Note di Calcolo delle Probabilità e Statistica Matematica

LUIGI AMEDEO BIANCHI

Versione: 3 giugno 2025

INDICE

INTRODUZIONE

Queste note intendono coprire quanto presentato nell'insegnamento *Calcolo delle Probabilità e Statistica Matematica* tenuto nel secondo semestre dell'anno accademico 2024/25 al Corso di Laurea triennale in Matematica. Si basano, espandendole, sulle note degli insegnamenti *Probabilità e Statistica* e *Probability* dei Corsi di Laurea Triennali in Informatica e in Ingegneria delle Comunicazioni ed Elettronica.

Parte di queste note (le prime lezioni) è stata pubblicata dalla casa editrice Scienza Express nella collana UMath. Tra le fonti cui queste note devono molto ci sono: le note del prof. Agostinelli, le lezioni passate del prof. Bonaccorsi, le lezioni del prof. Francesco Morandin dell'Università di Parma e il libro di probabilità di Sheldon Ross (nelle sue varie declinazioni).

Le note sono scritte in $\TeX_{\text{MACS}}$ (`https://www.texmacs.org`). Alcune delle immagini sono realizzate in TikZ, altre in R.

COSA È LA PROBABILITÀ?

Questo è un corso sulla Probabilità, quindi un buon punto di partenza sarebbe capire cos'è la Probabilità. La Probabilità è un ramo della Matematica che si occupa dell'incertezza. Se ci pensiamo per un attimo, questo suona strano: naturalmente associamo la matematica alla certezza, dopotutto, qualcosa che non può essere messo in discussione viene spesso definito *matematico*. Quindi prima di procedere, dobbiamo affrontare almeno due punti: cosa intendiamo per *incertezza* e come questo approccio che stiamo per approfondire sia matematico.

La Matematica è in generale un buon linguaggio per una descrizione sintetica dei fenomeni nel mondo reale. I primi esempi che ci vengono in mente sono quelli provenienti dalla Fisica, ma questo è solo parte del quadro, poiché possiamo utilizzare la matematica anche per trattare altre scienze, come quelle naturali o sociali. In tutti questi casi, in misura diversa, ci preoccupiamo di fenomeni (che spesso chiameremo *esperimenti*) che mostrano una certa variabilità (o *incertezza*) nei loro risultati. Queste incertezze possono essere legate a diverse prospettive: potrebbero esserci differenze nella misurazione dei risultati degli esperimenti, differenze nel tempo, differenze nello spazio, differenze tra casi o soggetti. In alcuni casi conosciamo le perturbazioni (o *rumore*) delle misurazioni, ma questo in generale non è sufficiente per distinguere il rumore dalla misurazione effettiva. La teoria della probabilità fornisce strumenti per gestire queste incertezze, in particolare per quantificarle e caratterizzarle.

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.

Una domanda immediata a questo punto potrebbe essere la seguente: *abbiamo davvero bisogno di una teoria matematica per questo*? *Non possiamo semplicemente fare affidamento sui nostri sensi*? È una domanda completamente legittima: dopotutto, la matematica riguarda tutto ciò che riguarda l'evitare calcoli non necessari, quindi se siamo in grado di procedere senza il costo di definire un'intera teoria coerente, tanto meglio.

Un'altra domanda potrebbe nascere da una visione della Matematica come qualcosa di più di uno strumento per applicazioni, più di un linguaggio per il mondo reale: qualcosa di bello e piacevole in sé, nella sua astrazione e nella sua forma. Come vedremo la Probabilità è (o può essere) anche questo: come spesso accade in matematica astrazione e applicazione sono due aspetti che convivono, legati in modo indissolubile, ma allo stesso tempo a nostra disposizione per poter scegliere tra i due quello che più ci attira e conquista, in modo che possa guidarci nello scoprire la Probabilità.

Esempio 1. Se lanciamo un dado bilanciato a sei facce, è più probabile che esca un 6 o un 1? È più probabile che, lanciandolo due volte, vediamo due 6 o due numeri diversi? In questo caso, la nostra intuizione ci aiuta abbastanza bene: nel primo caso il 1 è altrettanto probabile quanto il 6, mentre nel secondo caso (senza entrare troppo nei dettagli) ottenere due sei è meno probabile poiché abbiamo ``molto più'' risultati con due numeri diversi (vedremo che abbiamo 30 risultati con risultati diversi sui due dadi).

Esempio 2. È più probabile ottenere almeno un 6 lanciando un dado (lo stesso del primo esempio) 4 volte o è più probabile ottenere almeno quattro 6 lanciando il dado 24 volte?

Questo è più difficile da risolvere, senza alcuni strumenti tecnici che non abbiamo ancora, che è proprio il punto di questo esempio: possiamo comprendere il tipo di situazione, ma ci rendiamo conto che va oltre un calcolo rapido. Vedremo come risolvere questo e come calcolare la probabilità per entrambi i casi più avanti.

Non è un caso che questi primi due esempi, come molti altri in seguito, vengano dal mondo del gioco d'azzardo: in fondo, storicamente, l'origine della probabilità è proprio nei giocatori d'azzardo che tentavano di guadagnare qualche vantaggio sui propri avversari. In effetti l'Esempio 2 è ispirato a un problema storico, posto da Antoine Gombaud, Chevalier de Méré.

Esempio 3. È più probabile ottenere un 6 lanciando un dado quattro volte o ottenere almeno un doppio 6 lanciando due dadi 24 volte?

L'opinione di De Méré era che poiché ottenere un doppio sei con due lanci è 1/6 tanto probabile quanto ottenere un 6 con un solo lancio (cosa corretta, come vedremo), allora i due casi sopra dovrebbero essere ugualmente probabili, poiché stiamo lanciando i dadi doppi sei volte il numero di lanci del singolo dado. In questo caso si sbagliava.

In realtà, Gombaud chiese aiuto a Pascal e a Fermat per avere conferma della correttezza del suo ragionamento e che quindi gli altri giocatori stessero imbrogliandolo, ma i due matematici arrivarono alla risposta corretta, mostrando che non c'era alcuna evidenza di imbroglio. Questa non fu l'unica volta in cui Gombaud chiese supporto ai matematici su questioni di probabilità e matematica, ma non sempre apprezzò le risposte che ricevette.

Nel tempo, tuttavia, la Probabilità si è estesa un poco oltre il mondo del gioco d'azzardo ed è oggi uno strumento indispensabile in molti ambiti, oltre ad essere un oggetto di studio astratto.

Esempio 4. 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*.

Prima di passare a qualche contenuto più matematico, prendiamoci un minuto per discutere la storia e la filosofia della Probabilità. Prima di tutto: cosa intendiamo per storia e filosofia? La storia è la parte facile: quando è stata sviluppata la teoria della Probabilità da un punto di vista matematico. D'altra parte, sul lato filosofico, abbiamo almeno due domande da affrontare: quali idee e comprensioni del mondo hanno reso possibile anche solo iniziare a pensare alla probabilità e quali interpretazioni per la probabilità esistono (sia nel passato che oggi).

Per quanto riguarda la storia, come già accennato, la Probabilità è nata dal gioco d'azzardo. In particolare, Gerolamo Cardano è considerato il precursore, avendo scritto un libretto sulla probabilità (*De Ludo Aleae*) nel 1564. Il suo interesse era piuttosto pratico: aveva spesso problemi finanziari e cercava di superarli giocando d'azzardo (con successo). Infatti, non pubblicò mai il suo libro, che apparve solo dopo la sua morte. Lo studio del gioco d'azzardo era ancora la principale motivazione per Blaise Pascal, Pierre de Fermat e altri, che lo studiarono intorno al 1650. I primi approcci matematici (in un certo senso) rigorosi furono l'*Ars Conjectandi* di Jakob Bernoulli e la Doctrine of Chances di Abraham de Moivre, entrambi intorno al 1715 (ancora una volta, Bernoulli era già morto quando il suo libro fu pubblicato). Altri nomi importanti nella storia della Probabilità sono quelli di Pierre-Simon de Laplace e Karl Friedrich Gauss. Come accade per gran parte della matematica moderna, un fondamento formale e solido della teoria venne alla luce solo recentemente, attraverso il già menzionato lavoro di Andrei Kolmogorov.

Questa storia spiega anche alcune delle diverse prospettive filosofiche sulla Probabilità. La prima è l'*interpretazione classica* della Probabilità come rapporto tra il numero dei casi favorevoli e il numero totale dei casi possibili, assumendo che tali casi siano tutti ugualmente probabili. Questo è un approccio che funziona bene per problemi di gioco d'azzardo semplici (e non così semplici), ma impone alcune forti restrizioni sui fenomeni che possiamo trattare, come vedremo tra poco. Una prospettiva diversa è il *frequentismo*. In questo caso, la probabilità di un particolare risultato è definita come il limite della frequenza relativa dell'evento quando il fenomeno incerto viene osservato più volte nel tempo, assumendo che nulla cambi. Questa è una definizione molto sensata in molti casi. Tuttavia, ha alcuni limiti, poiché rende impossibile discutere la probabilità, ad esempio, dell'esito di una particolare elezione o partita sportiva: nessuna due elezioni sono uguali, anche se i candidati non cambiano, i tempi sono cambiati e così hanno alcune delle condizioni circostanti. Una possibile soluzione è l'interpretazione *soggettivista* della probabilità, che afferma, grosso modo, che una probabilità è solo una misura della credenza soggettiva su un fenomeno incerto. Questo è, come discuteremo meglio più avanti nel corso, affatto in contrasto con un trattamento matematico della probabilità: non è consentito assegnare la probabilità in

modo arbitrario, ci sono ancora regole da seguire, in particolare la credenza deve essere aggiornata con nuovi dati attraverso il teorema di Bayes, tanto che il soggettivismo è spesso classificato come interpretazione *bayesiana* della probabilità. Esistono anche altre interpretazioni della probabilità, come l'approccio *logico*, promosso da Rudolf Carnap.

Vale la pena sottolineare che queste interpretazioni non sono necessariamente mutuamente esclusive: molte persone ricorrono a una interpretazione o all'altra a seconda del contesto che stanno considerando, e molte altre hanno posizioni più sfumate. Inoltre, come vedremo, uno dei grandi vantaggi della teoria matematica di Kolmogorov è che è indipendente dall'interpretazione che diamo alla probabilità: ci dirà semplicemente come manipolarla da un punto di vista matematico. Se vogliamo possiamo leggerla come teorema di caratterizzazione della probabilità, senza un corrispondente teorema di unicità. 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, come abbiamo visto, da posizioni filosofiche.

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 pensare che la probabilità traduce (o rappresenta) l'incertezza in termini matematici permettendoci così di usare quest'ultima per studiare rigorosamente situazioni non deterministiche.

Questo insegnamento ha nel nome, oltre alla probabilità, 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.

Ora, dopo questa introduzione storica e filosofica, torniamo a un po' di contesto per le definizioni e vediamo ancora un esempio.

Esempio 5. Ci sono, in una borsa di velluto scuro, 60 monete, tutte della stessa dimensione: 20 di esse sono fatte di ottone (di densità 8,5 $\mathrm{g \cdot cm^{-3}}$), 20 sono fatte di acciaio (7,8 $\mathrm{g \cdot cm^{-3}}$), e 20 sono fatte di oro (19,2 $\mathrm{g \cdot cm^{-3}}$). Cosa possiamo dire sulla probabilità di estrarre una moneta d'oro?

C'è molto da analizzare in questa domanda, in particolare cosa significa *estrarre una moneta* in questo problema? Si potrebbe pensare che sia una domanda oziosa, ma vedremo che non è il caso. Ad esempio:

i. potremmo prendere una moneta dalla borsa senza né guardarla né pesarla
ii. potremmo svuotare la borsa su una superficie e prendere la moneta che si trova in cima alla pila risultante
iii. potremmo guardare e pesare le monete e scegliere quella che ci aspettiamo sia fatta d'oro
iv. continuiamo ad estrarre monete una per una dalla borsa e ci fermiamo quando pensiamo di averne presa una d'oro (o quando arriviamo all'ultima moneta).

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 la probabilità classica: 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?

Il terzo e quarto caso ``forzano'' una prospettiva soggettiva sul problema, e dovremmo tenere conto non solo delle proprietà fisiche delle monete, ma anche della nostra fiducia nelle nostre capacità percettive.

Ancora una volta, la morale è *fare attenzione* quando si legge o si (ri)enuncia un problema. Abbiamo bisogno di definire in modo più chiaro la situazione che consideriamo[1]. 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.

1. Naturalmente ci saranno convenzioni, per evitare di specificare tutti i parametri di ciascun esperimento.

CAPITOLO 1

STATISTICA DESCRITTIVA

Si dice spesso che la Matematica è il linguaggio della Scienza. Un particolare significato che si può dare a questa affermazione è il seguente: ogni scienza sperimentale raccoglie dati e questi dati vanno poi elaborati e interpretati. La Statistica fornisce metodi per farlo.

1.1. PRIME DEFINIZIONI

La nozione di partenza per un'indagine statistica è quella di popolazione.

DEFINIZIONE 1.1. *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 1.2. 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,
- le possibili partite di Go,
- le civiltà sottomesse dai Vogon.

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*. Nel primo caso abbiamo misure sull'intera popolazione e vogliamo *descrivere* alcune caratteristiche della popolazione stessa a partire da queste misure, calcolandone opportune funzioni che riassumano tutte le informazioni in una quantità ridotta di numeri o indicatori. Molto spesso cerchiamo, con la Statistica descrittiva, di riassumere le informazioni contenute nei dati in maniera più compatta, solitamente usando pochi numeri.

Esempio 1.3. Nel censimento della popolazione italiana viene raccolto, per ogni persona, il sesso. L'informazione completa sarebbe una stringa (un vettore) di lunghezza circa sessanta milioni (ossia pari alla popolazione residente in Italia al momento del censimento) i cui elementi sono "Maschio" o "Femmina". Spesso sono informazioni più utili il conto delle occorrenze dei due sessi o la loro proporzione sul totale.

La Statistica inferenziale, 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.

In Statistica siamo solitamente interessati a una o più caratteristiche di ogni individuo, ad esempio al sesso, all'altezza, al livello di educazione raggiunto…

DEFINIZIONE 1.4. *Le caratteristiche che misuriamo prendono il nome di* variabili*, i valori che assumono si chiamano* valori, 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 (significativamente) 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 e possiamo dare un senso alla differenza tra valori, ma non al loro rapporto,*
- rapporto, *se lo 0 è fissato in modo naturale e ha senso considerare il rapporto tra valori.*

Convenzione 1.5. Spesso identificheremo un individuo con la misurazione (il dato) della sua caratteristica di interesse, con un abuso di notazione che però è chiarito dal contesto.

Esempio 1.6. Vediamo alcuni esempi di variabili dei diversi tipi. La temperatura in Celsius o Fahrenheit è l'esempio standard di variabile numerica continua con scala intervallo, assieme alle coordinate in qualunque sistema di riferimento o la direzione rispetto a un "Nord" fissato. Viceversa, massa ed energia sono esempi di variabili numeriche continue in scala rapporto.

Variabili qualitative nominali possono essere il colore dei capelli o il tipo di mezzo di trasporto usato, mentre sono ordinali i livelli di educazione o i ranking tra università, così come i voti o i giudizi di gradimento di un corso.

Il tipo di una variabile è rilevante per determinare quali operazioni sono ammesse (o comunque ben definite) su di essa. Nei linguaggi come R è opportuno specificare il tipo di una variabile, come vedremo, perché fa sì che certe operazioni siano segnalate come errate. È infatti possibile che una variabile qualitativa (il possedere o meno un'automobile, ad esempio) sia *codificata* in maniera numerica (ad esempio 0/1 per no/sì). Ingenuamente si potrebbe pensare di sommare o in generale trattare come *numeri* queste modalità, ma ciò è in generale scorretto.

DEFINIZIONE 1.7. *Data una popolazione* $\{x_1,\dots,x_n\}$ *di taglia (o cardinalità o numerosità)* n *di osservazioni di una variabile (numerica o qualitativa) avente un numero finito di livelli, che codifichiamo con i numeri naturali* $\{1,\dots,K\}$*, la* frequenza assoluta *o* effettivo *del k-simo livello il numero naturale* $N_k := \#\{j \in \{1,\dots,n\} : x_j = k\}$ *per* $k \in \{1,\dots,K\}$*, ossia la cardinalità dell'insieme degli individui che assumono il valore k.*

La frequenza relativa *del k-simo livello è la quantità (razionale non negativa)* $p_k := \frac{N_k}{n}$*, ossia la proporzione di osservazioni all'interno della popolazione che assumono il valore k.*

Ci sono alcune osservazioni immediate che si possono fare (e che possono essere usati come *sanity check* di eventuali dati raccolti). La somma delle frequenze assolute deve essere uguale alla taglia della popolazione: $N_1 + \cdots + N_K = n$. Di conseguenza le frequenze relative sommano a 1.

Nel caso di una variabile continua[1] non è pratico definire in questo modo le frequenze assolute e relative. Infatti la variabile ha un numero potenzialmente infinito (più che numerabile) di valori possibili e quindi per la popolazione moltissimi livelli con 1 sola osservazione. Un possibile modo di risolvere questa situazione è suddividere l'intervallo $[\bigwedge_{i=1}^n x_i, \bigvee_{i=1}^n x_i]$, ossia il più piccolo intervallo che contiene tutti i valori assunti dalla (variabile nella) popolazione in un numero finito K di sottointervalli consecutivi (disgiunti) che lo ricoprano, $I_1, \ldots, I_K$, detti *classi*. A questo punto definiamo per $k \in \{1,\ldots,K\}$ le frequenze assolute associate agli intervalli $N_k := \#\{j \in \{1,\ldots,n\} : x_j \in I_k\}$ e le frequenze relative analogamente a prima.

1.2. RAPPRESENTAZIONI GRAFICHE

Per variabili con un numero finito di livelli è possibile fare un diagramma a barre o *barplot*. Se si tratta di una variabile ordinata è opportuno preservare l'ordine nell'asse delle ascisse. Se si tratta di una variabile numerica (discreta) è opportuno indicare tutte i livelli, anche quelli che hanno frequenza nulla o estremamente ridotta.

L'altezza delle barre è proporzionale alle frequenze (assolute o relative) dei vari livelli.

Esempio 1.8. Vogliamo rappresentare le informazioni sulla scuola di provenienza di chi ha risposto al questionario durante la prima lezione. Abbiamo i seguenti dati (già riassunti per livello): Liceo Scientifico (55), Istituto Tecnico (3), Altro Liceo (3), Altro Istituto (0). Li possiamo rappresentare in un grafico a barre con il seguente codice R `barplot(height = x$numeri, names.arg = x$scelta)` ottenendo l'immagine (veramente di base) del grafico a barre, in

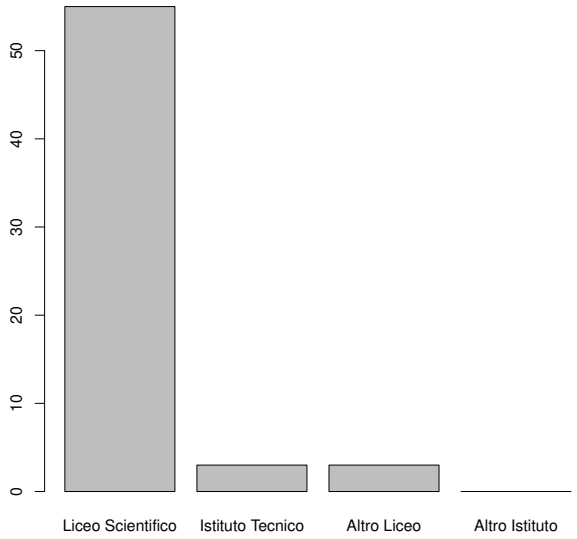


Figura 1.1. Un esempio di grafico a barre

1. Sembra rimanere fuori il caso di una variabile numerica discreta ma con un insieme infinito di valori. Cosa possiamo dire in quel caso?

Per variabili numeriche continue usiamo la medesima idea mostrata poco fa di suddividere l'intervallo dei valori assunti in sottointervalli e poi costruiamo su ciascun sottointervallo un rettangolo che abbia area proporzionale proporzionale alla corrispondente frequenza relativa. Otteniamo in questo modo un istogramma.

Osservazione 1.9. L'altezza delle barre di un istogramma è proporzionale alle frequenze relative (e assolute) solo nel caso in cui tutti i sottointervalli I_k abbiano la medesima ampiezza.

Questa scelta di proporzionalità delle aree può apparire inizialmente strana, ma come vedremo meglio in seguito è del tutto naturale, dal momento che ci garantisce un importante invariante. Osserviamo anche a seconda di come decidiamo di suddividere l'intervallo "originale" in sottointervalli possiamo ottenere rappresentazioni grafiche molto diverse tra loro, anche se descrivono la medesima popolazione. Non esiste un algoritmo o una regola fissa per determinarle, ma si tratta di *scelte di modello*. Un punto di partenza per l'esplorazione di queste possibili suddivisioni è dato dalla *Regola di Sturges* che, per una popolazione di n osservazioni, suddivide l'intervallo dei valori assunti in $1 + \frac{\log n}{\log 2}$ sottointervalli di uguale ampiezza. Non bisogna poi dimenticare di stabilire la convenzione su quale dei due estremi di ciascun intervallo sia aperto e quale chiuso e quale comportamento ci si aspetta dagli intervalli più esterni.

Esempio 1.10. Vediamo invece le risposte al problema 3 della prima lezione (quando è stato proposto la prima volta). In questo caso le risposte sono probabilità, quindi ci aspettiamo una variabile continua a valori nell'intervallo $[0,1]$. In effetti è presente una risposta 25, che decidiamo di rimuovere dal grafico (è interessante vedere cosa succede a lasciarla, sia nella scelta automatica degli intervalli, sia nella rappresentazione generale). Due possibili istogrammi sono i seguenti, ottenuti con la funzione R `hist`, nel primo caso con le suddivisioni (`breaks`) di default, nel secondo con suddivisioni specificate dall'utente.

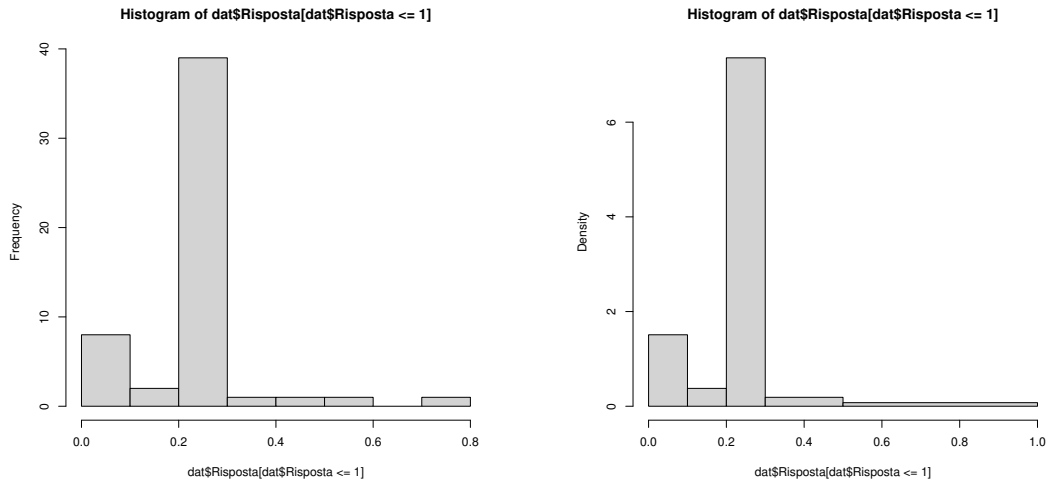


Figura 1.2. Due istogrammi sugli stessi dati. Che differenze si possono notare?

Se per una popolazione abbiamo 2 variabili (ad esempio altezza e peso) può essere interessante rappresentarle entrambe. In questo caso abbiamo i diagrammi di dispersione o scatterplot in cui su un asse rappresentiamo una variabile e sull'altro la seconda. Nel caso di variabili numeriche ogni individuo viene rappresentato dalle sue coordinate cartesiane. Nel caso di variabili non numeriche è sempre possibile codificare numericamente i livelli e successivamente rappresentare cartesianamente. Nel primo di questi casi ha senso considerare la distanza (euclidea) tra gli individui, nel secondo non necessariamente.

Esistono in realtà molti altri aspetti della rappresentazione visiva di dati (data visualisation o dataviz), alcuni dei quali verranno toccati più avanti. Vale la pena sottolineare che nella maggior parte dei casi le scelte fatte nella visualizzazione come la scala o il range degli assi non sono scelte neutrali, ma influenzano significativamente la percezione che si ha di quei dati.

1.3. INDICI DI CENTRALITÀ

Data una popolazione possiamo essere interessati a caratterizzarne in qualche senso il centro. L'accento in questo caso è su "in qualche senso", perché esistono molti modi di individuare un centro. Si parla in generale di indici di centralità. In generale il tipo di indici definiti dipenderà dalle operazioni ammesse sulle variabili e quindi, in ultima battuta, dal tipo delle variabili stesse.

DEFINIZIONE 1.11. *Chiamiamo* moda *di una variabile il livello di massima frequenza assoluta. Nel caso di una variabile continua possiamo parlare di moda delle classi* (*cioè dei sottointervalli*).

Nonostante la formulazione usata nella definizione di moda, essa non è necessariamente unica. La moda individua come "centro" della distribuzione il livello più frequente, ossia quello più comune. È un indice ben definito per tutti i tipi di variabile, anche se "funziona meglio" con quelle qualitative o numeriche discrete. In R base non è disponibile una funzione per calcolare la moda, ma è possibile scriversi del codice[2] che lo faccia oppure caricare un pacchetto che la contenga.

DEFINIZIONE 1.12. *Per una variabile ordinata una* mediana *è il minimo livello tale per cui almeno metà della popolazione è minore o uguale a esso e almeno metà della popolazione è maggiore o uguale a esso.*

Osserviamo che anche se "minore o uguale" ci fa pensare ai numeri questo indice è ben definito per ogni variabile ordinata, non solo per quelle numeriche. Osserviamo anche che se abbiamo una popolazione di taglia dispari $n = 2m+1$ allora (almeno) l'elemento $(m+1)$-simo della popolazione ordinata è una mediana. Non è detto che sia l'unico perché potrebbero esserci più osservazioni con lo stesso livello e tutte queste sono quindi mediane. Se invece la popolazione ha taglia pari $n = 2\,m$ allora la mediana è l'elemento m-simo della popolazione ordinata. Per variabili numeriche alle volte per una popolazione di taglia pari si considera come mediana la media aritmetica tra gli elementi m ed $m+1$ della popolazione ordinata. In R la mediana è calcolata dalla funzione `median`.

L'indice di centralità più noto è quello più restrittivo in termini di variabili ammesse.

DEFINIZIONE 1.13. *La* media (aritmetica) *di una variabile numerica* $\{x_1, \dots, x_n\}$ *è la quantità*

$$\bar{x} := \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

La funzione R che calcola la media è `mean`.

Osservazione 1.14. Media e mediana non sono necessariamente uguali, anzi sono quasi sempre diverse tra loro.

2. Una funzione che calcola la moda in R (proposta su stackoverflow) è la seguente (attenzione che restituisce solamente la prima occorrenza della moda, in caso di dati multimodali… può essere un esercizio interessante modificare il codice per considerare tutte le occorrenze, appoggiandosi alla funzione `tabulate` e alla funzione `match`).

```
Mode <- function(x, na.rm = FALSE) {
  if(na.rm){x = x[!is.na(x)]}
  ux <- unique(x)
  return(ux[which.max(tabulate(match(x, ux)))])
}
```

Osservazione 1.15. Se di una variabile (numerica) non conosciamo tutte le osservazioni ma solamente i vari livelli $z_1, \dots, z_K$ e le corrispondenti frequenze assolute $N_1, \dots, N_K$ o relative $p_1, \dots, p_K$, possiamo calcolare la media aritmetica in maniera del tutto equivalente come

$$\bar{x} = \sum_{k=1}^{K} p_k z_k = \frac{1}{n} \sum_{k=1}^{K} (n p_k) z_k = \frac{1}{n} \sum_{k=1}^{K} N_k z_k = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Può succedere di volere o dovere trasformare una variabile osservata. Molto spesso queste trasformazioni sono lineari. In questo caso ci interessa sapere come cambiano i nostri indici di centralità. Dati $a, b \in \mathbb{R}$, sia $y = a x + b$ la trasformazione lineare che stiamo considerando. Allora: la moda della popolazione y è la trasformazione lineare della moda, la mediana di y è la trasformazione lineare della mediana e la media di y è la trasformazione della media:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} (a x_i + b) = a \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) + b \left(\frac{1}{n} \sum_{i=1}^{n} 1 \right) = a \bar{x} + b.$$

Se invece abbiamo due popolazioni $y_1, \dots, y_m$ e $z_1, \dots, z_l$, la loro unione $x_1, \dots, x_{m+l}$ (ammesso che abbia senso) ha come media

$$\bar{x} = \frac{1}{m+l} \sum_{i=1}^{m+l} x_i = \frac{1}{m+l} \left(\sum_{i=1}^{m} y_i + \sum_{i=1}^{l} z_i \right) = \frac{1}{m+l} (m \bar{y} + l \bar{z}) = \frac{m}{m+l} \bar{y} + \frac{l}{m+l} \bar{z},$$

ossia la media pesata in base alla numerosità relativa delle due popolazioni delle due medie aritmetiche. Non esiste un risultato analogo per la moda e per la mediana.

A differenza della media, moda e mediana sono poco sensibili a perturbazioni dei dati (si dice che sono indici *robusti*).

Esempio 1.16. Il reddito medio a Villazzano (5025 dichiaranti) è 25653 euro, il reddito mediano è 21000 euro. Si trasferisce a Villazzano una persona dal reddito dichiarato di 6.8 milioni di euro. Possiamo chiederci come cambino i due indici. La mediana rimarrà sostanzialmente identica, la media invece diventa 27000 euro.

Esempio 1.17. Nelle risposte al problema 2 della prima lezione (dopo il lavoro di gruppo), qualche persona ha inserito 67, probabilmente pensando a 67%. Di conseguenza la media della risposta (che ricordiamo doveva essere una probabilità) è 2.23, mentre la mediana è 0.666.

1.4. INDICI DI DISPERSIONE

La varianza è la quantità $\sigma^2 := \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$. La deviazione standard è la radice quadrata della varianza. Il coefficiente di variazione CV è il rapporto tra la deviazione standard e la media, in valore assoluto (con la convenzione che valga infinito se la media è nulla).

Anche per la varianza abbiamo delle forme alternative che coinvolgono le frequenze assolute o relative:

$$\sigma^2 = \sum_{k=1}^{K} p_k (z_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^{K} N_k (z_k - \bar{x})^2.$$

Ma la forma alternativa più importante è

$$\begin{aligned} \sigma^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i^2 - 2 x_i \bar{x} + \bar{x}^2) \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \bar{x}^2 - 2 \bar{x} \frac{1}{n} \sum_{i=1}^{n} x_i \\ &= \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \end{aligned}$$

ossia è la media dei quadrati meno il quadrato della media.

Se i dati vengono sottoposti a una traslazione, varianza (e deviazione standard) rimangono invariate. Se i dati vengono riscalati (moltiplicati per una costante $a \in \mathbb{R}$) la varianza cambia in maniera quadratica (di a^2) mentre la deviazione standard varia di un fattore $|a|$.

Per le variabili categoriche (ma anche per tutte le altre) abbiamo l'indice di variazione definito come $v := 1 - \frac{f_m}{n} = 1 - p_m$ dove f_m è la frequenza assoluta della moda e p_m la sua frequenza relativa. In alternativa possiamo considerare l'entropia (di Shannon[3]) $h := -\sum_{k=1}^{K} p_k \log(p_k)$ (dove il logaritmo può essere in una base a scelta e tale base determina l'unità di misura).

Abbiamo quindi costruito indici di variabilità basati sulla media (e quindi solo per variabili numeriche) e sulla moda (per tutte le variabili). Ora cerchiamo di fare lo stesso appoggiandoci alla mediana, cercando quindi di coprire variabili numeriche e categoriche ordinate. Chiamiamo *quantile di ordine* α (per $\alpha \in (0,1)$) il livello $x_{(i)}$, ossia l'elemento della popolazione ordinata (cosa indicata dalle parentesi) in posizione $i = \lfloor \alpha (n+1) \rfloor$ (anche questo, come la mediana che stiamo generalizzando è soggetto a convenzioni non condivise globalmente: alcuni testi usano invece $i = \lceil \alpha (n+1) \rceil$ o un'opportuna interpolazione, soprattutto nel caso di dati numerici, tra i due livelli corrispondenti). Moralmente l'idea è che una proporzione α dei dati (ossia αn osservazioni) sono minori (o uguali) a quello individuato. Alcuni quantili speciali sono quelli corrispondenti ad $\alpha = 0.5$ che è la mediana (detto anche secondo quartile o cinquantesimo percentile), $\alpha = 0.25$ che è il primo quartile (o venticinquesimo percentile) e $\alpha = 0.75$ che è il terzo quartile (o settantacinquesimo percentile).

Osserviamo a questo punto che una misura di dispersione (per variabili numeriche) potrebbe essere il *range* ossia l'ampiezza dell'intervallo (o l'intervallo stesso se vogliamo allargarci alle categoriche ordinate). Questo però è spesso troppo poco informativo. Una versione più raffinata è quindi l'intervallo interquartile, ossia quello compreso tra il primo e il terzo quartile (che contiene la metà "centrale" dei dati osservati). Nel caso di variabili numeriche possiamo considerarne anche l'ampiezza.

Una rappresentazione grafica molto utile di una popolazione numerica (o ordinale, anche se è meno comune) è mediante i cosiddetti boxplot. In questa rappresentazione la variabile di interesse è indicata verticalmente (solitamente) e vengono tracciate una scatola (box) avente come estremo inferiore il primo quartile e come superiore il terzo quartile. Viene tracciata al suo interno la mediana, come segmento orizzontale. All'esterno della scatola vengono indicati (con una T e una T rovesciata) gli estremi (rispettivamente superiore e inferiore) dell'intervallo in cui la variabile assume i valori[4]. Il boxplot è particolarmente utile quando vogliamo confrontare la stessa variabile su popolazioni diverse.

Esempio 1.18. Uno studio sulla biodiversità marina vuole capire l'impatto di alcuni fitofarmaci sulle Posidonie (piante sottomarine). Vengono individuati diversi campioni di Posidonie, di cui viene misurata la densità (fasci per metro quadrato), prima e dopo un trattamento con uno dei due fitofarmaci considerati (o con nessuno dei due).

Di ognuno di essi si ha la densità di fasci per metro quadro prima del trattamento `initial`, il tipo di trattamento `treatment` (i valori 1 e 2 indicano i due diversi fitofarmaci, 3 è il gruppo di controllo che non ha ricevuto alcun trattamento) e la densità dopo il trattamento `final`. Nell'immagine qui sotto vediamo il boxplot della variabile `final` separatamente per le tre popolazioni individuate dai tre trattamenti ricevuti.

3. A questo si lega tutta la parte di teoria della comunicazione/dell'informazione secondo Shannon, che viene in parte coperta nell'insegnamento di Calcolo delle Probabilità II.

4. Questo non è sempre vero: alcuni software (ad esempio R) di default non si allontanano troppo dai quartili: se ci sono valori più estremi vengono marcati come outlier.

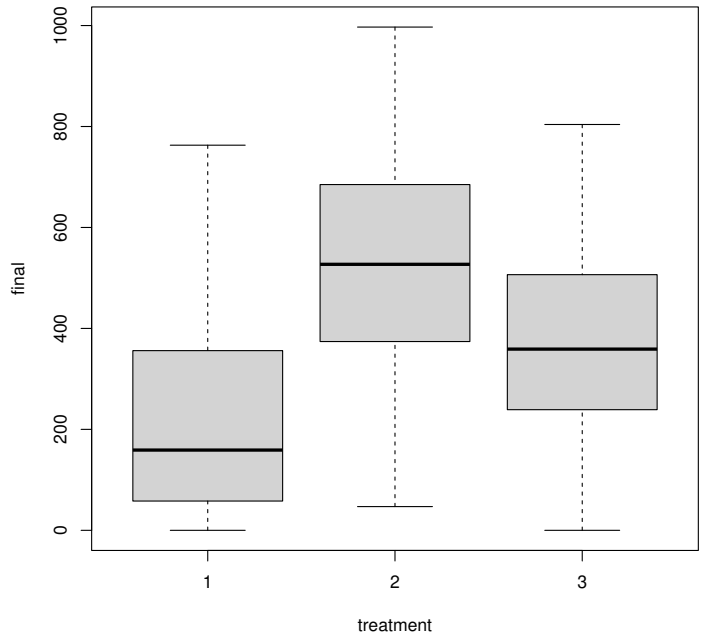


Figura 1.3. Boxplot

1.5. OLTRE LA MEDIA ARITMETICA

[TBC Medie p-sime (escono in maniera naturale in certi contesti). Disuguaglianza tra le medie. A differenza di media, moda e mediana qui sappiamo come sono ordinate.]

1.6. PROBLEMI

Problema 1. Cosa succede ai vari indici nel caso di trasformazioni non lineari?

Problema 2. Abbiamo riformulato moda e media in termini di frequenze. Possiamo fare lo stesso per la mediana (e i quantili in generale)? In che modo?

Parte I

Calcolo delle probabilità

CAPITOLO 2

PRIMI PASSI NELLA PROBABILITÀ

Come punto di partenza prendiamo la nozione di *esperimento aleatorio*.

DEFINIZIONE 2.1. *Un esperimento si dice* aleatorio *o* casuale *se, coi dati 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.

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 2.2. *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 2.3. 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, rosso e nero. Estrarre una biglia significa prenderne una dall'urna, senza guardarla finché non è stata tirata fuori.

Possibili esperimenti aleatori in questo contesto possono essere:

i. estraiamo una singola biglia dall'urna. In questo caso $\Omega = \{B, N, R\}$, avendo abbreviato i colori con le iniziali dei loro nomi (in italiano);

ii. estraiamo 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 è costituito dalle coppie ordinate

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

iii. estraiamo 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[1].

Osservazione 2.4. 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?

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] 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.

Esempio 2.5. Nell'estrazione con reimmissione vista nell'Esempio 2.3 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.

Altrimenti potremmo essere interessati a tutte le estrazioni singole che non contengono una biglia bianca. Non stiamo più parlando di singoli esiti, ma di collezioni di esiti.

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[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, $E_4 = \{3\}$…

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

Non dimentichiamo però il nostro obiettivo: vogliamo definire una misura di incertezza, che chiameremo *probabilità* nel contesto degli esperimenti aleatori. 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. Questo è quello che abbiamo chiamato approccio *classico* alla probabilità.

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*.

2. L'equivalenza in termini probabilistici di esperimenti aleatori verrà affrontata più avanti, nel Capitolo 6.

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.

Dovremo contare elementi di sottoinsiemi di Ω, quindi 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 (ma vedremo che questo non ci aiuterà davvero con la probabilità). 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[4].

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.

2.1. I TRE PRINCIPI DELLA COMBINATORIA

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 2.7. *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 2.8. 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 2.9. 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 2.10. *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.

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.

PROPOSIZIONE 2.11. *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 2.12. 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 2.13. *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 2.14. 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 2.15. 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 2.16. 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[5] 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.

5. 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.

Osservazione 2.17. 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 2.18. 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 2.19. *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:

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

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 2.20. *Con n insiemi $A_1, \dots, A_n$ 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 2.21. 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).

2.2. 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 $n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1$, 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 2.22. 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 2.23. 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:

$$\text{ANAGRAM}_1\text{M}_2\text{I}=\text{ANAGRAM}_2\text{M}_1\text{I}.$$

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 2.24. 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[6].

Esempio 2.25. 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 2.26. 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.

6. 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.

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 2.27. 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:

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

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 2.28. 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 2.29. 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 2.28, 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.

2.3. COMBINAZIONI E COEFFICIENTE BINOMIALE

Passiamo a un problema diverso, anche se l'ambientazione è analoga a quella dell'Esempio 2.28.

Esempio 2.30. 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 2 sottoinsiemi, di k ed $n-k$ elementi: 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 2.31. 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 2.32. *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 7. □

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[7] (o di Pascal[8]), 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 2.33. 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.

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

2.4. 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 2.34. Se nel Dipartimento di Matematica 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, Delladio, Pagani, Perotti e Zunino?

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 2.35. 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.

7. Niccolò Fontana detto Tartaglia (1499 circa – 1557).
8. Blaise Pascal (1623 – 1662).

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%.

2.5. 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 2.36. 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 2.37. 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[9]. 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.

9. In realtà per Andrea non sono molto favorevoli.

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,

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

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.

2.6. PROBLEMI

Problema 3. Quante sono le partizioni di un insieme di cardinalità $n=10$?

Problema 4. Quante sono le funzioni biiettive da un insieme A di cardinalità finita in se stesso (dette anche automorfismi)?

Problema 5. Quante sono le funzioni iniettive da un insieme A a un insieme B, entrambi di cardinalità finita?

Problema 6. Quante sono le funzioni suriettive da un insieme A a un insieme B, entrambi di cardinalità finita?

Problema 7. (PROPOSIZIONE 2.32) 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. $\binom{n}{k}+\binom{n}{k+1}=\binom{n+1}{k+1}$.

Problema 8. Qual è la probabilità che in un'aula con 70 studenti almeno 2 abbiano lo stesso compleanno?

CAPITOLO 3
UNA NUOVA PROBABILITÀ

Nel Capitolo 2 abbiamo iniziato a introdurre gli ingredienti fondamentali per parlare di probabilità: esperimenti aleatori (il contesto), esiti (risultati possibili dell'esperimento), insieme universo (insieme di tutti gli esiti) ed eventi, sottoinsiemi dell'insieme universo.

In particolare ora mettiamo la nostra attenzione su questi ultimi, perche, come accennato, sono quelli su cui vogliamo definire la probabilità. Finora non abbiamo dato una definizione formale di eventi, ci siamo limitati a dire che gli eventi sono sottoinsiemi dell'universo Ω. Ma sono tutti i sottoinsiemi o solo alcuni?

3.1. 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 3.1. *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 3.1 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 3.1 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 3.2. *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 3.3. 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 3.

Esempio 3.4. 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 3.1 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 3.5. *Una famiglia $\mathcal{F}$ di sottoinsiemi di un insieme Ω è una* tribù (*o* σ-algebra[1]) *se valgono tutte le seguenti proprietà:*

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

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

1. 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. Una discussione in merito è in Appendice A.2. Nella pratica tra i probabilisti il termine σ-algebra è sdoganato.

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 3.6. 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.

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 3.7. *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 3.8. *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. □

Il risultato precedente di estende a intersezioni finite, a intersezioni numerabili ma anche a intersezioni su famiglie arbitrariamente grandi. L'idea è la medesima: un evento che appartiene all'intersezione delle algebre (tribù) appartiene a ciascuna algebra (tribù), quindi anche il suo complementare appartiene a tutte le algebre (tribù) e dunque alla loro intersezione e lo stesso per l'unione (numerabile).

Potrebbe a questo punto sorgere la domanda se un risultato analogo valga per le unioni di tribù.

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 3.9. *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 3.10. 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 3.11. *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.

Esempio 3.12. 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. $(0,1)$

Procediamo in ordine sparso. Anzi, proprio dal fondo: consideriamo l'unione (numerabile) di tutti gli intervalli:

$$\bigcup_{n \in \mathbb{N}} \left[\frac{1}{2^{n+1}}, \frac{1}{2^n}\right) = (0,1).$$

Con quest'ultima idea possiamo osservare che

$$\bigcup_{n \in \mathbb{N}^+} \left[\frac{1}{2^{n+1}}, \frac{1}{2^n}\right) = \left(0, \frac{1}{2}\right)$$

che quindi è nella tribù generata.

Anche l'intervallo chiuso $[0,1] = (0,1) \cup \{0,1\}$ è nella tribù generata. Inoltre, similmente a $\left(0, \frac{1}{2}\right)$, anche $\left(0, \frac{1}{4}\right)$ è nella tribù, quindi deve esserci anche

$$(0,1) \cap \left(0, \frac{1}{4}\right)^c = (0,1) \cap \left((-\infty, 0] \cup \left[\frac{1}{4}, +\infty\right)\right) = \left[\frac{1}{4}, 1\right).$$

Per ottenere questo risultato sarebbe bastato anche osservare che $\left[\frac{1}{4}, 1\right) = \left[\frac{1}{4}, \frac{1}{2}\right) \cup \left[\frac{1}{2}, 1\right)$.

Invece non sembra possibile ottenere $\left[\frac{1}{4}, 1\right]$, perché non siamo in grado di "separare" 0 e 1. Similmente non possiamo avere $\{0\}$ né $\{1\}$. Se avessimo $\left[0, \frac{1}{2}\right]$, allora avremmo anche $\{0\}$, $\left\{\frac{1}{2}\right\}$ e $\{1\}$, ma anche altre implicazioni sono vere.

3.2. 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 3.13. *Dati un insieme Ω e una*[2] *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 A. Kolmogorov[3], che ci dice quali sono le proprietà che deve soddisfare una funzione per essere accettabile come funzione di probabilità.

DEFINIZIONE 3.14. *Assegnato uno spazio probabilizzabile $(\Omega, \mathcal{F})$, una funzione $P\colon \mathcal{F} \to \mathbb{R}$ si dice* funzione o misura[4] 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 $P(\bigcup_{i=1}^{+\infty} E_i) = \sum_{i=1}^{+\infty} P(E_i)$* ($\sigma$-additività).

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

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 3.15. *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 3.16. 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 3.17. Prendiamo ora $\Omega = \{\clubsuit, \blacklozenge, \heartsuit, \spadesuit\}$ e $\mathcal{F} = \mathcal{P}(\Omega)$. Della probabilità P sappiamo quanto segue:

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

2. 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.

3. Andrei Kolmogorov (1903 – 1987).

4. 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 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.

Esempio 3.18. 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.

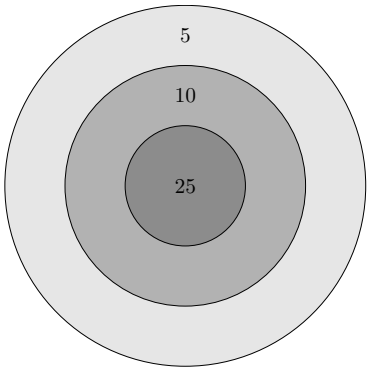


Figura 3.1. Freccette

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?

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 3.19. Prendiamo $\Omega = \{0,1\}$, $\mathcal{F} = \mathcal{P}(\Omega)$, cioè lo stesso spazio probabilizzabile visto nell'Esempio 3.16. 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.

3.3. PROPRIETÀ DELLA (MISURA DI) PROBABILITÀ

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 3.14 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 3.20. *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 3.14 e la terza identità dal terzo assioma in versione finita. Da questa identità ricaviamo $P(\emptyset) = 0$. □

PROPOSIZIONE 3.21. *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 3.14. □

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 3.22. In un ``Gratta e vinci''[5] 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 3.23. *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 3.14, per la disuguaglianza la non negatività del primo assioma. □

Questo risultato dice formalmente quanto avevamo visto nell'Esempio 4, 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 3.24. *L'immagine della funzione di probabilità è contenuta nell'intervallo unitario* $[0,1]$.

Dimostrazione. Segue immediatamente dal fatto che, per ogni evento $E\in\mathcal{F}$, vale la catena di inclusioni $\emptyset\subseteq E\subseteq\Omega$ e dalla monotonia (Proposizione 3.23). □

Vediamo ora qualcosa che apparentemente abbiamo già incontrato: la probabilità dell'unione di due eventi. Questa volta, tuttavia, lasciamo cadere la richiesta che avevamo nella versione finita del terzo assioma nella Definizione 3.14, ossia non chiediamo che i due eventi siano tra loro disgiunti.

Esempio 3.25. 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?

Cominciamo osservando che, siccome stiamo trattando una probabilità, il valore minimo è sicuramente maggiore o uguale di 0 (e minore o uguale di 1). A priori, in effetti, l'intersezione potrebbe essere vuota, portando al valore 0 per la probabilità. Tuttavia questo equivarrebbe a dire che i due insiemi sono mutualmente esclusivi, caso in cui potremmo sommare le due probabilità che conosciamo: $\frac{17}{24}+\frac{5}{6}=\frac{37}{24}$. Ci accorgiamo però facilmente di aver sbagliato qualcosa, dal momento che questa quantità è maggiore di 1: chiaramente siamo contando qualcosa più volte, l'intersezione dei due insiemi.

5. 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*.

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 $p\geqslant\frac{13}{24}$.

Anche se non è richiesto dal problema, possiamo dire qualcosa del valore massimo possibile per questa probabilità? Una prima idea potrebbe essere vedere quali sono le conseguenze di chiedere che $\frac{37}{24}-p\geqslant 0$, ma questo ci dà una stima abbastanza irrilevante: $p\leqslant\frac{37}{24}$. Avevamo già più informazioni dal fatto che $p\leqslant 1$. Possiamo fare di meglio osservando che per ogni coppia di eveni E, F abbiamo $E\cap F\subseteq E$ e allo stesso tempo $E\cap F\subseteq F$. Di conseguenza, per monotonia deve essere $P(E\cap F)\leqslant P(E)$ e $P(E\cap F)\leqslant P(F)$ ossia $P(E\cap F)\leqslant\min(P(E),P(F))$, che in questo caso ci dà una stima dall'alto pari a $p\leqslant\frac{17}{24}$.

PROPOSIZIONE 3.26. *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[6] 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. □

Esempio 3.27. 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%.

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.

Abbiamo una conseguenza immediata della Proposizione 3.26.

COROLLARIO 3.28. *Possiamo maggiorare la probabilità dell'unione di due eventi con la somma delle probabilità dei due eventi:*

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

Questa proprietà prende il nome di sub-additività.

Osservazione 3.29. La stima data dal Corollario 3.28 non sempre ci dà informazioni importanti. In particolare potrebbe essere maggiore di 1, come abbiamo visto nell'Esempio 3.25. In casi come questo la miglior stima sarebbe semplicemente 1.

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

Riguardiamo la Proposizione 3.26 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 3.26 possa essere estesa a un generico numero n di eventi.

PROPOSIZIONE 3.30. *Sia $(E_i)_{i=1}^n$ una famiglia finita di eventi. Allora*

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i=1}^{n} P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \cdots + (-1)^{n+1} P\left(\bigcap_{i=1}^{n} E_i\right). \tag{3.1}$$

Dimostrazione. La dimostrazione segue dalla Proposizione 3.26 e dal principio di inclusione-esclusione. □

Possiamo generalizzare a questo caso il Corollario 3.28.

COROLLARIO 3.31. *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 3.32. *La probabilità dell'unione di un numero finito di eventi può essere stimata dall'alto troncando il secondo membro della (3.1) 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[7].

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 3.31 al caso numerabile, detta anche disuguaglianza di Boole[8].

PROPOSIZIONE 3.33. *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:

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

7. Carlo Emilio Bonferroni (1892 – 1960).

8. 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 3.14 e per la disuguaglianza la Proposizione 3.23. □

3.4. PROBLEMI

Problema 9. 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?

Problema 10. 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?

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

Problema 12. 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?

Problema 13. 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 3.2. Quanti sono i percorsi di lunghezza minima che i ladri hanno a disposizione?

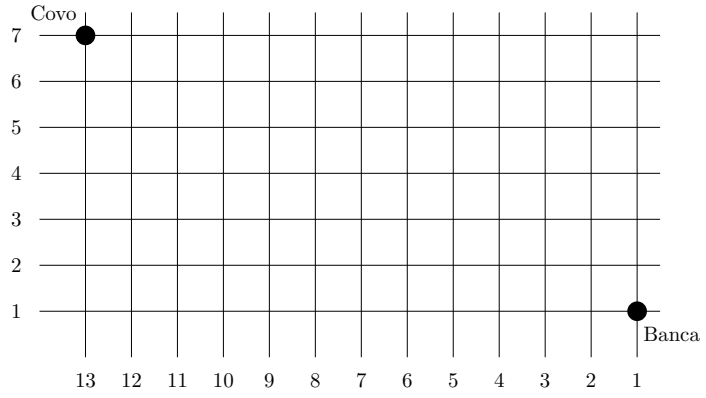


Figura 3.2. La fuga dei malfattori

Problema 14. Nelle stesse ipotesi del problema precedente, la polizia ha avuto una soffiata e ha piazzato due posti di blocco, come indicato in Figura 3.3: 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?

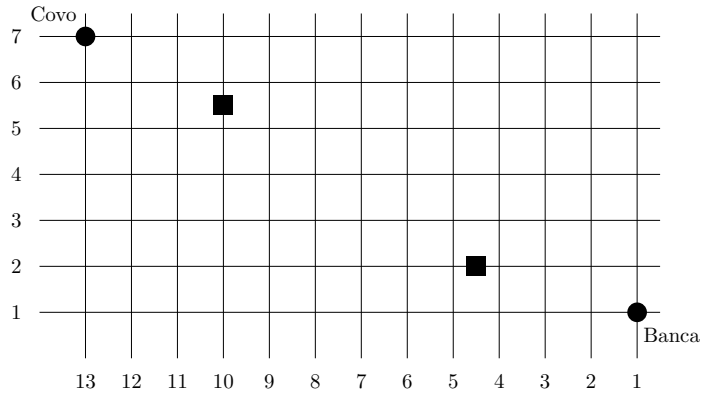


Figura 3.3. Malfattori in fuga con posti di blocco

Problema 15. Sia $\mathcal{C}$ la classe dei sottoinsiemi C di $\mathbb{N}$ per i quali esiste la "densità"

$$\lim_{m\to+\infty} m^{-1}\#\{k : 1 \leqslant k \leqslant m; k \in C\}.$$

Si potrebbe pensare a questa densità (quando è definita) come a una misura di probabilità che un numero scelto a caso appartenga a C. In realtà questo non funziona. Uno dei problemi è il seguente: esistono insiemi A e B in $\mathcal{C}$ la cui intersezione non è in $\mathcal{C}$. Mostrare un esempio di tali insiemi.

Problema 16. Mostrare, con un opportuno controesempio, che l'unione di tribù non è in generale una tribù. Lo stesso risultato vale per le algebre?

Problema 17. Dimostrare che se l'unione di due tribù è un'algebra, allora è una tribù".

Problema 18. Mostrare che l'unione di due tribù è una tribù se e solo se una delle due tribù è contenuta nell'altra[10].

10. Si parla in questo caso di sottotribù, dal momento che è contenuta, ma ha anche la struttura di tribù.

CAPITOLO 4

PROBABILITÀ CONDIZIONATA

Abbiamo citato, nell'introduzione, come una delle ragioni di essere della probabilità sia dare una misura di informazione o di incertezza. È pertanto naturale che vogliamo studiare come cambi questa misura al variare delle informazioni in nostro possesso. Sembra ragionevole che, in presenza di più informazioni, la nostra valutazione possa (e debba) cambiare, riflettendo così l'aggiornamento della nostra conoscenza. Questo è il linguaggio della scienza sperimentale: non c'è alcuna certezza (qualunque teoria può essere falsificata), ma una successione di risultati sperimentali a favore di una teoria aumenterà la nostra convinzione che tale teoria, che riesce a predire quei risultati, sia una buona teoria.

Supponiamo allora, 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.

Osservazione 4.1. In questo modo continuiamo a seguire la traccia della probabilità come una misura di informazione: se abbiamo nuovi dati possiamo e dobbiamo aggiornare la probabilità ossia la misura di incertezza che assegniamo. Come vedremo questo rende la probabilità condizionata il linguaggio del 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.

DEFINIZIONE 4.2. *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 4.1 possiamo vedere un'illustrazione in termini di insiemi della definizione.

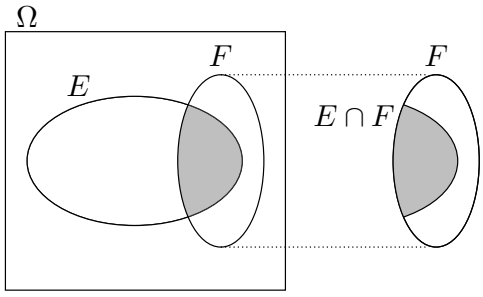


Figura 4.1. Nel condizionamento F è il ``nuovo'' universo

Esempio 4.3. 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 4.4. Vale la pena osservare che, nell'Esempio 4.3, non c'è alcun tipo di legame *temporale* o *causale* tra i due eventi che consideriamo: non è necessario che uno dei due causi l'altro o che uno accada prima dell'altro. Siamo interessati esclusivamente all'*informazione* che un evento porta riguardo a un altro. Nell'Esempio 4.3 il lancio del dado è già avvenuto, il risultato c'è già, ma Rosalia non ha alcuna informazione (e noi con lei) tranne quello che Stefano le dice.

Osservazione 4.5. Ci sono alcune cose cui prestare attenzione mentre prendiamo confidenza con la probabilità condizionale. Come prima cosa: $E|F$ **non** è un evento, non è definito come un insieme in $\mathcal{F}$ (in effetti $|$ non è un'operazione tra insiemi). È opportuno pensare a F come al parametro di una famiglia di funzioni. 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 4.6. 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:

$$P(E \cap F) = P(E|F) \cdot P(F) = P(F|E) \cdot P(E). \tag{4.1}$$

La doppia identità (4.1) prende anche il nome di *teorema (o regola) del prodotto*. Osserviamo che nella (4.1) 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). Pertanto, 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 4.7. *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 (4.1) vista sopra, oppure l'identità incontrata in precedenza:

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

Esempio 4.8. 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 3.26 e, per valutare $P(S\cap A)$, l'indipendenza.

Esempio 4.9. 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[1].

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 4.10. Prendiamo, in uno spazio di probabilità $(\Omega,\mathcal{F},P)$, due eventi E ed F tali che valgano $0<P(F)<1$ 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.

Esempio 4.11. Abbiamo due urne, una con 2 biglie bianche e 2 biglie nere, l'altra con 4 biglie nere. Non abbiamo modo di vedere cosa ci sia dentro le urne ed esternamente sono indistinguibili. Se estraiamo due biglie da un'urna scelta a caso tra le due, qual è la probabilità che siano entrambe nere?

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*.

Dal momento che stiamo estraendo due biglie assieme, ha senso considerare le coppie non ordinate di biglie come esiti, cioè $\Omega = \{\{B,B\},\{B,N\},\{N,N\}\}$. Tuttavia in questo modo assegnare le probabilità non è banale, perché le informazioni in nostro possesso sono codificate in modo diverso: dipende sia dall'urna sia dal colore delle biglie. Una buona scelta di modello può essere

$$\Omega = \{(U_1,\{B,B\}),(U_1,\{B,N\}),(U_1,\{N,N\}),(U_2,\{N,N\})\}.$$

Ci sono alcuni eventi di cui conosciamo la probabilità: E_1 "scegliamo la prima urna", ossia l'insieme costituito da tutti gli esiti della forma $(U_1,\cdot)$ e, in maniera simile, E_2 "scegliamo la seconda urna", l'insieme di tutti gli esiti della forma $(U_2,\cdot)$, ossia, il solo esito $(U_2,\{N,N\})$. Sappiamo dai dati del problema che $P(E_1) = P(E_2) = \frac{1}{2}$. Definiamo inoltre altri tre eventi: C_1 "entrambe le biglie sono nere"$=\{(U_1,\{N,N\}),(U_2,\{N,N\})\}$, C_2 "una biglia è nera e l'altra è bianca"$=\{(U_1,\{B,N\})\}$ e C_3 "entrambe le biglie sono bianche"$=\{(U_1,\{B,B\})\}$. Possiamo calcolare le probabilità condizionate:

$$P(C_1|E_1) = \frac{1}{6}; \qquad\qquad P(C_1|E_2) = 1.$$

La seconda è immediata (se siamo nella seconda urna siamo certi di estrarre due biglie nere). La prima segue, per esempio, da un ragionamento combinatorio: nella prima urna ci sono $\frac{4!}{2!\,2!} = 6$ combinazioni delle 4 biglie, tutte equiprobabili, tra cui l'unica che ci interessa è quella in cui le prime due biglie (ossia quelle estratte) sono nere.

Tuttavia la probabilità cercata è quella dell'evento C_1, e finora abbiamo calcolato quello che riuscivamo dai dati, ma non la probabilità di quello specifico evento. Possiamo però usare un'idea che abbiamo già incontrato nella dimostrazione della Proposizione 3.23: dal momento che E_1 e E_2 determinano una partizione di Ω,

$$\begin{aligned} P(C_1) &= P(C_1 \cap E_1) + P(C_1 \cap E_2) \\ &= P(C_1|E_1)\,P(E_1) + P(C_1|E_2)\,P(E_2) \\ &= \frac{1}{6}\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{7}{12}. \end{aligned}$$

Avendo introdotto la nozione di probabilità condizionale possiamo usarla per rendere rigorosa una rappresentazione diffusa delle probabilità, che può essere utile per descrivere alcuni modelli: la rappresentazione ad albero. Vediamo com'è nel caso dell'Esempio 4.11 in Figura 4.2.

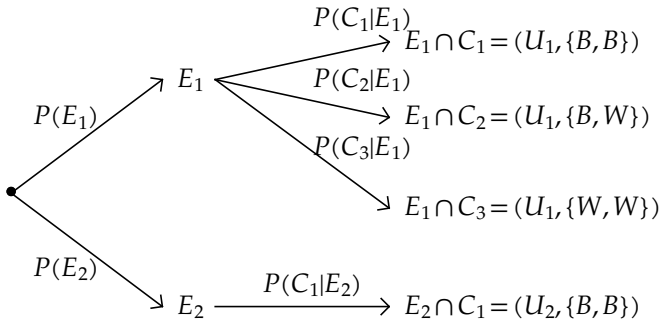


Figura 4.2. Rappresentazione ad albero

Le foglie e i nodi di questo albero sono eventi, mentre i rami dell'albero connettono un evento con una partizione dell'evento stesso, solitamente in forma di intersezione dell'evento con una partizione dell'intero insieme universo. In questo esempio, la partizione di primo livello è $\{E_1, E_2 = E_1^c\}$, mentre la partizione di secondo livello è $\{C_1, C_2, C_3\}$. I rami sono etichettati con la probabilità condizionata degli eventi che il ramo raggiunge dati gli eventi da cui il ramo parte. Per calcolare la probabilità di un nodo o di una foglia è sufficiente moltiplicare le probabilità associate ai rami che collegano la radice dell'albero al nodo di interesse. Ciò equivale a scrivere le probabilità delle intersezioni di eventi come il prodotto di oggetti incrementalmente condizionati. In Figura 4.2 la probabilità di $(U_1,\{N,N\}) = E_1 \cap C_3$ è $P(E_1)\cdot P(C_3|E_1)$.

Nell'Esempio 4.10 abbiamo incontrato un'idea interessante, poi ripresa anche nell'Esempio 4.11: 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 4.12. *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* $F \in \mathcal{F}$*,*

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

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 $F \cap E_i \subseteq E_i$. □

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 $P(F \mid E_i)$ 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 4.13. 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

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

Grazie alla formula di fattorizzazione possiamo riscrivere la probabilità cercata come

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

Esempio 4.14. 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 $E\subseteq F$, 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 $E\subseteq F^c$, 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 4.15. 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 $\mathcal{F}=\mathcal{D}(\Omega)$. 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 4.16. *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 4.17. 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 4.18. 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 4.19. 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

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

DEFINIZIONE 4.20. *In uno spazio di probabilità $(\Omega, \mathcal{F}, P)$ gli eventi $E_1, \ldots, E_n$ sono indipendenti rispetto a P se per qualunque scelta di indici (senza ripetizioni) $i_1, \ldots, i_m$ in $\{1, \ldots, n\}$ (con $m \leqslant n$) vale*

$$P\left(\bigcap_{j=1}^{m} E_{i_j}\right) = \prod_{j=1}^{m} P(E_{i_j}).$$

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

Esempio 4.21. 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 $a = P(D)$, $b = P(E)$, $c = P(F)$,

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

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

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

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

$$\begin{aligned} d\,e\,c + d'\,e' - d'\,e'\,c &= a\,b \\ &= (d\,c + d' - d'\,c)\,(e\,c + e' - e'\,c) \\ &= d\,e\,c^2 + d\,e'\,c - d\,e'\,c^2 + d'\,e\,c \\ &\quad + d'\,e' - d'\,e'\,c - d'\,e\,c^2 - d'\,e'\,c + d'\,e'\,c^2. \end{aligned}$$

Possiamo semplificare un po' di termini, arrivando a

$$d\,e\,c^2 + d\,e'\,c - d\,e'\,c^2 + d'\,e\,c - d'\,e\,c^2 - d'\,e'\,c + d'\,e'\,c^2 - d\,e\,c = 0$$

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.

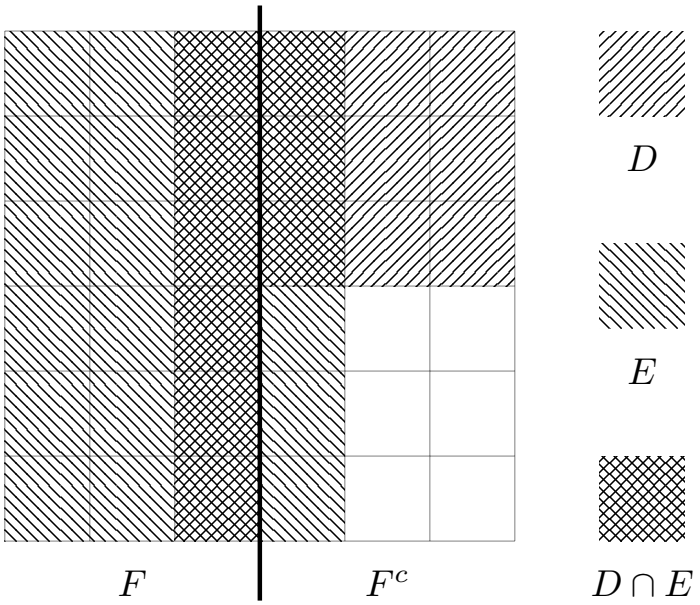


Figura 4.3. Un controesempio

Abbiamo allora tutti gli ingredienti per costruire un controesempio, rappresentato in Figura 4.3. 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.

Guardando però le probabilità degli eventi 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.

4.1. 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 4.22. Da una recente indagine[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

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

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

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 4.23. Tra i concorrenti delle Olimpiadi della Matematica[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:

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

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

$$P(B \cap ♀) = P(♀|B) \cdot P(B)$$

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

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

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

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

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

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

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

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

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 4.24. (BAYES[4]) *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:

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

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

COROLLARIO 4.25. *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*

$$P(E|F) = \frac{P(F|E) \cdot P(E)}{\sum_{i \in I} P(F|E_i) \cdot P(E_i)}. \tag{4.2}$$

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 4.26. 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:

$$\begin{aligned} P(M|+) &= \frac{P(+|M) \cdot P(M)}{P(+)} \\ &= \frac{P(+|M) \cdot P(M)}{P(+|M) \cdot P(M) + P(+|M^c) \cdot P(M^c)}, \end{aligned} \tag{4.3}$$

dove abbiamo usato anche l'identità (4.2) 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 (4.3), abbiamo allora

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

4. Thomas Bayes (1702 – 1761).

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[5] 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.

5. Un'esposizione divulgativa di alcuni risultati (dovuti a Kahneman e Tverski) in questo senso è nel volume *Thinking Fast and Slow* (*Pensieri lenti e pensieri veloci*) di D. Kahneman.

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 4.26, 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!

Ovviamente le cose sono un po' diverse nel mondo astratto: se potessimo fare infiniti esperimenti, potremmo raggiungere la (quasi) certezza, ossia probabilità 1 (o viceversa probabilità 0). Questo in effetti succede nel caso di alcuni teoremi, come vedremo più avanti. Tuttavia per dire che un evento ha probabilità 1 si usa comunemente la locuzione "quasi certo", sottolineando con il quasi che non è impossibile che non avvenga, anche se la probabilità che non avvenga è 0 (da vedersi in questo contesto come infinitesimamente piccola, più piccola di qualunque numero positivo.

4.1.1. 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 4.27. Torniamo al caso di Jacopo, incontrato all'Esempio 4.26. 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 4.26, solo che ora abbiamo due eventi rispetto ai quali condizioniamo. Allora

$$\begin{aligned} P(M|+_2\cap+_1) &= \frac{P(M\cap+_2\cap+_1)}{P(+_2\cap+_1)} = \frac{P(+_2|M\cap+_1)\,P(M\cap+_1)}{P(+_2\cap+_1)} \\ &= \frac{P(+_2|M\cap+_1)\,P(M|+_1)\,P(+_1)}{P(+_2\cap+_1)} = \frac{P(+_2|M\cap+_1)\,P(M|+_1)\,P(+_1)}{P(+_2|+_1)\,P(+_1)} \\ &= \frac{P(+_2|M\cap+_1)\,P(M|+_1)}{P(+_2|+_1)} = \frac{P(+_2|M\cap+_1)\,P(+_1|M)\,P(M)}{P(+_2|+_1)\,P(+_1)} \\ &= \frac{P(+_2|M\cap+_1)}{P(+_2|M\cap+_1)\,P(M|+_1)+P(+_2|M^c\cap+_1)\,P(M^c|+_1)}\cdot \\ &\quad \cdot\frac{P(+_1|M)\,P(M)}{P(+_1|M)\,P(M)+P(+_1|M^c)\,P(M^c)}, \end{aligned} \tag{4.4}$$

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[6].

Se ora torniamo alla (4.4) 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 4.26, 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 4.26) 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

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

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 4.27, possiamo fare alcune osservazioni generali.

Osservazione 4.28. 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.

6. 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 4.21.

Osservazione 4.29. 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 (4.4).

4.2. VARIAZIONI

Ora che abbiamo introdotto la nozione di indipendenza, possiamo studiare alcuni interessanti elementi di una qualunque tribù.

DEFINIZIONE 4.30. *Sia $(E_n)_{n\in\mathbb{N}}$ una successione di eventi in una tribù $\mathcal{F}$. Allora*

$$\limsup_n E_n := \bigcap_{k\in\mathbb{N}} \left(\bigcup_{n\geqslant k} E_n \right)$$
$$\liminf_n E_n := \bigcup_{k\in\mathbb{N}} \left(\bigcap_{n\geqslant k} E_n \right).$$

Entrambi questi oggetti sono eventi, ossia appartengono a $\mathcal{F}$, perché ogni unione (risp. intersezione) tra parentesi nella definizione è unione (risp. intersezione) numerabile di elementi della tribù ed è quindi in $\mathcal{F}$. Ma allora è in $\mathcal{F}$ anche la loro intersezione (risp. unione) numerabile.

Più interessante è il significato probabilistico. Se guardiamo $\bigcup_{n\geqslant k} E_n$, si tratta dell'evento che contiene gli esiti che stanno in almeno uno degli E_n per $n \geqslant k$. D'altra parte nel momento in cui prendiamo un'intersezione di eventi stiamo considerando quegli esiti che stanno in tutti gli eventi intersecati, quindi $\limsup_n E_n$ è l'evento che contiene quegli esiti ω per cui per ogni $k \geqslant 0$ esiste un $n \geqslant k$ tale che $\omega \in E_n$, ossia quegli esiti che accadono infinite volte negli E_n (che si verificano frequentemente).

Viceversa, l'evento $\liminf_n E_n$ raccoglie quegli eventi che si verificano definitivamente nella successione E_n, ossia che da un certo indice in poi si verifica sempre.

Avendo queste definizioni possiamo vedere due risultati in cui abbiamo eventi di probabilità 1.

TEOREMA 4.31. (PRIMO LEMMA DI BOREL[7]-CANTELLI[8]) *Sia $(E_n)_{n\in\mathbb{N}}$ una successione di eventi in una tribù $\mathcal{F}$. Se $\sum_{n\in\mathbb{N}} P(E_n) < +\infty$, allora con probabilità 1 solo un numero finito degli E_n si avvera.*

Dimostrazione. Basta dimostrare che

$$P\left(\limsup_n E_n\right) = 0$$

per la caratterizzazione del limsup data sopra. Chiamiamo $H_k = \bigcup_{n\geqslant k} E_n$, allora per la Proposizione 3.33

$$P(H_k) = P\left(\bigcup_{n\geqslant k} E_n\right) \leqslant \sum_{n\geqslant k} P(E_n).$$

Inoltre $\limsup_n E_n = \bigcap_{k\in\mathbb{N}} H_k$, ma anche $H_1 \supseteq H_2 \supseteq H_3 \supseteq \cdots$, quindi

$$P\left(\limsup_n E_n\right) = \lim_{k\to+\infty} P(H_k) \leqslant \lim_{k\to+\infty} \sum_{n=k}^{+\infty} P(E_n) = 0$$

7. Émile Borel (1871 – 1956).

8. Francesco Paolo Cantelli (1875 – 1966).

in cui l'ultima uguaglianza segue dal criterio di Cauchy di convergenza delle serie. □

TEOREMA 4.32. (SECONDO LEMMA DI BOREL-CANTELLI) *Sia $(E_n)_{n\in\mathbb{N}}$ una successione di eventi in una tribù $\mathcal{F}$. Se gli E_n sono tutti indipendenti e $\sum_{n\in\mathbb{N}} P(E_n) = +\infty$ allora $P(\limsup_n E_n) = 1$ (ossia sono frequentemente veri).*

Dimostrazione. Ricordiamo che $\limsup_n E_n = \bigcap_{k\in\mathbb{N}} H_k$, con $H_k = \bigcup_{n\geqslant k} E_n$. Allora, siccome gli eventi E_n^c sono tra loro indipendenti, in quanto complementari di eventi indipendenti,

$$P(H_k) = 1 - P(H_k^c) = 1 - P\left(\bigcap_{n\geqslant k} E_n^c\right) = 1 - \prod_{n\geqslant k} P(E_n^c) = 1 - \prod_{n\geqslant k} (1 - P(E_n)).$$

Osserviamo che se abbiamo una successione $(x_n)_{n\in\mathbb{N}}$ a valori in $[0,1)$, siccome per ogni $x \geqslant 0$ vale $1 - x \leqslant e^{-x}$, abbiamo $\prod_{n\geqslant k}(1-x_n) \leqslant \prod_{n\geqslant k} e^{-x_n}$, da cui $\prod_{n\geqslant k}(1-x_n) \leqslant e^{-\sum_{n\geqslant k} x_n}$.

Nel nostro caso $x_n = P(E_n)$, quindi $0 \leqslant \prod_{n\geqslant k}(1-P(E_n)) \leqslant e^{-\sum_{n\in\mathbb{N}} P(E_n)} \leqslant e^{-\infty} = 0$, quindi abbiamo

$$P(H_k) = 1 - \prod_{n\geqslant k} (1 - P(E_n)) = 1.$$

A questo punto la conclusione segue dalla catena di inclusioni $H_1 \supseteq H_2 \supseteq H_3 \supseteq \cdots$, da cui

$$1 = \lim_{n\to+\infty} P(H_n) = P\left(\bigcap_{n\in\mathbb{N}} H_n\right) = P\left(\limsup_n E_n\right)$$

che conclude la dimostrazione. □

4.3. PROBLEMI

Problema 19. Preso un mazzo ben mescolato di carte, numerate da 1 a 40, Francesco estrae una carta a caso, la guarda e la reinserisce nel mazzo, rimuovendo però tutte quelle con un numero strettamente maggiore di quello ottenuto. Mescolato ciò che resta del mazzo, lo passa a Silvia, che estrae a sua volta una carta a caso.

Se Silvia estrae la carta 32, qual è la probabilità che la carta estratta da Francesco fosse la numero 36?

Se Silvia estrae la carta 32, qual è la carta che più probabilmente ha estratto Francesco?

Problema 20. Un'urna contiene 3 biglie bianche e 2 biglie nere. Armida estrae 3 biglie dall'urna, le guarda e rimette nell'urna una delle biglie del colore più rappresentato nella sua estrazione. A questo punto Bartolomeo estrae una biglia, che si rivela essere nera. Quali sono le probabilità che:

1. Armida non abbia estratto alcuna biglia nera?
2. Armida abbia estratto esattamente una biglia nera?
3. Armida abbia estratto esattamente due biglie nere?

Problema 21. Sia $(p_n)_{n\in\mathbb{N}}$ una successione di numeri reali con $0 \leqslant p_n < 1$ per ogni n. Sia inoltre $S := \sum_{n\in\mathbb{N}} p_n \in [0, +\infty]$. Dimostrare che $\prod_{n\in\mathbb{N}}(1-p_n) = 0$ se e solo se $S = +\infty$. (Suggerimento: mostrare che se $S < 1$ allora $\prod_{n\in\mathbb{N}}(1-p_n) \geqslant 1 - S$.)

Problema 22. (LEMMI DI FATOU[9]) Sia $(E_n)_{n\in N}$ una successione di eventi nello spazio di probabilità $(\Omega, \mathcal{F}, P)$. Allora

$$\begin{aligned} P\left(\liminf_{n\in\mathbb{N}} E_n\right) &\leqslant \liminf_{n\in\mathbb{N}} P(E_n) \\ P\left(\limsup_{n\in\mathbb{N}} E_n\right) &\geqslant \limsup_{n\in\mathbb{N}} P(E_n). \end{aligned}$$

9. Pierre Joseph Louis Fatou (1878 – 1929).

Problema 23. (S. ROSS) Ho chiesto alla persona che abita accanto a me di innaffiare la mia amata pianta di Dracena, che è un po' delicata, mentre io sono in vacanza. Stimo che, senza acqua, la probabilità che la pianta muoia sia 0.8, mentre la probabilità che muoia anche se innaffiata sia 0.15. Stimo inoltre che la probabilità che la persona si ricordi di innaffiarla sia del 90%.

1. Qual è la probabilità che la pianta sia ancora viva al mio ritorno?
2. Se la pianta al mio ritorno fosse morta, con che probabilità stimo che la persona si sia dimenticata di innaffiarla?

Problema 24. Un algoritmo per individuare le email di spam classifica come "sospette" le email che contengono certe parole chiave. Per allenare l'algoritmo usiamo 100 email, 60 delle quali sono spam. Tra le email di spam, il 90% è classificato come sospetto, mentre solamente il 2.5% delle mail non di spam sono classificate come sospette.

1. Qual è la probabilità che una generica email sia classificata come sospetta?
2. Qual è la probabilità che una email classificata come sospetta sia effettivamente spam?
3. Qual è la probabilità che una email classificata come non sospetta sia effettivamente spam?

CAPITOLO 5

COSTRUIRE PROBABILITÀ

Finora abbiamo discusso delle proprietà di tutte le misure di probabilità e, nella maggior parte degli esempi, ci è stata fornita una misura di probabilità per eseguire i calcoli. Tuttavia, potremmo chiederci come dovremmo costruire uno spazio di probabilità e, in particolare, una misura, per un esperimento casuale che vogliamo modellare. Esiste una ricetta per costruire una funzione in modo che sia una misura di probabilità?

In effetti, definire una funzione di probabilità a partire dalla definizione sembra un po' intimidatorio: innanzitutto, dobbiamo assegnare un valore per la funzione calcolato in ogni evento e, come abbiamo visto, tali σ-algebre possono avere un numero enorme di elementi (e in generale dovremmo anche scegliere una particolare σ-algebra per il nostro modello). Inoltre, non possiamo assegnare tali valori liberamente: dobbiamo assicurarci che le scelte che facciamo siano compatibili con gli assiomi e le proprietà delle misure di probabilità.

D'altra parte, abbiamo bisogno di un modo per assegnare tali misure di probabilità se vogliamo creare il nostro modello matematico. In questo capitolo vedremo alcuni modi per farlo, a seconda dell'insieme campione Ω (in realtà principalmente in base alla sua cardinalità e alla sua rappresentazione come prodotto cartesiano). In molti casi, a meno di opportuni isomorfismi, il nostro lavoro sarà così compiuto e potremo dedicarci alla soluzione del problema. In altri avremo qualche idea in più per costruire quello che ci occorre. È però importante sottolineare che quelle che vedremo nelle prossime pagine non sono le uniche strategie possibili e non sono necessariamente le migliori in uno specifico contesto, dal momento che devono essere il più generali possibile. Sono in ogni caso un buon compromesso tra semplicità e specializzazione.

5.1. 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 come tribù $\mathcal{F}$ l'insieme delle parti di Ω, cioè $\mathcal{F} = \mathcal{P}(\Omega)$. 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[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 5.1. 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.

Esempio 5.2. 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 5.3. 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.

5.2. 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.

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 potenzialmente pericolosa, specie quando si muovono i primi passi nella probabilità, perché genera ambiguità.

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[2]. Consideriamo insiemi di numeri reali. Quelli che ci possono venire in mente di solito 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 $[0,1]$, interseche-remo 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.

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* (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). Mostriamo però più sotto, nell'Esempio 5.5, che non può coincidere con l'insieme delle parti (nel caso $\Omega = [0,1]$).

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[3]. 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 5.4. 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 $P([a,b]) = b - a$.

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.

3. 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.

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], ma non è l'unica.

La misura di probabilità uniforme su $[0,1]$ è ben definita sui Boreliani. Possiamo quindi usarla per mostrare che un insieme non appartiene ai Boreliani.

Esempio 5.5. (VITALI[5]) Vogliamo costruire un insieme di elementi di $\mathcal{P}([0,1])$ che non appartiene alla tribù dei Boreliani $\mathcal{B}([0,1])$.

Consideriamo le coppie di numeri reali nell'intervallo $[0,1]$ con la relazione R definita da xRy se e solo se $x - y \in \mathbb{Q}$. La relazione R è una relazione di equivalenza (cosa che si verifica facilmente) e, di conseguenza, determina delle classi di equivalenza, ad esempio:

- la classe dei numeri razionali, cioè dei numeri in relazione R con 0;
- la classe dei numeri in relazione R con π eccetera.

Osserviamo che le classi sono necessariamente in quantità più che numerabile, infatti hanno ciascuna una quantità numerabile di elementi, visto che $\mathbb{Q}$ è numerabile, ma la loro unione è più che numerabile. Siccome un'unione numerabile di insiemi numerabili è essa stessa numerabile, le classi devono essere in quantità più che numerabile. Resta il fatto che ogni classe ha una quantità numerabile di elementi.

Costruiamo ora l'insieme V prendendo un rappresentante per ogni classe di equivalenza (in altre parole $V = {}^{[0,1]}/_{R}$. Per $x \in \mathbb{Q}$ definiamo $x + V := \{x + v | v \in V\}$. L'unione $\bigcup_{x\in\mathbb{Q}} (x+V)$ è numerabile e disgiunta. Per definizione è anche un ricoprimento di $[0,1]$.

Supponiamo (per assurdo) che $V \in \mathcal{B}([0,1])$ e cerchiamo di calcolarne la probabilità uniforme. Osserviamo che questa probabilità P è invariante per traslazione (sui Boreliani), quindi $P(x+V) = k$ per qualche costante k indipendente da x. Allora

$$1 = P([0,1]) = P\left(\bigcup_{x\in\mathbb{Q}} (x+V)\right) = \sum_{x\in\mathbb{Q}} P(x+V) = \sum_{k\in\mathbb{Q}} k$$

e cadiamo in un assurdo. Infatti se $k = 0$ la somma è uguale a 0 e non a 1, ma se $k > 0$ la somma è uguale a $+\infty$ e non a 1.

Siccome gli insiemi $x + V$ non possono avere P costante, non sono Boreliani.

Se vogliamo che la probabilità di un intervallo dipenda solo dai suoi estremi, non dobbiamo necessariamente considerare la misura di Lebesgue, possiamo considerare una qualunque 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 5.4 è stata ottenuta scegliendo $F(x) = x$ (la funzione identità) nell'intervallo $[0,1]$. Chiaramente ci sono altre scelte possibili, come vedremo ora.

Esempio 5.6. Prendiamo la funzione $F\colon [0,1] \to \mathbb{R}$ definita da

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

4. Henri Léon Lebesgue (1875 – 1941).

5. Giuseppe Vitali (1875 – 1932).

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 5.4. 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 5.6, 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 3.23, 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 $\mathcal{B}$)[6]. La sua cardinalità è anche in questo caso quella del continuo.

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[7].

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

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

Osservazione 5.7. 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[8].

Esempio 5.8. Consideriamo la funzione $F\colon \mathbb{R} \to \mathbb{R}$ definita da

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

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:

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

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

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[9].

5.2.1. 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 $\mathcal{A}$.

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

9. Incontreremo di nuovo le variabili aleatorie nel Capitolo 6.

TEOREMA 5.9. (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 $\mathcal{A}$.*

Dimostrazione. Lasciata a un successivo corso di teoria della misura. □

Osservazione 5.10. In realtà possiamo chiedere che $\mathcal{A}$ sia una famiglia di sottoinsiemi di Ω chiusa rispetto all'intersezione e il teorema vale allo stesso modo (grazie al teorema della classe monotona).

5.3. 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 4), possiamo descriverli tutti assieme come spazio prodotto, portandoci dietro quello che sappiamo sulle varie componenti. Per farci un'idea, vediamo qualche esempio.

Esempio 5.11. 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 5.12. 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 5.2. 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 5.11, è ragionevole prendere il prodotto cartesiano $\Omega = \Omega_1 \times \Omega_2$.

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

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

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

$$P(E_1 \times E_2) = P_1(E_1) \cdot P_2(E_2). \tag{5.1}$$

Questo giustifica anche la notazione $P = P_1 \otimes P_2$. Si può obiettare che la (5.1) 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[10].

Esempio 5.13. Tornando all'Esempio 5.11, 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 5.12: 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 5.14. 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?

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

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

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,

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

10. 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.

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 5.2 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 5.15. 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''[11]. 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. I cilindri stanno nella tribù, dal momento che ne sono i generatori. La probabilità di questo cilindro è

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

5.4. VARIAZIONI

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 5.16. Supponiamo di avere un dado a 6 facce, di cui sappiamo che $P(1) = \frac{1}{6}$. Se volessimo procedere come visto nella Sezione 5.1 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[12] 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.

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 5.17. Due persone decidono di incontrarsi per cema. Stanno cercando di capire se il Fato abbia deciso che debbano essere una coppia, quindi per rendere più sfidante (per il Fato) l'appuntamento, si mettono d'accordo nel modo seguente: ciascuno di loro si impegna ad arrivare in Piazza Fiera in un qualunque momento tra le 19 e le 20 e a restare in attesa per 5 minuti. Passati questi 5 minuti (o allo scoccare delle 20) se ne andrà. Con che probabilità le due persone si incontreranno?

11. 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 9.

12. György Pólya (1887 – 1985).

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 una delle due persone. 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 la seconda persona arrivi tra le 19.21 e le 19.41 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 le due persone separatamente e stiamo trascurando il fatto che sono disposte ad aspettare. Se sapessimo che la prima persona arriva alle 19.13, allora la probabilità che si incontrino sarebbe uguale alla probabilità che la seconda arrivi nei 5 minuti precedenti alle 19.13 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 della prima persona e su quello delle ordinate quello della seconda. Le coordinate interne al quadrato rappresentano le combinazioni di arrivi delle due persone. 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[13]. Possiamo spostarci orizzontalmente o verticalmente di 5 minuti, rispetto alla diagonale, cioè considerare la diagonale ingrassata, colorata in grigio chiaro nella Figura 5.1. Questa superficie rappresenta tutte le coppie di orario d'arrivo per cui le due persone 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 delle due persone. Questo rapporto è $\frac{11}{36}$.

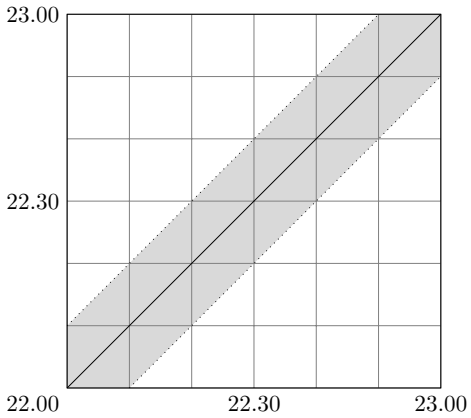


Figura 5.1. Incontro in Piazza Fiera

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 ciascuna delle due persone 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

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

13. 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.

Quando poi passiamo a considerare insieme le due persone, 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 5.18. 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.

5.5. PROBLEMI

Problema 25. Prendendo a caso due punti su un segmento, lo si divide in tre parti. Con che probabilità questi tre segmenti possono formare un triangolo?

Problema 26. Mostrare che l'insieme dei numeri razionali nell'intervallo $[0,1]$ appartiene alla tribù dei Boreliani (e quindi è misurabile) e ha misura (di Lebesgue) nulla.

Problema 27. Trovare un sottoinsieme dell'intervallo $[0,1]$ di cardinalità più che numerabile, ma di misura (di Lebesgue) nulla.

Problema 28. (G. LETTA) Si consideri una successione di lanci di una moneta equilibrata. Qual è la probabilità di ottenere testa per la prima volta in un lancio dispari? Fissato un numero reale $0 \leqslant p \leqslant 1$, è possibile determinare un insieme A di numeri interi strettamente positivi, in modo che la probabilità di ottenere testa per la prima volta in un lancio con indice appartenente ad A sia esattamente p?

Problema 29. (G. LETTA) Due giocatori lanciano una moneta equilibrata: il primo n volte e il secondo $n+1$ volte. Vince chi ottiene più teste, e in caso di parità vince chi ha fatto meno lanci. Si tratta di un gioco equilibrato[17]?

17. Ossia: i due giocatori hanno la medesima probabilità di vincere il gioco?

CAPITOLO 6

VARIABILI ALEATORIE

Nei precedenti Capitoli 3 e 5 abbiamo visto quali sono le componenti della descrizione matematica di un esperimento aleatorio, ossia le componenti di uno spazio di probabilità, ma anche come costruirli. Abbiamo inoltre visto che la rappresentazione non è unica. Da un lato questa non unicità ci può dare un senso di ricchezza, dall'altro, però, vorremmo una opportuna nozione di equivalenza a garantirci che rappresentazioni diverse diano luogo alla medesima probabilità (intesa come numero). Uno strumento fondamentale a questo scopo sono le variabili aleatorie, argomento di questo capitolo.

6.1. VARIABILI ALEATORIE

Cominciamo da qualcosa di già visto.

Esempio 6.1. 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 6.1. 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 6.2. Un'azienda produce calcolatori. Il costo di produzione di un singolo calcolatore è pari a 1000 €, mentre il prezzo di vendita è 1999 €. 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 $1999-1000\,n>0$, ossia l'azienda guadagna solo se $n=1$, per $n\geqslant 2$ va in perdita. Quindi la probabilità di guadagno è 0.9, quella di andare in rosso è $0.0\bar{9}=0.1$.

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 $3\cdot 1900-n\cdot 1000>0$, ossia $n<6$. 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:

$$\Omega_G^1=\{999,-1,-1001,-2001,\dots\},\qquad \Omega_G^3=\{2997,1998,999,-1,-1001,\dots\}$$

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:

$$P_G^1(999) = 0.9, P_G^1(-1) = 0.09, \dots$$

e (nel caso dei 3 calcolatori)

$$P_G^3(2997) = 0.729, P_G^3(1998) = 0.2187, P_G^3(999) = 0.04374, \dots$$

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

DEFINIZIONE 6.3. *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 6.4. La funzione S dell'Esempio 6.1 che calcola la somma dei risultati dei due dadi (cioè delle due componenti di ogni elemento ω) è una variabile aleatoria.

Osservazione 6.5. 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 6.6. 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 6.7. 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 6.8. *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 6.9. *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 6.10. 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 6.11. Nel momento in cui abbiamo una probabilità sui risultati dei due dadi nell'Esempio 6.1 (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 6.12. Non sempre è facile capire se una funzione sia o meno una variabile aleatoria: la branca della matematica che si occupa (anche) di questo è la *Teoria della misura*. Tuttavia ci viene in aiuto, nel caso $\Omega = \mathbb{R}$ e $\mathcal{F} = \mathcal{B}$, il seguente teorema.

TEOREMA 6.13. *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.*

Dimostrazione. Il caso X continua è immediato dalle definizioni di $\mathcal{B}$ e di continuità. Per gli altri due casi è necessaria un po' più di accortezza[1]. Supponiamo che X sia non decrescente. Consideriamo per ogni $a \in \mathbb{R}$ la semiretta $[a, +\infty)$ (che sta nei Boreliani) e la sua controimmagine mediante X, cioè $X^{-1}([a, +\infty)) = \{y \in \mathbb{R} : X(y) \geqslant a\} = I$. Se b appartiene a questa controimmagine, allora ogni $c \geqslant b$ appartiene a I, poiche $X(c) \geqslant X(b) \geqslant a$. Allora I può essere solamente uno dei seguenti insiemi: $\emptyset$, $(\inf I, +\infty)$, $[\inf I, +\infty)$ o $\mathbb{R}$, tutti appartenenti ai Boreliani. □

1. Una forte tentazione potrebbe essere quella di osservare (correttamente) che le discontinuità di una funzione monotona sono un insieme di cardinalità al più numerabile, per poi farne seguire che esiste una funzione continua che coincide con quella di partenza quasi certamente (falso) e usare il risultato sulle funzioni continue per concludere.

Osservazione 6.14. 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 nell'Osservazione 5.7).

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

Esempio 6.15. 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), come nel Problema 28.

Lo spazio di probabilità che descrive questo esperimento è $(\Omega,\mathcal{F},P)$, con $\Omega=\{T,C\}^{\mathbb{N}^+}$ 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

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

Quindi

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

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 $x=\frac{2}{3}$.

Abbiamo introdotto le *variabili aleatorie* come 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 6.16. (**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 6.17. (**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 6.18. (**Variabili aleatorie semplici**) Una volta che abbiamo le variabili aleatorie indicatrici, ne possiamo considerare delle combinazioni lineari[2]. Per esempio, dati $E, F \in \mathcal{F}$, possiamo prendere $X = I_E - 3 I_F$.

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

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

2. Questo dovrebbe richiamare memorie del corso di Analisi.

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? Queste domande sono il Problema 30.

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 $(\mathbb{R},\mathcal{B})$.

DEFINIZIONE 6.19. *Dati uno spazio di probabilità $(\Omega,\mathcal{F},P)$ e una variabile aleatoria $X:(\Omega,\mathcal{F})\to(\mathbb{R},\mathcal{B})$, 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 6.20. Tornando alla situazione descritta nell'Esempio 6.1, cioè la somma delle facce di due dadi bilanciati indipendenti, possiamo ora scrivere $P_S(\{7\})$ per indicare $P(S\in\{7\})$.

DEFINIZIONE 6.21. *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:*

$$\begin{aligned}X&:(\Omega,\mathcal{F},P)\to(\mathbb{R},\mathcal{B})\\ Y&:(\tilde\Omega,\tilde{\mathcal{F}},\tilde P)\to(\mathbb{R},\mathcal{B}).\end{aligned}$$

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 6.22. 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 6.23. 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 6.24. 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 $(\mathbb{R}, \mathcal{B})$.

6.2. FUNZIONI DI RIPARTIZIONE

Abbiamo concluso che, grazie alle variabili aleatorie, 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 6.25. *Data una variabile aleatoria X sullo spazio di probabilità $(\Omega, \mathcal{F}, P)$, la* funzione di ripartizione *o* funzione cumulativa[3] *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 6.26. Consideriamo una variabile aleatoria degenere $X \equiv c$. La sua funzione di ripartizione è

$$F_X(y) = P(X \leqslant y) = \begin{cases} 1 & \text{se } y \geqslant c \\ 0 & \text{se } y < c \end{cases}$$

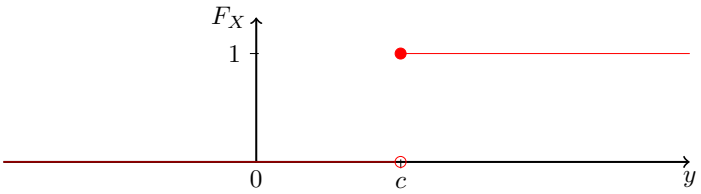


Figura 6.1. Funzione di ripartizione della v.a. degenere $X \equiv c$

Esempio 6.27. Consideriamo ora la variabile aleatoria indicatrice dell'evento $E \in \mathcal{F}$:

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

La funzione di ripartizione di questa variabile aleatoria è

$$F_{I_E}(y) = P(I_E \leqslant y) = \begin{cases} 0 & \text{se } y < 0 \\ P(E^c) & \text{se } 0 \leqslant y < 1 \\ 1 & \text{se } y \geqslant 1 \end{cases}$$

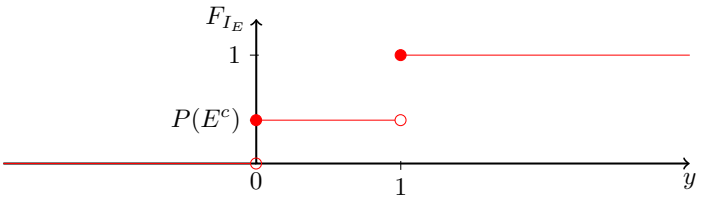


Figura 6.2. Funzione di ripartizione della v.a. indicatrice I_E

Esempio 6.28. 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

$$F_{D_4}(y) = P(D_4 \leqslant y) = \begin{cases} 0 & \text{se } y < 1 \\ 1/4 & \text{se } 1 \leqslant y < 2 \\ 2/4 & \text{se } 2 \leqslant y < 3 \\ 3/4 & \text{se } 3 \leqslant y < 4 \\ 1 & \text{se } y \geqslant 4 \end{cases}$$

3. Ci sono anche altre varianti, che nascono dal termine inglese, *cumulative distribution function*, da cui viene anche l'abbreviazione *cdf*.

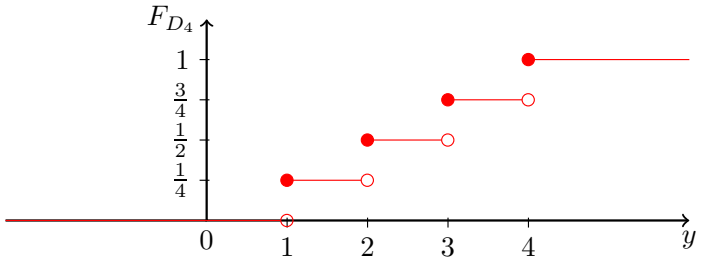


Figura 6.3. Funzione di ripartizione della v.a. di un dado a 4 facce D_4

Esempio 6.29. 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 (o 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* $[0,1]$.

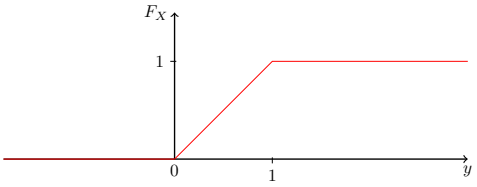


Figura 6.4. Funzione di ripartizione della v.a. uniforme su $[0,1]$

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 6.30. 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 6.31. *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,

$$\begin{aligned} \lim_{x \to -\infty} F_X(x) &= \lim_{x \to -\infty} P(X \leqslant x) \\ &= \lim_{x \to -\infty} P(\{\omega \in \Omega : X(\omega) \leqslant x\}) \\ &= P\left(\bigcap_{n \in \mathbb{N}} \{\omega \in \Omega : X(\omega) \leqslant -n\} \right) \\ &= P(\emptyset) = 0 \end{aligned}$$

e, in modo del tutto analogo,

$$\begin{aligned} \lim_{x \to +\infty} F_X(x) &= \lim_{x \to +\infty} P(X \leqslant x) \\ &= \lim_{x \to +\infty} P(\{\omega \in \Omega : X(\omega) \leqslant x\}) \\ &= P\left(\bigcup_{n \in \mathbb{N}} \{\omega \in \Omega : X(\omega) \leqslant n\} \right) \\ &= P(X^{-1}(\mathbb{R})) = P(\Omega) = 1. \end{aligned}$$

In un generico punto interno $x_0 \in \mathbb{R}$ abbiamo, per il limite da destra,

$$\begin{aligned} \lim_{x \to x_0^+} F_X(x) &= \lim_{x \to x_0^+} P(X \leqslant x) \\ &= P\left(\bigcap_{n \in \mathbb{N}^+} \left\{ \omega \in \Omega : X(\omega) \leqslant x_0 + \frac{1}{n} \right\} \right) \\ &= P(X \leqslant x_0) = F_X(x_0) \end{aligned}$$

e per quello da sinistra

$$\begin{aligned} \lim_{x \to x_0^-} F_X(x) &= \lim_{x \to x_0^-} P(X \leqslant x) \\ &= P\left(\bigcup_{n \in \mathbb{N}^+} \left\{ \omega \in \Omega : X(\omega) \leqslant x_0 - \frac{1}{n} \right\} \right) \\ &= P(X < x_0) \\ &= P(X \leqslant x_0) - P(X = x_0) \\ &= F_X(x_0) - P(X = x_0) \end{aligned}$$

in cui abbiamo usato la seguente osservazione: se $\omega \in \bigcup_{n \in \mathbb{N}^+} \left\{ X(\omega) \leqslant x_0 - \frac{1}{n} \right\}$, allora $X(\omega) \leqslant x$ per qualche $x < x_0$ e dunque $X(\omega) < x_0$. □

Osservazione 6.32. 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 6.33. Come già osservato in precedenza, a partire da una funzione di ripartizione F_X possiamo calcolare la probabilità in un qualunque Boreliano. Ad esempio:

- $P(X \in (a,b)) = F_X(b) - F_X(a) - P(X = b) = \lim_{x \to b^-} F_X(x) - F_X(a)$
- $P(X \in [a,b]) = F_X(b) - F_X(a) + P(X = a) = F_X(b) - \lim_{x \to a^-} F_X(x)$

– $P(X<a) = \lim_{x\to a^-} F_X(x)$
– $P(X>b) = 1 - F_X(b)$

e così via.

6.3. 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 6.34. *Una variabile aleatoria che può assumere al più un numero finito o numerabile di valori si dice* variabile aleatoria discreta.

Osservazione 6.35. 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[4]. I punti di discontinuità sono i valori che la variabile aleatoria può assumere.

DEFINIZIONE 6.36. *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 6.37. Il fatto che X sia chiamata *variabile aleatoria assolutamente continua* segue dal fatto che in quel caso F_X è una funzione assolutamente continua, condizione più stretta rispetto al solo essere continua.

Inoltre, potrebbe non essere chiarissimo il *motivo* di questa richiesta addizionale per la funzione di ripartizione, come mai non ci limitiamo a considerare le variabili aleatorie continue nella loro generalità. Per questo basta avere un po' di pazienza: lo capiremo tra qualche pagina.

DEFINIZIONE 6.38. *Una variabile aleatoria che non sia né discreta né (assolutamente) continua*[5] *di dice* variabile aleatoria mista.

Osservazione 6.39. 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.

Esempio 6.40. 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 derivata.

4. Non è possibile che ci siano più di una quantità numerabile di discontinuità. Chiamiamo D l'insieme dei punti di discontinuità. Per ogni $x \in D$ abbiamo $F(x^-) < F(x^+)$, quindi possiamo scegliere un numero razionale q_x tale che $F(x^-) < q_x < F(x^+)$. Dal momento che F è una funzione di ripartizione è non decrescente, quindi per $x \neq y$ in D abbiamo $q_x \neq q_y$. La funzione $x \mapsto q_x$ da D in $\mathbb{Q}$ è iniettiva e, dal momento che $\mathbb{Q}$ è numerabile, D è al più numerabile.

5. Alcuni libri non includono le variabili aleatorie continue ma non assolutamente continue tra le variabili aleatorie miste. Questa è una scelta del tutto legittima, solitamente dettata dalla comodità per quello che ci si prefigge di fare. Noi ci concentreremo sulle assolutamente continue e non sulle continue, quindi releghiamo queste ultime tra le miste.

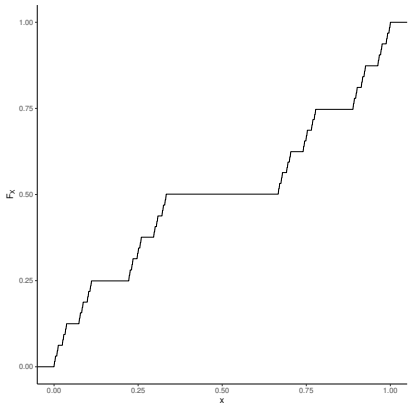


Figura 6.5. La funzione di ripartizione della variabile aleatoria di Cantor

Esempio 6.41. 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.

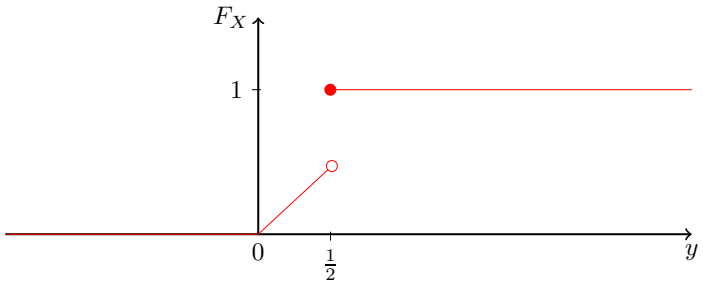


Figura 6.6. Funzione di ripartizione di una variabile aleatoria mista

6.3.1. Variabili aleatorie discrete

Ci concentriamo ora sulle variabili aleatorie discrete.

DEFINIZIONE 6.42. *Sia X una variabile aleatoria discreta. Chiamiamo* densità discreta *o* funzione di massa di probabilità (*a volte abbreviata con* pmf) *la funzione $p_X : \mathbb{R} \to [0,1]$, $p_X(x) := P(X = x)$. L'insieme $\mathcal{R}_X = \{x_i\}_{i \in I}$, al più numerabile, dei possibili valori assunti da X (e quindi punti in cui p_X è non nulla) prende il nome di* supporto di X *o di p_X.*

TEOREMA 6.43. (PROPRIETÀ DELLA PMF) *Sia X una variabile aleatoria discreta e sia p_X la sua densità discreta. Allora:*

i. *per ogni $x \in \mathbb{R}$, $p_X(x) \geqslant 0$*

ii. *per ogni $x \in \mathbb{R}$, $p_X(x) \leqslant 1$*

iii. *per ogni $x \in \mathcal{R}_X^c$, $p_X(x) = 0$*

iv. $\sum_{x \in \mathcal{R}_X} p_X(x) = 1$

v. se $E \in \mathcal{B}$, *allora* $P_X(E) = \sum_{x \in \mathcal{R}_X \cap E} p_X(x) = \sum_{x \in \mathcal{R}_X} \mathbb{1}_E(x)\, p_X(x)$.

Dimostrazione. Come prima cosa osserviamo che le proprietà i, ii e iii sono immediate dalla Definizione 6.42. Inoltre, le somme in iv e v sono ben definite, perché $\mathcal{R}_X$ è al più numerabile. Mostriamo la v e la iv seguirà come caso particolare. Abbiamo

$$\begin{aligned} P_X(E) &= P(X \in E) = P(\{\omega \in \Omega : X(\omega) \in E\} \cap \Omega) \\ &= P\left(\{\omega \in \Omega : X(\omega) \in E\} \cap \bigcup_{x \in \mathcal{R}_X} \{\omega \in \Omega : X(\omega) = x\}\right) \\ &= P\left(\bigcup_{x \in \mathcal{R}_X} (\{\omega \in \Omega : X(\omega) \in E\} \cap \{\omega \in \Omega : X(\omega) = x\})\right) \\ &= \sum_{x \in \mathcal{R}_X} P((X \in E) \cap (X = x)) \\ &= \sum_{x \in \mathcal{R}_X} \mathbb{1}_E(x)\, p_X(x). \end{aligned}$$

Come detto, la iv si ottiene nel caso particolare $E = \mathbb{R}$. □

Osservazione 6.44. Se X è una variabile aleatoria discreta di densità discreta p_X, allora la sua funzione di ripartizione F_X è tale che

$$F_X(y) = \sum_{x \in \mathcal{R}_X} \mathbb{1}_{(-\infty, y]}(x)\, p_X(x).$$

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.

6.3.2. 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 p_X sarebbe identicamente nulla.

DEFINIZIONE 6.45. *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 6.46. (PROPRIETÀ DELLA PDF) *Sia X una variabile aleatoria assolutamente continua e sia f_X la sua densità. Allora:*

i. *per ogni* $x \in \mathbb{R}$, $f_X(x) \geqslant 0$
ii. *per ogni* $x \in \mathcal{R}_X^c$, $f_X(x) = 0$
iii. $\int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = 1$
iv. $\int_a^b f_X(x)\, \mathrm{d}x = F_X(b) - F_X(a)$.

Dimostrazione. Le prime due proprietà seguono direttamente dalla definizione. Inoltre

$$\int_{-\infty}^{+\infty} f_X(x)\, \mathrm{d}x = \lim_{x \to +\infty} F_X(x) = 1$$

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 6.47. Le proprietà i, ii e iii nel Teorema 6.46 sono la controparte delle proprietà i, iii e iv del Teorema 6.43. Con una breve riflessione possiamo anche notare che la proprietà iv nel Teorema 6.46 corrisponde alla proprietà v nel Teorema 6.43, dal momento che tutti gli eventi nella tribù dei Boreliani $\mathcal{B}$ sono generati da intervalli.

Tuttavia non esiste per le densità di probabilità una proprietà corrispondente alla ii del Teorema 6.43. 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 6.48. 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 2ε 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.

Questo ci dice anche che la Definizione 6.45 non è precisissima nel dire "la" funzione. Ancora una volta, lasciamo questi dettagli a un futuro corso.

Esempio 6.49. (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 $\mathbb{R} \setminus \{0,1\}$:

$$f_X(x) = \begin{cases} 0 & \text{se } x < 0 \\ 1 & \text{se } 0 < x < 1 \\ 0 & \text{se } x > 1 \end{cases}$$

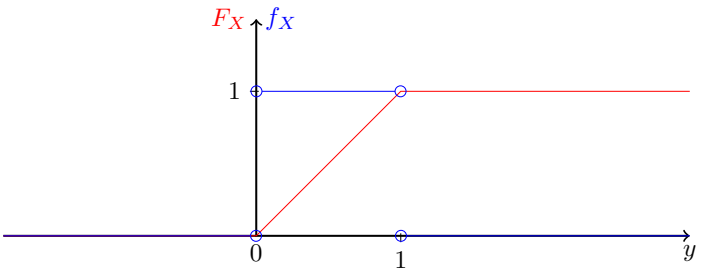


Figura 6.7. Funzione di ripartizione e densità della v.a. uniforme su $[0,1]$

Osservazione 6.50. Come la funzione di massa di probabilità p_X non ha senso per una variabile aleatoria continua, così la densità f_X non ha senso per le variabili aleatorie discrete.

6.4. PROBLEMI

Problema 30. In uno spazio di probabilità $(\Omega, \mathcal{F}, P)$, dati $E, F \in \mathcal{F}$, sia $X = I_E - 3\,I_F$. Per $A \in \mathcal{B}$, quanti sono i valori che può assumere la probabilità $P(X \in A)$? Quali sono? Quanti ne dobbiamo calcolare?

Problema 31. Consideriamo lo spazio di Borel $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ e la funzione reale

$$F(x) = \begin{cases} 1 - \exp(-2x) & \text{se } x \geqslant 0 \\ 0 & \text{altrimenti.} \end{cases}$$

1. Determinare se F è una funzione di ripartizione. Se non la è trovare una funzione simile che lo sia.
2. Qual è la probabilità che una variabile aleatoria di legge F cada nell'intervallo $(-0.65, 0.65]$?
3. Quali sono le probabilità che cada nell'intervallo $[0.65, 0.65]$ e in $[0.005, 0.65]$?
4. Qual è la probabilità che sia in $(0.005, 0.10] \cup (0.345, 2.717] \cup (1.98, 3]$?

Problema 32. (S. ROSS) La funzione di ripartizione di X è definita come segue:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{2} & 0 \leqslant x < 1 \\ \frac{2}{3} & 1 \leqslant x < 2 \\ \frac{11}{12} & 2 \leqslant x < 3 \\ 1 & x \geqslant 3 \end{cases}$$

1. Tracciare il grafico di F.
2. Calcolare $P\left(X > \frac{1}{2}\right)$.
3. Calcolare $P(2 < X \leqslant 4)$.
4. Calcolare $P(X < 3)$.
5. Calcolare $P(X = 1)$.

Problema 33. Sia $\mathcal{M}(\Omega, \mathcal{F})$ l'insieme delle variabili aleatorie[6] dallo spazio probabilizzabile $(\Omega, \mathcal{F})$ a valori reali. Dimostrare che $\mathcal{M}(\Omega, \mathcal{F})$ è chiuso rispetto alla somma, al prodotto per scalari, al prodotto, ai limiti su successioni e alla composizione con funzioni continue, ossia:

1. se $X, Y \in \mathcal{M}(\Omega, \mathcal{F})$ e $\lambda \in \mathbb{R}$, allora $X + Y \in \mathcal{M}(\Omega, \mathcal{F})$, $\lambda X \in \mathcal{M}(\Omega, \mathcal{F})$ e $X \cdot Y \in \mathcal{M}(\Omega, \mathcal{F})$;
2. se $(X_n)_{n \in \mathbb{N}} \subset \mathcal{M}(\Omega, \mathcal{F})$ allora $\inf_{n \in \mathbb{N}} X_n \in \mathcal{M}(\Omega, \mathcal{F})$, $\sup_{n \in \mathbb{N}} X_n \in \mathcal{M}(\Omega, \mathcal{F})$, $\liminf_{n \in \mathbb{N}} X_n \in \mathcal{M}(\Omega, \mathcal{F})$ e $\limsup_{n \in \mathbb{N}} X_n \in \mathcal{M}(\Omega, \mathcal{F})$
3. se $X \in \mathcal{M}(\Omega, \mathcal{F})$ e $f \colon \mathbb{R} \to \mathbb{R}$ è continua, allora $f \circ X \in \mathcal{M}(\Omega, \mathcal{F})$.

6. Il fatto che sia denotato con $\mathcal{M}$ viene dal fatto che le variabili aleatorie altro non sono che funzioni *misurabili*, ma siccome questo è un corso di Probabilità e non di Teoria della Misura, le chiamiamo variabili aleatorie.

CAPITOLO 7

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.

7.1. 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[1]? Vediamolo in un esempio.

Esempio 7.1. 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 p_{D_4} è

$$p_{D_4}(x) = \begin{cases} \frac{1}{4} & \text{se } x \in \{1,2,3,4\} \\ 0 & \text{altrimenti.} \end{cases}$$

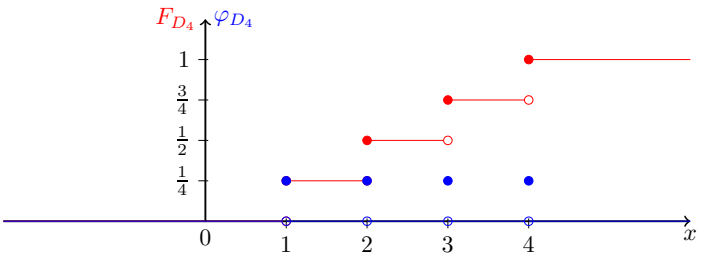


Figura 7.1. Cdf e pmf della v.a. di un dado a 4 facce D_4

Sia ora $Y = 2D_4 + 3$, quali sono i valori che può assumere[2]? Li possiamo riassumere in tabella:

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.

1. Il fatto che sia una variabile aleatoria segue dal Problema 33.

2. Attenzione in questo caso $2D_4$ non è da intendersi uguale al lancio di due dadi a 4 facce: quello sarebbe $D_4 + \widetilde{D_4}$ dal momento che avremmo due esperimenti aleatori sottostanti, per quanto identicamente distribuiti. Con $2D_4$ indichiamo la situazione in cui abbiamo un solo dado il cui risultato viene raddoppiato.

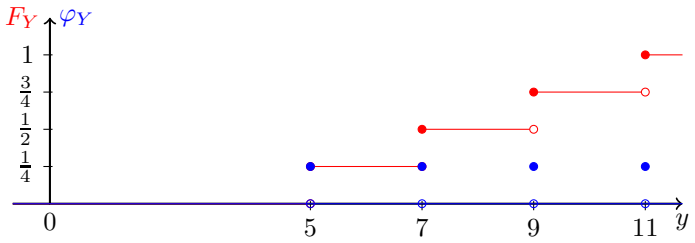


Figura 7.2. Cdf e pmf della v.a. $Y = 2D_4 + 3$

Possiamo infatti osservare che per la funzione di densità discreta

$$p_Y(y) = P(Y = y) = P(2D_4 + 3 = y) = P\left(D_4 = \frac{y-3}{2}\right) = p_{D_4}\left(\frac{y-3}{2}\right)$$

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 7.2. Sia X la variabile aleatoria uniforme sull'intervallo $[0,1]$. Ne conosciamo già sia la funzione di ripartizione F_X (Esempio 6.29), sia la densità f_X (Esempio 6.49).

Sia $Y = 2X + 3$. Non possiamo andarci a calcolare come prima i valori possibili, prendendo gli elementi 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'immagine dell'insieme dei numeri reali x tali che $f_{2X+3}(2x+3) \neq 0$.

Possiamo allora convincerci (vedremo i dettagli per assicurarcene in seguito) che il supporto di Y sia il supporto di X dilatato e traslato: $\mathcal{R}_Y = 2\,\mathcal{R}_X + 3$, in cui 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:

$$f_Y(x) = \begin{cases} c & x \in (3,5) \\ 0 & \text{altrimenti} \end{cases}.$$

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 allora

$$F_Y(y) = P(Y \leqslant y) = P(2X + 3 \leqslant y) = P\left(X \leqslant \frac{y-3}{2}\right) = F_X\left(\frac{y-3}{2}\right)$$

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,

$$f_Y(y) = \frac{1}{2} f_X\left(\frac{y-3}{2}\right),$$

in cui possiamo notare che, rispetto al caso discreto, compare un coefficiente.

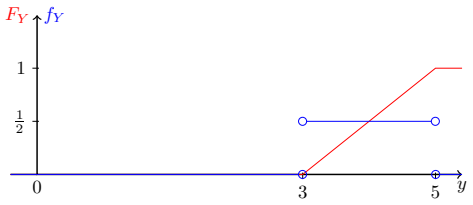


Figura 7.3. Cdf e pdf di una trasformazione lineare della v.a. uniforme su $[0,1]$

In generale vale il seguente risultato.

PROPOSIZIONE 7.3. *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 $a<0$*

$$F_Y(y) = \begin{cases} 1 - F_X\left(\frac{y-b}{a}\right) & \text{se } X \text{ è ass. continua} \\ 1 - F_X\left(\frac{y-b}{a}\right) + p_X\left(\frac{y-b}{a}\right) & \text{se } X \text{ è discreta.} \end{cases}$$

Inoltre, se X è continua

$$f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right),$$

mentre se è discreta $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$.

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

$$F_Y(y) = 1 - F_X\left(\frac{y-b}{a}\right) + p_X\left(\frac{y-b}{a}\right).$$

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,

$$p_Y(y)=P(Y=y)=P(aX+b=y)=P\left(X=\frac{y-b}{a}\right)=p_X\left(\frac{y-b}{a}\right)$$

in cui il fatto che ci sia l'uguaglianza rende insignificante il segno di a. □

7.1.1. 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 7.4. Consideriamo la funzione $f(x)=\mathrm{e}^{-x}$ 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

$$\int_{-\infty}^{+\infty}f(x)\,\mathrm{d}x=\int_0^1\mathrm{e}^{-x}\,\mathrm{d}x=1-\mathrm{e}^{-1}<1.$$

Allora la funzione

$$f_X(x)=\begin{cases}(1-\mathrm{e}^{-1})^{-1}\mathrm{e}^{-x} & x\in(0,1)\\ 0 & \text{altrimenti}\end{cases}$$

rappresentata in Figura 7.4 è una densità di probabilità.

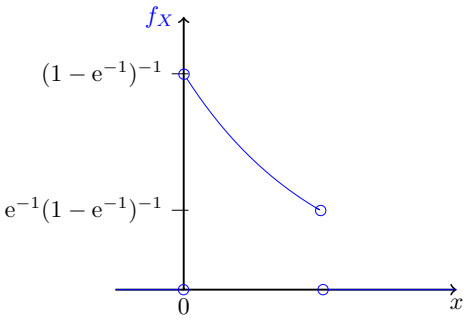


Figura 7.4. Densità di una trasformazione lineare della v.a. uniforme su $[0,1]$

Esempio 7.5. 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

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \begin{cases} 0 & x < 0 \\ \frac{1}{\pi} x & 0 \leqslant x < \pi \\ 1 & x \geqslant \pi. \end{cases}$$

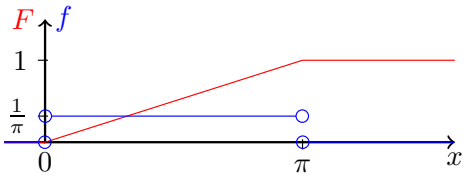


Figura 7.5. Funzione di ripartizione e densità della v.a. uniforme su $[0, \pi]$

Esempio 7.6. Sia ora $f(x) = c\,\mathrm{e}^{-|x|}$ 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,

$$1 = \int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = \int_{-\infty}^{+\infty} c\,\mathrm{e}^{-|x|}\mathrm{d}x = c\left(\int_{-\infty}^{0} \mathrm{e}^{x}\,\mathrm{d}x + \int_0^{+\infty} \mathrm{e}^{-x}\,\mathrm{d}x\right) = c\,[\mathrm{e}^{x}]_{-\infty}^{0} + c\,[-\mathrm{e}^{-x}]_0^{+\infty} = 2c,$$

da cui $c = 1/2$, quindi $f(x) = \frac{1}{2}\mathrm{e}^{-|x|}$ è una densità di probabilità. La corrispondente funzione di ripartizione è

$$F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t = \frac{1}{2}\int_{-\infty}^{x} \mathrm{e}^{-|t|}\mathrm{d}t = \begin{cases} \frac{1}{2}\int_{-\infty}^{x} \mathrm{e}^{-t}\,\mathrm{d}t = \frac{1}{2}\mathrm{e}^{x} & x < 0 \\ \frac{1}{2}\int_{-\infty}^{0} \mathrm{e}^{t}\,\mathrm{d}t + \frac{1}{2}\int_0^{x} \mathrm{e}^{-t}\,\mathrm{d}t = 1 - \frac{1}{2}\mathrm{e}^{-x} & x > 0. \end{cases}$$

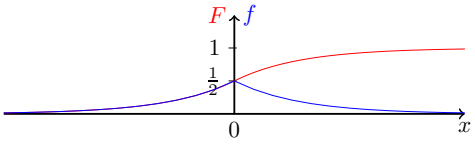


Figura 7.6. Densità di una trasformazione lineare della v.a. uniforme su $[0, 1]$

7.2. 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 p_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:

$$p_Y(y) = \sum_{x \in g^{-1}(\{y\})} p_X(x).$$

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[3]:

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 7.7. Sia X una variabile aleatoria (assolutamente continua) di densità

$$f_X(x) = \begin{cases} \mathrm{e}^{-x} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$$

e sia $g : \mathbb{R} \to \mathbb{R}$ la funzione $g(x) = \mathrm{e}^{-x}$. Vogliamo determinare la legge della variabile aleatoria $Y = \mathrm{e}^{-X}$.

Cominciamo dalla definizione:

$$F_Y(y) = P(Y \leqslant y) = P(\mathrm{e}^{-X} \leqslant y) = \begin{cases} P(-X \leqslant \log y) = P(X \geqslant -\log y) & y > 0 \\ 0 & y \leqslant 0 \end{cases}$$

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:

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = \int_0^x \mathrm{e}^{-t}\,\mathrm{d}t = 1 - \mathrm{e}^{-x}, \qquad x > 0$$

e identicamente nulla altrimenti. Allora

$$F_Y(y) = \begin{cases} 0 & y \leqslant 0 \\ 1 - (1 - \mathrm{e}^{-(-\log y)}) = y & 0 < y < 1 \\ 1 & y \geqslant 1 \end{cases}$$

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 $y > 0$)

$$\begin{aligned} f_Y(y) &= F_Y'(y) = \frac{\mathrm{d}}{\mathrm{d}y}\,(1 - F_X(-\log y)) \\ &= -F_X'(-\log y)\,\frac{\mathrm{d}}{\mathrm{d}y}\,(-\log y) = -f_X(-\log y)\left(-\frac{1}{y}\right) \\ &= \frac{1}{y} f_X(-\log y) \\ \text{per } -\log y > 0 \Leftrightarrow y < 1 &= \frac{1}{y}\mathrm{e}^{-(-\log y)} = 1. \end{aligned}$$

3. 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 7.8. Sia X una variabile aleatoria (assolutamente continua) di densità

$$f_X(x) = \begin{cases} \mathrm{e}^{-x} & x>0 \\ 0 & x\leqslant 0 \end{cases}$$

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

$$\begin{aligned} P(|1-X|\leqslant\sqrt{z}) &= P(-\sqrt{z}\leqslant 1-X\leqslant\sqrt{z}) \\ &= P(1+\sqrt{z}\geqslant X\geqslant 1-\sqrt{z}) \\ &= F_X(1+\sqrt{z}) - F_X(1-\sqrt{z}) \\ &= (1-\mathrm{e}^{-(1+\sqrt{z})})\,\mathbb{1}_{\{1+\sqrt{z}>0\}} - (1-\mathrm{e}^{-(1-\sqrt{z})})\,\mathbb{1}_{\{1-\sqrt{z}>0\}} \end{aligned}$$

e, riassumendo,

$$F_Z(z) = \begin{cases} 0 & z\leqslant 0 \\ \mathrm{e}^{-(1-\sqrt{z})} - \mathrm{e}^{-(1+\sqrt{z})} & 0<z<1 \\ 1-\mathrm{e}^{-(1+\sqrt{z})} & z\geqslant 1. \end{cases}$$

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 $z>0$,

$$\begin{aligned} f_Z(z) &= f_X(1+\sqrt{z})\frac{\mathrm{d}}{\mathrm{d}z}(1+\sqrt{z}) - f_X(1-\sqrt{z})\frac{\mathrm{d}}{\mathrm{d}z}(1-\sqrt{z}) \\ &= \frac{1}{2\sqrt{z}}\,(f_X(1+\sqrt{z}) + f_X(1-\sqrt{z})). \end{aligned}$$

Ora possiamo andare a inserire la forma esplicita di f_X, facendo attenzione al suo dominio, in particolare nel secondo addendo. Abbiamo

$$f_Z(z) = \begin{cases} 0 & z<0 \\ \frac{1}{2\sqrt{z}}\,(\mathrm{e}^{-(1+\sqrt{z})} + \mathrm{e}^{-(1-\sqrt{z})}) & 0<z<1 \\ \frac{1}{2\sqrt{z}}\,\mathrm{e}^{-(1+\sqrt{z})} & z>1. \end{cases}$$

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 7.9. (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 7.10. Cerchiamo di capire cosa significa la condizione $P(g'(X)=0)=0$ nel Teorema 7.9. 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. 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.

Osserviamo infine che potremmo rilassare ulteriormente le ipotesi: infatti per tutti gli $x\in\mathbb{R}$ per cui $f_X(x)=0$ (anche intervalli o semirette) non abbiamo alcuna restrizione sul comportamento di g, potrebbe addirittura non essere definita, senza che questo causi alcun problema.

Dimostrazione. (TEOREMA 7.9) Cominciamo considerando il caso in cui g sia una funzione C^1 strettamente crescente. Allora g è invertibile e la sua inversa è a sua volta strettamente crescente e dunque conserva le disuguaglianze. Dunque

$$\begin{aligned}
P(a\leqslant Y\leqslant b) &= P(g^{-1}(a)\leqslant g^{-1}(Y)\leqslant g^{-1}(b))\\
&= P(g^{-1}(a)\leqslant X\leqslant g^{-1}(b))\\
&= \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)\,\mathrm{d}x\\
&= \int_a^b f_X(g^{-1}(y))\cdot\frac{1}{g'(g^{-1}(y))}\,\mathrm{d}y\\
&= \int_a^b f_Y(y)\,\mathrm{d}y
\end{aligned}$$

in cui abbiamo sfruttato il cambio di variabile $y=g(x)$, da cui $x=g^{-1}(y)$ e $\mathrm{d}x=\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\,\mathrm{d}y$ (sfruttando la stretta crescenza di g e g^{-1}) e il teorema della funzione inversa, per cui

$$\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)=\frac{1}{g'(g^{-1}(y))}.$$

Nel caso g sia strettamente descrescente il procedimento è del tutto analogo, con la sola differenza della comparsa di un valore assoluto nel cambio di misura di integrazione e di un (doppio) cambio di estremi. Possiamo allora riassumere i due casi in uno solo: se g è C^1 e strettamente monotona, allora

$$f_Y(y)=\frac{f_X(g^{-1}(y))}{|g'(g^{-1}(y))|}.$$

Osserviamo però che quanto abbiamo fatto è vero anche se l'ipotesi di appartenenza a C^1 e quella di monotonia stretta valgono all'interno di un intervallo. Possiamo quindi ragionare separatamente sui vari intervalli in cui $g'\neq 0$ e concludere grazie a quanto notato nell'Osservazione Osservazione 7.10. □

Vediamo come usare il Teorema 7.9 in un esempio.

Esempio 7.11. Rimettiamoci nello stesso caso dell'Esempio 7.8. 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 7.9. 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 7.9

$$f_Z(z) = \begin{cases} 0 & z<0 \\ \frac{f_X(1)}{|g'(1)|} = +\infty & z=0 \\ \frac{f_X(1-\sqrt{z})}{|g'(1-\sqrt{z})|} + \frac{f_X(1+\sqrt{z})}{|g'(1+\sqrt{z})|} & z>0. \end{cases}$$

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

$$f_Z(z) = \begin{cases} 0 & z<0 \\ \frac{e^{-(1-\sqrt{z})}}{2\sqrt{z}} + \frac{e^{-(1+\sqrt{z})}}{2\sqrt{z}} & 0<z<1 \\ \frac{e^{-(1+\sqrt{z})}}{2\sqrt{z}} & z>1. \end{cases}$$

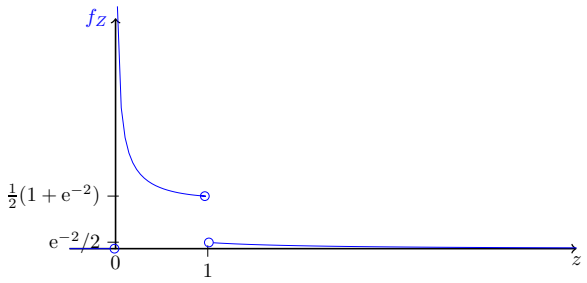


Figura 7.7. Densità della variabile aleatoria Z

7.3. PROBLEMI

Problema 34. Abbiamo una variabile aleatoria assolutamente continua X la cui densità di probabilità è

$$f_X(x) = \begin{cases} e^{-(x+c)} & x \geq 3.3 \\ 0 & x<3.3 \end{cases}$$

così come una funzione $g(t) = t^{1/2}$ definita per $t \geq 0$.

1. Quali valori può assumere la costante c?
2. Sia $Y = g(X)$. Qual è la probabilità che Y sia minore o uguale di 2.2?
3. Qual è il valore della funzione di densità di probabilità di Y in 2.3?

Problema 35. Sia X una variabile aleatoria assolutamente continua uniforme sull'intervallo $[0,1]$, ossia di funzione di densità costantemente 1 in $(0,1)$ e nulla altrove. Sia inoltre $g\colon \mathbb{R}^+ \to \mathbb{R}$ definita da $g(x) = \log(x^{-1})$. Qual è la distribuzione di $Y = g(X)$?

Problema 36. La variabile aleatoria X descrive il risultato del lancio di un dado bilanciato a 12 facce. La variabile aleatoria Y descrive il risultato di un particolare attacco in un qualche gioco che usa dadi a 12 facce, usando il risultato del lancio X e molti altri modificatori ed è definita come $Y = f(X)$, con

$$f(x) = 2x^3 - 21x^2 + 60x - 7.$$

1. Quanto vale $P(Y = 45)$?

2. Quanto vale $P(Y < 45)$?

CAPITOLO 8
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 8.1. *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*

$$\mathcal{R}_V = \mathcal{R}_{X,Y} = \mathcal{R}_X \times \mathcal{R}_Y = \{(x,y) \in \mathbb{R}^2 : x \in \mathcal{R}_X, y \in \mathcal{R}_Y\}.$$

In modo del tutto analogo possiamo definire i vettori aleatori di dimensione $d \geqslant 1$, *detti anche* variabili aleatorie multivariate.

Osservazione 8.2. 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 realtà, come osservato già in altri passaggi, è solo una questione di nomenclatura definire variabili aleatorie solamente quelle funzioni a valori in $(\mathbb{R}, \mathcal{B})$: potremmo anche definire le variabili aleatorie a valori in un qualunque spazio misurabile, di cui il caso reale univariato e quello reale multivariato non sono che esepi particolari.

DEFINIZIONE 8.3. *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 8.4. 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 8.5. *Data una variabile aleatoria doppia (X,Y) si dice* funzione di ripartizione di X condizionata a Y *la funzione $F_{X|Y}(x\,|\,y) := \frac{F_{X,Y}(x,y)}{F_Y(y)}$.*

Questa funzione è la probabilità dell'evento che immaginiamo: $F_{X|Y}(x\,|\,y) = P(X \leqslant x\,|\,Y \leqslant y)$. Non ci sorprende dunque quello che vediamo ora.

DEFINIZIONE 8.6. *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 8.7. *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 8.8. *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. □

Per quanto già visto in precedenza ne segue immediatamente una versione in termini delle funzioni di ripartizione condizionate.

PROPOSIZIONE 8.9. *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 8.10. 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 nella Definizione 4.20.

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). Per avere enunciati e risultati più specifici, possiamo concentrarci sui casi nei quali sappiamo quale tipo di variabile aleatoria abbiamo.

8.1. VETTORI ALEATORI DISCRETI

In questa sezione consideriamo variabili aleatorie bivariate (X,Y) in cui sia X sia Y sono variabili aleatorie discrete. In questo modo possiamo far entrare in gioco le densità discrete.

DEFINIZIONE 8.11. *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 $p_{X,Y} \colon \mathbb{R}^2 \to [0,1]$ definita da*

$$p_{X,Y}(x,y) = P(X = x, Y = y).$$

Si dice inoltre densità discreta condizionale di X data Y *la funzione*

$$p_{X|Y}(x\,|\,y) = \begin{cases} P(X = x\,|\,Y = y) & y \in \mathcal{R}_Y \\ 0 & y \in \mathcal{R}_Y^c. \end{cases}$$

Osservazione 8.12. Dalla definizione di densità discreta congiunta ricaviamo immediatamente le seguenti proprietà:

- per ogni (x,y) in $\mathbb{R}^2$, $0 \leqslant p_{X,Y}(x,y) \leqslant 1$
- $p_{X,Y}(x,y) = 0$ sui valori impossibili, cioè se $x \in \mathcal{R}_X^c$ o $y \in \mathcal{R}_Y^c$
- vale l'identità

$$\sum_{(x,y)\in\mathbb{R}^2} p_{X,Y}(x,y) = 1.$$

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 $p_{X,Y}$ è identicamente nulla. In effetti possiamo scrivere

$$\sum_{(x,y)\in\mathbb{R}^2} p_{X,Y}(x,y) = \sum_{x\in\mathcal{R}_X, y\in\mathcal{R}_Y} p_{X,Y}(x,y) = 1.$$

Vediamo ora alcune proprietà delle coppie di variabili aleatorie discrete.

PROPOSIZIONE 8.13. *Sia (X,Y) una coppia di variabili aleatorie discrete. Valgono le seguenti uguaglianze:*

i. per ogni $(x,y) \in \mathbb{R}^2$,

$$F_{X,Y}(x,y) = \sum_{(\xi,\eta)\in\mathcal{R}_{X,Y}} \mathbb{1}_{\{\xi\leqslant x\}} \mathbb{1}_{\{\eta\leqslant y\}} p_{X,Y}(\xi,\eta)$$

ii. per ogni $(x,y) \in \mathbb{R}^2$, $p_{X,Y}(x,y) = p_{X|Y}(x|y)\, p_Y(y)$

iii. per ogni $x \in \mathbb{R}$, $p_X(x) = \sum_{y\in\mathcal{R}_Y} p_{X,Y}(x,y)$

iv. le variabili aleatorie X e Y sono indipendenti se e solo se, per ogni $(x,y) \in \mathcal{R}_{X,Y}$, $p_{X,Y}(x,y) = p_X(x)\, p_Y(y)$

v. le variabili aleatorie X e Y sono indipendenti se e solo se, per ogni $(x,y) \in \mathcal{R}_{X,Y}$, $p_X(x) = p_{X|Y}(x|y)$ e $p_Y(y) = p_{Y|X}(y|x)$.

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

$$\begin{aligned} p_{X|Y}(x|y)\, p_Y(y) &= P(X=x|Y=y)\, P(Y=y) \\ &= \frac{P(X=x, Y=y)}{P(Y=y)} P(Y=y) \\ &= p_{X,Y}(x,y). \end{aligned}$$

iii. Iniziamo riscrivendo il secondo membro, sfruttando l'identità appena mostrata:

$$\begin{aligned} \sum_{y\in\mathcal{R}_Y} p_{X,Y}(x,y) &= \sum_{y\in\mathcal{R}_Y} p_{Y|X}(y|x)\, p_X(x) \\ &= p_X(x) \sum_{y\in\mathcal{R}_Y} p_{Y|X}(y|x) \\ &= p_X(x) \sum_{y\in\mathcal{R}_Y} P(Y=y|X=x) \\ &= p_X(x), \end{aligned}$$

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, Proposizione 8.9.

v. Segue dalla iv. e dalla ii. □

Osservazione 8.14. Vale la pena osservare che, se è nota $p_{X,Y}$, la proprietà iv. è molto pratica per verificare (o confutare) l'indipendenza di X e Y.

Esempio 8.15. 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:

$$p_{Y|X}(y\,|\,0)=\begin{cases}1/6 & y\in\{1,\dots,6\}\\ 0 & \text{altrimenti}\end{cases}\qquad p_{Y|X}(y\,|\,1)=\begin{cases}1/8 & y\in\{1,\dots,8\}\\ 0 & \text{altrimenti,}\end{cases}$$

che potremmo anche scrivere in modo più compatto come

$$p_{Y|X}(y\,|\,x)=\begin{cases}1/6 & x=0, y\in\{1,\dots,6\}\\ 1/8 & x=1, y\in\{1,\dots,8\}\\ 0 & \text{altrimenti.}\end{cases}$$

Poi andiamo a ricavarci la densità discreta congiunta, ricordando che $p_{X,Y}(x,y)=p_{Y|X}(y\,|\,x)\,p_X(x)$:

$$p_{X,Y}(x,y)=\begin{cases}1/12 & x=0, y\in\{1,\dots,6\}\\ 1/16 & x=1, y\in\{1,\dots,8\}\\ 0 & \text{altrimenti.}\end{cases}$$

A questo punto abbiamo tutti gli ingredienti necessari per calcolare la densità discreta di Y, sommando su tutti i possibili valori di X:

$$p_Y(y)=\sum_{x\in\mathcal{R}_X}p_{X,Y}(x,y)=\begin{cases}7/48 & y\in\{1,\dots,6\}\\ 1/16 & y\in\{7,8\}\\ 0 & \text{altrimenti.}\end{cases}$$

Ora 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 $p_{X,Y}(x,y)=p_X(x)\,p_Y(y)$, ma

$$p_X(x)\,p_Y(y)=\begin{cases}7/96 & x=0, y\in\{1,\dots,6\}\\ 7/96 & x=1, y\in\{1,\dots,6\}\\ 1/32 & x=1, y\in\{7,8\}\\ 0 & \text{altrimenti}\end{cases}\neq p_{X,Y}(x,y),$$

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 8.16. (SOMMA DI VARIABILI ALEATORIE DISCRETE) *Siano X e Y due variabili aleatorie sullo stesso spazio di probabilità $(\Omega,\mathcal{F},P)$ con densità congiunta $p_{X,Y}$. La loro somma ha densità discreta*

$$p_{X+Y}(z)=\sum_{x\in\mathcal{R}_X}p_{X,Y}(x,z-x).$$

Dimostrazione. Dalle definizioni abbiamo

$$\begin{aligned}p_{X+Y}(z) &= P(X+Y=z)\\ &= P\left(\bigcup_{x\in\mathcal{R}_X}\{X=x, X+Y=z\}\right)\\ &= P\left(\bigcup_{x\in\mathcal{R}_X}\{X=x, Y=z-x\}\right)\\ &= \sum_{x\in\mathcal{R}_X}P(X=x, Y=z-x)\\ &= \sum_{x\in\mathcal{R}_X}p_{X,Y}(x,z-x)\end{aligned}$$

e abbiamo così la densità discreta della variabile aleatoria $Z = X + Y$. □

Osservazione 8.17. Se le variabili aleatorie X e Y sono indipendenti, allora

$$p_{X+Y}(z) = \sum_{x \in \mathcal{R}_X} p_X(x)\, p_Y(z-x).$$

Esempio 8.18. 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, $p_{S,X}(s,x)$ e la densità discreta condizionata di S data X, $p_{S|X}(s\,|\,x)$.

Partendo dalla definizione,

$$\begin{aligned} p_{S,X}(s,x) &= P(S=s, X=x) = P(X+Y=s, X=x) \\ &= P(Y=s-x, X=x) = p_{X,Y}(x, s-x) = p_X(x)\, p_Y(s-x), \end{aligned}$$

dove nell'ultimo passaggio abbiamo sfruttato il fatto che X e Y siano indipendenti.

Passiamo alla densità discreta condizionata,

$$p_{S|X}(s\,|\,x) = \frac{p_{S,X}(s,x)}{p_X(x)} = p_Y(s-x).$$

Finora non abbiamo usato la particolare forma delle densità discrete di X e Y:

$$p_X(x) = p_Y(x) = \begin{cases} 1/10 & x \in \{1, \ldots, 10\} \\ 0 & \text{altrimenti,} \end{cases}$$

se andiamo a scriverle nelle identità ottenute sopra, abbiamo

$$p_{S,X}(s,x) = \begin{cases} 1/100 & s \in \{2, \ldots, 20\}, x \in \{1 \vee (s-10), \ldots, 10 \wedge (s-1)\} \\ 0 & \text{altrimenti} \end{cases}$$

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,

$$p_{S|X}(s\,|\,x) = \begin{cases} 1/10 & x \in \{1, \ldots, 10\}, s \in \{x+1, \ldots, x+10\} \\ 0 & \text{altrimenti.} \end{cases}$$

Per quanto riguarda la densità discreta di S abbiamo

$$p_S(s) = \sum_{x=1}^{10} p_{S,X}(s,x) = \begin{cases} 1/100 & s = 2 \\ 2/100 & s = 3 \\ \vdots & \vdots \\ 10/100 & s = 11 \\ \vdots & \vdots \\ 1/100 & s = 20 \\ 0 & s \in \{2, \ldots, 20\}^c. \end{cases}$$

8.2. 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 8.19. *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 8.20. 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 8.21. *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 8.22. Anche nel caso assolutamente continuo abbiamo l'analogo dell'Osservazione 8.12, 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 $p_{X,Y}$, 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 in Sezione 7.1.1 per il caso unidimensionale.

Esempio 8.23. Sia $f_{X,Y}(x,y) = \mathrm{e}^{-x}$ 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,

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)\,\mathrm{d}y = \int_0^x \mathrm{e}^{-x}\,\mathrm{d}y = x\,\mathrm{e}^{-x}.$$

Se vogliamo anche la funzione di ripartizione di X, possiamo ottenerla come

$$F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t = 1 - (x+1)\,\mathrm{e}^{-x}$$

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 8.24. *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 8.25. 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 8.26. 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 8.27. Due variabili aleatorie X e Y, assolutamente continue, hanno densità congiunta

$$f_{X,Y}(x,y)=6\,\mathrm{e}^{-2x}\mathrm{e}^{-3y}$$

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,

$$f_X(x)=\int_{-\infty}^{+\infty}f_{X,Y}(x,y)\,\mathrm{d}y=\begin{cases}\int_0^{+\infty}6\,\mathrm{e}^{-2x}\mathrm{e}^{-3y}\,\mathrm{d}y=2\,\mathrm{e}^{-2x} & x>0\\ 0 & x\leqslant 0\end{cases}$$

e per Y

$$f_Y(y)=\int_{-\infty}^{+\infty}f_{X,Y}(x,y)\,\mathrm{d}x=\begin{cases}\int_0^{+\infty}6\,\mathrm{e}^{-2x}\mathrm{e}^{-3y}\,\mathrm{d}x=3\,\mathrm{e}^{-3y} & y>0\\ 0 & y\leqslant 0.\end{cases}$$

A questo punto non ci resta che verificare l'indipendenza confrontando il prodotto delle densità marginali con la densità congiunta:

$$f_X(x)\,f_Y(y)=2\,\mathrm{e}^{-2x}3\,\mathrm{e}^{-3y}=6\,\mathrm{e}^{-2x}\mathrm{e}^{-3y}=f_{X,Y}(x,y).$$

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 8.28. *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 $y=-x+z$, evidenziati in grigio nella Figura 8.1, quindi

$$P((X,Y)\in E)=\iint_E f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x=\int_{-\infty}^{+\infty}\int_{-\infty}^{z-x}f_{X,Y}(x,y)\,\mathrm{d}y\,\mathrm{d}x.$$

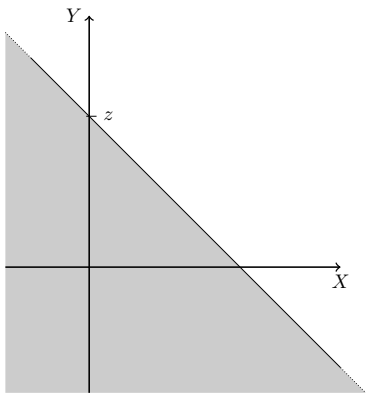


Figura 8.1. L'evento E è quello in grigio in figura

Infine, visto che siamo interessati alla densità, deriviamo in z e abbiamo concluso. □

Esempio 8.29. 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 $y > x$, in grigio in Figura 8.2.

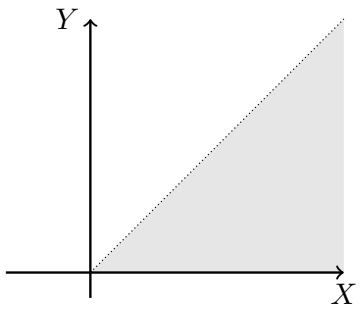


Figura 8.2. In grigio il supporto di $f_{X,Y}$

Sappiamo, dalla Proposizione 8.28, 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 8.1). 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 8.3.

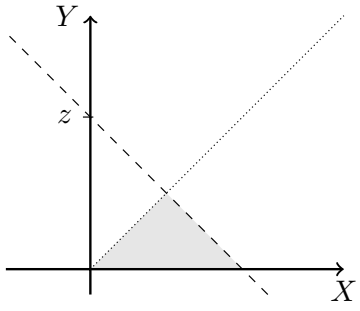


Figura 8.3. In grigio il dominio di integrazione di $f_{X,Y}$

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

$$\begin{aligned} F_{X+Y}(z) &= \int_0^{\frac{z}{2}} \int_y^{z-y} \mathrm{e}^{-x}\,\mathrm{d}x\,\mathrm{d}y = \int_0^{\frac{z}{2}} [-\mathrm{e}^{-x}]_y^{z-y}\,\mathrm{d}y \\ &= \int_0^{\frac{z}{2}} \mathrm{e}^{-y} - \mathrm{e}^{-(z-y)}\,\mathrm{d}y = [-\mathrm{e}^{-y}]_0^{z/2} - \mathrm{e}^{-z}[\mathrm{e}^{y}]_0^{z/2} \\ &= 1 - \mathrm{e}^{-\frac{z}{2}} - \mathrm{e}^{-\frac{z}{2}} + \mathrm{e}^{-z} = \left(1-\mathrm{e}^{-\frac{z}{2}}\right)^2 \end{aligned}$$

per ogni $z \geqslant 0$. A questo punto, per avere f_{X+Y} possiamo derivare in z.

Esempio 8.30. Siano X, Y due variabili aleatorie indipendenti e identicamente distribuite[1] con densità $f(t) = \mathrm{e}^{-t}$ 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

$$f_{X,Y}(x,y) = \mathrm{e}^{-x}\mathrm{e}^{-y}$$

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 8.4.

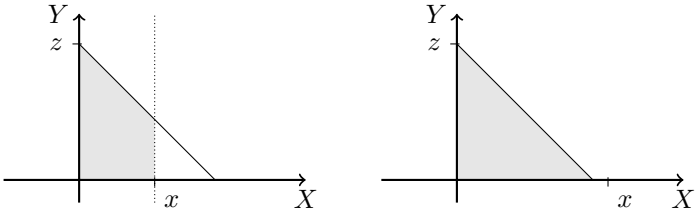


Figura 8.4. Il dominio di integrazione E in grigio, a sinistra se $0 < x < z$, a destra se $x \geqslant z$

Allora, se $x \geqslant z$,

$$F_{X,S}(x,z) = \int_0^z \mathrm{e}^{-t} \int_0^{z-t} \mathrm{e}^{-u}\,\mathrm{d}u\,\mathrm{d}t = \int_0^z \mathrm{e}^{-t}(1 - \mathrm{e}^{-z}\mathrm{e}^{t})\,\mathrm{d}t = 1 - \mathrm{e}^{-z} - z\mathrm{e}^{-z}.$$

Se invece $0 < x < z$,

$$F_{X,S}(x,z) = \int_0^x \mathrm{e}^{-t} \int_0^{z-t} \mathrm{e}^{-u}\,\mathrm{d}u\,\mathrm{d}t = \int_0^x \mathrm{e}^{-t}(1 - \mathrm{e}^{-z}\mathrm{e}^{t})\,\mathrm{d}t = 1 - \mathrm{e}^{-x} - x\mathrm{e}^{-z}.$$

Ora possiamo ricavare la densità congiunta derivando in x e z:

$$f_{X,S}(x,z) = \frac{\partial^2 F_{X,S}}{\partial x\,\partial z} = \begin{cases} \mathrm{e}^{-z} & 0 < x < z \\ 0 & x \geqslant z. \end{cases}$$

Vogliamo determinare $f_{X|S}(x\,|\,z) = \frac{f_{X,S}(x,z)}{f_S(z)}$, ma ci occorre ancora f_S,

$$f_S(z) = \int_{\mathbb{R}} f_{X,S}(x,z)\,\mathrm{d}x = \int_0^z \mathrm{e}^{-z}\,\mathrm{d}x = z\mathrm{e}^{-z}.$$

Quindi abbiamo, per $0 < x < z$,

$$f_{X|S}(x\,|\,z) = \frac{\mathrm{e}^{-z}}{z\mathrm{e}^{-z}} = \frac{1}{z}.$$

1. Questa sarà una richiesta molto comune, motivo per cui esiste una versione abbreviata di *indipendente e identicamente distribuita*, ossia la sigla *i.i.d.*.

8.3. 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 8.31. Tra le studentesse e gli studenti dell'Università di Otnert, il 52% studiano materie umanistiche e il 48% studiano materie scientifiche. Il tempo di studio al giorno per chi studia materie scientifiche è distribuito in modo uniforme tra 155 e 180 minuti, mentre per chi studia materie umanistiche è distribuito in modo uniforme tra 143 e 166 minuti[2]. 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

$$p_X(x) = \begin{cases} 0.48 & x = 0 \\ 0.52 & x = 1 \\ 0 & x \in \{0,1\}^c. \end{cases}$$

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 $(x, y) \in \mathbb{R}^2$,

$$f_{Y|X}(y\,|\,x) = \begin{cases} 1/25 & y \in [155, 180], x = 0 \\ 1/23 & y \in [143, 166], x = 1 \\ 0 & \text{altrimenti.} \end{cases}$$

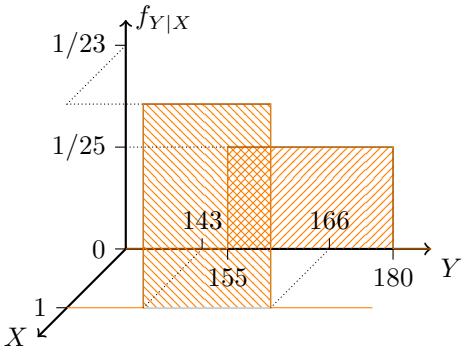


Figura 8.5. Densità condizionale di Y data X. Entrambi i rettangolini hanno area 1

2. Questo significa che se in un dato giorno scegliamo a caso una persona che studia nell'ambito umanistico, il tempo che dedicherà allo studio sarà una variabile aleatoria distribuita in modo uniforme, mentre per un'altra persona dell'ambito umanistico sarà una diversa variabile aleatoria, sebbene identicamente distribuita.

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",

$$f_{X,Y}(x,y) = f_{Y|X}(y|x)\,p_X(x),$$

un ibrido tra una densità congiunta e una densità discreta congiunta,

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{25}\cdot 0.48 = 0.0192 & y \in [155,180], x = 0 \\ \frac{1}{23}\cdot 0.52 = 0.0226087 & y \in [143,166], x = 1 \\ 0 & \text{altrimenti.} \end{cases}$$

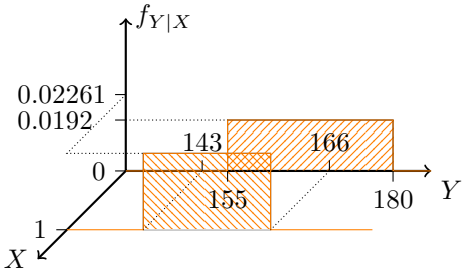


Figura 8.6. Densità congiunta di X e Y. La somma delle aree dei rettangolini è 1

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

$$P(X = x \,|\, Y < 160) = \frac{P(X = x, Y < 160)}{P(Y < 160)} = \frac{\int_{143}^{160} f_{X,Y}(x,y)\,\mathrm{d}y}{F_Y(160)} \approx \begin{cases} 0.2 & x = 0 \\ 0.8 & x = 1. \end{cases}$$

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[3]. Tuttavia possiamo usare la densità condizionata "ibrida"

$$f_{X|Y}(x\,|\,160) = \frac{f_{X,Y}(x,160)}{f_Y(160)} = \frac{f_{X,Y}(x,160)}{0.0418087} \approx \begin{cases} 0.46 & x = 0 \\ 0.54 & x = 1. \end{cases}$$

3. 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.

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.

8.4. PROBLEMI

Problema 37. In un gioco da tavolo un giocatore lancia tre dadi: due con 8 facce e uno con quattro facce. I valori ottenuti coi tre dadi sono poi sommati per determinare i danni inflitti a una creatura che il giocatore sta combattendo (all'interno del gioco). Se il giocatore ottiene almeno 13 sommando i risultati dei 3 lanci, la creatura viene sconfitta.

1. Qual è la probabilità che il giocatore sconfigga la creatura?
2. Supponiamo ora che il giocatore abbia sconfitto la creatura. Qual è la probabilità che il dado a 4 facce sia stato determinante per il suo successo (ossia che senza quel dado il risultato sarebbe stato insufficiente a sconfiggere la creatura)?

CAPITOLO 9

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 aleatorie, 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, come abbiamo già discusso nel Capitolo 6, in particolare nell'Osservazione 6.24.

9.1. 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 9.1. *Una variabile aleatoria discreta X si dice* Bernoulliana (*o* variabile aleatoria di Bernoulli[1]) di parametro p, *con* $p \in [0,1]$ *se ha densità discreta*

$$p_X(x) = \begin{cases} p & x = 1 \\ 1-p & x = 0 \\ 0 & \textit{altrimenti,} \end{cases}$$

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 6.17 e 6.27, in questo caso indicatrice dell'evento "successo", che ha probabilità p. Abbiamo anche visto una rappresentazione della sua funzione di ripartizione in Figura 6.2. Nello scrivere l'esperimento aleatorio come variabile casuale abbiamo codificato "successo" con 1 e "insuccesso" con 0.

In realtà l'abbiamo incontrata altre volte, spesso con la scelta di parametro $p = 0.5$: è il lancio di una moneta. Se $p = 0.5$, la moneta è equa, altrimenti è non bilanciata.

1. Jakob Bernoulli (1655 – 1705) uno dei molti matematici della famiglia. È legato alla probabilità dalla sua opera *Ars Conjectandi*, pubblicata postuma nel 1713.

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 $P(\{1\}) = p$.

9.2. BINOMIALI

Consideriamo ora il caso in cui abbiamo n variabili aleatorie Bernoulliane, indipendenti e identicamente distribuite[2] (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 p-moneta (cioè in n ripetizioni di un esperimento Bernoulliano).

DEFINIZIONE 9.2. *Diciamo che una variabile aleatoria discreta X è una* binomiale di parametri n e p, *con $n \in \mathbb{N}^+$ 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 9.3. Come suggerito dalla notazione introdotta in Definizione 9.1, una Bernoulliana è una binomiale di parametri $n = 1$ e p, cioè la somma di una sola Bernoulliana di parametro p.

Mentre in Definizione 9.1 avevamo identificato le variabili aleatorie Bernoulliane mediante la loro legge, ossia la loro densità discreta o funzione di ripartizione, nel caso della Definizione 9.2 abbiamo caratterizzato le binomiali a partire da altre variabili aleatorie. Come prima cosa, quindi, andiamo a ricavare la legge di una binomiale di parametri n e p.

PROPOSIZIONE 9.4. *Se $X \sim \text{bin}(n,p)$ con $n \in \mathbb{N}^+$ e $p \in [0,1]$, allora*

$$p_X(k) = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & k \in \{0,\dots,n\} \\ 0 & \text{altrimenti} \end{cases} \qquad F_X(x) = \sum_{k=0}^{\lfloor x \rfloor} \binom{n}{k} p^k (1-p)^{n-k}.$$

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 $p_X(k)$. Ricordiamo che, dalle definizioni,

$$p_X(k) = P(X=k) = P\left(\sum_{i=1}^{n} X_i = k\right)$$

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

$$p_X(k) = P(X=k) = P\left(\bigcup_{I_k \in \mathcal{I}_k} E_{I_k}\right) = \sum_{I_k \in \mathcal{I}_k} P(E_{I_k}).$$

2. In questo caso "identicamente distribuite" equivale a dire che hanno tutte il medesimo parametro p.

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

$$p_X(k) = \sum_{I_k\in\mathcal{I}_k} P(E_{I_k}) = p^k(1-p)^{n-k}\sum_{I_k\in\mathcal{I}_k} 1$$

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 nella Sezione 2.3, di contare in quanti modi diversi possiamo scegliere k indici tra n, cioè $\binom{n}{k}$. In conclusione, per $k\in\{0,\dots,n\}$, $p_X(k) = \binom{n}{k}p^k(1-p)^{n-k}$.

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:

$$F_X(x) = \sum_{k=0}^{\lfloor x\rfloor} p_X(k) = \sum_{k=0}^{\lfloor x\rfloor}\binom{n}{k}p^k(1-p)^{n-k},$$

concludendo così la dimostrazione. □

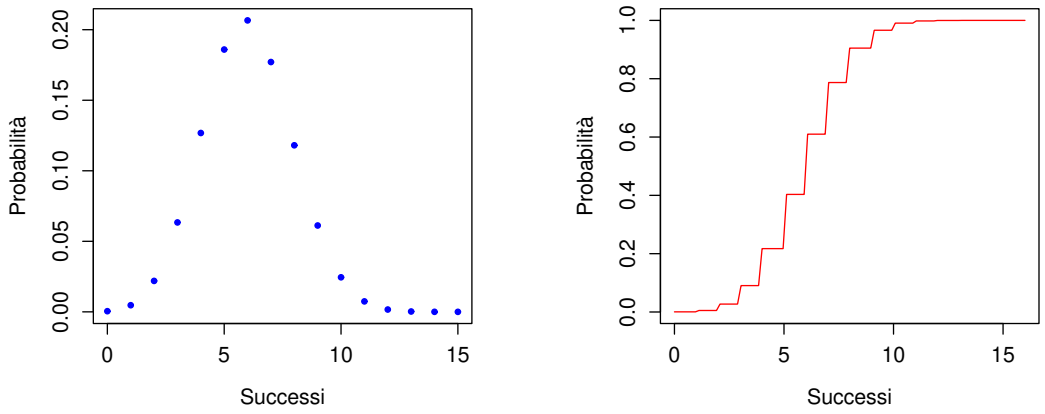


Figura 9.1. Densità discreta (sinistra) e funzione di ripartizione (destra) di bin(15, 0.4)

Osservazione 9.5. Nella dimostrazione abbiamo osservato che per una variabile aleatoria binomiale X, $p_X(k) = P(\sum_{i=1}^n X_i = k)$, con le $X_i \sim \text{bin}(1,p)$. Queste ultime sono tutte variabili aleatorie discrete, quindi potremmo farne ricorsivamente la somma, come visto nella Proposizione 8.16. Tuttavia questa strategia si presta più facilmente a errori e, nella sostanza, è analoga a quanto visto nella dimostrazione.

Osservazione 9.6. 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,

$$F_X(x) = \sum_{k=0}^{\lfloor x\rfloor} p_X(k) = \sum_{k=0}^{n} p_X(k) = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = (p+(1-p))^n = 1.$$

Esempio 9.7. **(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 $N > 1$,

$$\begin{aligned} P(N>1) &= \sum_{k=2}^{15} P(N=k) = 1 - P(N=0) - P(N=1) \\ &= 1 - p_N(0) - p_N(1) \\ &= 1 - \binom{15}{0}(1-0.02)^{15} - \binom{15}{1}0.02\,(1-0.02)^{14} \approx 3.5\% \end{aligned}$$

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

$$p_R(1) = \binom{4}{1}0.035\,(1-0.035)^3 \approx 12.7\%.$$

9.2.1. 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 p, il numero `size` di tentativi (quello che nella Definizione 9.2 abbiamo indicato con n) e la probabilità `prob` di successo di ogni tentativo (quella che nella Definizione 9.2 abbiamo indicato con p).

Per calcolare la densità discreta $p_X(11)$ di una variabile aleatoria $X \sim \text{bin}(44, 0.2)$ in R useremo la funzione `dbinom(x = 11, size = 44, prob = 0.2)`. Se nominiamo i parametri, possiamo anche passarli in ordine diverso da quello standard (ad esempio l'istruzione precedente è equivalente a `dbinom(size = 44, prob = 0.2, x = 11)`). In alternativa possiamo passarli anche senza nominarli, ma in questo caso devono essere nell'ordine predefinito, senza alcun parametro mancante in mezzo: `dbinom(11,44,0.2)`.

Funzione di ripartizione La funzione di ripartizione (cdf) per una binomiale in R 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[3] 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`.

3. La lettera `q` viene dal termine *quantile* che essa rappresenta, che definiremo in seguito.

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 9.8. Avremmo potuto rispondere alle domande nell'Esempio 9.7 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)
```

9.3. LO SCHEMA DI BERNOULLI

Nella Sezione 9.2 abbiamo introdotto le binomiali come somma finita di variabili aleatorie Bernoulliane indipendenti e identicamente distribuite. 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).

Se ci pensiamo bene, abbiamo già visto un esperimento di questo tipo negli Esempi 5.15 e 6.15. Possiamo generalizzarli, 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 6.15. Lo spazio degli esiti Ω è l'insieme delle successioni a valori in $\{0,1\}$, quindi $\Omega = \{0,1\}^{\mathbb{N}^+}$.

Stiamo considerando uno spazio prodotto infinito (visti nella Sezione 5.3), 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 9.9. 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 (vedi Esempio 6.15).

9.4. 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[4]: 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 9.10. *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 in maniera compatta $X \sim \text{geom}(p)$.*

Dalla definizione possiamo ricavare immediatamente la densità discreta di $X \sim \text{geom}(p)$:

$$p_X(k) = (1-p)^k p \tag{9.1}$$

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 9.11. 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 di densità discreta data in (9.1) 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 $p_X(k) = \mathbb{1}_{\{k=0\}}$. Se $p < 1$, allora

$$\sum_{k=0}^{+\infty} p_X(k) = \sum_{k=0}^{+\infty} (1-p)^k p = p \sum_{k=0}^{+\infty} (1-p)^k = p \cdot \frac{1}{1-(1-p)} = 1,$$

4. 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.

in cui abbiamo usato il fatto che la serie $\sum_{k=0}^{+\infty}(1-p)^k$ è geometrica[5] di ragione $1-p$ compresa tra 0 e 1.

Osservazione 9.12. 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:

$$p_{T_1}(k) = (1-p)^{k-1}p$$

per $k \in \mathbb{N}^+$. 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 9.10 il dominio era $\mathbb{N}$, per il primo successo il dominio è $\mathbb{N}^+$.

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 in (9.1) la densità discreta di una variabile aleatoria X geometrica di parametro p, adesso vediamo com'è fatta la sua funzione di ripartizione :

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \sum_{k=0}^{\lfloor x \rfloor} p_X(k) = p\sum_{k=0}^{\lfloor x \rfloor}(1-p)^k = 1-(1-p)^{\lfloor x \rfloor+1} & x \geqslant 0. \end{cases}$$

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

$$P(X > n) = 1 - F_X(n) = (1-p)^{n+1} \tag{9.2}$$

e, per $n \in \mathbb{N}$, la funzione di ripartizione è quindi

$$F_X(n) = P(X \leqslant n) = 1 - (1-p)^{n+1} \tag{9.3}$$

con l'estensione a $\mathbb{R}$ data dalle proprietà di una funzione di ripartizione discreta.

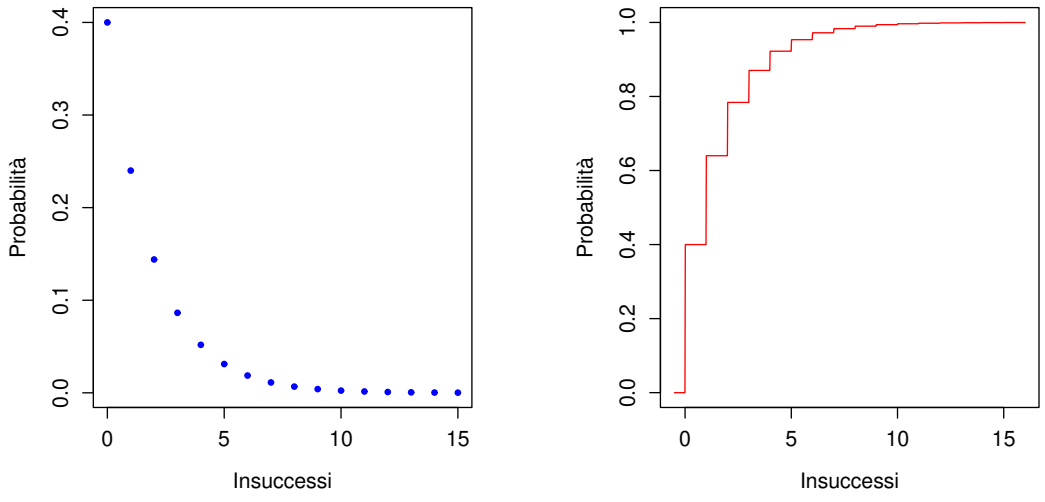


Figura 9.2. Densità discreta (sinistra) e funzione di ripartizione (destra) di geom(0.4)

PROPOSIZIONE 9.13. *Una variabile aleatoria geometrica X gode della proprietà di assenza di memoria, cioè per ogni $n, k \in \mathbb{N}$*

$$P(X \geqslant n+k \mid X \geqslant n) = P(X \geqslant k). \tag{9.4}$$

5. Qualche informazione in più sulla serie geometrica è disponibile in Appendice A.3

Dimostrazione. Partiamo dalla definizione di probabilità condizionata,

$$\begin{aligned} P(X \geqslant n+k \mid X \geqslant n) &= \frac{P((X \geqslant n+k) \cap (X \geqslant n))}{P(X \geqslant n)} \\ &= \frac{P(X \geqslant n+k)}{P(X > n-1)} \\ &= \frac{(1-p)^{n+k}}{(1-p)^n} \\ &= (1-p)^k \\ &= P(X \geqslant k), \end{aligned}$$

in cui abbiamo usato ripetutamente la (9.2). □

Osservazione 9.14. Da dove viene il nome *assenza di memoria* di questa proprietà? Possiamo leggere la (9.4) 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 9.15. Se nel Superenalotto il 67 non esce da 65 estrazioni[6], 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 67) 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 67 al Superenalotto" è una Bernoulliana di parametro $\frac{1}{15}$. Consideriamo lo schema di Bernoulli corrispondente e, in particolare, la probabilità che 67 esca alla prossima estrazione sapendo che non è uscito nelle prime 65. Chiamiamo X la variabile aleatoria che descrive l'ultima estrazione in cui non esce 67: la nostra richiesta è allora

$$P(X = 65 \mid X \geqslant 65) = \frac{P(X = 65)}{P(X > 65-1)} = \frac{\left(1 - \frac{1}{15}\right)^{65} \frac{1}{15}}{\left(1 - \frac{1}{15}\right)^{65+1}} = \frac{1}{15},$$

cioè è identica alla probabilità che 67 esca al primo tentativo.

Analogamente, la probabilità che esca per la prima volta tra almeno altre 30 estrazioni è

$$P(X \geqslant 65+30 \mid X \geqslant 65) = P(X \geqslant 30) = \left(1 - \frac{1}{15}\right)^{30} \approx 12.6\%.$$

La morale di questo esempio è che in termini di probabilità è assolutamente irrilevante che il 67 sia "in ritardo" da 65 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%).

9.4.1. 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 p e `prob` è la probabilità di successo di ogni tentativo (che abbiamo indicato con p).

6. Dati al 30 marzo 2025, non che sia rilevante, come vedremo.

La funzione di ripartizione 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 9.16. Avremmo potuto rispondere alle domande nell'Esempio 9.15 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)
```

9.5. 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 9.17. *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

$$p_X(b)=\begin{cases}\frac{\binom{m}{b}\binom{n}{k-b}}{\binom{n+m}{k}} & \max\{0,k-n\}\leqslant b\leqslant\min\{k,m\}\\ 0 & \text{altrimenti.}\end{cases}$$

Osservazione 9.18. Grazie a quanto visto nella dimostrazione della Proposizione 9.32, abbiamo immediatamente che

$$\sum_{b=0}^{k}p_X(b)=\sum_{b=\max\{0,k-n\}}^{\min\{k,m\}}p_X(b)=\frac{\sum_{b=\max\{0,k-n\}}^{\min\{k,m\}}\binom{m}{b}\binom{n}{k-b}}{\binom{n+m}{k}}=1.$$

Esempio 9.19. 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 $D \sim \text{hyp}(5, 10, 400-10)$.

9.5.1. Massima verosimiglianza (divagazione)

Nella realtà, però, in una situazione come quella dell'Esempio 9.19 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

$$P(M=m \mid D=b) = \frac{P(D=b \mid M=m)\, P(M=m)}{P(D=b)}, \tag{9.5}$$

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 (9.5) 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 $\{0, 1, \dots, t\}$, ossia che per ogni $m \in \{0, 1, \dots, t\}$, $P(M=m) = \frac{1}{t+1}$.

Il problema di massimizzare la (9.5) è diventato allora trovare il valore $\overline{m} \in \{0, 1, \dots, t\}$ che massimizza l'ipergeometrica:

$$\overline{m} = \text{argmax}_{m \in \{0,1,\dots,t\}} \frac{\binom{m}{b}\binom{t-m}{k-b}}{\binom{t}{k}} = \text{argmax}_m \binom{m}{b}\binom{t-m}{k-b}.$$

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)]
```

9.5.2. 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 p, `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 9.17).

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 9.20. 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[7] 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 20 "buone"). Vogliamo $p(2) \approx 14\%$ (possiamo calcolarla in R con la seguente funzione `dhyper(x = 2, m = 20, n = 32, k = 2)`).

Ora calcoliamo la probabilità che entrambe le carte siano assi, usando un'ipergeometrica di parametri $k = 2$, $m = 4$, $n = 48$: $p(2) \approx 0.5\%$ (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 $k = 2$, $m = 16$, $n = 36$: $p(2) \approx 9\%$ (in linguaggio R abbiamo in questo caso `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 9.21. 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?

7. 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.

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%.

Vediamo ora un legame tra le variabili aleatorie con distribuzione ipergeometrica e quelle con distribuzione binomiale.

PROPOSIZIONE 9.22. *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 9.23. 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 p_{X_i} converge alla densità discreta p_X, ossia in un qualche senso[8] le variabili ipergeometriche tendono a una binomiale o, meglio, le leggi delle ipergeometriche tendono alla legge binomiale.

9.6. POISSON

Abbiamo rotto il ghiaccio con le sequenze di variabili aleatorie. Vediamone ora altre che entrano in gioco nel seguente problema.

Esempio 9.24. In una partita di Premier League vengono segnati in media[9] 2.5 gol a partita[10]. 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.

Cosa succede se continuiamo a sviluppare una successione di questo tipo? Converge a qualcosa? E se sì, a cosa converge?

8. Vedremo meglio più avanti i concetti di convergenza per variabili aleatorie, nel Capitolo 14.

9. Non abbiamo ancora definito il concetto di media per una variabile aleatoria, anche se non manca molto, lo incontreremo nel Capitolo 10. 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.

10. Dati del 14.04.2021.

DEFINIZIONE 9.25. *Diciamo che una variabile aleatoria discreta X è* di Poisson[11] (o Poissoniana) di parametro λ, *con λ numero reale positivo, se ha densità discreta*

$$p_X(k) = \begin{cases} \frac{\lambda^k}{k!} e^{-\lambda} & k \in \mathbb{N} \\ 0 & \textit{altrimenti.} \end{cases} \tag{9.6}$$

In questo caso scriviamo $X \sim \text{Pois}(\lambda)$.

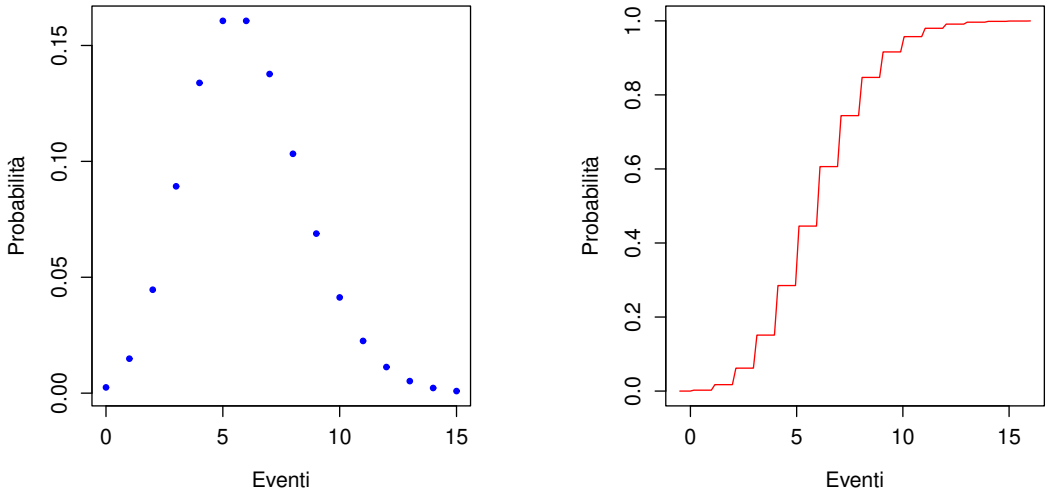


Figura 9.3. Densità discreta (sinistra) e funzione di ripartizione (destra) per Pois(6)

Osservazione 9.26. La funzione p_X definita in (9.6) 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

$$\sum_{k \in \mathcal{R}_X} p_X(k) = \sum_{k \in \mathbb{N}} \left(\frac{\lambda^k}{k!} e^{-\lambda} \right) = e^{-\lambda} \sum_{k=0}^{+\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1.$$

Osservazione 9.27. 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 10).

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[12]);

11. Siméon Denis Poisson (1781 – 1840).

12. 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.

– 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 9.24 e rendere matematicamente solido quanto detto prima sul comportamento limite della successione di binomiali.

PROPOSIZIONE 9.28. *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*

$$\lim_{n\to+\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

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 9.29. In termini di variabili aleatorie stiamo dicendo che se abbiamo una successione di variabili aleatorie binomiali

$$X_n \sim \text{bin}(n, p_n)$$

e una variabile aleatoria $X \sim \text{Pois}(\lambda)$, con $\lambda = \lim_{n\to+\infty} n \cdot p_n$, allora la successione delle densità discrete p_{X_n} converge alla densità discreta p_X, ossia la variabile aleatoria di Poisson è il "limite" delle variabili aleatorie binomiali[13].

9.6.1. 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 p 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 9.24, 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`.

13. Il termine *limite* è tra virgolette, perché non abbiamo (ancora) introdotto un concetto di limite per variabili aleatorie. Sarà un argomento del Capitolo 14.

9.7. 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 (Proposizioni 8.16 e 8.28) 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 9.30. *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 9.31. 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 8.16, abbiamo

$$\begin{aligned} p_S(k) &= \sum_{j\in\mathcal{R}_X} p_X(j)\, p_Y(k-j) \\ &= \sum_{j=0}^{+\infty} (1-p)^j p\, \mathbb{1}_{\{k-j\geqslant 0\}} (1-p)^{k-j} p \\ &= \sum_{j=0}^{k} p^2 (1-p)^k \\ &= (k+1)\, p^2 (1-p)^k. \end{aligned}$$

Questa non è la densità discreta di una geometrica, ma si tratta di un altro modello noto: le binomiali negative (approfondite in Sezione 9.8). 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 binomiali negative escono come somma di geometriche. Abbiamo già visto una famiglia di distribuzioni definite come somma di altre che quindi è un'ottima candidata alla riproducibilità. Infatti abbiamo definito le binomiali come somme di Bernoulliane ed è lecito chiedersi pertanto se le binomiali siano riproducibili.

PROPOSIZIONE 9.32. *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 $X \sim \mathrm{bin}(n,p)$ e $Y \sim \mathrm{bin}(m,p)$, tra loro indipendenti. Dalla Proposizione 8.16 abbiamo

$$p_{X+Y}(l) = \sum_{k=0}^{n} p_X(k)\, p_Y(l-k).$$

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

$$\begin{aligned} p_{X+Y}(l) &= \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{m}{l-k} p^{l-k} (1-p)^{m-(l-k)} \\ &= \sum_{k=0}^{n} p^l (1-p)^{n+m-l} \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{n}{k}\binom{m}{l-k} \\ &= p^l (1-p)^{n+m-l} \sum_{k=0}^{n} \mathbb{1}_{0\leqslant l-k\leqslant m} \binom{n}{k}\binom{m}{l-k} \end{aligned}$$

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 2.32. □

Come passo successivo possiamo chiederci se anche la famiglia delle ipergeometriche sia riproducibile. Un dubbio sulla possibilità che questo sia vero può venirci dal fatto che, a differenza del caso delle binomiali, 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.

Non rimane che un modello da affrontare: la distribuzione di Poisson. Per via del suo legame con la distribuzione binomiale, non ci dovrebbe sorprendere troppo che la famiglia delle Poisson sia riproducibile.

PROPOSIZIONE 9.33. *Le variabili aleatorie Poissoniane sono riproducibili.*

Dimostrazione. Vogliamo mostrare che, date due variabili aleatorie indipendenti $X_1 \sim \text{Pois}(\lambda_1)$ e $X_2 \sim \text{Pois}(\lambda_2)$, anche la loro somma ha legge Poissoniana. Consideriamone la densità discreta:

$$\begin{aligned} p_{X_1+X_2}(n) &= \sum_{k=0}^{n} p_{X_1}(k)\, p_{X_2}(n-k) \\ &= \sum_{k=0}^{n} \frac{n!}{n!} \frac{\lambda_1^k}{k!} \mathrm{e}^{-\lambda_1} \frac{\lambda_2^{n-k}}{(n-k)!} \mathrm{e}^{-\lambda_2} \\ &= \left[\sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k} \right] \frac{\mathrm{e}^{-(\lambda_1+\lambda_2)}}{n!} \\ &= \frac{(\lambda_1+\lambda_2)^n}{n!} \mathrm{e}^{-(\lambda_1+\lambda_2)} \end{aligned}$$

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 $X_1 + X_2 \sim \text{Pois}(\lambda_1+\lambda_2)$. □

9.8. 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 $n \in \mathbb{N}^+$

$$T_n = \inf \left\{ i \geqslant 1 : \sum_{k=1}^{i} \omega_k = n \right\},$$

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 9.34. *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 $X \sim \text{NB}(n,p)$.*

Anche in questo caso iniziamo a ricavare, dalla definizione, la funzione di densità discreta di $X \sim \text{NB}(n,p)$, per $k \geqslant 0$

$$\begin{aligned} p_X(k) &= P(X=k) = P(T_n = k+n) \\ &= P\left(\omega_{k+n} = 1, \sum_{i=1}^{k+n-1} \omega_i = n-1 \right) \\ &= p \binom{k+n-1}{n-1} p^{n-1} (1-p)^k \\ &= \binom{k+n-1}{n-1} p^n (1-p)^k \end{aligned}$$

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 9.35. 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 9.36. 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

$$\sum_{k=0}^{3} p_X(k) = \sum_{k=0}^{3} \binom{k+2}{2} \left(\frac{1}{4}\right)^3 \left(\frac{3}{4}\right)^k \approx 17\%.$$

Come accennato in precedenza, le binomiali negative sono strettamente imparentate alle geometriche. Le geometriche non sono riproducibili, ma se osserviamo cosa succede alla somma (magari iterata) di geometriche di parametro comune p possiamo convincerci che le binomiali negative possano essere riproducibili.

PROPOSIZIONE 9.37. *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 esplicito, oppure usando il fatto che una binomiale negativa è somma di geometriche. In questo secondo modo, osserviamo che ottenere il $(k+h)$-simo successo equivale ad attendere il k-simo successo e poi aspettare l'h-simo successo (come se avessimo azzerato il contatore). □

Osservazione 9.38. 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.

9.8.1. 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 p, `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 9.36 usando il seguente codice R: `pnbinom(q = 3, size = 3, prob = 0.25, lower.tail = TRUE)`.

9.9. PROBLEMI

Problema 38. (S. ROSS) Un moderno aereo civile è in grado di restare in volo se almeno la metà dei suoi motori è in funzione. Supponiamo che ogni motore abbia, indipendentemente dagli altri, una probabilità p di funzionare correttamente. È più sicuro un aereo a quattro motori o a due motori?

Problema 39. In una trasmissione digitale vengono inviati 100 kilobit[14]. Ciascun bit può essere corrotto dal rumore della trasmissione e venire ricevuto scorrettamente con probabilità stimata 0.0003.

1. Qual è il numero complessivo di bit errati che è più probabile vengano ricevuti?
2. Qual è la probabilità di ricevere esattamente quel numero di bit errati?

Problema 40. (P. BALDI) Una fabbrica di prodotti dolciari produce merendine su due linee di produzione, A e B, nelle proporzioni del 30% e 70% rispettivamente. La linea A ha una proporzione di prodotti imperfetti del 10%, contro 17% per B.

1. Qual è la probabilità che una merendina scelta a caso nella produzione della fabbrica sia imperfetta?
2. Le merendine vengono vendute in confezioni da 10 pezzi, tutti prodotti dalla medesima linea. Una di queste confezioni viene ispezionata e risulta contenere esattamente una sola merendina imperfetta. Da quale linea è più probabile che vengano le merendine nella confezione?

Problema 41. (P. BALDI) Un collezionista ha già raccolto 60 delle 100 figurine di un album. Egli acquista una bustina contenente 6 figurine (tutte diverse tra loro), tra cui naturalmente ce ne possono essere alcune che già possiede. Qual è la probabilità che tra le figurine appena acquistate ce ne siano almeno 5 che già possiede?

Problema 42. Siano $X_1 \sim \text{Pois}(\lambda_1)$ e $X_2 \sim \text{Pois}(\lambda_2)$ due variabili aleatorie indipendenti e sia S la loro somma. Determinare la legge di X_1 condizionata a S.

14. Un *kilobit* è una stringa di 1024 bit.

CAPITOLO 10

SPERANZA MATEMATICA

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 il valore atteso (o media), ossia un valore *deterministico* (cioè un numero) che ci dà in un certo senso[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.

Da lontano il concetto di media per una variabile aleatoria e per una collezione di dati (come visto nel Capitolo 1) non è troppo diverso. Vedremo poi, parlando di statistica inferenziale, che le due nozioni sono molto legate. Non sono però la medesima cosa, come risulta immediato dal fatto che in un caso abbiamo una collezione di dati (media) e nell'altro una variabile aleatoria (valore atteso).

In queste note, non volendoci appoggiare troppo alla teoria della misura, consideriamo separatamente il caso delle variabili aleatorie discrete e quello delle variabili aleatorie assolutamente continue. Questo dà (in apparenza) due modi diversi di calcolare esplicitamente il valore atteso. Nel momento in cui si vedono le cose sotto le lenti della teoria della misura, queste differenze spariscono, anche se al costo di avere rappresentazioni meno immediate dal punto di vista pratico e di perdere un po di vista il legame con la media di dati empirici.

10.1. VARIABILI ALEATORIE DISCRETE

Cominciamo col vedere la definizione di valore atteso per le variabili aleatorie discrete.

DEFINIZIONE 10.1. *Una variabile aleatoria discreta X ha* speranza finita (*indicato con $X \in \mathcal{L}^1$*) *se*

$$\sum_{x \in \mathcal{R}_X} |x| p_X(x) < +\infty.$$

Se $X \in \mathcal{L}^1$, il valore atteso, la speranza matematica *o* la media *di X è il baricentro della sua distribuzione, ossia*

$$\mathbb{E}[X] = E[X] = \sum_{k \in \mathcal{R}_X} k p_X(k).$$

Possiamo notare che la speranza è una media pesata dei possibili valori k assunti da X, i cui pesi sono le corrispondenti probabilità $p_X(k) = P(X = k)$. Da un punto di vista fisico è il centro di massa della distribuzione, il baricentro, appunto.

Osservazione 10.2. La condizione di speranza finita garantisce che $E[X] < +\infty$ per il criterio di convergenza assoluta. Se $X \geqslant 0$ e $\sum_{k \in \mathcal{R}_X} k p_X(k)$ diverge a $+\infty$ diciamo che X ha speranza infinita[2].

1. Vedremo più avanti che non è il solo.

2. Possiamo chiaramente anche considerare il caso in cui la speranza sia $-\infty$. Nella pratica quello che si fa è spezzare la variabile aleatoria nella sua parte positiva e nella sua parte negativa, entrambe non negative, e "riciclare" quanto visto finora. Il solo caso scomodo è quello in cui sia la parte positiva sia la parte negativa vanno a $+\infty$.

Possiamo quindi avere variabili aleatorie discrete con speranza finita (positiva, nulla o negativa), con speranza infinita (Esempio 10.3), ma anche non definita (la serie che la definisce potrebbe non convergere[3]). In generale, nel seguito, considereremo solamente variabili aleatorie $X \in \mathcal{L}^1$.

Esempio 10.3. (**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 $p_X(2^n) = P(T_1 = n)$, 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

$$E[X] = \sum_{x \in \mathcal{R}_X} x\, p_X(x) = \sum_{n \in \mathbb{N}^+} 2^n \cdot \frac{1}{2^n} = \sum_{n \in \mathbb{N}^+} 1 = +\infty.$$

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.

La Definizione 10.1 richiede di pesare ogni possibile risultato con la sua probabilità. Un'immediata generalizzazione, allora, è quella in cui consideriamo un'*altra* misura di probabilità come peso, in particolare la probabilità dei risultati 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*. A questo punto possiamo andare oltre e condizionare a eventi speciali, quale ad esempio il valore assunto da un'altra variabile aleatoria,

$$E[X|Y = y] = \sum_{x \in \mathcal{R}_X} x \cdot P(X = x|Y = y) = \sum_{x \in \mathcal{R}_X} x \cdot p_{X|Y}(x|y),$$

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 questo argomento, per quanto interessante ed estremamente utile, verrà solamente accennato in questo corso, nella Sezione 10.5.

Esempio 10.4. Sia $X \sim \text{bin}(1, p)$, calcoliamone la speranza. Dalla definizione abbiamo

$$E[X] = \sum_{k=0}^{1} k \cdot p_X(k) = 0 \cdot (1-p) + 1 \cdot p = p.$$

Avendo la definizione, possiamo usarla per calcolare la media di altre distribuzioni note (ad esempio quelle introdotte nel Capitolo 9) e, in generale, di una qualunque variabile aleatoria discreta. Tuttavia per farlo abbiamo bisogno di conoscere la densità discreta della variabile aleatoria (ossia, come abbiamo osservato, l'intera legge). Allo stesso tempo vorremmo usare la media per avere almeno qualche informazione di una variabile aleatoria di cui non conosciamo la legge, quindi introduciamo alcune proprietà del valore atteso che danno delle scorciatoie per calcolare il valore atteso di altre variabili aleatorie.

3. 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.

TEOREMA 10.5. *Siano X una variabile aleatoria di densità discreta p_X e $Y = g(X)$. Allora $Y \in \mathcal{L}^1$ se e solo se*

$$\sum_{k \in \mathcal{R}_X} |g(k)| p_X(k) < +\infty.$$

In questo caso

$$E[Y] = \sum_{k \in \mathcal{R}_X} g(k)\, p_X(k).$$

Dimostrazione. Iniziamo dalla condizione di speranza finita e sfruttiamo un risultato sulle trasformazioni di variabili aleatorie discrete visto nel Capitolo 7,

$$\begin{aligned}\sum_{y \in \mathcal{R}_Y} |y| \cdot p_Y(y) &= \sum_{y \in \mathcal{R}_Y} |y| \cdot \sum_{x \in g^{-1}(\{y\})} p_X(x) \\ &= \sum_{y \in \mathcal{R}_Y} \sum_{x \in g^{-1}(\{y\})} |g(x)| \cdot p_X(x) \\ &= \sum_{x \in \mathcal{R}_X} |g(x)| \cdot p_X(x),\end{aligned}$$

in cui abbiamo approfittato del fatto che se $x \in g^{-1}(\{y\})$, allora $y = g(x)$ e che $\mathcal{R}_Y = g(\mathcal{R}_X)$. Per concludere osserviamo che se abbiamo la convergenza della serie con i moduli, la seconda parte della tesi segue. □

Notiamo che per calcolare $E[Y]$ mediante il teorema precedente non abbiamo bisogno di calcolare esplicitamente p_Y. Possiamo generalizzare il Teorema 10.5 al caso dei vettori aleatori in qualunque dimensione finita (anche se vedremo enunciato e dimostrazione solo in dimensione 2).

TEOREMA 10.6. *Siano (X, Y) un vettore aleatorio di variabili aleatorie discrete con densità congiunta $p_{X,Y}$ e sia $Z = g(X, Y)$, per qualche funzione $g : \mathbb{R}^2 \to \mathbb{R}$. Allora $Z \in \mathcal{L}^1$ se e solo se*

$$\sum_{j \in \mathcal{R}_X} \sum_{k \in \mathcal{R}_Y} |g(j,k)| \cdot p_{X,Y}(j,k) < +\infty$$

e in tal caso

$$E[Z] = \sum_{j \in \mathcal{R}_X} \sum_{k \in \mathcal{R}_Y} g(j,k) \cdot p_{X,Y}(j,k).$$

Dimostrazione. Come prima cosa riscriviamo la densità discreta di Z:

$$p_Z(l) = P(Z = l) = P(g(X,Y) = l) = \sum_{(j,k) : g(j,k) = l} P((X,Y) = (j,k)) = \sum_{(j,k) : g(j,k) = l} p_{X,Y}(j,k).$$

Ora possiamo osservare che

$$\sum_{l \in \mathcal{R}_Z} |l| p_Z(l) = \sum_{l \in \mathcal{R}_Z} |l| \sum_{(j,k) : g(j,k) = l} p_{X,Y}(j,k) = \sum_{(j,k) \in \mathcal{R}_X \times \mathcal{R}_Y} |g(j,k)| p_{X,Y}(j,k)$$

in cui abbiamo sfruttato il fatto che i termini sono tutti non negativi (grazie al valore assoluto) e possiamo quindi riarrangiare la somma e il fatto che g sia una funzione (quindi per ogni coppia (j,k) ci sia un solo valore $l = g(j,k)$.

Anche nel caso multidimensionale, come in quello unidimensionale, la forma della speranza segue immediatamente. □

Esempio 10.7. Siano X e Y due d20 indipendenti tra loro. Sia $Z = \min(X, Y)$. Qual è il valore atteso della variabile aleatoria Z[4]?

4. Detta anche da alcuni "tiro con svantaggio".

Siamo nelle ipotesi del Teorema 10.6, quindi

$$\begin{aligned} E[Z] &= \sum_{j=1}^{20}\sum_{k=1}^{20} \min(j,k)\, p_{X,Y}(j,k) = \frac{1}{400}\sum_{j=1}^{20}\sum_{k=1}^{20} \min(j,k) \\ &= \frac{1}{400}\sum_{j=1}^{20}\left(\sum_{k=1}^{j} k + \sum_{k=j+1}^{20} j\right) = \frac{1}{400}\sum_{j=1}^{20}\left(\frac{j(j+1)}{2} + (20-j)\, j\right) \\ &= \frac{1}{400}\sum_{j=1}^{20}\left(-\frac{j^2}{2} + \frac{41}{2} j\right) = \frac{1}{800}\sum_{j=1}^{20}(j(41-j)) \\ &= \frac{5740}{800} = \frac{287}{40} = 7.175, \end{aligned}$$

mentre un normale d20 ha media 10.5.

PROPOSIZIONE 10.8. *Il valore atteso (per variabili aleatorie discrete) gode delle seguenti proprietà.*

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 10.6, con $g(x,y) = ax+y+b$,

$$\begin{aligned} E[aX+Y+b] &= E[g(X,Y)] = \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} g(j,k)\, p_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} (aj+k+b)\, p_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X} aj \sum_{k\in\mathcal{R}_Y} p_{X,Y}(j,k) + \sum_{k\in\mathcal{R}_Y} k \sum_{j\in\mathcal{R}_X} p_{X,Y}(j,k) + b\sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} p_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X} aj\, p_X(j) + \sum_{k\in\mathcal{R}_Y} k\, p_Y(k) + b \\ &= aE[X] + E[Y] + b, \end{aligned}$$

in cui, nella penultima uguaglianza, abbiamo marginalizzato la densità discreta congiunta.

Prodotto. Anche in questo caso il nostro riferimento è il Teorema 10.6, con $g(x,y) = x\cdot y$,

$$\begin{aligned} E[XY] &= E[g(X,Y)] = \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} g(j,k)\, p_{X,Y}(j,k) \\ &= \sum_{j\in\mathcal{R}_X}\sum_{k\in\mathcal{R}_Y} jk\, p_X(j)\, p_Y(k) \\ &= \sum_{j\in\mathcal{R}_X} j\, p_X(j) \sum_{k\in\mathcal{R}_Y} k\, p_Y(k) = E[X]\,E[Y], \end{aligned}$$

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,

$$E[X] = \sum_{k\in\mathcal{R}_X} k\, p_X(k) \geqslant 0$$

perché $p_X \geqslant 0$ 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 10.9. *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 10.10. 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

$$Y = -1 + 2 \cdot \mathrm{bin}\left(1, \frac{1}{2}\right)$$

(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.

Osservazione 10.11. Se una variabile aleatoria discreta X è in $\mathcal{L}^1$, ossia se ha speranza finita, anche la sua trasformazione nonlineare $|X|$ ha speranza finita. Infatti la condizione di appartenenza a $\mathcal{L}^1$ di $|X|$ coincide con la condizione di appartenenza a $\mathcal{L}^1$ di X. Possiamo quindi dire che X ha speranza finita se e solo se $|X|$ ha speranza finita.

PROPOSIZIONE 10.12. *Se X è una variabile aleatoria discreta a speranza finita, allora $|E[X]| \leqslant E[|X|]$.*

Dimostrazione. Il risultato segue dal Teorema 10.5, con $g(x) = |x|$, dopo aver invocato la disuguaglianza triangolare:

$$|E[X]| = \left| \sum_{x \in \mathcal{R}_X} x\, p_X(x) \right| \leqslant \sum_{x \in \mathcal{R}_X} |x\, p_X(x)| = \sum_{x \in \mathcal{R}_X} |x|\, p_X(x) = E[|X|]$$

sfruttando il fatto che $p_X \geqslant 0$. □

Osservazione 10.13. Prima di enunciare il prossimo risultato, richiamiamo il concetto di *convessità* di una funzione reale: $f \colon \mathbb{R} \to \mathbb{R}$ è *convessa* se per ogni $x, y \in \mathbb{R}$ e ogni $\alpha \in [0,1]$ abbiamo

$$\alpha f(x) + (1-\alpha) f(y) \geqslant f(\alpha x + (1-\alpha) y)$$

in cui nel primo membro stiamo interpolando in ordinata e nel secondo in ascissa.

TEOREMA 10.14. (DISUGUAGLIANZA DI JENSEN[5]) *Se X è una variabile aleatoria discreta e f è una funzione convessa tali che $E[X]$ e $E[f(X)]$ esistono finite o uguali a $+\infty$, allora $E[f(X)] \geqslant f(E[X])$.*

Dimostrazione. Consideriamo separatamente i vari casi.

***X* e $f(X) \in \mathcal{L}^1$.** Chiamiamo $\mu := E[X]$. Sia inoltre $y = \alpha x + \beta$ una[6] delle rette tangenti al grafico di f passante per il punto $(\mu, f(\mu))$. Per convessità di f, per ogni $x \in \mathbb{R}$ abbiamo $f(x) \geqslant \alpha x + \beta$, quindi vale la disuguaglianza $f(X) \geqslant \alpha X + \beta$ per ogni $\omega \in \Omega$, con X variabile aleatoria. Allora

$$E[f(X)] \geqslant \alpha E[X] + \beta = \alpha \mu + \beta = f(\mu) = f(E[X])$$

sfruttando la linearità della speranza.

5. Johan Jensen (1859 – 1925).

6. Come mai *una* e non *la*?

X o $f(X) \notin \mathcal{L}^1$. Se è $f(X)$ ad essere non negativa con speranza $+\infty$ non abbiamo nulla da dimostrare. Se invece $X \geqslant 0$ e $E[X] = +\infty$ ma $E[f(X)] < +\infty$, vogliamo dimostrare che vale la disuguaglianza $f(E[X]) \leqslant E[f(X)]$, ossia $\lim_{x\to+\infty} f(x) \leqslant E[f(X)]$.

La funzione f è convessa, quindi abbiamo due possibilità:

- se $\lim_{x\to+\infty} f(x) < +\infty$ allora f è non crescente
- se $\lim_{x\to+\infty} f(x) = +\infty$ allora $f(x) \geqslant \varepsilon x$ definitivamente, per qualche $\varepsilon > 0$.

Nel primo caso abbiamo, per ogni $x \in \mathbb{R}$, $f(x) \geqslant \lim_{y\to+\infty} f(y)$ e allora anche a livello di variabile aleatoria $f(X) \geqslant \lim_{y\to+\infty} f(y)$ (ci basterebbe quasi certamente) e possiamo concludere usando il Corollario 10.9: $E[f(X)] \geqslant \lim_{y\to+\infty} f(y) = f(E[X])$.

Nel secondo caso possiamo osservare che enumerando gli elementi di $\mathcal{R}_X$, per ogni $N \in \mathbb{N}$

$$+\infty = E[X] = \sum_{x\in\mathcal{R}_X} x\, p_X(x) = \sum_{i=N}^{+\infty} x_i p_X(x_i)$$

perché non ci interessa dove "partiamo" con la somma, eliminando un numero finito di addendi la somma rimane infinita. Allora

$$\begin{aligned} E[f(X)] &= \sum_{x\in\mathcal{R}_X} f(x)\, p_X(x) = \sum_{i=0}^{+\infty} f(x_i)\, p_X(x_i) \geqslant C + \sum_{i=N}^{+\infty} f(x_i)\, p_X(x_i) \\ &\geqslant C + \sum_{i=N}^{+\infty} \varepsilon\, x_i p_X(x_i) = C + \varepsilon \sum_{i=N}^{+\infty} x_i p_X(x_i) = +\infty \end{aligned}$$

e necessariamente $E[f(X)] = +\infty$.

Questo conclude la dimostrazione. □

Osservazione 10.15. Se g è una funzione concava, ossia se $-g$ è una funzione convessa, la disuguaglianza di Jensen diventa $E[g(X)] \leqslant g(E[X])$.

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 Sezione 10.3, queste proprietà valgono anche per la speranza di variabili aleatorie assolutamente continue.

10.2. VALORE ATTESO DI ALCUNE VARIABILI ALEATORIE NOTE

Calcoliamo ora la speranza dei modelli di variabili aleatorie discrete che abbiamo definito nel Capitolo 9.

Bernoulliane Come abbiamo già visto nell'Esempio 10.4, 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 9.24, 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,

$$p_X(k) = \frac{\lambda^k}{k!} \mathrm{e}^{-\lambda}.$$

Per ricavare la speranza di X, usiamo la definizione di valore atteso,

$$\begin{aligned} E[X] &= \sum_{k\in\mathcal{R}_X} k\cdot p_X(k) = \sum_{k=0}^{+\infty} k\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} = \sum_{k=1}^{+\infty} k\frac{\lambda^k}{k!}\mathrm{e}^{-\lambda} \\ &= \sum_{k=0}^{+\infty} \lambda\frac{\lambda^{k-1}}{(k-1)!}\mathrm{e}^{-\lambda} = \lambda\sum_{h=0}^{+\infty}\frac{\lambda^h}{h!}\mathrm{e}^{-\lambda} = \lambda\sum_{h=0}^{+\infty} p_X(h) = \lambda, \end{aligned}$$

in cui abbiamo usato la proprietà delle densità discrete per cui la somma sul supporto è 1.

Ipergeometriche Sia $X\sim\mathrm{hyp}(k,m,n)$. 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

$$p_X(b) = \frac{\binom{m}{b}\binom{n}{k-b}}{\binom{n+m}{k}},$$

per $b\in\{\max\{0,k-n\},\ldots,\min\{k,m\}\}$. 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 $i\in\{2,\ldots,k\}$, 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,

$$p_{Y_i}(x) = \begin{cases} \frac{m}{m+n} & x=1 \\ \frac{n}{m+n} & x=0 \end{cases}$$

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 $X\sim\mathrm{geom}(p)$. La densità discreta è $p_X(k)=p\,(1-p)^k$, ma non avremo bisogno di usarla esplicitamente. Per la speranza abbiamo infatti

$$\begin{aligned} E[X] &= \sum_{k\in\mathcal{R}_X} k p_X(k) = \sum_{k=0}^{+\infty} k P(X=k) \\ &= \sum_{k=1}^{+\infty}\sum_{i=0}^{k-1} P(X=k) = \sum_{i=0}^{+\infty}\sum_{k=i+1}^{+\infty} P(X=k) \\ &= \sum_{i=0}^{+\infty} P(X>i) = \sum_{i=0}^{+\infty}(1-p)^{i+1} \\ &= (1-p)\frac{1}{1-(1-p)} = \frac{1-p}{p}, \end{aligned}$$

in cui abbiamo usato la (9.2) 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 $X \sim \mathrm{NB}(n,p)$, e $Y_i \sim \mathrm{geom}(p)$ 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 10.16. Bisogna prestare attenzione che, a seconda della definizione data di geometrica (ossia a seconda che il suo supporto sia $\mathbb{N}$ o $\mathbb{N}^+$) e binomiale negativa, cambia il valore della speranza. In particolare l'istante di primo successo di uno schema di Bernoulli di parametro p ha valore atteso $\frac{1}{p}$.

10.3. 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 10.17. *Una variabile aleatoria assolutamente continua X di densità f_X ha* speranza finita *(indicato con $X \in \mathcal{L}^1$) se*

$$\int_{\mathbb{R}} |x|\, f_X(x) < +\infty.$$

Se $X \in \mathcal{L}^1$, il valore atteso, la speranza matematica *o* la media *di X è il baricentro della sua distribuzione, ossia*

$$E[X] = \int_{\mathbb{R}} x f_X(x)\,\mathrm{d}x.$$

Anche nel caso assolutamente continuo, come già nel caso discreto, possiamo considerare il caso in cui la speranza sia $\pm\infty$, ma anche quello in cui non è definita (se l'integrale non esiste).

Valgono nel caso di variabili aleatorie assolutamente continue, risultati analoghi ai Teoremi 10.5 e 10.6, che ci permettono di calcolare la speranza della trasformazione di una variabile aleatoria o di una funzione di un vettore aleatorio.

TEOREMA 10.18. *Siano X una variabile aleatoria assolutamente continua di densità f_X e $Y = g(X)$ una sua trasformazione. Allora $Y \in \mathcal{L}^1$ se e solo se*

$$\int_{\mathbb{R}} |g(x)|\, f_X(x)\,\mathrm{d}x < +\infty.$$

In questo caso

$$E[Y] = \int_{\mathbb{R}} g(x)\, f_X(x)\,\mathrm{d}x.$$

Dimostrazione. Del tutto analoga a quella vista per il caso discreto. □

TEOREMA 10.19. *Siano (X,Y) una coppia di variabili aleatorie assolutamente continue di densità congiunta $f_{X,Y}$ e $Z = g(X,Y)$ per qualche funzione $g\colon \mathbb{R}^2 \to \mathbb{R}$. Allora $Z \in \mathcal{L}^1$ se e solo se*

$$\iint_{\mathbb{R}^2} |g(x,y)|\, f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y < +\infty$$

e in tal caso

$$E[Z] = \iint_{\mathbb{R}^2} g(x,y)\, f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y.$$

Dimostrazione. Del tutto analoga a quella vista per il caso discreto. □

Possiamo estendere i Teoremi 10.6 e 10.19 al caso di vettori aleatori misti, visti nella Sezione 8.3.

TEOREMA 10.20. *Siano X una variabile aleatoria discreta e Y una variabile aleatoria assolutamente continua e che il vettore (X,Y) abbia densità mista*[7] *$f_{X,Y}$. Sia inoltre $Z = g(X,Y)$, per qualche funzione g: $\mathbb{R}^2 \to \mathbb{R}$. Allora $Z \in \mathcal{L}^1$ se e solo se*

$$\sum_{x \in \mathcal{R}_X} \int_{\mathbb{R}} |g(x,y)|\, f_{X,Y}(x,y)\, \mathrm{d}y < +\infty$$

e in tal caso

$$E[Z] = \sum_{x \in \mathcal{R}_X} \int_{\mathbb{R}} g(x,y)\, f_{X,Y}(x,y)\, \mathrm{d}y.$$

Dimostrazione. Mette insieme il caso discreto e quello continuo. □

Osservazione 10.21. 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.

Per dimostrare quindi che $E[V] = (E[X], E[Y])$, usiamo (supponendo che sia X sia Y siano assolutamente continue) il Teorema 10.19: 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 10.8 per la speranza di variabili aleatorie discrete. Possiamo quindi enunciare il seguente risultato più generale.

PROPOSIZIONE 10.22. *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. Le idee sono le stesse del caso discreto. □

Esempio 10.23. Sia X una variabile aleatoria di densità $f_X(x) = \mathrm{e}^{-x}$ 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:

$$E[X] = \int_0^{+\infty} x\,\mathrm{e}^{-x}\,\mathrm{d}x = -[x\,\mathrm{e}^{-x}]_0^{+\infty} + \int_0^{+\infty} \mathrm{e}^{-x}\,\mathrm{d}x = [-\mathrm{e}^{-x}]_0^{+\infty} = 1.$$

7. Abbiamo parlato della densità congiunta mista nella Sezione 8.3.

Passiamo ora a $E[X^{1/2}]$ e usiamo il Teorema 10.18, con $g(t) = \sqrt{t}$,

$$\begin{aligned} E[X^{1/2}] &= \int_0^{+\infty} \sqrt{x}\, \mathrm{e}^{-x}\, \mathrm{d}x \\ &= \int_0^{+\infty} \frac{\xi}{\sqrt{2}} \mathrm{e}^{-\xi^2/2} \xi\, \mathrm{d}\xi = \frac{1}{\sqrt{2}} \int_0^{+\infty} \xi^2 \mathrm{e}^{-\xi^2/2} \mathrm{d}\xi \\ &= \left[\frac{\xi}{\sqrt{2}} \mathrm{e}^{-\frac{\xi^2}{2}} \right]_0^{+\infty} + \frac{1}{\sqrt{2}} \int_0^{+\infty} \mathrm{e}^{-\frac{\xi^2}{2}} \mathrm{d}\xi = 0 + \frac{\sqrt{\pi}}{2}, \end{aligned}$$

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 $\int_0^{+\infty} \mathrm{e}^{-x^2/2} \mathrm{d}x$ è particolarmente importante in probabilità, come vedremo più avanti. Qualche informazione in più su come calcolarlo è in Appendice A.4.

10.4. DISUGUAGLIANZA DI MARKOV

Nell'introduzione di questo capitolo, abbiamo detto che vogliamo usare la media 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.

PROPOSIZIONE 10.24. (DISUGUAGLIANZA DI MARKOV[8]) *Sia X una variabile aleatoria non negativa di media finita. Allora, per ogni $a > 0$*

$$P(X \geqslant a) \leqslant \frac{E[X]}{a}. \tag{10.1}$$

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 10.25. 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.

Osservazione 10.26. Esistono altre varianti della disuguaglianza di Markov (10.1), 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. Una di queste è proposta come Problema 43 in fondo a questo capitolo.

10.5. SPERANZA CONDIZIONATA

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 E, ossia $P_E(\cdot) = P(\cdot \mid E)$ e dunque la densità (discreta) della variabile aleatoria di interesse condizionata a tale evento.

Per prima cosa formalizziamo un po' meglio una proprietà che abbiamo già usato sopra (nella dimostrazione della disuguaglianza di Markov (Proposizione 10.24).

8. 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.

PROPOSIZIONE 10.27. *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].

COROLLARIO 10.28. *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 10.29. Risultati analoghi valgono anche per variabili aleatorie assolutamente continue, con l'accortezza di usare le densità al posto delle densità discrete.

Osservazione 10.30. 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

$$E[X|Y] = \begin{cases} E[X|Y=0] & \text{con probabilità } 1-p \\ E[X|Y=1] & \text{con probabilità } p. \end{cases}$$

10.6. PROBLEMI

Problema 43. Sia X una variabile aleatoria in $\mathcal{L}^1$ (ossia con speranza finita). Sia inoltre $a>0$. Allora

$$P(X\geqslant a) \leqslant \inf_{t>0} E[\mathrm{e}^{tX}]\,\mathrm{e}^{-ta}$$

Problema 44. (S. ROSS) Il trasporto di 148 alunni di una scuola a un certo campo sportivo avviene mediante 4 autobus, sui quali salgono 40, 33, 25 e 50 alunni. Scegliamo un alunno a caso e indichiamo con X il numero totale di alunni saliti sul suo autobus, lui compreso. Scegliamo poi, indipendentemente, uno dei quattro autisti e indichiamo con Y il numero totale di alunni saliti sul bus del quale è alla guida. Quale sarà maggiore tra $E[X]$ e $E[Y]$? Perché?

Problema 45. A un parco divertimenti, una delle attrazioni è un distributore di palline, che contiene palline numerate da 1 a 6. A ciascuna pallina è associato un premio, di valore variabile. Per giocare è necessario un gettone. Le probabilità che il distributore restituisca una data pallina sono riportate in Tabella 10.1.

9. 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.

	1	2	3	4	5	6
p	0.27	0.13	0.09	0.25	$*$	0.16

Tabella 10.1. Probabilità di uscita delle varie palline

Sia X la variabile aleatoria che rappresenta il numero della pallina ottenuta da un giocatore, $X \in \{1, \ldots, 6\}$.

1. Qual è il valore atteso di X?
2. Due amiche, Alex e Beatrice, sono al parco divertimenti. Vorrebbero entrambe giocare al gioco descritto sopra, ma è la fine della serata e hanno solamente un gettone in due. Dal momento che il gettone è perfettamente bilanciato e ha le due facce decorate in modo diverso, lo lanciano per decidere chi delle due giocherà al gioco. Se Alex sceglie la faccia decorata con un fiore, qual è la probabilità che Beatrice vinca il premio associato alla pallina numero 5?

CAPITOLO 11

VARIANZA E COVARIANZA

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. Tra i momenti gioca un ruolo molto importante il momento centrato secondo, che prende il nome di varianza.

11.1. VARIANZA DI UNA VARIABILE ALEATORIA

DEFINIZIONE 11.1. *Per ogni $n \in \mathbb{N}^+$, diciamo che una variabile aleatoria ha momento n-simo finito e scriviamo $X \in \mathcal{L}^n$ se $E[|X^n|] < +\infty$ e in tal caso chiamiamo* momento n-simo *di X il numero reale $E[X^n]$.*

Chiamiamo inoltre momento centrato n-simo *di X il numero reale $E[(X - E[X])^n]$.*

Osservazione 11.2. 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

$$E[X - E[X]] = E[X] - E[X] = 0.$$

Osserviamo inoltre che $\mathcal{L}^p$ è uno spazio vettoriale, quindi se $X \in \mathcal{L}^p$, anche $X - E[X] \in \mathcal{L}^p$, dal momento che $E[X]$ è una costante.

DEFINIZIONE 11.3. *Data una variabile aleatoria $X \in \mathcal{L}^1$, 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 anche della varianza: se la media rappresenta il *baricentro* di una distribuzione di probabilità, la varianza ne è il *momento di inerzia*.

PROPOSIZIONE 11.4. *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 11.4 possiamo farlo calcolando la media $E[X]$ di X e il valore atteso della variabile aleatoria X^2, usando il Teorema 10.18 (o il Teorema 10.5, se X è discreta).

Possiamo ora mostrare alcune proprietà della varianza.

PROPOSIZIONE 11.5. *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 11.6. 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 11.5 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 11.7. Nel prossimo risultato, però, affrontiamo almeno una situazione particolare.

PROPOSIZIONE 11.7. *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 11.4,

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

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:

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = 1^2 \cdot \frac{1}{2} + 0^2 \cdot \frac{1}{2} - \left(\frac{1}{2}\right)^2 = \frac{1}{4},$$

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. Serve allora un altro indicatore.

DEFINIZIONE 11.8. *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 11.9. 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 $\sigma_X = \frac{1}{\sqrt{12}} \approx 0.29$.

11.2. VARIANZA DI ALCUNE VARIABILI ALEATORIE NOTE

Calcoliamo ora la varianza (e quindi la deviazione standard) di alcuni[1] modelli di variabili aleatorie discrete che abbiamo definito nel Capitolo 9.

Bernoulliane Sia $X \sim \mathrm{bin}(1,p)$. Allora

$$\begin{aligned}\mathrm{Var}[X] &= E[X^2] - E[X]^2 \\ &= p - p^2 = p\,(1-p).\end{aligned}$$

Binomiali Sia $X \sim \mathrm{bin}(n,p)$. La variabile aleatoria X è la somma di n Bernoulliane indipendenti e identicamente distribuite Y_i di legge $\mathrm{bin}(1,p)$, quindi

$$\mathrm{Var}[X] = \mathrm{Var}\left[\sum_{i=1}^{n} Y_i\right] = \sum_{i=1}^{n} \mathrm{Var}[Y_i] = np\,(1-p),$$

in cui abbiamo usato la Proposizione 11.7 e la forma della varianza per le Bernoulliane.

Geometriche Sia ora $X \sim \mathrm{geom}(p)$. 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,

$$p_Y(k) = P(Y = k) = 1 \cdot (1-p)^k \cdot p,$$

1. Non ricaviamo qui la varianza delle ipergeometriche, rimandandola all'Esempio 11.40 perché per farlo abbiamo bisogno della covarianza, che introdurremo solamente nella Sezione 11.7.

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

$$\begin{aligned}
E[X] &= \sum_{k=0}^{+\infty} k P(X=k) \\
&= \sum_{k=0}^{+\infty} (k P(X=k \mid \omega_1=0) P(\omega_1=0) + k P(X=k \mid \omega_1=1) P(\omega_1=1)) \\
&= \sum_{k=0}^{+\infty} k P(X=k \mid \omega_1=0) (1-p) + \sum_{k=0}^{+\infty} k P(X=k \mid \omega_1=1) p \\
&= E[X \mid \omega_1=0] (1-p) + E[X \mid \omega_1=1] p \\
&= \sum_{k=0}^{+\infty} k P(Y+1=k \mid \omega_1=0) (1-p) + 0 \cdot P(X=0 \mid \omega_1=1) p \\
&= E[Y+1 \mid \omega_1=0] (1-p) \\
&= E[Y+1] (1-p) \\
&= E[Y] (1-p) + 1 - p,
\end{aligned}$$

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 $X \sim \text{NB}(n,p)$ è la somma di n variabili aleatorie geometriche indipendenti e identicamente distribuite $Y_i \sim \text{geom}(p)$. Allora

$$\text{Var}[X] = \frac{n(1-p)}{p^2}.$$

Poissoniane Sia ora $X \sim \text{Pois}(\lambda)$. Sappiamo che $E[X]=\lambda$, quindi per ricavare $\text{Var}[X]$ ci basta calcolare il momento secondo $E[X^2]$. Possiamo provare a farlo usando il Teorema 10.5, 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

$$\mathrm{Var}[X] = E[X^2] - E[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

11.3. DISUGUAGLIANZA DI CHEBYCHEV

Avendo introdotto la varianza, possiamo pensare di ottenere stime sulla distribuzione di una variabile aleatoria di cui conosciamo solamente speranza e varianza, migliorando la disuguaglianza di Markov (Proposizione 10.24), in cui veniva usata solamente la speranza.

PROPOSIZIONE 11.10. (DISUGUAGLIANZA DI CHEBYCHEV) *Sia X una variabile aleatoria con varianza finita. Allora, per ogni $a > 0$*

$$P(|X - E[X]| \geqslant a) \leqslant \frac{\mathrm{Var}[X]}{a^2} \tag{11.1}$$

o equivalentemente

$$P(|X - E[X]| \geqslant a \cdot \sqrt{\mathrm{Var}[X]}) \leqslant \frac{1}{a^2}. \tag{11.2}$$

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 (10.1), sfruttando il fatto che la variabile aleatoria $(X - E[X])^2$ è non negativa. La forma equivalente (11.2) segue dalla (11.1) sostituendo ad a il numero reale positivo[2] $a\sqrt{\mathrm{Var}[X]}$ □

Grazie alla disuguaglianza di Chebychev (11.2) 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.

11.4. VARIANZA CONDIZIONATA [∗]

Come abbiamo introdotto la speranza condizionata nella Sezione 10.5, possiamo definire anche la varianza condizionata: sia F un evento, allora

$$\mathrm{Var}[X|F] = E[(X - E[X|F])^2|F] = \begin{cases} \sum_x (x - E[X|F])^2 p_{X|F}(x|F) \\ \int (x - E[X|F])^2 f_{X|F}(x|F)\,\mathrm{d}x. \end{cases}$$

Nel caso particolare in cui F è determinato da una variabile aleatoria Y (che supponiamo discreta)

$$\mathrm{Var}[X|Y=y] = \begin{cases} \sum_x (x - E[X|Y=y])^2 p_{X|Y}(x|y) \\ \int (x - E[X|Y=y])^2 f_{X|Y}(x|y)\,\mathrm{d}x, \end{cases}$$

in cui l'ultima densità congiunta è mista. Anche $\mathrm{Var}[X|Y]$ può essere vista come una variabile aleatoria che eredita la casualità da Y.

PROPOSIZIONE 11.11. *Date due variabili aleatorie X e Y vale la seguente identità*

$$\mathrm{Var}_X[X] = E_Y[\mathrm{Var}_X[X|Y]] + \mathrm{Var}_Y[E_X[X|Y]],$$

2. In realtà c'è il caso in cui $\mathrm{Var}[X] = 0$, ma allora $X = E[X]$ e la disuguaglianza non è molto interessante.

detta scomposizione (o fattorizzazione) della varianza.

Dimostrazione. Riscriviamo il secondo membro dell'uguaglianza

$$\begin{aligned} E_Y[\mathrm{Var}_X[X|Y]] + \mathrm{Var}_Y[E_X[X|Y]] &= E_Y[E_X[X^2|Y] - E_X[X|Y]^2] + E_Y[E_X[X|Y]^2] - (E_Y[E_X[X|Y]])^2 \\ &= E_Y[E_X[X^2|Y]] - E_X[X]^2 \\ &= E_X[X^2] - E_X[X]^2 = \mathrm{Var}_X[X] \end{aligned}$$

in cui abbiamo usato la linearità della speranza e, due volte, il Corollario 10.28. □

11.5. MOMENTI DI ORDINE SUPERIORE [∗]

Proseguendo oltre la varianza, vediamo alcuni indicatori associati ai momenti centrati di ordine superiore a 2.

DEFINIZIONE 11.12. *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 e una coda più lunga a destra, se $\mathrm{sk}[X] < 0$ ha una "scentratura" verso destra e una coda più lunga a sinistra.

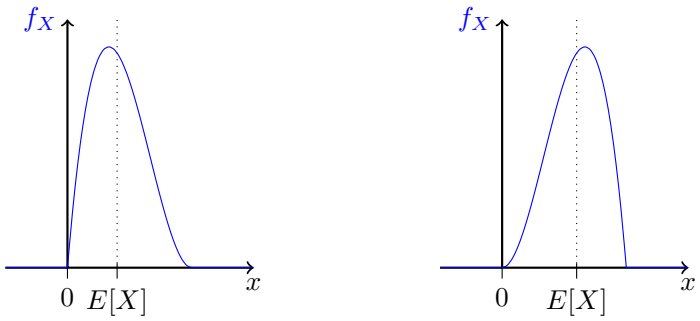


Figura 11.1. Due variabili aleatorie continue. Quella di sinistra ha skewness positiva, quella di destra ha skewness negativa.

DEFINIZIONE 11.13. *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 (Sezione 12.3).

11.6. ALTRI INDICATORI DI UNA DISTRIBUZIONE

Possiamo osservare che gli indicatori di una variabile aleatoria finora introdotti, speranza e varianza e altri momenti, non esauriscono le opzioni disponibili. Vediamo alcuni indicatori particolarmente importanti.

DEFINIZIONE 11.14. *Chiamiamo* mediana *di una variabile aleatoria X un numero m_X tale che*

$$P(X \leqslant m_X) = P(X \geqslant m_X). \tag{11.3}$$

Osservazione 11.15. 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 (11.3) in termini della funzione di ripartizione di X: m_X è tale che $F_X(m_X) = 1 - F_X(m_X) + P(X = m_X)$., Se $P(X = m_X) = 0$ ossia in particolare nel caso in cui X sia continua, $F_X(m_X) = \frac{1}{2}$. 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 in ogni caso 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 11.2.

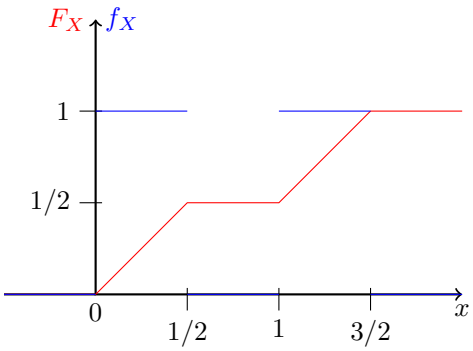


Figura 11.2. Esempio di non unicità della mediana: tutti i punti dell'intervallo $\left[\frac{1}{2}, 1\right]$ sono mediane.

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 11.16. 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[3] $(0,1)$ è una mediana di X.

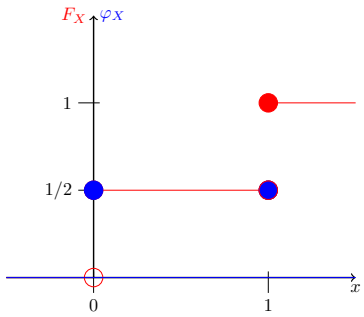


Figura 11.3. Tutti i punti in $(0,1)$ sono mediane.

3. Gli estremi non sono inclusi, come mai?

Esempio 11.17. 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 11.4 $F_X^{-1}\left(\left\{\frac{1}{2}\right\}\right) = \emptyset$.

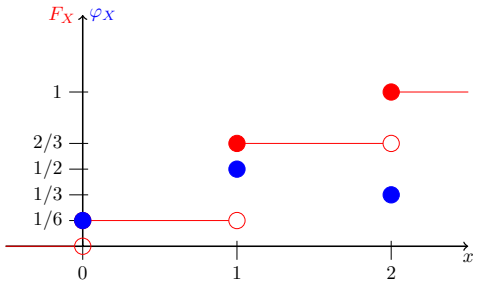


Figura 11.4. La variabile aleatoria X non ammette mediana.

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[4]. 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 11.14.

Dal momento che può essere utile avere un concetto di mediana definito per ogni variabile aleatoria, possiamo darne una definizione indebolita.

DEFINIZIONE 11.18. *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 11.19. 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 (11.3). La variabile aleatoria X nell'Esempio 11.17 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 11.20. *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*

$$Q_X(p) = \inf\{x \in \mathbb{R} : F_X(x) \geqslant p\}. \tag{11.4}$$

4. 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, $1 - F_X$ (più eventualmente $P(X = \cdot)$ nel punto) 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".

Osservazione 11.21. Per $p=\frac{1}{2}$ abbiamo qualcosa di molto simile alla mediana, ma in questo caso viene sempre scelto un solo valore[5]. Non solo, dal momento che la funzione di ripartizione F_X è sempre continua a destra, l'inf nella (11.4) 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 11.22. *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

$F_X(1)=P(X\leqslant 1)\approx 40\%$ 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 11.23. 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 11.24. *Chiamiamo* moda *di una variabile aleatoria X un numero $x\in\mathcal{R}_X$ tale che*

- *se X è discreta, p_X è massima in x, cioè $x\in\operatorname{argmax}_y p_X(y)$*
- *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 11.25. Vediamo alcuni esempi di variabili aleatorie e delle loro mode.

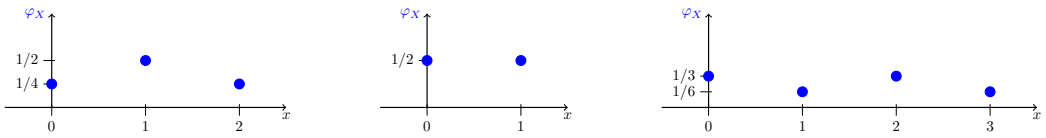


Figura 11.5. Tre variabili discrete. Nella prima la moda è 1, nella seconda sia 0 sia 1 sono mode, nella terza sia 0 sia 2 sono mode.

5. Quale?

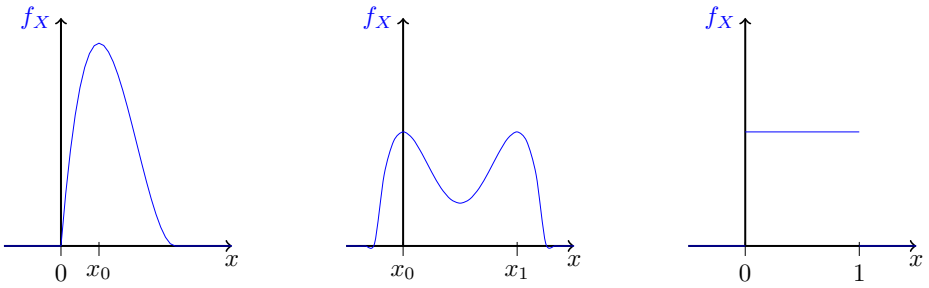


Figura 11.6. Tre variabili assolutamente continue. Nella prima la moda è x_0, nella seconda sono mode x_0 e x_1, nella terza tutti i punti tra 0 e 1 sono mode.

Osservazione 11.26. 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 11.7.

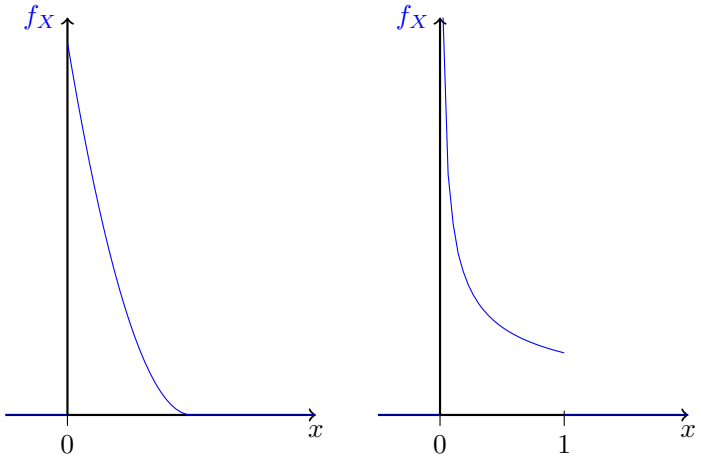


Figura 11.7. Due variabili assolutamente continue. Nella prima la moda è 0, ma la derivata è nulla per $x<0$ e per $x>1$. Nella seconda la moda è in corrispondenza dell'asintoto verticale in 0 (la densità in corrispondenza della moda è infinita).

Questo non è il solo modo in cui le variabili aleatorie assolutamente continue si comportano male riguardo alla moda. Infatti siccome la densità è definita a meno dei punti (o meglio, stiamo considerando classi di equivalenza in cui densità che differiscono su insiemi di misura di Lebesgue nulla sono nella stessa classe) nel punto "di massimo" potremmo prendere un valore diverso della densità senza impattare la variabile aleatoria, ma rendendo la moda non più tale. Possiamo però risolvere la questione riconducendoci meglio al caso discreto, usando il teorema fondamentale del calcolo: per una variabile aleatoria assolutamente continua di densità f si dice moda ogni numero m che appartiene all'insieme

$$\operatorname{argmax}_x \lim_{\varepsilon\to 0} \frac{1}{\varepsilon}\int_x^{x+\varepsilon} f_X(t)\,\mathrm{d}t.$$

Il limite può esistere solo da destra o da sinistra (in altre parole per ε positivo o negativo).

Osservazione 11.27. 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 (11.3) 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 11.28. 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.

11.7. 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 11.29. *Date due variabili aleatorie X e Y, chiamiamo* covarianza *di X e Y la quantità*

$$\mathrm{Cov}[X,Y] = E[(X-E[X])\,(Y-E[Y])]. \tag{11.5}$$

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[6], spesso più pratica da usare

$$\mathrm{Cov}[X,Y] = E[XY] - E[X]\,E[Y]. \tag{11.6}$$

PROPOSIZIONE 11.30. *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 $\mathrm{Cov}[X,Y] = 0$.

6. La dimostrazione di questo fatto è un semplice esercizio, proposto come Problema 47.

Dimostrazione. La prima proprietà segue dalla definizione di covarianza e dalla commutatività del prodotto in $\mathbb{R}$:

$$\mathrm{Cov}[X,Y] = E[(X-E[X])(Y-E[Y])] = E[(Y-E[Y])(X-E[X])] = \mathrm{Cov}[Y,X].$$

Per quanto riguarda la seconda, osserviamo che se X e Y sono indipendenti, allora vale l'identità $E[XY] = E[X]E[Y]$, quindi usando la (11.6) abbiamo

$$\mathrm{Cov}[X,Y] = E[XY] - E[X]E[Y] = 0. \qquad \square$$

DEFINIZIONE 11.31. *Se due variabili aleatorie X e Y hanno covarianza nulla* (*cioè* $\mathrm{Cov}[X,Y]=0$), *diciamo che sono* scorrelate.

Osservazione 11.32. È importante ricordare che non vale il viceversa della seconda proprietà nella Proposizione 11.30: non è necessariamente vero che due variabili aleatorie scorrelate siano indipendenti, anche se due variabili indipendenti sono anche scorrelate.

Esempio 11.33. Siano X e Y due variabili aleatorie di legge congiunta

$$p_{X,Y}(x,y) = \begin{cases} \frac{1}{4} & (x,y) \in \{(-1,-1),(1,-1)\} \\ \frac{1}{2} & (x,y) = (0,1) \\ 0 & \text{altrimenti.} \end{cases}$$

Possiamo ricavarci le leggi marginali di X e Y:

$$p_X(x) = \begin{cases} \frac{1}{4} & x \in \{-1,1\} \\ \frac{1}{2} & x = 0 \\ 0 & \text{altrimenti} \end{cases} \qquad p_Y(y) = \begin{cases} \frac{1}{2} & y \in \{-1,1\} \\ 0 & \text{altrimenti.} \end{cases}$$

A questo punto possiamo facilmente verificare che X e Y non sono indipendenti, infatti

$$p_{X,Y}(x,y) \neq p_X(x)\cdot p_Y(y) = \begin{cases} \frac{1}{8} & (x,y) \in \{(-1,-1),(-1,1),(1,-1),(1,1)\} \\ \frac{1}{4} & (x,y) \in \{(0,-1),(0,1)\} \\ 0 & \text{altrimenti.} \end{cases}$$

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 (11.6), possiamo usare i teoremi noti sulla speranza della funzione di un vettore aleatorio (Teoremi 10.6, 10.19 e 10.20), 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

$$\mathrm{Cov}[X,Y] = \sum_{x\in\mathcal{R}_X}\sum_{y\in\mathcal{R}_Y} xy\,p_{X,Y}(x,y) - \sum_{x\in\mathcal{R}_X} x p_X(x) \sum_{y\in\mathcal{R}_Y} y p_Y(y)$$

nel caso siano entrambe discrete o ancora

$$\mathrm{Cov}[X,Y] = \sum_{x\in\mathcal{R}_X} \int_{\mathbb{R}} xy f_{X,Y}(x,y)\,\mathrm{d}y - \sum_{x\in\mathcal{R}_X} x p_X(x) \int_{\mathbb{R}} y f_Y(y)\,\mathrm{d}y$$

nel caso X sia discreta e Y sia assolutamente continua.

PROPOSIZIONE 11.34. *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. Procediamo in ordine.

i. Iniziamo sfruttando la Proposizione 11.4,

$$\begin{aligned}
\mathrm{Var}[X+Y] &= E[(X+Y)^2] - (E[X+Y])^2 \\
&= E[X^2] + 2E[XY] + E[Y^2] - E[X]^2 - 2E[X]E[Y] - E[Y]^2 \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,(E[XY] - E[X]E[Y]) \\
&= \mathrm{Var}[X] + \mathrm{Var}[Y] + 2\,\mathrm{Cov}[X,Y].
\end{aligned}$$

ii. Usiamo più volte la definizione e le proprietà di linearità della speranza:

$$\begin{aligned}
\mathrm{Cov}[aX+bY,Z] &= E[(aX+bY-aE[X]-bE[Y])\,(Z-E[Z])] \\
&= E[(aX-aE[X])\,(Z-E[Z]) + (bY-bE[Y])\,(Z-E[Z])] \\
&= E[a\,(X-E[X])\,(Z-E[Z]) + b\,(Y-E[Y])\,(Z-E[Z])] \\
&= a\,\mathrm{Cov}[X,Z] + b\,\mathrm{Cov}[Y,Z].
\end{aligned}$$

iii. Usiamo più volte la definizione di covarianza e la linearità sui coefficienti. □

Osservazione 11.35. La Proposizione 11.34 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

$$\mathrm{Var}\left[\sum_{i=1}^n a_i X_i\right] = \mathrm{Cov}\left[\sum_{i=1}^n a_i X_i, \sum_{i=1}^n a_i X_i\right] = \sum_{i=1}^n \sum_{j=1}^n a_i a_j \mathrm{Cov}[X_i, X_j]. \tag{11.7}$$

DEFINIZIONE 11.36. *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 (11.7) 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 11.37. *Valgono le seguenti disuguaglianze,*

$$-\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]} \leqslant \mathrm{Cov}[X,Y] \leqslant \sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}. \tag{11.8}$$

Dimostrazione. Possiamo riformulare la (11.8) come

$$(\mathrm{Cov}[X,Y])^2 \leqslant \mathrm{Var}[X]\,\mathrm{Var}[Y].$$

Ora per ogni $a \in \mathbb{R}$ abbiamo

$$p(a) := a^2 E[Z^2] - 2a\,E[ZW] + E[W^2] = E[(aZ+W)^2] \geqslant 0.$$

Abbiamo cioè un polinomio (in a) di secondo grado che è sempre non negativo, ossia che ha al più una radice reale, ossia il cui discriminante è non positivo:

$$0 \geqslant \frac{\Delta}{4} = (E[ZW])^2 - E[Z^2]\,E[W^2].$$

Prendiamo come Z e W le variabili X e Y centrate, ossia $Z = X - E[X]$ e $W = Y - E[Y]$. Abbiamo

$$(\mathrm{Cov}[X,Y])^2 = (E[(X-E[X])\,(Y-E[Y])])^2 \leqslant E[(X-E[X])^2]\,E[(Y-E[Y])^2] = \mathrm{Var}[X]\,\mathrm{Var}[Y]. \qquad \square$$

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 11.38. 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 11.39. *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 (11.8), 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 11.40. Avendo definito la covarianza, possiamo ora calcolare la varianza delle ipergeometriche. Sia $X \sim \mathrm{hyp}(k,m,n)$.

Come abbiamo visto nella Sezione 10.2 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 (11.7) 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

$$\mathrm{Cov}[Y_i, Y_j] = \frac{n^2 - n}{(n+m)\,(n+m-1)} - \frac{n^2}{(n+m)^2} = \frac{\cancel{n^3} + n^2 m - \cancel{n^2} - n\,m - \cancel{n^3} - n^2 m + \cancel{n^2}}{(n+m)^2\,(n+m-1)}.$$

Ora possiamo mettere assieme il tutto,

$$\begin{aligned} \mathrm{Var}[X] &= k \cdot \frac{n\,m}{(n+m)^2} - k\,(k-1)\,\frac{n\,m}{(n+m)^2\,(n+m-1)} \\ &= \frac{k\,n\,m}{(n+m)^2} \left(1 - \frac{k-1}{n+m-1}\right). \end{aligned}$$

Un'ultima osservazione: nel caso con reimmissione in un'urna di uguale composizione, la varianza sarebbe $\frac{k\,n\,m}{(n+m)^2}$, 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.

11.8. PROBLEMI

Problema 46. Una squadra di basket prende parte a una stagione di 60 partite. Di queste, 32 sono con squadre di fascia A e 28 con squadre di fascia B. I risultati delle partite sono tutti indipendenti tra loro. Le probabilità di vittoria sono del 50% contro una squadra di fascia A e del 70% contro una squadra di fascia B. Sia X il numero totale di vittorie ottenute durante la stagione e siano inoltre X_A e X_B il numero di vittorie contro squadre di fascia A e B rispettivamente.

1. Qual è la distribuzione di X?
2. Che tipo di variabili aleatorie sono X_A e X_B?
3. Qual è il legame tra X, X_A e X_B?
4. Quanto vale, approssimativamente, la probabilità che ci siano almeno 40 vittorie?

Problema 47. Dimostrare che, date due variabili aleatorie X e Y vale l'identità

$$\mathrm{Cov}[X, Y] = E[X\,Y] - E[X]\,E[Y].$$

CAPITOLO 12

MODELLI ASSOLUTAMENTE CONTINUI

In analogia a quanto fatto nel Capitolo 9 per le variabili aleatorie discrete, vediamo ora alcuni modelli di variabili aleatorie assolutamente continue.

12.1. UNIFORMI

Le abbiamo già incontrate più volte, ma diamone comunque una definizione.

DEFINIZIONE 12.1. *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 nel modo seguente:

$$1=\int_{\mathbb{R}} f_X(x)\,\mathrm{d}x=\int_a^b c\,\mathrm{d}x=c\,(b-a),$$

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 12.2. La variabile aleatoria uniforme su $[1,3]$ è rappresentata in Figura 12.1.

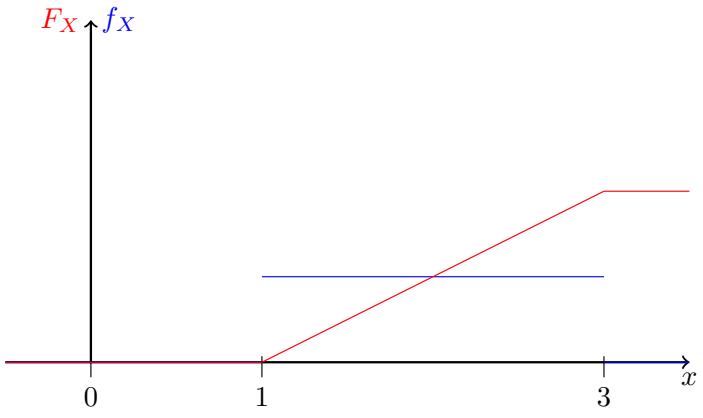


Figura 12.1. Funzione di ripartizione e di densità di $X\sim\text{unif}(1,3)$.

12.1.1. 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)`.

12.1.2. Indicatori per le uniformi

Possiamo calcolare gli indicatori visti nei Capitoli 10 e 11 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, per evidenti ragioni di simmetria.
- 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 12.3. 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
$$P(X > 10 \mid X > 5) = \frac{P(X > 10, X > 5)}{P(X > 5)} = \frac{1 - F_X(10)}{1 - F_X(5)} = \frac{1}{3}\cdot\frac{3}{2} = \frac{1}{2}.$$

Questa variabile aleatoria è ben definita e i risultati ottenuti sopra sono matematicamente solidi, ma non del tutto soddisfacenti dal punto di vista della modellistica. Ci torneremo nell'Esempio 12.6.

Osservazione 12.4. 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 F^{-1}. Non è sempre il modo computazionalmente più efficiente.

12.2. ESPONENZIALI

Anche se non l'abbiamo ancora definita, è una variabile aleatoria che abbiamo incontrato spesso in esempi ed esercizi.

DEFINIZIONE 12.5. *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 nella Sezione 7.1.1, ricaviamo che $c = \lambda$. La funzione di ripartizione di $X \sim \exp(\lambda)$ è

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

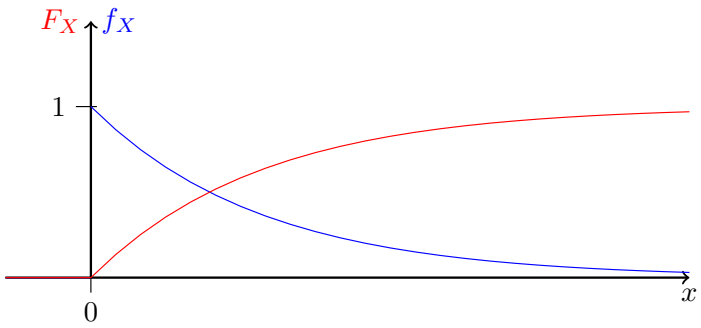


Figura 12.2. Funzione di ripartizione e di densità di $X \sim \exp(1)$.

In Figura 12.2 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 $\lambda < 1$.

12.2.1. 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)`.

12.2.2. 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 12.6. Siamo sempre alla fermata dell'autobus come nell'Esempio 12.3 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

$$P(X>5) = 1 - F_X(5) = \mathrm{e}^{-\frac{2}{15}\cdot 5} \approx 0.51.$$

Per la seconda, invece,

$$P(X>10 \mid X>5) = \frac{1-F_X(10)}{1-F_X(5)} = \mathrm{e}^{-\frac{4}{3}} \cdot \mathrm{e}^{\frac{2}{3}} = \mathrm{e}^{-\frac{2}{3}} = P(X>5) \approx 0.51.$$

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 12.3: qui possiamo anche osservare che la probabilità di aspettare più di mezz'ora (nulla nel caso della distribuzione uniforme) è $P(X>30) = 1 - F_X(30) = \mathrm{e}^{-4} \approx 0.02$. Da un punto di vista modellistico quale delle due variabili aleatorie ci sembra migliore?

La risposta alla seconda domanda nell'Esempio 12.6 ci suggerisce che anche per le esponenziali, così come per le geometriche, valga la proprietà di assenza di memoria.

PROPOSIZIONE 12.7. *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:

$$\begin{aligned} P(X>s+t \mid X>s) &= \frac{P(X>s+t, X>s)}{P(X>s)} \\ &= \frac{1-F_X(s+t)}{1-F_X(s)} \\ &= \mathrm{e}^{-\lambda(s+t)} \cdot \mathrm{e}^{\lambda s} \\ &= \mathrm{e}^{-\lambda t} = 1 - F_X(t) = P(X>t) \end{aligned}$$

semplicemente usando le proprietà dell'esponenziale. □

Osservazione 12.8. 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.

12.3. 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 12.9. *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 12.10. 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.4).

Inoltre possiamo osservare alcune proprietà della densità di una Gaussiana standard, rappresentata in Figura 12.3:

- 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 $\left(\sqrt{e^9\,2\pi}\right)^{-1} \approx 0.004$.

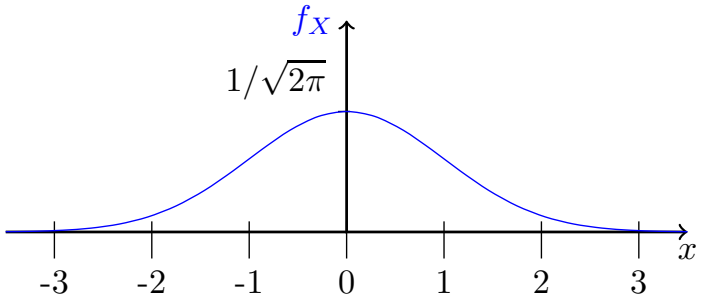


Figura 12.3. Densità di una normale standard

La funzione di ripartizione di una variabile aleatoria normale standard X è

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,dt =: \Phi(x).$$

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).

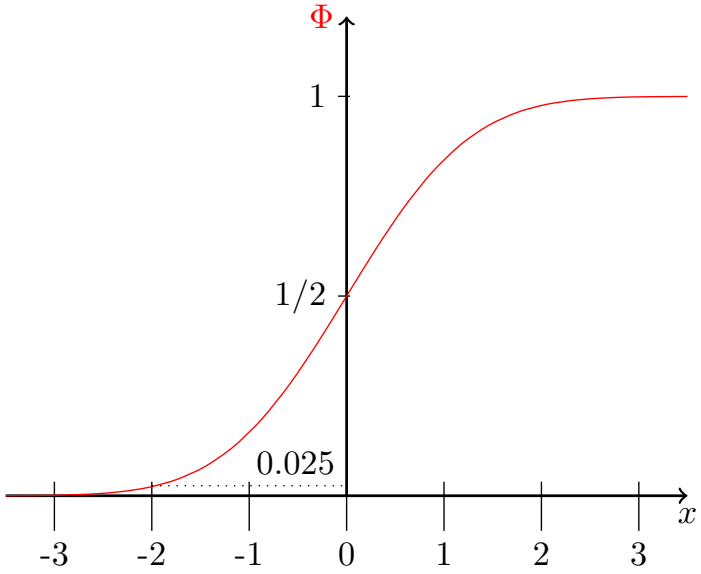


Figura 12.4. Funzione di ripartizione di una normale standard

Osservazione 12.11. Anche Φ, come già la densità, ha alcune proprietà interessanti, che possiamo anche vedere nella Figura 12.4

- 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 variabile aleatoria 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 $\Phi^{-1}(p)=-\Phi^{-1}(1-p)$.

12.3.1. 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 12.12. *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 12.13. 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

$$F_X(x)=P(X\leqslant x)=P(\sigma Z+\mu\leqslant x)=P\Big(Z\leqslant\frac{x-\mu}{\sigma}\Big)=F_Z\Big(\frac{x-\mu}{\sigma}\Big)=\Phi\Big(\frac{x-\mu}{\sigma}\Big)$$

così come la sua densità

$$f_X(x)=\frac{1}{\sigma}f_Z\Big(\frac{x-\mu}{\sigma}\Big)=\frac{1}{\sqrt{2\pi\sigma^2}}\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.\tag{12.1}$$

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.

12.3.2. 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 12.14. 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 12.15. 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 (12.1) o nella Figura 12.5.

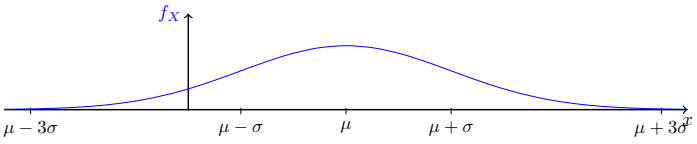


Figura 12.5. Densità di una normale

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 12.6

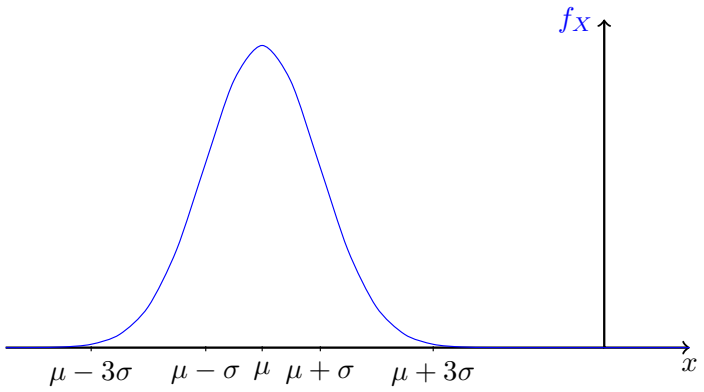


Figura 12.6. Densità di un'altra normale

PROPOSIZIONE 12.16. *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,*

$$X_1 + X_2 \sim \mathcal{N}\Big(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\Big)$$

ossia la loro somma è una Gaussiana di media la somma delle medie e di varianza la somma delle varianze.

Dimostrazione. Si può fare in molti modi, ad esempio scrivendo esplicitamente la convoluzione (ossia la densità di $f_{X_1+X_2}$ in termini di f_{X_1} e f_{X_2}) e riarrangiandola in modo che abbia la forma della densità di una Gaussiana dei parametri cercati. Vedremo però una dimostrazione molto più immediata nel prossimo capitolo. □

Osservazione 12.17. 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:

$$p = P(X \leqslant x) = P\Big(Z \leqslant \frac{x-\mu}{\sigma}\Big) = \Phi\Big(\frac{x-\mu}{\sigma}\Big)$$

e ora approfittiamo del fatto che la funzione Φ è invertibile per avere

$$\Phi^{-1}(p) = \Phi^{-1}\Big(\Phi\Big(\frac{x-\mu}{\sigma}\Big)\Big) = \frac{x-\mu}{\sigma}$$

che, scritta esplicitamente in x, ci dà $x = \sigma\Phi^{-1}(p) + \mu$.

12.3.3. 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)`.

12.3.4. 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 12.18. *Diciamo che un vettore aleatorio* $(X_1, \dots, X_n)$ *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

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\boldsymbol{x}\cdot\boldsymbol{x}} = \frac{1}{\sqrt{(2\pi)^n}} e^{-\frac{1}{2}\boldsymbol{x}^t\boldsymbol{x}}$$

in cui abbiamo usato la notazione $\boldsymbol{x} = (x_1, \ldots, x_n)$ per compattezza.

DEFINIZIONE 12.19. *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*[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

$$f_{X_1,X_2}(x_1, x_2) = \frac{1}{\sqrt{(2\pi)^2 \sigma_1^2 \sigma_2^2 (1-\rho^2)}} \mathrm{e}^{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right)}$$

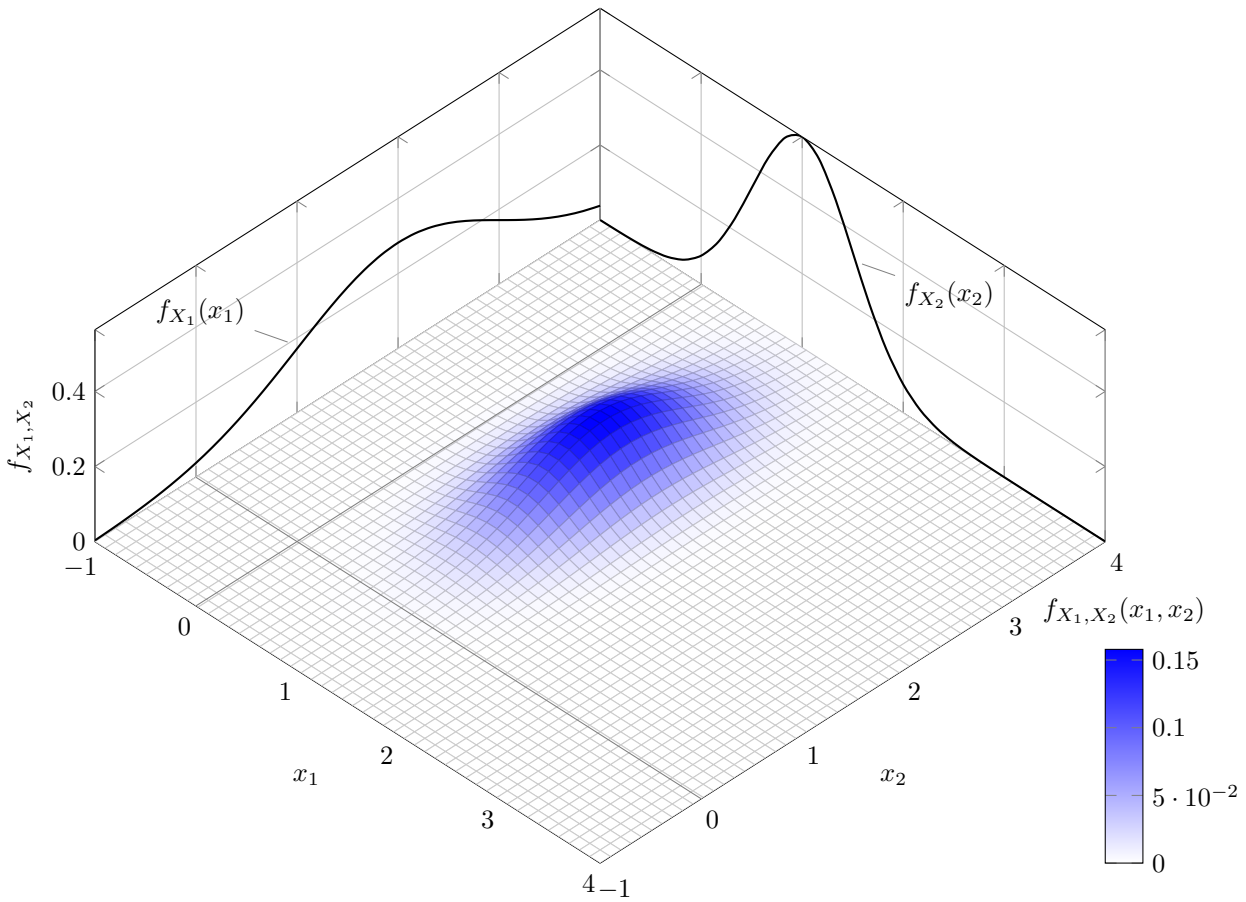


Figura 12.7. Una Gaussiana bivariata ($\mu_1 = 1$, $\sigma_1 = 0.5$, $\mu_2 = 2$, $\sigma_2 = 1$, $\rho = 0$)

1. Vogliamo un vettore colonna, per moltiplicare meglio. In particolare anche X è un vettore aleatorio colonna.

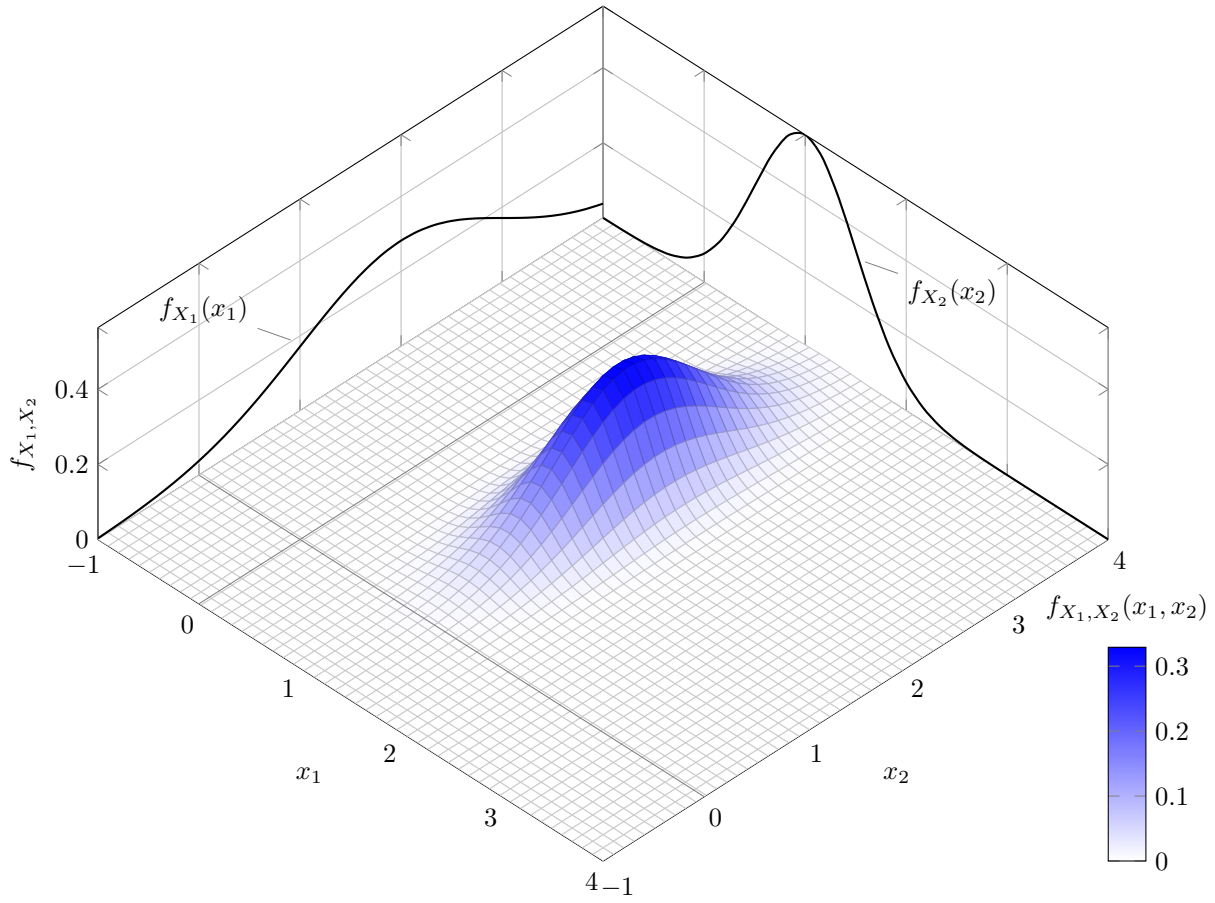


Figura 12.8. Una Gaussiana bivariata ($\mu_1 = 1$, $\sigma_1 = 0.5$, $\mu_2 = 2$, $\sigma_2 = 1$, $\rho = -0.75$)

Osservazione 12.20. Possiamo lasciar cadere l'ipotesi di indipendenza nella Proposizione 12.16. Prese due qualunque Gaussiane $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ che siano congiuntamente Gaussiane, ossia tali che (X_1, X_2) sia una Gaussiana bivariata e 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

$$\begin{aligned}\mathrm{Var}[X_1 + X_2] &= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\,\mathrm{Cov}[X_1, X_2] \\ &= \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\,\mathrm{Corr}[X_1, X_2]\sqrt{\mathrm{Var}[X_1]\,\mathrm{Var}[X_2]}.\end{aligned}$$

12.4. 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 12.21. *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 $X \sim \chi_n^2$.*

Non è semplicissimo ricavare la densità di una chi quadro con n gradi di libertà, tuttavia se $X \sim \chi_n^2$, 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 12.9 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.

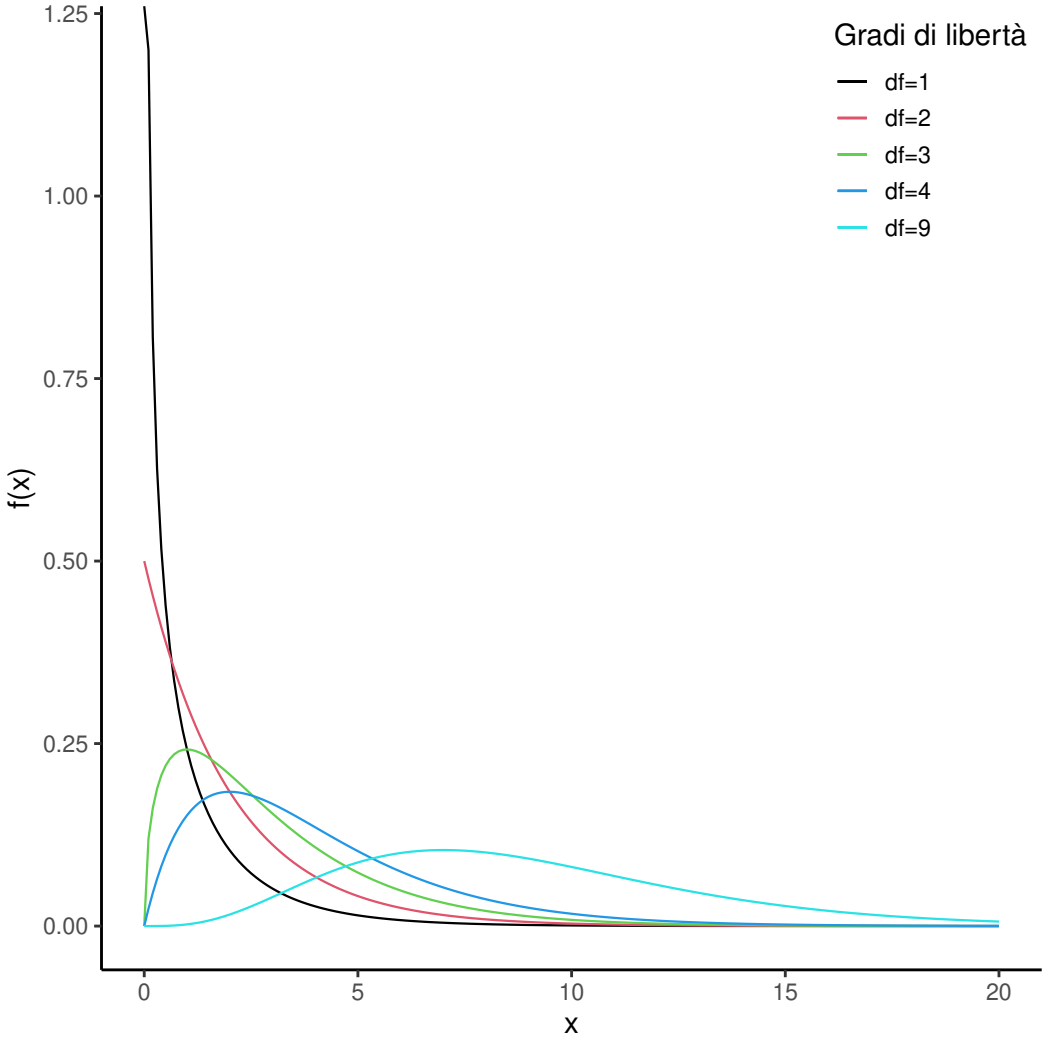


Figura 12.9. Densità di alcune chi quadro, al variare dei gradi di libertà

12.4.1. 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)`.

12.4.2. 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,

$$\mathrm{Var}[Y] = \mathrm{Var}\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n \mathrm{Var}[X_i] = 2n.$$

12.5. t DI STUDENT

DEFINIZIONE 12.22. *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 12.10 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.

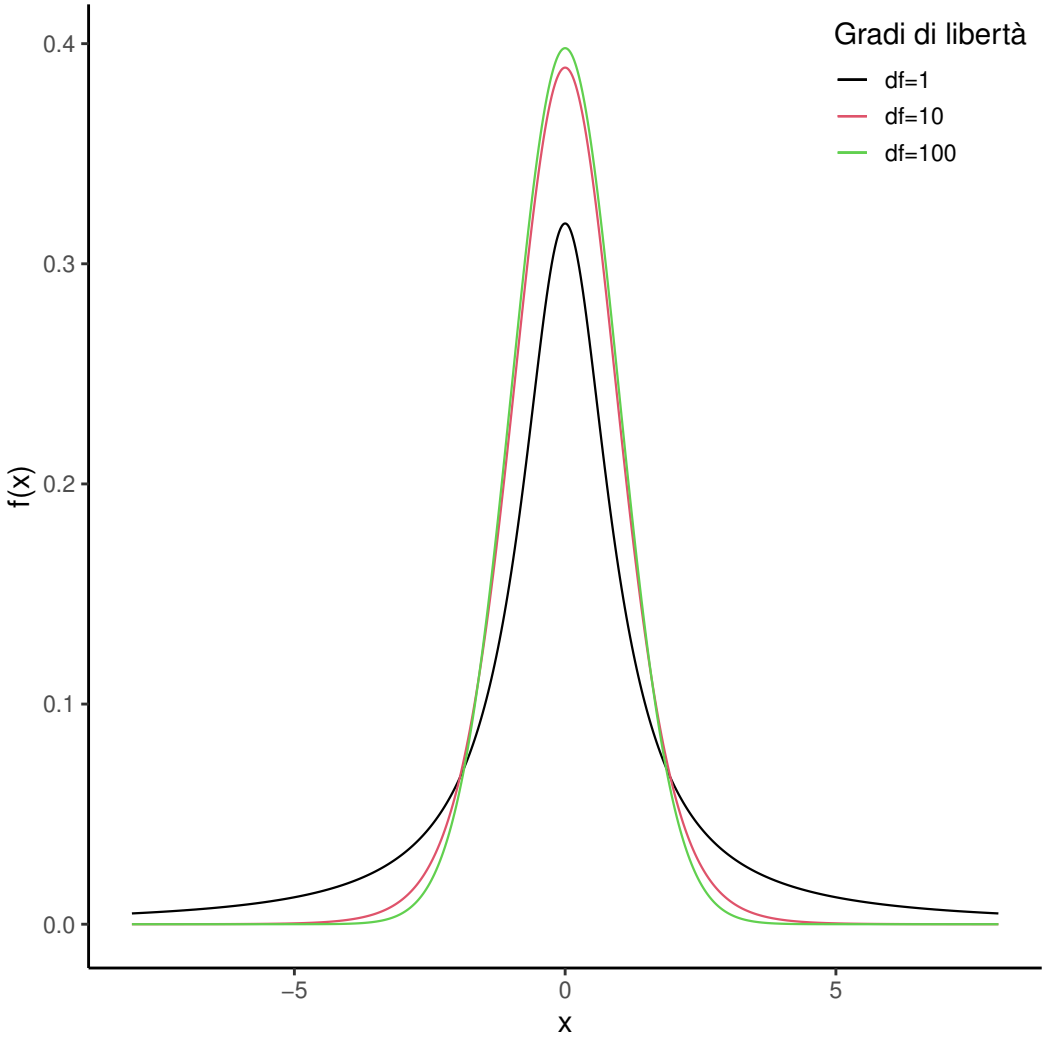


Figura 12.10. Densità di alcune t di Student, al variare dei gradi di libertà

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 $\text{Var}[X] = \frac{n}{n-2}$.

Osservazione 12.23. 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 $\text{Var}\left[\frac{W}{n}\right] \to 0$ 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.

12.5.1. 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)`.

12.6. PROBLEMI

Problema 48. È stato osservato che il tempo trascorso tra il passaggio di due veicoli successivi sotto una videocamera del traffico ha una distribuzione esponenziale di media 4 minuti. Chiamiamo T la corrispondente variabile aleatoria.

1. Qual è la probabilità che il tempo trascorso tra il passaggio di due veicoli successivi sia minore di 1.1 minuti?
2. Qual è l'intervallo di tempo t (in minuti) tale per cui siamo certi al 90% che il tempo trascorso tra il passaggio di due veicoli sia maggiore di t minuti?
3. Sapendo che sono già trascorsi 1.7 minuti dal passaggio dell'ultimo veicolo, qual è la probabilità che si debba attendere al più altri 5 minuti per il passaggio del veicolo successivo?

Problema 49. Siano $X \sim \exp(\lambda)$ e $Y = \alpha X$, (con $\alpha > 0$). Com'è distribuita Y?

Problema 50. Nella Proposizione 12.7 abbiamo mostrato che le variabili aleatorie esponenziali godono della proprietà di assenza di memoria. Dimostrare che non esistono altre variabili aleatorie assolutamente continue con tale proprietà.

Problema 51. (T. M. MILLS) Siano $X \sim \exp(\lambda)$ con $\lambda > 0$ e $Y = F_X(X)$. Com'è distribuita Y?

Problema 52. Una ditta deve asfaltare un tratto di strada lungo 1778.973 m. Dall'esperienza dei lavori precedenti si sa che il tratto di strada che la ditta riesce ad asfaltare in un giorno si distribuisce come una normale di media 50 m e deviazione standard $\sigma = 2.15$ m.

1. Con che probabilità serviranno più di $n = 36$ giorni per terminare i lavori?
2. Quanto dovrebbe essere lungo il tratto di strada affinché 36 giorni siano sufficienti per asfaltarla tutta con una probabilità maggiore o uguale al 62.2%?

Problema 53. È possibile dare una definizione alternativa di Gaussiane multivariate rispetto alla Definizione 12.19. Consideriamo il caso bidimensionale. Due variabili aleatorie X e Y si dicono congiuntamente Gaussiane se ogni combinazione lineare di X e Y è una Gaussiana. Mostrare che le due definizioni sono equivalenti.

Problema 54. Alla luce del Problema 53, dimostrare il claim nell'Osservazione 12.20. Come esercizio bonus discutere l'importanza che X e Y siano congiuntamente Gaussiane (ossia trovare un controesempio nel caso in cui X e Y siano entrambe Gaussiane ma non lo siano congiuntamente).

CAPITOLO 13

FUNZIONE GENERATRICE DEI MOMENTI

I momenti, che abbiamo introdotto in precedenza nei Capitoli 10 e 11, sono in realtà uno strumento molto potente per caratterizzare le distribuzioni di probabilità, in particolare quando conosciamo tutti i momenti di una variabile aleatoria. A questo scopo introduciamo la funzione generatrice dei momenti, una funzione che, come vedremo, ha alcune proprietà particolarmente interessanti e utili.

Prima di continuare con il capitolo, però, facciamo un breve commento. La funzione generatrice dei momenti in generale è meno studiata di un oggetto simile, ma per certi aspetti radicalmente diverso, ossia la *funzione caratteristica* di una variabile aleatoria, definita da $\varphi_X(t) := E[e^{itX}]$. Quest'ultima è più flessibile e, per certi aspetti, "migliore" rispetto alla funzione generatrice dei momenti. Tuttavia richiede (in realtà è) la trasformata di Fourier e necessita dell'analisi complessa, quindi le viene preferita in questo primo corso la funzione generatrice dei momenti. Non solo, dal punto di vista delle implementazioni numeriche la funzione generatrice dei momenti è migliore rispetto alla funzione caratteristica.

13.1. DEFINIZIONE

DEFINIZIONE 13.1. *Sia X una variabile aleatoria. La* funzione generatrice dei momenti (MGF) *di X è definita come $M_X(t) = E[e^{tX}]$.*

Osservazione 13.2. Nel caso in cui X sia una variabile aleatoria discreta o assolutamente continua, possiamo esplicitare la forma della funzione generatrice dei momenti, grazie ai Teoremi 10.5 e 10.18. Abbiamo, nel caso discreto,

$$M_X(t) = \sum_{x \in \mathcal{R}_X} e^{tx} \varphi_X(x)$$

e nel caso assolutamente continuo

$$M_X(t) = \int_{\mathbb{R}} e^{tx} f_X(x)\, dx.$$

Notiamo che queste somme e questi integrali potrebbero non convergere, quindi diciamo che la funzione generatrice dei momenti di una variabile aleatoria X esiste se per ogni $\varepsilon > 0$, $M_X(t)$ è finita per ogni $t \in (-\varepsilon, +\varepsilon)$, ossia in un intorno di 0.

Esempio 13.3. Siano $Y \sim \text{bin}(1, p)$ e $X = 2Y + 3$. Allora X assume i valori 3 e 5 con probabilità, rispettivamente, $1 - p$ e p. La funzione generatrice dei momenti di X è

$$M_X(t) = E[e^{tX}] = \sum_{x \in \{3,5\}} e^{tx} \varphi_X(x) = (1-p)\, e^{3t} + p\, e^{5t}.$$

Esempio 13.4. Consideriamo una variabile aleatoria assolutamente continua, un'uniforme su [0, 1] $Z \sim \text{unif}(0, 1)$. La funzione generatrice dei momenti di Z è

$$M_Z(t) = E[e^{tZ}] = \int_0^1 e^{tz} 1\, dz = \frac{e^t - 1}{t}.$$

13.2. PROPRIETÀ

PROPOSIZIONE 13.5. *Siano X, Y due variabili aleatorie indipendenti, a, b, c numeri reali. Allora*

$$M_{aX+bY+c}(t) = M_X(at)\, M_Y(bt)\, \mathrm{e}^{ct}.$$

Dimostrazione. Abbiamo

$$\begin{aligned} M_{aX+bY+c}(t) &= E[\mathrm{e}^{t(aX+bY+c)}] = E[\mathrm{e}^{taX}\mathrm{e}^{tbY}\mathrm{e}^{tc}] \\ [\text{indipendenza}] &= E[\mathrm{e}^{(at)X}]\, E[\mathrm{e}^{(bt)Y}]\, E[\mathrm{e}^{ct}] \\ &= M_X(at)\, M_Y(bt)\, \mathrm{e}^{ct} \end{aligned}$$

concludendo così la dimostrazione. □

Non abbiamo ancora affrontato una questione abbastanza centrale: perché si chiama *funzione generatrice dei momenti*? In che modo M_X genera i momenti di X? Attraverso le derivate valutate in $t = 0$.

Consideriamo il caso in cui X sia una variabile aleatoria discreta. Allora calcolando la derivata prima di $M_X(t)$ in t abbiamo

$$M'_X(t) = \frac{\mathrm{d}}{\mathrm{d}t} E[\mathrm{e}^{tX}] = \frac{\mathrm{d}}{\mathrm{d}t} \sum_{x \in \mathcal{R}_X} \mathrm{e}^{tx} \varphi_X(x) = \sum_{x \in \mathcal{R}_X} \frac{\mathrm{d}}{\mathrm{d}t} (\mathrm{e}^{tx} \varphi_X(x)) = \sum_{x \in \mathcal{R}_X} x\, \mathrm{e}^{tx} \varphi_X(x)$$

che valutata in $t = 0$ dà $M'_X(0) = \sum_{x \in \mathcal{R}_X} x\, \mathrm{e}^{0x} \varphi_X(x) = \sum_{x \in \mathcal{R}_X} x\, \varphi_X(x) = E[X]$.

Passiamo allora alla derivata seconda,

$$M''_X(t) = \frac{\mathrm{d}}{\mathrm{d}t} M'_X(t) = \sum_{x \in \mathcal{R}_X} \frac{\mathrm{d}}{\mathrm{d}t} (x\, \mathrm{e}^{tx} \varphi_X(x)) = \sum_{x \in \mathcal{R}_X} x^2 \mathrm{e}^{tx} \varphi_X(x)$$

che valutata in $t = 0$ dà $M''_X(0) = \sum_{x \in \mathcal{R}_X} x^2 \varphi_X(x) = E[X^2]$.

PROPOSIZIONE 13.6. *Siano X una variabile aleatoria e $M_X(t)$ la sua funzione generatrice dei momenti. Allora la derivata n-sima di M_X valutata in $t = 0$ è il momento n-simo di X, $M_X^{(n)}(0) = E[X^n]$.*

Dimostrazione. Abbiamo, usando la rappresentazione in serie di potenze della funzione esponenziale e la linearità della speranza

$$M_X^{(n)}(t) = \frac{\mathrm{d}^n}{\mathrm{d}t^n} E[\mathrm{e}^{tX}] = \frac{\mathrm{d}^n}{\mathrm{d}t^n} E\left[\sum_{k=0}^{+\infty} \frac{(tX)^k}{k!}\right] = \sum_{k=0}^{+\infty} \frac{\mathrm{d}^n}{\mathrm{d}t^n} \left(\frac{t^k}{k!} E[X^k]\right) = \sum_{k=0}^{+\infty} \left(\frac{\mathrm{d}^n}{\mathrm{d}t^n} t^k\right) \frac{1}{k!} E[X^k].$$

Ora dobbiamo trattare la derivata della potenza, che è

$$\frac{\mathrm{d}^n}{\mathrm{d}t^n} t^k = \begin{cases} \frac{k!}{(k-n)!} t^{k-n} & n \leqslant k \\ 0 & n > k \end{cases},$$

quindi andando a sostituire

$$M_X^{(n)}(t) = \sum_{k=n}^{+\infty} \frac{k!}{(k-n)!} t^{k-n} \frac{1}{k!} E[X^k] = E[X^n] + \sum_{k=n+1}^{+\infty} \frac{t^{k-n}}{(k-n)!} E[X^k].$$

Valutando questa espressione in $t = 0$ otteniamo la tesi, dal momento che tutti i termini dell'ultima somma si annullano. □

Passiamo ora al risultato che rende così utile e importante la funzione generatrice dei momenti: caratterizza in modo univoco una distribuzione. Se due variabili aleatorie hanno la medesima distribuzione allora hanno necessariamente la stessa funzione generatrice dei momenti. Ora vediamo che vale anche il viceversa.

TEOREMA 13.7. *Siano X e Y due variabili aleatorie e M_X e M_Y le loro rispettive funzioni generatrici dei momenti. Se esiste $\varepsilon > 0$ tale che $M_X(t) = M_Y(t)$ su $(-\varepsilon, +\varepsilon)$, allora $X \sim Y$.*

Dimostrazione. Diamo la dimostrazione solamente in un caso particolare, ossia quello delle variabili aleatorie discrete a supporto finito $\{0,\dots,n\}$. In questa situazione dobbiamo mostrare che sul supporto $\{0,\dots,n\}$ vale $\varphi_X = \varphi_Y$.

Per ipotesi abbiamo $M_X(t) = M_Y(t)$ in un intorno di 0. Per la definizione di funzione generatrice dei momenti nel caso discreto, questo equivale a scrivere

$$\sum_{x\in\{0,\dots,n\}} \mathrm{e}^{tx}\varphi_X(x) = \sum_{x\in\{0,\dots,n\}} \mathrm{e}^{tx}\varphi_Y(x)$$

che riarrangiando i termini è

$$\sum_{x\in\{0,\dots,n\}} (\varphi_X(x) - \varphi_Y(x))\,\mathrm{e}^{tx} = 0.$$

Possiamo interpretare questa scrittura come un polinomio di grado n in e^t, avente per ogni grado k coefficiente $\varphi_X(k) - \varphi_Y(k)$. Questa quantità è nulla per tutti i valori di t in $(-\varepsilon, +\varepsilon)$, una quantità più che numerabile, ma allo stesso tempo, vista come polinomio non può avere che n radici, a meno che i coefficienti non siano tutti nulli. Quindi $\varphi_X(x) = \varphi_Y(x)$ per ogni $x \in \{0,...,n\}$. □

Osservazione 13.8. La dimostrazione proposta si adatta facilmente a qualunque variabile aleatoria discreta a supporto finito. Se vogliamo la dimostrazione nel caso più generale abbiamo bisogno di strumenti che esulano da questo corso: la dimostrazione più usata sfrutta le proprietà delle funzioni olomorfe, dopo aver esteso la definizione a $\mathbb{C}$. Il fatto è che il lavoro necessario per tale dimostrazione è comparabile con quello per la trasformata di Fourier (o funzione caratteristica, nel caso delle variabili aleatorie). In effetti la funzione generatrice dei momenti di X è la trasformata di Laplace di $-X$.[LUIGI: Due approcci alla dimostrazione con referenze bibliografiche. Nel primo (vedi Billingsley, Probability and Measure, Sec 30) si mostra che la finitezza di M_X in un intorno di 0 implica che i momenti di X non crescono troppo velocemente, quindi F_X è determinata da $(E[X^k])_{k\in\mathbb{N}}$ che a sua volta e determinata da M_X. Il secondo (vedi per esempio Curtiss, Annals of Mathematical Statistics 13 430:433, ma anche esercizio 26.7 del Billingsley) è mostrare che M_X è analitica e può essere estesa alla striscia $(-\varepsilon,\varepsilon)\times \mathrm{i}\,\mathbb{R} \subseteq \mathbb{C}$ in modo che $M_X(z) = E[\mathrm{e}^{zX}]$ in modo che in particolare $M_X(\mathrm{i}\,t) = \varphi_X(t)$ per ogni $t \in \mathbb{R}$, con φ la funzione caratteristica, ossia la trasformata di Fourier della variabile aleatoria. A questo punto si conclude usando il fatto che φ determina F.]

Osservazione 13.9. La vera potenza del Teorema 13.7 è nel permetterci di caratterizzare una distribuzione a partire dalla sua funzione generatrice dei momenti. Vedremo nella prossima sezione alcuni risultati già ottenuti e altri solo enunciati con le loro dimostrazioni mediante le funzioni generatrici dei momenti.

13.3. MGF IN AZIONE

Esempio 13.10. Calcoliamo la funzione generatrice dei momenti per una variabile aleatoria di Poisson di parametro λ. Allora

$$M_X(t) = \sum_{n=0}^{+\infty} \mathrm{e}^{tn}\varphi_X(k) = \sum_{n=0}^{+\infty} \mathrm{e}^{tn}\frac{\lambda^n}{n!}\mathrm{e}^{-\lambda} = \mathrm{e}^{-\lambda}\sum_{n=0}^{+\infty}\frac{(\lambda\,\mathrm{e}^t)^n}{n!} = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}.$$

Possiamo allora ricavare il momento primo calcolando la derivata prima di M_X in 0

$$M_X'(0) = \frac{\mathrm{d}}{\mathrm{d}t}\,\mathrm{e}^{\lambda(\mathrm{e}^t-1)}\bigg|_{t=0} = \lambda\,\mathrm{e}^t\,\mathrm{e}^{\lambda(\mathrm{e}^t-1)}|_{t=0} = \lambda$$

e poi il momento secondo come derivata seconda (per poi sottrarre il momento primo al quadrato e avere la varianza)

$$M_X''(0) = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\,\mathrm{e}^{\lambda(\mathrm{e}^t-1)}\bigg|_{t=0} = \lambda\,\mathrm{e}^{\lambda(\mathrm{e}^t-1)+t}\,(\lambda\,\mathrm{e}^t+1)|_{t=0} = \lambda^2+\lambda$$

da cui $\mathrm{Var}[X] = \lambda^2 + \lambda - \lambda^2 = \lambda$.

Osservazione 13.11. Avendo la forma della funzione generatrice dei momenti di una variabile aleatoria di Poisson, possiamo dimostrare in modo alternativo la riproducibilità di tali variabili aleatorie. Se $X \sim \mathrm{Pois}(\lambda)$ e $Y \sim \mathrm{Pois}(\mu)$ indipendenti, consideriamo $X+Y$. Abbiamo

$$M_{X+Y}(t) = M_X(t)\, M_Y(t) = \mathrm{e}^{\lambda(\mathrm{e}^t-1)}\, \mathrm{e}^{\mu(\mathrm{e}^t-1)} = \mathrm{e}^{(\lambda+\mu)(\mathrm{e}^t-1)}$$

che è la funzione generatrice dei momenti di una variabile aleatoria di Poisson di parametro $\lambda+\mu$.

Esempio 13.12. Calcoliamo ora la funzione generatrice dei momenti per una variabile aleatoria Gaussiana di parametri μ e σ, $X \sim \mathcal{N}(\mu, \sigma)$. Abbiamo

$$M_X(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{tx} f_X(x)\, \mathrm{d}x = \int_{-\infty}^{+\infty} \mathrm{e}^{tx} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\, \mathrm{d}x.$$

Per continuare conviene fare un cambio di variabile: $z = \frac{x-\mu}{\sigma}$ (ossia proprio la standardizzazione), da cui $x = \sigma z + \mu$. Allora

$$M_X(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{t(\sigma z+\mu)} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{z^2}{2}} \sigma\, \mathrm{d}z = \mathrm{e}^{t\mu} \int_{-\infty}^{+\infty} \mathrm{e}^{t\sigma z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}}\, \mathrm{d}z$$

e per continuare abbiamo bisogno di raccogliere in modo furbo gli esponenziali dentro l'integrale,

$$\mathrm{e}^{t\sigma z} \mathrm{e}^{-\frac{z^2}{2}} = \mathrm{e}^{-\frac{1}{2}(z-t\sigma)^2 + \frac{1}{2}\sigma^2 t^2} = \mathrm{e}^{-\frac{1}{2}(z-t\sigma)^2} \mathrm{e}^{\frac{1}{2}\sigma^2 t^2}.$$

A questo punto

$$M_X(t) = \mathrm{e}^{t\mu} \int_{-\infty}^{+\infty} \mathrm{e}^{t\sigma z} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{z^2}{2}}\, \mathrm{d}z = \mathrm{e}^{t\mu} \mathrm{e}^{\frac{1}{2}\sigma^2 t^2} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{(z-\sigma t)^2}{2}}\, \mathrm{d}z$$

in cui possiamo osservare che la traslazione nell'ultimo integrale è irrilevante e lascia l'integrale uguale a 1, quindi $M_X(t) = \mathrm{e}^{t\mu} \mathrm{e}^{\frac{1}{2}\sigma^2 t^2}$.

Possiamo ora derivare una volta per ricavare il momento primo di X

$$M_X'(0) = \frac{\mathrm{d}}{\mathrm{d}t} \mathrm{e}^{t\mu} \mathrm{e}^{\frac{1}{2}\sigma^2 t^2} \Big|_{t=0} = (\mu + \sigma^2 t^2)\, \mathrm{e}^{t\mu} \mathrm{e}^{\frac{1}{2}\sigma^2 t^2} \Big|_{t=0} = \mu$$

e poi il momento secondo come derivata seconda (per poi sottrarre il momento primo al quadrato e avere la varianza)

$$M_X''(0) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \mathrm{e}^{t\mu} \mathrm{e}^{\frac{1}{2}\sigma^2 t^2} \Big|_{t=0} = (\mu^2 + \sigma^2 + \sigma^4 t^2 + 2\mu\sigma^2 t)\, \mathrm{e}^{t\mu} \mathrm{e}^{\frac{\sigma^2 t^2}{2}} \Big|_{t=0} = \mu^2 + \sigma^2$$

da cui $\mathrm{Var}[X] = \mu^2 + \sigma^2 - \mu^2 = \sigma^2$.

Anche per le Gaussiane possiamo usare la forma esplicita per dimostrare la riproducibilità.

PROPOSIZIONE 13.13. (PROPOSIZIONE 12.16) *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,*

$$X_1 + X_2 \sim \mathcal{N}\Big(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\Big)$$

ossia la loro somma è una Gaussiana di media la somma delle medie e di varianza la somma delle varianze.

Dimostrazione. Date $X_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ e $X_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ indipendenti, abbiamo per $X_1 + X_2$

$$M_{X_1+X_2}(t) = M_{X_1}(t)\, M_{X_2}(t) = \mathrm{e}^{t\mu_1} \mathrm{e}^{\frac{1}{2}\sigma_1^2 t^2} \mathrm{e}^{t\mu_2} \mathrm{e}^{\frac{1}{2}\sigma_2^2 t^2} = \mathrm{e}^{t(\mu_1+\mu_2)} \mathrm{e}^{\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2}$$

che è la funzione generatrice dei momenti di una variabile aleatoria Gaussiana di media $\mu_1 + \mu_2$ e varianza $\sigma_1^2 + \sigma_2^2$ (cioè deviazione standard $\sqrt{\sigma_1^2 + \sigma_2^2}$). □

Osservazione 13.14. In generale se la funzione generatrice dei momenti ha forma esponenziale, la famiglia sarà riproducibile, dal momento che il prodotto di esponenziali è l'esponenziale della somma.

13.4. VETTORI ALEATORI

Chiaramente possiamo estendere la nozione di funzione generatrice dei momenti anche al caso di vettori aleatori. Lo vediamo nel caso particolare delle Gaussiane multivariate.

Sia $Z = (Z_1, \dots, Z_n)$ un vettore di variabili aleatorie normali standard i.i.d. Allora per $\boldsymbol{t} \in \mathbb{R}^n$,

$$M_Z(\boldsymbol{t}) = E[\mathrm{e}^{\boldsymbol{t}^t Z}] = E[\mathrm{e}^{\sum_{i=1}^n t_i Z_i}] = E\left[\prod_{i=1}^n \mathrm{e}^{t_i Z_i}\right] = \prod_{i=1}^n E[\mathrm{e}^{t_i Z_i}]$$

ossia è il prodotto delle funzioni generatrici dei momenti delle componenti, quindi

$$M_Z(\boldsymbol{t}) = \prod_{i=1}^n \mathrm{e}^{\frac{t_i^2}{2}} = \mathrm{e}^{\sum_{i=1}^n \frac{t_i^2}{2}} = \mathrm{e}^{\frac{1}{2}\boldsymbol{t}^t \boldsymbol{t}}.$$

Se passiamo al caso generale abbiamo, per $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, che la funzione generatrice dei momenti è

$$M_X(\boldsymbol{t}) = M_{\boldsymbol{\Sigma} Z + \boldsymbol{\mu}}(\boldsymbol{t}) = \mathrm{e}^{\boldsymbol{t}^t \boldsymbol{\mu}} \mathrm{e}^{\frac{1}{2}(\boldsymbol{\Sigma}\boldsymbol{t})^t \boldsymbol{\Sigma}\boldsymbol{t}}.$$

Anche in questo caso, quindi, torniamo a una funzione generatrice dei momenti che ha la medesima forma della funzione generatrice dei momenti di una variabile aleatoria gaussiana unidimensionale.

Questo ci permette di dare un'altra caratterizzazione delle Gaussiane multivariate, che sarà particolarmente utile in corsi più avanzati: Z è un n-vettore Gaussiano se per ogni $\boldsymbol{a} \in \mathbb{R}^n$, $\boldsymbol{a}^t Z$ è una Gaussiana unidimensionale.

13.5. PROBLEMI

Problema 55. Calcolare la funzione generatrice dei momenti per una variabile aleatoria binomiale di parametri n e p.

Problema 56. Calcolare la funzione generatrice dei momenti per una variabile aleatoria esponenziale di parametro λ.

Problema 57. Calcolare la funzione generatrice dei momenti per una variabile aleatoria uniforme sull'intervallo $[a, b]$.

CAPITOLO 14

TEOREMI LIMITE

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 definite 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 potremo avvicinare alcuni risultati (i teoremi limite) che ci garantiscono sotto opportune ipotesi, la convergenza di alcune particolari successioni di variabili aleatorie.

14.1. CONVERGENZA DI VARIABILI ALEATORIE

DEFINIZIONE 14.1. *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 $P(\lim_{n\to+\infty} X_n = X) = 1$. Possiamo anche caratterizzare la convergenza quasi certa mediante il limsup di eventi (Definizione 4.30): $X_n \xrightarrow[n\to+\infty]{\text{q.c.}} X$ se per ogni $\varepsilon > 0$,

$$P\Big(\limsup_{n\to+\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\Big) = 0.$$

Osservazione 14.2. Il concetto di convergenza quasi certa è molto forte[1]: 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 14.3. *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 14.4. 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.

Possiamo dare una caratterizzazione alternativa della convergenza in probabilità.

1. Anche se non è la nozione più forte possibile, per le variabili aleatorie: se le vediamo come funzioni reali avremmo anche la convergenza *certa* o *puntuale*, ossia che vale per tutti gli $\omega \in \Omega$ e non solo per un evento di probabilità 1. Tuttavia questa nozione di convergenza non è molto naturale (in un certo senso è troppo forte) per le variabili aleatorie.

PROPOSIZIONE 14.5. *Una successione $(X_n)_{n\in\mathbb{N}}$ di variabili aleatorie converge in probabilità alla variabile aleatoria X se e solo se ogni sottosuccessione di $(X_n)_{n\in\mathbb{N}}$ ammette una sottosuccessione che converge quasi certamente a X.*

Dimostrazione. Supponiamo che $X_n \xrightarrow[n\to+\infty]{P} X$. Allora esiste una sottosuccessione $(X_{n_k})_k$ tale che per ogni $k>0$ vale $P(|X_{n_k}-X|>k^{-1})\leqslant 2^{-k}$. Sia A_k l'evento $A_k=\{\omega\in\Omega\,|\,|X_{n_k}(\omega)-X(\omega)|>k^{-1}\}$. Allora $\sum_{k>0}P(A_k)<+\infty$ e, grazie al primo lemma di Borel-Cantelli (Teorema 4.31) solo un numero finito degli A_k si verifica o, equivalentemente, un numero infinito degli A_k non si verifica. Pertanto con probabilità 1 per ogni $k>0$, tranne al più un numero finito, vale la disuguaglianza $|X_{n_k}-X|\leqslant k^{-1}$, quindi $X_{n_k}\xrightarrow[k\to+\infty]{\text{q.c.}} X$ (immediato usando la caratterizzazione con il limsup).

Supponiamo ora che X_n non converga in probabilità a X. Allora esistono δ, ε e una sottosuccessione $(X_{n_k})_k$ tali che per ogni k abbiamo $P(|X_{n_k}-X|>\varepsilon)\geqslant\delta$. Questa sottosuccessione non può contenere alcuna sottosuccessione che coverga quasi certamente. □

Con questa caratterizzazione non ci sorprende la seguente "gerarchia" di convergenza.

PROPOSIZIONE 14.6. *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. Se $X_n\xrightarrow[n\to+\infty]{\text{q.c.}}X$, allora l'insieme $N=\{\omega\in\Omega:\lim_{n\to+\infty}X_n(\omega)\neq X(\omega)\}$ ha probabilità nulla. Preso $\varepsilon>0$ consideriamo la successione di insiemi $A_k=\bigcup_{j\geqslant k}\{|X_j-X|>\varepsilon\}$. Questi insiemi formano una successione decrescente $(A_k\supseteq A_{k+1}\supseteq A_{k+2}\supseteq\cdots)$ che ammette come limite $A_\infty=\bigcap_{k\geqslant 1}A_k$. In particolare abbiamo che $\lim_{k\to+\infty}P(A_k)=P(A_\infty)$. Mostriamo ora che $P(A_\infty)=0$. Per ogni $\omega\notin N$, $\lim_{n\to+\infty}X_n(\omega)=X(\omega)$, quindi definitivamente (per k abbastanza grande) non sta in A_k e quindi nemmeno in A_∞. Questo significa che $A_\infty\subseteq N$, quindi $P(A_\infty)\leqslant P(N)=0$ e

$$P(|X_k-X|>\varepsilon)\leqslant P(A_k)\searrow 0.$$ □

Non abbiamo ancora finito con le nozioni di convergenza. Ora vediamo un altro tipo di convergenza che, pur avendo un corrispettivo nella visione puramente "analitica" della convergenza, assume qui un nuovo significato.

DEFINIZIONE 14.7. *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). Osserviamo anche che non ci sono motivi per non considerare la convergenza in L^p…

PROPOSIZIONE 14.8. *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 (10.1) 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. □

Osservazione 14.9. 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.

Vediamo ora un esempio di successione di variabili aleatorie che converge quasi certamente e in probabilità ma non in media quadratica.

Esempio 14.10. Consideriamo per $n>0$ la successione di variabili aleatorie $X_n = n\,\mathbb{1}_{\left(0,\frac{1}{n}\right)}$. Questa successione converge quasi certamente (e dunque anche in probabilità) alla variabile aleatoria identicamente nulla. Tuttavia $E[|X_n-0|^2]=n$ non converge a 0, quindi la successione non converge alla variabile aleatoria nulla in L^2.

Ora tocca a un esempio in cui abbiamo convergenza in media quadratica (e quindi in probabilità), ma non quasi certa. In particolare non vale il viceversa della Proposizione 14.6.

Esempio 14.11. Consideriamo per $n>0$ la successione di variabili aleatorie indipendenti, $(X_n)_{n\in\mathbb{N}^+}$ con $X_n \sim \text{bin}\left(1,\frac{1}{n}\right)$. Allora per ogni $0<\varepsilon<1$ vale

$$P(|X_n|>\varepsilon) = P(X_n=1) = \frac{1}{n} \xrightarrow[n\to+\infty]{} 0,$$

dunque $X_n \xrightarrow[n\to+\infty]{P} 0$. Inoltre (fatto che implica la convergenza in probabilità) $X_n \xrightarrow[n\to+\infty]{L^2} 0$ perché $E[|X_n-0|^2] = E[X_n] = \frac{1}{n} \xrightarrow[n\to+\infty]{} 0$.

Tuttavia non è vero che $X_n \xrightarrow[n\to+\infty]{\text{q.c.}} 0$, infatti gli eventi $A_n = \{X_n=1\}$ sono indipendenti e

$$\sum_{n=1}^{+\infty} P(A_n) = \sum_{n=1}^{+\infty} \frac{1}{n} = +\infty.$$

Allora il secondo lemma di Borel-Cantelli (Teorema 4.32) ci dà che $P(\text{limsup } A_n)=1$, ossia che gli A_n sono frequentemente veri, dunque $P(X_n \to 0)=0$.

DEFINIZIONE 14.12. *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 $x\in\mathbb{R}$ $\lim_{n\to+\infty} P(X_n\leqslant x) = \tilde P(X\leqslant x)$, ossia se $\lim_{n\to+\infty} F_{X_n}(x) = F_X(x)$ in ogni punto di continuità di F_X.*

Osservazione 14.13. 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 $(\tilde\Omega,\tilde{\mathcal{F}},\tilde P)$. In realtà non è nemmeno necessario considerare le variabili della successione nel medesimo spazio (anche se possiamo considerare uno spazio prodotto). C'è un caso in cui anche la convergenza in probabilità e quella in L^2 possono essere considerate per successioni di variabili aleatorie che non sono in un unico spazio di probabilità, ossia quando la variabile limite è una variabile degenere, cioè una costante. In questo caso infatti la variabile degenere sta in ciascuno degli spazi di probabilità (perché sta in tutti), quindi abbiamo $P_n(|X_n-c|>\varepsilon)$ e $\mathrm{E}_n[|X_n-c|^2]$ che sono ben definiti.

PROPOSIZIONE 14.14. *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. Dato un qualunque $a \in \mathbb{R}$, fissiamo $\varepsilon > 0$. Allora

$$\begin{aligned} F_{X_n}(a) &= P(X_n \leqslant a) = P(X_n \leqslant a, X \leqslant a+\varepsilon) + P(X_n \leqslant a, X > a+\varepsilon) \\ &= P(X_n \leqslant a \,|\, X \leqslant a+\varepsilon)\, P(X \leqslant a+\varepsilon) + P(X_n \leqslant a, X > a+\varepsilon) \\ &\leqslant P(X \leqslant a+\varepsilon) + P(X_n < X - \varepsilon) \\ &\leqslant F_X(a+\varepsilon) + P(|X_n - X| > \varepsilon) \end{aligned}$$

in cui abbiamo messo in evidenza alcuni termini che sappiamo controllare grazie alla convergenza in probabilità. In modo analogo

$$\begin{aligned} F_X(a-\varepsilon) &= P(X \leqslant a-\varepsilon, X_n \leqslant a) + P(X \leqslant a-\varepsilon, X_n > a) \\ &= P(X \leqslant a-\varepsilon \,|\, X_n \leqslant a)\, P(X_n \leqslant a) + P(X \leqslant a-\varepsilon, X_n > a) \\ &\leqslant P(X_n \leqslant a) + P(X < X_n - \varepsilon) \\ &\leqslant F_{X_n}(a) + P(|X_n - X| > \varepsilon). \end{aligned}$$

Mettendo assieme le disuguaglianze ottenute,

$$F_X(a-\varepsilon) - P(|X_n - X| > \varepsilon) \leqslant F_{X_n}(a) \leqslant F_X(a+\varepsilon) + P(|X_n - X| > \varepsilon).$$

Passando al limite abbiamo, grazie alla convergenza in probabilità di X_n a X,

$$F_X(a-\varepsilon) \leqslant \liminf_{n\to+\infty} F_{X_n}(a) \leqslant \limsup_{n\to+\infty} F_{X_n}(a) \leqslant F_X(a+\varepsilon)$$

da cui, siccome vale per ogni $\varepsilon > 0$, $\lim_{n\to+\infty} F_{X_n}(a) = F_X(a)$. □

In generale non è vero il viceversa, nemmeno se $(X_n)_{n\in\mathbb{N}}$ e X sono definite sullo stesso spazio(Problema 58).

Osservazione 14.15. Possiamo riassumere i legami tra i vari concetti di convergenza di variabili aleatorie con lo schema in Figura 14.1. Ci sono casi in cui è possibile invertire le implicazioni, sotto opportune ipotesi (ad esempio Problema 59), ma vanno oltre i contenuti di questo corso.

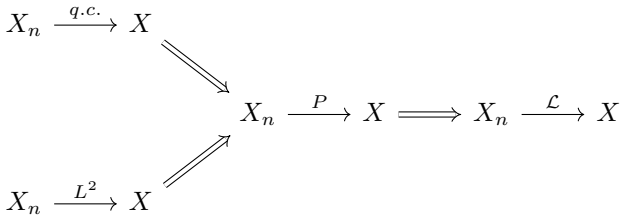


Figura 14.1. Gerarchia delle convergenze di variabili aleatorie

14.2. TEOREMI LIMITE

Cominciamo con qualche richiamo di risultati già visti.

PROPOSIZIONE 14.16. *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 14.17. È interessante notare come si comportino i risultati della Proposizione 14.16 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 14.17.

TEOREMA 14.18. (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 14.16 e la disuguaglianza di Chebychev (11.1): 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 14.19. Il fatto che il Teorema 14.18 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 14.6, 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

$$\frac{S_n}{n} \xrightarrow[n\to+\infty]{P} \frac{1}{2} \qquad \Longleftrightarrow \qquad \frac{S_n}{n} - \frac{1}{2} \xrightarrow[n\to+\infty]{P} 0 \qquad \Longleftrightarrow \qquad \frac{S_n - \frac{n}{2}}{n} \xrightarrow[n\to+\infty]{P} 0.$$

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 14.2.

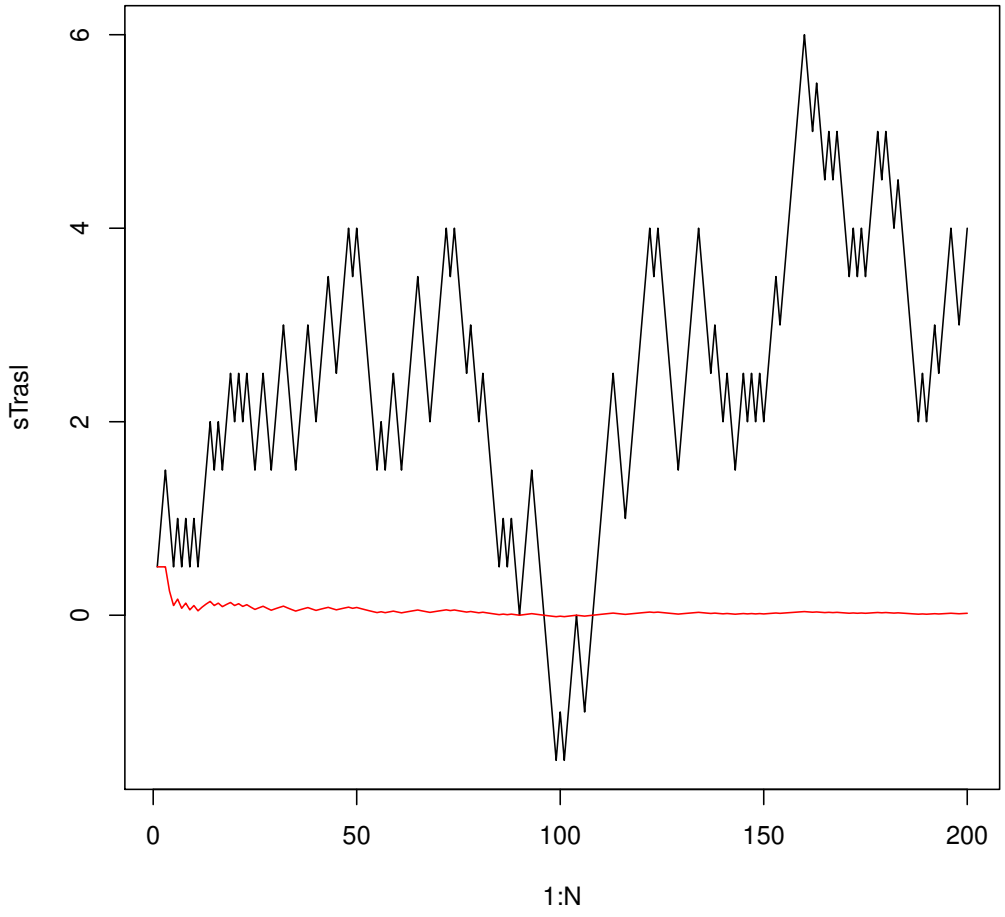


Figura 14.2. Una realizzazione di 200 lanci di una moneta. In nero la quantità $S_n - \frac{n}{2}$, che oscilla e non converge, in rosso $\frac{S_n - \frac{n}{2}}{n}$ che converge (molto rapidamente) a 0.

Vediamo anche il codice usato per generare la Figura 14.2:

```
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 14.20. 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 9.15) 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 9.15), 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 14.21. (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. Diamo la dimostrazione di un risultato più debole: richiediamo che le X_n siano indipendenti *e identicamente distribuite* (e quindi abbiano stessa media e stessa varianza) e che ammettano funzione generatrice dei momenti $M_X(t)$ finita in un intorno di 0.

In questo caso vediamo come è fatta la funzione generatrice dei momenti di $S_n^* := \frac{S_n - n\mu}{\sigma\sqrt{n}}$. Possiamo riscrivere

$$S_n^* = \frac{\sum (X_i - \mu)}{\sqrt{n}\,\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^*$$

ossia la somma delle standardizzate $X_i^* := \frac{X_i - \mu}{\sigma}$ (riscalata di $\frac{1}{\sqrt{n}}$). Grazie alla Proposizione 13.5 la funzione generatrice dei momenti di S_n^* è allora $M_{S_n^*}(t) = \left(M_{X^*}\left(\frac{t}{\sqrt{n}}\right)\right)^n$, dove M_{X^*} è la funzione generatrice dei momenti comune a tutte le X_i^*, identicamente distribuite dal momento che lo sono le X_i. Siccome per ipotesi esiste $M_X(t)$ finita in un intorno di 0, allora esiste anche $M_{X^*}(t)$ finita in un intorno di 0, dal momento che $M_{X^*}\left(\frac{t}{\sqrt{n}}\right) = M_X\left(\frac{t}{\sigma\sqrt{n}}\right) e^{-\frac{\mu t}{\sqrt{n}\sigma}}$.

Ora vediamo $M_{X^*}(t)$ scritta come serie di Taylor in un intorno di 0

$$\begin{aligned} M_{X^*}(t) &\approx M_{X^*}(0) + M'_{X^*}(0)\,t + M''_{X^*}(0)\,\frac{t^2}{2!} \\ &= 1 + E[X^*]\,t + E[(X^*)^2]\,\frac{t^2}{2!}. \end{aligned}$$

A questo punto osserviamo che $E[X^*] = E\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}(E[X] - \mu) = 0$ e che

$$E[(X^*)^2] = \mathrm{Var}[X^*] = \mathrm{Var}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma^2}\mathrm{Var}[X] = 1$$

e mettendo tutto assieme abbiamo

$$M_{S_n^*}(t) = \left(M_{X^*}\left(\frac{t}{\sqrt{n}}\right)\right)^n \approx \left(1 + \frac{\left(\frac{t}{\sqrt{n}}\right)^2}{2}\right)^n = \left(1 + \frac{\frac{t^2}{2}}{n}\right)^n \to \mathrm{e}^{\frac{t^2}{2}}$$

ossia il limite delle funzioni generatrici dei momenti delle S_n^* è la funzione generatrice dei momenti di una normale standard. □

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[2]. Possiamo però leggerci qualcosa di più: abbiamo che (la variabile aleatoria) $\left|S_n - \frac{n}{2}\right|$ 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 14.1, l'andamento di alcune grandezze al crescere di n, nell'esempio guida del processo di Bernoulli con la moneta bilanciata.

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 14.1. 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 $\left|S_n - \frac{n}{2}\right|$ dell'ordine di $\sqrt{n}$.

Osservazione 14.22. 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,

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \sim \mathcal{N}(0,1)$$

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

$$\frac{S_n}{n} \sim \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \text{oppure} \qquad S_n \sim \mathcal{N}(n\mu, \sigma\sqrt{n}).$$

2. E il suo nome: si chiama normale perché è la norma.

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 14.23. (**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

$$\frac{S_n - n\,E[X_1]}{\sqrt{n}\,\sqrt{\text{Var}[X_1]}} \dot\sim \mathcal{N}(0,1),$$

ossia che

$$P\left(\frac{S_n - n\,E[X_1]}{\sqrt{n}\,\sqrt{\text{Var}[X_1]}} \leqslant x\right) \simeq \Phi(x). \tag{14.1}$$

Cerchiamo di riscrivere (14.1) in modo da mettere in evidenza S_n e ottenere qualcosa della forma $P(S_n \leqslant y)$: manipolando il primo membro della (14.1) abbiamo

$$P(S_n \leqslant x\cdot\sqrt{n}\,\sqrt{\text{Var}[X_1]} + n\,E[X_1]) \simeq \Phi(x),$$

quindi vogliamo determinare x tale che

$$x\cdot\sqrt{n}\,\sqrt{\text{Var}[X_1]} + n\,E[X_1] = y,$$

cioè

$$x = \frac{y - n\,E[X_1]}{\sqrt{n}\,\sqrt{\text{Var}[X_1]}}.$$

Saremmo potuti arrivare allo stesso risultato ricordando che $S_n \dot\sim \mathcal{N}(y - n\,E[X_1], \sqrt{n}\,\sqrt{\text{Var}[X_1]})$ e usando le proprietà di standardizzazione di una Gaussiana, per cui

$$P(S_n \leqslant y) = F_{S_n}(y) \simeq \Phi\left(\frac{y - n\,E[X_1]}{\sqrt{n}\,\sqrt{\text{Var}[X_1]}}\right).$$

Quanto fatto finora non usa il contesto specifico del problema che stiamo considerando, ma ora andiamo a sostituire i valori specifici:

$$n = 10^6, \qquad E[X_1] = 0, \qquad \text{Var}[X_1] = \frac{10^{-20}}{12}, \qquad y = \pm 0.5\cdot 10^{-7},$$

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 14.24. Nell'Esempio 14.23 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[3]:

- $X_i \sim \mathcal{N}$: in questo caso $n = 1$, grazie alla riproducibilità
- $X_i \sim \mathrm{unif}$: in questo caso $n \geqslant 5$ dà di solito buoni risultati
- $X_i \sim \exp$ o $X_i \sim \mathrm{geom}$: abbiamo bisogno di $n \geqslant 15$ (sono molto dissimili da delle normali)
- $X_i \sim \chi^2$: possiamo usare la riproducibilità $\chi^2_n \sim \mathcal{N}(n, \sqrt{2n})$ 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,

$$\mathrm{bin}(n,p) \sim \mathcal{N}(np, \sqrt{np\,(1-np)}).$$

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,

$$\mathrm{Pois}(\lambda) \sim \mathcal{N}\left(\lambda, \sqrt{\lambda}\right)$$

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 $\tilde{\lambda} = \frac{\lambda}{n}$.

14.3. PROBLEMI

Problema 58. Mostrare (con un controesempio) che date una successione $(X_n)_n$ e una variabile aleatoria X sul medesimo spazio di probabilità tali che $X_n \xrightarrow{\mathcal{L}} X$, non è necessariamente vero che $X_n \xrightarrow{P} X$.

Problema 59. Mostrare che date una successione $(X_n)_n$ e una variabile aleatoria degenere $X = c \in \mathbb{R}$ sul medesimo spazio di probabilità tali che $X_n \xrightarrow{\mathcal{L}} X$, allora $X_n \xrightarrow{P} X$.

Problema 60. Consideriamo una successione $(X_n)_{n\in\mathbb{N}}$ di variabili aleatorie geometriche di parametri $(p_n)_{n\in\mathbb{N}}$, con $p_n \in (0,1)$ per ogni n. Dimostrare che se la successione $(p_n)_{n\in\mathbb{N}}$ è $\left(\frac{\lambda}{n}\right)_n$ per qualche $\lambda > 0$, allora $\frac{X_n}{n} \xrightarrow[n\to+\infty]{\mathcal{L}} \exp(\lambda)$. Possiamo rilassare questa condizione?

3. Queste sono regole molto spannometriche, giusto per avere un'idea nella pratica.

Problema 61. Il numero di molecole di sodio in 1 ml di acqua minerale può essere descritto da una variabile aleatoria di Poisson con media 200. Lo strumento con cui analizziamo quest'acqua può ricevere solo multipli interi positivi di 1 ml (capacità della pipetta con cui viene caricato).

1. Determinare la probabilità che 10 ml di acqua contengano più di 2000 molecole, utilizzando l'approssimazione normale e la correzione di continuità.
2. Qual è l'errore assoluto tra la probabilità ottenuta con l'approssimazione calcolata al punto precedente e il valore esatto?
3. Determinare la minima quantità di ml di acqua da analizzare nello strumento affinché la probabilità che ci siano almeno 2400 molecole di sodio sia non inferiore a 0.83, utilizzando l'approssimazione normale e la correzione di continuità.

Problema 62. Nella nota località Vason sul Monte Bondone, la quantità di neve che cade in un giorno, durante l'inverno, può essere descritta come una variabile aleatoria di media m cm e deviazione standard s cm. Qual è la probabilità che in g giorni cadano tra a e b centimetri di neve?

Parte II

Statistica

CAPITOLO 15

STIME PUNTUALI

Con questo capitolo iniziamo l'esplorazione della Statistica Inferenziale, 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 $X \approx \mathcal{N}$). Calcoliamo *stime* di parametri (o momenti o altre quantità) e facciamo verifiche della compatibilità delle ipotesi con i 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.

15.1. INTRODUZIONE ALLA STATISTICA

Alla base della Probabilità avevamo lo spazio degli esiti, in Statistica questo ruolo è preso dalla popolazione. Richiamiamo la definizione, già vista nel Capitolo 1, di popolazione.

DEFINIZIONE 15.1. *Chiamiamo* popolazione (di riferimento) *un insieme costituito da elementi* (*distinti*)*, sui quali conduciamo la nostra indagine. Chiamiamo tali elementi* esemplari, individui o unità statistiche.

Come già accennato, la differenza tra Statistica descrittiva, trattata nel Capitolo 1, e Statistica inferenziale è sostanzialmente che nel primo caso abbiamo dati su tutta la popolazione, nel secondo abbiamo dati solo su un sottoinsieme della popolazione e cerchiamo di ricostruire (inferire, appunto) la distribuzione dei dati per l'intera popolazione.

DEFINIZIONE 15.2. *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 è costituita dagli 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, di cui è la controparte non probabilistica).

Osservazione 15.3. 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 15.4. 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 15.5. *Una* statistica *è una funzione calcolabile a partire dalla misurazione del campione.*

Esempio 15.6. 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 di interesse è inferiore a una certa soglia c: $\inf\{i \in \{1, \dots, n\} : x_i < c\}$.

Osservazione 15.7. 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. Tuttavia, anche per motivi storici, non rispetteremo sempre questa regola.

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 quasi esclusivamente di Statistica parametrica.

Osservazione 15.8. 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 15.9. 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.

15.2. STIMATORI E STIME

DEFINIZIONE 15.10. *Lo* stimatore *di un parametro è 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.

NOTAZIONE 15.11. *In parziale contraddizione con quanto detto prima, è in realtà abbastanza comune usare $\hat{\vartheta}$ per indicare lo stimatore (ossia la funzione, la variabile aleatoria) del parametro ϑ, invece della maiuscola Θ. In particolare questo succede per la media μ, dal momento che la corrispondente maiuscola sarebbe M.*

Esempio 15.12. 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 14.18): 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 15.13. 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 14.21) ci dice infatti che

$$\overline{X} - \mu \sim \mathcal{N}\left(0, \frac{\sigma}{\sqrt{n}}\right).$$

Osservazione 15.14. 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[1], tra quelli che possiamo calcolare coi dati a nostra disposizione.

DEFINIZIONE 15.15. *Uno stimatore Θ di un parametro ϑ è:*

- corretto o non distorto (unbiased), *se* $E[\Theta] = \vartheta$

1. Non abbiamo ancora specificato rispetto a quale metrica intendiamo misurare la bontà degli stimatori. Ce ne sono infatti diverse, come vedremo tra poco.

– distorto (biased), *se* $E[\Theta] \neq \vartheta$*; in questo caso il valore* $E[\Theta] - \vartheta$ *è la* distorsione o bias.
Se $\lim_{n\to+\infty} E[\Theta_n] = \vartheta$*, allora* Θ *è* asintoticamente non distorto.

Osservazione 15.16. Dire che uno stimatore è distorto o 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'erroneità della stima. Possiamo considerare altre funzioni di errore, che penalizzino un maggiore allontanamento dal valore "vero" del parametro.

DEFINIZIONE 15.17. *L'*errore quadratico medio (mean square error, MSE) *di uno stimatore* Θ *del parametro* ϑ *è la quantità*

$$\mathrm{MSE}[\Theta] = E[(\Theta - \vartheta)^2].$$

Osservazione 15.18. Possiamo scrivere l'errore quadratico medio in modo leggermente diverso, in analogia a quanto visto per la varianza (anche questo è un momento secondo):

$$\begin{aligned}
\mathrm{MSE}[\Theta] &= E[(\Theta - \vartheta)^2] = E[(\Theta - E[\Theta] + E[\Theta] - \vartheta)^2] \\
&= E[(\Theta - E[\Theta])^2] + E[(E[\Theta] - \vartheta)^2] + 2\,E[(\Theta - E[\Theta])\,(E[\Theta] - \vartheta)] \\
&= \mathrm{Var}[\Theta] + (\mathrm{bias})^2 + 2\,(E[\Theta] - E[\Theta])\,(E[\Theta] - \vartheta) \\
&= \mathrm{Var}[\Theta] + (\mathrm{bias})^2.
\end{aligned}$$

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 15.19. *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* $n \to +\infty$*,* Θ *è* consistente in media quadratica.

Il prossimo risultato ci dà una condizione sufficiente per la consistenza di uno stimatore.

PROPOSIZIONE 15.20. *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 15.18,

$$\mathrm{MSE}[\Theta_n] = \mathrm{Var}[\Theta_n] + (E[\Theta_n] - \vartheta)^2$$

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 14.8). □

Osservazione 15.21. Uno stimatore può essere corretto ma non consistente. Ad esempio se le variabili aleatorie $(X_1, \ldots, X_n)$ sono indipendenti e identicamente distribuite, allora ciascuna X_i è uno stimatore non distorto della media μ, poiché $E[X_i] = \mu$ per ogni $i \in \{1, \ldots, n\}$.

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.

15.2.1. Alcuni stimatori

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 14.21) 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 14.18)

$$S_{*n}^2 = \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 \xrightarrow[n\to+\infty]{P} E[(X_i-\mu)^2] = \mathrm{Var}[X_i].$$

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,

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i=1}^n X_i^2 - \hat{\mu}^2\right] &= E\left[\frac{1}{n}\sum_{i=1}^n (X_i^2-\mu^2) - (\hat{\mu}^2-\mu^2)\right] \\ &= \frac{1}{n}n\sigma^2 - \mathrm{Var}[\hat{\mu}] \\ &= \sigma^2\cdot\frac{n-1}{n}, \end{aligned} \tag{15.1}$$

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 (15.1) 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 15.22. 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 15.23. Possiamo prendere, come stimatori della deviazione standard, $S = \sqrt{S^2}$ oppure $S_* = \sqrt{S_*^2}$. È possibile però mostrare (Problema 63) 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).

15.3. 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.

15.3.1. Metodo dei momenti

DEFINIZIONE 15.24. *Il k-simo* momento campionario *è la variabile aleatoria*

$$\hat{\mu}^k = \hat{\mu}^k_n = \frac{1}{n}\sum_{i=1}^{n} X_i^k.$$

Il k-simo momento della popolazione *è il numero $\mu^k = E[X_i^k]$.*

Osservazione 15.25. 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 15.26. *Lo* stimatore col metodo dei momenti *del parametro scalare ϑ è 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, lo stimatore col metodo dei momenti *del parametro vettoriale ϑ è 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 15.27. 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 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 $\hat{\mu}_{\text{mom}}$ tale che

$$\mu^1(\hat{\mu}_{\text{mom}}) = \hat{\mu},$$

ossia, siccome in questo caso la funzione μ^1 è l'identità, $\hat{\mu}_{\text{mom}} = \hat{\mu}$, cioè lo stimatore di μ con il metodo dei momenti è la media campionaria.

15.3.2. Metodo di massima verosimiglianza

Partiamo sempre dalla funzione di densità congiunta $f_X(x_1,\dots,x_n|\vartheta)$, 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 15.28. *Lo* stimatore di massima verosimiglianza[2] *del parametro ϑ è la quantità $\hat{\vartheta}_{\text{MLE}}$ che soddisfa*

$$\hat{\vartheta}_{\text{MLE}} = \operatorname{argmax}_\theta f(x_1,\dots,x_n|\theta) = \operatorname{argmax}_\theta \log(f(x_1,\dots,x_n|\theta)).$$

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. Notiamo anche che non abbiamo specificato in che spazio sia il parametro di interesse. In particolare può essere scalare, ma anche vettoriale in uno spazio di dimensione maggiore.

Esempio 15.29. 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 densità congiunta

$$f(x_1,\dots,x_n|\vartheta) = f(x_1,\dots,x_n|\mu,\sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{2\pi}\right)^{\frac{n}{2}} \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}.$$

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)

$$\begin{aligned}
\frac{\partial}{\partial\mu}\log(f(x_1,\dots,x_n|\mu,\sigma^2)) &= -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(-2)(x_i-\mu) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu) \\
\frac{\partial}{\partial\sigma^2}\log(f(x_1,\dots,x_n|\mu,\sigma^2)) &= -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2
\end{aligned}$$

2. In inglese si chiama *maximum likelihood estimator*, da cui la sigla MLE.

e azzerarle[3]

$$\begin{cases} \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)=0 \\ -\frac{n}{\sigma^2}+\frac{1}{(\sigma^2)^2}\sum_{i=1}^n (x_i-\mu)^2=0 \end{cases} \quad \text{da cui} \quad \begin{cases} \mu=\frac{1}{n}\sum_{i=1}^n x_i \\ \sigma^2=\frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2. \end{cases}$$

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 15.30. 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

$$f(x_1,\dots,x_n|p)=p_X(x_1,\dots,x_n|p)=p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

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

$$\hat{p}_{\text{MLE}}=\frac{1}{n}\sum_{i=1}^n X_i=\overline{X}.$$

15.4. PROBLEMI

Problema 63. Dimostrare che lo stimatore $S=\sqrt{S^2}$ della deviazione standard è distorto ma consistente. Suggerimento: S sottostima la deviazione standard.

Problema 64. Data 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, determinare gli stimatori con il metodo dei momenti dei parametri μ e σ^2, cioè $\hat{\vartheta}_{\text{mom}}=(\hat{\mu}_{\text{mom}},\hat{\sigma}^2_{\text{mom}})$.

Problema 65. Consideriamo una popolazione uniforme sull'intervallo $[-a,a]$ e un campione $(X_1,\dots,X_n)$ di dimensione n, con $X_i \sim \text{unif}(-a,a)$. Stimare il parametro $\vartheta=a$ usando il metodo dei momenti.

Problema 66. Consideriamo ora una popolazione esponenziale di parametro ignoto λ, da cui estraiamo un campione $(X_1,\dots,X_n)$ di variabili indipendenti. Calcolare lo stimatore di massima verosimiglianza $\hat{\lambda}_{\text{MLE}}$ per il parametro λ.

Problema 67. Calcolare lo stimatore di massima verosimiglianza per il parametro λ di una popolazione di Poisson.

3. Dovremmo controllare che i punti così ottenuti siano di massimo globale e non siano punti di minimo, di sella o di massimo locale.

Problema 68. Determinare lo stimatore di massima verosimiglianza per il parametro a di una popolazione uniforme su $[-a, a]$.

Problema 69. Determinare lo stimatore di massima verosimiglianza per i parametri a e b di una popolazione uniforme su $[a, b]$.

CAPITOLO 16

INTERVALLI DI FIDUCIA O 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[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 (*fiducia* o *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.

16.1. 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 15.2.1 uno stimatore per la media e uno per la varianza[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':

$$(n-1)\, S_n^2 = \sum_{i=1}^{n} (X_i - \hat{\mu})^2 = \sum_{i=1}^{n} X_i^2 - n\,\hat{\mu}^2.$$

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.

2. Ci concentriamo sulla varianza e non sulla deviazione standard perché la prima ha uno stimatore corretto e consistente.

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[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.

COROLLARIO 16.1. *Sia $(X_1,\ldots,X_n)$ 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 16.2. *Una* funzione ancillare per un parametro ϑ *è una variabile aleatoria la cui legge sia nota a priori*[4] *e che dipenda dai dati, da parametri noti e da ϑ, unico parametro non noto.*

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.

Parliamo anche di quantità pivot, *se lasciamo cadere la richiesta di dipendenza da un solo parametro incognito.*

Esempio 16.3. 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.

16.2. MEDIA DI UNA NORMALE DI VARIANZA NOTA

Abbiamo un campione $(X_1,\dots,X_n)$ estratto da una popolazione Gaussiana di media μ (che vogliamo stimare) e varianza σ^2 che assumiamo nota. Abbiamo già osservato nel Capitolo Capitolo 15 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 $\mathbb{R}$

$$P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}} \leqslant b\right) = \Phi(b), \qquad P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}} \geqslant a\right) = 1-\Phi(a).$$

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),

$$P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}} \leqslant x\right) = \beta \iff x = \Phi^{-1}(\beta)$$
$$P\left(\frac{\overline{X}_n-\mu}{\sqrt{\sigma^2/n}} \geqslant y\right) = \beta \iff y = \Phi^{-1}(1-\beta) = -\Phi^{-1}(\beta).$$

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[5], per $\alpha \in (0,1)$.

16.2.1. 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:

$$P(\mu < A) = \frac{\alpha}{2} \qquad P(\mu > B) = \frac{\alpha}{2}.$$

4. Dire che la legge è nota a priori significa che la distribuzione della variabile aleatoria non dipende dai parametri.

5. La nostra fiducia o confidenza è un numero reale in $(0,1)$, ma se vogliamo pensarla come probabilità dobbiamo farlo con un po' di cautela.

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

$$P\left(\overline{X}_n - \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} \leqslant \mu \leqslant \overline{X}_n + \Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}} \right) = 1-\alpha.$$

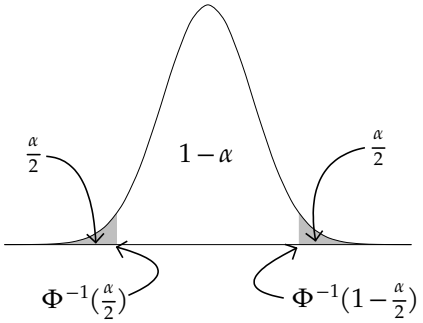
Gli estremi dell'intervallo, come accennato in precedenza, sono statistiche: dipendono dal campione e da parametri prefissati (in questo caso α). Quindi, avendo realizzato il campione, gli estremi saranno dei numeri.

DEFINIZIONE 16.4. *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 $\alpha \in (0,1)$, l'*intervallo di confidenza bilaterale a livello $1-\alpha$ per la media μ *è 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

$1-\alpha$	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 16.5. 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. La confidenza è una probabilità se cambiamo spazio di probabilità: prendendo campioni diversi dalla medesima popolazione con una certa probabilità conterranno la media vera. Ma nel momento in cui abbiamo estratto un campione la probabilità collassa in 0 o 1.

Esempio 16.6. 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 $\bar{x}=104.7$ (se avessimo avuto i dati, avremmo potuto calcolarla in R come `xbar <- mean(x)`, con `x` il vettore dei dati).

Allora l'intervallo di confidenza a livello 95% per μ è

$$\begin{aligned}\left(\bar{x}+\Phi^{-1}(0.025)\cdot\frac{\sigma}{\sqrt{16}},\bar{x}+\Phi^{-1}(0.975)\cdot\frac{\sigma}{\sqrt{16}}\right) &= \left(104.7-1.96\cdot\frac{3}{4},104.7+1.96\cdot\frac{3}{4}\right)\\ &= (103.23,106.17).\end{aligned}$$

Possiamo calcolarlo con l'aiuto di R:

```
conf <- 0.95 # livello di fiducia
sigma <- 3   # deviazione standard nota
xbar + qnorm(1/2+conf/2)*sigma/sqrt(length(x))*c(-1,+1)
```

La forma dei quantili nel codice è leggermente diversa da quella scritta sopra, ma basta osservare che il livello di fiducia è $1-\alpha$ quindi $1-\frac{\alpha}{2}=\frac{2-1+\text{conf}}{2}=\frac{1}{2}+\frac{\text{conf}}{2}$. Osserviamo anche che abbiamo usato il *recycling* in R e la simmetria della funzione quantile di una distribuzione Gaussiana standard.

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^2 nota. 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

$$\overline{X}_n+\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}-\overline{X}_n+\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}=2\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}$$

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

$$l=2\,\Phi^{-1}\left(1-\frac{\alpha}{2}\right)\sqrt{\frac{\sigma^2}{n}}$$

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,

$$n = \left\lceil \frac{4 \cdot (\Phi^{-1}(1-\frac{\alpha}{2}))^2 \cdot \sigma^2}{l^2} \right\rceil. \tag{16.1}$$

Esempio 16.7. 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 (16.1) 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.

16.2.2. 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

$$A = \overline{X}_n - \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}} \qquad B = \overline{X}_n + \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}.$$

DEFINIZIONE 16.8. *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 $\alpha \in (0,1)$, l'*intervallo di confidenza unilaterale destro (rispettivamente sinistro) a livello $1-\alpha$ per la media μ *è la semiretta*

$$\left(\overline{X}_n - \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}, +\infty\right) \qquad (\textit{rispettivamente} \quad \left(-\infty, \overline{X}_n + \Phi^{-1}(1-\alpha)\sqrt{\frac{\sigma^2}{n}}\right) \quad).$$

Esempio 16.9. 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 $+\infty$), in R

```
x <- c(3.35,3.73,3.14,4.37,4.28,2.91,2.96,1.94,2.29)
xbar <- mean(x)
conf <- 0.9
ll <- xbar - qnorm(conf)*sigma/sqrt(length(x))
print(ll)
```

L'intervallo di confidenza sinistro al 99% ha come estremo destro

$$\bar{x} + \Phi^{-1}(0.99) \cdot \sqrt{\frac{1}{9}} = 3.22 + 2.326 \cdot \frac{1}{3} = 3.995$$

```
conf <- 0.99
ul <- xbar + qnorm(conf)*sigma/sqrt(length(x))
print(ul)
```

16.2.3. Visualizzare gli intervalli di confidenza

Riprendiamo il discorso fatto sopra e generiamo 100 campioni da una popolazione normale standard. Ciascun campione è costituito da 500 individui.

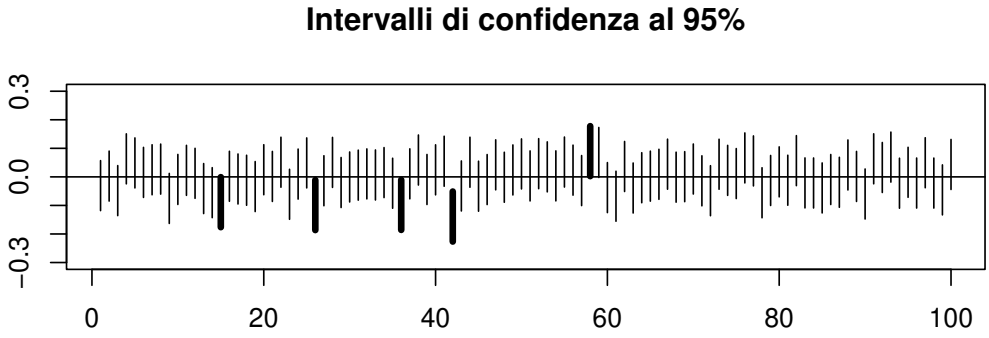


Figura 16.1. Intervalli di confidenza al 95% per la media di una popolazione Gaussiana di media teorica 0, costruiti su 100 campioni di numerosità 500 ciascuno. In nero sono evidenziati quelli che non contengono il valore teorico.

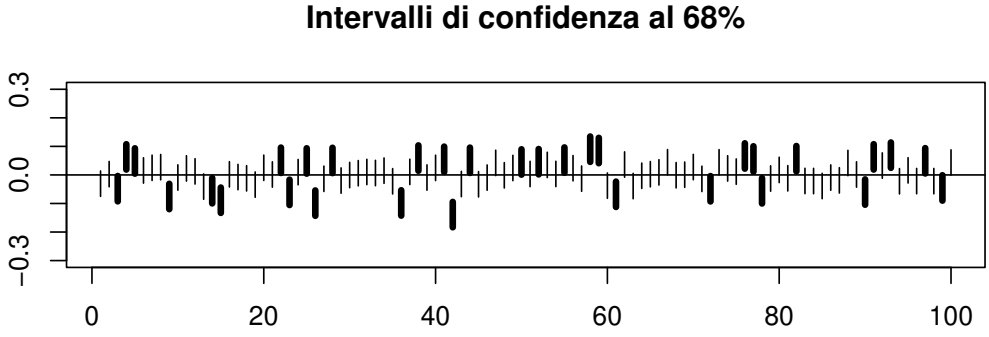


Figura 16.2. Intervalli di confidenza al 68% per la media di una popolazione Gaussiana di media teorica 0, costruiti su 100 campioni di numerosità 500 ciascuno. In nero sono evidenziati quelli che non contengono il valore teorico. Vale la pena osservare che gli intervalli sono ben più piccoli di quelli a livello 95% costruiti sui medesimi campioni in Figura 16.1.

Il codice che genera queste due figure è il seguente.

```
# funzione per rappresentare graficamente gli intervalli di
# confidenza
# Tratta da Ugarte et al., Probability and Statistics with R
interval.plot <- function(ll, ul, conf = 0.95){
  y0 <- ll
  y1 <- ul
  n <- length(ll)
 plot(y0, type = 'n', ylim = c(-0.3,0.3), xlab = "", ylab = "",
       main = paste0("Intervalli di confidenza al ", conf*100,
"%"))
  segments((1:n), y0, (1:n), y1, lwd = 1 + (y0>0|y1<0)*3)
  abline(h = 0)
  s <- sum(ll<=0 &ul>=0)
  print(paste0("Numero di intervalli che contengono 0: ", s))
}
# Codice principale
set.seed(127) # imposto un seme per riproducibilità
m <- 100 # numero di campioni
n <- 500 # numerosità di ciascun campione
conf <- c(0.95,0.68) # livelli di confidenza di interesse
M <- matrix(rnorm(m*n), nrow = m, ncol = n, byrow = TRUE)
  # genero i miei campioni,
  # mettendone ciascuno su una riga di una matrice
Mbar <- M %*% rep(1,n)/n
  # calcolo la media per ogni campione (riga) usando le
  # proprietà delle matrici %*% è il prodotto tra matrici
for(i in 1:length(conf)){
  ll <- Mbar + qnorm((1-conf[i])/2)*sqrt(1/n)
  ul <- Mbar + qnorm(1-(1-conf[i])/2)*sqrt(1/n)
  interval.plot(ll, ul, conf[i])
}
```

16.3. 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 16.10. 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).

Possiamo riscrivere tutto questo come codice R:

```
IC <- function(x, conf = .95, s = NA){
  # x il campione, conf il livello di fiducia (di default 0.95),
   # s la deviazione standard
   if(is.na(s)){ # se s non noto, usa statistica t
     s <- sd(x) # calcola stima puntuale di s
     return(mean(x) +
            qt(1/2 + conf/2, df = length(x) - 1)*
            s/sqrt(length(x))*c(-1,+1))
   }
   else{ # std dev nota
     return(mean(x) +
           qnorm(1/2 + conf/2) * s/sqrt(length(x)) * c(-1,+1))
   }
}
```

Possiamo anche scoprire che esiste una funzione built-in di R che fa (anche) questo lavoro, nel caso della varianza ignota: `t.test`. Se guardiamo l'help per questa funzione leggiamo

```
t.test(x, y = NULL,
       alternative = c("two.sided", ``less", ``greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

e ciò che ci interessa sono gli argomenti `x`, i dati, e `conf.level`, il livello di confidenza. L'output è decisamente più ricco di quello che ci occorre, ma include una stima puntuale per μ e il corrispondente intervallo di fiducia calcolato con la statistica t.

Chiudiamo osservando che il caso in cui σ è ignota è decisamente più frequente (e realistico) nell'uso della statistica.

Esempio 16.11. 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 16.12. 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 16.1. 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 $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

16.4. 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 [6].

Ci sono però alcuni casi speciali: il primo è 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,

$$\frac{S_{X,n}^2}{\sigma^2}(n-1) + \frac{S_{Y,m}^2}{\sigma^2}(m-1) \sim \chi^2(n-1) + \chi^2(m-1) \sim \chi^2(n+m-2).$$

6. 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

$$(\mu_X-\mu_Y)\in\left((\bar{x}-\bar{y})-F^{-1}_{t(n+m-2)}\left(1-\frac{\alpha}{2}\right)\sqrt{S^2_P\left(\frac{1}{n}+\frac{1}{m}\right)},(\bar{x}-\bar{y})+F^{-1}_{t(n+m-2)}\left(1-\frac{\alpha}{2}\right)\sqrt{S^2_P\left(\frac{1}{n}+\frac{1}{m}\right)}\right).$$

16.5. 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.

16.5.1. 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 14.21) 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 ritroviamo con una forma non così piacevole come quelle viste in precedenza, perché p compare anche a denominatore e per giunta sotto una radice quadrata. Poniamo per semplicità $u_\alpha := \Phi^{-1}(1-\frac{\alpha}{2})$; abbiamo

$$\left|\frac{Y_n - np}{\sqrt{np(1-p)}}\right| \leqslant u_\alpha$$

da cui, elevando al quadrato e moltiplicando,

$$\frac{1}{n}(Y_n - np)^2 \leqslant u_\alpha^2 p(1-p)$$

e raccogliendo i termini in p

$$(n+u_\alpha^2)p^2 - (2Y_n + u_\alpha^2)p + \frac{Y_n^2}{n} \leqslant 0.$$

Sappiamo che $\overline{X}_n = \frac{Y_n}{n}$ è uno stimatore della media. Poniamo quindi $\hat{p} = \frac{Y_n}{n} = \overline{X}_n$. Allora la disuguaglianza diventa, in p,

$$\frac{2n\hat{p} + u_\alpha^2 - \sqrt{u_\alpha^4 + 4u_\alpha^2 n\hat{p} - 4u_\alpha^2 n\hat{p}^2}}{2(n+u_\alpha^2)} \leqslant p \leqslant \frac{2n\hat{p} + u_\alpha^2 + \sqrt{u_\alpha^4 + 4u_\alpha^2 n\hat{p} - 4u_\alpha^2 n\hat{p}^2}}{2(n+u_\alpha^2)}.$$

In alternativa possiamo usare 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 poniamo nuovamente $\hat{p} = \frac{Y_n}{n} = \overline{X}_n$. 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

$$\frac{n\hat{p} - np}{\sqrt{n\hat{p}(1-\hat{p})}} \dot{\sim} \mathcal{N}(0,1).$$

A questo punto possiamo riprendere la strada iniziata prima:

$$P\left(-\Phi^{-1}\Big(1-\frac{\alpha}{2}\Big) \leqslant \frac{n\hat{p} - np}{\sqrt{n\hat{p}(1-\hat{p})}} \leqslant \Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\right) \approx 1-\alpha$$

che ora possiamo riscrivere in modo da avere un intervallo esplicito per p:

$$P\left(\hat{p} - \Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leqslant p \leqslant \hat{p} + \Phi^{-1}\Big(1-\frac{\alpha}{2}\Big)\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1-\alpha.$$

In modo simile possiamo procedere per gli intervalli unilaterali.

Osservazione 16.13. 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[7] 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.

7. Consideriamo qui il caso più approssimato, il secondo di quelli visti poco sopra, ma le considerazioni valgono allo stesso modo anche nel primo caso.

Esempio 16.14. 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 16.15. 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.

16.5.2. 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: $n \cdot \overline{X}_n \sim \text{Pois}(n\lambda)$, quindi

$$P(\overline{X}_n = k) = P\left(\sum_{i=1}^{n} X_i = nk\right) = \frac{(n\lambda)^{nk}}{(nk)!} e^{-n\lambda}.$$

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

$$F_{\text{Pois}(\lambda)}(k) = 1 - F_{\chi^2(2(k+1))}(2\lambda)$$

da cui possiamo ricavare[8] 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

$$\frac{\overline{X}_n - \lambda}{\sqrt{\lambda/n}} \dot{\sim} \mathcal{N}(0,1).$$

8. Non vediamo i dettagli, che non sono banali.

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 16.16. 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.

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[9]

```
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 16.17. 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.

9. Siccome generiamo il campione in modo aleatorio, gli estremi dell'intervallo saranno diversi in iterazioni diverse.

Osservazione 16.18. 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.

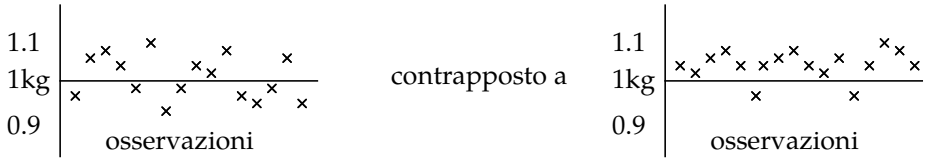
16.6. PROBLEMI

Gli esempi non visti a lezione possono essere usati come problemi.

CAPITOLO 17

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:



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 un buon candidato per 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 15 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 16 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. Questa può essere una buona idea (anche se non è la sola possibile), ma è molto imprecisa. Per rendere il ragionamento più rigoroso, iniziamo introducendo una terminologia dedicata.

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. Viceversa, l'ipotesi alternativa è la risposta in caso di test positivo, ossia se abbiamo evidenza statistica dai dati che che l'ipotesi nulla non sia compatibile coi dati raccolti. questa differenza filosofica tra l'ipotesi nulla e quella alternativa è abbastanza centrale per capire e interpretare i risultati di un test statistico e più avanti torneremo su questo argomento.

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, dividiamo lo spazio n-dimensionale dei potenziali campioni in due parti: la *regione critica* e la *regione di accettazione*. Se il campione (un vettore n-dimensionale, ossia un punto nello spazio n-dimensionale) cade all'interno della regione critica, allora rifiutiamo l'ipotesi nulla, scegliendo l'ipotesi alternativa come risposta. Se il campione giace al di fuori della regione critica (ossia all'interno della regione di accettazione) non rifiutiamo l'ipotesi nulla[1].

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 dobbiamo accettare che una procedura di questo tipo avrà sicuramente degli errori. In un test statistico possiamo 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 17.1.

	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$
H_0	ok	Errore di seconda specie
H_1	Errore di prima specie	ok

Tabella 17.1. 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 17.2.

	$\boldsymbol{H_0}$	$\boldsymbol{H_1}$
H_0	Vero negativo (TN)	Falso negativo (FN)
H_1	Falso positivo (FP)	Vero positivo (TP)

Tabella 17.2. 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[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.

1. È una questione di terminologia, ma per l'impatto che ha sulla comunicazione e sulla comprensione del valore di queste affermazioni sarebbe meglio dire *rifiutare* o *non rifiutare* l'ipotesi nulla rispetto a *non accettare* e *accettare*. In particolare quest'ultima, come vedremo tra poco, è un po' fuorviante come terminologia.

2. Vedremo alcuni dettagli in più tra qualche pagina.

17.1. IMPOSTARE TEST STATISTICI

Cominciamo considerando un test d'ipotesi bilaterale o a due code, ossia in cui l'ipotesi nulla è della forma $H_0: \vartheta = \vartheta_0$ e l'ipotesi alternativa $H_1: \vartheta \neq \vartheta_0$ può essere vista come composta da due code, appunto, $H_1: \vartheta < \vartheta_0 \vee \vartheta > \vartheta_0$. Come osserveremo in seguito i test unilaterali o a una sola coda non sono diversi nella sostanza, anche se nella forma presentano qualche differenza.

Algoritmo per un test bilaterale

1. Stabilire le ipotesi da testare. Nel caso bilaterale saranno della forma $H_0: \vartheta = \vartheta_0$ e $H_1: \vartheta \neq \vartheta_0$.
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 (o osservata) del test*, ossia la funzione ancillare valutata nel 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 $\mathrm{RA} = [a, b]$ e di conseguenza la regione critica $\mathrm{RC} = [a, b]^c$.
6. Rifiutare l'ipotesi nulla H_0 (e accettare l'alternativa H_1) se la statistica osservata giace nella regione critica, non rifiutare H_0 se la statistica osservata giace nella regione di accettazione.

Esempio 17.1. 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

$$Z_0 = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{5.96 - 5.38}{1/\sqrt{81}} = 5.22.$$

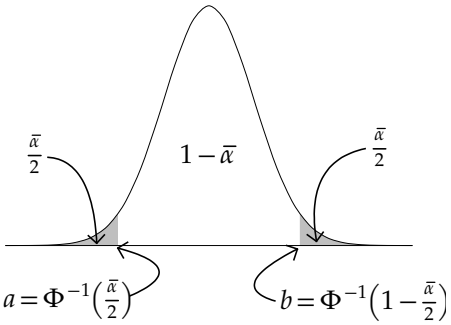
   Osserviamo che nel momento in cui abbiamo le osservazioni del campione (anzi, in realtà ci basta solamente il valore della media campionaria) questo è un numero, z_0.

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 $Z_0 \in \mathrm{RC}_Z = \mathrm{RA}_Z^c$ (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

$$\begin{aligned}
\alpha &= P(\text{dire } H_1 \,|\, \text{è vera } H_0) \\
&= P(Z_0 \notin \mathrm{RA}_Z \,|\, \text{è vera } H_0) \\
&= P(Z_0 \notin [a,b] \,|\, Z_0 \sim \mathcal{N}(0,1)) \\
&= P\left(Z_0 \notin \left[\Phi^{-1}\left(\frac{\bar{\alpha}}{2}\right), \Phi^{-1}\left(1-\frac{\bar{\alpha}}{2}\right)\right] \,|\, Z_0 \sim \mathcal{N}(0,1)\right) \\
&= \bar{\alpha}
\end{aligned}$$



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 $\bar{\alpha}$ (nell'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 17.1, 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.

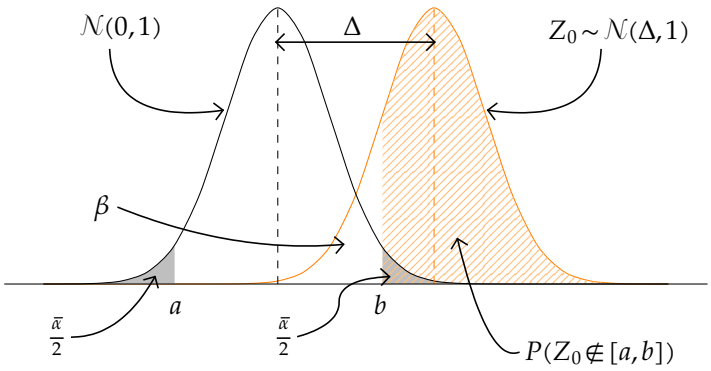


Figura 17.1. Probabilità che il test risponda H_1 se è vera H_1.

Per analizzarla meglio si può introdurre la *curva operativa caratteristica* (OCC) $\beta(\mu)$, come segue:

$$\begin{aligned}
\beta(\mu) &= P(\text{dire } H_0 | \text{la media è } \mu) = P(Z_0 \in \mathrm{RA}_Z | \text{la media è } \mu) \\
&= P\left(a \leqslant \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \leqslant b \middle| \text{la media è } \mu \right) \\
&= P\left(a + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leqslant \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leqslant b + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \text{la media è } \mu \right) \\
&= P\left(a + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \leqslant \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \leqslant b + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \text{la media è } \mu \right) \\
&= \Phi\left(b + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \right) - \Phi\left(a + \frac{\mu_0 - \mu}{\sqrt{\sigma^2/n}} \right).
\end{aligned}$$

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 $\overline{\alpha}$ e che sia sufficientemente potente da commettere errori di seconda specie con probabilità al più $\overline{\beta}$ se il valore vero della media è $\mu_T \neq \mu_0$ (cioè $\beta(\mu_T) \leqslant \overline{\beta}$), ci basta scegliere n sufficientemente grande affinché $\beta(\mu_T) \approx \overline{\beta}$. 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 17.2. 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 o a una coda, in cui l'ipotesi nulla è della forma $H_0: \vartheta \geqslant \vartheta_0$ e l'ipotesi alternativa è $H_1: \vartheta < \vartheta_0$ (o viceversa, $H_0: \vartheta \leqslant \vartheta_0$ e $H_1: \vartheta > \vartheta_0$): in questo caso la regione di accettazione sarà non più un intervallo, ma (in genere) una semiretta.

Un aspetto che vale la pena sottolineare nel caso di test unilaterali è il seguente: l'ipotesi alternativa è della forma $\vartheta > \vartheta_0$ (o simmetrica), mentre l'ipotesi nulla è spesso indicata con $\vartheta = \vartheta_0$ invece che con il complementare $\vartheta \leqslant \vartheta_0$. Questo perché nel momento in cui andiamo a considerare l'ipotesi nulla per esplicitare la distribuzione condizionata a tale ipotesi, prendiamo proprio il caso dell'uguaglianza, per alcuni motivi. Il primo è di comodità: si tratta di sostituire un valore, invece che un insieme di valori. Il secondo è che il valore di uguaglianza è quello "limite" verso l'ipotesi alternativa: in questo modo stiamo considerando il caso più favorevole per l'ipotesi nulla, che domina tutti i valori alla propria sinistra.

17.2. 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 $[\overline{\alpha}, 1]$. In altre parole, se il p-value è compreso tra $\overline{\alpha}$ e 1 non rifiutiamo l'ipotesi nulla H_0, mentre se è tra 0 e $\overline{\alpha}$, rifiutiamo H_0 e accettiamo l'ipotesi alternativa H_1.

Osservazione 17.3. 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 17.4. 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 sarà un numero non appena abbiamo la realizzazione del campione) 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.

Arricchiamo questo esempio con qualche numero. Supponiamo che il campione sia di taglia 64 e che l'ipotesi da testare sia $\mu = 5.5$. 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). Giusto per avere un'idea, la popolazione gaussiana da cui è stato estratto il campione aveva media 5 e deviazione standard 2.

Per 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 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)`.

In questo caso entra in gioco la funzione `t.test`: dall'help di R abbiamo

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

in cui `alternative` è il tipo di test che vogliamo condurre (in particolare qual è la forma dell'ipotesi alternativa), `mu` è quella che noi abbiamo chiamato μ_0 e `conf.level` corrisponde a $1-\bar{\alpha}$, ossia il complemento a 1 del livello di significatività del test.

Osservazione 17.5. 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 supponendo che sia vera l'ipotesi nulla.

Esempio 17.6. 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 funzione ancillare 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. Supponendo di avere il campione salvato come vettore `x`, chiamiamo `w0` la statistica test `var(x)*(length(x)-1)/4` (siamo nel caso $\sigma_0^2 = 4$). Ora 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 17.7. 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.

17.3. 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 $\Theta - \vartheta_0 \geqslant c$)[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 17.8. 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 $H_0 \colon \mu \leqslant \mu_0$ contro l'ipotesi alternativa $H_1 \colon \mu > \mu_0$. 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 $\overline{\alpha}$ ci basta allora considerare il caso peggiore, ossia in cui $\mu = \mu_0$:

$$\alpha = P\left(\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} > \frac{c + \mu_0 - \mu}{\sqrt{\sigma^2/n}} \middle| \mu \leqslant \mu_0 \right) \leqslant P\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} > \frac{c}{\sqrt{\sigma^2/n}} \right) \overset{!}{=} \overline{\alpha}.$$

Quindi

$$1 - \overline{\alpha} = P\left(\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \leqslant \frac{c}{\sqrt{\sigma^2/n}} \right)$$

e, approfittando del fatto che $\frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}}$ ha distribuzione normale standard, $c = \Phi^{-1}(1 - \overline{\alpha}) \sqrt{\sigma^2/n}$. In altre parole accettiamo l'ipotesi nulla $\mu \leqslant \mu_0$ se $\overline{X} \leqslant \mu_0 + \Phi^{-1}(1 - \overline{\alpha}) \sqrt{\sigma^2/n}$ e la rifiutiamo se invece abbiamo $\overline{X} > \mu_0 + \Phi^{-1}(1 - \overline{\alpha}) \sqrt{\sigma^2/n}$, se la significatività che abbiamo fissato è $\overline{\alpha}$.

Ovviamente anche in questo caso unilaterale possiamo ricavare il p-dei-dati, infatti rispondiamo H_0 se e solo se

$$Z_0 = \frac{\overline{X} - \mu_0}{\sqrt{\sigma^2/n}} \leqslant \Phi^{-1}(1 - \overline{\alpha}) \iff \Phi(Z_0) \leqslant 1 - \overline{\alpha} \iff \overline{\alpha} \leqslant 1 - \Phi(Z_0),$$

ossia il p-dei-dati è $1 - \Phi(Z_0)$.

Osservazione 17.9. 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 17.10. L'asimmetria tra ipotesi nulla e ipotesi alternativa è forse ancora più evidente nel caso di test unilaterali: se consideriamo due test speculari

3. Potremmo anche (se $\vartheta_0 > 0$) considerare il rapporto $\Theta/\vartheta_0 \leqslant c$, per un altro, opportuno, c.

$H_0: \vartheta \leqslant \vartheta_0$ $H_1: \vartheta > \vartheta_0$	$H_0: \vartheta \geqslant \vartheta_0$ $H_1: \vartheta < \vartheta_0$

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 17.11. 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:

$$\begin{aligned} P(W > W_0) &= P\left(W > \frac{s^2}{\sigma_0^2}(n-1)\right) = 1 - F^{-1}_{\chi^2(n-1)}\left(\frac{s^2}{\sigma_0^2}(n-1)\right) \\ &= 1 - F_{\chi^2(19)}\left(\frac{0.025}{0.0225}\cdot 19\right) = 1 - F_{\chi^2(19)}(21.11) \\ &= 1 - 0.669 = 0.331 \end{aligned}$$

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:

$$\begin{aligned} P(W < W_0) &= P\left(W < \frac{s^2}{\sigma_0^2}(n-1)\right) = F_{\chi^2(n-1)}\left(\frac{s^2}{\sigma_0^2}(n-1)\right) \\ &= F_{\chi^2(19)}(21.11) = 0.669. \end{aligned}$$

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 17.12. Supponiamo di voler fare un test sul parametro di una popolazione Bernoulliana, ad esempio per determinare se la probabilità di passare l'esame di *Calcolo delle Probabilità e Statistica Matematica* (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[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: $B \sim \text{bin}(n,p)$, per essa sappiamo che vale

$$P(B \geqslant a) = \sum_{k=\lceil a \rceil}^{n} P(B=k) = \sum_{k=\lceil a \rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k}. \tag{17.1}$$

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 (17.1) 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

$$P(B \geqslant n\bar{x} \,|\, p \leqslant p_0) \leqslant \sum_{k=n\bar{x}}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k}$$

cioè `pbinom(q, size = n, prob = p0, lower.tail = FALSE)`, con $q = n\bar{x} - 1$.

Se per esempio fossero passati all'esame 40 studenti su 50 del campione (cioè l'80% del campione, osserviamo che in questo caso $n = 50$, $\bar{x} = 0.8$ e $q = n\bar{x} - 1 = 39$), il p-dei-dati sarebbe 0.262, ossia non si potrebbe rifiutare l'ipotesi nulla che la probabilità di passare l'esame sia minore o uguale del 75%.

Osservazione 17.13. Nel caso Bernoulliano (come in altri casi) possiamo anche usare le statistiche approssimate che abbiamo visto grazie ai teoremi limite. Sappiamo infatti che

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.

$$Z = \frac{n\overline{X} - np}{\sqrt{np(1-p)}} \rightsquigarrow \mathcal{N}(0,1) \qquad\qquad Z_0 = \frac{n\overline{X} - np_0}{\sqrt{np_0(1-p_0)}}.$$

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. Per calcolarlo con R ricaviamo $q = \frac{n\bar{x} - np_0}{\sqrt{np_0(1-p_0)}}$ e usiamo la funzione `pnorm(q, lower.tail = TRUE)`.

Tuttavia dal momento che stiamo considerando una popolazione discreta e la stiamo approssimando con una continua, è opportuno introdurre (almeno nel calcolo del p-dei-dati) la correzione di continuità. Lo vediamo nel dettaglio nel prossimo esempio.

Esempio 17.14. Usiamo gli stessi dati dell'Esempio 17.12, ma ora ci mettiamo nei panni di chi tiene il corso e, salvo chiara evidenza statistica che l'esame sia troppo severo, non ha intenzione di modificare l'impostazione attuale delle prove d'esame.

In questo caso abbiamo come ipotesi nulla $H_0: p \geqslant p_0$ ossia $p \geqslant 0.75$ e come ipotesi alternativa $H_1: p < p_0$. Allora indicando con X la variabile aleatoria binomiale che conta il numero di successi (e che supponiamo di parametro $p_0 = 0.75$) il p-dei-dati è $P(X \leqslant 40)$. Introduciamo ora la correzione di continuità e passiamo poi al p-dei-dati approssimato. Abbiamo

$$\begin{aligned} p\text{-value} &= P(X \leqslant 40) \\ &= P(X \leqslant 40.5) \\ &= P\left(\frac{X - n\cdot p_0}{\sqrt{np_0(1-p_0)}} \leqslant \frac{40.5 - n\cdot p_0}{\sqrt{np_0(1-p_0)}}\right) \\ &= P\left(\frac{X - n\cdot p_0}{\sqrt{np_0(1-p_0)}} \leqslant \frac{40.5 - n\cdot p_0}{\sqrt{np_0(1-p_0)}}\right) \end{aligned}$$

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%.

Osservazione 17.15. Nel caso Bernoulliano, oltre a riscrivere in R quanto fatto in astratto, possiamo anche usare alcune funzioni predefinite. La prima è

```
binom.test(x, n, p = 0.5,
           alternative = c("two.sided", "less", "greater"),
           conf.level = 0.95)
```

che fa un test binomiale esatto. Tra gli argomenti `x` è il numero di successi o un vettore di lunghezza 2 contenente il numero di successi e di insuccessi, rispettivamente, `n` è il numero di esperimenti o tentativi, `p` è la probabilità soglia, che abbiamo chiamato p_0.

L'altra funzione conduce il test approssimato ed è

```
prop.test(x, n, p = NULL,
          alternative = c("two.sided", "less", "greater"),
          conf.level = 0.95, correct = TRUE)
```

con i medesimi argomenti della precedente. Per saperne di più su queste funzioni è possibile consultare l'help di R.

Grazie a queste funzioni possiamo anche studiare la bontà dei test svolti usando l'approssimazione Gaussiana, in particolare al variare della dimensione del campione analizzato.

17.4. TABELLE RIASSUNTIVE

Raccogliamo ora in alcune tabelle alcuni dei test più rilevanti.

H_0	H_1	Statistica test	Regione di accettazione (H_0)	p-dei-dati
$\mu=\mu_0$	$\mu\neq\mu_0$	$z_0=\dfrac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}}$	$-\Phi^{-1}\left(1-\dfrac{\bar{\alpha}}{2}\right)\leqslant z_0\leqslant\Phi^{-1}\left(1-\dfrac{\bar{\alpha}}{2}\right)$	$2\,(1-\Phi(\lvert z_0\rvert))$
$\mu\leqslant\mu_0$	$\mu>\mu_0$	$z_0=\dfrac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}}$	$z_0\leqslant\Phi^{-1}(1-\bar{\alpha})$	$1-\Phi(z_0)$
$\mu\geqslant\mu_0$	$\mu<\mu_0$	$z_0=\dfrac{\bar{x}-\mu_0}{\sqrt{\sigma^2/n}}$	$z_0\geqslant-\Phi^{-1}(1-\bar{\alpha})=\Phi^{-1}(\bar{\alpha})$	$\Phi(z_0)$

Tabella 17.3. 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-dei-dati
$\mu=\mu_0$	$\mu\neq\mu_0$	$t_0=\dfrac{\bar{x}-\mu_0}{\sqrt{s^2/n}}$	$-F^{-1}_{t(n-1)}\left(1-\dfrac{\bar{\alpha}}{2}\right)\leqslant t_0\leqslant F^{-1}_{t(n-1)}\left(1-\dfrac{\bar{\alpha}}{2}\right)$	$2\,(1-F_{t(n-1)}(\lvert t_0\rvert))$
$\mu\leqslant\mu_0$	$\mu>\mu_0$	$t_0=\dfrac{\bar{x}-\mu_0}{\sqrt{s^2/n}}$	$t_0\leqslant F^{-1}_{t(n-1)}(1-\bar{\alpha})$	$1-F_{t(n-1)}(t_0)$
$\mu\geqslant\mu_0$	$\mu<\mu_0$	$t_0=\dfrac{\bar{x}-\mu_0}{\sqrt{s^2/n}}$	$t_0\geqslant-F^{-1}_{t(n-1)}(1-\bar{\alpha})=F^{-1}_{t(n-1)}(\bar{\alpha})$	$F_{t(n-1)}(t_0)$

Tabella 17.4. 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-dei-dati
$\sigma^2=\sigma_0^2$	$\sigma^2\neq\sigma_0^2$	$w_0=\dfrac{s^2}{\sigma_0^2}(n-1)$	$F^{-1}_{\chi^2_{n-1}}\left(\dfrac{\bar{\alpha}}{2}\right)\leqslant w_0\leqslant F^{-1}_{\chi^2_{n-1}}\left(1-\dfrac{\bar{\alpha}}{2}\right)$	$2\,(F_{\chi^2_{n-1}}(w_0)\wedge 1-F_{\chi^2_{n-1}}(w_0))$
$\sigma^2\leqslant\sigma_0^2$	$\sigma^2>\sigma_0^2$	$w_0=\dfrac{s^2}{\sigma_0^2}(n-1)$	$w_0\leqslant F^{-1}_{\chi^2_{n-1}}(1-\bar{\alpha})$	$1-F_{\chi^2_{n-1}}(w_0)$
$\sigma^2\geqslant\sigma_0^2$	$\sigma^2<\sigma_0^2$	$w_0=\dfrac{s^2}{\sigma_0^2}(n-1)$	$w_0\geqslant F^{-1}_{\chi^2_{n-1}}(\bar{\alpha})$	$F_{\chi^2_{n-1}}(w_0)$

Tabella 17.5. 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.

17.5. CONFRONTO TRA MEDIE DI POPOLAZIONI NORMALI

Finora abbiamo sempre confrontato un parametro estratto da una popolazione con un valore di riferimento. Tuttavia è abbastanza frequente avere il problema di confrontare due popolazioni diverse, o meglio confrontarne i parametri. Come esempio guida prendiamo ancora una volta quello della media di una distribuzione normale, solo che abbiamo due popolazioni normali e ci chiediamo se ci sia evidenza statistica tale da dire che le loro medie sono diverse, oppure che la media di una sia maggiore della media dell'altra. Un caso concreto potrebbe essere il confronto della media di un indicatore biologico (normalmente distribuito) tra una popolazione cui è stato somministrato un certo farmaco e una cui è stato somministrato un altro farmaco (o un placebo).

Ci sono a priori due modi di raccogliere i dati (o meglio di selezionare i campioni). La prima possibilità è avere un campione da cui estraiamo *coppie* di dati, in cui il primo elemento appartiene alla prima popolazione e il secondo alla seconda popolazione. Si parla in questo caso di *dati appaiati*. Per chiarire un po' le idee vediamo qualche esempio.

Esempio 17.16. Dobbiamo capire se protegga meglio dai raggi solari la crema *Alsole* o la *Benabbronzo*. Un modo di raccogliere i dati in maniera appaiata potrebbe essere quello di selezionare un certo numero di individui e chiedere loro di applicare la crema A su un lato del corpo e la crema B sull'altro (tenendo traccia di quale è quale).

Vogliamo capire se un fertilizzante impatta la biodiversità del terreno in cui viene utilizzato. Possiamo allora individuare alcune porzioni di terreno, misurare la biodiversità in tali porzioni, spargere il fertilizzante e misurare nuovamente nelle medesime porzioni.

Per studiare meglio l'impatto dell'esposizione ai raggi cosmici o alla microgravità possiamo confrontare gli indicatori biologici di due gemelli, uno dei quali ha trascorso un anno sulla stazione spaziale internazionale e l'altro no.

In generale nel caso di dati appaiati possiamo confrontare la differenza tra due variabili minimizzando la variabilità dovuta ad altri fattori, non dipendenti dal campionamento.

Tuttavia, anche se preferibile in astratto, la raccolta di dati appaiati non è sempre possibile. A volte la si abbandona in favore di una maggiore semplicità del protocollo sperimentale, altre volte non è proprio realistica. Si parla in questo caso di *dati non appaiati*.

Esempio 17.17. Nel confronto tra le due creme solari si può decidere, per semplicità, che ad alcuni individui venga detto di usare una delle due creme e ad altri la seconda, senza dividere in due il corpo.

Negli studi farmaceutici uno stesso individuo non può prendere sia il farmaco sia il placebo, ci saranno individui che prendono il placebo e altri che prendono il farmaco. In questo caso i dati sono quindi necessariamente non appaiati (anche se una buona scelta del campione farà in modo che le popolazioni siano il più simili possibile).

17.5.1. Dati appaiati

Cominciamo allora dal caso di dati appaiati: per ogni individuo i abbiamo una coppia di osservazioni (X_i, Y_i) e possiamo definire delle variabili aleatorie ausiliarie $W_i = X_i - Y_i$. Stiamo supponendo che le popolazioni di riferimento (ossia le leggi delle X e delle Y) siano Gaussiane di media e varianza ignote. Allora anche W ha distribuzione Gaussiana e in particolare la sua media è la differenza tra la media di X e quella di Y. Ci siamo allora ricondotti a un problema che sappiamo risolvere: chiedersi se la media μ_X della prima popolazione sia uguale alla media μ_Y della seconda equivale a chiedere se la media μ_W delle differenze sia uguale a 0. Stiamo facendo un test d'ipotesi sulla media di una popolazione normale (la differenza delle due osservazioni) a varianza ignota. Possiamo quindi prendere $\overline{W} = \frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)$ e $S_W^2 = \frac{1}{n-1}\sum_{i=1}^{n}((X_i - Y_i) - \overline{W})^2$ e osservare che $\frac{\overline{W}}{\sqrt{S_W^2 \frac{1}{n}}} \sim t_{n-1}$. Abbiamo la funzione ancillare e tutto quello che ci occorre per condurre un test d'ipotesi sulla differenza tra le due popolazioni.

Stiamo facendo un test d'ipotesi e la statistica ancillare è ancora una volta una t. Questo ci fa pensare alla funzione R `t.test`. Dall'help di R abbiamo

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

in cui il parametro di interesse è questa volta `paired` che indica se i due campioni `x` e `y` sono da considerarsi appaiati. In questo caso dovremo specificare `paired = TRUE` per prevalere sul default, che è `FALSE`.

Esempio 17.18. Un gruppo di 28 pazienti viene sottoposto a una terapia sperimentale contro la Cespuglite. Per valutarne l'efficacia per ciascun paziente si misura un certo valore ematico prima e dopo la cura. Per ogni paziente si calcola quindi la differenza $x = x_p - x_d$ tra i valori ematici prima e dopo, di cui viene calcolata la media campionaria, pari a $\bar{x} = 0.16$ e la varianza campionaria $s^2 = 0.19$.

Non avendo informazioni più chiare su quale sia la "direzione" ottimale, ci limitiamo a indagare se ci sia evidenza di un effetto in seguito alla terapia, ossia se il valore medio sia cambiato. Abbiamo come ipotesi nulla $H_0: \mu = \mu_p - \mu_d = 0$ e come ipotesi alternativa $H_1: \mu \neq 0$.

Supponendo vera l'ipotesi nulla, la variabile $\frac{\bar{X}}{\sqrt{s^2/28}}$ ha distribuzione t di Student a 27 gradi di libertà. La probabilità di ottenere per la media campionaria valori che si scostano da 0 più di quello ottenuto è

$$P\left(\left|\frac{\bar{X}}{\sqrt{s^2/28}}\right| > \frac{0.16}{\sqrt{0.19/28}}\right) = P(|t_{27}| > 1.94),$$

ossia, riscrivendo in termini della funzione di ripartizione

$$\begin{aligned} P(t_{27} < -1.94) + P(t_{27} > 1.94) &= F_{t(27)}(-1.94) + 1 - F_{t(27)}(1.94) \\ &= 2\,(1 - F_{t(27)}(1.94)) \\ &= 2\,(1 - 0.97) \\ &= 0.06 \end{aligned}$$

Il test non risulta quindi significativo per livelli $\bar{\alpha} < 0.06$ e non possiamo rigettare l'ipotesi nulla: non c'è evidenza statistica (a livelli di significatività minori di 0.06) che la terapia abbia avuto alcun effetto.

In questo particolare esempio usare la funzione `t.test` non è indicato, dal momento che non abbiamo i dati espliciti ma già delle statistiche (o meglio delle stime) calcolate.

Esempio 17.19. (S. ROSS) Di recente nell'industria dei semiconduttori è stato introdotto un programma di sicurezza sul lavoro. Nella Tabella 17.6 sono riportate le medie settimanali delle ore-uomo perse a causa di incidenti, per 10 stabilimenti dalle caratteristiche simili, Le medie sono state calcolate nel corso di un mese prima e un mese dopo la riforma.

Stabilimento	Prima	Dopo	Differenza
1	30.5	23.0	−7.5
2	18.5	21.0	+2.5
3	24.5	22.0	−2.5
4	32.0	28.5	−3.5
5	16.0	14.5	−1.5
6	15.0	15.5	+0.5
7	23.5	24.5	+1.0
8	25.5	21.0	−4.5
9	28.0	23.5	−4.5
10	18.0	16.5	−1.5

Tabella 17.6. Medie settimanali di ore-uomo perse a causa di incidenti in 10 stabilimenti

Determiniamo a un livello di significatività del 5% se il programma di sicurezza è risultato efficace.

Se formalizziamo la richiesta in termini di un test d'ipotesi possiamo osservare che stiamo considerando dati appaiati (prima e dopo l'intervento, relativi ai medesimi individui) e dobbiamo chiederci quale sia la scelta giusta di ipotesi nulla e ipotesi alternativa. Indichiamo con μ_p la media prima dell'intervento, con μ_d la media dopo l'intervento e con μ_{diff} la media della differenza. Vorremmo avere evidenza statistica che il programma funzioni, quindi scegliamo come ipotesi nulla H_0: $\mu_{\text{diff}} = \mu_d - \mu_p \geqslant 0$ e come ipotesi alternativa H_1: $\mu_{\text{diff}} = \mu_d - \mu_p < 0$. In questo modo se l'evidenza statistica non è conclusiva diremo che non possiamo escludere che le ore-uomo perse siano (in media) aumentate dopo l'intervento.

Per procedere assumiamo che i dati siano distribuiti gaussianamente. Dovremmo in realtà verificare anche questa ipotesi. Graficamente non abbiamo molte informazioni (anche perché i dati sono molto pochi) e test più raffinati (come il test di Shapiro-Wilk) non escludono che i tre campioni (prima, dopo e differenza) siano estratti da popolazioni normali.

A questo punto possiamo usare R per la parte computazionale.

```
prima <- c(30.5, 18.5, 24.5, 32.0, 16.0, 15.0, 23.5, 25.5,
28.0, 18.0)
dopo <- c(23.0, 21.0, 22.0, 28.5, 14.5, 15.5, 24.5, 21.0, 23.5,
16.5)
diff <- dopo - prima
# facciamo ora qualche test di normalità, non affrontato a
# lezione
hist(prima)
qqnorm(prima)
qqline(prima)
shapiro.test(prima)
# I grafici non sono molto conclusivi, ma il test non esclude
# che i dati siano distribuiti normalmente
hist(dopo)
qqnorm(dopo)
qqline(dopo)
shapiro.test(dopo)
# Anche in questo caso i test non danno molte informazioni
hist(diff)
qqnorm(diff)
qqline(diff)
shapiro.test(diff)
# In questo caso c'è un po' di evidenza in più che la
# popolazione da cui è estratto diff sia gaussiana
v <- mean(diff)*sqrt(length(diff))/sd(diff)
pt(v, length(diff)-1)
```

Otteniamo un valore `v` della statistica circa uguale a -2.265949 per un p-dei-dati≈ 0.025. Dal momento che il p-value è inferiore alla soglia di significatività richiesta possiamo rigettare l'ipotesi nulla e dichiarare che nei dati c'è evidenza statistica a supporto della diminuzione delle ore-uomo perse per incidente[5].

Siccome si tratta di un test t possiamo anche usare direttamente la funzione R dedicata:

```
t.test(x = dopo, y = prima, paired = TRUE, alternative = "less")
# o equivalentemente usando la differenza
t.test(diff, alternative = "less")
```

5. Osserviamo che non siamo però in grado di affermare che la *causa* di questa diminuzione sia stata la riforma.

Osserviamo che, per i dati appaiati, non abbiamo avuto bisogno di fare ipotesi sulla varianza delle due popolazioni, che in particolare non deve essere uguale nei due casi. Usiamo infatti la varianza della differenza tra le osservazioni, approfittando del fatto che i dati sono appaiati.

17.5.2. Dati non appaiati, varianze note

Supponiamo ora che i campioni indipendenti $(X_i)_{i=1}^n$ e $(Y_j)_{j=1}^m$ siano estratti da due popolazioni $\mathcal{N}(\mu_X, \sigma_X)$ e $\mathcal{N}(\mu_Y, \sigma_Y)$, di cui conosciamo le deviazioni standard (e quindi le varianze). Anche in questo caso vogliamo testare se le due medie siano uguali (o se una sia maggiore o uguale dell'altra). Le possibili ipotesi nulle sono quindi H_0: $\mu_X = \mu_Y$ o H_0: $\mu_X \geqslant m_Y$. Le due popolazioni sono Gaussiane e sappiamo che $\bar{X}_n$ e $\bar{Y}_m$ sono due stimatori rispettivamente di μ_X e μ_Y. Allora $\bar{X}_n - \bar{Y}_m$ è uno stimatore della differenza tra le medie $\mu_X - \mu_Y$. Di questo stimatore conosciamo la distribuzione:

$$\bar{X}_n - \bar{Y}_m \sim \mathcal{N}\left(\mu_X - \mu_Y, \sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right)$$

dal momento che una combinazione lineare di variabili aleatorie Gaussiane indipendenti è a sua volta una Gaussiana. Ma allora possiamo standardizzare, ottenendo

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\sigma_X^2/n + \sigma_Y^2/m}} \sim \mathcal{N}(0,1)$$

e dal momento che le varianze sono note questa è una funzione ancillare per la quantità di interesse (ossia la differenza tra le medie).

Avendo la funzione ancillare possiamo impostare i test delle ipotesi a due e a una coda esattamente come nel caso della media di una singola popolazione Gaussiana di varianza nota.

17.5.3. Dati non appaiati, varianze ignote ma uguali

In molti casi, tuttavia, le varianze delle due popolazioni sono ignote, quindi abbiamo bisogno di una stima della deviazione standard (o della varianza) della differenza. Siamo ancora nel caso della differenza di due popolazione Gaussiane indipendenti e sappiamo che la varianza della differenza, come la varianza della somma, è la somma delle varianze. Per ciascuna delle due popolazioni abbiamo uno stimatore della varianza:

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \qquad S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2.$$

Se supponiamo di essere nel caso *omoschedastico*, ossia in cui le due varianze σ_X^2 e σ_Y^2 sono uguali, la varianza della differenza sarebbe $\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)$ e possiamo stimare σ^2 a partire da S_X^2 e S_Y^2 usando la cosiddetta *varianza pooled*:

$$S_p^2 = \frac{(n-1)\,S_X^2 + (m-1)\,S_Y^2}{n+m-2} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 + \sum_{j=1}^{m}(Y_j - \bar{Y}_m)^2}{n+m-2}.$$

Facciamo un passo indietro e ricordiamo che $(n-1)\frac{S_X^2}{\sigma^2} \sim \chi(n-1)$ e $(m-1)\frac{S_Y^2}{\sigma^2} \sim \chi(m-1)$, ma anche che la somma di due chi quadro è una chi quadro avente come gradi di libertà la somma dei gradi di libertà, quindi

$$(n-1)\frac{S_X^2}{\sigma^2} + (m-1)\frac{S_Y^2}{\sigma^2} = \frac{(n-1)\,S_X^2 + (m-1)\,S_Y^2}{\sigma^2} = (n+m-2)\frac{S_p^2}{\sigma^2} \sim \chi(n+m-2).$$

Mettendo assieme i pezzi abbiamo che

$$\frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{\sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)}}\sqrt{\frac{\sigma^2}{S_p^2}} = \frac{(\bar{X}_n - \bar{Y}_m) - (\mu_X - \mu_Y)}{\sqrt{S_p^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \sim t(n+m-2).$$

Anche in questo caso ci siamo quindi ricondotti a una funzione ancillare e a una situazione già nota e possiamo impostare i test delle ipotesi a due e a una coda come nel caso della media di una singola popolazione a varianza ignota, prestando attenzione al diverso numero di gradi di libertà e al diverso stimatore della varianza.

Ci rivolgiamo ancora una volta alla funzione R `t.test`. Dall'help di R abbiamo

```
t.test(x, y = NULL,
       alternative = c("two.sided", "less", "greater"),
       mu = 0, paired = FALSE, var.equal = FALSE,
       conf.level = 0.95, ...)
```

in cui il parametro di interesse è questa volta `var.equal` che indica se i due campioni `x` e `y` sono da considerarsi omoschedastici o no. In questo caso, volendo il caso omoschedastico, dovremo specificare `var.equal = TRUE` per prevalere sul default, che è `FALSE`.

Esempio 17.20. Vogliamo confrontare i valori di un parametro ottenibile dall'analisi delle urine e collegato alla funzionalità renale in due gruppi di lavoratori, esposti a un differente rischio di contaminazione chimica. I lavoratori del primo gruppo sono $n=8$ e presentano un valore di media $\bar{x}=0.84$ e varianza campionaria $s_X^2=0.20$. I lavoratori del secondo gruppo sono $m=12$ e presentano un valore di media $\bar{y}=1.31$ e varianza campionaria $s_Y^2=0.21$.

Vogliamo verificare se ci sia evidenza statistica che la media del parametro di funzionalità renale per il secondo gruppo sia più alta. Indichiamo con X la variabile Gaussiana che esprime il parametro di funzionalità renale dei lavoratori esposti a minor rischio di contaminazione chimica e con Y la variabile Gaussiana che esprime lo stesso parametro per i lavoratori esposti a rischio maggiore e supponiamo di sapere, per esperienza clinica, che le rispettive varianze σ_X^2 e σ_Y^2 sono approssimativamente uguali. Vogliamo verificare se $\mu_Y > \mu_X$.

Poniamo come ipotesi nulla $H_0: \mu_X \geqslant \mu_Y$ e come ipotesi alternativa $H_1: \mu_X < \mu_Y$. Supponendo vera l'ipotesi nulla allora la variabile $\frac{(\bar{X}-\bar{Y})}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}}$ sarà distribuita come una t di Student con 18 gradi di libertà. Lo stimatore pooled della varianza è

$$S_p^2 = \frac{(7\cdot 0.2 + 11\cdot 0.21)}{18} = 0.21.$$

Anche in questo caso vogliamo calcolare il p-dei-dati, ossia la probabilità di ottenere per la differenza $\bar{X}-\bar{Y}$ delle medie campionarie un valore minore di quello ottenuto, quindi

$$P\left(\frac{(\bar{X}-\bar{Y})}{\sqrt{S_p^2\left(\frac{1}{n}+\frac{1}{m}\right)}} < \frac{0.84-1.31}{\sqrt{0.21\left(\frac{1}{8}+\frac{1}{12}\right)}}\right) = P(t_{18} < -2.25).$$

Abbiamo allora come p-dei-dati $P(t_{18} < -2.25) = 0.019$. Adottando un livello di significatività maggiore o uguale a 0.02, come ad esempio $\bar{\alpha}=0.05$, possiamo rigettare l'ipotesi nulla: c'è evidenza statistica che la media del valore di funzionalità renale per la seconda popolazione sia più alto che nella prima.

17.5.4. Dati non appaiati, varianze ignote

Questo caso è molto più difficile e possiamo risolverlo usando gli strumenti a nostra disposizione solo nel caso di campioni molto grandi, in cui viene in nostro aiuto il Teorema Centrale del Limite, permettendoci di usare una funzione ancillare on distribuzione approssimativamente Gaussiana.

Non ci sorprende però che ci sia una via d'uscita in R, sempre grazie a `t.test`. Se infatti abbiamo due campioni e non specifichiamo che siano appaiati o a varianza uguale (quindi lasciamo `paired = FALSE` e `var.equal = FALSE`) otteniamo un test approssimato (con un'opportuna correzione) per permetterci di studiare questo caso.

17.6. CONFRONTO TRA PARAMETRI DI ALTRE POPOLAZIONI

Quando le dimensioni dei campioni sono sufficientemente grandi, possiamo usare i risultati già visti per il confronto tra due distribuzioni normali per confrontare i parametri di altre popolazioni, come quella Bernoulliana o Poissoniana. Tuttavia, quando la taglia del campione è ridotta, l'approssimazione data dal TLC è molto grezza, quindi i risultati sarebbero poco attendibili. Vediamo come è possibile procedere in questi casi con test (esatti) dedicati.

17.6.1. Bernoulli

In questo caso ci chiediamo se ci sia evidenza statistica che due popolazioni Bernoulliane abbiano parametro diverso, ossia che le probabilità di successo nell'una e nell'altra siano differenti. Per come abbiamo posto il problema stiamo confrontando un'ipotesi nulla della forma $H_0: p_1 = p_2$ con una alternativa della forma $H_1: p_1 > p_2$ o $H_1: p_1 \neq p_2$.

In generale avremo due campioni di numerosità n_1 e n_2 estratti dalle due popolazioni, che conterranno X_1 e X_2 successi. Rifiuteremo l'ipotesi nulla e accetteremo quella alternativa quando $\frac{X_1}{n_1}$ sarà molto diverso da $\frac{X_2}{n_2}$ (in una direzione specifica nel caso di test unilaterali, in qualunque direzione nel caso di test bilaterali).

Osserviamo che sotto l'ipotesi nulla, ossia nel caso in cui $p_1 = p_2 = p$, le due variabili aleatorie X_1 e X_2 sono distribuite come binomiali di parametri, rispettivamente, (n_1, p) e (n_2, p). Possiamo quindi considerare il numero totale di successi $X_1 + X_2$ sul numero totale di tentativi $n_1 + n_2$ ed esso avrà, per la riproducibilità delle binomiali, distribuzione binomiale di parametri $(n_1 + n_2, p)$. Supponiamo, sempre sotto ipotesi nulla, che $X_1 + X_2 = k$, allora la distribuzione di X_1 condizionata a questo evento sarà di tipo ipergeometrico di parametri (n_1, n_2, k). Infatti la situazione può essere descritta come una collezione di n_1 oggetti di un tipo e n_2 oggetti di un altro, da cui ne estraiamo k: la domanda diventa "quanti di questi sono del primo tipo?".

Allora se siamo nel caso $H_0: p_1 = p_2$ e $H_1: p_1 \neq p_2$, rifiuteremo l'ipotesi nulla se, considerando la variabile aleatoria $X \sim \text{hyper}(n_1, n_2, k)$ e osservando x_1 e x_2 si ha $P(X \leqslant x_1)$ molto piccola (caso in cui $p_1 < p_2$) o $P(X \geqslant x_1)$ molto piccola (caso in cui $p_1 > p_2$). In termini di regione critica abbiamo, a livello di significatività α, l'unione $[0, F_X^{-1}(\frac{\alpha}{2})] \cup [F_X^{-1}(1 - \frac{\alpha}{2}), n_1]$. Come p-dei-dati invece abbiamo $p = 2\min\{F_X(x_1), 1 - F_X(x_1 - 1)\}$.

Questo test ha vari nomi nella letteratura: test esatto di Fisher, test di Fisher-Irwin.

Esempio 17.21. TBA

Studi osservazionali [TBA]

17.6.2. Poisson

Abbiamo due variabili aleatorie di Poisson indipendenti, di parametri rispettivamente λ_1 e λ_2. Vogliamo valutare se, fissata una costante $c > 0$, ci sia evidenza statistica che $\lambda_2 \neq c\,\lambda_1$. In altre parole abbiamo come ipotesi nulla $H_0: \lambda_2 = c\,\lambda_1$ e come ipotesi alternativa $H_1: \lambda_2 \neq c\,\lambda_1$.

Per costruire il test usiamo una proprietà vista in precedenza (Esempio 8.40), ossia che date due Poissoniane, la distribuzione di una delle due nota la somma è una binomiale, in particolare date $X_1 \sim \text{Pois}(\lambda_1)$ e $X_2 \sim \text{Pois}(\lambda_2)$, la legge di $X_1 | X_1 + X_2 = n$ è $\text{bin}\left(n, \frac{\lambda_1}{\lambda_1 + \lambda_2}\right)$.

Nell'ambito del test avremo, sotto ipotesi nulla, che il secondo parametro della binomiale è $p_c = \frac{\lambda_1}{\lambda_1 + c\lambda_1} = \frac{1}{1+c}$. Questo ci permette di impostare la regione critica del test o di calcolare il p-dei-dati in maniera abbastanza semplice: se $X \sim \text{bin}\left(n, \frac{1}{1+c}\right)$ e osserviamo x_1 e $x_2 = n - x_1$ per le due popolazioni, allora rifiutiamo l'ipotesi nulla se $P(X \leqslant x_1)$ molto piccola o se $P(X \geqslant x_1)$ molto piccola, ossia se fissiamo significatività α, se $P(X \leqslant x_1) \leqslant \frac{\alpha}{2}$ o $P(X \geqslant x_1) \leqslant \frac{\alpha}{2}$, ossia se siamo nella regione critica $[0, F_X^{-1}(\frac{\alpha}{2})] \cup [F_X^{-1}(1 - \frac{\alpha}{2}), n]$. Come p-dei-dati invece abbiamo $p = 2\min\{F_X(x_1), 1 - F_X(x_1 - 1)\}$.

Rimane da capire come mai abbiamo introdotto la costante c. Siccome le Poisson sono riproducibili, la somma di Poisson è una Poisson di parametro la somma dei parametri. Se abbiamo k osservazioni dalla prima popolazione, possiamo vederle come un solo campione da una popolazione $\text{Pois}(k\lambda_a)$. Similmente per la seconda popolazione avremo $X_2 \sim \text{Pois}(h\lambda_b)$. Solitamente ci interessa sapere se le due popolazioni di partenza hanno il medesimo parametro, ossia se $\lambda_a = \lambda_b$. Tuttavia se $h \neq k$ mettendo assieme i campioni, confrontiamo cose un po' diverse. La c in questo caso misura proprio la proporzione delle due taglie dei campioni: $X_1 + X_2 \sim \text{Pois}((h+k)\lambda_a)$ e ponendo $\lambda_1 = h\lambda_a$, $\lambda_2 = k\lambda_a$, abbiamo che $\lambda_2 = \frac{k}{h}\lambda_1$.

Esempio 17.22. Ross TBA

17.7. PROBLEMI

Problema 70. Ricavare la curva operativa caratteristica per i test unilaterali della varianza in una popolazione normale. Usare i risultati ottenuti per approfondire la discussione dell'Esempio 17.11.

CAPITOLO 18

TEST CHI QUADRO

In questo capitolo parliamo ancora di test statistici, ma con un punto di vista un po' diverso rispetto a quanto visto nel Capitolo 17. Infatti vogliamo sfruttare quanto studiato finora per valutare alcune caratteristiche dei modelli probabilistici. Vorremo in particolare provare a rispondere a due domande:

1. Il modello probabilistico che abbiamo scelto descrive bene un fenomeno di cui abbiamo i dati? È *adatto* a descrivere questi dati?
2. Due variabili (o caratteristiche) di una popolazione sono effettivamente indipendenti tra loro?

In realtà lo strumento che usiamo per rispondere a entrambe queste domande è il medesimo, ossia il test del chi quadro.

18.1. TEST DI ADATTAMENTO

Il primo test che andiamo ad analizzare può essere applicato in condizioni abbastanza generali e permette di verificare se un insieme di dati è compatibile con un modello teorico.

Introduciamo il metodo con un esempio.

Esempio 18.1. Vogliamo indagare se un dado è truccato. Lo lanciamo 600 volte osservando che i numeri 1, 2, 3, 4, 5, 6 escono rispettivamente $n_1, n_2, n_3, n_4, n_5, n_6$ volte con

$$n_1 = 115, \quad n_2 = 97, \quad n_3 = 91, \quad n_4 = 101, \quad n_5 = 110, \quad n_6 = 86.$$

vogliamo testare l'ipotesi nulla H_0 secondo cui tutti i risultati sono equiprobabili con probabilità $p_i = 1/6$, per $i = 1, \dots, 6$ contro l'ipotesi alternativa H_1 che di fatto è definita come la negazione dell'ipotesi nulla. Stiamo quindi chiedendoci se ci sia evidenza statistica contro il bilanciamento del dado.

L'Esempio 18.1 illustra una situazione in cui abbiamo n esperimenti, che supponiamo indipendenti, che possono avere k possibili risultati con probabilità p_i, con $i = 1, \dots, k$. L'ipotesi che vogliamo testare è proprio la distribuzione dei possibili risultati, cioè il fatto che le k probabilità p_i assumano determinati valori che indichiamo con p_{0i}. Il test che andiamo a descrivere quindi permette di verificare l'ipotesi sulla distribuzione di una variabile casuale, anziché sul valore di un particolare parametro. L'ipotesi nulla è dunque della forma $H_0 = \{p_i = p_{0i}\}$, $i = 1, \dots, k$. Nel caso del dado (Esempio 18.1), $n = 600$, $k = 6$ e $p_{0i} = 1/6$, per $i = 1, .., 6$.

Per ognuno dei k possibili risultati, indicheremo con O_i il numero di esperimenti che hanno restituito il risultato i. Chiaramente O_i è una variabile casuale. Ne possiamo determinare la distribuzione, osservando che, se compiamo n esperimenti indipendenti e supponiamo vera l'ipotesi nulla $p_i = p_{0i}$ sarà una distribuzione binomiale $\text{bin}(n, p_{0i})$ e quindi $P(O_i = m) = \binom{n}{m} p_{0i}^m (1 - p_{0i})^{n-m}$. Per le proprietà della distribuzione binomiale, la media di O_i sarà $\mu_{O_i} = E[O_i] = n\, p_{0i}$. Notiamo quindi che, supponendo che l'ipotesi nulla sia vera, il numero atteso di esperimenti che danno come risultato i sarà pari a $n\, p_{0i}$. Indicheremo tale valore con il simbolo e_i.

Se il numero di prove n è abbastanza alto, La distribuzione di O_i sarà approssimabile con una distribuzione di Poisson, con uguale media e varianza $\sigma^2_{O_i} = np_{0i} = e_i$. Utilizzando il teorema limite centrale, potremmo approssimare la distribuzione di O_i con quella di una gaussiana con uguale media e varianza, quindi approssimativamente la variabile

$$\frac{O_i - \mu_{O_i}}{\sigma_{O_i}} = \frac{O_i - e_i}{\sqrt{e_i}}$$

avrà una distribuzione Gaussiana con media 0 e varianza 1. Lo stesso discorso è valido per tutte le variabili O_i indipendentemente da i.

Se andiamo a considerare la variabile $T = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$, possiamo dimostrare che si tratta una distribuzione chi quadrato, con $k-1$ gradi di libertà. L'intuizione è abbastanza immediata: tutti i termini sono gaussiane standard elevate al quadrato. Il motivo per cui abbiamo solamente $k-1$ gradi di libertà e non k è legato al vincolo sottinteso al nostro modello, ossia che $\sum_{i=1}^{k} O_i = n$ e quindi le O_i non sono tutte indipendenti tra loro.

Tornando a T, tale variabile dà una misura della bontà con cui i dati si adattano al modello descritto dall'ipotesi nulla: tanto più piccoli saranno i valori $(O_i - e_i)^2$ dei quadrati degli scarti tra valori attesi e valori osservati, tanto migliore sarà l'aderenza dei dati al modello.

Non dimentichiamo inoltre che per ottenere la distribuzione chi quadrato viene utilizzato il teorema limite centrale che consente di approssimare la distribuzione dei valori osservati con una Gaussiana, quindi per poter applicare tale test è necessario che i numeri in gioco siano "abbastanza grandi". Un criterio empirico che permette di discriminare tra il caso in cui l'approssimazione del limite centrale è valida e quella in cui non lo è prescrive di verificare che il numero n di esperimenti sia sufficientemente grande da avere $np_{0i} > 5$ per $i = 1, \dots, k$.

Vediamo come applicare questi risultato riprendendo l'Esempio 18.1. In questo caso $n = 600$, $k = 6$, $p_{0i} = 1/6$, quindi tutti i valori attesi e_i sono pari a $e_i = p_{0i} n = 100$. Se l'ipotesi nulla è vera, allora la variabile $T = \sum_{i=1}^{6} \frac{(O_i - 100)^2}{100}$ avrà una distribuzione chi-quadrato con 5 gradi di libertà. Il p-dei-dati cioè la probabilità di osservare valori "peggiori" di quelli osservati, è calcolabile come

$$P\left(\sum_{i=1}^{6} \frac{(O_i - 100)^2}{100} > \sum_{i=1}^{6} \frac{(o_i - 100)^2}{100} \right)$$

dove

$$o_1 = 115, \quad o_2 = 97, \quad o_3 = 91, \quad o_4 = 101, \quad o_5 = 110, \quad o_6 = 86.$$

Il valore di chi quadro che otteniamo dai dati è pari a

$$\frac{(115-100)^2 + (97-100)^2 + (91-100)^2 + (101-100)^2 + (110-100)^2 + (86-100)^2}{100} = 6.12.$$

Usando R o consultando la tavola del chi quadrato con 5 gradi di libertà possiamo concludere che il p-dei-dati è circa pari a 0.3. Dato abbiamo il 30% di probabilità di ottenere, se l'ipotesi nulla è vera, dati peggiori di quelli osservati non possiamo rigettare H_0: non abbiamo evidenza statistica per concludere che il dado è truccato.

18.2. TEST D'INDIPENDENZA

Nella sezione precedente abbiamo visto come un test d'ipotesi basato sulla variabile chi-quadrato permetta di verificare se un certo insieme di dati sia compatibile con una certa distribuzione. Vediamo ora come la stessa idea permetta di verificare l'indipendenza di due variabili discrete. Tale test risulta importante quando le variabili sotto esame non sono numeriche (come nell'esempio dell'altezza e della lunghezza dell'avambraccio degli individui di una certa popolazione) ma categoriche (ad esempio sesso e colore degli occhi degli individui di una certa popolazione). Tali variabili rivestono una particolare importanza nelle scienze della vita ed è importante saper capire se sono correlate.

Consideriamo il seguente esempio.

Esempio 18.2. (M. C. WHITLOCK, D. SCHLUTER) Il ciclo vitale di diversi parassiti necessita il passaggio del parassita stesso da un ospite di un certo tipo a quello di tipo differente[1]. Un particolare tipo di trematoide trascorre parte del suo ciclo vitale all'interno del pesce *fundulus parvipennis*. Per completare il ciclo il parassita ha la necessità che il pesce venga mangiato da un uccello, dove il trematode raggiungerà lo stadio di adulto maturo. Pare che il parassita, che si incista nella scatola cranica del pesce, sia in grado di modificare il comportamento di quest'ultimo rendendolo più facile alla predazione. Sembra infatti che i pesci infestati abbiano maggior tendenza a sostare in prossimità della superficie dell'acqua, dove gli uccelli possono vederli e catturarli.

Per testare se c'è correlazione tra le variabili X "grado di infestazione del pesce da parte del parassita" e Y "predazione da parte degli uccelli" è stata osservato un gruppo di 141 pesci situati in una grande vasca all'aperto e suddivisi in tre categorie: (1) non infestati, (2) lievemente infestati, (3) fortemente infestati. La vasca è stata lasciata accessibile a diversi tipi di uccelli e sono stati registrati i dati relativi al numero di pesci catturati a seconda del livello di predazione, organizzati nella Tabella 18.1

	(1)	(2)	(3)	totale righe
Mangiati	1	10	37	48
Non mangiati	49	35	9	93
Totale colonne	50	45	46	141

Tabella 18.1. Dati del problema

Possiamo concludere che la facilità con cui il pesce viene predato è collegata al suo livello di infestazione da parte del parassita?

Vediamo di formalizzare il problema in generale, guidandoci con l'esempio appena fatto. Indichiamo con X una variabile discreta che può assumere n possibili valori. Nell'Esempio 18.2 la variabile X indica il grado di infestazione dei parassiti. Notiamo che abbiamo suddiviso il grado di infestazione in tre categorie, etichettate con gli indici (1), (2) e (3) ma tali numeri non hanno significato e servono solo a denominare le categorie (avremmo equivalentemente potuto utilizzare le lettere A, B, C). Analogamente indichiamo con Y un'ulteriore variabile discreta che può assumere m valori. Nell'Esempio 18.2 Y assume due valori corrispondenti alle categorie "mangiati", $Y=1$, e "non mangiati", $Y=2$. Anche in questo caso i valori 1 e 2 servono solo ad etichettare le categorie (avremmo equivalentemente potuto utilizzare i simboli M e NM).

Indichiamo con $\varphi_{X,Y}$ la distribuzione congiunta delle variabili X e Y, definita da

$$\varphi_{X,Y}(i,j) = P(X=i, Y=j), \qquad i=1,\dots,n, \quad j=1,\dots,m$$

Per definizione di indipendenza, X e Y sono (statisticamente) indipendenti se per ognuna delle $n\,m$ coppie (i,j), con $i=1,\dots,n$, $j=1,\dots,m$, vale

$$\varphi_{X,Y}(i,j) = \varphi_X(i)\,\varphi_Y(j) \tag{18.1}$$

Dove

$$\varphi_X(i) = P(X=i) = \sum_{j=1}^{m} P(X=i, Y=j) = \sum_{j=1}^{m} \varphi_{X,Y}(i,j),$$

$$\varphi_Y(j) = P(Y=j) = \sum_{i=1}^{n} P(X=i, Y=j) = \sum_{i=1}^{n} \varphi_{X,Y}(i,j)$$

sono le distribuzioni marginali.

1. Ne è un esempio il toxoplasma, il cui sviluppo necessita del passaggio tra più tipi di ospiti, tra cui i topi e i gatti. Tale passaggio avviene tramite l'ingestione del topo infetto da parte del gatto. Ricordiamo che il toxoplasma può infettare anche l'uomo tramite l'ingestione da di alimenti contaminati da feci di gatti. Questo tipo di parassita risulta pericoloso per le donne in gravidanza in quanto può causare malformazioni nel nascituro. Sembra che i topi infestati dal toxoplasma siano più facile preda dei gatti. Pare infatti che il parassita riesca ia modificare il comportamento del topo che infesta rendendolo "meno timoroso".

Il test del chi-quadro va a testare l'ipotesi nulla di indipendenza delle variabili X e Y

$$H_0 = \{\varphi_{X,Y}(i,j) = \varphi_X(i)\,\varphi_Y(j),\ i = 1,\dots,n, j = 1,\dots,m\}$$

contro l'ipotesi alternativa in cui la condizione di indipendenza (18.1) non sia soddisfatta almeno per una coppia di valori (i,j), ossia

$$H_1 = \{\exists (i,j) : \varphi_{X,Y}(i,j) \neq \varphi_X(i)\,\varphi_Y(j)\}.$$

Il ragionamento che porta alla costruzione del test è analogo a quello che abbiamo fatto nella sezione precedente. Consideriamo un campione casuale semplice di numerosità campionaria pari a N, in cui i dati relativi alle variabili X e Y sono appaiati. Raccoglieremo quindi N coppie (x_k, y_k) con $k = 1,\dots,N$, e $x_k \in \{1,\dots,n\}$, $y_k \in \{1,\dots,m\}$ e indicheremo con o_{ij} il numero di coppie in cui $(x_k, y_k) = (i,j)$. Necessariamente tali valori soddisferanno il vincolo $\sum_{i=1}^{n}\sum_{j=1}^{m} o_{ij} = N$.

Di fatto i valori o_{ij} possono essere considerati come realizzazioni di $n \cdot m$ variabili casuali O_{ij}, con $i = 1,\dots,n$ e $j = 1,\dots,m$.

L'ipotesi nulla del test del chi-quadrato prevede una precisa distribuzione congiunta delle variabili X e Y data da $\varphi_{X,Y}(i,j) = p_0(i,j)$. Supponendo che tale ipotesi sia vera e guardando agli N elementi del campione casuale come a N esperimenti indipendenti, allora ognuna delle variabili O_{ij} avrà una distribuzione binomiale

$$P(O_{ij} = M) = \binom{N}{M} p_0(i,j)^M (1 - p_0(i,j))^{N-M}, \quad M = 0,\dots,N,$$

e per numerosità campionaria sufficientemente alta sarà approssimabile con una Gaussiana di media $\mu_{O_{ij}} = N p_0(i,j)$ e varianza $\sigma^2_{p_0(i,j)} = N p_0(i,j)$. Quindi

$$\frac{(O_{ij} - N p_0(i,j))^2}{N p_0(i,j)} \tag{18.2}$$

avrà una distribuzione chi quadrato con un grado di libertà.

L'ipotesi nulla prevede inoltre l'indipendenza delle variabili X e Y e quindi i valori $p_0(i,j)$ dovranno soddisfare la relazione $p_0(i,j) = p_0(i)\,p_0(j)$ e quindi la variabile (18.2) diviene

$$\frac{(O_{ij} - N p_0(i)\,p_0(j))^2}{N p_0(i)\,p_0(j)} \tag{18.3}$$

Notiamo subito una differenza rispetto al test del chi quadro utilizzato nel caso dell'Esempio 18.1 descritto nella sezione precedente. Nell'Esempio 18.1 infatti l'ipotesi nulla *specificava i valori esatti delle probabilità* $p_0(i,j)$ mentre ora, nel test d'indipendenza che stiamo per costruire, l'ipotesi nulla contiene solo le *relazioni* $p_0(i,j) = p_0(i)\,p_0(j)$ *che le probabilità* $p_0(i,j)$ *devono soddisfare*, ma non i loro precisi valori. Per poter effettuare il test dobbiamo sostituire i valori $p_0(i)$ e $p_0(j)$ con le loro stime $\hat{p}_0(i)$ e $\hat{p}_0(j)$ ottenute in base ai dati raccolti, date da

$$\hat{P}_0(i) = \frac{\sum_j O_{ij}}{N} \equiv \frac{O_i^X}{N}, \quad \hat{P}_0(j) = \frac{\sum_i O_{ij}}{N} \equiv \frac{O_j^Y}{N}$$

e le variabili che di fatto siamo in grado di studiare a partire dai dati raccolti sono

$$\frac{(O_{ij} - N\frac{O_i^X}{N}\frac{O_j^Y}{N})^2}{N\frac{O_i^X}{N}\frac{O_j^Y}{N}} = \frac{(O_{ij} - \frac{O_i^X O_j^Y}{N})^2}{\frac{O_i^X O_j^Y}{N}}. \tag{18.4}$$

Notiamo che la variabile O_i^X indica il numero di coppie in cui il valore della variabile X è pari a i, mentre O_j^Y indica il numero di coppie in cui il valore della variabile Y è pari a j.

Il test del chi quadrato per l'indipendenza di due variabili X e Y si basa sul seguente risultato: se è valida l'ipotesi nulla di indipendenza di X e Y allora la variabile

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(O_{ij} - \frac{O_i^X O_j^Y}{N})^2}{\frac{O_i^X O_j^Y}{N}} \tag{18.5}$$

ha una distribuzione chi-quadrato con $(n-1)(m-1)$ gradi di libertà.

Non dimostriamo questo risultato, notiamo solo che il numero di gradi di libertà $(n-1)(m-1)$ è inferiore al numero di termini nella sommatoria (pari a nm) in quanto, ancora una volta, tali addendi non sono fra loro indipendenti.

Vediamo ora come applicare tale risultato all'Esempio 18.2. In questo caso X è la variabile "grado di infestazione" mentre Y è la variabile "mortalità per predazione da parte degli uccelli" che hanno rispettivamente $n=3$ e $m=2$ possibili valori (o livelli o categorie). Dalla Tabella 18.1 possiamo dedurre che i valori osservati sono:

$$o_{11}=1, \quad o_{12}=49, \quad o_{21}=10, \quad o_{22}=35, \quad o_{31}=37, \quad o_{32}=9,$$

inoltre

$$o_1^X=50, \quad o_2^X=45, \quad o_3^X=46, \quad o_1^Y=48, \quad o_2^Y=93.$$

Da questi valori otteniamo

$$\begin{aligned}\chi^2_{oss} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(o_{ij}-\frac{o_i^X o_j^Y}{N}\right)^2}{\frac{o_i^X o_j^Y}{N}} = \frac{\left(o_{11}-\frac{o_1^X o_1^Y}{141}\right)^2}{\frac{o_1^X o_1^Y}{141}} + \frac{\left(o_{112}-\frac{o_1^X o_2^Y}{141}\right)^2}{\frac{o_1^X o_2^Y}{141}} + \\ &+ \frac{\left(o_{21}-\frac{o_2^X o_1^Y}{141}\right)^2}{\frac{o_2^X o_1^Y}{141}} + \frac{\left(o_{22}-\frac{o_2^X o_2^Y}{141}\right)^2}{\frac{o_2^X o_2^Y}{141}} + \frac{\left(o_{31}-\frac{o_3^X o_1^Y}{141}\right)^2}{\frac{o_3^X o_1^Y}{141}} + \frac{\left(o_{32}-\frac{o_3^X o_2^Y}{141}\right)^2}{\frac{o_3^X o_2^Y}{141}} = 69.5\end{aligned}$$

Andiamo ora a calcolare il p-dei-dati. Supponendo che sia vera l'ipotesi nulla calcoliamo la probabilità di ottenere per la variabile chi-quadro (18.5), che ha $(3-1)(2-1)=2$ gradi di libertà, valori maggiori di quelli osservati, cioè:

$$P\left(\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{\left(O_{ij}-\frac{O_i^X O_j^Y}{N}\right)^2}{\frac{O_i^X O_j^Y}{N}} > 69.5\right) = P(\chi_2 > 69.5)$$

Il valore di questa probabilità è bassissimo, inferiore a 10^{-4}. Possiamo quindi rigettare l'ipotesi nulla di indipendenza fra le variabili X e Y.

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[1].

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[2] (come mostrato nel Problema 3).

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[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[4]-BERNSTEIN[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 $\#A = \#B$.*

2. Eric Temple Bell (1883 – 1960).
3. Augustus De Morgan (1806 – 1871).
4. Georg Cantor (1845 – 1918).
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[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[7] i numeri reali in $[0,1]$.

6. Sono anche suriettive e in particolare l'una è l'inversa dell'altra.

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.

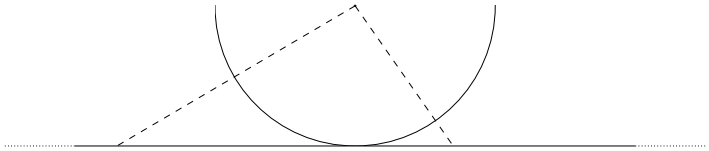


Figura A.1. Una biiezione tra $(0,1)$ e $\mathbb{R}$

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. TRIBÙ O σ-ALGEBRA?

Il[8] termine *σ-algebra* è nato nell'ambito delle teorie delle algebre booleane. Nel libro *Lectures on Boolean Algebras*, P. R. Halmos definisce un'*algebra booleana* come un insieme non vuoto, munito di due operazioni vinarie

$$(a,b) \mapsto a \wedge b, \qquad (a,b) \mapsto a \vee b$$

e da un'operazione a un sol argomento $a \mapsto a'$, godenti di opportune proprietà.

Egli definisce quindi un *campo d'insiemi* (*field of sets*) come una classe non vuota di parti di un insieme X, la quale sia stabile per le operazioni di intersezione binaria, di unione binaria e di passaggio al complementare rispetto all'intero insieme X. Osserva poi subito che un campo d'insiemi diventa una particolare algebra booleana quando si prendano in esso come operazioni booleane l'intersezione binaria, l'unione binaria e il passaggio al complementare. Ma si affretta ad aggiungere:

> *"This does not, however, justify the conclusion (it is false) that set-theoretic intersection, union and complement are the only possible operations that convert a class of sets into a Boolean algebra."*

8. Questo testo è tratto dalle note del prof. Giorgio Letta dell'Università di Pisa.

È proprio questo fatto che rende necessario introdurre una locuzione speciale (*field of sets*) per indicare quelle speciali algebre booleane per le quali, non soltanto è vero che gli elementi sono parti di un fissato insieme, ma è vero anche che le operazioni booleane $\wedge$ e $\vee$ coincidono con le operazioni insiemistiche di intersezione (binaria) e di unione (binaria) (e di conseguenza l'operazione booleana $'$ coincide con l'operazione insiemistica di passaggio al complementare).

Per analoga ragione, dopo aver introdotto le *σ-algebre* (*booleane*) (ossia quelle speciali algebre booleane nelle quali ogni insieme numerabile di elementi ammetta un estremo inferiore e un estremo superiore), Halmos introduce i *σ-campi* (*d'insiemi*), ossia quegli speciali campi (d'insiemi) che sono stabili per le operazioni di intersezione numerabile e di unione numerabile. I σ-campi sono speciali σ-algebre booleane: precisamente sono quelle speciali σ-algebre booleane per le quali, non soltanto è vero che gli elementi sono parti di un fissato insieme, ma è vero anche che le operazioni booleane di estremo inferiore e di estremo superiore numerabili coincidono con le operazioni insiemistiche di intersezione e di unione numerabili.

Bourbaki si è limitato a sostituire la locuzione *σ-field* (*of sets*) con una locuzione più semplice: *tribu*[9]. Ma è evidente la necessità di non confondere questa nozione con quella di σ-algebra (booleana). La confusione tra *campo* (*d'insiemi*) e *algebra* (*booleana*) è meno grave: infatti ogni algebra booleana è isomorfa a un campo d'insiemi. Ma l'enunciato analogo, ove si sostituisca algebra con σ-algebra e campo con σ-campo è falso!

A.3. 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.4. L'INTEGRALE GAUSSIANO

Ci interessano gli integrali definiti

$$I_1 = \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x \qquad I_2 = \int_{0}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x \qquad I = \int_{0}^{+\infty} \mathrm{e}^{-x^2}\,\mathrm{d}x$$

Osserviamo che $I_1 = 2\,I_2$ 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

$$f(t) = \int_0^t \mathrm{e}^{-x^2}\,\mathrm{d}x \qquad g(t) = \int_0^1 \frac{\mathrm{e}^{-t^2(1+x^2)}}{1+x^2}\,\mathrm{d}x$$

9. Si osservi che a questa parola francese basta aggiungere un accento per ottenere la corrispondente parola italiana: *tribù*.

e osserviamo che

$$(f(t))^2 + g(t) = c,$$

per qualche costante c, infatti

$$\frac{\mathrm{d}}{\mathrm{d}t}(f(t))^2 = 2f(t)f'(t) = 2\int_0^t \mathrm{e}^{-x^2}\mathrm{d}x\,\mathrm{e}^{-t^2} = 2\int_0^t \mathrm{e}^{-x^2-t^2}\mathrm{d}x$$

ma, allo stesso tempo, se facciamo il cambio di variabili $x = yt$,

$$2\int_0^t \mathrm{e}^{-x^2-t^2}\mathrm{d}x = \int_0^1 2\mathrm{e}^{-t^2(y^2+1)}\,t\,\mathrm{d}y.$$

D'altro canto,

$$g'(t) = \int_0^1 \frac{\mathrm{e}^{-t^2(1+x^2)}(-2t)(1+x^2)}{1+x^2}\mathrm{d}x = -\int_0^1 2\mathrm{e}^{-t^2(1+x^2)}\mathrm{d}x.$$

Quindi per ogni t, $(f(t))^2 + g(t) = (f(0))^2 + g(0)$ e quindi

$$\begin{aligned} I^2 &= \lim_{t\to+\infty}(f(t))^2 = g(0) - \lim_{t\to+\infty} g(t) \\ &= \int_0^1 \frac{1}{1+x^2}\mathrm{d}x - \lim_{t\to+\infty}\int_0^1 \frac{\mathrm{e}^{-t^2(1+x^2)}}{1+x^2}\mathrm{d}x \\ &= \Big[\operatorname{atan}(x)\Big]_0^1 - 0 \\ &= \frac{\pi}{4} \end{aligned}$$

da cui abbiamo $I = \sqrt{\pi}/2$.

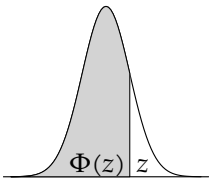
[TBC] Altri modi per calcolarlo si possono trovare in questo documento: https://kconrad.math.uconn.edu/blurbs/analysis/gaussianintegral.pdf.

APPENDICE B

TAVOLE

Tavole della funzione di ripartizione per una distribuzione normale standard

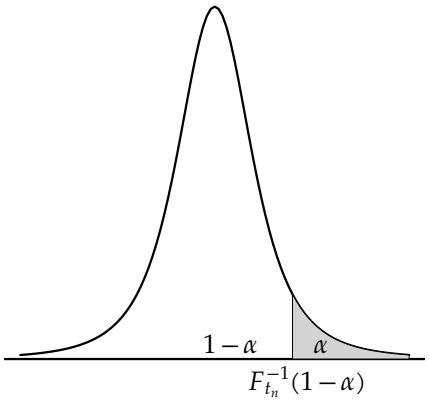
$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$



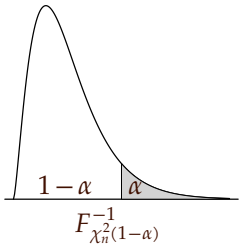
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	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 $\Phi(-0.78) \simeq 0.2177$.

APPENDICE C

ESERCIZI ULTERIORI

In queste pagine ci sono esercizi che potrebbero essere in un esame. Chiaramente l'esame non sarà necessariamente di questa forma, con questo numero di esercizi eccetera. In particolare val la pena notare che le prime proposte non contengono esercizi di Statistica.

C.1. UN POSSIBILE ESAME (PROBABILITÀ)

Problema 71. Siano X e Y due variabili aleatorie esponenziali indipendenti di medie rispettivamente uguali a 1 e 0.2857143.

1. Determinare la seguente probabilità: $P(X < 0.57 \leq Y)$.
2. Determinare la seguente probabilità: $P(3.5 \cdot Y \leq X)$.
3. Determinare la legge di Y/X.

Problema 72. In quanti modi diversi è possibile scrivere 23 come somma di 8 interi positivi (tenendo conto dell'ordine: $2+1$ e $1+2$ sono due modi diversi di scrivere 3 come somma di due interi positivi)?

Problema 73. Due amici giocano al seguente gioco: lanciano una moneta 11 volte e contano il numero di teste. Per ogni testa, estraggono (senza sostituzione) da un'urna contenente biglie numerate da 1 a 90.

1. Se la probabilità che la moneta esca testa è 0.55, qual è la probabilità che venga estratto il numero 41?
2. Se vengono estratte 3 biglie, qual è il valore più probabile per la probabilità che la moneta esca testa?
3. Se vengono estratte 5 biglie, qual è la probabilità che vengano estratte in ordine decrescente?

Problema 74. Il server di proprietà di un gruppo di ricerca impiega un tempo medio di 3.53 giorni (con deviazione standard 1.38) per eseguire una simulazione particolarmente difficile.

1. Se il gruppo di ricerca vuole eseguire simulazioni per 41 differenti dati iniziali, quanto tempo ci vorrà (in media) per eseguirle tutte?
2. Qual è la probabilità che siano necessari più di 144 giorni per eseguire tutte le 41 simulazioni?

C.2. UN POSSIBILE ESAME

Problema 75. Un'urna contiene 10 biglie nere e 7 bianche.

1. Se estraiamo 15 biglie *senza reinserimento*, qual è la probabilità di vedere 9 biglie nere e 6 bianche, sapendo che la nona biglia estratta è nera?
2. Se estraiamo 15 biglie *con reinserimento*, qual è la probabilità di vedere 9 biglie nere e 6 bianche, sapendo che la nona biglia estratta è nera?

Problema 76. Sono date 8 variabili aleatorie indipendenti e identicamente distribuite $X_1, \dots, X_8$, con funzione densità di probabilità $f(x) = 1$ in $[0,1]$ e 0 altrimenti. Sia $M = \min\{X_1, \dots, X_8\}$.

1. Qual è la probabilità che $M \geq 0.672$?
2. Qual è il valore atteso di M?

Problema 77. La funzione di densità congiunta di due variabili aleatorie (assolutamente continue) X e Y è $f_{X,Y}(x,y) = \alpha\, x\, y$ per $0 \leq x \leq 1$ e $0 \leq y \leq \sqrt{x}$ e 0 altrimenti.

1. Determina il valore di α.
2. Determina se X e Y sono indipendenti o meno.
3. Calcola la funzione di densità condizionata per X dato $Y = y$.

C.3. PROVA IN ITINERE (2024/04/15)

Problema 78. La variabile aleatoria X ha come funzione di densità

$$f_X(x) = \begin{cases} c\,x\,(1-x) & 0 \leqslant x \leqslant 1 \\ 0 & \text{altrimenti.} \end{cases}$$

Determina c. Calcola media, mediana, moda e varianza di X e traccia il grafico (qualitativo) della funzione di ripartizione. Sia $Y = 2X - 1$. Calcola media e varianza di Y. Scrivi la funzione di densità di Y e tracciane il grafico (qualitativo).

Problema 79. Quello che segue è un problema tratto da un esame passato, con la risposta data da una persona che ha preso parte. Controlla se la risposta è corretta e, nel caso non fosse così, indica tutti gli errori che hai individuato nel ragionamento. Fornisci anche una soluzione corretta che superi i limiti che hai trovato. I valori calcolati sono corretti (nel senso che gli errori, se presenti, non sono nei conti).

> *Problema:* In una pasticceria ci sono 7 cannoncini alla crema, 6 allo zabaglione e 5 alla nutella. Bisogna preparare un vassoio con 7 paste. Se i cannoncini sono scelti uniformemente a caso, qual è la probabilità che nel vassoio ci siano almeno 2 cannoncini per tipo?
>
> *Soluzione:* Ci sono 7 cannoncini alla crema C, 6 allo zabaglione Z, e 5 alla nutella N, quindi in tutto abbiamo 18 cannoncini. Per lo spazio di probabilità scegliamo come Ω l'insieme di tutte le possibili cinquine di cannoncini, come tribù l'insieme delle parti, visto che l'insieme Ω è finito, come probabilità quella uniforme discreta, classica che dà la medesima probabilità a ogni quintupla.
>
> Cominciamo con il conteggio dei casi totali: abbiamo $\binom{18}{7} = 31824$ modi di scegliere 7 cannoncini tra 18.
>
> Per i casi favorevoli, abbiamo $\binom{6}{2} = 15$ modi di scegliere le due Z, $\binom{7}{2} = 21$ modi di scegliere le due C, $\binom{5}{2} = 10$ modi di scegliere le due N e $\binom{12}{1} = 12$ modi di scegliere il rimanente cannoncino, dal momento che può essere di uno qualunque dei tre tipi e che ne sono rimasti $(6-2) + (7-2) + (5-2) = 12$.
>
> La probabilità richiesta è allora
>
> $$\frac{\#F}{\#T} = \frac{\binom{7}{2} \cdot \binom{6}{2} \cdot \binom{5}{2} \cdot \binom{12}{1}}{\binom{18}{7}} = \frac{37800}{31824}.$$

Problema 80. In una versione semplificata del Monopoli, ogni turno si tirano due dadi e si avanza di un numero di caselle uguale alla somma dei risultati dei due dadi. Inoltre, se i dadi hanno lo stesso risultato, si tirano nuovamente i dadi e si avanza ancora (continuando a tirare e avanzare senza limiti a priori, finché i dadi non danno punteggi diversi). Di quanto si avanza mediamente in un turno?

Problema 81. In un recente festival di tecnologia e intrattenimento, sono state installate due diverse stazioni che distribuiscono automaticamente gadget. La prima stazione, *Alpha*, dà gadget seguendo una distribuzione di Poisson di media 5 gadget per visitatore. La seconda stazione, *Beta*, offre premi "esclusivi", e dà gadget secondo lo stesso tipo di distribuzione, ma con media 2 gadget per visitatore. I visitatori scelgono la stazione *Alpha* con probabilità $p = 0.8$, altrimenti la stazione *Beta*.

1. Determina la distribuzione del numero di gadget dati per visitatore e calcolarne la media.
2. Sapendo che un visitatore ha ricevuto esattamente 3 gadget, qual è la probabilità che essi siano stati distribuiti dalla stazione *Beta*?

C.4. PROVA IN ITINERE (2024/06/06)

Problema 82. Un quadrato Q è inscritto in una circonferenza C. Sia p la probabilità che un punto preso uniformemente a caso nel cerchio delimitato da C sia strettamente interno a Q.

1. Qual è il valore di p?
2. Qual è la probabilità che, scegliendo uniformemente a caso e indipendentemente tra loro 10 punti nel cerchio delimitato da C almeno 3 non giacciano all'interno di Q? (Nel caso il risultato dipenda da p e tu non lo abbia calcolato in 1., puoi lasciare indicato il parametro p.)
3. Sia x_0 un punto sulla *circonferenza* C fissato. Scegliamo uniformemente a caso un altro punto sulla circonferenza C. Qual è la lunghezza media della corda che unisce questi due punti? (Nel caso il risultato dipenda da p e tu non lo abbia calcolato in 1., puoi lasciare indicato il parametro p. Può essere utile passare alle coordinate polari e ricordare l'identità $\frac{1-\cos x}{2} = \sin^2(\frac{x}{2})$.)

Problema 83. Quello che segue è un problema tratto da un esame passato, con la risposta data da una persona che ha preso parte. Controlla se la risposta è corretta e, nel caso non fosse così, indica tutti gli errori che hai individuato nel ragionamento. Fornisci anche una soluzione corretta che superi i limiti che hai trovato.

> *Problema:* È data una sequenza di variabili aleatorie iid di media 1 e varianza 1.
>
> 1. Calcolare la probabilità che la somma delle prime 12 variabili della sequenza sia strettamente maggiore di 15, supponendo che siano **discrete**.
> 2. Calcolare la probabilità che la somma delle prime 12 variabili della sequenza sia strettamente maggiore di 15, supponendo che siano tutte esponenziali.
> 3. Vengono osservati i seguenti dati: $(12, 16, 19, 12, 20, 5, 11, 10, 8, 13)$, di cui il primo è la somma delle prime 12 variabili (realizzate) della sequenza, il secondo la somma delle successive 12 e così via. Calcolare un intervallo di confidenza al 90% per la media.
>
> *Soluzione:* Nella prima domanda dobbiamo prestare attenzione al fatto che le variabili sono discrete, quindi il fatto che sia maggiore e non maggiore o uguale è rilevante. Ciò premesso, non avendo altre informazioni usiamo il Teorema Centrale del Limite, sfruttando le informazioni sulla media e sulla varianza
>
> $$P(S_{12} > 15) = 1 - P(S_{12} \leqslant 15) \simeq 1 - \Phi\left(\frac{15 - 12 \cdot 1}{\sqrt{12 \cdot 1}}\right) \approx 19.3\%.$$
>
> Per quanto riguarda la seconda domanda non abbiamo bisogno di usare il TLC dal momento che le variabili esponenziali sono riproducibili, di parametro uguale alla somma dei parametri quindi la probabilità richiesta può essere calcolata in R (ricordando che le esponenziali sono continue, quindi $\geqslant 15$ o > 15 è uguale) come segue: `pexp(15, rate = 12, lower.tail= FALSE)` che dà un risultato piccolissimo, prossimo a 0.

Per l'intervallo di confidenza usiamo di nuovo il TLC per approssimare con una Gaussiana. Usiamo una statistica normale perché è nota la varianza, anche se il campione è piccolo (10 osservazioni). La media campionaria puntuale è 12.6, ottenuta con R `sum(c(12,16,19,12,20,5,11,10,8,13))/10`. L'intervallo di confidenza è $12.6 \pm \Phi^{-1}(0.95)\frac{1}{10} = (12.44, 12.76)$

Problema 84. Il dataset `iris` presente in ogni installazione di R contiene misurazioni di diverse specie di fiori. L'obiettivo dell'esercizio è quello di analizzare le lunghezze del sepalo per due specie di iris, `setosa` e `versicolor`, delle tre presenti nel dataset.

1. Confrontare (anche graficamente) le distribuzioni della variabile `Sepal.Length` per le due specie, spiegando quanto fatto e commentando i risultati ottenuti.
2. C'è evidenza statistica che le lunghezze medie dei sepali nelle due specie considerate siano differenti? Spiegare il procedimento seguito e commentare eventuali ipotesi fatte.

C.5. APPELLO D'ESAME (2024/06/24)

Problema 85. Un laboratorio deve valutare il contenuto di caffeina nel tè verde di una certa marca. Per fare questo, è stato considerato un campione di 20 bustine di tè e misurato il contenuto di caffeina in ciascuna bustina. I valori risultanti dalle analisi, in milligrammi per bustina, sono i seguenti:

$$35, 40, 38, 37, 39, 36, 35, 40, 38, 41, 34, 35, 37, 36, 40, 39, 37, 33, 36, 42.$$

Le leggi in vigore vietano la commercializzazione del tè per valori medi superiori a 39 milligrammi per bustina, quindi il dipartimento legale dell'azienda chiede un'analisi statisica sui dati sopra individuati che risponda ad alcuni quesiti.

1. Qual è una stima (con il 95% di fiducia) della probabilità che una bustina scelta a caso nell'intera produzione abbia contenuto di caffeina oltre la soglia legale?
2. C'è evidenza statistica che il prodotto sia a norma di legge?
3. Osservando altrettanti dati (estratti dalla medesima popolazione) e considerando il campione complessivo ottenuto unendo i dati nuovi ai precedenti, come cambieranno le risposte alle due domande sopra?

Problema 86. Sia X una variabile aleatoria continua con densità $f(t) = c$ per $4 < t < 8$.

1. Dopo aver determinato c, calcolare media, deviazione standard, moda e mediana di X e la probabilità $P(X > 5 | X < 6.5)$.
2. Ricavare legge, media e mediana della variabile aleatoria $Y := |X - 5|$.
3. Trovare una trasformazione $g\colon \mathbb{R} \to \mathbb{R}$ tale che $Z = g(X)$ sia una normale standard.

Problema 87. Un edicolante vende un quotidiano a due tipi di clienti: quelli con abbonamento e gli altri. Ogni giorno il numero di copie che vengono chieste dai primi è una variabile aleatoria X che può assumere, con la medesima probabilità, i valori 0, 1, 2, 3 e 4. Il numero Y di copie chieste dai secondi ha invece distribuzione di Poisson di media 2.5. È possibile assumere le due variabili indipendenti.

1. Sia $S = X + Y$ il numero totale di copie vendute. Calcolare media e varianza di S e il valore della funzione di massa di probabilità di S nei punti 0, 2 e 5.
2. Questa mattina l'edicolante ha ricevuto solamente 7 copie. Con che probabilità non saranno sufficienti?
3. Venerdì scorso ha venduto 5 copie. Com'è la distribuzione della coppia X, Y, conoscendo questa informazione?

C.6. APPELLO D'ESAME (2024/07/15)

Problema 88. Di una variabile aleatoria X sappiamo che la funzione di densità è nulla sui reali negativi e ha la forma $f(x) = \frac{c}{(1+x)^4}$ per $x \geqslant 0$.

1. Determinare i possibili valori di c.
2. Calcolare moda e mediana di X.
3. Determinare la legge di $Y = X + 1$ e di $Z = aX$. Le variabili aleatorie Y e Z sono tra loro indipendenti?

Problema 89. La dirigenza di una fabbrica di cioccolato, nonostante le pressioni del temibile sindacato OLU, non vuole introdurre la settimana lavorativa ridotta, sostenendo che la maggior parte di chi lavora non desidera un tale cambiamento. Per risolvere la questione viene effettuato un sondaggio su una piccola parte della forza lavoro: a 380 dipendenti viene chiesto se siano favorevoli o meno alla settimana corta. Delle risposte, 191 sono favorevoli alla nuova politica, le altre contrarie.

Sia la dirigenza, sia il sindacato convocano una conferenza stampa ed entrambe le parti dichiarano che i dati mostrano, a un livello di significatività del 5%, che la maggior parte di chi lavora nella fabbrica è dalla loro parte.

1. Quale delle due parti ha ragione da un punto di vista statistico?
2. Determinare il numero minimo di dipendenti favorevoli affinché ci sia evidenza statistica che la forza lavoro sia favorevole alla settimana corta.
3. Determinare il numero massimo di dipendenti favorevoli affinché ci sia evidenza statistica che la forza lavoro non sia favorevole alla settimana corta.
4. Commentare i risultati ottenuti.

Problema 90. La professoressa Mari riceve, in media, 4 email al giorno. Indicando con i numeri interi positivi $k \in \{1, \dots, 7\}$ i giorni della settimana, la variabile aleatoria X_k rappresenta il numero di email ricevute nel giorno k. Assumiamo che le variabili X_k siano indipendenti e identicamente distribuite secondo una legge di Poisson.

1. Le ipotesi fatte sulle X_k sono ragionevoli?
2. Sapendo che il terzo giorno della settimana ($k = 3$) ha ricevuto almeno una mail, qual è la probabilità che ne riceva esattamente altre 2 nello stesso giorno?
3. Qual è il valore atteso di email ricevute nel corso di una settimana?
4. Sapendo che nel weekend (giorni 6 e 7) ha ricevuto complessivamente $n \in \mathbb{N}$ email, qual è il valore atteso del numero di email ricevute la domenica?
5. Qual è il coefficiente di correlazione tra le variabili aleatorie che contano le mail ricevute sabato e domenica e quelle ricevute domenica e lunedì?

C.7. APPELLO D'ESAME (2024/08/29)

Problema 91. Una strada in uscita dalla Città Senza Nome (CSN) è tristemente nota per la frequenza degli incendi che si sviluppano ai suoi margini (anche a causa dell'incuria dei guidatori fumatori). La CSN decide pertanto di costruire una stazione dei vigili del fuoco sulla strada, ma vuole farlo in maniera ottimale.

Dai dati raccolti negli anni si ipotizza che nel tratto di competenza della CSN della strada (lungo L leghe) la densità di probabilità di incendi sia minima all'inizio della strada e cresca linearmente con coefficiente angolare 0.5 fino al termine.

1. Quanto vale il minimo h della densità di probabilità?
2. Quanto è lungo al più il tratto di competenza della CSN della strada?
3. A che punto deve essere costruita la stazione se vogliamo che ci sia la medesima probabilità che i vigili debbano intervenire in ciascuna delle due direzioni (verso la città e verso i confini comunali)?

4. Uscita dal territorio comunale la strada continua all'infinito nel territorio del Paese Senza Nome (PSN). Anche il PSN decide di costruire una stazione dei vigili del fuoco lungo la parte di propria competenza (ossia la parte non "coperta" dalla CSN). Dai dati raccolti la densità di probabilità degli incendi è, lungo la strada a partire dai confini comunali della CSN) un'esponenziale di parametro $\lambda = \frac{1}{4}$. Dove deve costruire il PSN la centrale dei vigili del fuoco se vuole minimizzare la distanza media percorsa dai mezzi per intervenire a spegnere un incendio?
5. Credi che sia meglio utilizzare la distanza media o la misura usata al punto 3 nella scelta della posizione della stazione? (Giustifica matematicamente la tua scelta di modello.)

Problema 92. Il corso di Fisica 1 tenuto dal temuto professor Tassi è frequentato da 70 studenti. Ogni volta che c'è un ricevimento, ciascuno studente e ciascuna studentessa decide di approfittare di questa occasione con una probabilità uguale a 0.01, indipendentemente dalle altre persone.

1. In media quante persone si presentano a un dato ricevimento?
2. Con che probabilità non si presenta nessuno a un dato ricevimento?
3. Durante l'anno il professor Tassi offre 50 ricevimenti. Se indichiamo con V i ricevimenti in cui non si presenta alcuna persona, quali sono valore atteso e varianza di V?
4. Stimare la probabilità che in almeno 30 ricevimenti non si presenti alcuna persona spiegando le ipotesi fatte e confrontando con il risultato "esatto" ottenuto in R.

Problema 93. Un ingegnere automobilistico sospetta che il consumo medio di carburante di certe auto sia diverso da quello dichiarato (pari a 17 mpg, miglia per gallone). I dati sono raccolti nel dataset `mtcars` disponibile in R. Rispondere alle seguenti richieste, esplicitando eventuali ipotesi necessarie e discutendo se sono soddisfatte o meno.

1. Dare una stima puntuale del consumo medio.
2. Dare una stima intervallare a un livello di fiducia del 98% del consumo medio.
3. Impostare un opportuno test statistico al 2% di significatività, commentando le scelte fatte e i risultati ottenuti.

Soluzione. Per quanto riguarda il primo punto la richiesta è semplicemente di dare una stima puntuale della media, che non richiede ipotesi sulla popolazione, se non l'indipendenza degli elementi del campione. Abbiamo

```
data(mtcars)
summary(mtcars)
mean(mtcars$mpg)
```

in cui già nell'output della funzione summary abbiamo il valore della media, 20.09062.

Per calcolare l'intervallo di fiducia dobbiamo controllare che siano soddisfatte (almeno approssimativamente) le ipotesi di normalità della popolazione:

```
qqnorm(mtcars$mpg)
qqline(mtcars$mpg, col = "red")
hist(mtcars$mpg)
```

La popolazione non è propriamente normale, anche per il numero relativamente ridotto di elementi che costituiscono il campione, tuttavia la possiamo considerare approssimativamente normale. Per il calcolo dell'intervallo di fiducia possiamo scrivere del codice dedicato, usando la statistica t dal momento che non è nota la varianza, di cui possiamo solamente avere una stima campionaria. Siccome nel punto successivo vogliamo fare un test d'ipotesi a sua volta basato sulla statistica t, possiamo fare entrambe le cose con un'unica funzione: `t.test(mtcars$mpg, mu = 17, conf.level = 0.98, alternative = "two.sided")` che ci restituisce quanto segue

```
One Sample t-test

data:  mtcars$mpg
t = 2.9008, df = 31, p-value = 0.006788
alternative hypothesis: true mean is not equal to 17
98 percent confidence interval:
 17.47733 22.70392
sample estimates:
mean of x
 20.09062
```

Nel leggere i risultati vediamo che l'intervallo richiesto è $(17.47733, 22.70392)$ e che dato il p-dei-dati pari a 0.006788 possiamo rifiutare l'ipotesi nulla che la media sia uguale a 17 e accettare l'alternativa che sia differente.

In questo test d'ipotesi l'alternativa è simmetrica, ma avendo visto che la media campionaria è circa 20 avremmo forse potuto impostare un test che mostrasse come la media sia maggiore di quanto dichiarato, ossia con ipotesi alternativa $\mu > \mu_0 = 17$. In questo caso il codice non è molto diverso, `t.test(mtcars$mpg, mu = 17, conf.level = 0.98, alternative = "greater")` e come p-dei-dati abbiamo 0.003394 che ci spinge anche in questo caso a rigettare l'ipotesi nulla e a sostenere l'alternativa, ossia che la media sia significativamente maggiore di 17.

C.8. PROVA IN ITINERE (2025/04/17)

Problema 94. Ogni persona ha, per quanto riguarda il colore degli occhi, un aspetto genetico (genotipo) e un aspetto fenomenologico (fenotipo). I due sono legati tra loro.

I genotipi associati al colore degli occhi sono $\{MM, AA, MA, AM\}$. Si sviluppa il fenotipo A (i.e. si hanno gli occhi azzurri) se e solo se si ha il genotipo AA. Si sviluppa il fenotipo M (occhi marroni) se e solo se si ha uno tra i genotipi $\{MA, AM, MM\}$.

Secondo le leggi di Mendel, il genotipo della prole (biologica) di due individui è equamente distribuito tra le possibili quattro combinazioni $(X_i Y_j)$, dove X_i e Y_i rappresentano, rispettivamente, il primo e il secondo elemento del genotipo del genitore G_i $(i \in \{1,2\})$. A titolo di esempio, si considerino le seguenti tabelle.

	M	M
A	AM	AM
M	MM	MM

	M	A
A	AM	AA
M	MM	MA

Tabella C.1. Sinistra: possibili esiti per genitori con genotipi AM e MM. Il figlio avrà genotipo AM con probabilità $\frac{2}{4}$, mentre avrà genotipo MM con probabilità $\frac{2}{4}$. Destra: possibili esiti per genitori con genotipi AM e MA. Il figlio avrà genotipo AM, AA, MM, MA con probabilità $\frac{1}{4}$.

Supponiamo di vivere in una popolazione omogenea, ovvero che $P(AA) = P(AM) = P(MA) = P(MM) = 1/4$, per ogni individuo della popolazione.

Denotiamo G_1 e G_2 i genitori.

1. Se un individuo ha gli occhi azzurri e G_1 ha gli occhi azzurri, qual è la probabilità che anche G_2 abbia gli occhi azzurri?
2. Se un individuo ha gli occhi marroni e G_1 ha gli occhi azzurri, qual è la probabilità che G_2 abbia gli occhi marroni?
3. Se un individuo ha gli occhi azzurri, qual è la probabilità che almeno uno dei due genitori abbia gli occhi marroni?

Problema 95. Sia $f_X\colon \mathbb{R} \to \mathbb{R}$ definita come

$$f_X(x) := \begin{cases} c\,(x^2 + \alpha x) & \text{se } x \in [0,1], \\ 0 & \text{altrimenti,} \end{cases}$$

ove $c \geq 0$ e $\alpha \in \mathbb{R}$.

1. Per quali valori dei parametri α, c risulta che f_X è una densità di probabilità?
2. Sia X la variabile aleatoria con densità f_X. Determina, se esistono, i valori dei parametri α e c per cui la media di X vale 0.7.
3. Utilizzando i valori ottenuti al punto 2., definiamo $Y := X^2$. Determina la funzione densità f_Y e la funzione di ripartizione F_Y di Y.
4. Quanto valgono $P\big(Y \geq \frac{\pi}{3}\big)$ e $P\big(Y \in [0, \frac{1}{2}]\big)$?

Problema 96. In una fabbrica di graffette vengono usati due macchinari diversi. Il primo produce 1000 graffette all'ora, mentre il secondo le produce a un ritmo variabile. Entrambi i macchinari producono in media ogni ora 50 graffette difettose. Chiamiamo X e Y le variabili aleatorie che descrivono il numero di graffette difettose prodotte dal primo e dal secondo macchinario, rispettivamente, nell'arco di una certa ora.

1. Che distribuzione possiamo ipotizzare per X e Y? Come mai?
2. Sia Z il numero totale di graffette difettose prodotte in un'ora. Calcolane il momento primo e momento centrato secondo.
3. Ogni giorno un'ora è dedicata al controllo qualità: tutte le graffette prodotte da entrambi i macchinari vengono controllate. Il conteggio determina che ci sono 160 graffette difettose. Quanto è sorprendente questo numero?