremo dell'intervallo (quello che abbiamo indicato con a) e `max` il secondo estremo (b nella definizione vista prima).

Abbiamo poi la funzione `punif(q, min=0, max = 1, lower.tail = TRUE)` che ci permette di calcolare, in `q`, la funzione di ripartizione, con il consueto parametro per determinare quale coda ci interessa.

La funzione quantile è `qunif(p, min = 0, max = 1, lower.tail = TRUE)` e il generatore casuale è `runif(n, min = 0, max = 1)`.

10.1.2. Indicatori per le uniformi

Possiamo calcolare gli indicatori per $X \sim \text{unif}[a,b]$.

- Speranza: $E[X] = \int_a^b x \cdot \frac{1}{b-a}\,dx = \frac{1}{b-a}\cdot\frac{b^2-a^2}{2} = \frac{a+b}{2}$, come ci saremmo aspettati.
- Varianza: $\text{Var}[X] = \int_a^b x^2 \cdot \frac{1}{b-a}\,dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$.
- Mediana: coincide con la media.
- Moda: qualunque valore in (a,b).
- Skewness: $\text{sk}[X] = \frac{E[(X-E[X])^3]}{\text{Var}[X]^{3/2}} = 0$, per simmetria della distribuzione, oppure calcolando gli integrali,
$$\int_a^b \left(x - \frac{a+b}{2}\right)^3 dx = \int_{\frac{a-b}{2}}^{\frac{b-a}{2}} y^3\,dy = \frac{2}{b-a}\int_{-1}^{+1} z^3\,dz = 0,$$
in cui abbiamo fatto i due cambi di variabile $y = x - \frac{a+b}{2}$ e $z = \frac{b-a}{2}y$ e abbiamo sfruttato il fatto che z^3 è dispari e integrata in un dominio simmetrico rispetto a 0.
- Kurtosis: $\text{kr}[X] = \frac{E[(X-E[X])^4]}{\text{Var}[X]^2} = \frac{\frac{1}{b-a}\cdot\frac{(b-a)^5}{80}}{\frac{(b-a)^4}{144}} = \frac{9}{5}$.

Esempio 10.3. Siamo in attesa alla fermata dell'autobus, che (in teoria) passa ogni 15'. Possiamo rappresentare il tempo che passiamo alla fermata tra il nostro arrivo e la salita sull'autobus come una variabile aleatoria uniforme $X \sim \text{unif}[0,15]$.

Qual è la probabilità di aspettare più di 5'? Qual è la probabilità che, avendo aspettato (senza successo) 5', ne dobbiamo aspettare ancora più di 5?

Sappiamo che la funzione densità è $f_X(x) = \frac{1}{15}\mathbb{1}_{[0,15]}(x)$ e dunque che la funzione di ripartizione è
$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{15} & 0 \leqslant x < 15 \\ 1 & x \geqslant 15. \end{cases}$$
La prima domanda ci chiede di calcolare
$$P(X > 5) = 1 - F_X(5) = 1 - \frac{5}{15} = \frac{2}{3}.$$
La seconda, invece, chiede
$$P(X > 10 \mid X > 5) = \frac{P(X > 10, X > 5)}{P(X > 5)} = \frac{1 - F_X(10)}{1 - F_X(5)} = \frac{1}{3}\cdot\frac{3}{2} = \frac{1}{2}.$$

Osservazione 10.4. Per generare realizzazioni di una variabile aleatoria di distribuzione assegnata F, possiamo generare realizzazioni di una distribuzione uniforme su $[0,1]$ (che non a caso è quella di default in R) e calcolarne la funzione quantile. Non è sempre il modo computazionalmente più efficiente.

10.2. ESPONENZIALI

Anche se non l'abbiamo ancora definita, è una variabile aleatoria che abbiamo incontrato spesso in esempi ed esercizi.

DEFINIZIONE 10.5. *Diciamo che una variabile aleatoria X è* esponenziale di parametro $\lambda > 0$ *se ha densità*

$$f_X(x) = \begin{cases} 0 & x < 0 \\ c \cdot e^{-\lambda x} & x \geqslant 0. \end{cases}$$

In questo caso scriviamo $X \sim \exp(\lambda)$ o $X \sim \text{expo}(\lambda)$. Il parametro λ prende anche il nome di intensità *o* rate *dell'esponenziale.*

Da quanto visto sulle costanti di rinormalizzazione, ricaviamo che $c = \lambda$. La funzione di ripartizione di $X \sim \exp(\lambda)$ è

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-\lambda x} & x \geqslant 0. \end{cases}$$

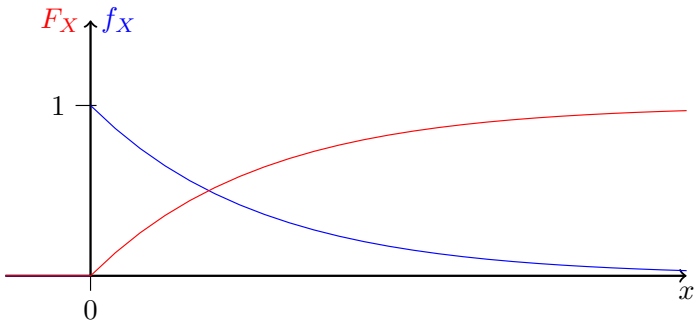


Figura 10.2. Funzione di ripartizione e di densità di $X \sim \exp(1)$.

In Figura 10.2 sono rappresentate le funzioni di ripartizione e di densità dell'esponenziale di rate 1. Al variare di λ abbiamo comportamenti leggermente diversi. In particolare: l'intercetta sulle ordinate è λ e la pendenza delle curve è maggiore se $\lambda > 1$ e minore se $\lambda < 1$.

10.2.1. Esponenziali in R

La famiglia delle funzioni associate all'esponenziale in R prende il nome `exp`. Abbiamo dunque la densità `dexp(x, rate = 1)`, la funzione di ripartizione `pexp(q, rate = 1, lower.tail = TRUE)` e le funzioni quantile `qexp(p, rate = 1, lower.tail = TRUE)` e generatore casuale `rexp(n, rate = 1)`.

10.2.2. Indicatori per le esponenziali

Possiamo calcolare gli indicatori per $X \sim \exp(\lambda)$.

- Speranza: $E[X] = \int_a^b x \cdot \lambda \cdot e^{-\lambda x} \, dx = \lambda \left[\frac{e^{-\lambda x}}{\lambda^2} (\lambda x - 1) \right]_0^{+\infty} = \frac{1}{\lambda}$, ossia il reciproco del rate.
- Varianza: $\text{Var}[X] = \int_a^b x^2 \cdot \lambda \cdot e^{-\lambda x} \, dx - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$, integrando per parti.
- Mediana: dal momento che F_X è monotona strettamente crescente per $x > 0$, dobbiamo risolvere $1 - e^{-\lambda x} = \frac{1}{2}$, cioè $-\lambda x = \log\left(\frac{1}{2}\right)$, da cui $x = \log(2) \cdot \lambda^{-1}$.
- Moda: è il punto di massimo di f_X, ossia 0.
- Skewness: $\text{sk}[X] = \frac{E[(X - E[X])^3]}{\text{Var}[X]^{3/2}} = 2$.
- Kurtosis: $\text{kr}[X] = \frac{E[(X - E[X])^4]}{\text{Var}[X]^2} = 9$.

Esempio 10.6. Siamo sempre alla fermata dell'autobus come nell'Esempio 10.3 e il tempo medio di attesa è ancora una volta $\frac{15}{2}$. Questa volta, però, ipotizziamo che il tempo di attesa per l'arrivo dell'autobus sia distribuito come un'esponenziale.

Qual è la probabilità di aspettare più di 5'? Qual è la probabilità che, avendo aspettato (senza successo) 5', ne dobbiamo aspettare ancora più di 5?

Sapendo la media, possiamo ricavare immediatamente il rate dell'esponenziale: $\lambda = \frac{2}{15}$. Per rispondere alla prima domanda dobbiamo calcolare

$$P(X>5) = 1 - F_X(5) = e^{-\frac{2}{15}\cdot 5} \approx 0.51.$$

Per la seconda, invece,

$$P(X>10 \mid X>5) = \frac{1-F_X(10)}{1-F_X(5)} = e^{-\frac{4}{3}} \cdot e^{\frac{2}{3}} = e^{-\frac{2}{3}} = P(X>5) \approx 0.51.$$

Il fatto di aver aspettato 5' non fa diminuire la probabilità che dobbiamo aspettarne altri 5.

La descrizione è abbastanza diversa da quella vista nell'Esempio 10.3: qui possiamo anche osservare che la probabilità di aspettare più di mezz'ora (nulla nel caso della distribuzione uniforme) è $P(X>30) = 1 - F_X(30) = e^{-4} \approx 0.02$.

Lezione 19

La risposta alla seconda domanda nell'Esempio 10.6 ci suggerisce che anche per le esponenziali, così come per le geometriche, valga la proprietà di assenza di memoria.

PROPOSIZIONE 10.7. *Se X è una variabile aleatoria esponenziale, allora ha assenza di memoria, ossia per s, $t>0$*

$$P(X>s+t \mid X>s) = P(X>t).$$

Dimostrazione. Basta sfruttare la forma della funzione di ripartizione:

$$\begin{aligned} P(X>s+t \mid X>s) &= \frac{P(X>s+t, X>s)}{P(X>s)} \\ &= \frac{1-F_X(s+t)}{1-F_X(s)} \\ &= e^{-\lambda(s+t)} \cdot e^{\lambda s} \\ &= e^{-\lambda t} = 1 - F_X(t) = P(X>t) \end{aligned}$$

semplicemente usando le proprietà dell'esponenziale. □

Osservazione 10.8. Le variabili aleatorie esponenziali sono la controparte continua delle geometriche. Possiamo usarle per descrivere i tempi d'attesa di eventi casuali con assenza di memoria, ossia che non diventano più probabili solo perché sono "in ritardo". È anche possibile caratterizzare le esponenziali come limite di variabili aleatorie geometriche.

Esempio 10.9. Se $X \sim \exp(\lambda)$ e $Y = \alpha X$, (con $\alpha > 0$) allora $Y \sim \exp\left(\frac{\lambda}{\alpha}\right)$. Infatti sappiamo dalla Proposizione 6.3 che

$$f_Y(y) = \frac{1}{\alpha} f_X\left(\frac{y}{\alpha}\right) = \frac{\lambda}{\alpha} e^{-\lambda \frac{y}{\alpha}}.$$

10.3. GAUSSIANE O NORMALI

È la famiglia più nota e diffusa di variabili aleatorie (vedremo nel prossimo capitolo una delle ragioni), dalla caratteristica forma "a campana" della funzione di densità. Possiamo averla senza parametri (normale standard) oppure con parametri espliciti.

DEFINIZIONE 10.10. *Diciamo che una variabile aleatoria X è una* normale (o Gaussiana) standard *se ha densità*

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}.$$

Scriviamo in questo caso $X \sim \mathcal{N}(0,1)$.

Osservazione 10.11. Consideriamo i vari elementi della funzione densità e capiamone il ruolo:

- il termine e^{-x^2} ci dà la forma a campana
- il coefficiente $\frac{1}{2}$ a esponente semplifica la derivazione
- il coefficiente $\frac{1}{\sqrt{2\pi}}$ è la costante di rinormalizzazione (vedi anche Appendice A.3).

Inoltre possiamo osservare alcune proprietà della densità di una Gaussiana standard, rappresentata in Figura 10.3:

- è simmetrica rispetto all'asse $x=0$, quindi $f_X(-x)=f_X(x)$
- ha massimo in $x=0$, con valore $1/\sqrt{2\pi} \approx 0.4$
- ha flessi in $x=\pm 1$ e in tali punti ha valore $(\sqrt{e\,2\pi})^{-1} \approx 0.24$
- in ± 2 ha valore $(e^2\sqrt{2\pi})^{-1} \approx 0.05$
- in ± 3 ha valore $\left(\sqrt{e^9\,2\pi}\right)^{-1} \approx 0.004$.

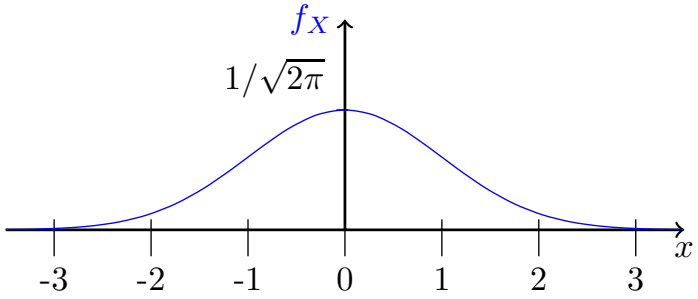


Figura 10.3. Densità di una normale standard

La funzione di ripartizione di una variabile aleatoria normale standard X è

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx =: \Phi(x).$$

Non è un integrale con soluzione algebrica (anche se siamo in grado di calcolarlo in 0 e $\pm\infty$). Per sapere il valore di Φ in un certo punto, possiamo usare le tavole (riportate in molti libri e anche qui in Appendice B), oppure usare un software (ad esempio R).

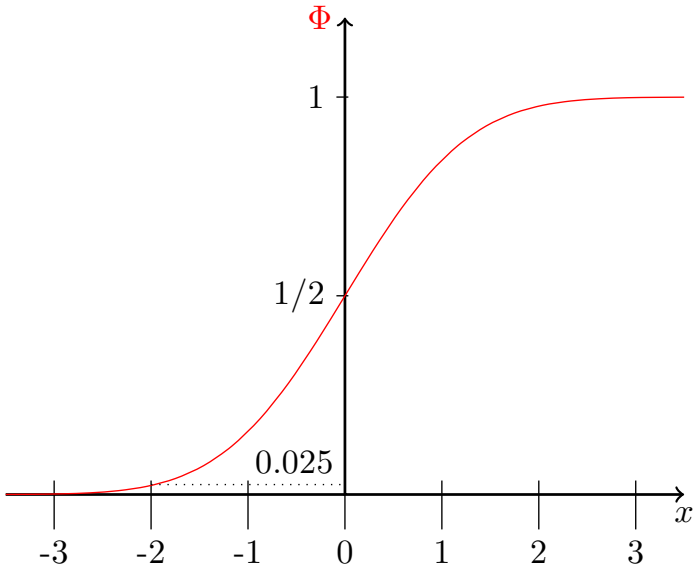


Figura 10.4. Funzione di ripartizione di una normale standard

Osservazione 10.12. Anche Φ, come già la densità, ha alcune proprietà interessanti, che possiamo anche vedere nella Figura 10.4

- è simmetrica rispetto al punto $\left(0,\frac{1}{2}\right)$, quindi $\Phi(-x)=1-\Phi(x)$
- in $x=0$ vale $\frac{1}{2}$
- in $x=-2$ vale circa 0.0228, in $x=2$ vale circa 0.9772
- in $x=-3$ vale circa 0.0013, in $x=3$ vale circa 0.9987.

In particolare abbiamo che, per $X\sim\mathcal{N}(0,1)$, $P(X\in(-3,3))\approx 0.997$ e $P(X\in(-2,2))\approx 0.95$. La funzione di densità è non nulla su tutto $\mathbb{R}$ e quindi X può assumere valori su tutto $\mathbb{R}$, ma in realtà le realizzazioni saranno con alta probabilità concentrate in un intervallo centrato in 0 e di larghezza 6.

La funzione quantile di una normale standard è l'inversa Φ^{-1} della funzione di ripartizione Φ. Abbiamo

$$\Phi^{-1}(p)=x\iff\Phi(x)=p\iff P(X\leqslant x)=p.$$

La funzione quantile (definita sull'intervallo $[0,1]$) è simmetrica rispetto al punto $\left(\frac{1}{2},0\right)$, quindi $\Phi^{-1}(p)=-\Phi^{-1}(1-p)$.

10.3.1. Indicatori per la normale standard

Sia $X\sim\mathcal{N}(0,1)$. Cominciamo col calcolarne la speranza:

$$E[X]=\int_{-\infty}^{+\infty}xf_X(x)\,\mathrm{d}x=\int_0^{+\infty}-xf_X(-x)\,\mathrm{d}x+\int_0^{+\infty}xf_X(x)\,\mathrm{d}x=\int_0^{+\infty}0\cdot f_X(x)\,\mathrm{d}x=0.$$

In altre parole, grazie alla simmetria rispetto a $x=0$ abbiamo che $E[X]=0$. Osserviamo che anche mediana e moda sono in $x=0$.

Passiamo ora alla varianza:

$$\mathrm{Var}[X]=\int_{-\infty}^{+\infty}x^2\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x=\left[-x\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\right]_{-\infty}^{+\infty}+\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}\mathrm{d}x=1,$$

in cui abbiamo integrato per parti e usato il fatto che f_X è una densità di probabilità. Possiamo anche calcolare skewness e kurtosis: ricaviamo $\mathrm{sk}[X]=0$ (per simmetria) e $\mathrm{kr}[X]=3$.

Possiamo però osservare che, pur non avendo dichiarato che la normale standard non ha parametri, l'abbiamo definita come $\mathcal{N}(0,1)$ e, ora, potremmo avere qualche sospetto su cosa siano quei due numeri.

DEFINIZIONE 10.13. *Sia $Z\sim\mathcal{N}(0,1)$ una normale standard. Chiamiamo* Gaussiana (o normale) di parametri $\mu\in\mathbb{R}$ e $\sigma\in\mathbb{R}_0^+$ *una variabile aleatoria X tale che $X=\sigma Z+\mu$. In questo caso scriviamo $X\sim\mathcal{N}(\mu,\sigma)$.*

Osservazione 10.14. Possiamo facilmente ricavare densità e funzione di ripartizione di una Gaussiana di parametri μ e σ, infatti è per definizione una trasformazione (lineare) di una variabile aleatoria di cui conosciamo la legge: $X=\sigma Z+\mu$, ossia $Z=\frac{X-\mu}{\sigma}$, cioè Z è centrata e standardizzata. A partire dalla legge di Z possiamo scrivere la funzione di ripartizione di X

$$F_X(x)=P(X\leqslant x)=P(\sigma Z+\mu\leqslant x)=P\left(Z\leqslant\frac{x-\mu}{\sigma}\right)=F_Z\left(\frac{x-\mu}{\sigma}\right)=\Phi\left(\frac{x-\mu}{\sigma}\right)$$

così come la sua densità

$$f_X(x)=\frac{1}{\sigma}f_Z\left(\frac{x-\mu}{\sigma}\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.\tag{10.1}$$

In particolare possiamo calcolare la funzione di ripartizione usando la stessa funzione (o le stesse tavole) definite per la funzione di ripartizione Φ della normale standard.

10.3.2. Indicatori per una normale

Se $X \sim \mathcal{N}(\mu, \sigma)$, possiamo ricavare facilmente i suoi indicatori usando il fatto che è una trasformazione lineare di una normale standard:

$$E[X] = E[\sigma Z + \mu] = \sigma E[Z] + \mu = \mu$$

per la speranza (ma anche per la mediana e la moda) e

$$\mathrm{Var}[X] = \mathrm{Var}[\sigma Z + \mu] = \sigma^2 \mathrm{Var}[Z] = \sigma^2$$

per la varianza. I due parametri che caratterizzano una distribuzione normale sono la sua media e la sua deviazione standard (o equivalentemente la sua varianza σ^2).

Skewness e kurtosis sono invariate, rispetto a una normale standard: $\mathrm{sk}[X] = 0$, dal momento che la simmetria rispetto al valore atteso non è venuta meno, e $\mathrm{kr}[X] = \frac{3\sigma^4}{\sigma^4} = 3$.

Osservazione 10.15. Si può equivalentemente usare la varianza come secondo parametro di una normale. Questo è comodo perché semplifica alcune scritture, in particolare quella per la somma di due Gaussiane, come vedremo tra poco, e nel momento in cui passiamo alle Gaussiane multivariate, nelle quali entra in gioco la matrice di covarianza. In queste note abbiamo scelto di usare la deviazione standard σ e non la varianza σ^2 per allinearci alla convenzione usata da R.

Osservazione 10.16. Siccome abbiamo trasformato linearmente una normale standard, la funzione di densità sarà una traslazione e dilatazione della densità di una normale standard, come possiamo vedere nella (10.1) o nella Figura 10.5.

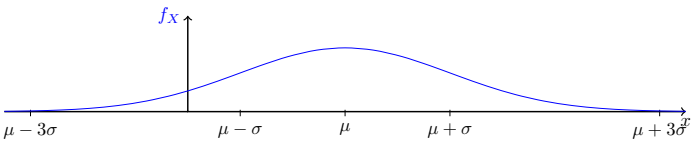


Figura 10.5. Densità di una normale

In μ abbiamo il massimo e in $\mu \pm \sigma$ abbiamo i due flessi, e al di fuori dell'intervallo $[\mu - 3\sigma, \mu + 3\sigma]$ la densità è molto piccola (pur non essendo nulla). Questo è importante in fase di modellizzazione: da un lato non è del tutto corretto usare una Gaussiana per descrivere un fenomeno aleatorio in cui sappiamo che i valori sono necessariamente all'interno di un intervallo, ma se vogliamo (o dobbiamo) farlo, è necessario che controlliamo almeno che le realizzazioni della variabile aleatoria che scegliamo cadano con altissima probabilità all'interno dell'intervallo di interesse, cosa che è codificata nei parametri μ e σ.

Naturalmente, al variare di μ e σ varieranno il centro della densità, l'altezza del massimo e la concentrazione della distribuzione, come vediamo ad esempio in Figura 10.6

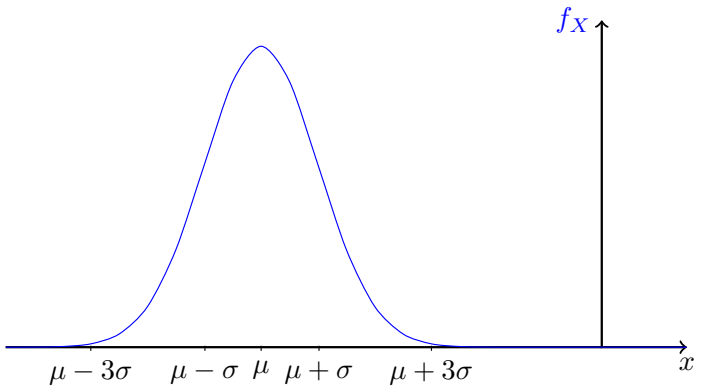


Figura 10.6. Densità di un'altra normale

PROPOSIZIONE 10.17. *La famiglia delle distribuzioni Gaussiane è riproducibile. Date* $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ *e* $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ *indipendenti,*

$$X_1 + X_2 \sim \mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$$

ossia la loro somma è una Gaussiana di media la somma delle medie e di varianza la somma delle varianze.

Dimostrazione. [TBA] □

Osservazione 10.18. Possiamo lasciar cadere l'ipotesi di indipendenza. Prese due qualunque Gaussiane $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ aventi correlazione $\rho = \mathrm{corr}[X_1, X_2]$, la loro somma è una Gaussiana di media $\mu_1 + \mu_2$ e di varianza $\sigma_1^2 + \sigma_2^2 + 2\,\sigma_1\,\sigma_2\,\rho$, ossia uguale alla somma delle varianze, infatti

$$\begin{aligned} \mathrm{Var}[X_1 + X_2] &= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\,\mathrm{Cov}[X_1, X_2] \\ &= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\,\mathrm{Corr}[X_1, X_2]\,\sqrt{\mathrm{Var}[X_1]\,\mathrm{Var}[X_2]}. \end{aligned}$$

Osservazione 10.19. Come abbiamo visto, possiamo calcolare la funzione di ripartizione di una Gaussiana appoggiandoci alla funzione di ripartizione Φ di una Gaussiana standard. Questa trasformazione $X \to \frac{X-\mu}{\sigma}$ prende il nome di standardizzazione ed entra in gioco non solo per la funzione di ripartizione, ma anche per la sua inversa, la funzione quantile Φ^{-1}. Se stiamo cercando il p-quantile di $X \sim \mathcal{N}(\mu, \sigma)$, il problema che vogliamo risolvere è trovare $x \in \mathbb{R}$ tale che $P(X \leqslant x) = p$. Cominciamo riscrivendo questa identità mediante la standardizzazione, come già fatto in precedenza:

$$p = P(X \leqslant x) = P\left(Z \leqslant \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

e ora approfittiamo del fatto che la funzione Φ è invertibile per avere

$$\Phi^{-1}(p) = \Phi^{-1}\left(\Phi\left(\frac{x-\mu}{\sigma}\right)\right) = \frac{x-\mu}{\sigma}$$

che, scritta esplicitamente in x, ci dà $x = \sigma\,\Phi^{-1}(p) + \mu$.

10.3.3. Gaussiane in R

Le variabili aleatorie Gaussiane o normali sono descritte in R nella famiglia `norm`. In particolare, la densità di una normale è `dnorm(x, mean = 0, sd = 1)`, in cui `mean` è la media e `sd` è la deviazione standard. Osserviamo anche che, se non passiamo valori per questi due parametri, di default R considererà la normale standard.

La funzione di ripartizione è `pnorm(q, mean=0, sd=1, lower.tail=TRUE)`, mentre la sua inversa (la funzione quantile) è `qnorm(p, mean = 0, sd = 1, lower.tail = TRUE)`.

Infine, per generare numeri casuali distribuiti secondo una Gaussiana, possiamo usare la funzione `rnorm(n, mean = 0, sd = 1)`.

10.3.4. Normali multivariate

Abbiamo visto che possiamo considerare vettori aleatori invece di variabili aleatorie e che questi sono caratterizzati dalla loro densità congiunta (nel caso assolutamente continuo). Vediamo ora un caso particolare, quello delle normali multivariate.

DEFINIZIONE 10.20. *Diciamo che un vettore aleatorio* $(X_1, \ldots, X_n)$ *è una* normale standard multivariata *o* vettore aleatorio normale standard *se le sue componenti* X_i *sono indipendenti e identicamente distribuite come normali standard, ossia* $X_i \sim \mathcal{N}(0,1)$*. In questo caso scriviamo* $X \sim \mathcal{N}(\mathbf{0}, \mathrm{Id})$*, con* $\mathbf{0}$ *il vettore* n*-dimensionale di soli 0 e* Id *la matrice identità* $n \times n$*.*

Osserviamo che in questo caso la densità congiunta è

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n}}\, e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} = \frac{1}{\sqrt{(2\pi)^n}}\, e^{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{x}} = \frac{1}{\sqrt{(2\pi)^n}}\, e^{-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{x}}$$

in cui abbiamo usato la notazione $\boldsymbol{x} = (x_1, \ldots, x_n)$ per compattezza.

DEFINIZIONE 10.21. *Diciamo che un vettore aleatorio $X = (X_1, \ldots, X_n)$ è una* normale multivariata *o* vettore aleatorio normale (non degenere) *se esistono un n-vettore aleatorio standard Z, un n-vettore colonna*[10.1] $\boldsymbol{\mu}$ *e una matrice $n \times n$ $\boldsymbol{M}$ tali che $X = \boldsymbol{M} Z + \boldsymbol{\mu}$. In questo caso scriviamo $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, dove $\boldsymbol{\Sigma} = \boldsymbol{M}\boldsymbol{M}^t$.*

In questo caso la densità congiunta è

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{1}{\sqrt{|\boldsymbol{\Sigma}|(2\pi)^n}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}.$$

Il vettore $\boldsymbol{\mu}$ è il vettore media e la matrice $\boldsymbol{\Sigma}$ è la matrice di covarianza, con determinante $|\boldsymbol{\Sigma}|$. Nel caso particolare $n = 2$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ e

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix},$$

dove σ_i è la deviazione standard di X_i (e dunque σ_i^2 ne è la varianza), $\rho = \mathrm{corr}[X_1, X_2]$ e quindi $\rho\,\sigma_1\sigma_2 = \mathrm{Cov}[X_1, X_2]$. Allora

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2 (1-\rho^2)}} \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)}$$

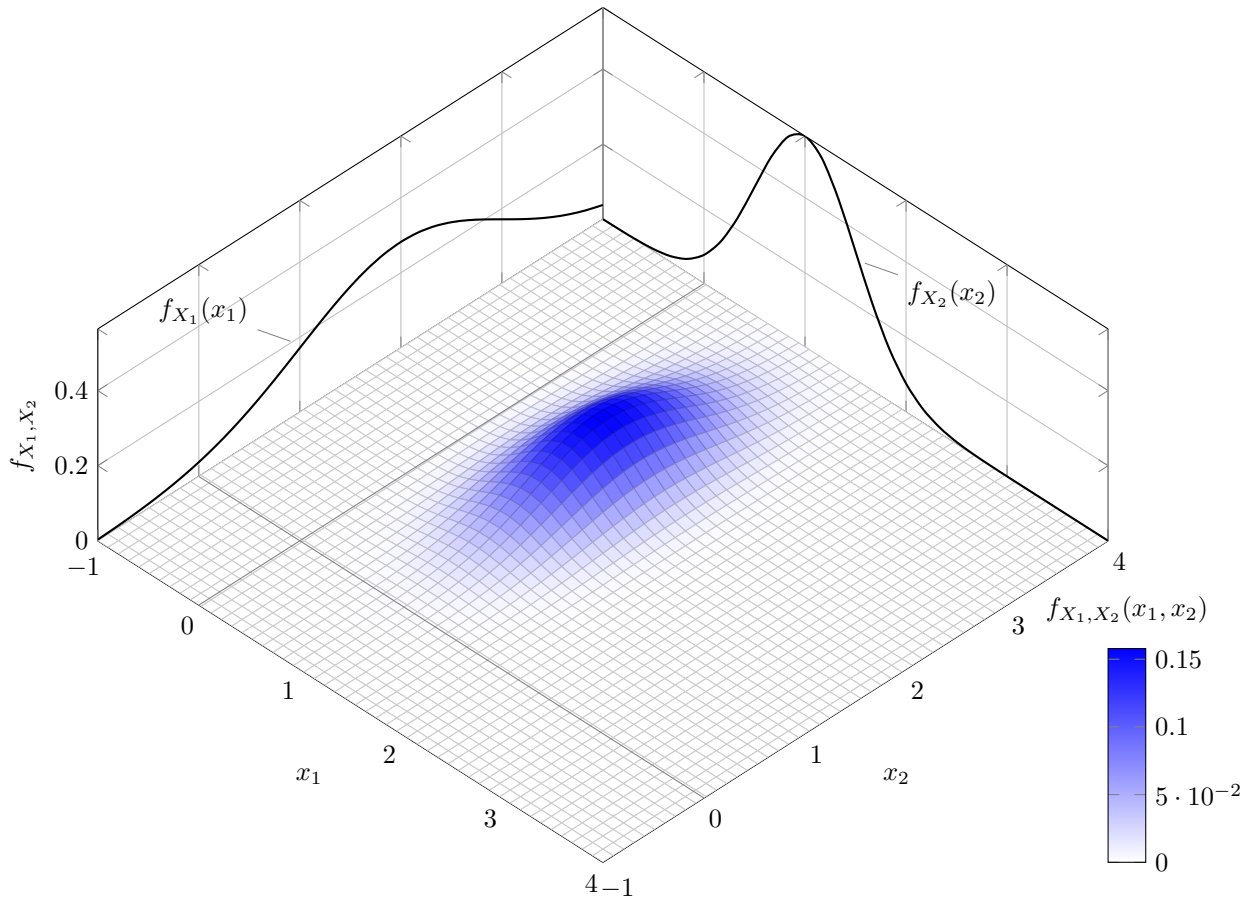


Figura 10.7. Una Gaussiana bivariata ($\mu_1 = 1$, $\sigma_1 = 0.5$, $\mu_2 = 2$, $\sigma_2 = 1$, $\rho = 0$)

10.1. Vogliamo un vettore colonna, per moltiplicare meglio. In particolare anche X è un vettore aleatorio colonna.

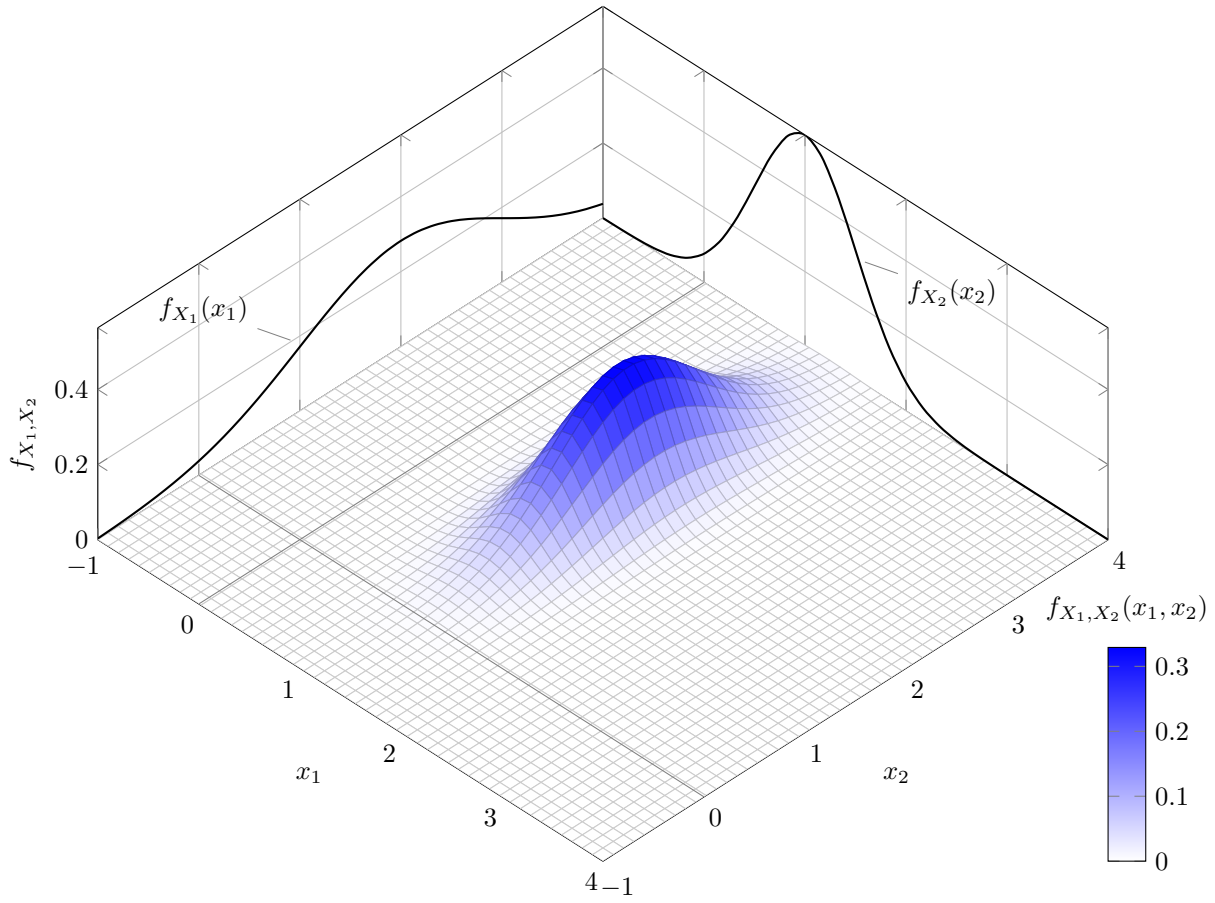


Figura 10.8. Una Gaussiana bivariata ($\mu_1 = 1$, $\sigma_1 = 0.5$, $\mu_2 = 2$, $\sigma_2 = 1$, $\rho = -0.75$)

10.4. CHI QUADRO

Partiamo da una variabile aleatoria normale standard: $X \sim \mathcal{N}(0,1)$. Qual è la legge di $Y = X^2$? Abbiamo (per $y \geqslant 0$)

$$F_Y(y) = P(X^2 \leqslant y) = P(-\sqrt{y} \leqslant X \leqslant \sqrt{y}) = P(X \leqslant \sqrt{y}) - P(X \leqslant -\sqrt{y}) = 2\,\Phi(\sqrt{y}) - 1$$

usando la proprietà di Φ per cui $-\Phi(-\sqrt{y}) = -(1 - \Phi(\sqrt{y}))$. Per la funzione densità (per $y > 0$),

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}[2\,\Phi(\sqrt{y}) - 1] = \frac{1}{\sqrt{y}} f_X(\sqrt{y}) = \frac{1}{\sqrt{2\pi y}} \mathrm{e}^{-\frac{y}{2}}.$$

Prendiamo ora due normali standard tra loro indipendenti X_1 e X_2 e siano $Y_1 = X_1^2$ e $Y_2 = X_2^2$ i loro quadrati. Qual è la legge della variabile aleatoria $Z = Y_1 + Y_2$?

Abbiamo

$$F_Z(z) = P(X_1^2 + X_2^2 \leqslant z) = \iint_A f_{X_1,X_2}(x_1,x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2 = \iint_A \frac{1}{2\pi} \mathrm{e}^{-\frac{1}{2}(x_1^2+x_2^2)}\,\mathrm{d}x_1\,\mathrm{d}x_2$$

in cui abbiamo usato nell'ultima uguaglianza l'indipendenza tra X_1 e X_2. Qual è il dominio di integrazione A? È l'insieme dei punti (x_1, x_2) del piano $\mathbb{R}^2$ tali che $x_1^2 + x_2^2 \leqslant z$, ossia (per $z \geqslant 0$)

il cerchio centrato in $(0,0)$ e di raggio $\sqrt{z}$. Allora (passando a coordinate polari, ossia raggio e angolo)

$$F_Z(z) = \iint_A \frac{1}{2\pi} e^{-\frac{1}{2}r^2} r \, d\vartheta \, dr = \int_0^{\sqrt{z}} \frac{1}{2\pi} r e^{-\frac{1}{2}r^2} 2\pi \, dr = \left[-e^{-\frac{1}{2}r^2}\right]_0^{\sqrt{z}} = 1 - e^{-\frac{z}{2}}$$

e $f_Z(z) = F'_Z(z) = \frac{1}{2} e^{-\frac{z}{2}}$ per $z \geqslant 0$, ossia $Z \sim \exp\left(\frac{1}{2}\right)$.

DEFINIZIONE 10.22. *Se una variabile aleatoria X è la somma dei quadrati di n variabili aleatorie Gaussiane standard indipendenti, la chiamiamo* chi quadro con n gradi di libertà *(spesso indicati con* df*) e la indichiamo con $X \sim \chi^2(n)$ o $X \sim \chi^2_n$.*

Non è semplicissimo ricavare la densità di una chi quadro con n gradi di libertà, tuttavia se $X \sim \chi^2_n$, allora $f_X(x) = c_n x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$, dove c_n è un'opportuna costante di rinormalizzazione (che dipende da n). In Figura 10.9 possiamo vedere le funzioni densità di alcune chi quadro, al variare dei gradi di libertà: il loro comportamento cambia abbastanza e, al crescere di n, assomiglia sempre di più a quello di una Gaussiana.

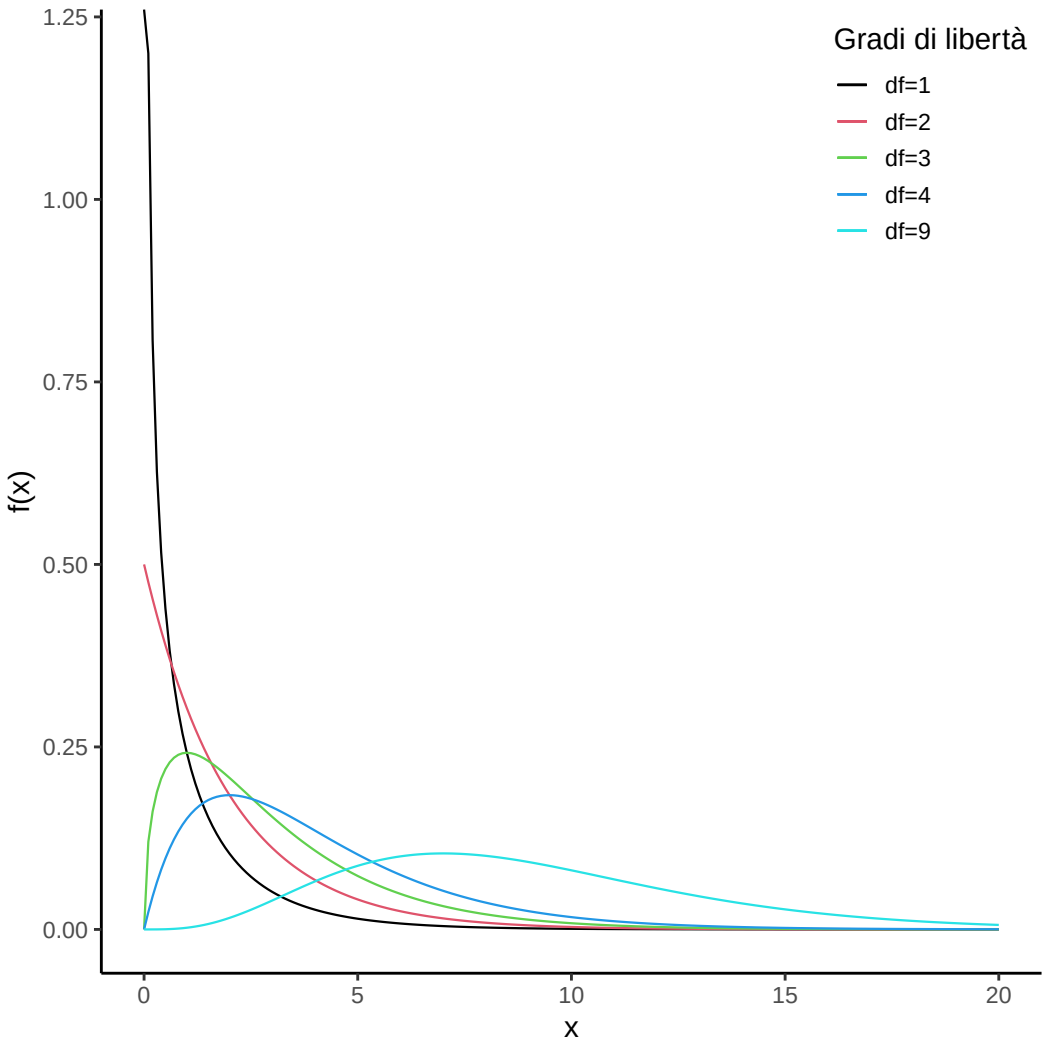


Figura 10.9. Densità di alcune chi quadro, al variare dei gradi di libertà

10.4.1. Chi quadro in R

Non abbiamo detto esplicitamente chi siano la funzione di ripartizione e la funzione quantile di una χ^2_n, perché la loro forma non è semplice. Come già visto nel caso della normale, ci appoggiamo alle tavole o alle funzioni di R.

La famiglia delle chi quadro in R è `chisq`. La densità è `dchisq(x, df)`, in cui è necessario specificare il numero di gradi di libertà `df`. Per la funzione di ripartizione abbiamo `pchisq(q, df, lower.tail = TRUE)` in cui dobbiamo prestare attenzione a passare il parametro `lower.tail` sempre con il nome, dal momento che la funzione prevede un ulteriore parametro (`ncp = 0`, del quale non ci interessiamo) tra `df` e `lower.tail`. Per la funzione quantile (che ci sarà molto utile in statistica) abbiamo `qchisq(p, df, lower.tail = TRUE)`, con la stessa accortezza vista per `pchisq`. Il generatore casuale è `rchisq(n, df)`.

10.4.2. Indicatori delle chi quadro

Per definizione le chi quadro sono riproducibili: se $X \sim \chi^2_n$ e $Y \sim \chi^2_m$, allora $X+Y \sim \chi^2_{n+m}$. Questo ci aiuta nel calcolo dei momenti, permettendoci di calcolarli in modo ricorsivo. Cominciamo dunque dal caso $n=1$. Abbiamo $X \sim \chi^2_1$, cioè $X = Z^2$, con $Z \sim \mathcal{N}(0,1)$. Allora $E[X] = E[Z^2] = \mathrm{Var}[Z] = 1$ e

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = E[Z^4] - 1 = \mathrm{kr}[Z] - 1 = 2.$$

Se ora $Y \sim \chi^2_n$, allora $Y = \sum_{i=1}^n X_i$ con le X_i indipendenti e identicamente distribuite $X_i \sim \chi^2_1$. Allora

$$E[Y] = E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E[X_i] = n$$

e per la varianza, grazie all'indipendenza,

$$\mathrm{Var}[Y] = \mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathrm{Var}[X_i] = 2n.$$

10.5. t DI STUDENT

DEFINIZIONE 10.23. *Diciamo che una variabile aleatoria X è distribuita come una t* di Student con n gradi di libertà *se esistono $Z \sim \mathcal{N}(0,1)$ e $W \sim \chi^2_n$ indipendenti tali che $X = \frac{Z}{\sqrt{W/n}}$. Scriviamo in questo caso $X \sim t(n)$ o $X \sim t_n$.*

È una variante della normale standard, ma con le code molto più pesanti (ossia la probabilità di essere lontani dal centro è maggiore). Anche in questo caso (come per le normali e per le chi quadro) la legge è abbastanza difficile da scrivere esplicitamente, ma possiamo usare le tavole oppure R. Possiamo però osservare che le t ereditano la simmetria delle normali standard, quindi se $X \sim t_n$, allora $f_X(-x) = f_X(x)$, $F_X(-x) = 1 - F_X(x)$ e, per la funzione quantile, $F_X^{-1}(p) = -F_X^{-1}(1-p)$. Anche per le t, come per le Gaussiane, nelle tavole sono riportati solo "metà" dei valori.

In Figura 10.10 vediamo le densità di alcune t di Student al variare del numero di gradi di libertà. Possiamo osservare che al crescere di n il comportamento è sempre più simile a quello di una Gaussiana.

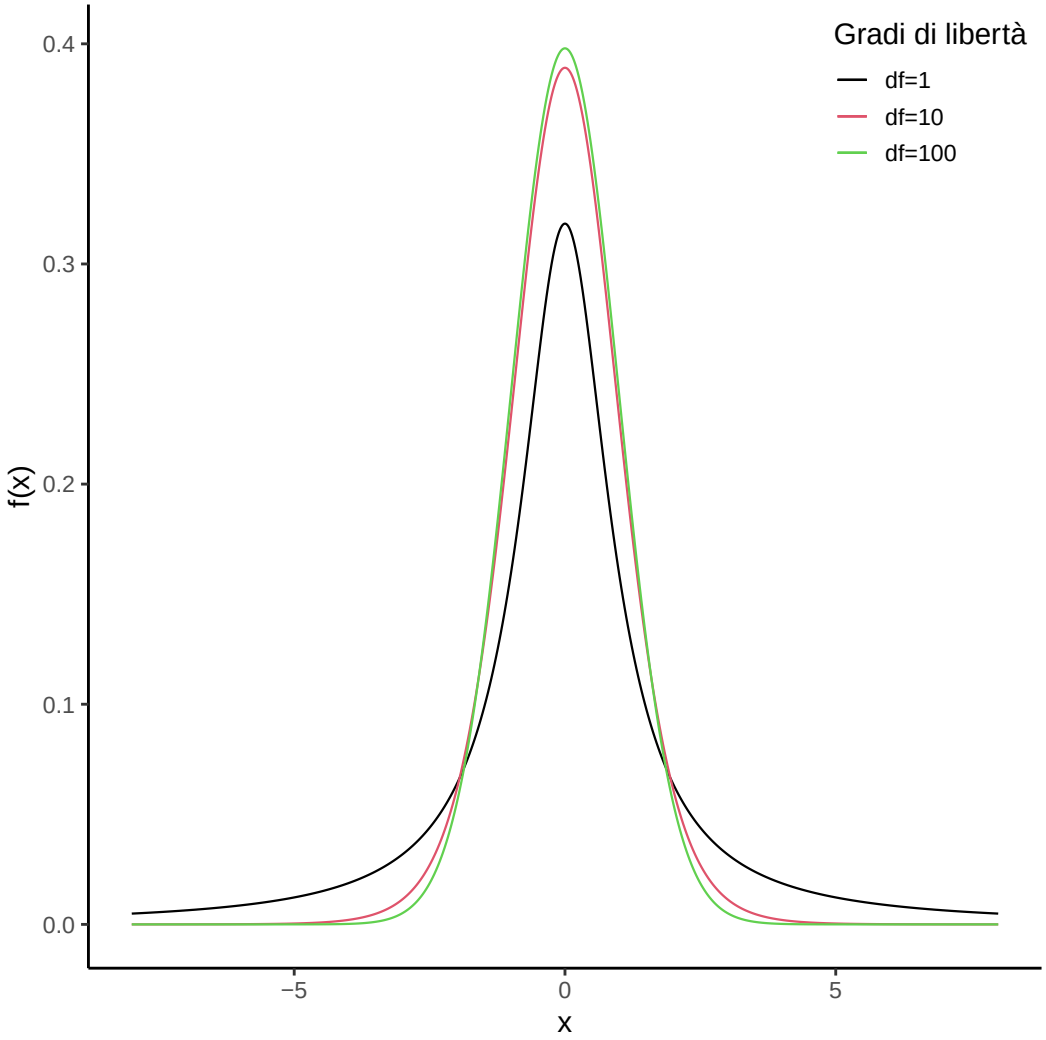


Figura 10.10. Densità di alcune t di Student, al variare dei gradi di libertà

Osserviamo che la speranza di $X \sim t_n$ è $E[X] = 0$ per simmetria rispetto all'origine. La varianza non è definita per $n = 1$, è infinita per $n = 2$, mentre per $n > 2$ è $\frac{n}{n-2}$.

Osservazione 10.24. Dalla definizione se $X \sim t_n$, allora $X = \frac{Z}{\sqrt{W/n}}$ con Z normale standard e W chi quadro con n gradi di libertà. Possiamo allora studiare (anche se per ora solo qualitativamente) il comportamento di $\frac{W}{n}$ al crescere di n: $E\left[\frac{W}{n}\right] = 1$, fatto che non ci stupisce, visto che $E[W] = n$. Inoltre $\mathrm{Var}\left[\frac{W}{n}\right] \to 0$ al tendere di n a $+\infty$, quindi possiamo dire con più confidenza che al crescere del numero n dei gradi di libertà t_n tende in qualche senso a una normale standard.

Questo entra in gioco nella consultazione delle tavole: la Gaussiana non ha la sua tavola delle funzioni quantile, ma compare in quella delle t di Student come caso con infiniti gradi di libertà.

10.5.1. t di Student in R

Come già detto le funzioni in R sono abbastanza cruciali per poter manipolare le t di Student, dal momento che non abbiamo una forma esplicita della legge. La famiglia delle t in R è `t`. La densità è `dt(x, df)`, in cui è necessario specificare il numero di gradi di libertà `df`. Per la funzione di ripartizione abbiamo `pt(q, df, lower.tail = TRUE)` in cui dobbiamo prestare attenzione a passare il parametro `lower.tail` sempre con il nome, dal momento che la funzione prevede un ulteriore parametro (`ncp`, del quale non ci interessiamo) tra `df` e `lower.tail`. Per

la funzione quantile (che ci sarà molto utile in statistica) abbiamo `qt(p, df, lower.tail = TRUE)`, con la stessa accortezza vista per `pt`. Il generatore casuale è `rt(n, df)`.

CAPITOLO 11

TEOREMI LIMITE

Lezione 20

In questo capitolo vogliamo dare un significato rigoroso a un concetto che abbiamo toccato in precedenza: data una successione $(X_n)_{n\in\mathbb{N}}$ di variabili aleatorie su uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, cosa significa dire che $\lim_{n\to+\infty} X_n = X$, ossia passare al limite? Come vedremo, ci sono diverse nozioni di convergenza di variabili aleatorie.

Una volta viste queste nozioni vedremo alcuni risultati (i teoremi limite) che ci garantiscono sotto opportune ipotesi, la convergenza di alcune particolari successioni di variabili aleatorie.

11.1. CONVERGENZA DI VARIABILI ALEATORIE

DEFINIZIONE 11.1. *Siano $(\Omega, \mathcal{F}, P)$ uno spazio di probabilità, X una variabile aleatoria su tale spazio e $(X_n)_{n\in\mathbb{N}}$ una successione di variabili aleatorie sullo stesso spazio. Diciamo che* $(X_n)_n$ converge quasi certamente *(o* puntualmente*) a X e scriviamo $X_n \xrightarrow[n\to+\infty]{q.c.} X$ se esiste un evento $E \in \mathcal{F}$ con $P(E) = 1$ tale che per ogni esito $\omega \in E$, $\lim_{n\to+\infty} X_n(\omega) = X(\omega)$.*

Osserviamo che il limite $\lim_{n\to+\infty} X_n(\omega) = X(\omega)$ è il limite di una successione di numeri reali. Per indicare la convergenza quasi certa possiamo anche usare la scrittura $P(\lim_{n\to+\infty} X_n = X) = 1$.

Osservazione 11.2. Il concetto di convergenza quasi certa è molto forte: stiamo chiedendo che la successione di funzioni converga puntualmente per quasi ogni $\omega \in \Omega$. È un tipo di convergenza molto difficile da verificare direttamente.

DEFINIZIONE 11.3. *Siano $(X_n)_{n\in\mathbb{N}}$ una successione di variabili aleatorie e X una variabile aleatoria sul medesimo spazio di probabilità $(\Omega, \mathcal{F}, P)$. Diciamo che* $(X_n)_n$ converge in probabilità *a X e scriviamo $X_n \xrightarrow[n\to+\infty]{P} X$ se, per ogni $\varepsilon > 0$, $\lim_{n\to+\infty} P(|X_n - X| > \varepsilon) = 0$.*

Anche in questo caso ci siamo ricondotti al limite di una successione di numeri reali, ma in modo diverso: ogni $|X_n - X|$ è una variabile aleatoria, di cui chiediamo la probabilità di essere maggiore di ε, probabilità che è un numero reale (tra 0 e 1).

Osservazione 11.4. A differenza della convergenza quasi certa, la convergenza in probabilità guarda il comportamento globale della successione di variabili aleatorie. Dobbiamo infatti controllare che gli esiti $\omega \in \Omega$ per cui $|X_n(\omega) - X(\omega)| > \varepsilon$ siano un insieme di probabilità che, al tendere di n all'infinito, converge a 0.

DEFINIZIONE 11.5. *Siano $(X_n)_{n\in\mathbb{N}}$ una successione di variabili aleatorie e X una variabile aleatoria sul medesimo spazio di probabilità $(\Omega, \mathcal{F}, P)$. Diciamo che* $(X_n)_n$ converge in media quadratica *(o in L^2) a X e scriviamo $X_n \xrightarrow[n\to+\infty]{L^2} X$ se $\lim_{n\to+\infty} E[|X_n - X|^2] = 0$.*

Ancora una volta abbiamo espresso una convergenza di variabili aleatorie (e quindi di funzioni) in termini di una convergenza di numeri reali: ogni $|X_n - X|$ è una variabile aleatoria, di cui chiediamo che convergano i momenti secondi (che sono numeri reali).

PROPOSIZIONE 11.6. *La convergenza in media quadratica implica la convergenza in probabilità, ossia se $X_n \xrightarrow[n\to+\infty]{L^2} X$, allora $X_n \xrightarrow[n\to+\infty]{P} X$.*

Dimostrazione. Prendiamo $\varepsilon > 0$. Dalla disuguaglianza di Markov (9.1) abbiamo per ogni n

$$\begin{aligned} P(|X_n - X| \geqslant \varepsilon) &= P(|X_n - X|^2 \geqslant \varepsilon^2) \\ &\leqslant \frac{E[|X_n - X|^2]}{\varepsilon^2}. \end{aligned}$$

Ora possiamo prendere il limite per $n \to +\infty$:

$$\lim_{n\to+\infty} P(|X_n - X| > \varepsilon) \leqslant \frac{\lim_{n\to+\infty} E[|X_n - X|^2]}{\varepsilon^2} = 0$$

in cui l'ultima uguaglianza è garantita dalla convergenza in media quadratica. □

PROPOSIZIONE 11.7. *La convergenza quasi certa implica la convergenza in probabilità, ossia se* $X_n \xrightarrow[n\to+\infty]{q.c.} X$, *allora* $X_n \xrightarrow[n\to+\infty]{P} X$.

Dimostrazione. [TBA] □

Osservazione 11.8. Viene naturale, a questo punto, chiedersi quale sia "più forte" tra le convergenze in L^2 e quasi certa, ossia se ce ne sia una delle due che implica l'altra. In realtà si può mostrare che le due convergenze non sono confrontabili: esistono successioni di variabili aleatorie che convergono quasi certamente ma non in L^2 e, viceversa, successioni che convergono in L^2 ma non quasi certamente.

DEFINIZIONE 11.9. *Siano* $(X_n)_{n\in\mathbb{N}}$ *una successione di variabili aleatorie su uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$ *e* X *una variabile aleatoria sullo spazio di probabilità* $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$. *Diciamo che* $(X_n)_n$ converge in legge (o in distribuzione o debolmente) *a* X *e scriviamo* $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$ *o* $X_n \xrightarrow[n\to+\infty]{d} X$, *se per ogni* $x \in \mathbb{R}$ $\lim_{n\to+\infty} P(X_n \leqslant x) = P(X \leqslant x)$, *ossia se* $\lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$.

Osservazione 11.10. Questa nozione è chiaramente più debole della convergenza in probabilità (e quindi delle altre due). In particolare non è necessario che la successione $(X_n)_n$ e il suo limite X siano nello stesso spazio di probabilità, come sottolineato nella definizione prendendo $(\Omega, \mathcal{F}, P)$ e $(\tilde{\Omega}, \tilde{\mathcal{F}}, \tilde{P})$.

PROPOSIZIONE 11.11. *La convergenza in probabilità implica la convergenza in legge, ossia se* $X_n \xrightarrow[n\to+\infty]{P} X$, *allora* $X_n \xrightarrow[n\to+\infty]{\mathcal{L}} X$.

Dimostrazione. [TBA] □

Osservazione 11.12. Possiamo riassumere i legami tra i vari concetti di convergenza di variabili aleatorie con lo schema in Figura 11.1. Ci sono casi in cui è possibile invertire le implicazioni, sotto opportune ipotesi, ma vanno oltre i contenuti di questo corso.

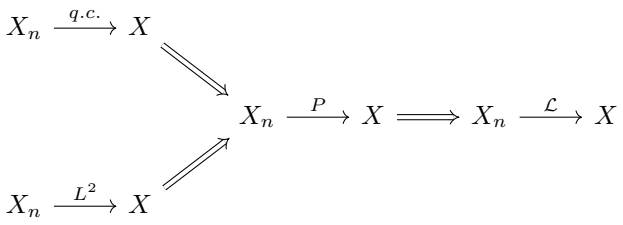


Figura 11.1. Gerarchia delle convergenze di variabili aleatorie

11.2. TEOREMI LIMITE

Cominciamo con qualche richiamo di risultati già visti.

PROPOSIZIONE 11.13. *Siano* $X_1, \ldots, X_n$ *variabili aleatorie indipendenti di media comune* μ *e di varianza comune* σ^2. *Sia inoltre* S_n *la variabile aleatoria somma,* $S_n = \sum_{i=1}^n X_i$. *Allora*

$$E\left[\frac{S_n}{n}\right]=\mu \qquad e \qquad \mathrm{Var}\left[\frac{S_n}{n}\right]=\frac{\sigma^2}{n}.$$

Dimostrazione. Sappiamo che la speranza è lineare, quindi (senza necessità dell'ipotesi di indipendenza)

$$E\left[\frac{S_n}{n}\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n}\sum_{i=1}^{n}E[X_i]=\mu.$$

Per la varianza abbiamo invece bisogno dell'indipendenza,

$$\mathrm{Var}\left[\frac{S_n}{n}\right]=\frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n}X_i\right]=\frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i]=\frac{\sigma^2}{n},$$

concludendo la dimostrazione. □

Osservazione 11.14. È interessante notare come si comportino i risultati della Proposizione 11.13 al crescere di n: la speranza $E\left[\frac{S_n}{n}\right]$ converge a μ per $n\to+\infty$ (addirittura è costantemente uguale a μ per ogni n), mentre la varianza $\mathrm{Var}\left[\frac{S_n}{n}\right]$ converge a 0 per $n\to+\infty$. Abbiamo allora per ogni n una variabile aleatoria $\frac{S_n}{n}$ che mantiene il suo centro in μ e che si restringe sempre di più, fino a essere costantemente uguale alla sua media al limite.

È arrivato il momento di uno dei risultati di probabilità più citati (solitamente a sproposito), che rende rigoroso (in termini di convergenza di variabili aleatorie) quanto detto nell'Osservazione 11.14.

TEOREMA 11.15. (LEGGE DEBOLE DEI GRANDI NUMERI) *Sia $(X_n)_{n\in\mathbb{N}}$ una successione di variabili aleatorie indipendenti, ciascuna di media μ e varianza finita σ^2. Sia inoltre $S_n=\sum_{i=1}^{n}X_i$ la variabile aleatoria somma parziale delle X_i. Allora la variabile aleatoria $\frac{S_n}{n}$ converge in probabilità a μ, ossia per ogni $\varepsilon>0$*

$$\lim_{n\to+\infty}P\left(\left|\frac{S_n}{n}-\mu\right|>\varepsilon\right)=0.$$

Dimostrazione. Sfruttiamo la Proposizione 11.13 e la disuguaglianza di Chebychev (9.2): sia infatti $\varepsilon>0$, allora

$$\begin{aligned}P\left(\left|\frac{S_n}{n}-\mu\right|>\varepsilon\right) &= P\left(\left|\frac{S_n}{n}-E\left[\frac{S_n}{n}\right]\right|>\varepsilon\right)\\ &\leqslant \frac{\mathrm{Var}\left[\frac{S_n}{n}\right]}{\varepsilon^2}\\ &= \frac{\sigma^2}{n\varepsilon^2}.\end{aligned}$$

Passando al limite per $n\to+\infty$, l'ultimo termine converge a 0. □

Osservazione 11.16. Il fatto che il Teorema 11.15 si chiami "Legge *debole* dei grandi numeri" suggerisce che ci siano altri enunciati, più forti. Così è, in effetti: esiste anche la legge *forte* dei grandi numeri che dà sotto ipotesi meno restrittive un risultato più forte, ossia garantisce la convergenza quasi certa (che, come abbiamo visto nella Proposizione 11.7, implica in particolare la convergenza in probabilità). In questo corso dovremo però accontentarci della legge debole dei grandi numeri, senza enunciare (o dimostrare) altre varianti.

Vediamo ora cosa dice (e cosa *non* dice) la legge debole dei grandi numeri. Prendiamo, come esempio guida, un processo di Bernoulli di parametro $\frac{1}{2}$, ossia infiniti lanci consecutivi di una moneta bilanciata. Le X_i sono indipendenti e identicamente distribuite, $X_i\sim\mathrm{bin}\left(1,\frac{1}{2}\right)$. Inoltre la variabile aleatoria "somma parziale" S_n conta il numero di 1 (ossia di successi) nei primi n lanci, quindi è una binomiale di parametri n e $p=\frac{1}{2}$: $S_n\sim\mathrm{bin}\left(n,\frac{1}{2}\right)$. La legge debole dei grandi numeri ci dice che $\frac{S_n}{n}\xrightarrow[n\to+\infty]{P}\frac{1}{2}=E[X_1]$.

Questo risultato viene spesso (erroneamente) letto come

$$S_n \sim \frac{n}{2} \qquad \text{o, peggio,} \qquad S_n \longrightarrow \frac{n}{2}.$$

Entrambe queste scritture dovrebbero insospettirci in partenza: non sono precise (nel primo caso) o non hanno proprio senso (nel secondo caso: se stiamo passando al limite non può esserci un n dopo il limite).

Cerchiamo di scrivere meglio la prima, $S_n \sim \frac{n}{2}$, in modo che abbia più significato. Abbiamo

$$\frac{S_n}{n} \xrightarrow[n\to+\infty]{P} \frac{1}{2} \iff \frac{S_n}{n} - \frac{1}{2} \xrightarrow[n\to+\infty]{P} 0 \iff \frac{S_n - \frac{n}{2}}{n} \xrightarrow[n\to+\infty]{P} 0.$$

Non dobbiamo fraintendere l'ultima leggendola come $S_n - \frac{n}{2} \to 0$: questa è falsa, non in modo grossolano come la $S_n \longrightarrow \frac{n}{2}$, ma falsa ugualmente. Infatti quello che noi sappiamo dalla legge debole dei grandi numeri è che $S_n - \frac{n}{2}$ cresce più lentamente di n, non che decresce. Anzi, se facciamo qualche esperimento, possiamo vedere che la quantità $S_n - \frac{n}{2}$ *cresce* al crescere di n (all'incirca come $\sqrt{n}$, come vedremo tra poco).

La moneta che stiamo lanciando non ha idea di cosa sia uscito, quindi non cerca di bilanciare il *numero* di teste e croci (ossia di mandare $S_n - \frac{n}{2}$ a 0), ma bilancia la *frequenza* sul totale: il rapporto di teste sul totale dei lanci tende a $\frac{1}{2}$, ma sono possibili sbilanciamenti molto ampi sul numero. Vediamo una rappresentazione di questa situazione nella Figura 11.2.

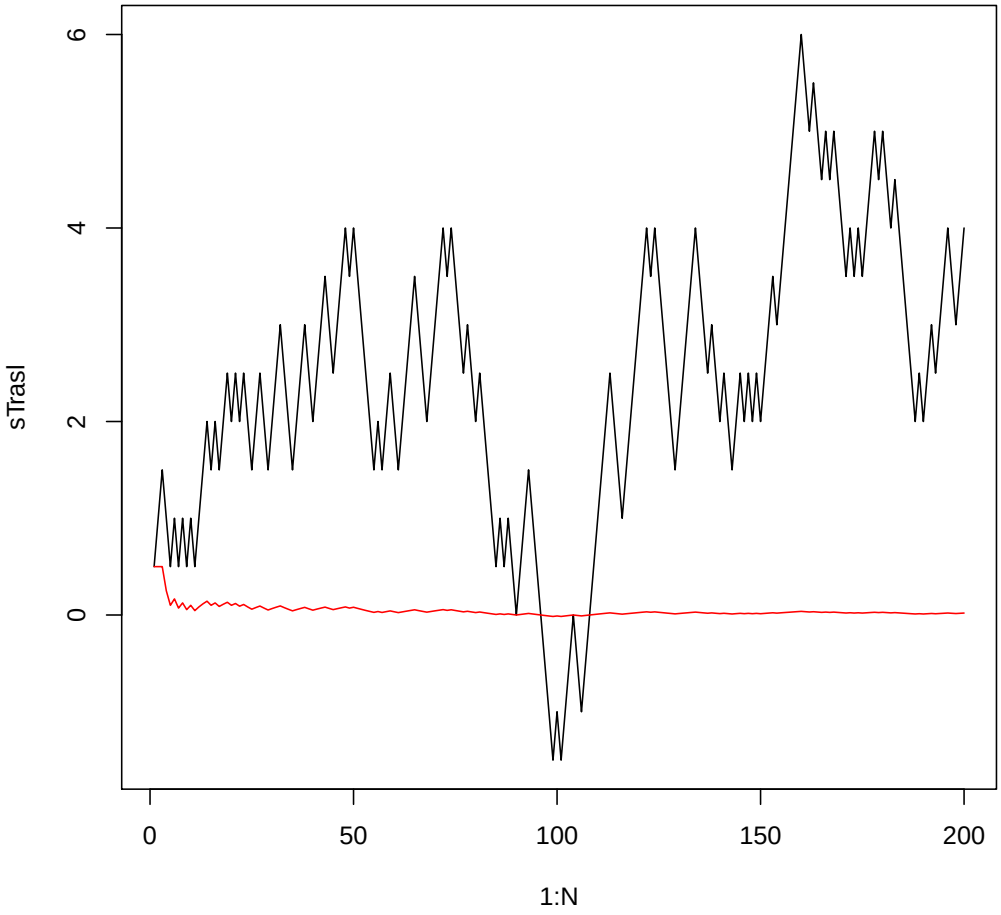


Figura 11.2. Una realizzazione di 200 lanci di una moneta. In nero la quantità $S_n - \frac{n}{2}$, che oscilla e non converge, in rosso $\frac{S_n - \frac{n}{2}}{n}$ che converge (molto rapidamente) a 0.

Vediamo anche il codice usato per generare la Figura 11.2:

```
N <- 200 # lunghezza dei vettori
x <- rbinom(N, size = 1, prob = 1/2) # lanci della moneta
uni <- rep(1, N) # N-vettore di soli 1
M <- matrix(1, nrow = N, ncol = N) # matrice NxN di soli 1
M[upper.tri(M)] <- 0 # che trasformiamo in una matrice
                     # triangolare inferiore di soli 1
                     # (diagonale inclusa)
s <- M %*% x # vettore dei valori di Sn, ottenuto mediante
             # moltiplicazione di matrici (e vettori)
sTrasl <- s - 1/2 * M %*% uni # vettore Sn-n/2, di nuovo via
                              # moltiplicazione di matrici
# Senza passare da s avremmo potuto scrivere
# sTrasl <- M %*% (x - 1/2*uni)
plot(1:N, sTrasl, type = "l")
lines(1:N,sTrasl/(M%*%uni), col = "red")
```

in cui abbiamo usato una rappresentazione geometrica (matrici) per evitare cicli `for`.

Esempio 11.17. Un numero al Superenalotto in media uscirà ogni $\frac{90}{6} = 15$ estrazioni. Infatti possiamo vedere la successione di estrazioni come un processo di Bernoulli (come già visto nell'Esempio 8.15) in cui a ogni estrazione abbiamo probabilità $\frac{6}{90}$ di successo (ossia di vedere uscire il numero scelto). Sappiamo anche che se prendiamo n estrazioni, ci aspettiamo in media $n \cdot \frac{6}{90}$ successi. Noi vogliamo trovare n per cui abbiamo in media 1 successo, quindi $n = \frac{90}{6} = 15$.

Questo però non significa che se il nostro numero manca (o "ritarda") da un po' allora è "più probabile che esca, per la legge dei grandi numeri". La probabilità non cambia (di nuovo, come visto nell'Esempio 8.15), quello che succede per la legge dei grandi numeri è che la *frequenza* con cui il nostro numero esce tenderà a $\frac{1}{15}$.

TEOREMA 11.18. (TEOREMA CENTRALE DEL LIMITE) *Sia $(X_n)_{n\in\mathbb{N}}$ una successione di variabili aleatorie indipendenti, ciascuna di media μ e varianza finita σ^2. Sia inoltre $S_n = \sum_{i=1}^{n} X_i$ la variabile aleatoria somma parziale delle X_i. Allora*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow[n\to+\infty]{\mathcal{L}} \mathcal{N}(0,1) \qquad \textit{cioè} \qquad \lim_{n\to+\infty} P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leqslant x\right) = \Phi(x).$$

Dimostrazione. [TBA] □

Il teorema centrale del limite ci dice che $\frac{S_n - n\mu}{\sigma\sqrt{n}}$ converge in legge (o distribuzione) a una normale standard. Qualunque sia la distribuzione originaria di ciascuna delle X_i, giustificando l'importanza della distribuzione normale[11.1]. Possiamo però leggerci qualcosa di più: abbiamo che (la variabile aleatoria) $|S_n - \frac{n}{2}|$ va all'infinito come $\sqrt{n}$. Non può andare più rapidamente, altrimenti avremmo un'esplosione all'infinito, ossia una distribuzione limite con varianza infinita, ma nemmeno più lentamente, altrimenti la distribuzione limite sarebbe concentrata in 0 con varianza nulla.

Vediamo, nella Tabella 11.1, l'andamento di alcune grandezze al crescere di n, nell'esempio guida del processo di Bernoulli con la moneta bilanciata.

11.1. E il suo nome: si chiama normale perché è la norma.

n	$\sqrt{n}$	$E[S_n]=\frac{n}{2}$	$S_n-\frac{n}{2}$	$\frac{S_n-\frac{n}{2}}{n}$
10	3.16...	5	$(-3.16, 3.16)$	$\left(-\frac{1}{3}, \frac{1}{3}\right)$
100	10	50	$(-10, 10)$	$\left(-\frac{1}{10}, \frac{1}{10}\right)$
10^4	100	5000	$(-100, 100)$	$\left(-\frac{1}{100}, \frac{1}{100}\right)$
10^8	10^4	$5\cdot 10^7$	$(-10^4, 10^4)$	$\left(-\frac{1}{10^4}, \frac{1}{10^4}\right)$

Tabella 11.1. Confronto tra gli ordini di grandezza degli intervalli in cui assumono valore $S_n-\frac{n}{2}$ e $\frac{S_n-\frac{n}{2}}{n}$ al crescere di n, usando il comportamento asintotico di $|S_n-\frac{n}{2}|$ dell'ordine di $\sqrt{n}$.

Osservazione 11.19. Nella pratica non useremo sostanzialmente mai il teorema centrale del limite come limite, ossia usando quanto visto nell'enunciato

$$\lim_{n\to+\infty} P\left(\frac{S_n-n\mu}{\sigma\sqrt{n}}\leqslant x\right)=\Phi(x).$$

Infatti non avremo mai infinite realizzazioni di un esperimento (e quindi infinite variabili aleatorie X_i di cui fare la somma).

Il teorema servirà invece per avere delle approssimazioni: quando n è "sufficientemente grande" abbiamo

$$P\left(\frac{S_n-n\mu}{\sigma\sqrt{n}}\leqslant x\right)\approx\Phi(x).$$

Scriviamo anche, in questo caso,

$$\frac{S_n-n\mu}{\sigma\sqrt{n}}\dot{\sim}\mathcal{N}(0,1)$$

per dire che la distribuzione di $\frac{S_n-n\mu}{\sigma\sqrt{n}}$ è approssimativamente normale standard. Possiamo riscrivere questa distribuzione approssimata anche come

$$\frac{S_n}{n}\dot{\sim}\mathcal{N}\left(\mu,\frac{\sigma}{\sqrt{n}}\right) \qquad \text{oppure} \qquad S_n\dot{\sim}\mathcal{N}(n\mu,\sigma\sqrt{n}).$$

Resta però una domanda: quand'è che n è sufficientemente grande? Lasciamola un momento da parte e vediamo un esempio di applicazione del teorema centrale del limite.

Esempio 11.20. (Baldi) Un calcolatore somma 10^6 numeri, con un errore di arrotondamento in ciascuna operazione. I singoli errori sono indipendenti e hanno distribuzione uniforme sull'intervallo $[-0.5\cdot 10^{-10}, 0.5\cdot 10^{-10}]$. Qual è la probabilità che l'errore assoluto finale sia minore di $0.5\cdot 10^{-7}$?

Come prima cosa formuliamo il problema in termini di variabili aleatorie: per $i\in\{1,\dots,n\}$ abbiamo $X_i\sim\text{unif}(-0.5\cdot 10^{-10}, 0.5\cdot 10^{-10})$, tutte indipendenti che rappresentano gli errori fatti in ciascuna operazione. Abbiamo inoltre $S_n=\sum_{i=1}^n X_i$ che rappresenta l'errore totale e, dai dati del problema, sappiamo anche $n=10^6$.

Ora pensiamo un momento a cosa ci interessa: non vogliamo calcolare la legge "vera" di S_n, ma vogliamo calcolare

$$P(|S_n|<0.5\cdot 10^{-7})=P(S_n<0.5\cdot 10^{-7})-P(S_n<-0.5\cdot 10^{-7})$$

e siamo quindi interessati a (un'approssimazione di) $P(S_n\leqslant y)$, per qualche y. Dal teorema centrale del limite (nell'enunciato approssimato) sappiamo che

$$\frac{S_n-nE[X_1]}{\sqrt{n}\sqrt{\text{Var}[X_1]}}\dot{\sim}\mathcal{N}(0,1),$$

ossia che

$$P\left(\frac{S_n-nE[X_1]}{\sqrt{n}\sqrt{\text{Var}[X_1]}}\leqslant x\right)\simeq\Phi(x). \tag{11.1}$$

Cerchiamo di riscrivere (11.1) in modo da mettere in evidenza S_n e ottenere qualcosa della forma $P(S_n \leqslant y)$: manipolando il primo membro della (11.1) abbiamo

$$P(S_n \leqslant x \cdot \sqrt{n}\sqrt{\mathrm{Var}[X_1]} + n\,E[X_1]) \simeq \Phi(x),$$

quindi vogliamo determinare x tale che

$$x \cdot \sqrt{n}\sqrt{\mathrm{Var}[X_1]} + n\,E[X_1] = y,$$

cioè

$$x = \frac{y - n\,E[X_1]}{\sqrt{n}\sqrt{\mathrm{Var}[X_1]}}.$$

Saremmo potuti arrivare allo stesso risultato ricordando che $S_n \dot{\sim} \mathcal{N}(y - n\,E[X_1], \sqrt{n}\sqrt{\mathrm{Var}[X_1]})$ e usando le proprietà di standardizzazione di una Gaussiana, per cui

$$P(S_n \leqslant y) = F_{S_n}(y) \simeq \Phi\left(\frac{y - n\,E[X_1]}{\sqrt{n}\sqrt{\mathrm{Var}[X_1]}}\right).$$

Quanto fatto finora non usa il contesto specifico del problema che stiamo considerando, ma ora andiamo a sostituire i valori specifici:

$$n = 10^6, \qquad E[X_1] = 0, \qquad \mathrm{Var}[X_1] = \frac{10^{-20}}{12}, \qquad y = \pm 0.5 \cdot 10^{-7},$$

quindi

$$\begin{aligned} P(|S_n| < 0.5 \cdot 10^{-7}) &= P(S_n < 0.5 \cdot 10^{-7}) - P(S_n < -0.5 \cdot 10^{-7}) \\ &= F_{S_n}(0.5 \cdot 10^{-7}) - F_{S_n}(-0.5 \cdot 10^{-7}) \\ &\simeq 2\,\Phi\left(\frac{0.5 \cdot 10^{-7}}{10^3 \cdot \frac{1}{\sqrt{12}} \cdot 10^{-10}}\right) - 1 \\ &\simeq 2\,\Phi(1.75) - 1 \\ &= 1.9108 - 1 \end{aligned}$$

e la probabilità cercata è all'incirca 91%.

Osservazione 11.21. Nell'Esempio 11.20 siamo passati da $P(S_n < y)$ a $P(S_n \leqslant y)$ senza porci troppi problemi, grazie al fatto che le variabili aleatorie coinvolte erano assolutamente continue. Se però sommiamo variabili aleatorie discrete un singolo punto può avere probabilità non nulla, quindi possiamo commettere errori (anche significativi) se non prestiamo attenzione nell'uso del teorema centrale del limite per le approssimazioni.

Per fortuna c'è un facile accorgimento (che prende il nome di *correzione di continuità*) che ci viene in aiuto in questo caso: se S_n è una somma di variabili aleatorie discrete, allora

$$F_{S_n}(x) \simeq \Phi\left(\frac{x + \frac{1}{2} - n\,E[X_1]}{\sqrt{n}\sqrt{\mathrm{Var}[X_1]}}\right),$$

in cui quel termine $\frac{1}{2}$ che compare al numeratore in Φ è la correzione di continuità.

Torniamo alla domanda che ci eravamo posti prima: quanto deve essere grande n per avere una buona approssimazione? La risposta non è unica e dipende dalla distribuzione delle X_i, in particolare dalla loro "forma". Vediamo alcuni casi:

- $X_i \sim \mathcal{N}$: in questo caso $n = 1$, grazie alla riproducibilità
- $X_i \sim$ unif: in questo caso $n \geqslant 5$ dà di solito buoni risultati
- $X_i \sim$ exp o $X_i \sim$ geom: abbiamo bisogno di $n \geqslant 15$ (sono molto dissimili da delle normali)
- $X_i \sim \chi^2$: possiamo usare la riproducibilità $\chi^2_n \dot{\sim} \mathcal{N}(n, \sqrt{2n})$ e l'approssimazione è buona per $n \geqslant 25$ quindi se sommiamo $\chi^2(1)$ ne occorrono almeno 25, se sommiamo $\chi^2(9)$ ne bastano circa 3.

Abbiamo lasciato fuori due casi importanti, che meritano di essere considerati a parte: binomiale e Poisson.

Binomiale. È necessario che la distribuzione non sia troppo sbilanciata, quindi che p sia "lontano" dagli estremi 0 e 1. In tal caso possiamo usare il teorema centrale del limite per avere un'approssimazione della distribuzione stessa,

$$\text{bin}(n,p) \overset{\cdot}{\sim} \mathcal{N}(np, \sqrt{np(1-p)}).$$

La condizione su p (lontano dagli estremi) dipende da n, come regola di massima si chiede che $np(1-p) \gtrsim 3$.

Poisson. Anche in questo caso abbiamo la riproducibilità che ci viene in aiuto,

$$\text{Pois}(\lambda) \overset{\cdot}{\sim} \mathcal{N}\left(\lambda, \sqrt{\lambda}\right)$$

per $\lambda \gtrsim 30$. Possiamo infatti vedere una Poisson di parametro λ (ricordiamo che λ è sia la media sia la varianza, per una Poisson) come la somma di n Poisson di parametro $\tilde{\lambda} = \frac{\lambda}{n}$.

Parte II

Statistica

CAPITOLO 12

STIME PUNTUALI

Lezione 21

Con questo capitolo iniziamo l'esplorazione della Statistica, costruendo sulle fondamenta di Probabilità. Nello studio della Probabilità abbiamo usato un livello abbastanza alto di astrazione: le ipotesi alla base dei modelli erano "assolute" e alle volte impossibili da verificare in casi applicati. Abbiamo visto come calcolare probabilità e momenti, che sono tutti valori deterministici e certi.

In Statistica abbiamo invece dati *reali* e ipotesi ragionevoli (anche se approssimate, come ad esempio $X \sim \mathcal{N}$). Calcoliamo *stime* di parametri (o momenti o altre quantità) e facciamo verifiche della compatibilità delle ipotesi con in dati a nostra disposizione. Ma in questo caso i valori (ad esempio dei parametri) hanno margini di incertezza, non sono certi come lo erano in Probabilità.

12.1. INTRODUZIONE ALLA STATISTICA

Alla base della Probabilità avevamo lo spazio degli esiti, in Statistica questo ruolo è preso dalla popolazione.

DEFINIZIONE 12.1. *Chiamiamo* popolazione (di riferimento) *un insieme costituito da elementi distinti, sui quali conduciamo la nostra indagine. Chiamiamo tali elementi* esemplari, individui o unità statistiche.

Esempio 12.2. Sono esempi di popolazione di riferimento:

- la popolazione mondiale,
- gli animali ospitati in uno zoo,
- gli studenti che frequentano un corso,
- le aziende in una determinata provincia,
- i prodotti di uno stabilimento.

In Statistica siamo interessati alle misure di (alcune) caratteristiche degli individui, dette *dati*. Vogliamo usare questi dati per avere informazioni riguardo all'intera popolazione. Abbiamo però davanti a noi una biforcazione nella Statistica, proprio a questo punto: da un lato abbiamo la *Statistica descrittiva*, dall'altro la *Statistica inferenziale*. Di queste, la prima non chiama in gioco la Probabilità: abbiamo misure sull'intera popolazione e vogliamo *descrivere* alcune caratteristiche dell'intera popolazione a partire da queste misure, calcolandone opportune funzioni che riassumano tutte le informazioni in una quantità ridotta di numeri o indicatori.

La seconda, invece, entra in campo quando non abbiamo dati sull'intera popolazione, ma solamente su un suo sottoinsieme, detto *campione*. Vogliamo fare affermazioni sull'intera popolazione, ma abbiamo bisogno di ricavarle (o *inferirle*) dalle informazioni sul campione, usando opportuni modelli probabilistici.

La Statistica descrittiva e quella inferenziale hanno forti legami, ma sono ben distinte. Da un lato i metodi e le tecniche che si usano sono molto differenti, dall'altro quando ci restringiamo al campione (considerandolo come nuova popolazione) e calcoliamo funzioni delle grandezze misurate, stiamo facendo Statistica descrittiva.

DEFINIZIONE 12.3. *Chiamiamo* campione *un sottoinsieme della popolazione di riferimento.*

Il campione è la controparte statistica degli eventi (che erano opportuni sottoinsiemi dello spazio degli esiti). Per gli eventi chiedevamo fossero soddisfatte alcune ipotesi astratte, cioè che la loro famiglia fosse una tribù. Per i campioni abbiamo alcune richieste, che però non sono così rigidamente definite. Prima di arrivare a queste caratteristiche, però, cerchiamo di rispondere a una domanda: perché concentrarci su un campione? Una prima ragione è la praticità: la popolazione può essere molto grande, oppure le misurazioni che prendiamo richiedono la distruzione degli esemplari (pensiamo ad esempio ai crash test degli autoveicoli). Ci sono poi anche ragioni di costo e di etica.

Vorremmo che il campione del quale raccogliamo le misurazioni sia il più possibile rappresentativo della popolazione di riferimento, ma è difficile dare una definizione assoluta di cosa significhi rappresentativo. Ci sono anche modi diversi di scegliere un campione, nella pratica. Ciascun modo ha un costo (decrescente nell'elenco qui sotto) e caratteristiche peculiari:

- campionamento casuale semplice,
- campionamento casuale stratificato (nel quale vogliamo preservare alcune caratteristiche della popolazione),
- campionamento a grappoli (ad esempio se la popolazione sono gli scolari della Provincia Autonoma di Trento, i grappoli potrebbero esse singole classi che scegliamo a caso, ma come unità); può essere a uno o due stadi, a seconda che all'interno dei grappoli facciamo o meno un campionamento,
- campionamento selettivo,
- campionamento per convenienza o disponibilità,
- campionamento per quote (da non confondere con il campionamento stratificato).

DEFINIZIONE 12.4. *Le caratteristiche che misuriamo prendono il nome di* variabili, *i valori che assumono si chiamano* livelli o modalità.

Le variabili possono essere di tipo

- qualitativo o categorico, *se sono aggettivi o simili, in particolare*
  - nominale, *se non hanno un ordinamento naturale,*
  - ordinale, *se hanno un ordinamento naturale;*
- quantitativo o numerico, *se sono grandezze descritte da numeri, in particolare*
  - discreto, *se sono descritte da numeri interi,*
  - continuo, *se sono descritte da numeri reali.*

Nel caso di variabili numeriche, la scala di misurazione può essere di tipo

- intervallo, *se lo 0 è fissato in modo arbitrario,*
- rapporto, *se lo 0 è fissato in modo naturale.*

Esempio 12.5. Vediamo alcuni esempi di variabili dei diversi tipi.

Osservazione 12.6. Nell'ambito della Statistica inferenziale, possiamo identificare un esemplare con le misure a esso associate. In questo modo possiamo vedere la popolazione come la distribuzione (non nota) di una variabile aleatoria.

Esempio 12.7. Una ditta produce bulloni di 7 mm di diametro. Un bullone è accettabile se il suo diametro è compreso tra 6.5 mm e 7.5 mm.

Prendiamo un bullone e misuriamo il suo diametro effettivo. Possiamo vedere questo come un esperimento aleatorio e possiamo descrivere il diametro come una variabile aleatoria di densità f_X.

Il problema diventa allora come utilizzare le misurazioni del diametro di alcuni bulloni per inferire la distribuzione f_X e poter prendere decisioni sull'intera popolazione, come per esempio ricalibrare la macchina, qualora il diametro medio fosse troppo piccolo o troppo grande. Procedendo in questo modo stiamo vedendo le misurazioni fatte sul campione come variabili aleatorie indipendenti e identicamente distribuite. La distribuzione comune è la distribuzione (non nota) dell'intera popolazione.

DEFINIZIONE 12.8. *Una* statistica *è una funzione calcolabile a partire dalla misurazione del campione.*

Esempio 12.9. Sono esempi di statistiche calcolate per un campione $(x_1, \dots, x_n)$

1. la media campionaria $\bar{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$
2. la varianza campionaria a media μ nota: $s_*^2 := \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu)^2$
3. la varianza campionaria a media ignota: $s^2 := \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$
4. il numero di misurazioni eccedenti una certa soglia c: $\#\{i \in \{1, \dots, n\} : x_i > c\}$
5. il primo esemplare del campione per cui la misurazione è inferiore a una certa soglia c: $\inf\{i \in \{1, \dots, n\} : x_i < c\}$.

Osservazione 12.10. Vale la pena prestare attenzione alla notazione che useremo: se abbiamo delle quantità, dei numeri, usiamo in genere la lettera minuscola. La lettera maiuscola denota invece le variabili aleatorie, ossia funzioni di esemplari (ignoti) del campione. In generale, dunque, s^2 e S^2 indicheranno cose diverse: la prima sarà un numero, la seconda una variabile aleatoria, il risultato di un esperimento aleatorio.

Ci sono molti modi in cui la distribuzione f_X può essere ignota, ma li possiamo dividere in due categorie. Nella prima categoria il modello è noto a meno di parametri. Ad esempio, sappiamo che X è una variabile aleatoria di Poisson, ma non ne conosciamo il parametro λ. In questo caso il nostro obiettivo è stimare il parametro (o i parametri) a partire dai dati. Parliamo quindi di *Statistica parametrica*. Nella seconda categoria, invece, il modello è completamente ignoto: in questo caso parliamo di *Statistica non parametrica*. In questo corso ci occuperemo esclusivamente di Statistica parametrica.

Osservazione 12.11. Parliamo di *modelli* per la popolazione perché non ci aspettiamo di conoscere con certezza la realtà: un modello è una ragionevole astrazione o approssimazione della verità. Inoltre sono *modelli statistici*, ossia modelli di variabili aleatorie (cioè una distribuzione) che ipotizziamo essere la legge comune all'intera popolazione. Questo modello sarà parametrico, ossia avremo per ipotesi la famiglia di appartenenza e vorremo determinarne i parametri.

Esempio 12.12. Se ipotizziamo che il passaggio degli autobus della linea 5 a Povo sia distribuito secondo una legge esponenziale, dovremo stimarne il parametro λ o, equivalentemente, il valore atteso a partire dalle misurazioni del campione.

12.2. STIMATORI E STIME

DEFINIZIONE 12.13. *Chiamiamo* stimatore *di un parametro una variabile aleatoria che sia una funzione del campione (ossia una statistica), il cui valore è "spesso vicino" al parametro che ci interessa.*

Il valore deterministico assunto dallo stimatore usando i dati osservati prende il nome di stima.

È importante sottolineare quanto appena detto nella definizione: lo stimatore è una *funzione*, in particolare una variabile aleatoria, che ha come argomenti le osservazioni. La stima è un *numero*, una quantità deterministica calcolata a partire dalle effettive misure fatte.

Esempio 12.14. Se il nostro campione (da un punto di vista astratto, prima di fare le misurazioni) è un vettore di n variabili aleatorie indipendenti e identicamente distribuite $(X_1, \dots, X_n)$, un parametro di interesse è il valore atteso comune $E[X_i] = \mu$, che supponiamo ignoto. Uno stimatore della media è la media campionaria (intesa come funzione) $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. Abbiamo una variabile aleatoria $\overline{X}$ e una quantità deterministica μ.

Il fatto che $\overline{X}$ sia uno stimatore (cioè $\overline{X} \approx \mu$) ci è suggerito dalla legge dei grandi numeri (Teorema 11.15): sappiamo infatti che

$$\frac{\sum_{i=1}^{n} X_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = (\overline{X})_n \xrightarrow[n\to+\infty]{P} \mu.$$

È legittimo chiedersi come cambi lo stimatore al crescere del numero delle osservazioni nel campione. Possiamo sottolineare questo aspetto indicando in modo esplicito la dipendenza da n a pedice: $\Theta_n = \hat{\vartheta}((X_i)_{i=1}^n)$.

In generale, nello stimare una quantità, ci aspettiamo di commettere un errore. Questo errore prende il nome di *errore di stima*, che vorremmo quantificare e controllare. Anche questo errore è una variabile aleatoria, quindi siamo interessati, se è possibile, ad averne la distribuzione.

Esempio 12.15. Se sapessimo che $X_1, \dots, X_n$ sono variabili aleatorie indipendenti e identicamente distribuite con media μ ignota, ma varianza σ^2 nota (ad esempio per le specifiche tecniche del macchinario), allora potremmo avere una distribuzione per l'errore commesso nello stimare μ con $\overline{X}$: il teorema centrale del limite (Teorema 11.18) ci dice infatti che

$$\overline{X} - \mu \sim \mathcal{N}\left(0, \frac{\sigma}{\sqrt{n}}\right).$$

Osservazione 12.16. Un parametro non ha necessariamente un unico stimatore. In particolare possiamo avere più stimatori, ottenuti a partire da statistiche (cioè da funzioni) diverse. Gli errori di stima a essi associati avranno in genere distribuzioni diverse tra loro. Vorremmo quindi individuare caratteristiche degli stimatori che ci permettano di scegliere quelli migliori[12.1], tra quelli che possiamo calcolare coi dati a nostra disposizione.

DEFINIZIONE 12.17. *Diciamo che uno stimatore Θ di un parametro ϑ è:*

- corretto o non distorto (unbiased), *se $E[\Theta] = \vartheta$*
- distorto (biased), *se $E[\Theta] \neq \vartheta$; in questo caso il valore $E[\Theta] - \vartheta$ si dice* distorsione o bias.

Se $\lim_{n\to+\infty} E[\Theta_n] = \vartheta$ diciamo che Θ è asintoticamente non distorto.

NOTAZIONE 12.18. *Spesso invece di Θ usiamo $\hat{\vartheta}$ o $\hat{\vartheta}(X)$ per indicare uno stimatore del parametro ϑ. In particolare lo faremo per la media μ, visto che la maiuscola greca sarebbe M.*

Osservazione 12.19. Dire che uno stimatore è biased significa che abbiamo un errore sistematico di sottostima o sovrastima. Questo errore può essere costante o dipendere dal valore del parametro o dalla numerosità del campione.

Il bias misura solamente un aspetto dell'errore. Possiamo considerare altre funzioni di errore, che penalizzino un maggiore allontanamento dal valore "vero" del parametro.

DEFINIZIONE 12.20. *Chiamiamo* errore quadratico medio (mean square error) *di uno stimatore Θ del parametro ϑ la quantità*

$$\mathrm{MSE}[\Theta] = E[(\Theta - \vartheta)^2].$$

12.1. Non abbiamo ancora specificato rispetto a quale metrica intendiamo misurare la bontà degli stimatori. Ce ne sono infatti diverse, come vedremo tra poco.

Osservazione 12.21. Possiamo scrivere l'errore quadratico medio in modo leggermente diverso, in analogia a quanto visto per la varianza (anche questo è un momento secondo):

$$\begin{aligned}\mathrm{MSE}[\Theta] &= E[(\Theta-\vartheta)^2] = E[(\Theta-E[\Theta]+E[\Theta]-\vartheta)^2] \\ &= E[(\Theta-E[\Theta])^2]+E[(E[\Theta]-\vartheta)^2]+2E[(\Theta-E[\Theta])\,(E[\Theta]-\vartheta)] \\ &= \mathrm{Var}[\Theta]+(\mathrm{bias})^2+2\,(E[\Theta]-E[\Theta])\,(E[\Theta]-\vartheta) \\ &= \mathrm{Var}[\Theta]+(\mathrm{bias})^2.\end{aligned}$$

In particolare se Θ è corretto il bias è nullo e $\mathrm{MSE}[\Theta]=\mathrm{Var}[\Theta]$, ma non è detto che conosciamo la varianza di Θ.

DEFINIZIONE 12.22. *Diciamo che uno stimatore Θ di un parametro ϑ è* consistente *se Θ_n converge in probabilità a ϑ per $n\to+\infty$. Se inoltre Θ_n converge in media quadratica a ϑ per $n\to+\infty$, diciamo che Θ è* consistente in media quadratica.

Il prossimo risultato ci dà una condizione sufficiente per la consistenza di uno stimatore.

PROPOSIZIONE 12.23. *Se Θ è asintoticamente non distorto e $\lim_{n\to+\infty}\mathrm{Var}[\Theta_n]=0$, allora Θ è uno stimatore consistente in media quadratica (e quindi anche consistente).*

Dimostrazione. Chiedere che Θ sia consistente in media quadratica significa chiedere che

$$\lim_{n\to+\infty} E[(\Theta_n-\vartheta)^2]=0$$

ossia che $\lim_{n\to+\infty}\mathrm{MSE}[\Theta_n]=0$. Ma per quanto visto nell'Osservazione 12.21,

$$\mathrm{MSE}[\Theta_n]=\mathrm{Var}[\Theta_n]+(E[\Theta_n]-\vartheta)^2$$

e la convergenza a 0 è assicurata dalle ipotesi, separatamente per i due addendi a secondo membro.

La consistenza segue dalla consistenza in media quadratica perché la convergenza in L^2 implica la convergenza in probabilità (Proposizione 11.6). □

Osservazione 12.24. Uno stimatore può essere corretto ma non consistente. Ad esempio se le variabili aleatorie $(X_1,\dots,X_n)$ sono indipendenti e identicamente distribuite, allora ciascuna X_i è uno stimatore non distorto della media μ, poiché $E[X_i]=\mu$ per ogni $i\in\{1,\dots,n\}$.

Tuttavia, posta $\sigma^2=\mathrm{Var}[X_i]\neq 0$ (comune a tutte le X_i), non possiamo avere convergenza in probabilità di alcuno di questi stimatori a μ, poiché, per qualche $\varepsilon>0$

$$P(|X_i-\mu|>\varepsilon)\neq 0$$

e, per ogni n, lo stimatore (X_i) è una variabile aleatoria di media μ e di varianza costante $\sigma^2\neq 0$. Non possiamo dunque avere convergenza in legge ad una costante μ e, a maggior ragione, non possiamo avere convergenza in probabilità.

12.2.1. Alcuni stimatori

Lezione 22

Assumiamo, per questa sottosezione, che le variabili aleatorie $X_1,\dots,X_n$ che costituiscono il campione siano indipendenti e identicamente distribuite di valore atteso comune $E[X_i]=\mu$ e di varianza comune $\mathrm{Var}[X_i]=\sigma^2$.

La *media campionaria* $\hat\mu=\hat\mu_n=\frac{1}{n}\sum_{i=1}^n X_i$ (alle volte indicata anche come $\overline{X}$) è uno stimatore corretto e consistente del valore atteso $E[X_1]=\mu$: questo segue dal teorema centrale del limite (Teorema 11.18) o dalle proprietà del valore atteso e della varianza:

$$\begin{aligned}E[\hat\mu] &= E\left[\frac{1}{n}\sum_{i=1}^n X_i\right]=\frac{1}{n}\sum_{i=1}^n E[X_i]=\frac{1}{n}\cdot n\cdot\mu=\mu \\ \mathrm{Var}[\hat\mu] &= \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}[X_i]=\frac{\sigma^2}{n}\xrightarrow[n\to+\infty]{}0.\end{aligned}$$

Per la varianza σ^2 possiamo usare lo stimatore $S_*^2 = S_{*n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$, ma dobbiamo conoscere la speranza μ. Questo stimatore è corretto

$$E[S_*^2] = \frac{1}{n}\sum_{i=1}^n E[(X_i-\mu)^2] = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}[X_i] = \mathrm{Var}[X_i]$$

ed è anche consistente, poiché per la legge dei grandi numeri (Teorema 11.15)

$$S_{*n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 \xrightarrow[n\to+\infty]{P} E[(X_i-\mu)^2] = \mathrm{Var}[X_i].$$

Se non conosciamo la speranza μ, la prima idea è di sostituire a μ lo stimatore $\hat{\mu}$. Tuttavia

$$\frac{1}{n}\sum_{i=1}^n (X_i-\hat{\mu})^2 = \frac{1}{n}\left(\sum_{i=1}^n X_i^2 - 2\sum_{i=1}^n X_i\hat{\mu} + n\hat{\mu}^2\right) = \frac{1}{n}\left(\sum_{i=1}^n X_i^2 - n\hat{\mu}^2\right)$$

e, se ne prendiamo il valore atteso,

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^n (X_i^2-\mu^2) - (\hat{\mu}^2-\mu^2)\right] \\ &= \frac{1}{n}n\sigma^2 - \mathrm{Var}[\hat{\mu}] \\ &= \sigma^2\cdot\frac{n-1}{n}, \end{aligned} \tag{12.1}$$

ossia abbiamo uno stimatore distorto, dal momento che

$$E\left[\frac{1}{n}\sum_{i=1}^n (X_i-\hat{\mu})^2\right] - \sigma^2 = -\frac{1}{n}\cdot\sigma^2 \neq 0.$$

Vale la pena notare, prima di proseguire, che questo stimatore, pur distorto, è consistente, ancora una volta per la legge dei grandi numeri.

La (12.1) ci suggerisce però la correzione da fare allo stimatore per renderlo non distorto: possiamo moltiplicarlo per $\frac{n}{n-1}$. Quindi uno stimatore per la varianza, se non conosciamo la speranza μ, è $S^2 = S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\hat{\mu})^2$. Questo è uno stimatore corretto, come possiamo facilmente verificare ripercorrendo quanto appena visto, e anche consistente, sempre per la legge dei grandi numeri.

Osservazione 12.25. Lo stimatore S^2 può essere scritto in forma matematicamente equivalente come

$$S^2 = S_n^2 = \frac{1}{n-1}\left(\sum_{i=1}^n X_i^2 - n\hat{\mu}^2\right),$$

ma da un punto di vista numerico o computazionale, questa forma è molto più instabile.

Osservazione 12.26. Possiamo prendere, come stimatori della deviazione standard, $S = \sqrt{S^2}$ oppure $S_* = \sqrt{S_*^2}$. È possibile però mostrare che entrambi questi stimatori sono consistenti ma distorti. In generale non esiste uno stimatore non distorto della deviazione standard valido indipendentemente dalla particolare distribuzione della popolazione (e quindi del campione).

12.2.2. Distribuzione degli stimatori

Vorremmo ora sfruttare meglio il fatto che gli stimatori siano variabili aleatorie e cercare di usare le loro proprietà per ottenere una valutazione degli errori di stima. Per farlo abbiamo però bisogno di conoscere la distribuzione di probabilità dello stimatore.

Consideriamo la seguente situazione: supponiamo che la popolazione abbia una distribuzione Gaussiana di parametri (ignoti) μ e σ. Stiamo quindi affermando che ogni X_i nel campione ha legge $\mathcal{N}(\mu,\sigma)$.

Abbiamo già visto, nella Sotto-sezione 12.2.1 uno stimatore per la media e uno per la varianza[12.2] adatti a questo caso: la media campionaria $\hat{\mu}$ (o $\overline{X}$) e la varianza campionaria (a media ignota) S^2. Ora siamo interessati a determinarne le distribuzioni. Per farlo, iniziamo sfruttando la riproducibilità delle Gaussiane: se tutte le X_i sono Gaussiane, anche $\hat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum X_i$ è Gaussiana e, in particolare ha valore atteso μ e varianza σ^2/n:

$$\hat{\mu}_n = \overline{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \text{ossia} \qquad \frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Sapevamo già, qualunque fosse la distribuzione della popolazione, che $E[\hat{\mu}_n] = \mu$ e $\mathrm{Var}[\hat{\mu}_n] = \frac{\sigma^2}{n}$, ma ora abbiamo l'informazione aggiuntiva che $\hat{\mu}$ ha distribuzione normale e quindi, sapendone anche i parametri, ne conosciamo completamente la legge.

Vogliamo fare lo stesso per lo stimatore S^2: determinarne la distribuzione nel caso di una popolazione Gaussiana. Iniziamo dalla definizione di S^2 e manipoliamola un po':

$$(n-1)\, S_n^2 = \sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \sum_{i=1}^{n} X_i^2 - n\,\hat{\mu}^2.$$

La scrittura a ultimo membro mette in evidenza che (modulo i coefficienti), S^2 è la somma di quadrati di Gaussiane indipendenti e identicamente distribuite più un'ulteriore Gaussiana al quadrato. Questo ci suggerisce un possibile legame con una variabile aleatoria chi quadro.

Per approfondire questo legame, andiamo a riscrivere S^2 in termini di normali standard:

$$\begin{aligned}(n-1)\, S_n^2 &= \sum_{i=1}^{n} X_i^2 - n\,\hat{\mu}_n^2 = \sum_{i=1}^{n} (X_i^2 + \mu^2) - n\,(\hat{\mu}_n^2 + \mu^2) \\ &= \sum_{i=1}^{n} (X_i^2 - 2\,X_i\,\mu + \mu^2) - \sum_{i=1}^{n} (\hat{\mu}_n^2 - 2\,X_i\,\mu + \mu^2) \\ &= \sum_{i=1}^{n} (X_i - \mu)^2 - n\,(\hat{\mu}_n - \mu)^2.\end{aligned}$$

Ora dividiamo primo e ultimo membro per σ^2

$$\frac{(n-1)\, S_n^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}\right)^2$$

e, osservando che ogni $\frac{X_i - \mu}{\sigma}$ è una normale standard, così come $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$, riscriviamo questa identità come

$$\frac{(n-1)\, S_n^2}{\sigma^2} + \left(\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2,$$

in cui a secondo membro abbiamo la somma di n Gaussiane standard indipendenti (ossia una χ^2 a n gradi di libertà) e a primo membro abbiamo $\frac{(n-1)\, S_n^2}{\sigma^2}$ più il quadrato di una normale standard (ossia una χ^2 a un grado di libertà). Allora, per la proprietà di riproducibilità delle chi quadro, deve essere[12.3]

$$\frac{(n-1)\, S_n^2}{\sigma^2} \sim \chi^2(n-1) \qquad \text{cioè} \qquad S_n^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1).$$

Riassumendo, abbiamo ottenuto che per una popolazione Gaussiana di speranza μ e varianza σ^2, lo stimatore media campionaria ha distribuzione Gaussiana $\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ e lo stimatore varianza campionaria a media ignota ha distribuzione $S_n^2 \sim \frac{\sigma^2}{n-1}\chi^2(n-1)$ e che queste variabili aleatorie sono tra loro indipendenti. Ne segue un importante risultato.

12.2. Ci concentriamo sulla varianza e non sulla deviazione standard perché la prima ha uno stimatore corretto e consistente.

12.3. Il risultato è vero, ma stiamo imbrogliando un po' nella giustificazione, infatti dovremmo mostrare che i termini a primo membro sono tra loro indipendenti, cosa che non facciamo in queste note.

COROLLARIO 12.27. *Sia $(X_1, \dots, X_n)$ un campione n- dimensionale estratto da una popolazione a distribuzione Gaussiana di speranza μ e varianza σ^2. Allora*

$$\frac{\hat{\mu}_n - \mu}{\sqrt{S_n^2/n}} \sim t(n-1).$$

Dimostrazione. È sufficiente manipolare la variabile aleatoria che stiamo considerando e usare quanto appena mostrato:

$$\frac{\hat{\mu}_n - \mu}{\sqrt{S_n^2/n}} = \frac{\hat{\mu}_n - \mu}{\sqrt{\sigma^2/n}} \cdot \sqrt{\frac{\sigma^2}{S_n^2}} = Z \cdot \frac{1}{\sqrt{S_n^2/\sigma^2}} = \frac{Z}{\sqrt{\frac{S_n^2}{\sigma^2}(n-1) \cdot \frac{1}{n-1}}} = \frac{Z}{\sqrt{\frac{W}{n-1}}},$$

con $Z \sim \mathcal{N}(0,1)$ e $W \sim \chi^2(n-1)$. Ora possiamo concludere osservando che quella a ultimo membro è precisamente la definizione di una t di Student a $n-1$ gradi di libertà. □

DEFINIZIONE 12.28. *Chiamiamo* funzione ancillare per un parametro ϑ *una variabile aleatoria la cui legge sia nota a priori*[12.4] *e che dipenda dai dati, da parametri noti e da ϑ, unico parametro non noto.*

Parliamo anche di quantità pivot, *se lasciamo cadere la richiesta di dipendenza da un solo parametro incognito.*

Esempio 12.29. Vediamo alcuni esempi di funzione ancillare (per una popolazione Gaussiana):

- $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$ è una funzione ancillare per il parametro μ se la deviazione standard σ è nota, infatti la legge $\mathcal{N}(0,1)$ non dipende dai parametri e la variabile aleatoria è funzione di μ (ignoto), di σ (nota) e della dimensione n del campione, oltre che dai dati;
- $\frac{\hat{\mu}_n - \mu}{\sqrt{S^2/n}} \sim t(n-1)$ è una funzione ancillare per il parametro μ se la deviazione standard σ non è nota;
- $\frac{S_n^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$ è una funzione ancillare per la varianza σ^2;
- $\frac{S_{*n}^2}{\sigma^2} n = \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2(n)$ è una funzione ancillare per la varianza σ^2 se la speranza μ è nota.

12.3. COSTRUIRE STIMATORI

Un problema interessante è quello di costruire stimatori per i parametri di nostro interesse in una popolazione. Abbiamo costruito alcuni stimatori nella sezione precedente, ma ora vogliamo studiare metodi più generali.

Come prima cosa pensiamo alla notazione. Dal momento che, come detto, ci occupiamo di problemi di statistica parametrica, vuol dire che, a meno dei parametri, conosciamo la forma della funzione di densità (o di densità discreta, se il modello è discreto). Se abbiamo un solo parametro ϑ da stimare, possiamo rendere esplicita la dipendenza della funzione densità da questo parametro usando la notazione $f_X(x \mid \vartheta)$. Se abbiamo più di un parametro (ad esempio per una distribuzione normale, che dipende dalla speranza μ e dalla varianza σ^2), possiamo prendere come ϑ il vettore di tutti i parametri (nell'esempio della normale $\vartheta = (\mu, \sigma^2)$).

Siccome per definizione il campione è un vettore di variabili aleatorie indipendenti e identicamente distribuite, esso avrà densità (eventualmente discreta) congiunta

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n \mid \vartheta) = f_X(x_1 \mid \vartheta) \cdots f_X(x_n \mid \vartheta).$$

Vogliamo sfruttare questo fatto per costruire degli stimatori per ϑ.

12.4. Dire che la legge è nota a priori significa che la distribuzione della variabile aleatoria non dipende dai parametri.

12.3.1. Metodo dei momenti

DEFINIZIONE 12.30. *Chiamiamo k-simo* momento campionario *la variabile aleatoria*

$$\hat{\mu}^k = \hat{\mu}^k_n = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Chiamiamo k-simo momento della popolazione *il numero* $\mu^k = E[X_i^k]$.

Osservazione 12.31. La statistica $\hat{\mu}^k$ è uno stimatore corretto di μ^k. Infatti

$$E[\hat{\mu}^k_n] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i^k\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i^k] = \frac{1}{n}\sum_{i=1}^{n} \mu^k = \mu^k.$$

In generale il momento k-simo di una popolazione che dipende da un parametro ϑ sarà una funzione *deterministica* di ϑ. Abbiamo infatti

$$\mu^k = \mu^k(\vartheta) = E[X_1^k] = \int_{-\infty}^{+\infty} x^k \cdot f_X(x \mid \vartheta) \cdot \mathrm{d}x.$$

DEFINIZIONE 12.32. *Chiamiamo* stimatore col metodo dei momenti *del parametro scalare ϑ la soluzione (se esiste) $\hat{\vartheta}_{\mathrm{mom}}$ dell'equazione*

$$\mu^1(\hat{\vartheta}_{\mathrm{mom}}) = \hat{\mu}^1.$$

Se ϑ è un vettore di lunghezza h di parametri, chiamiamo stimatore col metodo dei momenti *del parametro vettoriale ϑ la soluzione (se esiste) $\hat{\vartheta}_{\mathrm{mom}}$ del sistema h-dimensionale di equazioni*

$$\begin{cases} \mu^1(\hat{\vartheta}_{\mathrm{mom}}) = \hat{\mu}^1 \\ \cdots \\ \mu^h(\hat{\vartheta}_{\mathrm{mom}}) = \hat{\mu}^h. \end{cases}$$

Esempio 12.33. Consideriamo una popolazione Gaussiana e un campione $(X_1, \ldots, X_n)$ di dimensione n, con $X_i \sim \mathcal{N}(\mu, \sigma)$, di varianza σ^2 nota. Vogliamo stimare il parametro $\vartheta = \mu$ con il metodo dei momenti.

In questo caso abbiamo

$$\mu^1(\vartheta) = \mu^1(\mu) = E[X_1] = \mu \qquad \text{e, allo stesso tempo,} \qquad \hat{\mu}^1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = \hat{\mu}.$$

Lo stimatore dei momenti di μ è $\hat{\mu}_{\mathrm{mom}}$ tale che

$$\mu^1(\hat{\mu}_{\mathrm{mom}}) = \hat{\mu},$$

ossia, siccome in questo caso la funzione μ^1 è l'identità, $\hat{\mu}_{\mathrm{mom}} = \hat{\mu}$, cioè lo stimatore di μ con il metodo dei momenti è la media campionaria.

Esempio 12.34. Consideriamo una popolazione Gaussiana e un campione $(X_1, \ldots, X_n)$ di dimensione n, con $X_i \sim \mathcal{N}(\mu, \sigma)$, di varianza σ^2 ignota. Vogliamo stimare con il metodo dei momenti i parametri μ e σ^2, quindi cerchiamo $\hat{\vartheta}_{\mathrm{mom}} = (\hat{\mu}_{\mathrm{mom}}, \hat{\sigma}^2_{\mathrm{mom}})$.

Abbiamo

$$\begin{aligned} \mu^1(\vartheta) &= E[X_1] = \mu = \vartheta_1 \\ \mu^2(\vartheta) &= E[X_1^2] = \mu^2 + \sigma^2 = \vartheta_1^2 + \vartheta_2^2 \end{aligned} \qquad \text{e anche} \qquad \begin{aligned} \hat{\mu}^1 &= \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X} = \hat{\mu} \\ \hat{\mu}^2 &= \frac{1}{n}\sum_{i=1}^{n} X_i^2. \end{aligned}$$

Vogliamo risolvere il sistema

$$\begin{cases} \mu^1(\hat{\vartheta}_{\mathrm{mom}}) = \hat{\mu}^1 \\ \mu^2(\hat{\vartheta}_{\mathrm{mom}}) = \hat{\mu}^2 \end{cases} \iff \begin{cases} \hat{\mu}_{\mathrm{mom}} = \hat{\vartheta}_{\mathrm{mom},1} = \hat{\mu} \\ \hat{\mu}^2_{\mathrm{mom}} + \hat{\sigma}^2_{\mathrm{mom}} = \hat{\vartheta}^2_{\mathrm{mom},1} + \hat{\vartheta}^2_{\mathrm{mom},2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \end{cases}$$

quindi abbiamo

$$\begin{aligned}\hat{\mu}_{\text{mom}} &= \hat{\mu} = \overline{X} \\ \hat{\sigma}^2_{\text{mom}} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \hat{\mu}^2_{\text{mom}} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{n-1}{n} S^2.\end{aligned}$$

Allora lo stimatore dei momenti della speranza è anche in questo caso la media campionaria, mentre lo stimatore della varianza ottenuto col metodo dei momenti è lo stimatore *distorto* della varianza che avevamo già incontrato in precedenza.

Esempio 12.35. Consideriamo una popolazione uniforme sull'intervallo $[-a, a]$ e un campione $(X_1, \dots, X_n)$ di dimensione n, con $X_i \sim \text{unif}(-a, a)$. Vogliamo stimare il parametro $\vartheta = a$ usando il metodo dei momenti.

In questo caso la densità è $f_X(x \mid \vartheta) = f_X(x \mid a) = \frac{1}{2a}$ per ogni $x \in [-a, a]$ (e nulla altrimenti). Quindi se andiamo a scrivere μ^1 e $\hat{\mu}^1$ abbiamo

$$\hat{\mu}^1 = \frac{1}{n}\sum_{i=1}^{n} X_i = \hat{\mu} = \overline{X},$$

come già negli esempi precedenti, ma

$$\mu^1(\vartheta) = \mu^1(a) = E[X_1] = \int_{-a}^{a} \frac{1}{2a} \cdot x \, \mathrm{d}x = 0.$$

Non possiamo allora avere soluzioni (in a) per l'equazione $\overline{X} = 0$, dunque (in questo caso) non esiste lo stimatore dei momenti per $\vartheta = a$.

Lezione 23

12.3.2. Metodo di massima verosimiglianza

Partiamo sempre dalla funzione di densità congiunta $f_X(x_1, \dots, x_n \mid \vartheta)$, ma la leggiamo in un modo diverso: come *verosimiglianza* della n-upla di valori $(x_1, \dots, x_n)$ dato il parametro ϑ, cioè quanto è verosimile vedere proprio i valori $(x_1, \dots, x_n)$ se il parametro assume il valore ϑ. Possiamo pensare al problema in questo modo: vogliamo scegliere un valore per ϑ, quindi ha senso prendere quello che massimizza la verosimiglianza che $(x_1, \dots, x_n)$, i valori che osserviamo nel campione alla sua realizzazione, siano quelli assunti dalla variabile aleatoria di cui ϑ è parametro.

DEFINIZIONE 12.36. *Chiamiamo* stimatore di massima verosimiglianza[12.5] *del parametro ϑ la quantità $\hat{\vartheta}_{\text{MLE}}$ che soddisfa*

$$\hat{\vartheta}_{\text{MLE}} = \text{argmax}_\theta f(x_1, \dots, x_n \mid \theta) = \text{argmax}_\theta \log(f(x_1, \dots, x_n \mid \theta)).$$

Osserviamo che massimizzare la verosimiglianza (likelihood, in inglese) o massimizzarne il logaritmo è indifferente, per quanto riguarda il punto in cui il massimo è ottenuto (anche se cambia il valore), grazie al fatto che il logaritmo è una funzione monotona crescente.

Esempio 12.37. Consideriamo una popolazione Gaussiana e un campione $(X_1, \dots, X_n)$ di dimensione n, con $X_i \sim \mathcal{N}(\mu, \sigma)$, di varianza σ^2 ignota. Vogliamo stimare con il metodo di massima verosimiglianza i parametri μ e σ^2, quindi cerchiamo $\hat{\vartheta}_{\text{MLE}} = (\hat{\mu}_{\text{MLE}}, \hat{\sigma}^2_{\text{MLE}})$. In particolare il parametro da stimare è vettoriale.

Come prima cosa scriviamo esplicitamente la densita congiunta

$$f(x_1, \dots, x_n \mid \vartheta) = f(x_1, \dots, x_n \mid \mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sigma^n} \mathrm{e}^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

12.5. In inglese si chiama *maximum likelihood estimator*, da cui la sigla MLE.

Vogliamo trovare $\vartheta = (\mu, \sigma^2)$ che massimizza questa quantità. Data la forma esponenziale della funzione, prendiamone il logaritmo, che poi andremo a massimizzare

$$\log(f(x_1,\dots,x_n|\mu,\sigma^2)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

Per trovare il massimo di questa funzione al variare di μ e σ^2, possiamo calcolarne le derivate (parziali)

$$\begin{aligned}\frac{\partial}{\partial\mu}\log(f(x_1,\dots,x_n|\mu,\sigma^2)) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2)(x_i-\mu) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)\\ \frac{\partial}{\partial\sigma^2}\log(f(x_1,\dots,x_n|\mu,\sigma^2)) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2\end{aligned}$$

e azzerarle[12.6]

$$\begin{cases}\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0\\ -\frac{n}{\sigma^2}+\frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2=0\end{cases} \quad \text{da cui} \quad \begin{cases}\mu=\frac{1}{n}\sum_{i=1}^{n}x_i\\ \sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2.\end{cases}$$

Gli stimatori di massima verosimiglianza sono dunque $\hat{\mu}_{\text{MLE}} = \overline{X}$ e $\hat{\sigma}^2_{\text{MLE}} = \frac{n-1}{n}S^2$, ossia gli stessi stimatori ottenuti con il metodo dei momenti.

Esempio 12.38. Consideriamo una popolazione Bernoulliana e un campione $(X_1,..,X_n)$ di dimensione n, con $X_i \sim \text{bin}(1,p)$. Vogliamo stimare p usando il metodo di massima verosimiglianza.

Iniziamo scrivendo la funzione di verosimiglianza, che in questo caso è la funzione di densità discreta dato p, cioè

$$f(x_1,\dots,x_n|p) = \varphi_X(x_1,\dots,x_n|p) = p^{\sum_{i=1}^{n}x_i}(1-p)^{n-\sum_{i=1}^{n}x_i}.$$

Anche in questo caso ci conviene prenderne il logaritmo

$$\log(f(x_1,\dots,x_n|p)) = \sum_{i=1}^{n}x_i\cdot\log(p) + \left(n-\sum_{i=1}^{n}x_i\right)\log(1-p)$$

che poi deriviamo in p, ponendo la derivata uguale a 0

$$\frac{\mathrm{d}}{\mathrm{d}p}\log(f(x_1,\dots,x_n|p)) = \sum_{i=1}^{n}x_i\cdot\frac{1}{p} - \left(n-\sum_{i=1}^{n}x_i\right)\frac{1}{1-p} = 0$$

da cui

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

Lo stimatore di massima verosimiglianza per p è allora

$$\hat{p}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}X_i = \overline{X}.$$

Esempio 12.39. Consideriamo ora una popolazione esponenziale di parametro ignoto λ, da cui estraiamo un campione $(X_1,\dots,X_n)$ di variabili indipendenti. Vogliamo calcolare lo stimatore di massima verosimiglianza $\hat{\lambda}_{\text{MLE}}$ per il parametro λ.

La densità congiunta dato λ, che vediamo come funzione di verosimiglianza, è

$$f_X(x_1,\dots,x_n|\lambda) = \lambda^n \mathrm{e}^{-\lambda\sum_{i=1}^{n}x_i}.$$

12.6. Dovremmo controllare che i punti così ottenuti siano di massimo globale e non siano punti di minimo, di sella o di massimo locale.

Ne prendiamo il logaritmo, lo deriviamo in λ e poniamo la derivata uguale a 0

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log(f(x_1,\dots,x_n|\lambda)) = n\frac{1}{\lambda} - \sum_{i=1}^{n} x_i = 0$$

da cui otteniamo $\hat{\lambda}_{\mathrm{MLE}} = \overline{X}^{-1}$.

Esempio 12.40. Per lo stimatore di massima verosimiglianza del parametro di una popolazione di Poisson, prendiamo la densità discreta congiunta dato λ,

$$f_X(x_1,\dots,x_n|\lambda) = \frac{\lambda^{\sum_i x_i}}{\prod_i x_i!}\cdot \mathrm{e}^{-n\lambda}$$

ne prendiamo il logaritmo, ne facciamo la derivata rispetto a λ e la poniamo uguale a 0

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\log(f_X(x_1,\dots,x_n|\lambda)) = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0$$

cioè $\hat{\lambda}_{\mathrm{MLE}} = \overline{X}$.

Esempio 12.41. Lo stimatore di massima verosimiglianza per il parametro a di una popolazione uniforme su $[-a,a]$ è $\hat{a}_{\mathrm{MLE}} = \max(|\min_i(x_i)|, |\max_i(x_i)|)$. Infatti

$$f_X(x_1,\dots,x_n|a) = \frac{1}{(2a)^n}\prod_{i=1}^{n} \mathbb{1}_{[-a,a]}(x_i) = \frac{1}{(2a)^n}\,\mathbb{1}_{\{-a\leqslant \min_i(x_i)\leqslant \max_i(x_i)\leqslant a\}}.$$

Osserviamo però che in questo caso derivare non ci è di grande aiuto. Tuttavia il fattore $(2a)^{-n}$ è decrescente in a, ma il fattore $\mathbb{1}_{\{a\geqslant \max(|\min_i(x_i)|,|\max_i(x_i)|)\}}$ è nullo per $a < \max(|\min_i(x_i)|, |\max_i(x_i)|)$, quindi il massimo è in $\hat{a}_{\mathrm{MLE}} = \max(|\min_i(x_i)|, |\max_i(x_i)|)$.

Esempio 12.42. Lo stimatore di massima verosimiglianza per i parametri a e b di una popolazione uniforme su $[a,b]$ è il vettore $(\hat{a}_{\mathrm{MLE}}, \hat{b}_{\mathrm{MLE}}) = (\min_i(x_i), \max_i(x_i))$. Anche in questo caso partiamo dalla densità congiunta

$$f_X(x_1,\dots,x_n|a,b) = \frac{1}{(b-a)^n}\prod_{i=1}^{n} \mathbb{1}_{[a,b]}(x_i) = \frac{1}{(b-a)^n}\,\mathbb{1}_{\{a\leqslant \min_i(x_i)\leqslant \max_i(x_i)\leqslant b\}}$$

e osserviamo che è una funzione decrescente in b, purché $b \geqslant \max_i(x_i)$ ed è una funzione crescente in a, purché $a \leqslant \min_i(x_i)$.

CAPITOLO 13

INTERVALLI DI CONFIDENZA

Rimaniamo nel contesto della stima di parametri, ma vogliamo ora concentrarci su un particolare aspetto: l'errore di stima. Anche quando abbiamo uno stimatore corretto, nel momento in cui passiamo dallo stimatore alla stima, ossia nel momento in cui calcoliamo la statistica in funzione dei valori osservati, cioè della realizzazione del campione, commettiamo un errore e la stima, per quanto prossima, sarà diversa dal valore "teorico[13.1]" del parametro per la popolazione considerata.

Se conosciamo la distribuzione dell'errore di stima, ossia se abbiamo delle opportune funzioni ancillari, possiamo però calcolare non solo un valore numerico per la stima, cioè quella stima *puntuale* su cui ci siamo concentrati nel capitolo precedente, ma anche un margine d'errore. L'idea è quella di individuare un *range* di valori possibili per il parametro che stiamo stimando, all'interno del quale abbiamo un certo livello di sicurezza (*confidenza*, come vedremo tra qualche pagina) che si trovi il valore "teorico" del parametro.

Per semplicità studieremo gli intervalli di confidenza guidandoci con alcuni esempi specifici.

13.1. MEDIA DI UNA NORMALE DI VARIANZA NOTA

Abbiamo un campione $(X_1,...,X_n)$ estratto da una popolazione Gaussiana di media μ (che vogliamo stimare) e varianza σ^2 che assumiamo nota. Abbiamo già osservato nel Capitolo 12 che la media campionaria $\overline{X}_n$ è uno stimatore per la media μ, ma anche che

$$\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}}\sim\mathcal{N}(0,1).$$

Sapendo la distribuzione della variabile aleatoria $\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}}$, possiamo calcolare la probabilità che sia maggiore o minore di un qualche valore: per a e b in $\mathbb{R}$

$$P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}}\leqslant b\right)=\Phi(b),\qquad\qquad P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}}\geqslant a\right)=1-\Phi(a).$$

Specularmente, possiamo anche fissare una probabilità $\beta\in(0,1)$ e chiederci quali siano i numeri reali x e y per cui la variabile aleatoria sia minore o uguale di x con probabilità β o maggiore o uguale di y con probabilità β (ossia i quantili β e $1-\beta$, rispettivamente),

$$P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}}\leqslant x\right)=\beta\iff x=\Phi^{-1}(\beta)$$

$$P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}}\geqslant y\right)=\beta\iff y=\Phi^{-1}(1-\beta)=-\Phi^{-1}(\beta).$$

13.1. È preferibile usare l'attributo *teorico* a *vero*, perché in un certo senso il parametro della distribuzione non è vero nella realtà (che misuriamo), ma solamente nel modello, teorico appunto.

Rimettiamo a fuoco il problema che vogliamo risolvere: vogliamo determinare un range di valori in cui abbiamo un certo livello di fiducia o confidenza che giaccia il valore teorico del parametro μ. Chiamiamo $1-\alpha$ questo livello di confidenza[13.2], per $\alpha \in (0,1)$.

13.1.1. Intervalli bilaterali di confidenza

Supponiamo inoltre di non voler sbagliare troppo né in eccesso, né in difetto. In altre parole vogliamo che il range sia un intervallo $[A,B]$ e che la probabilità che μ sia minore di A sia $\frac{\alpha}{2}$, così come la probabilità che μ sia maggiore di B:

$$P(\mu < A) = \frac{\alpha}{2} \qquad\qquad P(\mu > B) = \frac{\alpha}{2}.$$

In questo modo $P(A \leqslant \mu \leqslant B) = 1-\alpha$.

Come mai parliamo di probabilità? Il parametro μ, per quanto ignoto, non è una variabile aleatoria, quindi saranno variabili aleatorie gli estremi A e B, come suggerito dalla scrittura maiuscola, anzi saranno statistiche, ossia variabili aleatorie dipendenti dal campione e da parametri fissati (ad esempio α). Andiamo infatti a riscrivere il tutto in modo da mettere in evidenza lo stimatore puntuale $\overline{X}_n$ della media μ,

$$\frac{\alpha}{2} = P(\mu < A) = P\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} > \frac{\overline{X}_n - A}{\sqrt{\sigma^2/n}} \right) \iff \frac{\overline{X}_n - A}{\sqrt{\sigma^2/n}} = -\Phi^{-1}\left(\frac{\alpha}{2}\right)$$

da cui, risolvendo in A,

$$A = \overline{X}_n + \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} = \overline{X}_n - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}$$

e, in maniera del tutto analoga,

$$\frac{\alpha}{2} = P(\mu > B) = P\left(\frac{\overline{X}_n - \mu}{\sqrt{\sigma^2/n}} < \frac{\overline{X}_n - B}{\sqrt{\sigma^2/n}} \right) \iff \frac{\overline{X}_n - B}{\sqrt{\sigma^2/n}} = \Phi^{-1}\left(\frac{\alpha}{2}\right)$$

da cui, risolvendo in B,

$$B = \overline{X}_n - \Phi^{-1}\left(\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} = \overline{X}_n + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}.$$

Allora abbiamo

$$P\left(\overline{X}_n - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} \leqslant \mu \leqslant \overline{X}_n + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} \right) = 1-\alpha.$$

Gli estremi dell'intervallo, come accennato in precedenza, sono statistiche: dipendono dal campione e da parametri prefissati (in questo caso α). Quindi, avendo realizzato il campione, gli estremi saranno dei numeri.

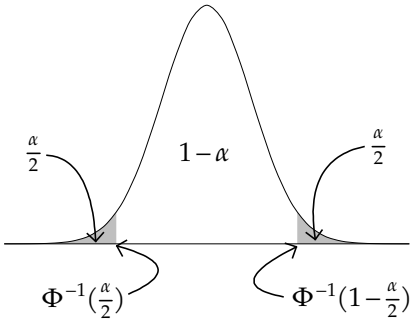
DEFINIZIONE 13.1. *Dato un campione $(X_1,\dots,X_n)$ estratto da una famiglia Gaussiana di media μ ignota e varianza σ^2 nota e fissato un numero $\alpha \in (0,1)$, chiamiamo* intervallo di confidenza bilaterale a livello $1-\alpha$ per la media μ *l'intervallo*

$$\left(\overline{X}_n - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}, \overline{X}_n + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} \right).$$

Valori tipici per $1-\alpha$ sono 90%, 95% e 99%. In questi casi abbiamo

13.2. La nostra fiducia o confidenza è un numero reale in $(0,1)$, ma come osserveremo più avanti non è una probabilità.

$1-\alpha$	α	$\frac{\alpha}{2}$	$1-\frac{\alpha}{2}$	$\Phi^{-1}(1-\frac{\alpha}{2})$
0.9	0.1	0.05	0.95	1.645
0.95	0.05	0.025	0.975	1.96
0.99	0.01	0.005	0.995	2.576



Osservazione 13.2. Come mai parliamo di *confidenza* e non di *probabilità* per questi intervalli? Il motivo è legato alla differenza tra stimatore e stima: il primo è una variabile aleatoria, la seconda è un numero. Per gli intervalli, finché sono scritti in termini degli stimatori possiamo parlare di probabilità, ma quando andiamo a sostituire le stime, ossia i valori calcolati a partire dalla realizzazione del campione, tutto è deterministico: non ha senso parlare di probabilità. In particolare, mentre possiamo dire che il parametro ϑ sta nell'intervallo aleatorio con probabilità $1-\alpha$, nel momento in cui gli estremi sono calcolati a partire dai dati o il parametro ϑ sta lì dentro, oppure non ci sta, non ci sono probabilità. Per questo motivo parliamo di confidenza.

Esempio 13.3. Supponiamo di avere un campione di taglia 16 estratto da una popolazione Gaussiana di media μ e varianza $\sigma^2=9$. Il valore della media campionaria calcolata su questo campione è $\bar{x}=104.7$.

Allora l'intervallo di confidenza a livello 95% per μ è

$$\begin{aligned}\left(\bar{x}-\Phi^{-1}(0.025)\cdot\frac{\sigma}{\sqrt{16}},\bar{x}+\Phi^{-1}(0.025)\cdot\frac{\sigma}{\sqrt{16}}\right) &= \left(104.7-1.96\cdot\frac{3}{4},104.7+1.96\cdot\frac{3}{4}\right)\\ &= (103.23,106.17).\end{aligned}$$

Per curiosità, la popolazione da cui è stato estratto il campione aveva media $\mu=105$.

Supponiamo ora di voler risolvere un problema leggermente diverso, sempre con una popolazione Gaussiana di media μ da stimare e varianza σ^2nota. Vogliamo sapere (prima di raccogliere le osservazioni) quale deve essere la numerosità n del campione per garantire che l'intervallo di confidenza bilaterale per la media μ a livello $1-\alpha$ non sia più ampio di una certa lunghezza prefissata l.

Come prima cosa, osserviamo che, per quanto scritto sopra, la larghezza dell'intervallo di confidenza a livello $1-\alpha$ è

$$\overline{X}_n+\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{\sigma^2}{n}}-\overline{X}_n+\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{\sigma^2}{n}}=2\,\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{\sigma^2}{n}}$$

e dipende da α, da σ^2 e da n, ma non da $\overline{X}_n$. Non solo, dei tre parametri che determinano la larghezza, solo n è variabile, perché la varianza e il livello di confidenza sono assegnati. Osserviamo che la larghezza dell'intervallo diminuisce al crescere di n.

Il problema che vogliamo risolvere è determinare n tale che

$$l=2\,\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{\sigma^2}{n}}$$

cioè, con qualche manipolazione algebrica,

$$n = \frac{4 \cdot (\Phi^{-1}(1-\frac{\alpha}{2}))^2 \cdot \sigma^2}{l^2}.$$

Dobbiamo però prestare attenzione al fatto che $n \in \mathbb{N}$, quindi in generale dovremo approssimare questa soluzione e, per garantire che l'intervallo sia sufficientemente ampio, dovremo farlo per eccesso,

$$n = \left\lceil \frac{4 \cdot (\Phi^{-1}(1-\frac{\alpha}{2}))^2 \cdot \sigma^2}{l^2} \right\rceil. \tag{13.1}$$

Esempio 13.4. Supponiamo di avere una popolazione Gaussiana di varianza $\sigma^2 = 4$ di cui vogliamo stimare la media μ. Vogliamo un intervallo di confidenza al 99% si ampiezza inferiore a 2 e vogliamo determinare la numerosità del campione che ci occorre.

Dalla (13.1) otteniamo

$$n = \left\lceil \frac{4 \cdot (\Phi^{-1}(0.995))^2 \cdot 4}{2^2} \right\rceil = \lceil 26.54 \rceil = 27.$$

Se avessimo voluto un intervallo di larghezza massima 1, allora

$$n = \left\lceil \frac{4 \cdot (2.576)^2 \cdot 4}{1^2} \right\rceil = \lceil 106.158 \rceil = 107,$$

per uno di ampiezza massima 0.5,

$$n = \left\lceil \frac{4 \cdot (2.576)^2 \cdot 4}{0.5^2} \right\rceil = \lceil 424.633 \rceil = 425.$$

Al dimezzarsi della larghezza massima dell'intervallo, il numero di osservazioni necessarie quadruplica. E questo non ci dovrebbe stupire.

Lezione 24

13.1.2. Intervalli unilaterali di confidenza

Facciamo un passo indietro: a volte potremmo essere interessati ad avere un range diverso rispetto a un intervallo, ad esempio potremmo voler avere una soglia sola, essere confidenti che la media sia al di sopra (o al di sotto) di un certo numero. In sostanza stiamo cercando A (una variabile aleatoria) tale che $P(\mu < A) = \alpha$, cioè tale che $P(\mu \geqslant A) = 1-\alpha$ (o analogamente una variabile aleatoria B tale che $P(\mu > B) = \alpha$, cioè $P(\mu \leqslant B) = 1-\alpha$).

L'impostazione però non cambia molto, rispetto a prima: abbiamo

$$A = \overline{X}_n - \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}} \qquad\qquad B = \overline{X}_n + \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}.$$

DEFINIZIONE 13.5. *Dato un campione $(X_1, \ldots, X_n)$ estratto da una famiglia Gaussiana di media μ ignota e varianza σ^2 nota e fissato un numero $\alpha \in (0,1)$, chiamiamo* intervallo di confidenza unilaterale destro (rispettivamente sinistro) a livello $1-\alpha$ per la media μ *la semiretta*

$$\left(\overline{X}_n - \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}, +\infty\right) \qquad \textit{(rispettivamente} \quad \left(-\infty, \overline{X}_n + \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}\right) \quad \textit{)}.$$

Esempio 13.6. Abbiamo le seguenti osservazioni, estratte da una popolazione Gaussiana di media μ ignota e varianza $\sigma^2 = 1$:

3.35 3.73 3.14 4.37 4.28 2.91 2.96 1.94 2.29

Vogliamo calcolare per la media l'intervallo di confidenza destro al 90% e l'intervallo di confidenza sinistro al 99%.

Possiamo calcolare la media campionaria, ad esempio usando il comando R `mean`:

```
mean(c(3.35,3.73,3.14,4.37,4.28,2.91,2.96,1.94,2.29))
[1] 3.218889
```

Allora l'intervallo di confidenza destro al 90% ha come estremo sinistro

$$\bar{x} - \Phi^{-1}(0.9) \cdot \sqrt{\frac{1}{9}} = 3.22 - 1.282 \cdot \frac{1}{3} = 2.793$$

(e come estremo destro $+\infty$).

L'intervallo di confidenza sinistro al 99% ha come estremo destro

$$\bar{x} + \Phi^{-1}(0.99) \cdot \sqrt{\frac{1}{9}} = 3.22 + 2.326 \cdot \frac{1}{3} = 3.995.$$

13.2. COSTRUIRE INTERVALLI DI CONFIDENZA

Vediamo ora un algoritmo per costruire intervalli di confidenza bilaterali per un parametro ϑ a livello di confidenza $1-\alpha$.

Algoritmo per intervalli di confidenza bilaterali

1. Determinare la migliore funzione ancillare $f(X)$ per il caso in considerazione.
2. Trovare i quantili della legge associata ai livelli di confidenza richiesti, ossia $\frac{\alpha}{2}$ e $1-\frac{\alpha}{2}$.
3. Ricavare dall'identità $P(a \leqslant f(X) \leqslant b) = 1-\alpha$ gli estremi a e b.
4. Scrivere l'intervallo (aleatorio) rispetto a ϑ, i cui estremi A e B saranno statistiche.

Esempio 13.7. Mettiamo in pratica l'algoritmo in un caso concreto: vogliamo stimare la media μ di una popolazione Gaussiana di varianza σ^2 ignota.

1. Il nostro parametro ϑ è la media μ. Siamo nel caso in cui la varianza è ignota, quindi
$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$
è la migliore candidata come funzione ancillare.
2. La legge associata è una t di Student a $n-1$ gradi di libertà, quindi i quantili che ci interessano sono $F^{-1}_{t_{n-1}}(\frac{\alpha}{2})$ e $F^{-1}_{t_{n-1}}(1-\frac{\alpha}{2})$. Possiamo calcolarli, ad α fissato, con R o con le tavole.
3. Abbiamo
$$P\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \geqslant a\right) = 1 - \frac{\alpha}{2} \iff P\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leqslant a\right) = \frac{\alpha}{2},$$
da cui $a = F^{-1}_{t_{n-1}}(\frac{\alpha}{2})$. Similmente,
$$P\left(\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leqslant b\right) = 1 - \frac{\alpha}{2} \iff b = F^{-1}_{t_{n-1}}\left(1-\frac{\alpha}{2}\right).$$
4. Dal punto precedente abbiamo
$$F^{-1}_{t_{n-1}}\left(\frac{\alpha}{2}\right) \leqslant \frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \leqslant F^{-1}_{t_{n-1}}\left(1-\frac{\alpha}{2}\right)$$
che possiamo scrivere esplicitamente, per μ,
$$\overline{X}_n - F^{-1}_{t_{n-1}}\left(1-\frac{\alpha}{2}\right) \cdot \frac{S_n}{\sqrt{n}} \leqslant \mu \leqslant \overline{X}_n - F^{-1}_{t_{n-1}}\left(\frac{\alpha}{2}\right) \cdot \frac{S_n}{\sqrt{n}} = \overline{X}_n + F^{-1}_{t_{n-1}}\left(1-\frac{\alpha}{2}\right) \cdot \frac{S_n}{\sqrt{n}},$$
in cui abbiamo sfruttato le proprietà di simmetria di $F^{-1}_{t_{n-1}}$ (attenzione che non tutte le statistiche ancillari hanno leggi simmetriche).

Esempio 13.8. Sempre usando l'algoritmo visto sopra, determiniamo l'intervallo di confidenza bilaterale per la varianza di una popolazione Gaussiana a media ignota.

1. La funzione ancillare in questo caso è $\frac{S_n^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$.
2. I quantili (non simmetrici!) sono $F^{-1}_{\chi^2_{n-1}}(\frac{\alpha}{2})$ e $F^{-1}_{\chi^2_{n-1}}(1-\frac{\alpha}{2})$, da calcolare con R o con le tavole.
3. Per gli estremi abbiamo
$$P\left(\frac{S_n^2}{\sigma^2}(n-1) \leqslant a\right) = \frac{\alpha}{2} \implies a = F^{-1}_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right)$$
$$P\left(\frac{S_n^2}{\sigma^2}(n-1) \leqslant b\right) = 1-\frac{\alpha}{2} \implies b = F^{-1}_{\chi^2_{n-1}}\left(1-\frac{\alpha}{2}\right).$$
4. Infine,
$$P\left(F^{-1}_{\chi^2_{n-1}}\left(\frac{\alpha}{2}\right) \leqslant \frac{S_n^2}{\sigma^2}(n-1) \leqslant F^{-1}_{\chi^2_{n-1}}\left(1-\frac{\alpha}{2}\right)\right) = 1-\alpha$$
che dà, esplicitato in σ^2,
$$P\left(\frac{S_n^2(n-1)}{F^{-1}_{\chi^2_{n-1}}(1-\frac{\alpha}{2})} \leqslant \sigma^2 \leqslant \frac{S_n^2(n-1)}{F^{-1}_{\chi^2_{n-1}}(\frac{\alpha}{2})}\right) = 1-\alpha$$
in cui vale la pena notare che, siccome σ^2 era al denominatore, $F^{-1}_{\chi^2_{n-1}}(1-\frac{\alpha}{2})$ è passato all'estremo sinistro e $F^{-1}_{\chi^2_{n-1}}(\frac{\alpha}{2})$ all'estremo destro.

Osservazione 13.9. Possiamo con qualche accortezza usare il medesimo algoritmo per trovare anche gli intervalli unilaterali destri o sinistri. Quello che cambia è la necessità di trovare solamente un quantile (e non due): dobbiamo però prestare attenzione a prendere quello giusto e calcolato al livello giusto.

Nel caso di una popolazione Gaussiana $\mathcal{N}(\mu, \sigma)$ abbiamo i seguenti intervalli di confidenza:

ϑ	note	Int. bilaterale	Int. sinistro	Int. destro
μ	σ^2 nota	$\overline{X}_n \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}$	$\left(-\infty, \overline{X}_n + \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}\right)$	$\left(\overline{X}_n - \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}, +\infty\right)$
μ	σ^2 ignota	$\overline{X}_n \pm F^{-1}_{t_{n-1}}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{S^2}{n}}$	$\left(-\infty, \overline{X}_n + F^{-1}_{t_{n-1}}(1-\alpha)\sqrt{\frac{S^2}{n}}\right)$	$\left(\overline{X}_n - F^{-1}_{t_{n-1}}(1-\alpha)\sqrt{\frac{S^2}{n}}, +\infty\right)$
σ^2	μ nota	$\left(\frac{S^2_{*n}n}{F^{-1}_{\chi^2_n}(1-\frac{\alpha}{2})}, \frac{S^2_{*n}n}{F^{-1}_{\chi^2_n}(\frac{\alpha}{2})}\right)$	$\left(0, \frac{S^2_{*n}n}{F^{-1}_{\chi^2_n}(\alpha)}\right)$	$\left(\frac{S^2_{*n}n}{F^{-1}_{\chi^2_n}(1-\alpha)}, +\infty\right)$
σ^2	μ ignota	$\left(\frac{S_n^2(n-1)}{F^{-1}_{\chi^2_{n-1}}(1-\frac{\alpha}{2})}, \frac{S_n^2(n-1)}{F^{-1}_{\chi^2_{n-1}}(\frac{\alpha}{2})}\right)$	$\left(0, \frac{S_n^2(n-1)}{F^{-1}_{\chi^2_{n-1}}(\alpha)}\right)$	$\left(\frac{S_n^2(n-1)}{F^{-1}_{\chi^2_{n-1}}(1-\alpha)}, +\infty\right)$

Tabella 13.1. Intervalli di confidenza per una popolazione Gaussiana a livello $1-\alpha$.

in cui $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $S^2_{*n} = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2$ e $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

13.3. INTERVALLI DI CONFIDENZA PER LA DIFFERENZA DI MEDIE

Consideriamo ora una situazione un po' diversa. Abbiamo due popolazioni, entrambe Gaussiane. Ci chiediamo quanto grande sia la differenza tra le loro medie.

Come prima cosa osserviamo che, avendo due popolazioni, avremo anche due campioni: $(X_i)_{i=1}^n$ e $(Y_j)_{j=1}^m$, con ciascuna $X_i \sim \mathcal{N}(\mu_X, \sigma_X)$ e ciascuna $Y_j \sim \mathcal{N}(\mu_Y, \sigma_Y)$. Osserviamo che, come sottolineato dalla notazione che abbiamo usato, i due campioni non sono necessariamente della stessa taglia.

Il nostro obiettivo è stimare la differenza tra la media della prima popolazione e quella della seconda, ossia $\mu_X - \mu_Y$. Uno stimatore di questa quantità (che in particolare è lo stimatore di massima verosimiglianza) è $\overline{X}_n - \overline{Y}_m$, quindi abbiamo una stima puntuale.

Se siamo però interessati a una stima intervallare, abbiamo bisogno di indagare più a fondo sulla distribuzione di $\overline{X}_n - \overline{Y}_m$. Come prima cosa osserviamo che $\overline{X}_n \sim \mathcal{N}\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right)$ e $\overline{Y}_m \sim \mathcal{N}\left(\mu_Y, \frac{\sigma_Y}{\sqrt{m}}\right)$. I due campioni sono estratti da popolazioni diverse, quindi li possiamo considerare indipendenti. Sapendo che combinazioni lineari di Gaussiane indipendenti sono a loro volta Gaussiane,

$$\overline{X}_n - \overline{Y}_m \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$

in cui vale la pena osservare che, anche se delle medie abbiamo la differenza, delle varianze abbiamo la somma, infatti

$$\begin{aligned} E[\overline{X}_n - \overline{Y}_m] &= E[\overline{X}_n] - E[\overline{Y}_m] = \mu_X - \mu_Y \\ \mathrm{Var}[\overline{X}_n - \overline{Y}_m] &= \mathrm{Var}[\overline{X}_n] + (-1)^2 \mathrm{Var}[\overline{Y}_m] = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}. \end{aligned}$$

Con le (ormai consuete) manipolazioni, abbiamo

$$\frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim \mathcal{N}(0,1).$$

A questo punto, se sappiamo σ_X e σ_Y, possiamo ricavare l'intervallo di confidenza: stiamo rifacendo quanto visto per la media di una Gaussiana a varianza nota. Avremo dunque che

$$(\mu_X - \mu_Y) \in \left((\bar{x} - \bar{y}) - \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}, (\bar{x} - \bar{y}) + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \right)$$

è un intervallo di confidenza bilaterale a livello $1 - \alpha$ (in cui abbiamo messo in evidenza le stime $\bar{x}$ e $\bar{y}$, ossia gli stimatori calcolati nei campioni realizzati. In modo analogo possiamo ricavare gli intervalli unilaterali.

Tuttavia non sempre sappiamo le varianze delle due popolazioni, siamo in grado di dire qualcosa nel caso in cui esse siano ignote? Nel caso di una singola Gaussiana abbiamo usato

$$\frac{\overline{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1),$$

per ottenere la quale abbiamo sfruttato la distribuzione χ^2 di S_n^2. Se proviamo a replicare questa strategia nel caso della differenza abbiamo

$$\frac{(\overline{X}_n - \overline{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{S_{X,n}^2/n + S_{Y,m}^2/m}},$$

per cui in generale abbiamo una distribuzione di $S_{X,n}^2/n + S_{Y,m}^2/m$ che non è semplice da ricavare e che dipende dalle due varianze, rendendoci quindi impossibile l'uso come funzione ancillare [13.3].

C'è però un caso speciale: il caso *omoschedastico*, cioè in cui le varianze delle due popolazioni, pur ignote, coincidono, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$. In questa situazione

$$\frac{S_{X,n}^2}{\sigma_X^2}(n-1) = \frac{S_{X,n}^2}{\sigma^2}(n-1) \sim \chi^2(n-1), \qquad \frac{S_{Y,m}^2}{\sigma_Y^2}(m-1) = \frac{S_{Y,m}^2}{\sigma^2}(m-1) \sim \chi^2(m-1)$$

e sommandole, grazie all'indipendenza dei due campioni e alla riproducibilità delle χ^2,

$$\frac{S_{X,n}^2}{\sigma^2}(n-1) + \frac{S_{Y,m}^2}{\sigma^2}(m-1) \sim \chi^2(n-1) + \chi^2(m-1) \sim \chi^2(n+m-2).$$

13.3. Ricordiamo che una funzione ancillare può avere al più un parametro ignoto, in questo caso ne avremmo due.

Possiamo allora scrivere

$$\begin{aligned}\frac{S^2_{X,n}}{\sigma^2}(n-1)+\frac{S^2_{Y,m}}{\sigma^2}(m-1) &= \frac{S^2_{X,n}(n-1)+S^2_{Y,m}(m-1)}{n+m-2}\cdot\frac{n+m-2}{\sigma^2}\\ &=: \frac{S^2_P}{\sigma^2}(n+m-2)\sim\chi^2(n+m-2),\end{aligned}$$

in analogia con quanto visto per una singola popolazione Gaussiana di varianza ignota. Abbiamo così introdotto lo stimatore S^2_P, detto *stimatore pooled della varianza*, che è una media pesata di S^2_X e S^2_Y di pesi dati dai gradi di libertà delle loro distribuzioni (ossia $\frac{n-1}{n+m-2}$ e $\frac{m-1}{n+m-2}$).

A questo punto possiamo ancora una volta continuare come nel caso della singola popolazione Gaussiana di varianza ignota e otteniamo

$$\begin{aligned}\frac{(\overline{X}_n-\overline{Y}_m)-(\mu_X-\mu_Y)}{\sqrt{S^2_P\left(\frac{1}{n}+\frac{1}{m}\right)}} &= \frac{(\overline{X}_n-\overline{Y}_m)-(\mu_X-\mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}}\cdot\frac{1}{\sqrt{\frac{S^2_P}{\sigma^2}(n+m-2)\cdot\frac{1}{n+m-2}}}\\ &\sim \mathcal{N}(0,1)\cdot\frac{1}{\sqrt{\chi^2(n+m-2)/_{n+m-2}}}\sim t(n+m-2).\end{aligned}$$

A questo punto individuare gli intervalli di confidenza è analogo a quanto visto nel caso di una Gaussiana con varianza ignota. In particolare, nel caso bilaterale abbiamo

$$(\mu_X-\mu_Y)\in\left((\bar{x}-\bar{y})-F^{-1}_{t(n+m-2)}\left(1-\frac{\alpha}{2}\right)\sqrt{S^2_P\left(\frac{1}{n}+\frac{1}{m}\right)},(\bar{x}-\bar{y})+F^{-1}_{t(n+m-2)}\left(1-\frac{\alpha}{2}\right)\sqrt{S^2_P\left(\frac{1}{n}+\frac{1}{m}\right)}\right).$$

13.4. INTERVALLI DI CONFIDENZA APPROSSIMATI

Oltre alle funzioni ancillari esatte, grazie al teorema centrale del limite abbiamo anche delle funzioni ancillari approssimate. Queste alle volte ci vengono in aiuto quando le funzioni ancillari esatte sono difficili o impossibili da usare. Dobbiamo però essere consapevoli che gli intervalli così ottenuti non saranno altrettanto precisi di quelli ottenuti con funzioni ancillari non approssimate.

13.4.1. Popolazione Bernoulliana

Consideriamo il caso di una popolazione Bernoulliana di parametro p. In altre parole stiamo dicendo che ogni individuo nella popolazione ha una determinata caratteristica con probabilità p. Alcuni esempi di caratteristiche di questo tipo possono essere "possedere un'automobile", "avere un certa caratteristica genetica uniformemente diffusa nella popolazione".

Vogliamo stimare il parametro p. Iniziamo osservando che ogni elemento del campione è una variabile aleatoria Bernoulliana, ossia assume il valore 1 con probabilità p e 0 con probabilità $1-p$. Abbiamo allora la variabile aleatoria $Y_n=\sum_{i=1}^n X_i$ che conta il numero di successi (e ha distribuzione binomiale).

Dal teorema centrale del limite (Teorema 11.18) sappiamo che per n sufficientemente grande

$$\frac{Y_n-np}{\sqrt{np(1-p)}}\dot{\sim}\mathcal{N}(0,1),$$

ossia abbiamo una funzione ancillare approssimata per il parametro p. Tuttavia non siamo in grado di usarla direttamente: sappiamo che

$$P\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\leqslant\frac{Y_n-np}{\sqrt{np(1-p)}}\leqslant\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right)=1-\alpha,$$

ma se proviamo a scriverlo esplicitamente in termini di p ci blocchiamo, perché p compare anche a denominatore e per giunta sotto una radice quadrata.

Abbiamo bisogno di una seconda approssimazione: sappiamo che p è la media di ciascuna variabile aleatoria estratta dalla popolazione (gli individui sono tutti Bernoulliani di parametro p) e sappiamo anche che $\overline{X}_n = \frac{Y_n}{n}$ è uno stimatore della media. Poniamo quindi $\hat{p} = \frac{Y_n}{n} = \overline{X}_n$. È una statistica calcolabile del campione e $\sqrt{np(1-p)} \approx \sqrt{n\hat{p}(1-\hat{p})}$ e dunque abbiamo una nuova funzione ancillare (ulteriormente) approssimata

$$\frac{n\hat{p} - np}{\sqrt{n\hat{p}(1-\hat{p})}} \rightsquigarrow \mathcal{N}(0,1).$$

A questo punto possiamo riprendere la strada iniziata prima:

$$P\left(-\Phi^{-1}\left(1-\frac{\alpha}{2}\right) \leqslant \frac{n\hat{p} - np}{\sqrt{n\hat{p}(1-\hat{p})}} \leqslant \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\right) \approx 1-\alpha$$

che ora possiamo riscrivere in modo da avere un intervallo esplicito per p:

$$P\left(\hat{p} - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqslant p \leqslant \hat{p} + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha$$

e analogamente per gli intervalli unilaterali.

Osservazione 13.10. Anche per le Bernoulliane ha senso chiedersi quanto grande debba essere la numerosità n del campione per garantire che l'ampiezza dell'intervallo (bilaterale) sia al di sotto di una certa soglia l. L'ampiezza è $2\,\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, ma dipende da $\hat{p}$ e quindi dalle osservazioni del campione, così come il corrispondente

$$n = \left\lceil \frac{4\,(\Phi^{-1}(1-\frac{\alpha}{2}))^2}{l^2}\hat{p}(1-\hat{p}) \right\rceil.$$

Qual è il problema? Che non sappiamo quanto vale $\hat{p}$ prima di iniziare a raccogliere i nostri dati. In questo caso una soluzione pratica è iniziare a raccogliere i dati e, dalle prime m misurazioni, stimare rozzamente p (e quindi anche $\hat{p}$) con $\overline{X}_m$ usando questo valore per stimare la numerosità necessaria del campione. A questo punto è possibile continuare a raccogliere gli ulteriori dati.

Esempio 13.11. Vogliamo stimare la proporzione di studenti che consulta libri in biblioteca e vorremmo avere un margine di incertezza (ossia metà dell'ampiezza dell'intervallo di confidenza) del 2.5% per un intervallo al 95%.

Iniziamo intervistando i primi 25 studenti, di cui 9 consultano libri in biblioteca. La nostra stima grossolana di p è $p^* = \frac{9}{25} = 0.36$. Questo ci suggerisce di intervistare in tutto

$$n = \left\lceil \frac{4\left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2}{l^2} p^*(1-p^*) \right\rceil = \left\lceil \frac{4 \cdot 1.96^2}{0.05^2} \cdot 0.36 \cdot 0.64 \right\rceil = 1417,$$

cioè altri 1392. Di questi 535 rispondono positivamente, per una stima puntuale di p uguale a $\hat{p} = \frac{535+9}{1392+25} \simeq 0.384$.

Il nostro intervallo di confidenza (approssimato) è quindi

$$\hat{p} \pm \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.384 \pm 1.96 \cdot \sqrt{\frac{0.384 \cdot 0.616}{1417}},$$

cioè $(0.359, 0.409)$, che ha ampiezza 0.05 e margine d'errore 0.025.

Osservazione 13.12. Stiamo approssimando a più livelli, quindi ci aspettiamo un po' di errore aggiuntivo. Tuttavia possiamo, come esercizio, chiederci cosa succeda se invece di p^* o $\hat{p}$ usassimo il vero valore di p nel determinare la numerosità del campione. In questo caso

$$n = \left\lceil \frac{4\left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2}{l^2} p\,(1-p) \right\rceil \leqslant \left\lceil \frac{\left(\Phi^{-1}(1-\frac{\alpha}{2})\right)^2}{l^2} \right\rceil,$$

poiché $p\,(1-p) \leqslant 1/4$. Questa approssimazione (che non dipende da p) è però sempre meno precisa quanto più p è vicino agli estremi 0 o 1.

13.4.2. Popolazione Poissoniana

Se abbiamo una popolazione Poissoniana di parametro λ che vogliamo stimare (ossia $\vartheta = \lambda$), abbiamo immediatamente uno stimatore per λ, che è la media della distribuzione, ossia la media campionaria $\overline{X}_n$. Inoltre, poiché la distribuzione Poissoniana è riproducibile, ne conosciamo anche la distribuzione: $n \cdot \overline{X}_n \sim \mathrm{Pois}(n\lambda)$, quindi

$$P(\overline{X}_n = k) = P\left(\sum_{i=1}^{n} X_i = nk\right) = \frac{(n\lambda)^{nk}}{(nk)!} \mathrm{e}^{-n\lambda}.$$

Grazie a questa informazione, possiamo costruire degli intervalli di confidenza per λ, ma con un po' di difficoltà. Sappiamo che i tempi d'attesa tra due eventi Poissoniani di media λ sono distribuiti secondo una legge esponenziale di intensità λ, che le variabili aleatorie esponenziali sono scalabili, cioè che $\alpha \cdot \exp(\lambda) \sim \exp\left(\frac{\lambda}{\alpha}\right)$, che $\exp\left(\frac{1}{2}\right) \sim \chi^2(2)$ e che la distribuzione χ^2 è riproducibile. Mettendo assieme queste informazioni abbiamo che

$$F_{\mathrm{Pois}(\lambda)}(k) = 1 - F_{\chi^2(2(k+1))}(2\lambda)$$

da cui possiamo ricavare[13.4] che un intervallo bilaterale di confidenza a livello $1-\alpha$ per λ è

$$\left(\frac{1}{2n} F^{-1}_{\chi^2(2n\overline{X}_n)}\left(\frac{\alpha}{2}\right), \frac{1}{2n} F^{-1}_{\chi^2(2n\overline{X}_n+2)}\left(1-\frac{\alpha}{2}\right)\right).$$

Data la forma non semplicissima, possiamo in alternativa accontentarci di un intervallo di confidenza approssimato, sfruttando il teorema centrale del limite, in questo caso

$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} \dot{\sim} \mathcal{N}(0,1).$$

Come nel caso della binomiale, tuttavia, abbiamo un fattore $\sqrt{\lambda}$ a denominatore che ci causa problemi, ma possiamo approssimarlo anche in questo caso con $\overline{X}_n$. A questo punto l'intervallo bilaterale di confidenza approssimato a livello $1-\alpha$ che otteniamo è

$$\left(\overline{X}_n - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\overline{X}_n}{n}}, \overline{X}_n + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\overline{X}_n}{n}}\right).$$

Esempio 13.13. Il numero di email di studenti ricevute da un docente nel corso di una giornata è ipotizzato essere distribuito come una Poisson di media λ ignota. Vengono contate le email ricevute giorno per giorno per 100 giorni. La media campionaria misurata è $\bar{x} = 5.04$. Qual è un intervallo di confidenza bilaterale al 95% per λ?

Calcoliamo come prima cosa l'intervallo di confidenza esatto: esso è

$$\left(\frac{1}{2\cdot 100} \cdot F^{-1}_{\chi^2(2\cdot 100\cdot 5.04)}(0.025), \frac{1}{200} \cdot F^{-1}_{\chi^2(2\cdot 100\cdot 5.04+2)}(0.975)\right) = (4.609536, 5.499841),$$

che ha centro $5.054688 \neq \bar{x}$ e ampiezza 0.8903049.

13.4. Non vediamo i dettagli, che non sono banali.

Passiamo invece all'intervallo approssimato: esso è

$$\left(5.04-\Phi^{-1}(0.975)\sqrt{\frac{5.04}{100}},5.04-\Phi^{-1}(0.975)\sqrt{\frac{5.04}{100}}\right)=(4.599989,5.480011),$$

che ha centro $5.04=\bar{x}$ e ampiezza 0.8800216.

Osserviamo in particolare che il primo, a differenza del secondo, non è simmetrico rispetto alla media campionaria $\bar{x}=5.04$.

Possiamo fare questi conti in R, usando il seguente codice[13.5]

```
n_days <- 100
lambda <- 5
alpha <- 0.05
x <- rpois(n_days, lambda)
x_bar <- mean(x)

# Intervallo corretto
ci_chi <- c(1/(2*n_days)*qchisq(alpha/2, df = 2*n_days*x_bar),
1/(2*n_days)*qchisq(1-alpha/2, df = 2*n_days*x_bar+2))

# Intervallo approssimato
ci_approx <- sapply(c(alpha/2, 1-alpha/2),function(x){x_bar +
qnorm(x)*sqrt(x_bar/n_days)})

# centri dei due intervalli
mean(ci_chi)
mean(ci_approx)

# ampiezza dei due intervalli (%*% è il prodotto matriciale)
ci_chi %*% c(-1,1)
ci_approx %*% c(-1,1)
```

Osservazione 13.14. Per campioni di numerosità elevata la differenza tra gli intervalli ottenuti nei due modi è trascurabile. Nel caso di campioni di numerosità ridotta la differenza è più significativa, come si può verificare adattando il codice appena visto.

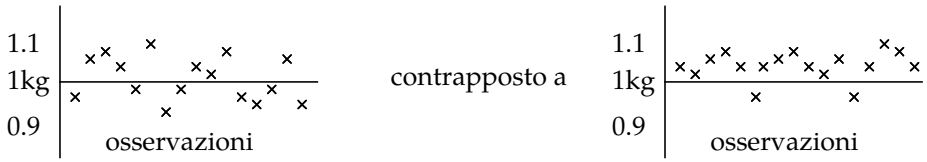
Osservazione 13.15. In generale, non solo per le distribuzioni di Poisson, ci possono essere più scelte possibili di intervalli di confidenza, sia esatti, sia approssimati: questo succede se si parte da quantità pivot (o statistiche ancillari) diverse, se si usano approssimazioni diverse e così via.

13.5. Siccome generiamo il campione in modo aleatorio, gli estremi dell'intervallo saranno diversi in iterazioni diverse.

CAPITOLO 14

TEST STATISTICI

L'idea che sta alla base dei test statistici è la seguente: abbiamo un'ipotesi (ad esempio "in media una confezione contiene 1 kg di pomodori") e vogliamo vedere se le osservazioni a nostra disposizione (i dati) supportano (cioè non contraddicono) questa ipotesi o se la contraddicono. Possiamo mostrare graficamente l'idea: Lezione 25



Nell'immagine a sinistra non possiamo escludere che il valore corrispondente alla retta orizzontale sia la media della popolazione da cui abbiamo estratto il campione, mentre nell'immagine a destra sembra poco plausibile che quel valore sia la media.

Come si lega questo con quanto abbiamo visto finora in Statistica? Abbiamo una popolazione sottostante che supponiamo avere una distribuzione comune, dipendente da un parametro (ad esempio la media). Pensiamo che la media della popolazione sia un certo valore μ_0 e vogliamo mettere alla prova questa ipotesi, usando i dati, ossia il campione estratto dalla popolazione.

Nel Capitolo 12 abbiamo visto che, a partire dal campione, possiamo stimare la media con la media campionaria $\overline{X}$, quindi una possibilità per testare la nostra ipotesi potrebbe essere la seguente: se il valore stimato $\overline{X}$ coincide con la nostra ipotesi μ_0, allora è vero che la popolazione ha proprio quella media, altrimenti no. Abbiamo però visto che la stima puntuale è troppo imprecisa per poter fare un ragionamento del genere.

Nel Capitolo 13 abbiamo però introdotto la stima intervallare: potremmo pensare di adattare quella. Se μ_0 è nell'intervallo di confidenza a un certo livello attorno a $\overline{X}$, allora non escludiamo che μ_0 possa davvero essere il valore della media, mentre se μ_0 giace al di fuori dell'intervallo di confidenza, escludiamo che sia il valore della media della popolazione.

Questo però è ancora molto impreciso, anche se ci dà un'idea di quello che vogliamo fare. Per rendere il ragionamento più rigoroso, iniziamo introducendo una terminologia più precisa.

Chiamiamo *ipotesi statistica* da verificare su una popolazione (o distribuzione) un'affermazione relativa a uno (o più) dei suoi parametri. Usando il termine ipotesi vogliamo sottolineare che a priori non sappiamo se questa affermazione sia vera oppure no. La forma che prende un'ipotesi statistica può variare: se il parametro di interesse è ϑ e ϑ_0 è un valore soglia (o target) fissato, sono esempi di ipotesi statistiche $\vartheta = \vartheta_0$, $\vartheta \geqslant \vartheta_0$ e così via.

Per fare un *test statistico*, come prima cosa stabiliamo due ipotesi: un'*ipotesi nulla*, denotata con H_0, che rappresenta il caso di default (ad esempio H_0: $\vartheta = \vartheta_0$) e un'*ipotesi alternativa*, denotata con H_1 o H_a, a essa complementare (nel nostro esempio H_1: $\vartheta \neq \vartheta_0$). L'ipotesi nulla è la risposta che diamo in caso di test negativo: accettiamo quella come risposta di default se non abbiamo evidenza (statistica) del contrario nei dati, se non possiamo escludere che l'ipotesi nulla sia vera. L'ipotesi alternativa è la risposta in caso di test positivo, ossia se abbiamo evidenza (statistica) che l'ipotesi nulla sia falsa. Torneremo su questo aspetto più avanti.

Il secondo passo in un test statistico è il seguente: mettiamo alla prova la nostra ipotesi (nulla) usando un campione estratto dalla popolazione. Per fare questo, determiniamo una *regione di accettazione* dello spazio n-dimensionale: se il campione (un vettore n-dimensionale, ossia un punto nello spazio n-dimensionale) cade all'interno della regione, allora accettiamo l'ipotesi nulla, altrimenti la rifiutiamo, scegliendo l'ipotesi alternativa[14.1]. Il complementare della regione di accettazione, ossia la porzione dello spazio n-dimensionale in cui rifiutiamo l'ipotesi nulla prende il nome di *regione critica*. Vedremo a breve come determinare queste regioni dello spazio (e da cosa dipendono), ma nel frattempo osserviamo il contatto con gli intervalli di confidenza suggerito poco sopra: una possibilità potrebbe essere quella di calcolarci un'opportuna statistica a partire dal campione e prendere come regione di accettazione un'intervallo, magari di confidenza. Dobbiamo però far entrare in gioco anche il nostro valore target.

Prima di continuare in concreto, però, abbiamo bisogno di altre considerazioni astratte. In particolare vogliamo pensare agli errori che possiamo commettere in un test statistico: possiamo infatti sbagliare in due modi: rifiutare l'ipotesi nulla H_0 quando questa è vera (*errore di prima specie*) oppure accettare l'ipotesi nulla quando questa è falsa (*errore di seconda specie*). Le quattro possibili situazioni sono rappresentate nella Tabella 14.1.

	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$
H_0	ok	Errore di seconda specie
H_1	Errore di prima specie	ok

Tabella 14.1. Nella colonna abbiamo la risposta del test, nella riga (in grassetto) la realtà.

Come abbiamo accennato prima, l'ipotesi nulla è anche detta *test negativo* e l'ipotesi alternativa *test positivo*, terminologia ereditata dai test clinici. Possiamo allora dare nomi diversi alle quattro possibili situazioni, riportati in Tabella 14.2.

	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$
H_0	Vero negativo (TN)	Falso negativo (FN)
H_1	Falso positivo (FP)	Vero positivo (TP)

Tabella 14.2. Nella colonna abbiamo la risposta del test, nella riga (in grassetto) la realtà.

Indichiamo con α la probabilità di commettere un errore di prima specie. Solitamente vorremo che α sia al di sotto di una certa soglia $\bar{\alpha}$ detta *livello di significatività*. Indichiamo invece con β la probabilità di commettere un errore di seconda specie. In un certo senso[14.2], le due quantità α e β misurano la "qualità" di un test: un test molto buono avrà sia α sia β molto piccoli, un test perfetto (che non esiste!) li avrà entrambi nulli.

Per tornare alla regione di accettazione, quello che vorremmo fare è fissare una soglia massima $\bar{\alpha}$ per la probabilità α di errori di prima specie (ed eventualmente assegnare qualche condizione su β) e a partire da ciò determinare la regione di accettazione e la regione critica. Inoltre, siccome quello che abbiamo a disposizione è un campione estratto dalla popolazione, ne calcoleremo una funzione (cioè una statistica) che useremo per il nostro test.

Ci sono diversi modi per procedere, che si differenziano tra loro per il bilanciamento della complessità tra la statistica da calcolare e quella della regione di accettazione.

14.1. IMPOSTARE TEST STATISTICI

Algoritmo per un test bilaterale

1. Stabilire le ipotesi da testare. Nel caso bilaterale saranno della forma $H_0: \vartheta = \vartheta_0$ e $H_1: \vartheta \neq \vartheta_0$.

14.1. Sarebbe meglio indicare le due alternative come *non rifiuto* dell'ipotesi nulla e *rifiuto* dell'ipotesi nulla. Come vedremo, *accettare* l'ipotesi nulla è un po' fuorviante come terminologia.

14.2. Vedremo alcuni dettagli in più tra qualche pagina.

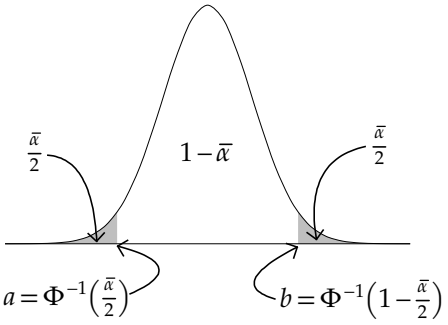
2. Fissare il livello di significatività $\bar{\alpha}$ (piccolo).
3. Determinare la funzione ancillare più adatta per il caso in considerazione.
4. Calcolare a partire dal campione la *statistica standard del test*, ossia la funzione ancillare per il valore soglia $\vartheta = \vartheta_0$.
5. Individuare i quantili a e b per per la statistica, come per un intervallo di confidenza a livello $1-\bar{\alpha}$. Questo determina la regione di accettazione $\mathrm{RA} = [a,b]$.
6. Accettare H_0 se la statistica standard è nella regione di accettazione, altrimenti rifiutare H_0 e accettare H_1.

Esempio 14.1. Mettiamo in pratica l'algoritmo in un caso concreto: un test sulla media μ di una popolazione Gaussiana di varianza nota $\sigma^2 = 1$. Il nostro campione ha taglia 81 e media campionaria $\bar{x} = 5.96$. Vogliamo sapere se queste osservazioni sono compatibili con una media teorica uguale a 5.38, a un livello di significatività pari al 5%.

1. Il nostro parametro incognito ϑ è la media μ. Fissato il valore soglia $\vartheta_0 = \mu_0$, in questo esempio $\mu_0 = 5.38$, le ipotesi sono H_0: $\mu = \mu_0 = 5.38$ e H_1: $\mu \neq \mu_0 = 5.38$.
2. Fissiamo il livello $\bar{\alpha} = 5\% = 0.05$.
3. La popolazione è Gaussiana e la varianza è nota. La funzione ancillare più adatta per μ è allora
$$\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1).$$
4. La statistica standard è
$$Z_0 = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{5.96 - 5.38}{1/\sqrt{81}} = 5.22.$$
Osserviamo che nel momento in cui abbiamo le osservazioni del campione (anzi, in realtà ci basta solamente il valore della media campionaria) questo è un numero, z_0.
5. Avendo fissato il livello di significatività $\bar{\alpha} = 0.05$, $1-\bar{\alpha} = 0.95$ e i quantili che ci interessano sono $a = \Phi^{-1}\left(\frac{\bar{\alpha}}{2}\right) = -\Phi^{-1}\left(1-\frac{\bar{\alpha}}{2}\right) = -\Phi^{-1}(0.975) = -1.96$ e $b = \Phi^{-1}\left(1-\frac{\bar{\alpha}}{2}\right) = \Phi^{-1}(0.975) = 1.96$. La regione di accettazione è quindi $\mathrm{RA}_Z = [-1.96, 1.96]$.
6. Siccome $Z_0 \notin \mathrm{RA}_Z$ (infatti $5.22 \notin [-1.96, 1.96]$) rifiutiamo l'ipotesi nulla e accettiamo l'ipotesi alternativa: abbiamo evidenza statistica che la media della popolazione da cui abbiamo estratto il campione non sia 5.38. Non possiamo escluderlo con certezza, ma è improbabile, in particolare c'è al più il 5% di probabilità che la media teorica sia 5.38.

A questo punto sorge spontaneo chiedersi *perché* quanto abbiamo visto funzioni. Vediamolo con qualche dettaglio: sia α la probabilità di un errore di prima specie, allora

$$\begin{aligned}
\alpha &= P(\text{dire } H_1 \mid \text{è vera } H_0) \\
&= P(Z_0 \notin \mathrm{RA}_Z \mid \text{è vera } H_0) \\
&= P(Z_0 \notin [a,b] \mid Z_0 \sim \mathcal{N}(0,1)) \\
&= P\left(Z_0 \notin \left[\Phi^{-1}\left(\frac{\bar{\alpha}}{2}\right), \Phi^{-1}\left(1-\frac{\bar{\alpha}}{2}\right)\right] \mid Z_0 \sim \mathcal{N}(0,1)\right) \\
&= \bar{\alpha}
\end{aligned}$$



in cui abbiamo usato che se è vera H_0, allora $\mu_0 = \mu$ e quindi

$$Z_0 = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \overset{H_0}{=\!=} \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$$

e che a e b sono proprio i quantili $\frac{\bar{\alpha}}{2}$ e $1 - \frac{\bar{\alpha}}{2}$ di una normale standard. Quindi, se H_0 è vera, la risposta del test è H_1 con probabilità minore o uguale al livello di significatività $\bar{\alpha}$ (mell'esempio appena fatto $\alpha = \bar{\alpha}$, ma in generale sarà $a \leqslant \bar{\alpha}$). Come nel caso degli intervalli di confidenza bilaterali, stiamo suddividendo la probabilità di errore tra lo sbagliare per eccesso e lo sbagliare per difetto.

Se invece è vera H_1, quale sarà la risposta del test? Se è vera H_1, allora $\mu_0 \neq \mu$, quindi

$$Z_0 = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} + \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1) + \Delta \sim \mathcal{N}(\Delta, 1)$$

con $\Delta \neq 0$, cioè la probabilità che la risposta del test sia H_1 è l'area tratteggiata in arancione nella Figura 14.1, la probabilità che Z_0 cada al di fuori dell'intervallo $[a, b]$. Ricordando la definizione di β come probabilità che il test risponda H_0 se è vera H_1, β è l'area non tratteggiata in arancione sotto la curva arancione.

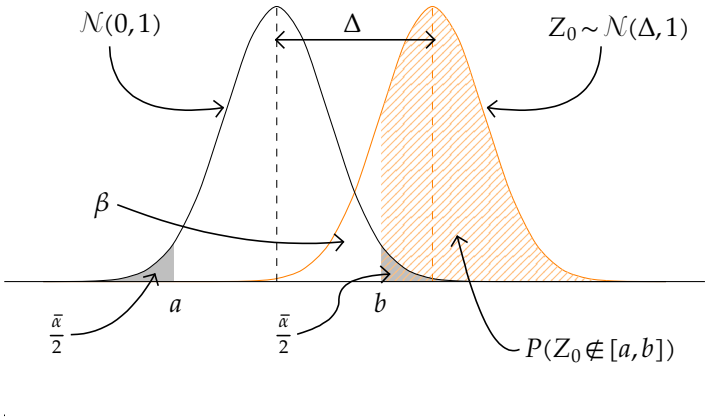


Figura 14.1. Probabilità che il test risponda H_1 se è vera H_1.

Per analizzarla meglio si può introdurre la *curva operativa caratteristica* (OCC) $\beta(\mu)$, come segue:

$$\begin{aligned}
\beta(\mu) &= P(\text{dire } H_0 | \text{la media è } \mu) = P(Z_0 \in \text{RA}_Z | \text{la media è } \mu) \\
&= P\left(a \leqslant \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \leqslant b \middle| \text{la media è } \mu \right) \\
&= P\left(a + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leqslant \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leqslant b + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \text{la media è } \mu \right) \\
&= P\left(a + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leqslant \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \leqslant b + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \text{la media è } \mu \right) \\
&= \Phi\left(b + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \right) - \Phi\left(a + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \right).
\end{aligned}$$

Chiamiamo il suo complemento a 1, ossia $1 - \beta(\mu)$, *funzione di potenza del test*. Per μ fissato, la potenza del test misura la probabilità di rifiutare H_0 (ossia rifiutare che la media sia μ_0) quando la media è μ.

Questo ci permette di fare alcune considerazioni. La funzione $\beta(\mu)$ dipende anche dalla numerosità del campione e, con gli altri parametri fissati, è decrescente in n. Allora, se vogliamo che il nostro test abbia significatività $\bar{\alpha}$ e che sia sufficientemente potente da commettere errori di seconda specie con probabilità al più $\bar{\beta}$ se il valore vero della media è $\mu_T \neq \mu_0$ (cioè $\beta(\mu_T) \leqslant \bar{\beta}$), ci basta scegliere n sufficientemente grande affinché $\beta(\mu_T) \approx \bar{\beta}$. Inoltre, se non guardiamo β, ma solamente α, come accade effettivamente il caso nell'algoritmo visto prima, accettare H_0 non significa che questa sia vera con alta probabilità, ma solamente che non abbiamo abbastanza evidenza per escluderla, cosa che è ben diversa da dire che i dati sostengono l'ipotesi nulla.

Osservazione 14.2. Una volta che abbiamo capito come funzionano le cose nel caso del test bilaterale per la media di una Gaussiana, possiamo facilmente passare a test bilaterali per altri parametri di altre popolazioni di cui conosciamo funzioni ancillari, eventualmente approssimate. Possiamo anche considerare test unilaterali, in cui l'ipotesi nulla è della forma $H_0\colon \vartheta \geqslant \vartheta_0$ e l'ipotesi alternativa è $H_1\colon \vartheta < \vartheta_0$ (o viceversa, $H_0\colon \vartheta \leqslant \vartheta_0$ e $H_1\colon \vartheta > \vartheta_0$): in questo caso la regione di accettazione sarà non più un intervallo, ma (in genere) una semiretta.

14.2. IL p-DEI-DATI

Un'altra possibile via per il test delle ipotesi è la seguente: calcoliamo una statistica, detta *p-dei-dati* o *p-value*, un po' più complicata rispetto alla statistica standard vista prima ma per cui la regione di accettazione è della forma $[\bar{\alpha}, 1]$. In altre parole, se il p-value è compreso tra $\bar{\alpha}$ e 1 accettiamo l'ipotesi nulla H_0, mentre se è tra 0 e $\bar{\alpha}$, rifiutiamo H_0 e accettiamo l'ipotesi alternativa H_1.

Osservazione 14.3. In questo caso, se il p-value è molto piccolo (ad esempio 10^{-4}), allora sceglieremo sempre H_1, mentre se è molto grande (ad esempio 0.3) accetteremo sempre H_0. Inoltre, a differenza delle regioni di accettazione ricavate prima, possiamo cambiare la soglia $\bar{\alpha}$ senza fare troppi conti: in questo caso infatti la dipendenza da $\bar{\alpha}$ è molto semplice, a differenza di quanto visto nel caso precedente. Possiamo quindi trattare meglio i casi limite o valutare al volo, senza bisogno di ricalcolare, se un'ipotesi è accettabile al 5% ma non all'1%.

L'idea alla base di questa costruzione è molto semplice: vogliamo sfruttare il fatto che le funzioni di ripartizione sono crescenti. Con il test precedente avevamo una regione di accettazione della forma $a \leqslant \Theta_0 \leqslant b$ per qualche statistica Θ calcolata in $\vartheta = \vartheta_0$ (Θ_0 è la statistica standard del test, nella nomenclatura usata prima), in cui $a = F_{\mathcal{L}}^{-1}\left(\frac{\bar{\alpha}}{2}\right)$ e $b = F_{\mathcal{L}}^{-1}\left(1 - \frac{\bar{\alpha}}{2}\right)$, in cui $\mathcal{L}$ è la legge della funzione ancillare associata alla statistica Θ. Dunque il test risponde con H_0 se e solo se $a \leqslant \Theta_0 \leqslant b$, ma qui entra in gioco l'idea, perché questa condizione è equivalente a $F_{\mathcal{L}}(a) \leqslant F_{\mathcal{L}}(\Theta_0) \leqslant F_{\mathcal{L}}(b)$, grazie alla monotonia crescente della funzione di ripartizione $F_{\mathcal{L}}$. Possiamo riscrivere la stessa condizione esplicitando la forma di a e b: $\frac{\bar{\alpha}}{2} \leqslant F_{\mathcal{L}}(\Theta_0) \leqslant 1 - \frac{\bar{\alpha}}{2}$, ossia

$$\begin{cases} \bar{\alpha} \leqslant 2F_{\mathcal{L}}(\Theta_0) \\ 2 - \bar{\alpha} \geqslant 2F_{\mathcal{L}}(\Theta_0) \end{cases}$$

da cui ricaviamo che accettiamo H_0 se e solo se $\bar{\alpha} \leqslant 2\min\left(F_{\mathcal{L}}(\Theta_0), 1 - F_{\mathcal{L}}(\Theta_0)\right)$. Il p-dei-dati è quindi $2\min\left(F_{\mathcal{L}}(\Theta_0), 1 - F_{\mathcal{L}}(\Theta_0)\right)$, una quantità numerica che possiamo ricavare dalla legge $\mathcal{L}$ della funzione ancillare associata alla statistica, dal valore target ϑ_0 del parametro ϑ che stiamo testando e dalla realizzazione del campione.

Esempio 14.4. Supponiamo di avere una popolazione Gaussiana di media μ che vogliamo testare contro un valore target μ_0 e di varianza ignota. In questo caso $H_0\colon \mu = \mu_0$ e $H_1\colon \mu \neq \mu_0$.

La statistica di riferimento in questo caso è quella per la media di una Gaussiana a σ ignota,

$$T = \frac{\overline{X}_n - \mu}{\sqrt{S^2/n}} \sim t(n-1).$$

La statistica test (che ricordiamo essere un numero) è $T_0 = (\overline{X}_n - \mu_0)\sqrt{\frac{n}{S^2}}$.

Per definizione il p-dei-dati è $2\min\left(F_{t(n-1)}(T_0), 1 - F_{t(n-1)}(T_0)\right)$, ma in questo caso possiamo sfruttare il fatto che $t(n-1)$ abbia delle proprietà di simmetria:

- se $T_0 < 0$, il minimo è $F_{t(n-1)}(T_0) = 1 - F_{t(n-1)}(-T_0)$,
- se $T_0 > 0$, il minimo è $1 - F_{t(n-1)}(T_0)$,

quindi possiamo semplificare il p-dei-dati: $2 - 2F_{t(n-1)}(|T_0|)$.

Ora non ci resta che calcolare questo numero in funzione dei parametri noti e della realizzazione del campione e confrontarlo con il livello di significatività fissato $\bar{\alpha}$: se il p-dei-dati è maggiore o uguale di $\bar{\alpha}$ accettiamo H_0, se è minore scegliamo H_1.

Abbiamo un campione di taglia 64 e vogliamo testare l'ipotesi che $\mu = 5.5$. La media campionaria è $\bar{x} = 5.213$ e la varianza campionaria è $s^2 = 3.684$. La statistica test è quindi $T_0 = -1.196$ e il p-dei-dati è 0.236, non possiamo quindi scartare l'ipotesi nulla che la media sia veramente 5.5, se non a un livello superiore al 23.6% (che sarebbe molto alto). La popolazione da cui è stato estratto il campione aveva media 5 e deviazione standard 2.

Se vogliamo usare R in un caso simile, supponiamo di avere il campione salvato nel vettore `x`. Allora la media campionaria è `mean(x)`, la varianza campionaria è $\mathrm{var}(x)$ e la statistica standard T_0 è, per $\mu_0 = 5.5$, `(mean(x)-5.5)/(sqrt(var(x)/length(x)))`. Supponendo di aver assegnato questo valore alla variabile `T_0`, possiamo calcolare il p-value usando la funzione `pt` (dal momento che la funzione ancillare è una t): `2-2*pt(abs(T_0),length(x)-1)`.

Osservazione 14.5. Possiamo vedere il p-dei-dati come la soglia critica della significatività: per tutti gli $\bar{\alpha} > p$-dei-dati rifiutiamo H_0, per tutti gli $\bar{\alpha} \leqslant p$-dei-dati accettiamo H_0. Da un altro punto di vista, il p-dei-dati è la probabilità di vedere un evento "altrettanto o più estremo" di quello osservato nel campione se è vera l'ipotesi nulla.

Lezione 26

Esempio 14.6. Consideriamo ora una popolazione Gaussiana di media e varianza ignote. Vogliamo un test statistico sul valore della varianza, $H_0: \sigma^2 = \sigma_0^2$ e $H_1: \sigma^2 \neq \sigma_0^2$.

Come prima cosa individuiamo la statistica di riferimento W e la statistica test W_0:

$$W = \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1) \qquad\qquad W_0 = \frac{S^2}{\sigma_0^2}(n-1).$$

Il p-dei-dati è, dalla definizione, $2\min\left(F_{\chi^2(n-1)}(W_0), 1 - F_{\chi^2(n-1)}(W_0)\right)$. In questo caso non abbiamo proprietà di simmetria che possiamo usare per semplificare la formulazione, ma siamo comunque in grado di calcolare questo valore a partire dai dati.

Supponiamo di avere un campione di taglia 49 per cui $s^2 = 3.744$. Se $\sigma_0^2 = 4$, allora la statistica test è $W_0 = 44.928$ e il p-dei-dati è 0.801. Non possiamo escludere che la varianza vera sia 4 (come effettivamente è nel campione da cui sono estratti i dati). Avessimo voluto testare l'ipotesi che la varianza fosse 2.5 avremmo avuto $W_0 = 71.885$ e un p-dei-dati corrispondente di 0.029. In questo caso a un livello di significatività del 5% rifiutiamo l'ipotesi nulla, ma a un livello di significatività dell'1% la accettiamo: non abbiamo in quest'ultimo caso abbastanza evidenza statistica per escludere che sia vera l'ipotesi nulla, che ricordiamo è la risposta "di default".

Anche in questo caso possiamo aiutarci con R per i calcoli. Se salviamo in `W0` la statistica test `var(x)*(length(x)-1)/4` (nel caso in cui $\sigma_0^2 = 4$), il calcolo del p-dei-dati si appoggia sulla funzione `pchisq`,

```
2*min(pchisq(W0,df=length(x)-1),1-pchisq(W0,df =length(x)-1))
```

Osservazione 14.7. Anche con il p-dei-dati è possibile impostare test statistici unilaterali, in cui l'ipotesi nulla è della forma $H_0: \vartheta \leqslant \vartheta_0$ (rispettivamente $H_0: \vartheta \geqslant \vartheta_0$) e l'ipotesi alternativa è della forma $H_1: \vartheta > \vartheta_0$ (rispettivamente $H_1: \vartheta < \vartheta_0$). Le idee sono sostanzialmente le stesse, ma come già nel caso degli intervalli di confidenza, dobbiamo stare attenti ai valori da considerare.

14.3. TEST STATISTICI UNILATERALI

Abbiamo visto, nel caso dei test bilaterali, che accettiamo l'ipotesi nulla $\vartheta = \vartheta_0$ se lo stimatore Θ non è troppo lontano da ϑ_0, né troppo più grande, né troppo più piccolo. Quanto sia "troppo" lo abbiamo quantificato in funzione della taglia del campione, del livello di significatività e di altri parametri, determinando così una regione di accettazione. Nel caso dei test unilaterali, l'ipotesi nulla è della forma $H_0: \vartheta \leqslant \vartheta_0$ (rispettivamente $H_0: \vartheta \geqslant \vartheta_0$) e l'ipotesi alternativa è della forma $H_1: \vartheta > \vartheta_0$ (rispettivamente $H_1: \vartheta < \vartheta_0$), quindi se Θ è lontano da ϑ_0, ma verso il basso, cioè è molto minore di ϑ_0 (rispettivamente molto maggiore di ϑ_0) non è un problema, non usciamo dalla regione di accettazione. Ci sarà allora una costante c (che dovremo determinare) tale che la regione di accettazione sarà della forma $\Theta - \vartheta_0 \leqslant c$ (rispettivamente $\Theta - \vartheta_0 \geqslant c$)[14.3].

Richiamiamo ora la definizione di α, cioè la probabilità di un errore di prima specie, ovvero la probabilità che il test risponda H_1 se è vera H_0. In questo caso unilaterale abbiamo

$$\alpha = P(\text{dire } H_1 \,|\, \text{è vera } H_0) = P(\Theta - \vartheta_0 > c \,|\, \vartheta \leqslant \vartheta_0)$$

e possiamo proseguire se abbiamo la distribuzione di Θ. Vediamo qualche esempio.

Esempio 14.8. Torniamo al caso della media di una popolazione Gaussiana di varianza nota. Allora $\vartheta = \mu$, $\Theta = \overline{X}$ e sappiamo anche che la statistica standard è

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1).$$

Vogliamo testare l'ipotesi nulla $H_0: \mu \leqslant \mu_0$ contro l'ipotesi alternativa $H_1: \mu > \mu_0$. Allora

$$\begin{aligned} \alpha &= P(\text{dire } H_1 \,|\, \text{è vera } H_0) = P(\overline{X} - \mu_0 > c \,|\, \mu \leqslant \mu_0) \\ &= P\left(\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} + \frac{\mu - \mu_0}{\sqrt{\sigma^2/n}} > \frac{c}{\sqrt{\sigma^2/n}} \middle| \mu \leqslant \mu_0 \right) \\ &= P\left(\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} > \frac{c + \mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \mu \leqslant \mu_0 \right). \end{aligned}$$

Sotto l'ipotesi nulla, $\mu \leqslant \mu_0$, quindi $c + \mu_0 - \mu \geqslant c$ e, il massimo di questa probabilità al variare di $\mu \leqslant \mu_0$ è nel caso $\mu = \mu_0$, perché la funzione $P(X > d)$ è decrescente in d. Per essere sicuri di avere una probabilità di errore di prima specie non superiore ad $\bar{\alpha}$ ci basta allora considerare il caso peggiore, ossia in cui $\mu = \mu_0$:

$$\alpha = P\left(\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} > \frac{c + \mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \mu \leqslant \mu_0 \right) \leqslant P\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} > \frac{c}{\sqrt{\sigma^2/n}} \right) \overset{!}{=} \bar{\alpha}.$$

Quindi

$$1 - \bar{\alpha} = P\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \leqslant \frac{c}{\sqrt{\sigma^2/n}} \right)$$

e, approfittando del fatto che $\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}}$ ha distribuzione normale standard, $c = \Phi^{-1}(1 - \bar{\alpha}) \sqrt{\sigma^2/n}$. In altre parole accettiamo l'ipotesi nulla $\mu \leqslant \mu_0$ se $\overline{X} \leqslant \mu_0 + \Phi^{-1}(1 - \bar{\alpha}) \sqrt{\sigma^2/n}$ e la rifiutiamo se invece abbiamo $\overline{X} > \mu_0 + \Phi^{-1}(1 - \bar{\alpha}) \sqrt{\sigma^2/n}$, se la significatività che abbiamo fissato è $\bar{\alpha}$.

Ovviamente anche in questo caso unilaterale possiamo ricavare il p-dei-dati, infatti rispondiamo H_0 se e solo se

$$Z_0 = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \leqslant \Phi^{-1}(1 - \bar{\alpha}) \iff \Phi(Z_0) \leqslant 1 - \bar{\alpha} \iff \bar{\alpha} \leqslant 1 - \Phi(Z_0),$$

14.3. Potremmo anche (se $\vartheta_0 > 0$) considerare il rapporto $\Theta/\vartheta_0 \leqslant c$, per un altro, opportuno, c.

ossia il p-dei-dati è $1-\Phi(Z_0)$.

Osservazione 14.9. Con il p-dei-dati quantifichiamo la probabilità (condizionata all'evento "l'ipotesi nulla è soddisfatta") di vedere una statistica del test "più estrema" di quella standard. I risultati "più estremi" sono quelli che ci aspettiamo si verifichino nel caso sia vera l'ipotesi alternativa, quindi nel caso dei test unilaterali, quelli nel verso dell'ipotesi alternativa (o equivalentemente di senso contrario all'ipotesi nulla).

Osservazione 14.10. L'asimmetria tra ipotesi nulla e ipotesi alternativa è forse ancora più evidente nel caso di test unilaterali: se consideriamo due test speculari

$H_0: \vartheta \leqslant \vartheta_0$ $H_1: \vartheta > \vartheta_0$	$H_0: \vartheta \geqslant \vartheta_0$ $H_1: \vartheta < \vartheta_0$

è possibile che accettiamo l'ipotesi nulla in entrambi (e non in uno solo dei due come ci potremmo aspettare). Il motivo di ciò è nel già citato diverso valore delle due ipotesi, nulla e alternativa. Infatti scegliamo l'ipotesi alternativa (rifiutando quindi l'ipotesi nulla) se abbiamo evidenza statistica a suo favore, ma scegliamo l'ipotesi nulla (ossia quella di default) se non abbiamo sufficiente evidenza statistica per rifiutarla.

Questo ci impone di prestare molta attenzione (soprattutto nel caso dei test unilaterali) alla scelta dell'ipotesi nulla. Essa dipenderà dal caso particolare che stiamo considerando e da quale vogliamo che sia la nostra risposta in caso non ci sia evidenza in un senso o nell'altro.

Esempio 14.11. Un'azienda produce microprocessori e sta considerando l'acquisto di una nuova linea di produzione. I microprocessori prodotti hanno una distribuzione Gaussiana e una linea di produzione è considerata affidabile se la performance dei processori prodotti in un certo benchmark ha deviazione standard non superiore a 0.15 ms. Da una prima produzione di prova della nuova linea, abbiamo un campione di taglia 20, da cui ricaviamo una varianza campionaria $s^2 = 0.025\,\text{ms}^2$. La nuova linea di produzione è affidabile o no?

Mettiamoci come prima cosa nei panni del *venditore*. Il nostro scopo è mostrare che il dato sia compatibile con il fatto che la linea di produzione sia affidabile, ossia che non c'è evidenza statistica che la linea non sia affidabile. Scegliamo allora come ipotesi del test statistico

$$H_0: \sigma^2 \leqslant (0.15)^2 = 0.0225 \qquad H_1: \sigma^2 > 0.0225.$$

Stiamo facendo un test statistico unilaterale sulla varianza di una distribuzione Gaussiana. La funzione ancillare è

$$W = \frac{S^2}{\sigma^2}(n-1) \sim \chi^2(n-1)$$

e la statistica test è $W_0 = s^2\,(n-1)\,\sigma_0^{-2}$. I valori di S^2 più vicini a 0 concordano con la nostra ipotesi nulla, quindi i valori "estremi" che ci interessano per determinare il p-dei-dati sono quelli per cui la χ^2 è maggiore della statistica test:

$$\begin{aligned} P(W > W_0) &= P\left(W > \frac{s^2}{\sigma_0^2}(n-1)\right) = 1 - F^{-1}_{\chi^2(n-1)}\left(\frac{s^2}{\sigma_0^2}(n-1)\right) \\ &= 1 - F^{-1}_{\chi^2(19)}\left(\frac{0.025}{0.0225}\cdot 19\right) = 1 - F^{-1}_{\chi^2(19)}(21.11) \\ &= 1 - 0.669 = 0.331 \end{aligned}$$

e siccome questo valore è relativamente alto, non rifiuteremo (e dunque accetteremo) l'ipotesi nulla, ossia che la linea di produzione sia affidabile, per ogni ragionevole significatività $\bar{\alpha}$.

Mettiamoci ora nei panni dell'*acquirente*. Il nostro scopo è avere evidenza statistica del fatto che la linea di produzione sia affidabile: vogliamo essere convinti che la linea sia affidabile, di default rispondiamo che non lo è. Scegliamo allora come ipotesi del test statistico

$$H_0: \sigma^2 \geqslant (0.15)^2 = 0.0225 \qquad H_1: \sigma^2 < 0.0225.$$

Stiamo sempre facendo un test statistico unilaterale sulla varianza di una distribuzione Gaussiana con la medesima funzione ancillare di prima, solo che ora i valori di S^2 più vicini a 0 non sono in accordo con la nostra ipotesi nulla, quindi i valori "estremi" che ci interessano per determinare il p-dei-dati sono quelli per cui la statistica è minore della statistica test:

$$\begin{aligned} P(W < W_0) &= P\left(W < \frac{s^2}{\sigma_0^2}(n-1)\right) = F^{-1}_{\chi^2(n-1)}\left(\frac{s^2}{\sigma_0^2}(n-1)\right) \\ &= F^{-1}_{\chi^2(19)}(21.11) = 0.669. \end{aligned}$$

Non abbiamo allora abbastanza evidenza statistica per rifiutare l'ipotesi nulla, ossia che la linea di produzione *non* sia affidabile.

Siamo allora in una situazione in cui i dati osservati non hanno abbastanza forza statistica per puntare nell'una o nell'altra direzione: non siamo in grado di rifiutare alcuna delle due ipotesi nulle, anche se esse sono apparentemente opposte. Come possiamo risolvere una situazione di stallo come questa? Raccogliendo nuovi dati.

Esempio 14.12. Supponiamo di voler fare un test sul parametro di una popolazione Bernoulliana, ad esempio per determinare se la probabilità di passare l'esame di Probabilità e Statistica (o equivalentemente la proporzione di studentesse e studenti che lo passano sul totale di chi lo ha in piano di studi) sia inferiore al 75%.

Mettiamoci nei panni dei rappresentanti degli studenti[14.4]: di default sostengono che l'esame sia troppo difficile e che la probabilità di passare sia minore o uguale al 75%. L'ipotesi nulla è quindi $H_0: p \leqslant p_0 = 0.75$, mentre l'ipotesi alternativa è $H_1: p > p_0$.

Sappiamo che uno stimatore per il parametro di una Bernoulliana è $\overline{X}$, inoltre, se moltiplichiamo $\overline{X}$ per n, abbiamo una variabile aleatoria che conta i successi all'interno del campione, di cui quindi sappiamo la distribuzione: è una variabile aleatoria binomiale di parametri n e p. Allora abbiamo la nostra statistica standard: $B \sim \mathrm{bin}(n,p)$, per essa sappiamo che vale

$$P(B \geqslant a) = \sum_{k=\lceil a \rceil}^{n} P(B=k) = \sum_{k=\lceil a \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{14.1}$$

Ricordiamoci che siamo interessati a controllare la probabilità di errori di prima specie, vogliamo cioè che la probabilità di rifiutare l'ipotesi nulla se essa è vera sia controllata dalla significatività fissata $\bar{\alpha}$. Come prima cosa, notiamo che con l'ipotesi nulla scelta i casi estremi che spingono a rifiutare H_0 sono quelli per cui $B = n\overline{X}$ è oltre una certa soglia c che dobbiamo determinare.

Osserviamo inoltre che la (14.1) è crescente se vista come funzione di p: infatti se la probabilità di successo in un singolo tentativo è maggiore, sarà maggiore anche la probabilità di ottenere almeno a successi in n tentativi, quindi sotto l'ipotesi nulla tale probabilità è massima per $p = p_0$, cioè

$$\alpha = P(\text{dire } H_1 \,|\, \text{è vera } H_0) = P(B \geqslant c \,|\, p \leqslant p_0) \leqslant \sum_{k=\lceil c \rceil}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

e se vogliamo che $\alpha \leqslant \bar{\alpha}$ dobbiamo prendere c tale che $\sum_{k=\lceil c \rceil}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} \leqslant \bar{\alpha}$, in particolare il più piccolo c tale per cui ciò valga, che denotiamo con $c_{\min}$. Allora accetteremo l'ipotesi nulla con un livello di significatività $\bar{\alpha}$ se $n\bar{x} < c_{\min}$.

Calcolare questo valore $c_{\min}$ non è semplicissimo, anche se fattibile usando qualche software, per esempio R. Tuttavia, da quanto abbiamo osservato sopra ricaviamo immediatamente quale sia il p-dei-dati: siccome vogliamo la probabilità di vedere eventi più estremi rispetto a quello misurato (ossia $n\bar{x}$) supponendo che H_0 sia vera, ci basta calcolare

$$P(B \geqslant n\bar{x} \,|\, p \leqslant p_0) \leqslant \sum_{k=n\bar{x}}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

14.4. Chiaramente esiste anche il punto di vista opposto, quello per cui in mancanza di evidenza statistica in contrario l'esame è da ritenersi di difficoltà adeguata, in cui sono scambiate ipotesi nulla e ipotesi alternativa rispetto al caso presentato nello svolgimento di questo esempio.

cioè `pbinom(q, size = n, prob = p0, lower.tail = FALSE)`, con $q = n\bar{x}$.

Se per esempio fossero passati all'esame 40 studenti su 50 del campione (cioè l'80% del campione), il p-dei-dati sarebbe 0.164, ossia non si potrebbe rifiutare l'ipotesi nulla che la probabilità di passare l'esame sia minore o uguale del 75%.

Osservazione 14.13. Nel caso Bernoulliano (come in altri casi) possiamo anche usare le statistiche approssimate che abbiamo visto grazie ai teoremi limite. Sappiamo infatti che

$$Z = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \rightsquigarrow \mathcal{N}(0,1) \qquad\qquad Z_0 = \frac{n\overline{X} - np_0}{\sqrt{np_0(1-p_0)}}.$$

In questo caso per il test delle ipotesi $H_0: p \leqslant p_0$ e $H_1: p > p_0$ a livello $\bar{\alpha}$ abbiamo che la regione di accettazione per Z_0 è della forma $(-\infty, b]$ con $b = \Phi^{-1}(1-\bar{\alpha})$ (infatti se $p \ll p_0$ la statistica Z_0 è sempre più piccola e vogliamo che cada nella regione di accettazione).

Se invece volessimo fare il test delle ipotesi $H_0: p \geqslant p_0$ e $H_1: p < p_0$ a livello $\bar{\alpha}$ abbiamo che la regione di accettazione per Z_0 è della forma $[a, +\infty)$ con $a = -\Phi^{-1}(1-\bar{\alpha}) = \Phi^{-1}(\bar{\alpha})$ (infatti se $p \gg p_0$ la statistica Z_0 è sempre più grande).

Possiamo anche calcolare il p-dei-dati (approssimato), che in questo caso è $\Phi(Z_0)$, perché il caso estremo che ci interessa è avere una probabilità più piccola di p_0 e quindi un risultato minore della statistica Z_0 calcolata dal campione. Usando gli stessi dati dell'Esempio 14.12 otteniamo un p-dei-dati approssimato uguale a 0.793, non è quindi possibile rifiutare l'ipotesi che la probabilità di passare l'esame sia maggiore o uguale al 75%.

14.4. TABELLE RIASSUNTIVE

Raccogliamo ora in alcune tabelle alcuni dei test più rilevanti.

H_0	H_1	Statistica test	Regione di accettazione (H_0)	p-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$z_0 = \dfrac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}}$	$-\Phi^{-1}\left(1-\dfrac{\bar{\alpha}}{2}\right) \leqslant z_0 \leqslant \Phi^{-1}\left(1-\dfrac{\bar{\alpha}}{2}\right)$	$2(1-\Phi(\lvert z_0\rvert))$
$\mu \leqslant \mu_0$	$\mu > \mu_0$	$z_0 = \dfrac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}}$	$z_0 \leqslant \Phi^{-1}(1-\bar{\alpha})$	$1-\Phi(\lvert z_0\rvert)$
$\mu \geqslant \mu_0$	$\mu < \mu_0$	$z_0 = \dfrac{\overline{x} - \mu_0}{\sqrt{\sigma^2/n}}$	$z_0 \geqslant -\Phi^{-1}(1-\bar{\alpha}) = \Phi^{-1}(\bar{\alpha})$	$\Phi(\lvert z_0\rvert)$

Tabella 14.3. Test delle ipotesi per la media μ di una popolazione Gaussiana di varianza σ^2 nota. Il campione ha taglia n, $\bar{x}$ è la media campionaria calcolata nel campione.

H_0	H_1	Statistica test	Regione di accettazione (H_0)	p-value
$\mu = \mu_0$	$\mu \neq \mu_0$	$t_0 = \dfrac{\overline{x} - \mu_0}{\sqrt{s^2/n}}$	$-F^{-1}_{t(n-1)}\left(1-\dfrac{\bar{\alpha}}{2}\right) \leqslant t_0 \leqslant F^{-1}_{t(n-1)}\left(1-\dfrac{\bar{\alpha}}{2}\right)$	$2(1-F_{t(n-1)}(\lvert t_0\rvert))$
$\mu \leqslant \mu_0$	$\mu > \mu_0$	$t_0 = \dfrac{\overline{x} - \mu_0}{\sqrt{s^2/n}}$	$t_0 \leqslant F^{-1}_{t(n-1)}(1-\bar{\alpha})$	$1-F_{t(n-1)}(\lvert t_0\rvert)$
$\mu \geqslant \mu_0$	$\mu < \mu_0$	$t_0 = \dfrac{\overline{x} - \mu_0}{\sqrt{s^2/n}}$	$t_0 \geqslant -F^{-1}_{t(n-1)}(1-\bar{\alpha}) = F^{-1}_{t(n-1)}(\bar{\alpha})$	$F_{t(n-1)}(\lvert t_0\rvert)$

Tabella 14.4. Test delle ipotesi per la media μ di una popolazione Gaussiana di varianza σ^2 ignota. Il campione ha taglia n, $\bar{x}$ è la media campionaria e s^2 la varianza campionaria, calcolate nel campione.

H_0	H_1	Statistica test	Regione di accettazione	p-value
$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$w_0 = \frac{s^2}{\sigma_0^2}(n-1)$	$F^{-1}_{\chi^2_{n-1}}\left(\frac{\overline{\alpha}}{2}\right) \leqslant w_0 \leqslant F^{-1}_{\chi^2_{n-1}}\left(1-\frac{\overline{\alpha}}{2}\right)$	$2\,(F_{\chi^2_{n-1}}(w_0) \wedge 1 - F_{\chi^2_{n-1}}(w_0))$
$\sigma^2 \leqslant \sigma_0^2$	$\sigma^2 > \sigma_0^2$	$w_0 = \frac{s^2}{\sigma_0^2}(n-1)$	$w_0 \leqslant F^{-1}_{\chi^2_{n-1}}(1-\overline{\alpha})$	$1 - F_{\chi^2_{n-1}}(w_0)$
$\sigma^2 \geqslant \sigma_0^2$	$\sigma^2 < \sigma_0^2$	$w_0 = \frac{s^2}{\sigma_0^2}(n-1)$	$w_0 \geqslant F^{-1}_{\chi^2_{n-1}}(\overline{\alpha})$	$F_{\chi^2_{n-1}}(w_0)$

Tabella 14.5. Test delle ipotesi per la varianza σ^2 di una popolazione Gaussiana (di media μ ignota). Il campione ha taglia n e s^2 è la varianza campionaria calcolata nel campione.

Parte III

Appendici

APPENDICE A
RICHIAMI

A.1. RICHIAMI DI TEORIA ELEMENTARE DEGLI INSIEMI

Un *insieme*, dal punto di vista matematico, è una collezione di oggetti detti *elementi*. Esso può essere caratterizzato *per estensione*, andando a elencarne tutti gli elementi. È il modo forse più naturale, ma è possibile solamente se l'insieme è finito ed è pratico solo se l'insieme ha pochi elementi. In alternativa, possiamo caratterizzare un insieme mediante le proprietà soddisfatte da tutti e soli i suoi elementi. In questo caso parliamo di definizione *intensiva*.

Se però dobbiamo lavorare con più di un insieme, ci piacerebbe avere un modo per confrontarli e identificarli. Diciamo che due insiemi A e B sono uguali e scriviamo $A = B$ se ciascuno è sottoinsieme dell'altro, $A \subseteq B$ e $B \subseteq A$, ossia se tutti gli elementi di A sono anche elementi di B e viceversa.

D'altra parte ci sono, soprattutto in combinatoria, occasioni in cui non ci interessa sapere quali sono gli elementi di un insieme, ma solamente quanti sono. Di conseguenza vogliamo identificare due insiemi che abbiano lo stesso numero di elementi (anche se gli elementi non sono gli stessi). Da questo punto di vista un insieme con sei foglie non è diverso da un insieme con sei palline o con sei punti. Questo è molto "matematico": estraiamo e astraiamo dagli oggetti solo quelle proprietà che ci interessano, ignorando tutte le altre.

Vogliamo contare il numero di elementi di un insieme, cioè conoscere la sua *cardinalità*. Useremo il simbolo $\#A$ per indicare la cardinalità di un insieme A. Questa può essere un numero (naturale) finito, ma anche infinito, sia numerabile, denotato con $\aleph_0$, sia pari al continuo, denotato con $2^{\aleph_0}$. Per il momento ci limitiamo a insiemi con un numero finito di elementi, cioè insiemi di cardinalità finita.

Torniamo agli insiemi e alle loro operazioni. Cominciamo con l'intersezione e l'unione. L'*intersezione* di due insiemi A e B è l'insieme, denotato con $A \cap B$ che contiene tutti gli elementi che appartengono sia ad A che a B, cioè

$$A \cap B = \{x : x \in A \wedge x \in B\}.$$

L'*unione* di due insiemi A e B è l'insieme, denotato con $A \cup B$ che contiene tutti gli elementi che appartengono ad almeno uno tra A e B, cioè

$$A \cup B = \{x : x \in A \vee x \in B\}.$$

Un'altra operazione è quella di differenza tra due insiemi: l'insieme $A \setminus B$ contiene tutti gli elementi che appartengono ad A, ma non a B. Viceversa l'insieme $B \setminus A$ contiene tutti gli elementi di B che non sono anche elementi di A,

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

C'è, per ogni insieme A, una particolare collezione di sue rappresentazioni diverse: la collezione delle partizioni di A. Dato un insieme A, una sua *partizione* è una famiglia $\mathcal{S}$ di sottoinsiemi di A tali che ogni elemento di A appartiene a uno e uno solo degli insiemi in $\mathcal{S}$. In altre parole, $\bigcup_{S \in \mathcal{S}} S = A$ ed è un'unione disgiunta: due insiemi non coincidenti $S, T \in \mathcal{S}$ hanno intersezione vuota. Vale la pena osservare che nella definizione di partizione non è richiesto che A sia non vuoto. L'insieme vuoto ha un'unica partizione, l'insieme vuoto stesso[A.1].

A.1. Si potrebbe fare un'osservazione filosofica che la partizione dell'insieme vuoto non è l'insieme vuoto stesso, ma è semplicemente a esso isomorfa: infatti nel secondo caso gli elementi che non sono nell'insieme vuoto sono loro stessi insiemi (i sottoinsiemi dell'insieme vuoto). Un bel grattacapo, che possiamo lasciare tranquillamente ai logici e ai teorici degli insiemi.

Per tornare verso la combinatoria e la probabilità, possiamo chiederci quante siano, per un insieme finito, le partizioni possibili. Se abbiamo un insieme finito di cardinalità n, il numero delle sue partizioni è B_n, l'n-esimo numero di Bell[A.2] (come mostrato nel Problema 1.1).

Abbiamo già introdotto alcune operazioni tra insiemi, unione, intersezione e differenza. Queste operazioni sono binarie, perché coinvolgono due insiemi. C'è però un'altra importante operazione unaria per gli insiemi: il complementare. Dato un insieme A contenuto nell'insieme universo Ω (cioè $A \subseteq \Omega$) il *complementare* di A è l'insieme denotato con A^c che contiene tutti gli elementi di Ω non contenuti in A.

- Intersezione e unione sono *idempotenti*, cioè $A \cap A = A$ e $A \cup A = A$.
- Intersezione e unione sono *commutative*, cioè $A \cap B = B \cap A$ e $A \cup B = B \cup A$.
- Intersezione e unione sono *associative*, cioè $A \cap B \cap C = (A \cap B) \cap C = A \cap (B \cap C)$ e $A \cup B \cup C = (A \cup B) \cup C = A \cup (B \cup C)$.
- Intersezione e unione sono *distributive* l'una rispetto all'altra, cioè $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ e $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- Il complementare è *involutorio*, ossia è l'operazione inversa di se stesso, cioè $(A^c)^c = A$.
- La differenza può essere scritta in termini di intersezioni e complementare, $A \setminus B = A \cap B^c$.

In realtà ci è sufficiente avere una sola delle operazioni tra unione e intersezione, perché grazie al risultato seguente possiamo scrivere l'operazione binaria rimanente in termini dell'operazione binaria che conosciamo e del complementare.

TEOREMA A.1. (LEGGI DI DE MORGAN[A.3]) *Se A e B sono due insiemi, valgono le seguenti identità:*

- $(A \cap B)^c = A^c \cup B^c$
- $(A \cup B)^c = A^c \cap B^c$.

Vediamo ora alcuni modi di scrivere la differenza simmetrica tra due insiemi $A \triangle B$, definita come $A \triangle B = (A \setminus B) \cup (B \setminus A)$. La dimostrazione del prossimo risultato è un esercizio di teoria degli insiemi, che richiede le leggi di De Morgan.

PROPOSIZIONE A.2. *Dati due insiemi A e B, i seguenti insiemi sono uguali: $A \triangle B$, $(A \cup B) \setminus (A \cap B)$, $(A \cup B) \cap (A \cap B)^c$, $(A \cup B) \cap (A^c \cup B^c)$, $(A \cap B^c) \cup (A^c \cap B)$, $A^c \triangle B^c$.*

Se, dati due insiemi A e B, esiste una funzione iniettiva $f: A \to B$, allora $\#A \leqslant \#B$. Se inoltre non esiste una funzione biettiva tra i due, possiamo dire che $\#A < \#B$. Attenzione: nel momento in cui iniziamo a maneggiare gli infiniti dobbiamo procedere con estrema cautela. Infatti non è necessariamente vero (all'interno della teoria degli insiemi) che valga la proprietà di tricotomia, ossia che dati due insiemi debba essere vera una delle seguenti: $\#A < \#B$, $\#A = \#B$ o $\#B < \#A$. Il problema si ha con i cardinali infiniti e, in particolare, la tricotomia equivale all'assioma della scelta.

Esempio A.3. Consideriamo l'insieme $2\mathbb{N}$ dei numeri naturali pari. Esso è un sottoinsieme proprio dell'insieme $\mathbb{N}$ dei numeri naturali. Tuttavia $2\mathbb{N}$ e $\mathbb{N}$ hanno la stessa cardinalità. Infatti possiamo prendere $f: \mathbb{N} \to 2\mathbb{N}$ tale che $f(n) = 2n$. La funzione f è biettiva: la sua inversa è f^{-1}: $2\mathbb{N} \to \mathbb{N}$ con $f^{-1}(2m) = m$. Allora i due insiemi $2\mathbb{N}$ e $\mathbb{N}$ sono equipotenti, ossia ci sono tanti numeri naturali pari quanti numeri naturali.

Dobbiamo quindi procedere con cautela, come mostrato anche dal seguente risultato.

TEOREMA A.4. (CANTOR[A.4]-BERNSTEIN[A.5]) *Dati due insiemi A e B, se esistono due funzioni iniettive f: $A \to B$ e $g: B \to A$, allora esiste almeno una funzione biettiva tra i due insiemi. In altri termini, se $\#A \leqslant \#B$ e $\#B \leqslant \#A$, allora $\#A = \#B$.*

A.2. Eric Temple Bell (1883 – 1960).

A.3. Augustus De Morgan (1806 – 1871).

A.4. Georg Cantor (1845 – 1918).

A.5. Felix Bernstein (1878 – 1956).

Questa affermazione sulla cardinalità di due insiemi sembra assolutamente ovvia. Ma, come abbiamo visto nell'Esempio A.3, l'uso degli infiniti può trarre in inganno. Perciò il teorema è necessario; la sua dimostrazione, comunque, non è per niente banale.

Esempio A.5. Dato un insieme A, le funzioni da A a $\{0,1\}$ formano un insieme di cardinalità $2^{\#A}$. Infatti una funzione da A a $\{0,1\}$ associa a ogni elemento di A uno tra 0 e 1 e la scelta per un elemento di A non influenza quella per gli altri. Quindi abbiamo 2 scelte per ciascun elemento e i fattori 2 devono essere moltiplicati tra loro. Gli elementi di A sono $\#A$, da cui il risultato.

In generale possiamo dire qualcosa di più: le funzioni da un insieme A a un insieme B formano un insieme di cardinalità $\#B^{\#A}$. Per questo motivo l'insieme delle funzioni da A a B si scrive B^A.

PROPOSIZIONE A.6. *Dato un insieme A di cardinalità eventualmente infinita (anche più che numerabile) l'insieme delle parti di A (o insieme potenza di A) $\mathcal{P}(A)$ ha cardinalità $\#\mathcal{P}(A) = 2^{\#A}$.*

Dimostrazione. Come detto in precedenza, per mostrare che un insieme ha una certa cardinalità, quello che possiamo fare è costruire una relazione biunivoca (o una codifica) dal nostro insieme a un insieme che sappiamo avere la cardinalità cercata. Sappiamo anche che un insieme che ha proprio $2^{\#A}$ elementi è l'insieme delle funzioni da A in $\{0,1\}$. Quello che ci resta da fare, dunque, è far vedere che i sottoinsiemi di A sono tanti quanti le funzioni da A in $\{0,1\}$. Per ogni sottoinsieme $S \subseteq A$ definiamo la funzione $f_S : A \to \{0,1\}$ come segue: $f_S(a) = 1_S(a)$. In sostanza, codifichiamo con un 1 la presenza dell'elemento nel sottoinsieme, con 0 la sua assenza. Viceversa per ogni funzione $f : A \to \{0,1\}$ possiamo definire $S_f = f^{-1}(1)$, cioè il sottoinsieme di Ω contenente tutti gli a per cui $f(a) = 1$. Si verifica facilmente che le relazioni $S \to f_S$ e $f \to S_f$ sono entrambe iniettive[A.6], quindi $\#\mathcal{P}(A) \leqslant 2^{\#A} \leqslant \#\mathcal{P}(A)$ (Teorema di Cantor-Bernstein) e abbiamo l'uguaglianza cercata. □

TEOREMA A.7. (CANTOR) *Non esiste alcuna funzione suriettiva da un insieme A al suo insieme delle parti $\mathcal{P}(A)$. In particolare, quindi, $\#A < \#\mathcal{P}(A)$.*

Dimostrazione. Cominciamo osservando che $\mathcal{P}(A)$ contiene una copia di A: la funzione $i : A \to \mathcal{P}(A)$ che manda ogni elemento di A nel suo singoletto è iniettiva, quindi $\#A \leqslant \#\mathcal{P}(A)$.

Procediamo ora per assurdo e supponiamo di avere una funzione $f : A \to \mathcal{P}(A)$ suriettiva. Consideriamo l'insieme $N = \{a \in A : a \notin f(a)\}$ degli elementi di A che non appartengono alla propria immagine mediante f. Dal momento che f è suriettiva su $\mathcal{P}(A)$, esiste un elemento $\alpha \in A$ tale che $f(\alpha) = N$. A questo punto abbiamo una contraddizione: se $\alpha \in N$, allora dalla definizione di N segue $\alpha \notin f(\alpha) = N$. Se invece $\alpha \notin N$, allora $\alpha \in f(\alpha) = N$. Dunque non può esistere una funzione suriettiva da A a $\mathcal{P}(A)$.

Se non possono esistere funzioni suriettive, non possono in particolare esistere funzioni biiettive: pertanto i due insiemi hanno cardinalità diversa e vale la disuguaglianza stretta. □

PROPOSIZIONE A.8. *L'insieme $\mathcal{P}(\mathbb{N})$ ha cardinalità uguale a quella dei numeri reali.*

Dimostrazione. L'idea di Cantor che sta alla base di questa dimostrazione è far vedere che possiamo identificare i sottoinsiemi dei numeri naturali con i numeri reali nell'intervallo $[0,1]$. Questo intervallo, a sua volta, ha tanti elementi quanti tutti i numeri reali.

Cominciamo con la prima parte. Per prima cosa, forti di quanto visto sopra, identifichiamo $\mathcal{P}(\mathbb{N})$ con l'insieme delle successioni binarie. Vogliamo interpretarle come rappresentazioni binarie dei numeri reali in $[0,1]$, considerandole come se fossero le cifre dopo la virgola. In questo modo abbiamo tutti e soli[A.7] i numeri reali in $[0,1]$.

A.6. Sono anche suriettive e in particolare l'una è l'inversa dell'altra.

A.7. In realtà stiamo un po' imbrogliando: come nel caso delle rappresentazioni decimali abbiamo il problema dei numeri che finiscono con un 9 periodico o con uno 0 periodico. Questi numeri vengono ``contati'' due volte, quindi abbiamo un problema simile con i numeri che finiscono con 1 periodico o con 0 periodico, che possiamo caratterizzare come i numeri con un numero finito di cifre. Tuttavia con un po' di accortezza siamo in grado di aggirare questo ostacolo, tenendo conto che questi numeri sono in quantità numerabile.

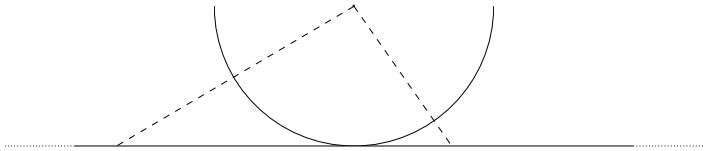


Figura A.1. Una biiezione tra $(0,1)$ e $\mathbb{R}$

Ora vogliamo identificare $[0,1]$ con l'intera retta reale. In realtà identifichiamo l'intervallo aperto $(0,1)$ con la retta reale. Per fare ciò deformiamo il segmento senza estremi $(0,1)$ in una semicirconferenza, anch'essa senza estremi. Prendiamo la retta reale e disegniamola in modo che sia tangente al punto medio della semicirconferenza. A questo punto possiamo tracciare le semirette uscenti dal centro della semicirconferenza che intersecano la semicirconferenza stessa, come rappresentato in Figura A.1. Ciascuna di esse incontra la retta reale in uno e un solo punto. Abbiamo così stabilito una relazione biunivoca tra ogni punto della semicirconferenza (e quindi ogni numero nell'intervallo $(0,1)$) e ogni punto della retta reale (cioè ogni numero reale). □

I risultati precedenti danno due informazioni interessanti riguardo ad alcuni insiemi che abbiamo appena visto.

COROLLARIO A.9. *L'insieme $\mathbb{R}$ dei numeri reali ha cardinalità $2^{\aleph_0}$ strettamente maggiore della cardinalità $\aleph_0$ dell'insieme dei numeri naturali.*

COROLLARIO A.10. *L'insieme $\mathcal{P}(\mathbb{R})$ ha cardinalità $2^{(2^{\aleph_0})}$, che in particolare è più grande di quella di $\mathbb{R}$.*

Dimostrazione. La prima parte segue dalla Proposizione A.6. La seconda parte dal Teorema A.7. □

Due ultime curiosità, prima di andare oltre la cardinalità. Ci sono infinite cardinalità infinite, di cui $\aleph_0$ non è che la prima. Osserviamo però che non abbiamo detto che la cardinalità $2^{\aleph_0}$ dei reali (detta anche *continuo* e indicata con $\mathfrak{c}$) sia il secondo numero cardinale infinito (cioè $\aleph_1$). Non lo abbiamo fatto perché non è (necessariamente) vero: è la famosa *ipotesi del continuo*.

A.2. SERIE ARITMETICA E SERIE GEOMETRICA

Per la serie aritmetica ci interessa sapere che

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2}.$$

Una serie si dice *geometrica* se è della forma $\sum_{k=0}^{+\infty} r^k$, per qualche $r \in \mathbb{R}$. Possiamo considerare a parte il caso $r=1$ (una somma di 1), per cui la somma diverge a $+\infty$. Per $r \neq 1$, consideriamo la somma troncata $s_n = \sum_{k=0}^{n} r^k$. Allora, moltiplicando s_n per r e sottraendo da s_n abbiamo

$$s_n - r s_n = \sum_{k=0}^{n} r^k - r^{k+1} = 1 - r^{n+1},$$

cioè $s_n = \frac{1-r^{n+1}}{1-r}$.

Se vogliamo il comportamento per $n \to +\infty$, osserviamo che per $|r|>1$, $|r^{n+1}| \to +\infty$, quindi la serie non converge, per $r=-1$ la serie oscilla tra 0 e 1 (quindi non converge), mentre per $|r|<1$, $r^{n+1} \to 0$ per $n \to +\infty$ e quindi $s_n \to s = (1-r)^{-1}$.

A.3. L'INTEGRALE GAUSSIANO

Ci interessano gli integrali definiti

$$I_1 = \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x \qquad I_2 = \int_{0}^{+\infty} e^{-\frac{x^2}{2}}\,\mathrm{d}x \qquad I = \int_{0}^{+\infty} e^{-x^2}\,\mathrm{d}x$$

Osserviamo che $I_1 = 2I_2$ perché la funzione $e^{-x^2/2}$ è simmetrica rispetto a $x = 0$. Inoltre $I_2 = I\sqrt{2}$, come si vede con un cambio di variabile. Ci basta allora calcolare l'integrale indefinito I.

Questo integrale si può calcolare in molti modi, pur non avendo e^{-x^2} una primitiva. Vediamone alcuni.

Iniziamo definendo le due funzioni ausiliarie

$$f(t) = \int_0^t e^{-x^2}\,\mathrm{d}x \qquad g(t) = \int_0^1 \frac{e^{-t^2(1+x^2)}}{1+x^2}\,\mathrm{d}x$$

e osserviamo che

$$(f(t))^2 + g(t) = \text{const},$$

infatti

$$\frac{\mathrm{d}}{\mathrm{d}t}(f(t))^2 = 2f(t)f'(t) = 2\int_0^t e^{-x^2}\,\mathrm{d}x\, e^{-t^2} = 2\int_0^t e^{-x^2-t^2}\,\mathrm{d}x$$

ma, allo stesso tempo, se facciamo il cambio di variabili $x = yt$,

$$2\int_0^t e^{-x^2-t^2}\,\mathrm{d}x = \int_0^1 2e^{-t^2(y^2+1)}\,t\,\mathrm{d}y.$$

D'altro canto,

$$g'(t) = \int_0^1 \frac{e^{-t^2(1+x^2)}\,(-2t)(1+x^2)}{1+x^2}\,\mathrm{d}x = -\int_0^1 2e^{-t^2(1+x^2)}\,\mathrm{d}x.$$

Quindi per ogni t, $(f(t))^2 + g(t) = (f(0))^2 + g(0)$ e quindi

$$\begin{aligned} I^2 &= \lim_{t\to+\infty} (f(t))^2 = g(0) - \lim_{t\to+\infty} g(t) \\ &= \int_0^1 \frac{1}{1+x^2}\,\mathrm{d}x - \lim_{t\to+\infty} \int_0^1 \frac{e^{-t^2(1+x^2)}}{1+x^2}\,\mathrm{d}x \\ &= \Big[\operatorname{atan}(x)\Big]_0^1 - 0 \\ &= \frac{\pi}{4} \end{aligned}$$

da cui abbiamo $I = \sqrt{\pi}/2$.

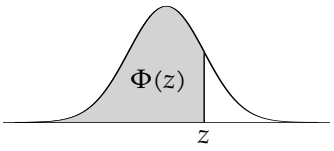
[TBC] Altri modi per calcolarlo si possono trovare in questo documento: https://kconrad.math.uconn.edu/blurbs/analysis/gaussianintegral.pdf.

APPENDICE B

TAVOLE

Tavole della funzione di ripartizione per una distribuzione normale standard

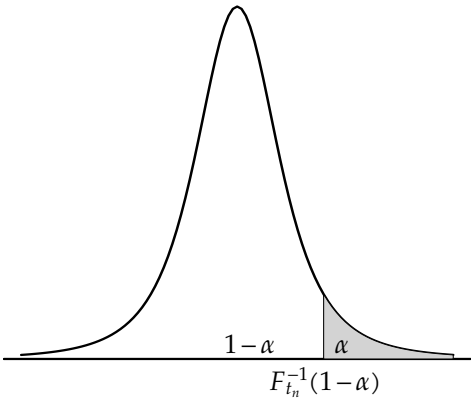
$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



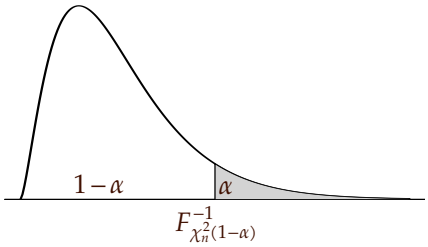
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Tavole dei quantili per una distribuzione t di Student

df α	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
$+\infty$	1.282	1.645	1.960	2.326	2.576



Tavole dei quantili per una distribuzione χ^2



df α	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.00004	0.00016	0.001	0.004	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	124.342	129.561	135.807	140.169

Come si leggono le tavole?

Cominciamo dalla tavola per Φ. Supponiamo di voler calcolare $\Phi(1.26)$. Allora cerchiamo nella prima colonna la riga corrispondente a 1.2 e, su quella riga, individuiamo la cella nella colonna corrispondente a 0.06. Abbiamo allora $\Phi(1.26) \simeq 0.8962$. Se invece volessimo calcolare $\Phi(-0.78)$, come prima cosa ricordiamo che $\Phi(-0.78) = 1 - \Phi(0.78)$ poi cerchiamo quest'ultimo nella tabella, all'incrocio tra la riga 0.7 e la colonna 0.08: $\Phi(0.78) \simeq 0.7823$, quindi $\Phi(-0.78) \simeq 0.2177$.