APPUNTI DI CALCOLO DELLE PROBABILITÀ

DARIO TREVISAN

Indice

Appunti del corso CPS 269AA A.A. 2017-2018, CdL in Informatica. Vi prego di segnalare errori di battitura, punti poco chiari ecc., via e-mail a dario.trevisan@unipi.it.

1. Aspetti qualitativi della probabilità

La probabilità è una teoria matematica che si occupa di concetti quali *aleatorietà*, *incertezza*, *plausibilità*. Prima di studiarne gli aspetti quantitativi, ossia il calcolo, è importante capire le caratteristiche qualitative delle situazioni che vogliamo descrivere servendoci in questa teoria. Questo sarà utile per evitare di cadere in problemi come falsi paradossi e veri errori concettuali, principalmente dovuti ad un atteggiamento che tende ad attribuire un significato fisico (ossia, *oggettivo*) alle probabilità.

Cominciamo col riflettere sui seguenti termini del linguaggio comune:

- *casuale*: fatto o accaduto senza metodo o decisione cosciente;
- *aleatorio*: che dipende dal lancio di un dado (*alea* in latino) o dal caso;
- *incerto*: non conosciuto o definito, inaffidabile.
- *plausibile*: che sembra ragionevolmente vero.

La teoria della probabilità ci permetterà di tradurre in numeri e quindi manipolare secondo delle regole determinate ed efficaci, la "quantità" di casualità (il grado di incertezza o plausibilità), che chiameremo appunto *probabilità*.

1.1. **I motivi dell'incertezza.**

Consideriamo i seguenti esempi e consideriamo i motivi per cui vi è incertezza:

(1) Siete incerti se domani pioverà nella località in cui vi trovate ora (evento futuro);
(2) Siete incerti se 100 anni fa pioveva nella località in cui vi trovate ora (evento passato); ma siete certi se ieri pioveva o no;
(3) Appena svegliati, con le finestre chiuse, siete incerti se fuori piove (evento presente); ma basta guardare fuori dalla finestra per capirlo.

Siamo quindi portati a ritenere che la distanza nel tempo tra osservatore (noi) e un evento (la pioggia) non è necessariamente un motivo di incertezza, ma può contribuire. In modo simile, la separazione spaziale non è necessariamente una causa di incertezza, ad esempio se vogliamo sapere il tempo in una località lontana, basta telefonare a qualcuno che si trova lì e chiedere (oppure consultare un sito web con il meteo in tempo reale).

Un altro aspetto dell'incertezza che appare evidente da esempi è che è *soggettiva*, ossia qualcosa potrebbe essere incerto per qualcuno ma essere una ovvietà (vera o falsa) per qualcun altro.

(4) Appena svegliati con le finestre chiuse, siamo incerti se fuori piove ma chiunque si trovi fuori casa e stia già andando al lavoro sa bene se piove o no.
(5) Uno sviluppatore di software è incerto se un utente userà o meno certi comandi della applicazione sviluppata, ma l'utente (sperabilmente) lo sa.
(6) Se una persona sceglie un numero da 0 a 9 e non ve lo comunica, questa sicuramente lo conosce ma voi siete incerti.

Ci sono, ovviamente, situazioni su cui tutti noi siamo incerti, ma questo di per sé non le rende speciali – possiamo immaginare qualche soggetto, in qualche situazione (futura o passata), per cui questo incertezza scomparirebbe.

(7) Prima di lanciare un dado, l'esito è incerto a tutti, a meno che non stia truccando il tiro. Dopo il lancio, l'esito è certo a tutti quelli che lo possono osservare.
(8) Siamo tutti incerti (chi più chi meno, a parte forse qualche fanatico) del fatto che su Marte ci sia stata vita. Un ipotetico osservatore nel lontano passato tuttavia potrebbe esserne stato sicuro, oppure la scoperta di un residuo fossile nel futuro potrebbe renderci certi del fatto che la vita sia in effetti esistita.

Se riflettiamo un po' di più sugli esempi sopra, vediamo che quello che accomuna queste situazioni e che potremmo accreditare come motivo di incertezza è la mancanza di informazione. Nelle situazioni che quindi vogliamo studiare l'incertezza è dovuta ad *informazione insufficiente* e, idealmente, se solo potessimo raccogliere abbastanza informazioni, potremmo diventare sicuri circa la verità o meno di ogni aspetto di queste. [1]

Osservazione 1. L'incertezza è dovuta ad *informazione insufficiente* ed è quindi naturalmente *soggettiva*, poiché l'informazione disponibile varia da soggetto a soggetto.

Esercizio 2. Costruite esempi di situazioni in cui l'incertezza dipende dal soggetto: in particolare di casi in cui qualcuno potrebbe essere sicuro di qualcosa, qualcun altro completamente incerto e un terzo soggetto (magari a torto) quasi sicuro dell'opposto, basandosi sulla informazione disponibile.

1.2. **Nuova informazione modifica l'incertezza.** Il fatto che l'incertezza non sia una proprietà "fisica" da prescrivere ad oggetti (ad esempio, una proprietà del dado lanciato come la massa, la composizione chimica), ma piuttosto una conseguenza dello stato dell'informazione di soggetti, è ovvio in virtù del fatto che nuova informazione modifica l'incertezza, pur mantenendo invariato lo stato fisico della situazione.

(1) Siamo incerti del fatto che ieri piovesse in una località lontana. Controlliamo i report del meteo. Diventiamo "più sicuri" circa il meteo di ieri in quella località.
(2) Siamo incerti del fatto che domani pioverà in una località. Controlliamo le previsioni del tempo. Diventiamo "meno incerti" circa il meteo di domani in quella località.
(3) Siamo incerti del fatto che la vita su Marte sia esistita. Una esplorazione trova delle tracce fossili. Diventiamo "quasi sicuri" che la vita sia esistita.

Notate comunque che *più* informazione non significa necessariamente che un soggetto diventa *meno* incerto:

[1] Questo approccio, esteso a *tutte* le situazioni naturali, porterebbe ad un punto di vista "deterministico" che da più di un secolo i fisici hanno mostrato non valere in certi contesti, però legati alla natura su scale molto piccole (fisica quantistica): per le applicazioni della probabilità nei nostri contesti, il punto di vista deterministico è una approssimazione corretta.

(4) Nuovi dati sperimentali possono confermare la validità di una teoria scientifica, ma possono anche minarne la validità, portando a molta incertezza in quell'ambito di ricerca.
(5) Solitamente siamo certi che il nostro/a compagno/a sia fedele, ma incontralo/a in compagnia di un/a ex potrebbe renderci molto più insicuri su questo fatto.

Ricapitolando: vogliamo studiare il grado di plausibilità di affermazioni sulla base di informazione incompleta, quindi in situazioni incerte. Abbiamo visto che tale plausibilità cambia in base a nuova informazione, in particolare a seconda del soggetto.

2. Teoria quantitativa

In questa sezione affrontiamo come tali aspetti si possano tradurre in un vero e proprio calcolo, che in particolare estende rendendo più "flessibile" la logica Booleana, che da questo punto di vista si occupa di situazioni di totale *assenza di incertezza*. Per avvicinarci alle regole di calcolo, descriviamo prima un approccio dovuto al fisico R.T. Cox (si veda il libro consigliato di E.T. Jaynes "Probability Theory" per una discussione più approfondita) e poi i tradizionali assiomi di Kolmogorov, che descrivono le basi della probabilità matematica moderna.

2.1. **Logica Booleana.**

Gli elementi fondamentali di cui si occupa la logica Booleana sono le *proposizioni*.

Definizione 3. Una proposizione è una affermazione di cui si può attribuire (almeno, in linea di principio) un ben determinato "valore di verità": Vero oppure Falso.

Spesso si indicano le proposizioni con lettere maiuscole A, B, C ...e il valore Vero con 1 e il valore Falso con 0. Usiamo la notazione $v(A) = 1$ se A è vero e $v(A) = 0$ se A è falso.

È facile costruire esempi di proposizioni usando la matematica:

(1) Ogni numero naturale è pari (Falso)
(2) I numeri primi sono infiniti (Vero)
(3) Ogni numero pari è la somma di due primi (Congettura di Goldbach, attualmente i matematici non sanno il suo valore di verità, ma è comunque una proposizione).

È facile costruire esempi che NON sono proposizioni usando il linguaggio comune, per via della sua naturale imprecisione:

(4) Oggi piove (dove? quando precisamente?)
(5) Mi piace la cioccolata (a chi?)...

A noi interessano situazioni intermedie, e studieremo proposizioni che descrivono situazioni reali (ripensate agli esempi delle sezioni precedenti).

La logica permette di stabilire un *calcolo* dei valori di verità tra proposizioni. Di solito questi sono rappresentati in tabelle, ma si possono anche

riassumere nelle seguenti identità:

$$\begin{aligned} v(A \wedge B) &= v(A) \cdot v(B) \\ v(\neg A) &= 1 - v(A) \\ v(A \vee B) &= v(A) + v(B) - v(A \wedge B) = v(A) + v(B) - v(A) \cdot v(B) \\ v(A \to B) &= v((\neg A) \vee B) = 1 - v(A)(1 - v(B)). \end{aligned}$$

Osservazione 4 (Proposizioni ed insiemi). Le operazioni tra proposizioni si possono anche rappresentare graficamente mediante diagrammi di Venn. Si associa ad ogni proposizione A un insieme che indichiamo con la stessa lettera. Conviene inoltre considerare questi insiemi come sottoinsiemi di un insieme "universo" Ω. Questa associazione è astratta, ma conveniente per avere un punto di vista "grafico" sul calcolo. Si ha ad esempio

$$\begin{aligned} A \wedge B \quad &\text{corrisponde a} \quad A \cap B \text{ (intersezione)} \\ A \vee B \quad &\text{corrisponde a} \quad A \cup B \text{ (unione)} \\ \neg A \quad &\text{corrisponde a} \quad A^c = \Omega \setminus A \text{ (complementare)} \end{aligned}$$

In una teoria logico-matematica, si stabiliscono delle premesse (assiomi) ossia una (o più) proposizioni vere fin dall'inizio (possiamo indicarle con I, oppure Ω) e poi si procede per *deduzione* ossia usando le regole di calcolo sopra, per ottenere nuove proposizioni vere (Teoremi). Quindi, il valore di verità di una proposizione A, anche in una teoria matematica, dipende dalle premesse: volendo evidenziare questo fatto si potrebbe scrivere

$$v(A|I) \in \{0, 1\}.$$

per il valore di verità di A assumendo che I sia vera.

2.2. **Il "teorema" di Cox.** In situazioni di incertezza, pur cercando di utilizzare tutta l'informazione I (una proposizione) che si ritiene vera, il valore di verità di alcune proposizioni A potrebbe non essere univocamente determinato, secondo le regole della logica deduttiva Booleana. In analogia con le teorie logico-matematiche, se l'informazione I è considerata come un "assioma", A non è necessariamente un "teorema".

Possiamo però introdurre un "grado di plausibilità" di una proposizione A sapendo che l'informazione I è veritiera, che denominiamo *probabilità di A sapendo I* e scriviamo

$$P(A|I).$$

Le probabilità di combinazioni di proposizioni si otterranno secondo opportune regole, come nel calcolo Booleano (anzi, estendendolo). Prima di elencare queste regole, che in realtà sono poche e semplici, in questa sezione descriviamo un approccio dovuto a R.T. Cox che si propone di "dimostrare" come queste siano in effetti conseguenze necessarie di alcuni "prerequisiti" che riteniamo qualitativamente irrinunciabili[2]. In effetti, almeno a partire dal XIX secolo, si è dibattuto sulla natura stessa della probabilità e della validità delle sue regole di calcolo (o di alcune sue conseguenze). Sapere che esse seguono in modo deduttivo da ipotesi ancora più evidenti può confortare chi avesse dubbi sulla loro validità.

[2]L'argomento è logico-deduttivo ma non una vera dimostrazione matematica.

Un altro motivo per cui introduciamo questi requisiti è che possono essere fornire degli indicatori di possibili errori nel calcolo: se nella risoluzione di un problema concreto ci rendiamo conto che essi sono violati, dobbiamo ritornare sui nostri passi e capire dove si trova un errore di calcolo o ragionamento!

Requisito 1 (Comparabilità)**.** Due probabilità devono sempre essere confrontabili, si deve sempre poter stabilire quale delle due sia maggiore. In termini matematici, la probabilità di A sapendo I è sempre un numero reale compreso tra 0 ed 1,

$$P(A|I) \in [0,1].$$

e $P(A|I) = 0$ indica un grado di fiducia nullo, ossia A si ritiene falsa, mentre $P(A|I) = 1$ indica un grado di fiducia certo, ossia A si ritiene vera. In linguaggio matematico, se $P(A|I) = 0$, diciamo che A è *trascurabile* (sapendo I), mentre se $P(A|I) = 0$, diciamo che A è *quasi certa* (sapendo I).

Il grado di fiducia dipende dall'informazione I che si suppone vera, e può cambiare drasticamente al cambiare di I. Un esempio molto semplice: $P(A|A) = 1$, ma $P(\neg A|A) = 0$.

Una probabilità non è mai *negativa* oppure *più grande di* 1. Molti errori si potrebbero evitare semplicemente accorgendosi che un calcolo o una formula non può valere perché potrebbe dare come risultati probabilità negative o più grandi di 1.

Osservazione 5 (quote)*.* Notate che invece di assumere valori in $[0,1]$, si potrebbero fare altre scelte per definire un grado di fiducia: nell'ambito delle scommesse, si preferisce parlare in termini di *quote decimali*, definite come $1/P(A|I) \in [1,\infty]$, oppure di *quote frazionali* (nel mondo anglosassone), definite come $(1/P(A|I)) - 1 \in [0,\infty]$. Esse corrispondono al fattore per cui deve essere moltiplicata una cifra giocata, nel caso di vincita, per ottenere rispettivamente il ricavo (quote decimali) o il guadagno (quote frazionali).

Requisito 2 (Buon senso)**.** Questo è il requisito più difficile da tradurre in termini matematici: le probabilità devono variare in modo qualitativamente consistente con le aspettative dettate dal buon senso (in tutte le situazioni immaginabili). Ad esempio: supponiamo di avere una informazione I e due proposizioni A, B, per cui sono assegnate le probabilità

$$P(A|I) \quad \text{e} \quad P(B|A \wedge I),$$

e supponiamo di ricevere una nuova informazione I' che aumenta il grado di fiducia in A, ma non cambia il grado di fiducia in B, sapendo A e I', ossia

$$P(A|I') \geq P(A|I) \quad \text{e} \quad P(B|A \wedge I') = P(B|A \wedge I).$$

Allora necessariamente il grado di fiducia della congiunzione A e B deve aumentare

$$P(A \wedge B|I') \geq P(A \wedge B|I).$$

Notiamo che altre regole, a prima vista molto simili e di "buon senso", sono invece da escludere. Ad esempio, supponiamo di avere una informazione I e due proposizioni A, B, per cui sono assegnate le probabilità

$$P(A|I) \quad \text{e} \quad P(B|I),$$

e supponiamo di ricevere una nuova informazione I' che aumenta sia il grado di fiducia in A sia quello di B, sapendo I, ossia

$$P(A|I') \geq P(A|I) \quad \text{e} \quad P(B|I') \geq P(B|I).$$

Tuttavia, non è necessariamente vero che il grado di fiducia della congiunzione A e B deve aumentare, ossia $P(A \wedge B|I') \geq P(A \wedge B|I)$ (trovate un esempio di una situazione realistica).

Cosa possiamo imparare da questa richiesta? Posti di fronte ad un problema da trattare con il calcolo delle probabilità, molto spesso l'intuizione già ci suggerisce qualitativamente una risposta, ad esempio: la probabilità di A è maggiore se conosco I' invece di I, mentre per B diminuisce, ecc. D'altra parte, calcoli e ragionamenti sbagliati ci possono portare a risposte in contraddizione con l'intuizione iniziale: a questo punto, conviene sempre rivedere ogni singolo passaggio e la sua correttezza – infatti spesso l'intuizione è corretta e il calcolo è sbagliato. Notiamo però che non sempre l'intuizione iniziale magari è corretta, anzi il calcolo corretto delle probabilità ci potrebbe confermare che l'intuizione era sbagliata, trovando così dei "paradossi"! In tal caso, è un buon esercizio allenare l'intuizione cercando di trovare il passaggio in cui l'intuizione viene a mancare.

Rimandiamo al libro di E.T. Jaynes per chi è interessato ad una descrizione più dettagliata di questo requisito: qui notiamo solamente che si tratta comunque di imporre la validità di *disuguaglianze* e non di formule precise su come si trasformano le probabilità (che sono le regole che cerchiamo di ottenere).

Requisito 3 (Razionalità)**.** Il calcolo delle probabilità deve essere il più possibile "consistente". Più precisamente:

i) Se la stessa probabilità $P(A|I)$ può essere ottenuta in modi diversi, il valore deve essere lo stesso.
ii) *Tutta e sola* l'informazione I deve essere utilizzata per il calcolo di $P(A|I)$, nulla di I deve essere tralasciato e nessuna nuova deve essere arbitrariamente introdotta.
iii) Se due informazioni I e I' descrivono situazioni corrispondenti (ad esempio, a meno di cambiare etichette, nomi, colori del tutto ininfluenti), allora pure le probabilità dovranno corrispondere.

In particolare, nel primo punto affermiamo anche che se proposizioni A e A' (e I, I') sono equivalenti dal punto di vista della logica Booleana, ossia se $v(A) = v(A')$ e $v(I) = v(I')$, allora si ha $P(A|I) = P(A'|I')$. Ad esempio, potremo scrivere uguaglianze del tipo

$$P(\neg(A \wedge B)|I) = P((\neg A) \vee (\neg B)|I)$$

e similmente

$$P(A|\neg(B \vee I)) = P(A|(\neg B) \wedge (\neg I)).$$

Il secondo ed il terzo punto sono difficili da mettere in forma matematicamente rigorosa. Tuttavia questo requisito può essere di grande aiuto nella risoluzione di problemi. Il primo punto si traduce nel fatto che se ci viene in mente più di un modo per calcolare una probabilità, tutti i risultati dovrebbero coincidere, o sicuramente c'è un errore in almeno uno dei modi. Il secondo punto ci mette in guardia dal trascurare "pezzi" dell'informazione

I, oppure di aggiungere ipotesi che magari semplificano il calcolo, ma non sono presenti (neppure implicitamente). Il terzo ci ricorda che possiamo spesso ricondurci a situazioni "modello" (ne vedremo nel corso) e in questo modo evitare di ripetere ragionamenti.

A questo punto si potrebbe argomentare la validità del seguente

Risultato 6 ("teorema" di R.T. Cox)**.** *L'unico modo di soddisfare i requisiti 1, 2, 3 descritti sopra è che la probabilità soddisfi le seguenti regole di calcolo:*

$$P(A|I) + P(\neg A|I) = 1 \quad \textit{(regola della somma)}$$

$$P(A \wedge B|I) = P(A|I)P(B|A \wedge I) \quad \textit{(regola del prodotto)}$$

per ogni possibile scelta di proposizioni A, B ed I.

In effetti, l'unicità è da intendere a meno di trasformazioni matematicamente semplici, simili ad esempio al passaggio da probabilità a quote di scommesse descritto sopra. Quello che colpisce di questo risultato è come le due regole di calcolo fondamentali (della somma e del prodotto) seguano dalla lista di proprietà "qualitative" descritte sopra. In effetti usando queste due e i requisiti possiamo *dedurre* la validità di (quasi) tutte le altre "regole" del calcolo delle probabilità.

Osservazione 7 (additività, due proposizioni incompatibili). Siano A, B due proposizioni *incompatibili*, ossia tali che se una è vera necessariamente l'altra è falsa, o più brevemente

$$A \wedge B \quad \text{è sempre sicuramente falsa,} \quad v(A \wedge B) = 0.$$

Ad esempio: $B = \neg A$, o anche $B = (\neg A) \wedge C$. Allora, grazie alle due regole sopra possiamo dedurre che, qualunque sia l'informazione I, la probabilità è *additiva*:

$$P(A \vee B|I) = P(A|I) + P(B|I).$$

Notiamo infatti che l'ipotesi $v(A \wedge B) = 0$ e le regole di calcolo Booleano ci permettono di dedurre che

$$v((\neg A) \wedge B) = v(B) - v(A \wedge B) = v(B),$$

quindi per il requisito 3i), abbiamo l'uguaglianza

$$P((\neg A) \wedge B|I) = P(B|I).$$

Ora usiamo le regole di somma e prodotto nel seguente modo:

$$\begin{aligned}
P(A \vee B|I) &= 1 - P(\neg(A \vee B)|I) \quad \text{(regola della somma)} \\
&= 1 - P((\neg A) \wedge (\neg B)|I) \\
&= 1 - P(\neg A|I)P(\neg B|(\neg A) \wedge I) \quad \text{(regola del prodotto)} \\
&= 1 - P(\neg A|I)\left[1 - P(B|(\neg A) \wedge I)\right] \quad \text{(regola del somma)} \\
&= 1 - P(\neg A|I) + P(\neg A|I)P(B|(\neg A) \wedge I) \\
&= P(A|I) + P((\neg A) \wedge B|I) \\
&= P(A|I) + P(B|I)
\end{aligned}$$

Si può generalizzare l'esempio sopra in diversi modi, ad esempio aumentando il numero di proposizioni (si dimostra per induzione matematica partendo dal caso di due).

Osservazione 8 (additività, n proposizioni a due a due incompatibili). Sia $n \geq 2$, A_1, A_2, $\dots A_n$ proposizioni *a due a due incompatibili*, ossia tali che

$$A_i \wedge A_j \text{ è falsa per ogni } i, j \in \{1, \dots, n\} \text{ con } i \neq j.$$

Allora vale

$$P(A_1 \vee A_2 \vee \dots \vee A_n | I) = P(A_1|I) + \dots + P(A_n|I) = \sum_{i=1}^{n} P(A_i|I).$$

Osservazione 9 (probabilità di $A \vee B$, caso generale). Cosa possiamo dire di $P(A \vee B|I)$ se A e B non sono incompatibili? In generale, vale la formula

$$P(A \vee B|I) = P(A|I) + P(B|I) - P(A \wedge B|I) \tag{1}$$

(che ricorda quella per il valore di verità vista sopra). Per dedurla dalle altre, basta notare che le proposizioni

$$A \wedge (\neg B), \quad (\neg A) \wedge B, \quad A \wedge B$$

sono a due a due incompatibili e la loro disgiunzione è $A \vee B$ (disegnate un diagramma di Venn per convincervene), quindi per l'additività si ha

$$\begin{aligned} P(A \vee B|I) &= P(A \wedge (\neg B)|I) + P((\neg A) \wedge B|I) + P(A \wedge B|I) \\ &= [P(A \wedge (\neg B)|I) + P(A \wedge B|I)] + [P((\neg A) \wedge B|I) + P(A \wedge B|I)] - P(A \wedge B|I). \end{aligned}$$

D'altra parte, $(A \wedge (\neg B)) \vee (A \wedge B) = A$ e similmente $((\neg A) \wedge B) \vee (A \wedge B) = B$, quindi

$$P(A \wedge (\neg B)|I) + P(A \wedge B|I) = P(A|I) \quad P((\neg A) \wedge B|I) + P(A \wedge B|I) = P(B|I).$$

Ci sono formule (dette di inclusione-esclusione) che permettono di trattare l'analogo di (1) quando si hanno n proposizioni, non necessariamente incompatibili.

Osservazione 10 (sub-additività). Dalla formula (1), siccome $P(A \wedge B|I) \geq 0$, otteniamo che in generale vale la disuguaglianza

$$P(A \vee B|I) \leq P(A|I) + P(B|I).$$

Ragionando per induzione, possiamo estendere la disuguaglianza anche per n proposizioni A_1, $\dots A_n$, ottenendo la *sub-additività* della probabilità:

$$P(A_1 \vee A_2 \vee \dots \vee A_n | I) \leq \sum_{i=1}^{n} P(A_i|I),$$

che a parole si può dire come *la probabilità che almeno una tra le proposizioni risulti vera è più piccola della somma delle singole probabilità.* Notiamo che il membro a sinistra è sempre più piccolo di 1, essendo una probabilità, quindi se le probabilità a destra sono molto grandi (e la somma supera 1), non è una disuguaglianza molto utile. Al contrario, se le probabilità a destra sono piccole, pure la somma potrà risultare piccola e quindi si ottiene un risultato interessante.

Nel caso estremo in cui $P(A_i|I) = 0$ per ogni i, ossia le A_i sono trascurabili (sapendo I) otteniamo che

$$P(A_1 \vee A_2 \vee \dots \vee A_n | I) \leq \sum_{i=1}^{n} P(A_i|I) = 0.$$

A parole: se diamo grado di fiducia nullo a n proposizioni, pure il fatto che almeno una di queste risulti vera avrà grado di fiducia nullo.

Osservazione 11 (regola del prodotto per n proposizioni). Sia $n \geq 2$, A_1, A_2, $\dots A_n$ proposizioni. Allora ragionando per induzione su n, si dimostra che la regola del prodotto permette di calcolare

(2)
$$P(A_1 \wedge A_2 \wedge \dots \wedge A_n | I) = P(A_1|I) \cdot P(A_2|A_1 \wedge I) \cdot \dots \cdot P(A_n|A_{n-1} \wedge A_{n-2} \wedge \dots \wedge A_1 \wedge I).$$

Osservazione 12 (monotonia). Siano A, B proposizioni tali che, in qualunque situazione A risulti vera, allora anche B è vera, ossia brevemente $A \to B$ è sempre vera, o $A \wedge (\neg B)$ è sempre falsa. Allora possiamo mostrare che $P(A|I) \leq P(B|I)$, usando le due proposizioni incompatibili $A \wedge B$, $A \wedge (\neg B)$

$$\begin{aligned} P(A|I) &= P(A \wedge B|I) + P(A \wedge (\neg B)|I) \\ &= P(A \wedge B|I) \\ &= P(B \wedge A|I) = P(B|I)P(A|B \wedge I) \quad \text{(regola del prodotto)} \\ &\leq P(B|I) \quad \text{perchè } P(A|B \wedge I) \leq 1. \end{aligned}$$

Un esempio che si trova spesso è del tipo $A = B \wedge C$, da cui si ottiene che

$$P(B \wedge C|I) \leq P(B|I) \quad \text{e anche} \quad P(B \wedge C|I) \leq P(C|I).$$

La proprietà di monotonia della probabilità è evidentemente in accordo con l'intuizione, però è facile formulare problemi in cui a prima vista si risponde nel modo opposto, ossia che $P(B \wedge C|I) > P(B|I)$. Un esempio famoso è il seguente

Esempio 13 (Linda[3]). Linda ha 31 anni, nubile, estroversa, brillante, laureata in economia, da studentessa molto impegnata politicamente e di ideologia anti-nucleare. Dovendo scommettere, quale delle seguenti affermazioni è più probabile?

A: Linda lavora in banca.
B: Linda è una femminista militante.
C: Linda lavora in banca ed è una femminista militante.

Siete stati tentati dal rispondere C? Ovviamente si ha $C = A \wedge B$, quindi C è da escludere (al massimo il dubbio può essere tra A e B).

2.3. **Gli assiomi di Kolmogorov.** Le regole di calcolo della probabilità, come la regola della somma, del prodotto e la proprietà di additività per proposizioni a due a due incompatibili, la monotonia, e altre che vedremo, erano note molto tempo prima dell'argomento di R.T. Cox. Tuttavia, non era completamente chiaro quale posizione avessero nell'ambito della matematica.

Un importante contributo è stato dato dal matematico A. Kolmogorov, il quale ha proposto una teoria *assiomatica* della probabilità, basandosi sulla corrispondenza tra proposizioni ed insiemi, e tra probabilità e misura, che poi è stata adottata sostanzialmente da tutti i matematici, anche per via del fatto che permette agevolmente di studiare limiti di problemi in cui intervengono "infinite proposizioni". In questa sezione, descriviamo brevemente

[3] https://it.wikipedia.org/wiki/Teoria_del_prospetto

il punto di vista di Kolmogorov, ma nel resto del corso manterremo comunque un approccio più intuitivo al calcolo delle probabilità, senza occuparci di discutere aspetti puramente matematici ad esso collegati.

Nell'approccio di Kolmogorov si sfrutta la corrispondenza tra proposizioni ed insiemi. Il primo passo consiste nel fissare un insieme "universo", tradizionalmente indicato con Ω, che nei problemi concreti rappresenta l'informazione di cui si dispone inizialmente, e quindi "vera" o accettata come tale. Successivamente si individua una collezione $\mathcal{A}$ di insiemi $A \subseteq \Omega$ che corrispondono alle proposizioni "interessanti" ai fini del problema, per i quali andremo a definire le probabilità $P(A|\Omega)$, rispetto all'informazione iniziale Ω. Non necessariamente *tutti* i sottoinsiemi di A devono appartenere alla collezione $\mathcal{A}$, ma è sufficiente che $\mathcal{A}$ sia una *algebra* (o, per trattare problemi con infiniti insiemi, una σ-algebra).

Definizione 14 (Algebra di eventi)**.** Fissiato un insieme Ω, una collezione $\mathcal{A}$ di insiemi $A \subseteq \Omega$ è detta algebra se

i) $\emptyset \in \mathcal{A}$, $\Omega \in \mathcal{A}$;
ii) per ogni $A \in \mathcal{A}$, l'insieme $A^c = \Omega \setminus A$ pure appartiene ad $\mathcal{A}$;
iii) per ogni A, B entrambi appartenenti ad $\mathcal{A}$, si ha $(A \cap B) \in \mathcal{A}$ e $(A \cup B) \in \mathcal{A}$

La collezione è detta σ-algebra se la terza condizione vale anche per unioni infinite numerabili di insiemi: se $(A_n)_{n=1}^{\infty}$ sono tali che $A_n \in \mathcal{A}$ per ogni $n \geq 1$, allora $(\bigcap_{n=1}^{\infty} A_n) \in \mathcal{A}$ e $(\bigcup_{n=1}^{\infty} A_n) \in \mathcal{A}$.

Gli insiemi $A \in \mathcal{A}$ sono detti *eventi*.

A prima vista l'idea di introdurre un'algebra $\mathcal{A}$ sembra una complicazione: perché non considerare direttamente *tutti* i sottoinsiemi di Ω? Ci sono due motivi. Il primo è "economico": nella teoria di Kolmogorov basta assegnare le probabilità solamente agli eventi $A \in \mathcal{A}$, quindi uno non si deve preoccupare degli insiemi che non vi appartengono, a volte con un notevole risparmio. Il secondo è propriamente matematico: nel caso di insiemi Ω infiniti (ad esempio, $\Omega = [0,1]$) è possibile dimostrare che, in alcuni casi, richiedere di definire $P(A|\Omega)$ con certe proprietà, per ogni $A \subseteq \Omega$, porta a contraddizioni. Perciò ci si accontenta di lavorare su una collezione $\mathcal{A}$, comunque sufficientemente ampia.

Il secondo passo consiste nella introduzione di una probabilità $P(\cdot|\Omega)$ rispetto alla informazione iniziale. Ecco la definizione secondo Kolmogorov.

Definizione 15 (Probabilità)**.** Sia Ω un insieme su cui è definita una $\mathcal{A}$ una algebra di eventi (o una σ-algebra). Si definisce come probabilità $P(\cdot|\Omega)$ una funzione

$$P(\cdot|\Omega) : \mathcal{A} \to [0,1], \quad A \mapsto P(A|\Omega)$$

tale che

i) $P(\emptyset|\Omega) = 0$, $P(\Omega|\Omega) = 1$;
ii) se A, $B \in \mathcal{A}$ sono eventi incompatibili, ossia $A \cap B = \emptyset$, allora

$$P(A \cup B|\Omega) = P(A|\Omega) + P(B|\Omega);$$

iii) (solo nel caso in cui $\mathcal{A}$ sia una σ-algebra) se $(A_n)_{n=1}^{\infty}$ sono eventi a due a due incompatibili, ossia $A_i \cap A_j = \emptyset$ per ogni $i \neq j$, allora

$$P(\bigcup_{n=1}^{\infty} A_n|\Omega) = \sum_{n=1}^{\infty} P(A_n|\Omega).$$

Spesso (quasi sempre) si omette di specificare Ω e si scrive $P(A) = P(A|\Omega)$.

L'approccio di Kolmogorov mette in risalto la proprietà di *additività* per eventi a due a due incompatibili, che abbiamo visto essere una conseguenza delle regole di somma e prodotto nell'approccio della sezione precedente. Mentre è chiaro che la regola di somma si ottiene scegliendo $B = A^c$, non è a prima vista chiaro come ottenere la regola del prodotto. In effetti, il terzo passo della teoria di Kolmogorov consiste nel trasformare la regola del prodotto in una *definizione* di probabilità condizionata.

Definizione 16 (Probabilità condizionata)**.** Sia Ω un insieme su cui è definita una $\mathcal{A}$ una algebra di eventi e una probabilità $P(\cdot|\Omega)$. Dati eventi A, $B \in \mathcal{A}$, se $P(A|\Omega) > 0$, si definisce $P(B|A \cap \Omega)$ come la quantità

$$P(B|A \cap \Omega) = \frac{P(A \cap B|\Omega)}{P(A|\Omega)}.$$

Confrontanto la definizione di Kolmogorov e la regola del prodotto (scritta in termini di eventi, ossia insiemi, invece di proposizioni)

$$P(A \cap B|I) = P(A|I)P(B|A \cap I),$$

notiamo che, nel caso $I = \Omega$, semplicemente abbiamo diviso ambo i membri per la quantità $P(A|I)$ (non nulla, per ipotesi), in modo tale che la regola valga *per definizione*, almeno nel caso di $I = \Omega$. In realtà, nella teoria di Kolmogorov, la regola vale qualunque sia $I \in \mathcal{A}$, purché sia $P(I \in \mathcal{A}|\Omega) > 0$. Basta infatti confrontare i due membri che si ottengono usando la definizione di Kolmogorov. Da un lato,

$$P(A \cap B|I) = \frac{P(A \cap B \cap I|\Omega)}{P(I|\Omega)}$$

dall'altro si ha

$$P(A|I)P(B|A \cap I) = \frac{P(A \cap I|\Omega)}{P(I|\Omega)} \cdot \frac{P(B \cap A \cap I|\Omega)}{P(A \cap I|\Omega)} = \frac{P(A \cap B \cap I|\Omega)}{P(I|\Omega)},$$

e quindi coincidono.

Osservazione 17 (Pro e contro della teoria di Kolmogorov). Abbiamo visto che la regola del prodotto e della somma quindi valgono nella teoria assiomatica di Kolmogorov, come pure l'additività per eventi a due a due incompatibili (per definizione) e pure la proprietà di monotonia, che sappiamo essere una conseguenza delle altre. Dal punto di vista della risoluzione di problemi pratici, quindi, la teoria di Kolmogorov non si differenzia molto dalla teoria "logica" descritta nelle sezioni precedenti. Le differenze si vedono invece nel momento in cui si devono dimostrare teoremi matematici che coinvolgono infinite variabili aleatorie (ossia limiti di famiglie finite): la teoria di Kolmogorov diventa molto flessibile, e utile. Evidenziamo alcuni punti:

i) La teoria richiede *sempre e comunque* di "costruire" un insieme Ω e una (σ-)algebra di eventi $\mathcal{A}$, e una probabilità "iniziale" $P(\cdot|\Omega)$, *prima* di risolvere il problema, ossia calcolare probabilità cercate del tipo $P(A|I)$. Questo è un aspetto positivo perché garantisce una certa coerenza, ed è un esercizio molto utile per i matematici, ma dal punto di vista pratico spesso non aggiunge molto alla comprensione del problema rispetto ad un approccio che privilegia il ruolo dei *sistemi di alternative* e *delle probabilità condizionate*, come cercheremo di evidenziare.
ii) La costruzione matematica di $P(\cdot|\Omega)$ è a volte un problema non banale, e spesso si appoggia a risultati molto profondi della *teoria della misura*. D'altra parte, tanti aspetti problematici nascono con passaggi al limite, e spesso nei problemi non è necessario, oppure si può trovare opportune "scorciatoie" (ad esempio, trovare prima una formula per la probabilità nel caso finito, e poi passare al limite).
iii) Si tende inevitabilmente a dare un ruolo "principale" alla probabilità $P(\cdot|\Omega)$ e subordinato a quelle condizionate $P(\cdot|I)$ rispetto ad altre informazioni, quando invece nella pratica a volte sono più interessanti le seconde. Inoltre, si tende ad associare un valore "oggettivo" e "immutabile" alla probabilità iniziale, quando invece la probabilità deve aggiornarsi sempre quando si ottiene nuova informazione.
iv) La distinzione tra "proposizioni" ed "eventi" permette di separare il problema concreto, reale, dalla trattazione matematica. Inoltre ragionare con insiemi (anche aiutandosi con diagrammi di Venn) più essere utile per evitare errori.

Nel seguito, adotteremo in modo implicito la teoria di Kolmogorov, immaginando di lavorare sempre in qualche insieme Ω con eventi $A \in \mathcal{A}$ e una probabilità iniziale. Però non ci preoccuperemo mai della costruzione di tali spazi, e useremo in modo interscambiabile i termini *proposizione*, *insieme* ed *evento*, indicandoli spesso con lettere maiuscole A, B, I, Ω, come pure le operazioni tra insiemi ($\cup$, $\cap$, $\cdot^c$) e proposizioni ($\vee$, $\wedge$, $\neg$).

3. Sistemi di alternative

Abbiamo visto che una proprietà importante della probabilità, sia che usiamo l'approccio "logico" di Cox o quello "insiemistico" di Kolmogorov, è l'additività per eventi *a due a due incompatibili*. Nella teoria di Kolmogorov, dati A_1, A_2, $\dots A_n$ eventi, essi si dicono *a due a due incompatibili* (o mutuamente esclusivi) se

$$A_i \cap A_j = \emptyset \quad \text{per ogni } i,\, j \in \{1, \dots, n\}, \text{ con } i \neq j.$$

In termini di proposizioni, significa che $A_i \wedge A_j$ è sicuramente falsa (se $i \neq j$). In questo caso, qualunque sia l'informazione I, vale la proprietà di additività

$$P(\bigcup_{i=1}^{n} A_i|I) = \sum_{i=1}^{n} P(A_i|I),$$

che abbiamo dedotto in precedenza (per induzione su n) dalla regola della somma e del prodotto.

Un caso speciale, ma molto utile, nei problemi è dato da una famiglia di eventi A_1, A_2, $\dots A_n$ a due a due incompatibili tali che *almeno* (e quindi necessariamente uno solo) tra questi è sempre vero. In formule,

$$\bigcup_{i=1}^{n} A_i = A_1 \cup A_2 \cup \dots \cup A_n = \Omega.$$

In tal caso si dice che gli eventi A_1, A_2, $\dots A_n$ sono un *sistema di alternative*.

Esempio 18 (alternative "semplici"). Dato un evento A, possiamo sempre costruire il sistema di alternative comprendente A e la sua negazione $\neg A = A^c = \Omega \setminus A$. Esempio:

$$A = \text{“oggi piove a Pisa”}, \quad \neg A = \text{“oggi non piove a Pisa”}.$$

Esempio 19. Supponiamo di avere davanti a noi una scatola chiusa che contiene un certo numero di palline al suo interno, che non possiamo vedere (questo sarà il "modello dell'urna", che studieremo meglio più avanti). Supponiamo di essere certi che tale numero è compreso tra 0 e 5 (ad esempio, sappiamo che un nostro amico di cui ci fidiamo ha messo appunto un tale numero di palline, che però non conosciamo). Allora possiamo considerare il sistema di alternative

$$A_i = \text{la scatola contiene } i \text{ palline}$$

con $i \in \{0, 1, 2, 3, 4, 5\}$.

Se abbiamo un sistema di alternative A_1, A_2, $\dots A_n$, dalla proprietà di additività, otteniamo che

$$\sum_{i=1}^{n} P(A_i|I) = P(\bigcup_{i=1}^{n} A_i|I) = P(\Omega|I) = 1.$$

Una proprietà molto utile di un sistema di alternative è la seguente, che permette di calcolare la probabilità di un evento B "decomponendo" a seconda della alternativa che si potrebbe presentare.

Proposizione 20 (decomposizione della probabilità)**.** *Dato un sistema di alternative A_1, A_2, $\dots A_n$ e un evento B (in generale diverso dalle alternative), si ha*

$$P(B|I) = \sum_{i=1}^{n} P(B|A_i \cap I) P(A_i|I). \tag{3}$$

Dimostrazione. Si ha

$$B = B \cap \Omega = B \cap (\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} (B \cap A_i),$$

e gli eventi $B \cap A_i$ sono a due a due incompatibili (perché?). Per l'additività

$$P(B|I) = P(\bigcup_{i=1}^{n} (B \cap A_i)|I) = \sum_{i=1}^{n} P(B \cap A_i|I),$$

che equivale alla tesi, usando la regola del prodotto

$$P(B \cap A_i|I) = P(A_i|I) P(B|A_i \cap I). \qquad \square$$

Attenzione: nella pratica, un errore molto comune è di calcolare le singole $P(B|A_i \cap I)$ e poi di sommarle, senza tenere conto del "peso" $P(A_i|I)$. Questo potrebbe essere originato da uno "scambio" tra

$$P(B \cap A_i|I) \quad \text{e} \quad P(B|A_i \cap I),$$

che però sono quantità diverse (proprio per il fattore $P(A_i|I)$).

Osservazione 21 (Alternative trascurabili). Dato un sistema di alternative $A_1, \dots A_n$, ci possiamo trovare in una situazione in cui una o più di queste è tale che $P(A_i|I) = 0$, ossia è trascurabile sapendo l'informazione I. Nella formula (3), allora, possiamo semplicemente omettere queste alternative. Ad esempio: supponiamo di avere le 5 alternative dell'esempio (19),

$$A_i = \text{la scatola contiene } i \text{ palline}$$

con $i \in \{0, 1, 2, 3, 4, 5\}$, ma di venire poi a sapere che la scatola non è vuota (ad esempio, pesandola): allora possiamo "eliminare" l'alternativa A_0 dal nostro ragionamento.

Un'altra semplificazione può accadere quando gli eventi $A_1, \dots A_n$ non sono propriamente *incompatibili*, ma per qualche motivo si riesce a dimostrare che

$$P(A_i \cap A_j|I) = 0 \quad \text{per ogni } i \neq j,$$

ossia le intersezioni sono *trascurabili* (sapendo I) e l'unione non è tutto Ω, ma si ha

$$P(\bigcup_{i=1}^{n} A_i|I) = 1,$$

ossia è *quasi certa* (sapendo I). In questo caso, possiamo trattarli come un vero e proprio sistema di alternative (fintanto che usiamo l'informazione I), ad esempio (3) vale pure in questo caso. Per dimostrarlo rigorosamente, basta costruire un vero sistema di alternative, ad esempio ponendo

$$C_1 = A_1, \quad C_2 = A_2 \setminus A_1, \quad \dots, \quad C_n = A_n \setminus (A_1 \cup A_2 \cup \dots A_{n-1})$$

e infine $C_{n+1} := \Omega \setminus (\bigcup_{i=1}^{n} A_i)$. Dato un qualunque B, ripetendo la dimostrazione di (2), otteniamo

$$P(B|I) = \sum_{i=1}^{n+1} P(B \cap C_i|I).$$

Siccome $B \cap C_i$ e $B \cap (C_i \setminus A_i)$ sono incompatibili, si ha

$$P(B \cap A_i|I) = P(B \cap C_i|I) + P(B \cap (A_i \setminus C_i)|I) = P(B \cap A_i|I)$$

perché

$$P(B \cap (A_i \setminus C_i)|I) \leq P(A_1 \cup \dots A_{i-1}|I) \leq \sum_{j=1}^{i-1} P(A_i|I) = 0.$$

Quindi possiamo scrivere

$$P(B|I) = \sum_{i=1}^{n+1} P(B \cap C_i|I) = \sum_{i=1}^{n+1} P(B \cap A_i|I).$$

e concludere la come nella dimostrazione di (2).

3.1. **Alberi e alternative.** Possiamo dare una rappresentazione grafica di un sistema di alternative mediante un grafo ad albero (che si ramifica da sinistra a destra), in cui nella "radice" abbiamo una informazione I, ciascuna foglia è una alternativa e ciascun "ramo" è "pesato" la probabilità $P(A_i|I)$ (figura 3.1). A questo punto, se ci interessa la probabilità di un evento B, sapendo I, possiamo aggiungere una ulteriore ramificazione da ciascuna foglia e pesarla con la probabilità $P(B|A_i \cap I)$, e la formula (3) ci dice che per calcolare la $P(B|I)$ dobbiamo sommare per ciascun "ramo" il prodotto dei "pesi" corrispondenti (figura 3.2).

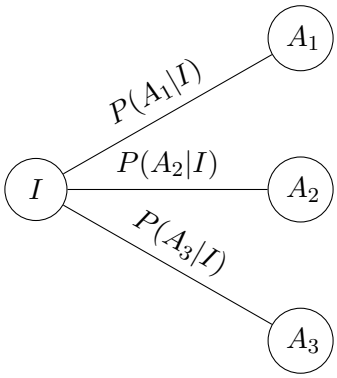


FIGURA 3.1. Albero corrispondente ad un sistema di 3 alternative A_1, A_2, A_3.

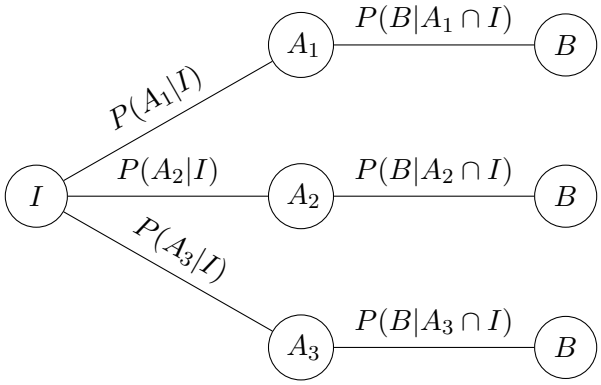


FIGURA 3.2. $P(B) = \sum_{i=1}^{3} P(B|A_i \cap I)P(A_i|I)$.

Questo ragionamento si può estendere anche quando si hanno diversi sistemi di alternative: ciascuna foglia A_i può a sua volta diventare una nuova radice per un qualunque altro sistema di alternative, e quindi le nuove alternative diventano foglie, e così via... Attenzione: in generale, il "peso" corretto da mettere nei rami è la probabilità che sia vero l'evento a destra sapendo *tutta l'informazione a sinistra, ottenuta "risalendo" il ramo fino alla radice*, e NON semplicemente la probabilità dell'evento a destra sapendo l'evento immediatamente a sinistra (figura 3.3). Dopo aver completato un albero (che può diventare anche molto complesso) se siamo interessati

alla probabilità di un evento B, basterà come prima aggiungere ad ogni foglia una ulteriore ramificazione e pesarla con la probabilità di B, sapendo tutta l'informazione a sinistra, e poi sommare su tutti i rami i prodotti dei pesi (figura 3.4). Notiamo anche che questo ultimo passaggio si può anche interpretare come l'introduzione del sistema di alternative B, B^c.

Notiamo infine che, se una o più alternative hanno peso nullo (ossia sono trascurabili), possiamo sempre "tagliare" il ramo corrispondente, ossia eliminarlo dal ragionamento e comportarci come se non esistesse affatto (ovviamente, prima di eliminarlo, sempre giustificare perché!). Allo stesso modo, se invece di un vero sistema di alternative si dispone di un sistema di alternative come descritto nell'Osservazione (21), possiamo comunque ragionare costruendo l'albero allo stesso modo.

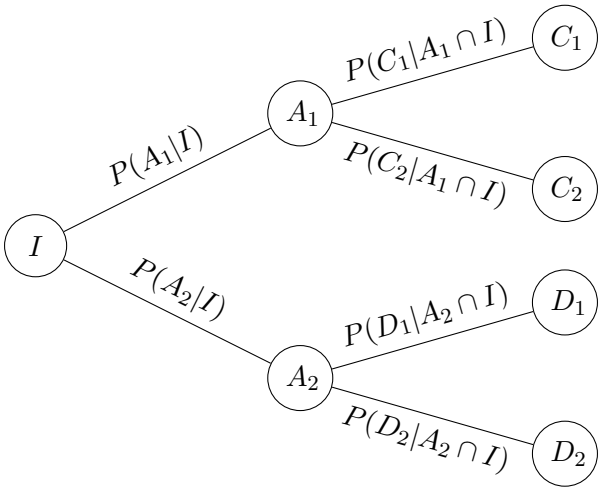


FIGURA 3.3. Albero costruito combinando diversi sistemi di alternative, precisamente i tre sistemi (A_1, A_2), (C_1, C_2) e (D_1, D_2). Notate che a partire da ogni nodo possiamo usare un sistema di alternative diverse.

3.2. **Probabilità uniforme.** Ora abbiamo quasi tutti gli strumenti per affrontare i primi problemi concreti usando il calcolo delle probabilità. Resta però una questione, sia che usiamo l'approccio di Cox o quello di Kolmogorov: come attribuire delle probabilità "iniziali" sulla base di una informazione ottenuta mediante il linguaggio naturale?

In generale, questo è un problema difficile, e più informazione iniziale abbiamo, più è difficile attribuire delle probabilità. Anzi, meno l'informazione di cui disponiamo *favorisce* un evento A rispetto all'alternativa A^c, più sicuri ci sentiamo nell'attribuire *eguale probabilità*

$$P(A|\Omega) = P(\neg A|\Omega) = \frac{1}{2}$$

siccome la somma deve essere 1 (pensiamo al lancio di una moneta).

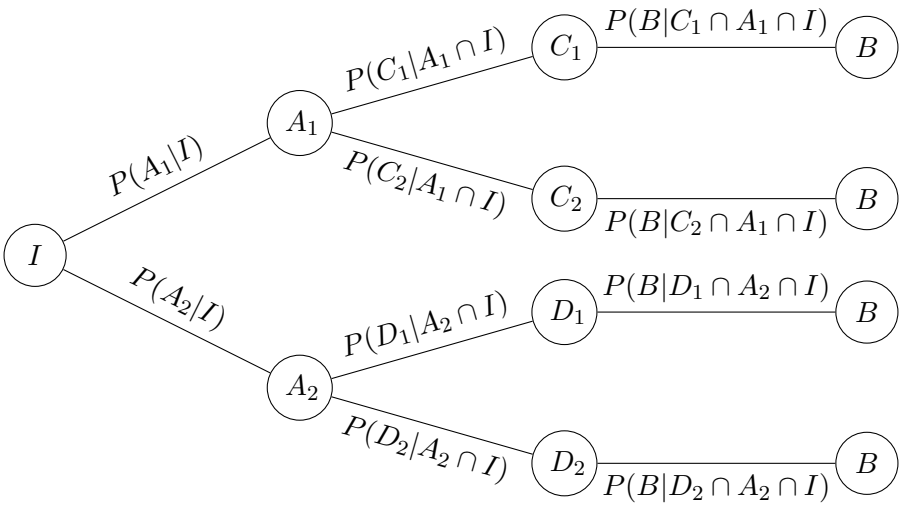


FIGURA 3.4. Per calcolare $P(B|I)$ basta sommare i prodotti dei pesi dei singoli rami, dalla radice I a ciascuna foglia.

Più in generale, supponiamo che il problema ci permetta di individuare un sistema di alternative $A_1, \ldots, A_n$, quindi sappiamo che

$$\sum_{i=1}^{n} P(A_i|\Omega) = 1,$$

e però *non abbiamo* ulteriore informazione per favorire una alternativa rispetto alle altre: ad esempio, è possibile fare un cambio di nomi/etichette in modo da scambiare le diverse alternative, senza modificare sostanzialmente il problema. A questo punto, pensando anche al Requisito 3iii), attribuiamo la stessa probabilità a ciascuna alternativa

$$P(A_1|\Omega) = P(A_2|\Omega) = \ldots = P(A_n|\Omega)$$

e la probabilità sarà detta *uniforme*

$$P(A_1|\Omega) = \ldots = P(A_n|\Omega) = \frac{1}{n}.$$

Esempio 22 (estrazione da un'urna). Abbiamo una scatola chiusa che contiene 4 palline, tutte indistinguibili al tatto ma numerate $1, 2, 3, 4$ in superficie. Una persona bendata estrae una pallina. Possiamo introdurre il sistema di alternative

$$A_i = \text{“la pallina con etichetta } i \text{ viene estratta”} \quad \text{per } i = 1, \ldots, 4.$$

In questa situazione possiamo immaginare di cambiare le etichette senza cambiare il problema, quindi attribuiamo probabilità uniforme

$$P(A_i|\Omega) = \frac{1}{4}.$$

Se un sistema di alternative ha probabilità uniforme, usando l'additività possiamo calcolare la probabilità che almeno una tra una sotto-famiglia di k alternative $A_{i_1}, \ldots, A_{i_k}$ si realizzi (possiamo pensare questa famiglia come

dei "casi favorevoli"). Si recupera allora la classica formula della probabilità come "casi favorevoli" su "casi possibili"

$$P(A_{i_1} \cup \ldots \cup A_{i_k}|\Omega) = \sum_{j=1}^{k} P(A_{i_j}|\Omega) = \frac{k}{n}.$$

Questo "metodo" con cui abbiamo attribuito probabilità uniforme a un sistema di alternative, è anche detto *principio di indifferenza* o principio di ragione insufficiente, e attribuito a P.S. Laplace (anche se non fu il primo a usare probabilità uniformi). Possiamo riassumerlo così: *la probabilità è il rapporto tra i casi favorevoli e i casi possibili, quando non vi sono informazioni sufficienti per preferire un caso rispetto ad un altro.*

Esercizio 23. Provate a giustificare se in queste situazioni si può invocare il "principio di indifferenza" di Laplace e ricondurci a probabilità uniformi su eventi che costituiscono un sistema di alternative. Ragionate sulla informazione che state (o non state) usando per applicare (o meno) il principio.

(1) "I possibili esiti dell'esame di CPS sono due: o lo supero o non lo supero"
(2) "Domani il sole sorgerà oppure non sorgerà"
(3) "Estraggo bendato una carta da un mazzo di 52"
(4) "Bendato, apro una pagina di un vocabolario di italiano e guardo la lettera iniziale cui si riferisce"
(5) "Chiedo il mese di nascita della prima persona (che non conosco) che incontro per strada"
(6) "Chiedo ad un amico di pensare ad un numero da 1 a 10 e provo ad indovinarlo"
(7) "Il numero di e-mail che riceverò nella prossima ora sarà (quasi) sicuramente un numero tra 0 e 100"
(8) "Una password di un utente è formata da 4 cifre, quindi sarà una tra le 10^4 possibili password"

Provate a costruire da voi altre situazioni (realistiche) e ragionate sulla validità del "principio di indifferenza".

4. Il modello dell'urna (I) estrazioni senza reimmissione

In questa sezione approfondiamo una situazione probabilistica fondamentale (a cui molti problemi si potranno ricondurre) riprendendo l'esempio dell'urna (19). Immaginiamo quindi di avere davanti a noi un'urna (una scatola, un vaso) di cui non vediamo il contenuto, che sappiamo essere di un numero (noto) N di palline tutte identiche tra loro, eccetto per un'etichetta. Ad esempio, possiamo supporre che siano numerate da 1 a N, ma, per semplificare ulteriormente la trattazione, immaginiamo di sapere solamente che un numero (noto) R tra queste sia colorata di rosso e il rimanente $B = N - R$ sia colorata invece di blu (ad esempio, se sono numerate, possiamo pensare che quelle numerate 1, 2, ..., R sono rosse, mentre quelle numerate $R+1$, $R+2$, ..., $R+B$ sono blu). Da questa "urna" immaginiamo di fare eseguire a qualcuno una successione di estrazioni (senza guardare!), prelevando una sola pallina per volta, *senza poi rimetterla dentro l'urna.* Inoltre, ad ogni

estrazione, possiamo venire a sapere o meno qual è il colore della pallina estratta: concretamente, ad esempio, la persona che effettua l'estrazione tiene nota in ordine dei colori delle pallina estratte, e può comunicarcelo oppure no.

Il numero di estrazioni *massimo* che la persona può fare è N. Per ogni $i \in \{1, 2, \dots, N\}$ possiamo introdurre il sistema di alternative

$$R_i = \text{“la pallina estratta all'estrazione } i \text{ è di colore rosso”},$$

$$R_i^c = B_i = \text{“la pallina estratta all'estrazione } i \text{ è di colore blu”}.$$

Inoltre, è comodo riassumere la descrizione del contenuto dell'urna introducendo l'informazione

$$I(N, R, B) = \text{“l'urna contiene } N \text{ palline di cui } R \text{ rosse e } B \text{ blu”}$$

In effetti, siccome $R+B = N$ si potrebbe tenere conto solamente del numero di palline rosse e quelle totali (oppure solo delle rosse e delle blu), ma questa “ridondanza” forse permette di capire meglio lo stato dell'urna.

4.1. **Prima estrazione.** Consideriamo la prima estrazione: anche se le alternative sono due R_1, B_1, è chiaro che l'informazione $I(N, R, B)$ favorisce l'una o l'altra, a seconda del numero di palline (pensate al caso in cui sono tutte rosse). Per calcolare $P(R_1|I(N, R, B))$ possiamo introdurre il sistema di alternative

$$A_i = \text{“la pallina estratta è la numero } i\text{”}$$

per $i \in \{1, \dots, N\}$. Abbiamo già visto che possiamo attribuirvi probabilità uniforme (il colore non favorisce alcuna pallina rispetto alle altre) e quindi

$$P(A_i|I(N, R, B)) = \frac{1}{N}.$$

D'altra parte, pensando ad esempio che le palline rosse sono quelle numerate da 1 ad R, troviamo

$$R_1 = \bigcup_{i=1}^{R} A_i$$

e quindi

$$P(R_1|I(N, R, B)) = \sum_{i=1}^{R} P(A_i|I(N, R, B)) = \frac{R}{N},$$

quindi la probabilità di estrarre una pallina di un certo colore (rosso) in un urna contenente N palline di cui R è data dal rapporto R/N. Similmente, oppure per differenza:

$$P(B_1|I(N, R, B)) = \frac{B}{N} = 1 - \frac{R}{N}.$$

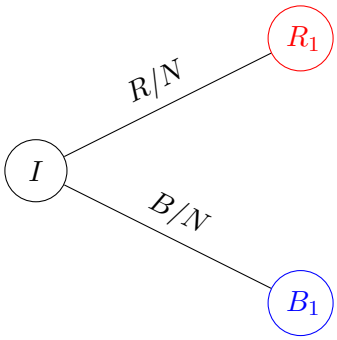


FIGURA 4.1. Albero relativo alla prima estrazione (con $I = I(N, R, B)$).

4.2. **Seconda estrazione.** Supponiamo che ci sia una seconda estrazione (supponiamo quindi che sia $N \geq 2$). Ci chiediamo quale sia la probabilità di estrarre una rossa (R_2). È chiaro però che questa può dipendere (e in effetti dipende) dal sapere l'esito della prima estrazione. Supponiamo di sapere che la prima pallina estratta è rossa: allora la seconda estrazione è un problema *del tutto equivalente* ad una prima estrazione da un'urna che contiene $N-1$ palline di cui $R-1$ rosse (perché abbiamo tolto una pallina rossa) e B blu. Più brevemente, stiamo dicendo che l'informazione R_1, ai fini della seconda estrazione, equivale all'informazione $I(N-1, R-1, B)$, e quindi

$$P(R_2|I(N,R,B) \cap R_1) = P(R_1|I(N-1,R-1,B)) = \frac{R-1}{N-1} \quad \text{e}$$

$$P(B_2|I(N,R,B) \cap R_1) = P(B_1|I(N-1,R-1,B)) = \frac{B}{N-1}.$$

Similmente, sapendo che la prima pallina estratta è blu, otteniamo che l'urna contiene $N-1$ palline, R rosse e $B-1$ blu, quindi

$$P(R_2|I(N,R,B) \cap B_1) = \frac{R}{N-1} \quad \text{e} \quad P(R_2|I(N,R,B) \cap B_1) = \frac{B-1}{N-1}.$$

Possiamo quindi estendere l'albero di Figura (4.1) come in Figura (4.2). Grazie a questo albero, possiamo allora rispondere anche alla domanda: qual è la probabilità di R_2 se non ci viene comunicato l'esito della prima estrazione (quindi rispetto all'informazione $I(N,R,B)$)? Basta sommare i prodotti dei (2) cammini che portano ad R_2: troviamo

$$\begin{aligned} P(R_2|I(N,R,B)) &= \frac{R}{N} \cdot \frac{R-1}{N-1} + \frac{B}{N} \cdot \frac{R}{N-1} = \frac{R}{N(N-1)} \cdot (R-1+B) \\ &= \frac{R(N-1)}{N(N-1)} = \frac{R}{N}, \end{aligned}$$

.

che è la stessa probabilità di estrarre rossa alla prima estrazione! Se non sappiamo che la prima estrazione è avvenuta, è come se la seconda giocasse il ruolo della prima...

Esercizio 24. Mostrare che, per ogni $i \in \{1, \dots, N\}$, si ha

$$P(R_i|I(N,R,B)) = \frac{R}{N}.$$

Suggerimento Si può ragionare per induzione: ad esempio, per calcolare $P(R_3|I(N,R,B))$, immaginiamo di sapere l'esito della prima estrazione: se la prima pallina è rossa,

$$P(R_3|I(N,R,B)\cap R_1) = P(R_2|I(N-1,R-1,B)) = \frac{R-1}{N-1}$$

perché la terza estrazione è equivalente ad una seconda estrazione (tenendo conto che il contenuto dell'urna è cambiato). Se è blu,

$$P(R_3|I(N,R,B)\cap B_1) = P(R_2|I(N-1,R,B)) = \frac{R}{N-1}$$

(disegnate l'albero associato a questo ragionamento). Ricomponendo queste due alternative, si trova

$$P(R_3|I(N,R,B)) = \frac{R}{N}\cdot\frac{R-1}{N-1} + \frac{B}{N}\frac{R}{N-1} = \frac{R}{N}.$$

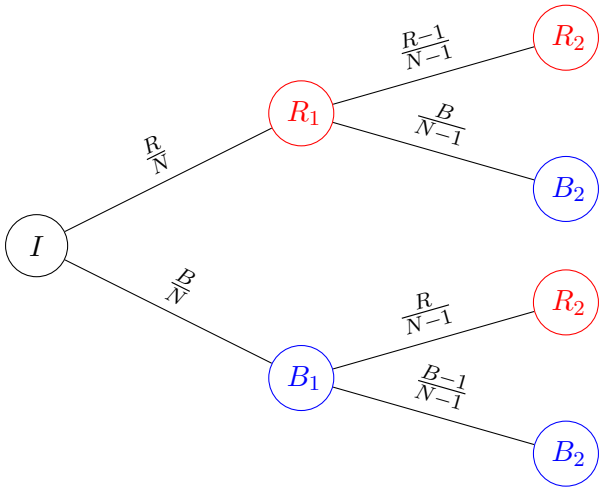


FIGURA 4.2. Albero relativo alla seconda estrazione (con $I = I(N,R,B)$).

4.3. **Estrazione di una specifica sequenza ordinata.** Supponiamo che vengano effettuate $n \le N$ estrazioni. Ci chiediamo quale sia la probabilità (rispetto all'informazione iniziale) di ottenere esattamente $r \le R$ palline rosse e $b \le B$ blu (con $b = n - r$) in una *specifica* sequenza ordinata. Ad esempio, con $n = 4$, $r = 2$ e $b = 2$, possiamo chiederci la probabilità di ottenere la sequenza R_1, B_2, B_3, R_4. Si tratta quindi di calcolare l'intersezione di tali eventi, e quindi per la regola del prodotto, (posta $I = I(N,R,B)$),

$$\begin{aligned} P(R_1\cap B_2\cap B_3\cap R_4|I) &= P(R_1|I)P(B_2|R_1\cap I)P(B_3|B_2\cap R_1\cap I)P(R_4|B_3\cap B_2\cap R_1\cap I) \\ &= \frac{R}{N}\cdot\frac{B}{N-1}\cdot\frac{B-1}{N-2}\cdot\frac{R-1}{N-3}, \end{aligned}$$

avendo ragionato per le estrazioni terza e quarta, in modo analogo a quanto fatto nella seconda estrazione. Mentre l'ordine di apparizione dei colori (rosso, blu, blu e poi rosso) è importante nella definizione dell'evento

$R_1 \cap B_2 \cap B_3 \cap R_4$, possiamo notare che la probabilità invece non dipende dall'ordine. Ad esempio, calcoliamo la probabilità della sequenza in cui prima appaiono le due rosse e poi le due blu:

$$P(R_1 \cap R_2 \cap B_3 \cap B_4|I) = P(R_1|I)P(R_2|R_1 \cap I)P(B_3|R_2 \cap R_1 \cap I)P(B_4|B_3 \cap R_2 \cap R_1 \cap I)$$
$$= \frac{R}{N} \cdot \frac{R-1}{N-1} \cdot \frac{B}{N-2} \cdot \frac{B-1}{N-3},$$

che è uguale a quella trovata prima (basta scambiare i fattori)

$$\frac{R}{N} \cdot \frac{B}{N-1} \cdot \frac{B-1}{N-2} \cdot \frac{R-1}{N-3} = \frac{R}{N} \cdot \frac{R-1}{N-1} \cdot \frac{B}{N-2} \cdot \frac{B-1}{N-3}.$$

Questo ragionamento si potrebbe fare in generale, e si ottiene il seguente risultato: *la probabilità di ottenere una specifica sequenza ordinata dipende solamente dal numero $r \leq R$ di palline rosse che contiene e il numero $b \leq B$ di palline blu.* Per calcolarla, basta ragionare nel caso della sequenza in cui escono prima tutte le rosse e poi tutte le blu. Si trova l'espressione

(4)
$$P(R_1R_2 \dots R_rB_{r+1}B_{r+2} \dots B_n|I(N,R,B)) =$$
$$= \frac{R \cdot (R-1) \cdot \ldots \cdot (R-r+1) \cdot B \cdot (B-1) \cdot \ldots \cdot (B-b+1)}{N \cdot (N-1) \cdot \ldots \cdot (N-n+1)},$$

in cui dobbiamo calcolare r fattori corrispondenti alle palline rosse, b fattori corrispondenti alle blu e a denominatore gli n fattori relativi alle palline "possibili".

4.4. **Legge ipergeometrica.** E se l'ordine in cui otteniamo le diverse palline non fosse importante? Precisamente, ci chiediamo quale sia la probabilità (rispetto all'informazione iniziale), effettuando $n \leq N$ estrazioni, di ottenere esattamente $r \leq R$ palline rosse e $b \leq B$ (con $b = n - r$). Equivalentemente, possiamo pensare di estrarre n palline in una sola volta e di chiederci la probabilità di ottenere r rosse e b blu.

Per calcolare questa probabilità, possiamo usare il risultato precedente: se σ è una possibile sequenza ordinata di n palline contenente esattamente r rosse e b blu e poniamo

$$A_\sigma = \text{"si estraggono le } n \text{ palline nella sequenza } \sigma\text{"},$$

allora gli eventi A_σ, al variare di $\sigma \in \Sigma$, dove Σ è l'insieme delle possibili sequenze con r rosse e b blu, sono a due a due incompatibili[4]. L'evento che ci interessa è

$$A = \text{"in } n \text{ estrazioni si ottengono } r \text{ rosse e } b \text{ blu"} = \bigcup_{\sigma \in \Sigma} A_\sigma$$

e quindi per la proprietà di additività,

$$P(A|I(N,R,B)) = \sum_{\sigma \in \Sigma} P(A_\sigma|I(N,R,B)).$$

D'altra parte, le probabilità $P(A_\sigma|I(N,R,B))$ sono tutte le stesse, e date dalla formula (4), quindi per concludere basterà moltiplicare la quantità in (4) per il *numero* delle possibli sequenze $\sigma \in \Sigma$. Queste sequenze sono

[4] ma non un sistema di alternative!

tante quante i sottoinsiemi dell'insieme $\{1, 2, \dots, n\}$ contenenti esattamente r elementi: infatti a ciascuna sequenza possiamo far corrispondere l'insieme delle r posizioni in cui la pallina è rossa (e le rimanenti saranno blu). È noto allora che tale numero è il coefficiente binomiale

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} = \frac{n!}{r!b!} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-r+1)}{r \cdot (r-1) \cdot \ldots \cdot 2 \cdot 1}.$$

In conclusione, troviamo che

$$\begin{aligned}
&P(\text{“in } n \text{ estrazioni si ottengono } r \text{ rosse e } b \text{ blu”} | I(N,R,B)) = \\
&= \binom{n}{r} \cdot \frac{R \cdot (R-1) \cdot \ldots \cdot (R-r+1) \cdot B \cdot (B-1) \cdot \ldots \cdot (B-b+1)}{N \cdot (N-1) \cdot \ldots \cdot (N-n+1)} \\
&= \frac{n!}{r!b!} \cdot \frac{R \cdot (R-1) \cdot \ldots \cdot (R-r+1) \cdot B \cdot (B-1) \cdot \ldots \cdot (B-b+1)}{N \cdot (N-1) \cdot \ldots \cdot (N-n+1)} \\
&= \frac{R \cdot (R-1) \cdot \ldots \cdot (R-r+1)}{r!} \cdot \frac{B \cdot (B-1) \cdot \ldots \cdot (B-b+1)}{b!} \cdot \frac{n!}{N \cdot (N-1) \cdot \ldots \cdot (N-n+1)} \\
&= \binom{R}{r} \cdot \binom{B}{b} / \binom{N}{n}.
\end{aligned}$$

Questa espressione è anche detta *legge ipergeometrica*. Un'altra interpretazione, in termini di casi favorevoli e casi possibili è la seguente: in n estrazioni da N palline i casi possibili, siccome l'ordine non è importante, sono $\binom{N}{n}$, mentre i favorevoli sono le estrazioni in cui ci sono esattamente r rosse b blu, quindi il prodotto $\binom{R}{r} \cdot \binom{B}{b}$.

Esempio 25. Supponiamo che sia $N = 9$, $R = 6$, $B = 3$ e sia $n = 3$, $r = 1$, $b = 2$. Allora la probabilità di ottenere in 3 estrazioni senza rimpiazzo esattamente 1 pallina rossa e due blu (o, equivalentemente, di trovare 1 pallina rossa e due blu estraendone 3) è

$$\binom{6}{1}\binom{3}{2} / \binom{9}{3} = \frac{6 \cdot 3 \cdot 3 \cdot 2}{9 \cdot 8 \cdot 7}.$$

Proviamo a ripetere il ragionamento visto sopra. Le sequenze in cui si può ottenere 1 rossa e due blu sono 3 (a seconda che la rossa sia prima, seconda o terza estratta), e ciascuna ha probabilità $6 \cdot 3 \cdot 2/(9 \cdot 8 \cdot 7)$. Otteniamo quindi la stessa probabilità, $3 \cdot 6 \cdot 3 \cdot 2/(9 \cdot 8 \cdot 7)$.

5. Probabilità “inversa”

Nella sezione precedente, ci siamo occupati perlopiù di probabilità che riguardavano eventi presenti o futuri, al più sapendo oppure no qualche informazione legata al passato (estrazioni precedenti). Questo tipo di problemi spesso è noto come probabilità “diretta”, per distinguere invece dalla probabilità “inversa”, che invece si occupa di determinare la probabilità che qualcosa nel passato sia accaduto, sapendo che una proposizione che riguarda il presente o il futuro è vera. Notiamo subito però che questa distinzione è completamente “artificiale”: nelle regole di calcolo della probabilità non c'è riferimento al tempo, o a cause ed effetti, e infatti le stesse regole permettono di risolvere entrambi i tipi di problemi.

Più precisamente il problema, è il seguente: siamo interessati ad una probabilità $P(B|A \cap I)$, ma conosciamo invece $P(A|B \cap I)$. Ad esempio, se siamo un giudice che deve decidere se un imputato è colpevole o no, e poniamo

(5)
$$B = \text{“l'imputato è colpevole”}, \quad A = \text{“l'imputato si trovava sulla scena del delitto”},$$

allora $P(A|B \cap I)$ sarà molto grande, ma non è la probabilità che ci interessa.

Come scambiare i ruoli di A e B? La seguente formula di Bayes è una regola utile allo scopo.

Proposizione 26 (Formula di Bayes)**.** *Siano A, B ed I eventi. Allora vale*

$$P(B|A \cap I) = \frac{P(A|B \cap I) \cdot P(B|I)}{P(A|I)} \tag{6}$$

(purché tutte le probabilità condizionate abbiano significato e $P(A|I) > 0$).

Prima di dimostrare la validità della formula, conviene osservare che si può leggere in due modi:

$$P(B|A \cap I) = P(A|B \cap I) \cdot \frac{P(B|I)}{P(A|I)}$$

oppure

$$P(B|A \cap I) = P(B|I) \cdot \frac{P(A|B \cap I)}{P(A|I)}.$$

Nel primo modo, ci permette di “scambiare” il ruolo di B con A, (in un certo senso è come se scambiassimo l'ipotesi con la tesi, un grave errore nella logica deduttiva, ma permesso nel calcolo delle probabilità!). Nel secondo modo, stiamo invece “aggiornando” la probabilità di B rispetto alla nuova informazione $A \cap I$: per farlo, basta moltiplicare la probabilità $P(B|I)$ per il termine

$$\frac{P(A|B \cap I)}{P(A|I)},$$

che è anche detto a volte rapporto di verosimiglianza.

In entrambi i punti di vista, la formula è utile per calcolare $P(B|A \cap I)$ se conosciamo le tre probabilità nel membro di destra: applicarla male, in molte circostanze, potrebbe semplicemente aumentare il numero di probabilità che vanno calcolate!

Dimostrazione Formula di Bayes. È una semplice conseguenza della regola del prodotto. Infatti, possiamo scrivere

$$P(A \cap B|I) = P(A|I) \cdot P(B|A \cap I)$$

ma anche, essendo $A \cap B = B \cap A$,

$$P(A \cap B|I) = P(B \cap A|I) = P(B|I) \cdot P(A|B \cap I).$$

Di conseguenza,

$$P(A|I) \cdot P(B|A \cap I) = P(B|I) \cdot P(A|B \cap I)$$

e dividendo ambo i membri per $P(A|I)$ (che è positiva per ipotesi) si trova la (6). □

Esempio 27. Nell'esempio del giudice, con A e B come in (5), supponiamo di aver stimato che

$$P(A|B \cap I) \approx 1,$$

ossia una probabilità molto alta. D'altra parte, se sappiamo che la probabilità che l'imputato si trovasse sulla scena del delitto è pure molto alta (ad esempio, ci passa tutti i giorni all'ora in cui il delitto si è compiuto), allora $P(A|I) \approx 1$. Ne deduciamo che il rapporto di verosimiglianza

$$\frac{P(A|B \cap I)}{P(A|I)} \approx 1,$$

quindi la probabilità che sia colpevole non cambia di molto, pur ammettendo la prova che si trovasse sulla scena del delitto. Se invece è molto improbabile che l'imputato si trovasse sulla scena del delitto, allora $P(A|I) \approx 0$ e quindi

$$\frac{P(A|B \cap I)}{P(A|I)} \approx \frac{1}{0}$$

è molto grande: la probabilità che sia colpevole viene amplificata, se riteniamo vera questa ipotesi.

Esempio 28. Torniamo al modello dell'urna della sezione precedente (estrazioni senza reimmissione). Possiamo usare la formula di Bayes per calcolare la probabilità che alla prima estrazione si trovi una pallina rossa ($R1$) sapendo che alla seconda è stata estratta una blu ($B2$):

$$\begin{aligned} P(R1|B2 \cap I(N,R,B)) &= P(B2|R1 \cap I(N,R,B)) \cdot \frac{P(R1|I(N,R,B))}{P(B2|I(N,R,B))} \\ &= \frac{B}{N-1} \cdot \frac{R}{N} \cdot \frac{N}{B} = \frac{R}{N-1}. \end{aligned}$$

Osserviamo che questa coincide con la probabilità di estrarre rossa alla seconda, sapendo che nella prima è stata estratta blu: ecco un altro esempio per cui, dal punto di vista della probabilità, l'ordine delle estrazioni non è rilevante (l'altro era il fatto che $P(R_i|I(N,R,B)) = R/N$ per ogni $i \in \{1,\dots,N\}$).

A volte si combina la formula di Bayes con la decomposizione della probabilità di un evento B rispetto a un sistema di alternative $A_1, \dots A_n$, per ottenere la probabilità che sia vera un'alternativa sapendo che B si è realizzato. Ad esempio, potremmo pensare che ci siano n indiziati per un delitto e, avendo acquisito una prova B, l'investigatore deve aggiornare tutte le probabilità dell'evento $A_i :=$ "l'indiziato i è colpevole". Si tratta semplicemente di applicare la formula di Bayes per ciascuna coppia A_i, B, ottenendo

$$P(A_i|B \cap I) = P(A_i|I) \cdot \frac{P(B|A_i \cap I)}{P(B|I)}$$

e di decomporre il denominatore usando il sistema di alternative

$$P(B|I) = \sum_{j=1}^{n} P(B|A_j \cap I) P(A_j|I).$$

In conclusione si trova questa identità, a volte detta di *probabilità delle cause*:

$$P(A_i|B \cap I) = P(A_i|I) \cdot \frac{P(B|A_i \cap I)}{\sum_{j=1}^{n} P(B|A_j \cap I)P(A_j|I)}.$$

Esercizio 29 (Formula di Bayes "parziale"). Mostrare che, per eventi A, B, I, J, si ha

$$P(B \cap J|A \cap I) = \frac{P(A \cap J|B \cap I)P(B|I)}{P(A|I)},$$

(purché le probabilità condizionate siano definite e $P(A|I) > 0$). Questo permette di scambiare soltanto una "parte" dell'evento di cui si calcola la probabilità, lasciando l'informazione J al suo posto.

Esempio 30 (Paradosso di Bertrand). Davanti a noi si trovano tre scatole indistinguibili dall'esterno, ciascuna contenente due palline. Una contiene due palline bianche, un'altra due palline nere e la terza una pallina bianca e una nera. Scegliamo una scatola ed estraiamo una pallina. Sapendo che la pallina estratta è bianca, qual è la probabilità che l'altra pallina nella scatola sia bianca?

Per risolvere il problema, introduciamo il sistema di alternative

$$BB = \text{"la scatola scelta contiene due palline bianche"}$$

$$NN = \text{"la scatola scelta contiene due palline nere"}$$

$$BN = \text{"la scatola scelta contiene una pallina bianca e una nera"},$$

e il sistema di alternative

$$E_B = \text{"la pallina estratta è bianca"}, \quad E_N = \text{"la pallina estratta è nera"}.$$

Sulla base del testo, diamo probabilità uniforme alle alternative BB, NN, BN rispetto all'informazione che abbiamo prima di fare l'estrazione (che indichiamo con Ω). Sapendo quale scatola è stata scelta, possiamo facilmente assegnare le probabilità ad E_B ed E_N (Figura 5.1). La probabilità richiesta si può esprimere usando gli eventi sopra come $P(BB|E_B)$. Possiamo usare la formula di Bayes,

$$P(BB|E_B) = P(E_B|BB) \cdot \frac{P(BB|\Omega)}{P(E_B|\Omega)},$$

e notiamo che l'unica probabilità che rimane da calcolare è $P(E_B|\Omega)$, che otteniamo dall'albero come

$$P(E_B|\Omega) = \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{3} \cdot \frac{3}{2} = \frac{1}{2}.$$

Concludiamo quindi che

$$P(BB|E_B) = 1 \cdot \frac{1}{3} \cdot \frac{2}{1} = \frac{2}{3}.$$

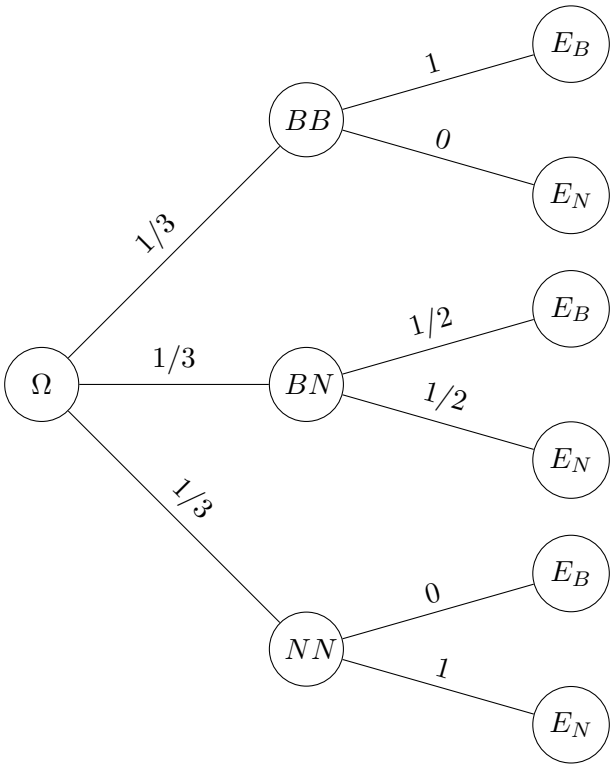


FIGURA 5.1. Albero corrispondente alla situazione del paradosso di Bertand.

6. Il modello dell'urna (II) estrazioni con reimmissione

In questa sezione riprendiamo la trattazione del problema delle estrazioni da un'urna, cambiandone però lo svolgimento. Supporremo infatti che, dopo ciascuna estrazione, la pallina estratta viene rimessa nella scatola. È chiaro che questo esperimento non cambia nulla ai fini della prima estrazione, per cui, riprendendo gli eventi

$$R_i = \text{“la pallina estratta all'estrazione } i \text{ è di colore rosso”},$$

$$R_i^c = B_i = \text{“la pallina estratta all'estrazione } i \text{ è di colore blu”}.$$

per $i \in \{1, 2, 3, \dots, n\}$ dove n è il numero di estrazioni che si effettuano (stavolta n può essere arbitrariamente grande), e

$$I(N, R, B) = \text{“l'urna contiene } N \text{ palline di cui } R \text{ rosse e } B \text{ blu”}$$

avremo di nuovo

$$P(R_1 | I(N, R, B)) = \frac{R}{N} \quad \text{e} \quad P(B_1 | I(N, R, B)) = \frac{B}{N}.$$

6.1. **Estrazioni successive.** Alla seconda estrazione, la situazione evidentemente cambia. Innanzitutto sappiamo che il numero e il tipo di palline all'interno dell'urna è lo stesso della prima, quindi abbiamo $I(N, R, B)$. Tuttavia, è lecito chiedersi se sapere l'esito della prima estrazione possa influenzare la probabilità della seconda. Ad esempio, se la prima pallina estratta è rossa e poi viene rimessa in cima alla scatola e la persona che estrae è pigra

tenderà a riprendere la pallina rossa appena pescata, oppure, se viene rimessa in fondo, magari la persona tenderà a pescare proprio dal fondo della scatola...

Una soluzione possibile è di rimettere la pallina all'interno e "agitare la scatola", in modo da rendere ancora più difficile determinare dove la pallina rimessa è finita. Per quanto pure questa procedura si possa criticare, è chiaro che l'effetto finale che vogliamo ottenere è che l'informazione della prima estrazione sia a tutti gli effetti *inutilizzabile* ai fini del calcolo della probabilità della seconda, quindi l'unica informazione utile è $I(N, R, B)$ ed è come effettuare la prima estrazione da un'urna contenente N palline di cui R rosse e B blu. In termini matematici, scriveremo allora (Figura 6.1)

$$P(R_2|R_1 \cap I(N, R, B)) = P(R_2|I(N, R, B)) = \frac{R}{N}.$$

e similmente, se la prima estrazione fosse blu,

$$P(R_2|B_1 \cap I(N, R, B)) = P(R_2|I(N, R, B)) = \frac{R}{N}.$$

Nella prossima sezione formalizzeremo meglio questo concetto, dicendo che gli eventi R_2, B_2, relativi alla seconda estrazione sono *indipendenti* da quelli relativi alla prima ossia R_1, B_2 (sapendo l'informazione $I(N, R, B)$).

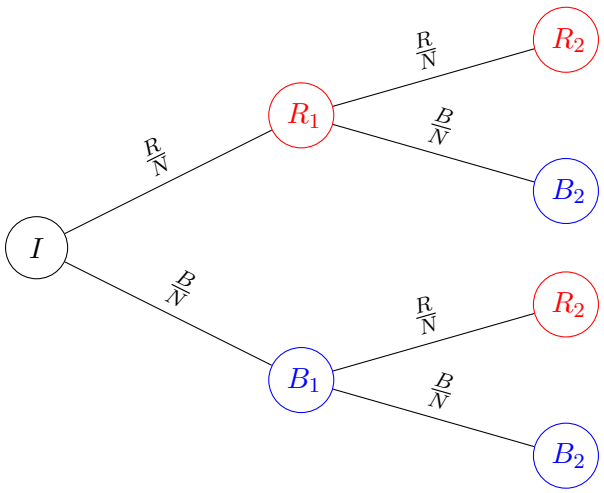


FIGURA 6.1. Albero relativo alla seconda estrazione (con $I = I(N, R, B)$).

Come possiamo ragionare con la terza, quarta ecc. estrazione? Di nuovo, ci si può chiedere se conoscere gli esiti di tutte le estrazioni precedenti possa influenzare il grado di fiducia di una estrazione successiva. Ad esempio, se l'urna contiene una pallina rossa e una blu, e vedessimo solamente palline rosse per un gran numero di estrazioni, chiunque avrebbe qualche ragionevole dubbio sul modo in cui le palline vengono estratte. Eppure, volendo trattare un esperimento ideale, supporremo comunque che *qualunque informazione* dalle passate estrazioni non possa cambiare il grado di fiducia dell'estrazione successiva, che quindi è calcolata come se fosse la prima

estrazione da un'urna contenente N palline di cui R rosse e B blu. Diremo che gli eventi relativi alle diverse estrazioni sono *indipendenti* (sapendo $I(N,R,B)$).

6.2. **Estrazione di una specifica sequenza ordinata.** Supponiamo che siano effettuate $n \geq 1$ estrazioni. Come nel caso delle estrazioni con reimmissione, ci chiediamo quale sia la probabilità (rispetto all'informazione iniziale) di ottenere esattamente r palline rosse e b blu (con $b = n-r$) in una *specifica* sequenza ordinata. Ad esempio, poniamo $n = 4$, $r = 2$ e $b = 2$, e ci chiediamo la probabilità di ottenere la sequenza R_1, B_2, B_3, R_4, ossia l'intersezione di tali eventi, e quindi per la regola del prodotto, (posta $I = I(N,R,B)$),

$$\begin{aligned} P(R_1 \cap B_2 \cap B_3 \cap R_4|I) &= P(R_1|I)P(B_2|R_1 \cap I)P(B_3|B_2 \cap R_1 \cap I)P(R_4|B_3 \cap B_2 \cap R_1 \cap I) \\ &= \frac{R}{N} \cdot \frac{B}{N} \cdot \frac{B}{N} \cdot \frac{R}{N}, \end{aligned}$$

avendo usato l'ipotesi di *indipendenza* delle estrazioni successive alle precedenti. Anche stavolta, è facile rendersi conto che l'ordine di apparizione dei colori non è importante ai fini del valore della probabilità: ad esempio,

$$\begin{aligned} P(R_1 \cap R_2 \cap B_3 \cap B_4|I) &= P(R_1|I)P(R_2|R_1 \cap I)P(B_3|R_2 \cap R_1 \cap I)P(B_4|B_3 \cap R_2 \cap R_1 \cap I) \\ &= \frac{R}{N} \cdot \frac{R}{N} \cdot \frac{B}{N} \cdot \frac{B}{N}, \end{aligned}$$

che è uguale a quella trovata prima (basta scambiare i fattori). Anche nel caso delle estrazioni con reimmissione, *la probabilità di ottenere una specifica sequenza ordinata dipende solamente dal numero r di palline rosse che contiene e il numero b di palline blu* e per calcolarla, basta ragionare nel caso della sequenza in cui escono prima tutte le rosse e poi tutte le blu. Stavolta però si trova

$$\begin{aligned} &P(R_1R_2 \ldots R_rB_{r+1}B_{r+2} \ldots B_n|I(N,R,B)) = \\ &\qquad = \frac{R^rB^b}{N^n} = \left(\frac{R}{N}\right)^r \left(1 - \frac{R}{N}\right)^{n-r}, \end{aligned} \tag{7}$$

in cui l'ultima scrittura evidenzia il ruolo del rapporto R/N e i numeri r ed n: se pensiamo ad estrarre una pallina rossa come un "successo" ed una blu come un "insuccesso", con probabilità di rispettivamente di $p = R/N$ ed $1 - p = 1 - R/N$, la probabilità di ottenere una specifica sequenza di r successi in n tentativi "indipendenti", risulta

$$p^r(1-p)^{n-r}.$$

6.3. **Legge binomiale.** Possiamo quindi porci la stessa domanda che nel caso di estrazioni senza rimpiazzo ci ha condotti alla legge ipergeometrica: e se l'ordine in cui otteniamo le diverse palline non fosse importante? Qual è la probabilità di ottenere, in n estrazioni con reimmissione, esattamente r palline rosse e b blu (in un qualunque ordine)?

Ripetendo il ragionamento, siccome abbiamo visto che la probabilità di ottenere una determinata sequenza ordinata dipende solamente dai numeri delle palline e non dall'ordinamento, anche stavolta basta moltiplicare (7) per il numero di sequenze ordinate lunghe n con r palline rosse e b blu, che

abbiamo già visto essere il coefficiente binomiale $\binom{n}{r} = \binom{r+b}{r}$. Troviamo allora la probabilità

$$P(\text{ “}r \text{ rosse e } b \text{ blu, estrazioni con reimmissione”}|I(N,R,B)) = \binom{r+b}{r}\cdot\frac{R^r B^b}{N^n}.$$

Se pensiamo ancora all'estrazione di una pallina rossa come un successo, con probabilità $p = R/N$, otteniamo che *la probabilità di ottenere esattamente r successi in n prove “indipendenti” (in un qualunque ordine)* è data da

$$P(\text{“}r \text{ successi in } n \text{ prove indipendenti”}|p = \text{prob. di un successo}) = \binom{n}{r}p^r(1-p)^{n-r},$$

che viene anche detta *legge binomiale*.

Esempio 31. Supponiamo che sia $N = 9$, $R = 6$, $B = 3$ e sia $n = 3$, $r = 1$, $b = 2$. Allora la probabilità di ottenere in 3 estrazioni con rimpiazzo esattamente 1 pallina rossa e due blu (o, equivalentemente, di trovare 1 pallina rossa e due blu estraendone 3) è

$$\binom{3}{1}\frac{6\cdot 3^2}{9^3} = \frac{3\cdot 6\cdot 3^2}{9^3} \approx 0,22.$$

Nel caso di estrazioni senza reimmissione avevamo trovato che la probabilità dello stesso evento è

$$\frac{3\cdot 6\cdot 3\cdot 2}{9\cdot 8\cdot 7} \approx 0,21.$$

7. Eventi indipendenti

Riflettiamo in modo più astratto sulla ipotesi che ci ha permesso di trattare il caso di estrazioni con reimmissione. Partendo dai due eventi $A = R_1$ e $B = R_2$, ci siamo chiesti in quale modo utilizzare l'informazione che nella prima estrazione l'esito fosse rosso, per calcolare la probabilità che pure la seconda pallina estratta fosse rossa:

$$P(B|A\cap I)$$

dove $I = I(N,R,B)$. Pur potendo immaginare diversi scenari che potessero favorire o sfavorire l'estrazione di una pallina rossa, rispetto ad un urna completamente nuova, in un certo senso ci siamo “arresi”, dicendo che l'urna è sufficientemente agitata per cui non riusciamo ad estrarre informazione utile dall'evento A (se non la stessa informazione I). Perciò abbiamo posto

$$P(B|A\cap I) = P(B|I),$$

e similmente abbiamo argomentato per più di due estrazioni.

Diamo ora una definizione matematica di questa ipotesi, detta di *indipendenza* probabilistica.

7.1. **Due eventi.** Dati eventi A, B, I, diremo che B è indipendente da A (rispetto alla informazione I) se vale

$$P(B|A\cap I) = P(B|I), \tag{8}$$

(purché tutte le probabilità condizionate abbiano senso, in particolare per Kolmogorov deve essere $P(A\cap I|\Omega) > 0$).

Come abbiamo visto, questo concetto va inteso come una ipotesi che inseriamo nella descrizione delle situazioni: codifica il fatto che una (possibilmente) nuova informazione A non modifica il grado di fiducia di B, rispetto alla informazione che già si possiede, I.

Come gli altri concetti della probabilità, pure l'indipendenza probabilistica però non va interpretata come un concetto di indipendenza "fisica": potrebbe addirittura essere che A sia una causa fisica di B, ma semplicemente la nostra parziale informazione I non ci permette di saperlo. Ad esempio, una persona poco istruita (I) potrebbe pensare che le maree B non abbiano nulla a che vedere con la posizione della luna A e quindi dire che queste siano indipendenti. Tuttavia, dopo osservazioni ed esperimenti, potrebbe acquisire nuova informazione J che invece le rende tutt'altro che indipendenti.

L'esempio sopra mostra anche che l'indipendenza di B da A dipende dall'informazione I che suppone vera: cambiare I potrebbe cambiare la validità dell'indipendenza.

Una proprietà interessante dell'indipendenza probabilistica è che essa è *simmetrica* nei ruoli di A e B: se B è indipendente da A allora A è indipendente da B (rispetto alla stessa informazione I). Basta infatti usare la formula di Bayes

$$P(A|B\cap I) = P(B|A\cap I)\cdot\frac{P(A|I)}{P(B|I)} = P(B|I)\cdot\frac{P(A|I)}{P(B|I)} = P(A|I).$$

Notiamo che questo è rigorosamente vero se $P(B|I) > 0$. Questa simmetria ha spinto i matematici a definire l'indipendenza tra A e B (rispetto ad I) tramite la validità dell'identità

$$P(A\cap B|I) = P(A|I)\cdot P(B|I) \tag{9}$$

che non richiede $P(A|I) > 0$ o $P(B|I) > 0$. Per vedere l'equivalenza con la definizione (8), se $P(A|I) > 0$, basta usare la regola del prodotto:

$$P(A|I)\cdot P(B|I) = P(A\cap B|I) = P(A|I)P(B|A\cap I)$$

e dividere ambo i membri per $P(A|I) > 0$. Per questo invece di dire B indipendente da A, si dice semplicemente che A e B sono eventi indipendenti.

Notiamo infine che la definizione (9) ricorda anche l'identità tra i valori di verità della logica Boole, per cui $v(A\wedge B) = v(A)\cdot v(B)$.

Osservazione 32 (Indipendenza e incompatibilità). Un errore purtroppo frequente è di confondere il concetto di eventi indipendenti con quello di eventi incompatibili. In realtà, più che simili sono due concetti completamente estranei l'uno all'altro: infatti l'indipendenza di B da A (rispetto I) afferma che l'informazione che A sia vera non modifica la probabilità di B, mentre l'incompatibilità di B da A ci permette di dedurre subito, se A è vera, che B è falsa, quindi si avrebbe

$$P(B|I) = P(B|A\cap I) = 0.$$

Quindi l'unico modo per cui due eventi indipendenti siano incompatibili è che (almeno) uno abbia probabilità nulla (rispetto ad I).

Esercizio 33. Siano A, B eventi indipendenti (rispetto ad I). Mostrare che A^c e B sono indipendenti (rispetto ad I). Dedurne che pure A^c e B^c sono indipendenti (rispetto ad I).

7.2. **Più di due eventi.** Passiamo ora alla definizione di indipendenza per più di due eventi, cominciando da tre, A_1, A_2, A_3. Una definizione ingenua sarebbe di chiedere che siano indipendenti a due a due, ossia

$$P(A_i|A_j \cap I) = P(A_i|I), \quad \text{per ogni } i, j \in \{1,2,3\} \text{ con } i \neq j. \tag{10}$$

Tuttavia, può accadere che l'informazione accumulata di due di questi possa cambiare il grado di fiducia sul terzo.

Esercizio 34. Un'urna contiene due palline, una rossa e una blu. Si effettuano due estrazioni con reimmissione, e poniamo $A_1 = R_1$, $A_2 = R_2$ come descritto nelle sezioni precedenti e infine

$A_3 =$ "le palline estratte sono entrambe rosse oppure entrambe blu" $= (R_1 \cap R_2) \cup (B_1 \cap B_2)$.

Mostrare che A_1, A_2 e A_3 sono a due a due indipendenti (rispetto ad $I = I(2,1,1)$) ma

$$P(A_3|A_1 \cap A_2 \cap I) \neq P(A_3|I).$$

Per questo, la definizione di *tre eventi indipendenti* richiede pure che

$$P(A_i|A_j \cap A_k \cap I) = P(A_i|I) \quad \text{per ogni } i, j, k \in \{1,2,3\} \text{ con } i \neq j,\ i \neq k.$$

Notiamo che se $j = k$, si recupera la definizione di eventi a due a due indipendenti. Si può anche scrivere la condizione in modo analogo alla (9): per ogni $i \neq j$ si ha

$$P(A_i \cap A_j|I) = P(A_i|I)P(A_j|I)$$

e inoltre

$$P(A_1 \cap A_2 \cap A_3|I) = P(A_1|I)P(A_2|I)P(A_3|I). \tag{11}$$

Passiamo al caso generale: come definire $n \geq 3$ eventi $A_1, \dots A_n$ indipendenti tra loro (rispetto all'informazione I)? Ci sono in realtà tanti modi, tutti equivalenti tra loro.

Ad esempio, un modo veloce (ma non molto trasparente) è di ragionare *ricorsivamente*: supponendo di aver definito $n-1$ eventi indipendenti, per definire $n \geq 3$ eventi basterà dire che *ogni sottofamiglia* di $n-1$ eventi presi tra questi deve risultare di eventi indipendenti, e inoltre vale l'analogo di (11), ossia

$$P(A_1 \cap A_2 \cap \ldots \cap A_n|I) = P(A_1|I) \cdot P(A_2|I) \cdot \ldots \cdot P(A_n|I).$$

Un modo apparentemente più complicato, ma in realtà equivalente, è il seguente: gli eventi $A_1, \dots A_n$ si dicono indipendenti tra loro (rispetto ad I) se, comunque si prendano due sottoinsiemi F, $G \subseteq \{1,2,\dots,n\}$ *disgiunti*, ossia tali che $F \cap G = \emptyset$, i due eventi

$$\bigcap_{i \in F} A_i \quad \text{e} \quad \bigcap_{j \in G} A_j$$

sono indipendenti tra loro (rispetto ad I). In particolare, possiamo scrivere

$$P(\bigcap_{i \in F} A_i | \bigcap_{j \in G} A_j \cap I) = P(\bigcap_{i \in F} A_i|I)$$

(se le probabilità condizionate sono ben definite). In parole più semplici: comunque accumuliamo informazione riguardo gli eventi in G, questa non modifica il grado di fiducia sugli eventi fuori di G, e neppure sulle loro combinazioni.

L'equivalenza tra questi due (e altre possibili caratterizzazioni note) si dimostrano per induzione su n, e noi evitiamo di farlo. Per esercizio, nel caso di $n = 3$, verificate che le due definizioni sono equivalenti.

8. Variabili aleatorie (discrete)

Spesso, nei problemi, dobbiamo ragionare con *quantità* sul cui valore siamo incerti. Ad esempio,

(1) La RAM libera di un calcolatore (mentre stiamo usando oppure progettando un programma),
(2) Il numero di amicizie sui social network di un utente di una applicazione che stiamo progettando,
(3) La temperatura massima di domani a Pisa,
(4) Il voto che riceveremo all'esame di CPS,
(5) Il numero di palline rosse estratte (in un qualunque ordine) in n estrazioni da un'urna con R palline rosse totali...

Poiché sappiamo (o immaginiamo) che la quantità ha in realtà un preciso valore, possiamo descrivere quantità aleatorie mediante sistemi di alternative. Ad esempio, nel caso delle palline rosse estratte, possiamo introdurre gli eventi

$$A_i := \text{“vengono estratte esattamente } i \text{ palline rosse”}$$

al variare di $i \in \{0, 1, \dots, n\}$. Chiaramente, possiamo anche variare i in un insieme più grande, ma gli eventi in quel caso avranno probabilità nulla in ogni caso.

Esercizio 35. Per ciascun esempio sopra, fornire un sistema di alternative che descriva la quantità aleatoria.

Un modo equivalente, ma che si rivela più utile ai fini del calcolo, è di definire una “variabile” X, che assume valori tra quelli possibili per la quantità incerta, ad esempio $\{0, 1, 2, \dots, n\}$ nel caso delle palline rosse, in modo tale che $X = i$ corrisponda all'evento “la quantità assume il valore i”. In formule, scriveremo

$$\{X = i\} = \text{“la quantità assume il valore } i\text{”}.$$

Nell'esempio delle palline, quindi $\{X = i\} = A_i$, quindi vediamo che si tratta semplicemente di una riscrittura. Tuttavia, questa presenta molti vantaggi quando dobbiamo effettuare operazioni matematiche su queste quantità. Ad esempio, se invece del numero di palline rosse X fossimo interessati al numero di palline blu estratte Y, basterà scrivere $Y = n - X$, e quindi, per $i \in \{0, 1, \dots, n\}$,

$$\{Y = i\} = \{n - X = i\} = \{X = n - i\}.$$

Il simbolo di uguaglianza nella scrittura $\{X = i\}$ ci permette di lavorare con X (o eventualmente con altre quantità aleatorie che appaiono) come se

fossero classiche variabili matematiche: ecco perché il termine "variabile" aleatoria.

Possiamo quindi introdurre la seguente definizione operativa di "variabile aleatoria". In questa sezione, ragioniamo sempre solamente nel caso in cui l'insieme dei valori E è finito o numerabile (ma discreto, ad esempio $\mathbb{N}$, $\mathbb{Z}$) E. Diciamo quindi che una variabile aleatoria X a valori in E è data da un sistema di alternative

$$\{X = e\} \quad \text{al variare di } e \in E.$$

Possiamo anche scrivere per brevità "la variabile aleatoria $X \in E$" (ma per evitare confusione, è sempre meglio specificare che X è una variabile aleatoria).

Anche se abbiamo iniziato la discussione trattando di quantità numeriche, quindi $E \subseteq \mathbb{R}$, l'insieme dei valori di X può essere anche dato da colori, $E = \{R, B\}$, vettori, $E \subseteq \mathbb{R}^2$, o altro, purché sia finito o numerabile.

Per ora abbiamo definito solamente la scrittura $\{X = e\}$, al variare di $e \in E$. Tuttavia, ricordando il significato che attribuiamo a tale scrittura, è facile estendere ad altre relazioni matematiche, come

$$\{X \in F\} = \bigcup_{e \in F} \{X = e\}, \qquad \text{dove } F \subseteq E,$$

oppure, nel caso in cui $E \subseteq \mathbb{R}$, per ogni $t \in \mathbb{R}$ definiamo

$$\{X \leq t\} = \bigcup_{\substack{e \in E \\ e \leq t}} \{X = e\} \quad \text{e} \quad \{X > t\} = \{X > t\}^c = \bigcup_{\substack{e \in E \\ e > t}} \{X = e\}$$

e analogamente

$$\{X \geq t\} = \bigcup_{\substack{e \in E \\ e \geq t}} \{X = e\} \quad \text{e} \quad \{X < t\} = \{X > t\}^c = \bigcup_{\substack{e \in E \\ e < t}} \{X = e\}.$$

Osservazione 36 (Variabili aleatorie nella definizione di Kolmogorov). Prima di procedere con altre proprietà delle variabili aleatorie, è meglio chiarire che la definizione operativa data sopra è leggermente meno precisa di quella nella teoria di Kolmogorov. Infatti, avendo specificato $(\Omega, \mathcal{A}, P(\cdot|\Omega))$, la definizione di Kolmogorov per una variabile aleatoria X a valori in E è una *funzione* $X : \Omega \to E$ tale che, per ogni $e \in E$, l'immagine inversa

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

sia un evento. Poiché X è una funzione, necessariamente recuperiamo che gli eventi $\{X = e\}$ sono un sistema di alternative, che poi è l'unica proprietà che si usa in pratica nei problemi. Siccome abbiamo deciso di non occuparci della costruzione di Ω, non ci occuperemo neppure di costruire la funzione $X : \Omega \to E$, e useremo le variabili aleatorie solamente con la definizione operativa data sopra.

Osserviamo che le variabili aleatorie si "comportano" in modo piuttosto immediato quando operiamo su di esse, proprio come se fossero variabili matematiche. Ad esempio

(1) Date variabili aleatorie $X \in E$, $Y \in F$, possiamo sempre definire la variabile coppia $(X, Y) \in E \times F$ mediante

$$\{(X,Y) = (e,f)\} = \{X = e \quad \text{e} \quad Y = f\} = \{X = e\} \cap \{Y = f\}$$

al variare di e in E, ed f nell'insieme F. Questo ovviamente si generalizza anche a triple o n-uple di variabili aleatorie.

(2) Data una variabile aleatoria $X \in E$ e una *funzione* $g : E \to G$, la variabile aleatoria $g(X)$ è definita mediante

$$\{g(X) = \ell\} = \left\{X \in g^{-1}(\ell)\right\} = \bigcup_{\substack{e \in E \\ g(e) = \ell}} \{X = e\}\,.$$

(3) Combinando i due esempi sopra, otteniamo che la somma $X + Y$, prodotto $X \cdot Y$ di variabili aleatorie (discrete) è ben definito. Ad esempio, se $X \in \mathbb{N}$, $Y \in \mathbb{N}$, allora troviamo che

$$\{X + Y = n\} = \bigcup_{k=0}^{n} (\{X = k\} \cap \{Y = n - k\}) = \bigcup_{k=0}^{n} (\{X = n - k\} \cap \{Y = k\})\,.$$

Esempio 37 (variabile indicatrice). Abbiamo quindi visto che le variabili aleatorie permettono di lavorare agevolmente con sistemi di alternative. Nel caso di una alternativa semplice A, A^c, si può associare una variabile aleatoria a valori in $\{0, 1\}$, che indica 1 se vale A e 0 se vale A^c. Indichiamo con 1_A la variabile aleatoria *indicatrice* dell'evento A. In simboli

$$\{1_A = 1\} = A \quad \text{e} \quad \{1_A = 0\} = A^c.$$

Viceversa, ogni variabile X a valori in $\{0, 1\}$ può essere sempre pensata come la variabile indicatrice dell'evento $A = \{X = 1\}$. Osserviamo inoltre che, se $X = 1_A$ e $Y = 1_B$ allora $XY = 1_A \cdot 1_B = 1_{A \cap B}$.

8.1. **Legge di una variabile aleatoria.** Finora abbiamo solamente definito le variabili aleatorie come un modo agevole per trattare quantità (perlopiù) numeriche il cui valore è incerto. Ovviamente tratteremo tale incertezza usando la probabilità. In particolare, si definisce *legge* (o distribuzione) di una variabile aleatoria $X \in E$, rispetto all'informazione I, la funzione a valori in $[0, 1]$, che associa ad ogni sottoinsieme $F \subseteq E$ la probabilità che X assuma un valore tra quelli di F, ossia

$$F \subseteq E \quad \mapsto \quad P(\{X \in F\}\,|I) = P(X \in F|I).$$

Un caso particolare è quando $F = \{e\}$, per cui $\{X \in F\} = \{X = e\}$. Si trova allora la *densità* (discreta) di X,

$$e \mapsto P(X = e|I).$$

Notiamo che si ha sempre $P(X = e|I) \in [0, 1]$ e

$$\sum_{e \in E} P(X = e|I) = P(X \in E|I) = 1.$$

Osservazione 38 (densità discreta e legge). In effetti, la densità discreta è sufficiente per conoscere la legge, usando l'additività per eventi a due a due incompatibili:

$$P(X \in F|I) = P(\bigcup_{f \in F} \{X = f\}\,|I) = \sum_{f \in F} P(X = f|I).$$

Un po' impropriamente, a volte chiamiamo legge quella che precisamente sarebbe la densità discreta di X.

Esempio 39 (variabili aleatorie costanti). Questo esempio è un po' banale, ma può essere utile. Supponiamo che l'informazione I sia tale per cui $P(X = e|I) = 1$ per un certo (e necessariamente uno solo) valore $e \in E$. Allora a tutti gli effetti la variabile X non è *aleatoria* ma possiamo identificarla con il suo valore e (attenzione però, rispetto ad altre informazioni X potrebbe assumere altri valori). Viceversa, possiamo sempre pensare un numero noto (e fissato) $e \in E$ come una variabile aleatoria che assume solamente il valore e, con probabilità 1.

Esempio 40 (legge Bernoulli). L'esempio più semplice (ma non banale) di una legge è quello di una variabile che può assumere due valori. Nel caso di una variabile indicatrice $X \in \{0, 1\}$, $X = 1_A$, siccome

$$P(X = 1|I) + P(X = 0|I) = P(A|I) + P(A^c|I) = 1,$$

tutta la legge è determinata dalla singola probabilità

$$p = P(X = 1|I) = P(A|I).$$

In tal caso diciamo che la variabile aleatoria $X \in \{0, 1\}$ ha legge *Bernoulli di parametro* p. Ad esempio, se lanciamo una moneta e poniamo $X = 1$ se esce testa, $X = 0$ se esce croce, allora X ha legge *Bernoulli di parametro* $1/2$.

Esempio 41 (legge uniforme). Un altro esempio semplice si trova quando la variabile aleatoria $X \in E$, dove E insieme finito di $\sharp E$ elementi, è tale per cui le alternative $\{X = e\}$ hanno probabilità uniforme, quindi

$$P(X = e|I) = \frac{1}{\sharp E}.$$

Ad esempio, se $X \in \{1, 2, \dots, 6\}$ indica l'esito del lancio di un dado, poniamo

$$P(X = i|\text{“prima del lancio”}) = \frac{1}{6}.$$

8.2. **Funzione di ripartizione e di sopravvivenza.** Data una variabile aleatoria $X \in E$ con $E \subseteq \mathbb{R}$, a volte si è più interessati alla probabilità che X sia maggiore di un dato valore e. Ad esempio, se X indica la durata di una lampadina, il costruttore vuole che $P(X > t|I)$ sia più grande possibile, per $t \in \mathbb{R}$.

A tale scopo, si definisce la *funzione di sopravvivenza* di una variabile aleatoria $X \in E$, con $E \subseteq \mathbb{R}$ (rispetto all'informazione I) come la funzione (definita per ogni $t \in \mathbb{R}$)

$$t \in \mathbb{R} \quad \mapsto \quad P(X > t|I).$$

In modo analogo, se si è interessati alla probabilità che $X \in E$ sia minore (o uguale) di un certo valore, si può studiare la *funzione di ripartizione* di X, definita come

$$t \in \mathbb{R} \quad \mapsto \quad P(X \leq t|I).$$

Nella prossima sezione daremo esempi di funzioni di sopravvivenza e ripartizione per alcune leggi discrete. Qui notiamo solamente che, per la proprietà di additività, si ha sempre

$$P(X > t|I) + P(X \leq t|I) = 1,$$

quindi conoscendo la funzione di ripartizione, possiamo trovare subito quella di sopravvivenza e viceversa.

Osservazione 42 (funzione di ripartizione e densità discreta). Siccome

$$\{X \leq t\} = \bigcup_{\substack{e \in E \\ e \leq t}} \{X = e\}$$

e gli eventi a destra sono a due a due incompatibili, possiamo ottenere la funzione di ripartizione conoscendo la densità, precisamente come

$$P(X \leq t|I) = \sum_{\substack{e \in E \\ e \leq t}} P(X = e|I).$$

In modo simile, per trovare la funzione di sopravvivenza, basterà sommare sugli $e \in E$ con $e > t$.

Si può anche procedere all'opposto. Supponiamo di voler trovare $P(X = e|I)$ conoscendo solamente la funzione di ripartizione. Allora, siccome $\{X \leq e\} = \{X < e\} \cup \{X = e\}$ sono due eventi incompatibili, si ha

$$P(X = e|I) = P(X \leq e|I) - P(X < e|I).$$

Per calcolare $P(X < e|I)$ possiamo notare che, essendo l'insieme E discreto, possiamo sempre trovare il più grande $\bar{e} \in E$ tale che $\bar{e} < e$. Avremo allora $\{X \leq \bar{e}\} = \{X < e\}$ e quindi

$$[P(X = e|I) = P(X \leq e|I) - P(X \leq \bar{e}|I).$$

8.3. **Valore atteso.** Quando $E \subseteq \mathbb{R}$ è un insieme di numeri, questo può essere anche molto grande, anche infinito, e quindi lavorare agevolmente la legge di una variabile aleatoria $X \in E$ può diventare difficile, ad esempio, se non abbiamo formule semplici. È possibile però introdurre delle quantità numeriche ben precise (non aleatorie) che *descrivono* alcune proprietà tipicamente interessanti di X, come ad esempio: stimare grandi possono essere i suoi valori, quanto incerti siamo sulla stima dei suoi valori, ecc. Lo stesso si può dire di due variabili aleatorie numeriche, X e Y: mentre la variabile (X, Y) e la sua legge può essere molto complicata, possiamo essere interessati a descrivere proprietà più semplici, come sapere che se X assume un valore "grande" anche Y sarà "grande" ecc.

La prima quantità che introduciamo è il valore atteso di una variabile aleatoria (discreta) $X \in E$, che fornisce una prima indicazione circa la grandezza del valore di X, basandoci sulla informazione I di cui disponiamo.

Definizione 43. Sia $E \subseteq \mathbb{R}$ un insieme discreto. Data una variabile aleatoria $X \in E$ si definisce il suo valore atteso (sapendo l'informazione I) il numero reale

$$\mathbb{E}[X|I] = \sum_{e \in E} e \cdot P(X = e|I).$$

La lettera $\mathbb{E}$ viene dall'inglese come abbreviazione di *Expected Value* e non ha nulla a che fare con l'insieme dei possibili valori $E \subseteq \mathbb{R}$ della variabile aleatoria X.

Osservazione 44 (caso E infinito). Nel caso in cui E sia infinito, la somma a destra va intepretata come una serie, ossia un limite di somme finite. In tal caso si richiede in realtà che la serie a destra sia (assolutamente) convergente), ossia

$$\sum_{e \in E} |e| \cdot P(X = e|I) < \infty.$$

A parte pochi esempi, tuttavia, lavoreremo con variabili aleatorie discrete a valori in un insieme finito E, quindi questo problema matematico non si pone.

Come specificato nella definizione, la quantità $\mathbb{E}[X|I]$ è un numero ben specifico (ne calcoleremo molti in seguito), in particolare non è una variabile aleatoria (anche se nessuno ci vieta di pensare un numero fissato come una variabile aleatoria che assume un solo valore con probabilità 1). Osserviamo inoltre che il valore atteso di X (sapendo I) dipende unicamente dalla densità discreta di X (sapendo I).

Esempio 45 (valore atteso di una Bernoulli). Supponiamo che la variabile aleatoria $X \in \{0, 1\}$ abbia legge Bernoulli di parametro $p \in [0, 1]$. Allora

$$\mathbb{E}[X|I] = 0 \cdot P(X = 0|I) + 1 \cdot P(X = 1|I) = 0 \cdot (1 - p) + 1 \cdot p = p.$$

Il valore atteso di una variabile indicatrice di un evento A coincide con la probabilità di A. Notiamo inoltre che se $p \neq 0$ e $p \neq 1$ il valore atteso *non* è uno dei possibili valori!

Esempio 46 (valore atteso, lancio di un dado). Supponiamo che la variabile aleatoria $X \in \{1, 2, \ldots, 6\}$ indichi l'esito del lancio di un dado. Allora, rispetto all'informazione precedente al lancio abbiamo legge uniforme e quindi

$$\begin{aligned}\mathbb{E}[X|I] &= 1 \cdot P(X = 1|I) + 2 \cdot P(X = 2|I) + \ldots + 6 \cdot P(X = 6|I) \\ &= \frac{1}{6}(1 + 2 + 3 + 4 + 5 + 6) = \frac{21}{6} = 3,5.\end{aligned}$$

Osserviamo ancora una volta che il valore atteso non è tra i possibili valori, ma riconosciamo che è una stima del valore "tipico" che si può ottenere dal lancio.

Il valore atteso è una operazione matematica che gode di tante utili proprietà per il calcolo.

Teorema 47 (proprietà del valore atteso)**.** *Valgono le seguenti proprietà:*

i) (decomposizione) Se $X \in E \subseteq \mathbb{R}$ è una variabile aleatoria e A_1, $\ldots A_n$ sono un sistema di alternative, allora

$$\mathbb{E}[X|I] = \sum_{i=1}^{n} \mathbb{E}[X|A_i \cap I] P(A_i|I).$$

ii) (funzione composta) Se $X \in G$ è una variabile aleatoria discreta (non necessariamente $G \subseteq \mathbb{R}$) e $f : G \to E \subseteq \mathbb{R}$ è una funzione, allora

$$\mathbb{E}[f(X)|I] = \sum_{e \in E} f(e) \cdot P(X = e|I).$$

iii) (linearità) Se X, $Y \in E \subseteq \mathbb{R}$ sono variabili aleatorie e $c \in \mathbb{R}$ è un numero (non aleatorio), si ha

$$\mathbb{E}[X + Y|I] = \mathbb{E}[X|I] + \mathbb{E}[Y|I] \quad e \mathbb{E}[cX|I] = c\mathbb{E}[X|I].$$

In particolare, $\mathbb{E}[c|I] = \mathbb{E}[c1_\Omega|I] = c$.

iv) (monotonia) Se X, $Y \in E \subseteq \mathbb{R}$ sono variabili aleatorie e $P(X \leq Y|I) = 1$, allora

$$\mathbb{E}[X|I] \leq \mathbb{E}[Y|I].$$

In particolare, se $X \geq 0$ (con probabilità 1 sapendo ad I) allora $\mathbb{E}[X|I] \geq 0$ e

$$|\mathbb{E}[X|I]| \leq \mathbb{E}[|X||I].$$

v) (disuguaglianza di Markov) Se $X \in E \subseteq \mathbb{R}$ è una variabile aleatoria, per ogni numero reale $c > 0$ (non aleatorio) si ha

$$P(|X| \geq c) \leq \frac{\mathbb{E}[|X||I]}{c}. \tag{12}$$

Osservazione 48 (unità di misura del valore atteso). Può essere comodo attribuire alle variabili X delle unità di misura (ad esempio, metri, se X rappresenta lunghezze, secondi se rappresenta un tempo ecc.). Dalla definizione di valore atteso, segue che $\mathbb{E}[X|I]$ ha la stessa unità di misura (siccome la probabilità non ha unità di misura). Questo stratagemma può essere utile per ricordare che nella disuguaglianza di Markov (12) c deve avere la stessa unità di misura di X e quindi nel membro a destra deve apparire a denominatore, in modo che il numero risulti privo di unità di misura, come la probabilità a sinistra

Osservazione 49 (concentrazione attorno ad $\mathbb{E}[X|I]$). La disuguaglianza di Markov (12) permette di esprimere meglio l'intuizione che una variabile aleatoria $X \in E$ assume valori "vicino" al valore atteso $\mathbb{E}[X|I]$ con grande probabilità. Infatti, possiamo applicarla con la variabile aleatoria $X - \mathbb{E}[X|I]$ invece di X, ottenendo

$$P(|X - \mathbb{E}[X|I]| \geq c) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X|I]||I]}{c} \leq \frac{2\mathbb{E}[|X||I]}{c},$$

dove abbiamo usato anche la monotonia e la disuguaglianza

$$|X - \mathbb{E}[X|I]| \leq |X| + |\mathbb{E}[X|I]| \leq |X| + \mathbb{E}[|X||I].$$

Possiamo inoltre passare alla probabilità dell'evento complementare $\{|X - \mathbb{E}[X|I]| < c\}$, ottenendo

$$P(|X - \mathbb{E}[X|I]| < c) \geq 1 - \frac{2\mathbb{E}[|X||I]}{c}.$$

Da questa disuguaglianza otteniamo che, se c è grande, la probabilità che X assuma un valore più vicino di c ad al suo valore atteso diventa grande.

Precisamente, la "grandezza" di c si può "misurare" in multipli di $\mathbb{E}\left[|X|\,|I\right]$: se $c = m\mathbb{E}\left[|X|\,|I\right]$, troviamo

$$[P\left(|X - \mathbb{E}\left[X|I\right]| < c\right) \geq 1 - \frac{2}{m}.$$

Dimostrazione delle proprietà del valore atteso. Seguiamo l'ordine in cui sono presentate.

i) (decomposizione) Questa proprietà segue dalla decomposizione della probabilità rispetto ad un sistema di alternative e dalla possibilità di scambiare due sommatorie: per ogni $e \in E$, scriviamo

$$P(X = e|I) = \sum_{i=1}^{n} P(X = e|A_i \cap I)P(A_i|I).$$

Per la definizione di valore atteso,

$$\begin{aligned}
\mathbb{E}\left[X|I\right] &= \sum_{e\in E} e \cdot P(X = e|I) \\
&= \sum_{e\in E} e \cdot \sum_{i=1}^{n} P(X = e|A_i \cap I)P(A_i|I) \\
&= \sum_{e\in E} \sum_{i=1}^{n} e \cdot P(X = e|A_i \cap I)P(A_i|I) \\
&= \sum_{i=1}^{n} \left(\sum_{e\in E} e \cdot P(X = e|A_i \cap I) \right) P(A_i|I) \\
&= \sum_{i=1}^{n} \mathbb{E}\left[X|A_i \cap I\right] P(A_i|I)
\end{aligned}$$

ii) (funzione composta) Usiamo la decomposizione rispetto al sistema di alternative $A_e = \{X = e\}$, per $e \in E$. Si trova

$$\begin{aligned}
\mathbb{E}\left[f(X)|I\right] &= \sum_{e\in E} \mathbb{E}\left[f(X)|\left\{X = e\right\} \cap I\right] P(X = e|I) \\
&= \sum_{e\in E} \mathbb{E}\left[f(e)|\left\{X = e\right\} \cap I\right] P(X = e|I) \quad \text{perché } f(X) = f(e), \text{ sapendo } \{X = e\} \\
&= \sum_{e\in E} f(e)P(X = e|I)
\end{aligned}$$

iii) (linearità) Notiamo intanto che $\mathbb{E}\left[c|I\right] = c$, quando $c \in \mathbb{R}$ è un numero (interpetato come una variabile aleatoria che assume un solo valore). Segue da questo la proprietà

$$\mathbb{E}\left[X + c|I\right] = \mathbb{E}\left[X|I\right] + c,$$

perché decomponiamo rispetto al sistema di alternative $\{X = e\}$,

$$\begin{aligned}\mathbb{E}[X + c|I] &= \sum_{e\in E} \mathbb{E}[X + c|\{X = e\} \cap I]\, P(X = e|I)\\ &= \sum_{e\in E} \mathbb{E}[e + c|\{X = e\} \cap I]\, P(X = e|I)\\ &= \sum_{e\in E} (e + c)P(X = e|I) = \sum_{e\in E} eP(X = e|I) + c\sum_{e\in E} P(X = e|I)\\ &= \mathbb{E}[X|I] + c,\end{aligned}$$

perché $\sum_{e\in E} P(X = e|I) = 1$. Date X e $Y \in E$ variabili aleatorie, decomponiamo rispetto al sistema $\{Y = f\}$, per ottenere

$$\begin{aligned}\mathbb{E}[X + Y|I] &= \sum_{f\in E} \mathbb{E}[X + f|\{Y = f\} \cap I]\, P(Y = f|I)\\ &= \sum_{f\in E} \mathbb{E}[X|\{Y = e\} \cap I]\, P(Y = f|I) + \sum_{f\in E} fP(Y = f|I)\\ &= \mathbb{E}[X|I] + \mathbb{E}[Y|I]\end{aligned}$$

La proprietà

$$\mathbb{E}[cX|I] = c\mathbb{E}[X|I]$$

si dimostra in modo simile, decomponendo rispetto al sistema di alternative $\{X = e\}$ e poi ricomponendo.

iv) (monotonia) Basta mostrare che, se $Z \geq 0$, con probabilità 1 (rispetto ad I), allora

$$\mathbb{E}[Z|I] \geq 0.$$

Infatti, applicando questo a $Z = Y - X$ si trova $\mathbb{E}[Y - X|I] = \mathbb{E}[Y|I] - \mathbb{E}[X|I] \geq 0$ e quindi la tesi. D'altra parte, se $P(Z \geq 0|I) = 1$, nella somma che definisce il valore atteso di Z, tutti possibili termini $e \in E$ con $e < 0$ sono moltiplicati per una probabilità nulla, $P(Z = e|I) \leq P(Z < 0|I) = 0$, quindi

$$\mathbb{E}[Z|I] = \sum_{\substack{e\in E\\ e\geq 0}} e \cdot P(Z = e|I) \geq 0$$

perché è una somma di termini non-negativi.

v) Markov Per dimostrare la proprietà di Markov, introduciamo il sistema di alternative $\{|X| \geq c\}$, $\{|X| > c\}$. Troviamo

$$\begin{aligned}\mathbb{E}[|X||I] &= \mathbb{E}[|X||\{|X| \geq c\} \cap I]\, P(|X| \geq c|I) + \mathbb{E}[|X||\{|X| < c\} \cap I]\, P(|X| < c|I)\\ &\geq \mathbb{E}[|X||\{|X| \geq c\} \cap I]\, P(|X| \geq c|I)\\ &\qquad \text{perché } \mathbb{E}[|X||\{|X| < c\} \cap I] \geq 0, \text{ essendo } |X| \geq 0, \text{ e } P(|X| < c|I) \geq 0,\\ &\geq \mathbb{E}[c|\{|X| \geq c\} \cap I]\, P(|X| \geq c|I)\\ &\qquad \text{per monotonia del valore atteso e usando l'informazione } \{|X| \geq c\},\\ &= cP(|X| \geq c|I).\end{aligned}$$

Leggendo dalla fine all'inizio, abbiamo quindi ottenuto che

$$cP(|X| \geq c|I) \leq \mathbb{E}[|X||I],$$

da cui la disuguaglianza segue dividendo per $c > 0$ ambo i membri. □

8.4. **Varianza.** La seconda quantità che introduciamo è la varianza di una variabile aleatoria $X \in E \subseteq \mathbb{R}$. Questo numero (sempre non-negativo) indica in un modo spesso facile da calcolare quanto il valore di X si discosta dal suo valore atteso $\mathbb{E}[X|I]$.

Definizione 50 (varianza)**.** Data una variabile aleatoria $X \in E \subseteq \mathbb{R}$, si definisce la sua varianza $\mathrm{Var}(X|I)$ (sapendo I) il numero reale non-negativo

$$\mathrm{Var}(X|I) = \mathbb{E}\left[(X - \mathbb{E}[X|I])^2|I\right].$$

Affinché la varianza sia definita, deve essere definito il valore atteso $\mathbb{E}[X|I]$, quindi nel caso di E infinito ci potrebbero essere dei problemi di convergenza di serie (ma noi non ce ne occupiamo).

Osservazione 51 (deviazione standard)*.* Notiamo che la varianza misura il *quadrato* dello scostamento di X rispetto al suo valore atteso, quindi ad esempio se X si misura in metri, la varianza si misura in metri quadri ecc. Per questo, se si vuole confrontare X con la sua varianza, bisogna passare alla *deviazione standard*, definita come

$$\sigma(X|I) = \sqrt{\mathrm{Var}(X|I)}.$$

Teorema 52 (Proprietà della varianza)**.** *Valgono le seguenti proprietà, per $X \in E \subseteq \mathbb{R}$ variabile aleatoria discreta:*

i) (espressione alternativa) vale l'identità

$$\mathrm{Var}(X|I) = \mathbb{E}\left[X^2|I\right] - (\mathbb{E}[X|I])^2.$$

ii) (quadraticità) se $\lambda, c \in \mathbb{R}$ sono costanti (rispetto ad I) si ha

$$\mathrm{Var}(\lambda X + c|I) = \lambda^2 \mathrm{Var}(X|I).$$

iii) (varianza nulla) $\mathrm{Var}(X|I) = 0$ *se e solo se* $P(X = \mathbb{E}[X|I]\,|I) = 1$*,*

iv) (disuguaglianza di Chebychev) per ogni costante $c > 0$ si ha

$$P(|X - \mathbb{E}[X|I]\,| \geq c|I) \leq \frac{\mathrm{Var}(X|I)}{c^2}.$$

Osservazione 53 (concentrazione attorno ad $\mathbb{E}[X|I]$)*.* Partendo dalla disuguaglianza di Chebychev, possiamo ragionare come nell'osservazione dopo la disuguaglianza di Markov, e ponendo $c = \varepsilon\sigma$ (dove σ è la deviazione standard di X sapendo I), otteniamo la disuguaglianza

$$P(|X - \mathbb{E}[X|I]\,| < \varepsilon\sigma|I) \geq 1 - \frac{1}{\varepsilon^2}. \tag{13}$$

Stavolta otteniamo che la probabilità che X sia vicino al valore atteso è grande se misurata in termini di multipli della deviazione standard, $m\sigma$. In modo molto più informale, si scrive spesso che X è bene approssimabile come il valore atteso più o meno la deviazione standard, e si scrive

$$X \approx \mathbb{E}[X|I] \pm \sqrt{\mathrm{Var}(X|I)} = \mathbb{E}[X|I] \pm \sigma(X|I).$$

Dimostrazione. *i) (espressione alternativa)* Sviluppiamo il quadrato

$$\begin{aligned}\operatorname{Var}(X|I) &= \mathbb{E}\left[(X-\mathbb{E}[X|I])^2|I\right]\\ &= \mathbb{E}\left[X^2-2\mathbb{E}[X|I]\,X+(\mathbb{E}[X|I])^2|I\right]\\ &= \mathbb{E}\left[X^2|I\right]-\mathbb{E}\left[2\mathbb{E}[X|I]\,X|I\right]+\mathbb{E}\left[(\mathbb{E}[X|I])^2|I\right] \quad \text{per linearità del valore atteso}\\ &= \mathbb{E}\left[X^2|I\right]-2\mathbb{E}[X|I]\,\mathbb{E}[X|I]+\mathbb{E}[X|I])^2 \quad \text{perché } \mathbb{E}[X|I] \text{ è una costante}\\ &= \mathbb{E}\left[X^2|I\right]-(\mathbb{E}[X|I])^2\end{aligned}$$

ii) (quadraticità) Ricordiamo che

$$\mathbb{E}[\lambda X+c|I] = \lambda\mathbb{E}[X|I]+c.$$

Si trova

$$\begin{aligned}\operatorname{Var}(\lambda X+c||I) &= \mathbb{E}\left[(\lambda X+c-\mathbb{E}[\lambda X+c|I])^2|I\right]\\ &= \mathbb{E}\left[(\lambda(X-\mathbb{E}[X|I]))^2|I\right]\\ &= \mathbb{E}\left[\lambda^2(X-\mathbb{E}[X|I])|I\right] = \lambda^2\operatorname{Var}(X|I)\end{aligned}$$

iii) (varianza nulla) È chiaro che, se $X=\mathbb{E}[X|I]$ allora nella definizione di varianza stiamo prendendo il valore atteso di una variabile aleatoria che è con probabilità 1 uguale a 0, e quindi la varianza è nulla. Viceversa, usando la regola del valore atteso di una variabile aleatoria composta $f(X)=(X-\mathbb{E}[X|I])^2$, troviamo

$$0=\operatorname{Var}(X|I)=\sum_{e\in E}(e-\mathbb{E}[X|I])^2P(X=e|I).$$

Siccome tutti gli addendi sono non-negativi, deve essere necessariamente

$$(e-\mathbb{E}[X|I])^2P(X=e|I)=0 \quad \text{per ogni possibile valore } e\in E,$$

quindi $P(X=e|I)>0$ solamente nel caso in cui $e=\mathbb{E}[X|I]$, e quindi $P(X=\mathbb{E}[X|I]\,|I)=1$. *iv) Chebychev* Possiamo ragionare in due modi: ripetendo la dimostrazione della disuguaglianza di Markov, stavolta con le alternative $\{|X-\mathbb{E}[X|I]\,|\geq c\}$ e $\{|X-\mathbb{E}[X|I]\,|<c\}$, oppure applicando direttamente la disuguaglianza di Markov con la variabile $(X-\mathbb{E}[X|I])^2$ al posto di X e c^2 al posto di c e notando che i due eventi

$$\{|X-\mathbb{E}[X|I]\,|>c\}, \quad \left\{(X-\mathbb{E}[X|I])^2>c^2\right\}$$

coincidono. □

Osservazione 54 (errori frequenti sulla varianza). Un errore molto frequente negli esercizi è di decomporre la varianza secondo un sistema di alternative $A_1,\dots A_n$, come il valore atteso:

$$\operatorname{Var}(X|I)=\sum_{i=1}^{n}\operatorname{Var}(X|A_i\cap I)\,P(A_i|I) \quad \leftarrow \text{NON è sepre VERA!}$$

Pensateci un momento: se questa formula fosse vera, si potrebbe decomporre con il sistema di alternative $\{X=e\}$, ma allora $\operatorname{Var}(X|\{X=e\}\cap I)=\operatorname{Var}(e|\{e\}\cap I)=0$ e quindi la varianza sarebbe sempre nulla! Invece, consigliamo di usare sempre l'espressione alternativa $\operatorname{Var}(X|I)=\mathbb{E}\left[X^2|I\right]-(\mathbb{E}[X|I])^2$ e calcolare separatamente i due valori attesi $\mathbb{E}[X|I]$ e $\mathbb{E}\left[X^2|I\right]$,

per i quali si certamente può decomporre rispetto ad un sistema di alternative.

Un altro errore è di calcolare la varianza di una somma $X + Y$ come la somma delle varianza

$$\operatorname{Var}(X + Y|I) = \operatorname{Var}(X|I) + \operatorname{Var}(Y|I) \leftarrow \text{NON è sempre VERA!}$$

Vedremo che questa proprietà è vera quando X e Y sono *indipendenti* (o più in generale non-correlate). Ma per rendersi conto che questa formula non può essere sempre vera, basta porre $X = Y$, così si troverebbe

$$2\operatorname{Var}(X|I) = \operatorname{Var}(X + X|I) = \operatorname{Var}(2X|I) = 4\operatorname{Var}(X|I).$$

e di nuovo la varianza dovrebbe essere sempre nulla!

8.5. **Covarianza.** Concludiamo con una terza quantità, stavolta riguardante due variabili aleatorie X, $Y \in E$.

Definizione 55 (covarianza)**.** Se X, $Y \in E \subseteq \mathbb{R}$ sono variabili aleatorie, si definisce la covarianza tra X ed Y (sapendo I) il numero reale

$$\operatorname{Cov}(X, Y|I) = \mathbb{E}\left[(X - \mathbb{E}[X|I])(Y - \mathbb{E}[Y|I])|I\right].$$

La definizione somiglia a quella di varianza, e in effetti se $X = Y$ si trova $\operatorname{Cov}(X, X|I) = \operatorname{Var}(X|I)$. Notiamo anche che $\operatorname{Cov}(X, Y|I) = \operatorname{Cov}(Y, X|I)$.

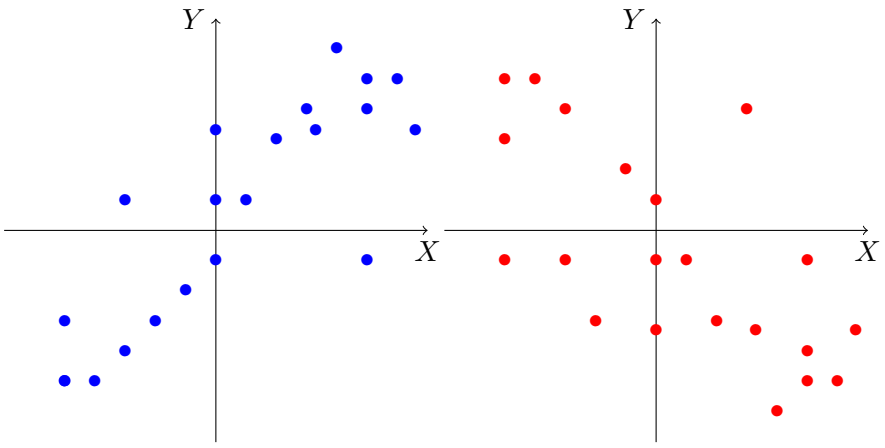
In generale, che cosa indica la covarianza? Più che la grandezza assoluta, un primo importante indicatore è il *segno* di $\operatorname{Cov}(X, Y|I)$. Infatti, se $\operatorname{Cov}(X, Y|I) > 0$ allora le due variabili si dicono positivamente correlate, se $\operatorname{Cov}(X, Y|I) < 0$ si dicono negativamente correlate e infine se $\operatorname{Cov}(X, Y|I) = 0$ si dicono non-correlate (o scorrelate). Cosa indica qualitativamente il fatto che X e Y siano positivamente correlate? Osserviamo che, affinché $\operatorname{Cov}(X, Y|I) > 0$ la variabile aleatoria $(X - \mathbb{E}[X|I])(Y - \mathbb{E}[Y|I])$ dovrà essere probabilmente più "positiva" che "negativa" (questa è una approssimazione, non è proprio così sempre). Ma il prodotto è positivo quando $X - \mathbb{E}[X|I]$ e $Y - \mathbb{E}[Y|I]$ sono entrambi positivi o entrambi negativi. Quindi possiamo dire (sempre con approssimazione) che X e Y sono positivamente correlate se, sapendo ad esempio che $X > \mathbb{E}[X|I]$, allora probabilmente anche $Y > \mathbb{E}[Y|I]$, e similmente sapendo che $X < \mathbb{E}[X|I]$, allora probabilmente anche $Y < \mathbb{E}[Y|I]$ (e viceversa, scambiando i ruoli di X e Y). In senso opposto invece se le variabili sono negativamente correlate. Infine, se sono non correlate, sapere se $X > \mathbb{E}[X|I]$ o $X < \mathbb{E}[X|I]$ dovrebbe lasciarci indifferenti circa il valore di Y (Figure 8.1). In effetti, questi ragionamenti sono approssimativi ma, in molte occasioni, abbastanza utili.

Esercizio 56 (Proprietà della covarianza). Se $X, Y \in E \subseteq \mathbb{R}$ sono variabili aleatorie discrete, allora

$$\operatorname{Cov}(X, Y|I) = \mathbb{E}[XY|I] - \mathbb{E}[X|I]\, EY|I,$$

e

$$\operatorname{Var}(X + Y|I) = \operatorname{Var}(X|I) + \operatorname{Var}(Y|I) + 2\operatorname{Cov}(X, Y|I).$$



(A) $\mathrm{Cov}\,(X, Y|I) > 0$ (B) $\mathrm{Cov}\,(X, Y|I) < 0$

FIGURA 8.1. Esempi di diverse correlazioni. I punti rappresentano i possibili valori (e, f) che le variabili (X, Y) possono assumere, e supponiamo probabilità uniforme sulle alternative $\{(X, Y) = (e, f)\}$. Gli assi cartesiani si intersecano nel punto $(\mathbb{E}\,[X|I]\,, \mathbb{E}\,[Y|I])$.

9. Esempi di leggi discrete

In questa sezione presentiamo alcune delle leggi di variabili aleatorie discrete che capita più spesso di incontrare nei problemi. Molto spesso, queste leggi sono in realtà famiglie di leggi, al variare di alcuni parametri naturali. Per quanto possibile, discuteremo di ciascuna di esse il significato intuitivo dei parametri e ne calcoleremo valore atteso e varianza (in alcuni casi, la funzione di ripartizione e di sopravvivenza).

9.1. **Legge Bernoulli.** Abbiamo già introdotto la legge Bernoulli, descritta da un parametro $p \in [0, 1]$, come la densità discreta di un variabile aleatoria a valori in $\{0, 1\}$, per cui

$$P(X = 1|X \text{ Bernoulli}(p)) = p \quad \text{e} \quad P(X = 0|X \text{ Bernoulli}(p)) = 1 - p.$$

Abbiamo già calcolato il valore atteso

$$\mathbb{E}\,[X|X \text{ Bernoulli}(p)] = 0 \cdot (1 - p) + 1 \cdot p = p.$$

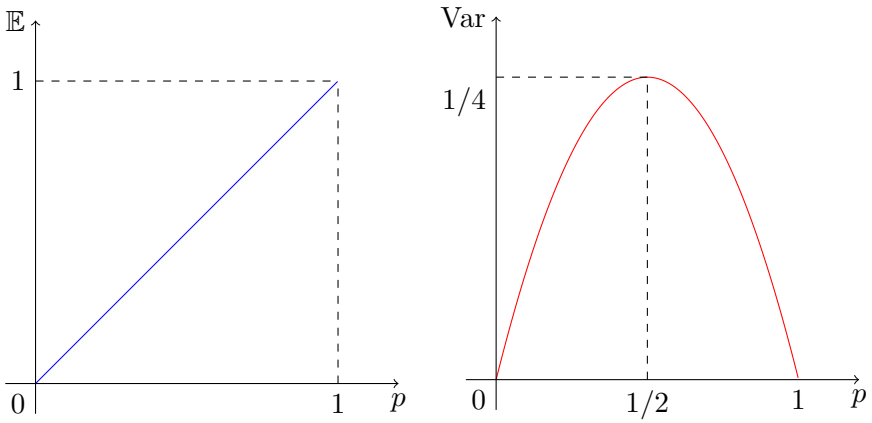
Per calcolare la varianza, usiamo la regola del valore atteso di una funzione composta, e troviamo prima

$$\mathbb{E}\left[X^2|X \text{ Bernoulli}(p)\right] = 0^2 \cdot (1 - p) + 1^2 \cdot p = p$$

Di conseguenza,

$$\mathrm{Var}\,(X|X \text{ Bernoulli}(p)) = p - p^2 = p(1 - p).$$

Nella figura 9.2 rappresentiamo, al variare del parametro $p \in [0, 1]$, il valore atteso e la varianza di una variabile X con legge Bernoulli di parametro p. Notiamo che nei valori estremi $p = 0$, $p = 1$ la varianza è nulla (perché X è costante), mentre la varianza è massima per $p = 1/2$, in accordo con l'intuizione che l'incertezza è massima se $p = 1/2$.



(A) $\mathbb{E}[X|\text{Bernoulli}(p)] = p$ (B) $\text{Var}(X|\text{Bernoulli}(p)) = p(1-p)$

FIGURA 9.1. Grafici del valore atteso e della varianza di una legge Bernoulli, al variare del parametro $p \in [0,1]$.

Esercizio 57. Descrivere la funzione di ripartizione e di sopravvivenza di una variabile avente legge Bernoulli di parametro p (in particolare, disegnarne i grafici).

9.2. **Legge uniforme (su un intervallo $\{1,\dots,n\}$).** Fissato un numero naturale $n \geq 1$, possiamo considerare una variabile aleatoria X a valori in $\{1,\dots,n\}$ avente *legge uniforme* sugli n elementi di tale intervallo, ossia

$$P(X = i | X \text{ unif. su } \{1,\dots,n\}) = \frac{1}{n},$$

per ogni elemento $i \in \{1,\dots,n\}$. Per calcolare valore atteso e varianza di X, usiamo i seguente risultato riguardanti la somma dei primi n numeri naturali positivi e dei quadrati dei primi n numeri naturali positivi

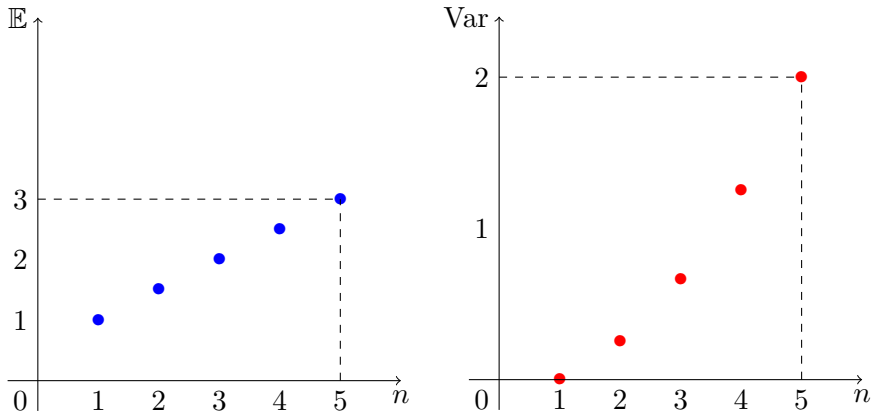
$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}, \quad \text{e} \quad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}.$$

Usando questi risultati, troviamo

$$\begin{aligned}\mathbb{E}[X|X \text{ unif. su } \{1,\dots,n\}] &= \sum_{i=1}^{n} i \cdot \frac{1}{n} \\ &= \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2},\end{aligned}$$

e, come passo intermedio per calcolare la varianza di X,

$$\begin{aligned}\mathbb{E}\left[X^2|X \text{ unif. su } \{1,\dots,n\}\right] &= \sum_{i=1}^{n} i^2 \cdot \frac{1}{n} \\ &= \frac{1}{n} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{(n+1)(2n+1)}{6}.\end{aligned}$$



(A) $\mathbb{E}[X|\{1,\dots,n\}] = \frac{n+1}{2}$ (B) $\mathrm{Var}(X|\{1,\dots,n\}) = \frac{n^2-1}{12}$.

FIGURA 9.2. Grafici del valore atteso e della varianza di una legge uniforme sull'insieme $\{1, 2\dots, n\}$, al variare del parametro n.

Usando l'espressione alternativa per la varianza, troviamo

$$\begin{aligned}\mathrm{Var}\left(X^2|X \text{ unif. su } \{1,\dots,n\}\right) &= \frac{(n+1)(2n+1)}{6} - \left(\frac{n+1}{2}\right)^2 \\ &= \frac{(n+1)}{12}\left(2(2n+1) - 3(n+1)\right) = \frac{(n+1)}{12}\cdot(n-1) \\ &= \frac{n^2-1}{12}.\end{aligned}$$

Vediamo quindi che, più grande è n maggiore è il valore atteso $(n+1)/2$, che si colloca geometricamente nel punto medio del segmento $[1, n]$, e pure è la varianza, che è un polinomio di secondo grado in n (in accordo con il fatto che deve essere quadratica).

Esercizio 58. Descrivere la funzione di ripartizione e di sopravvivenza di una variabile avente legge uniforme sull'insieme $\{1,\dots,n\}$ (in particolare, disegnarne i grafici).

Un semplice argomento di "traslazione" ci permette di calcolare valore atteso e varianza di variabili uniformi su un qualunque intervallo discreto.

Esempio 59. Supponiamo che una variabile aleatoria X abbia legge uniforme sull'insieme $\{10, 11, 12, \dots, 20\}$ (rispetto ad una informazione I). Allora la variabile $Y = X - 9$ ha legge uniforme sull'insieme $\{1, 2, \dots, 11\}$ e pertanto calcoliamo

$$\mathbb{E}[Y|I] = \frac{12}{2} = 6, \quad \mathrm{Var}(Y|I) = \frac{11^2-1}{12} = 10.$$

D'altra parte, abbiamo $X = Y + 9$, quindi

$$\mathbb{E}[X|I] = \mathbb{E}[Y+9|I] = 6{+}9 = 15, \quad \mathrm{Var}(X|I) = \mathrm{Var}(Y+9|I) = \mathrm{Var}(Y|I) = 10.$$

9.3. **Legge binomiale.** Riprendiamo la legge binomiale trovata nello studio delle estrazioni da un'urna con reimmissione. In quel caso abbiamo visto che, posta $p = R/N$ la probabilità di un successo in una singola estrazione, la probabilità di avere esattamente k successi in n tentativi (in un ordine qualunque) è data dalla formula

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

Diciamo quindi che una variabile aleatoria X, a valori in $\{0, 1, \dots, n\}$ ha legge binomiale di parametri (n, p) (e abbreviamo con $B(n, p)$) se vale

$$P(X = k | X \text{ è } B(n,p)) = \binom{n}{k} p^k (1-p)^{n-k}$$

per ogni $k \in \{0, 1, \dots, n\}$. Le variabili binomiali sono piuttosto frequenti: precisamente ogni volta che vogliamo contare il numero X di successi in una successione di n tentativi *indipendenti*, in cui ciascuno ha la stessa probabilità di successo $p \in [0, 1]$.

Notiamo che $B(1, p)$ coincide con la legge Bernoulli (un solo tentativo).

Nelle figure 9.3 e 9.4 raffiguriamo, al variare di n e p, il grafico della densità discreta di un variabile con legge binomiale $B(n, p)$. Siccome i parametri sono due, è comodo discutere cosa accade se teniamo fisso uno e modifichiamo l'altro. Ad esempio, tenendo fisso n, notiamo che per p vicino a 0 la densità si "concentra" verso il valore 0 (in accordo con il fatto che è più difficile avere successo), mentre per p vicino ad 1 si concentra verso il valore massimo, n. Ci possiamo aspettare che il valore atteso quindi seguirà lo stesso andamento, mentre la varianza tenderà a zero al tendere di p agli estremi. Se invece teniamo fisso p, ad esempio per $p = 1/2$, e facciamo crescere n, la densità si distribuisce su intervalli sempre più grandi, e tende a diventare piccola (quasi uniforme). Quindi ci aspettiamo che il valore atteso diventerà grande, e pure la varianza, un po' come accade con la legge uniforme.

Per calcolare in modo veloce valore atteso e varianza di X, ritorniamo al modello delle estrazioni dall'urna, e introduciamo delle variabili aleatorie ausiliarie $X_1, \dots, X_n$ indicatrici dell'evento "successo" al tentativo i, per $i \in \{1, \dots, n\}$, ossia

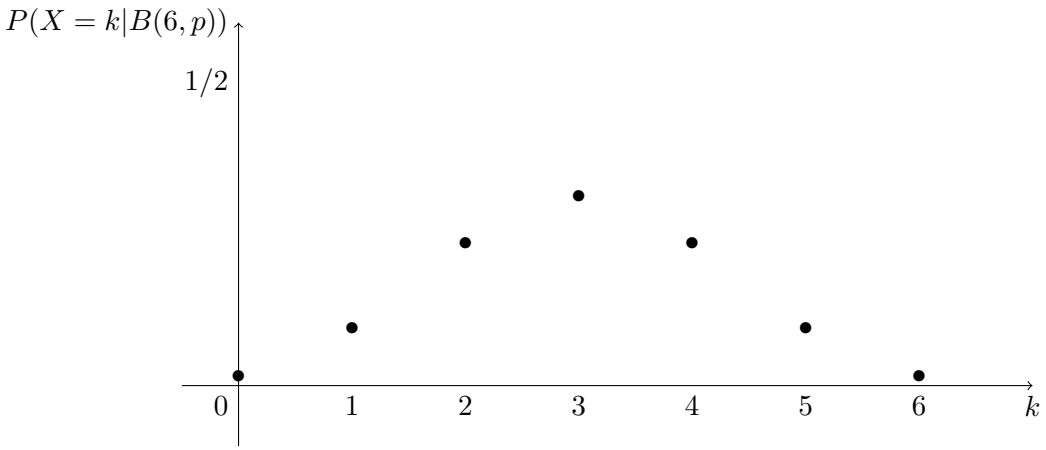
$$\{X_i = 1\} = R_i.$$

A questo punto, il numero totale dei successi in n tentativi è dato semplicemente dalla *somma* delle variabili $X_1, \dots X_n$, quindi possiamo rappresentare

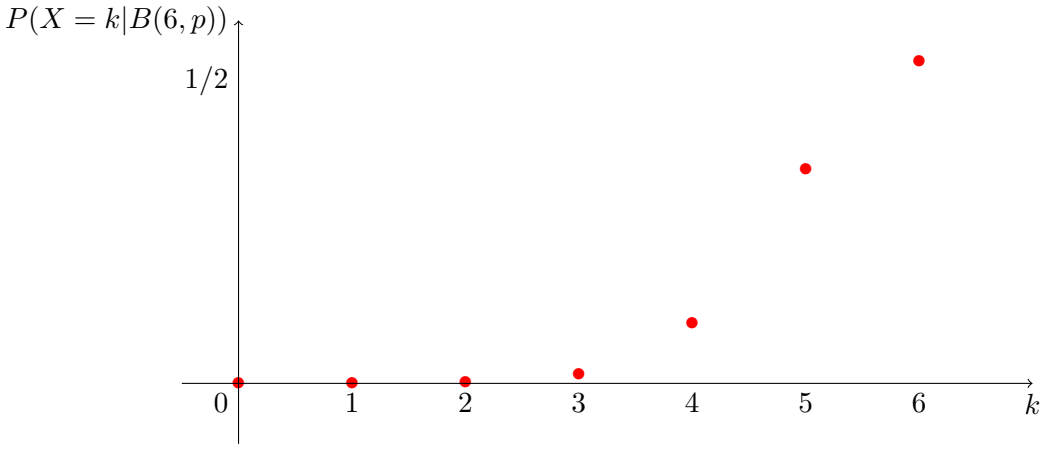
$$X = \sum_{i=1}^{n} X_i.$$

Per calcolare il valore atteso, usiamo il fatto che ciascuna X_i è Bernoulli di parametro $p = P(R_i|I)$, quindi per linearità del valore atteso abbiamo

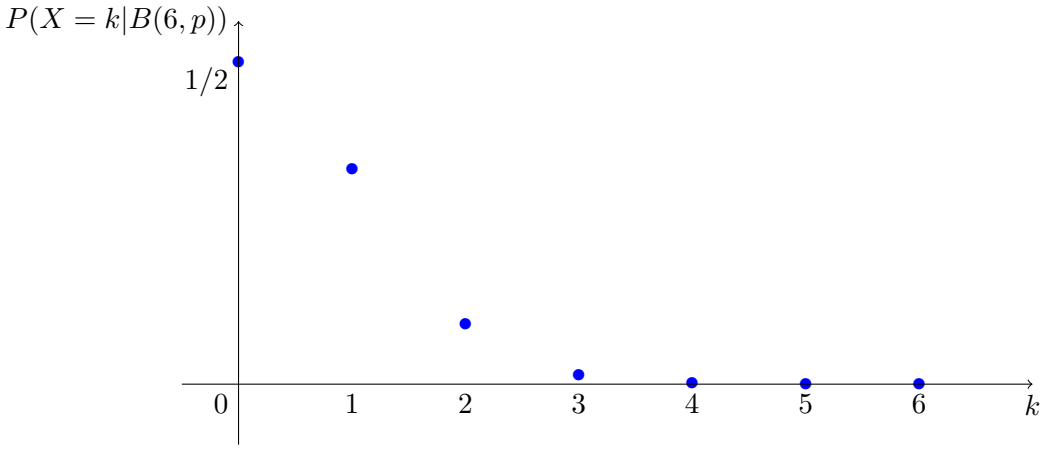
$$\mathbb{E}[X|I] = \sum_{i=1}^{n} \mathbb{E}[X_i|I] = \sum_{i=1}^{n} p = n \cdot p.$$



(A) Densità discreta $B(6, 1/2)$ (nero)



(B) Densità discreta $B(6, 9/10)$ (rosso)

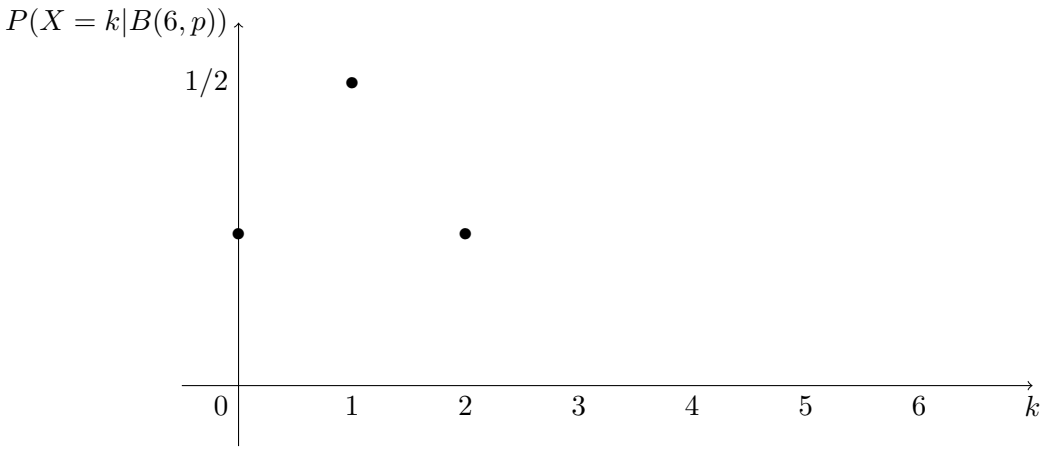


(C) Densità discreta $B(6, 1/10)$ (blu)

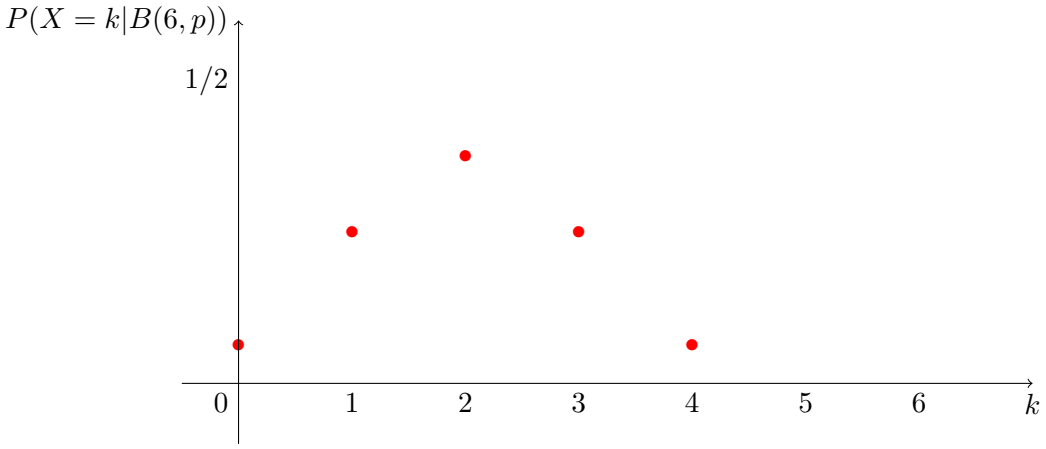
FIGURA 9.3. Grafici della densità binomiale al variare di p.

Per calcolare la varianza, calcoliamo al solito prima

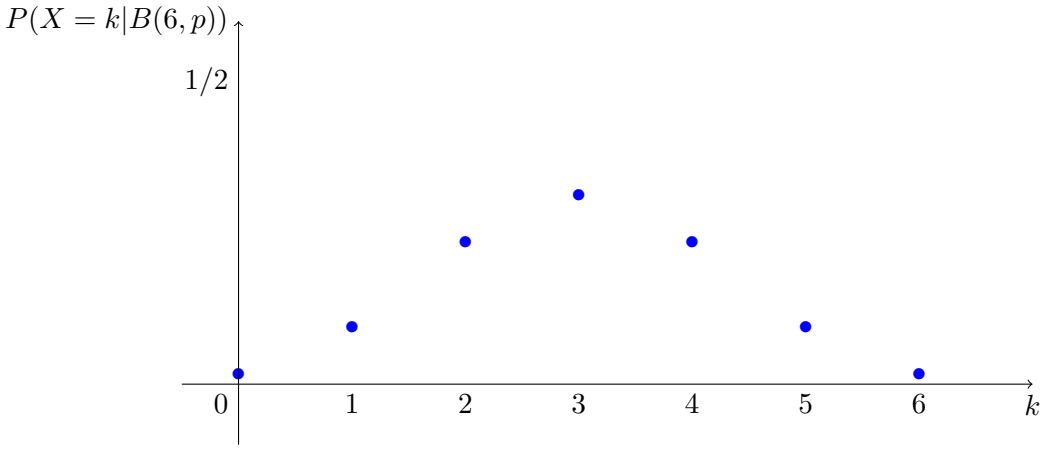
$$\mathbb{E}\left[X^2|I\right] = \mathbb{E}\left[(\sum_{i=1}^{n} X_i)^2|I\right]$$
$$= \mathbb{E}\left[\sum_{i,j=1}^{n} X_i X_j|I\right] = \sum_{i,j=1}^{n} \mathbb{E}\left[X_i X_j|I\right]$$



(A) Densità discreta $B(2, 1/2)$ (nero)



(B) Densità discreta $B(4, 1/2)$ (rosso)



(C) Densità discreta $B(6, 1/2)$ (blu)

FIGURA 9.4. Grafici della densità binomiale al variare di n.

Per calcolare $\mathbb{E}[X_i X_j | I]$ ricordiamo che essendo X_i, X_j indicatrici, il loro prodotto è la variabile indicatrice dell'intersezione e quindi, se $i \neq j$, usando

l'ipotesi di indipendenza tra gli eventi R_i, R_j, perché le estrazioni sono con reimmissione,

$$\mathbb{E}\left[X_iX_j|I\right] = P(R_i \cap R_j|I) = P(R_i|I)P(R_j|I) = p^2.$$

D'altra parte, se $i = j$, abbiamo semplicemente $X_i^2 = X_i$, quindi

$$\mathbb{E}\left[X_i^2|I\right] = p.$$

Dobbiamo ora contare quanti casi del primo e del secondo tipo si presentano nella somma $\sum_{i,j=1}^n \mathbb{E}\left[X_iX_j|I\right]$. Siccome i casi totali sono n^2 e quelli del tipo $i = j$ sono chiaramente n, troviamo

$$\mathbb{E}\left[X^2|I\right] = \sum_{i,j=1}^{n} \mathbb{E}\left[X_iX_j|I\right] = np + (n^2 - n)p^2.$$

Abbiamo quindi trovato che

$$\begin{aligned}\mathrm{Var}\left(X|X \text{ è } B(n,p)\right) &= np + (n^2 - n)p^2 - (np)^2 = np - np^2 \\ &= np(1-p).\end{aligned}$$

Notiamo un fatto interessante: dato che la varianza di ciascuna X_i, essendo Bernoulli, è $p(1-p)$, la varianza di X coincide con la somma delle varianze delle singole prove. Vedremo che questo fatto è una conseguenza dell'indipendenza delle varie prove.

9.4. **Legge Poisson.** A volte ci si trova in una situazione in cui si effettuano tante prove *indipendenti* ma la probabilità di successo è molto bassa. Nell'esempio delle estrazioni dall'urna con reimmissione, supponiamo che la probabilità di successo $p \in [0,1]$ (estrarre una rossa) sia molto piccola, però facciamo n un numero molto grande di estrazioni. Dopo le estrazioni, il numero "tipico" di palline rosse che avremo visto sarà

$$\mathbb{E}\left[X|X \text{ è } B(n,p)\right] = n \cdot p.$$

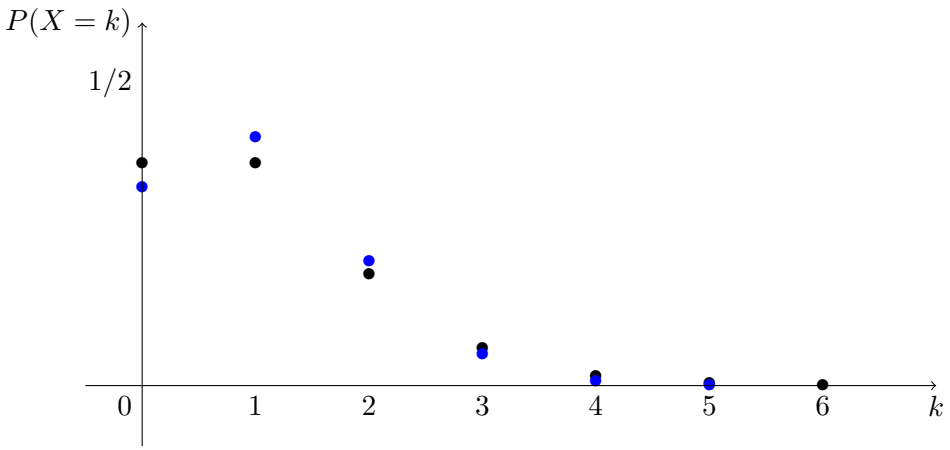
Ad esempio, se vogliamo aspettarci di vedere una pallina (in media), dovremo fare $n = 1/p$ estrazioni, quindi un numero molto grande.

In queste situazioni, diventa problematico lavorare direttamente con la densità binomiale di parametri n, p e un argomento matematico permette di semplificare notevolmente i calcoli. Più precisamente, se $p \in [0,1]$ è molto piccolo, n è molto grande in modo tale che il prodotto $n \cdot p$ sia vicino ad un numero reale $\lambda > 0$, allora con poco errore si può approssimare una variabile aleatoria X, avente densità binomiale $B(n,p)$, con una avente legge Poisson[5] di parametro λ, di cui ora diamo la definizione.

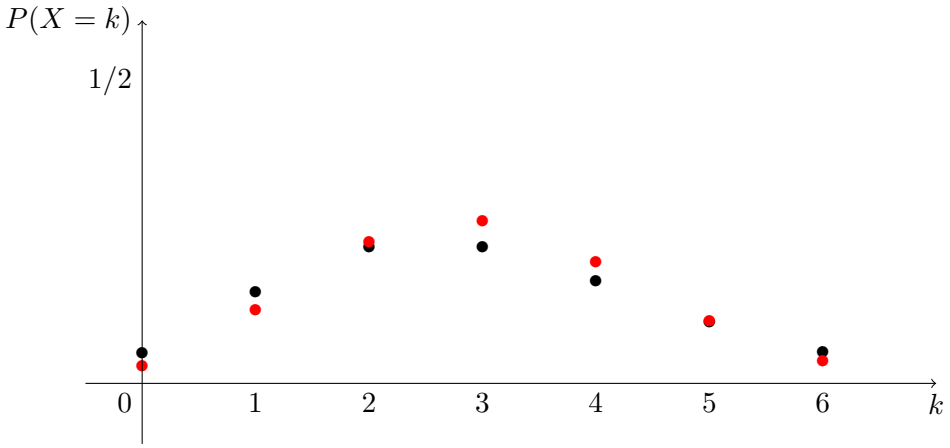
Diciamo che una variabile aleatoria $X \in \mathbb{N}$ ha legge (sapendo I) Poisson di parametro $\lambda > 0$ se, per ogni $k \in \mathbb{N}$, vale

$$P(X = k|I) = e^{-\lambda}\frac{\lambda^k}{k!}.$$

[5]in onore del matematico francese Siméon Denis Poisson



(A) Densità discreta Poisson(1) (nero) e $B(5, 1/5)$ (blu)



(B) Densità discreta Poisson(3) (nero) e $B(10, 3/10)$ (rosso)

FIGURA 9.5. Grafici della densità Poisson al variare di λ, al confronto con $B(n, p)$ con $np = \lambda$.

dove $k! = k(k-1)\cdot \ldots \cdot 1$, $(0! = 1)$. Il numero $e^{-\lambda}$ è solamente una costante necessaria affinché la somma delle probabilità su tutti i valori possibili sia 1,

$$\sum_{k=0}^{+\infty} P(X = k|I) = \sum_{k=0}^{+\infty} e^{-\lambda}\frac{\lambda^k}{k!} = 1,$$

una condizione che deve sempre essere vera quando si ha una densità discreta.

Nella figura 9.6 confrontiamo il grafico di densità Poisson di parametro λ con quelle di una Binomiale $B(n, p)$ con p piccolo ed n grande in modo che $np = \lambda$. Vediamo che la differenza non è molto rilevante. Questa approssimazione ci permette di "indovinare" subito valore atteso e varianza di una Poisson di parametro λ: dovendo essere

$$\mathbb{E}\left[X|\text{Poisson}(\lambda)\right] \approx \mathbb{E}\left[X|B(n,p),\ p \text{ piccolo},\ n \text{ grande } np = \lambda\right] = np$$

otteniamo che

$$\mathbb{E}\left[X|\text{Poisson}(\lambda)\right] = \lambda.$$

Per la varianza, pure

$$\mathrm{Var}\,(X|\mathrm{Poisson}(\lambda)) \approx \mathrm{Var}\,(X|B(n,p),\ p \text{ piccolo, } n \text{ grande } np=\lambda) = (np)(1-p)$$

otteniamo (approssimando $1-p$ con 1)

$$\mathrm{Var}\,(X|\mathrm{Poisson}(\lambda)) = \lambda.$$

Queste "ipotesi" sono confermate con un calcolo rigoroso, tramite la definizione di legge Poisson, che però eviteremo, perché richiederebbe di lavorare con serie numeriche.

9.5. **Legge geometrica.** Concludiamo questa esposizione di leggi discrete studiando ancora una volta un problema che proviene dal modello delle estrazioni con reimmissione.

Dati k, n, con $1 \leq k \leq n$, supponiamo di effettuare $n \geq 1$ estrazioni con rimpiazzo da un urna, in cui la probabilità di ottenere successo (pallina rossa) è $p \in [0,1]$. Qual è la probabilità che la *prima volta* che otteniamo un successo sia all'estrazione k?

Equivalentemente, cerchiamo la probabilità di ottenere la sequenza ordinata di lunghezza k in cui le prime $k-1$ palline sono blu e l'ultima è rossa, quindi

$$P(B1B2B3\ldots B(k-1)Rk|I) = (1-p)^{k-1}p.$$

Notiamo anche che la probabilità di non ottenere alcun successo in n tentativi è $(1-p)^n$. Per raccogliere questi risultati, vogliamo introdurre una variabile aleatoria $X \in \{1,\ldots,n\}$ in modo che

$$\{X=k\} = \text{"primo successo avviene all'estrazione } k\text{"}.$$

Tuttavia, dobbiamo anche tenere conto dell'alternativa in cui non ci sono successi, quindi si potrebbe introdurre un nuvoo simbolo $\{\infty\}$ all'insieme dei valori di X e porre

$$\{X=\infty\} = \text{"nessun successo nelle } n \text{ estrazioni"}.$$

Si trova quindi

$$P(X=k|I) = (1-p)^{k-1}p \quad \text{per ogni } k \in \{1,\ldots,n\},$$

e

$$P(X=\infty|I) = (1-p)^n.$$

Anche in questo caso, come per la binomiale, avremmo a che fare con due possiblii parametri: il numero totale di estrazioni $n \geq 1$ e la probabilità del singolo successo $p \in [0,1]$. Tuttavia, se il numero di estrazioni è molto grande, possiamo considerare direttamente il caso limite in cui n è infinito. In questo caso il parametro è solamente $p \in [0,1]$ e si trova che $X \in \{1,2,3,\ldots\}$ ha legge *geometrica*:

$$P(X=k|\mathrm{Geom}(p)) = (1-p)^{k-1}p \quad \text{per ogni } k \in \{1,2,3\ldots\}.$$

Osservazione 60 (Paradosso di Borel). 'Notiamo che la probabilità di non estrarre mai una pallina rossa in n estrazioni, pari a $(1-p)^n$, tende a 0 nel limite di n infinitamente grande. Possiamo interpretare questo fatto nel seguente modo: *per quanto piccola sia la probabilità p di un evento, effettuando un numero sufficientemente grande di tentativi indipendenti questo*

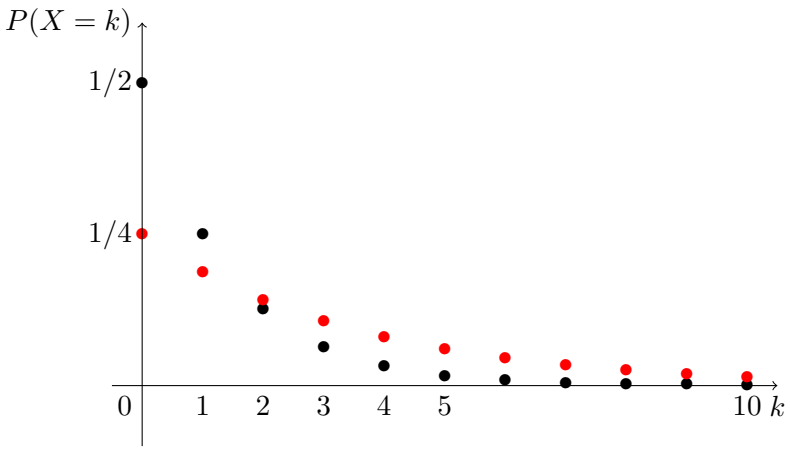


FIGURA 9.6. Grafici di densità Geometrica, $p = 1/2$ (nero), $p = 1/4$ (rosso).

prima o poi dovrebbe avvenire. Il condizionale "dovrebbe" è obbligatorio, almeno per due ragioni:

(1) se la probabilità p è molto piccola, il numero "tipico" di tentativi (vedremo che è $1/p$) può essere talmente grande da superare ogni limite imposto dalla realtà fisica ;
(2) le previsioni della probabilità riguardano il grado di fiducia basandoci sull'informazione che si possiede, che potenzialmente è soggetta a cambiamento o rivelarsi errata.

Alla luce di queste ragioni, possiamo riformulare il risultato anche nel seguente modo, che forse suona meno paradossale: *se dopo tantissimi $(1/p)$ esperimenti (che riteniamo indipendenti tra loro) l'evento non si realizza, dovremmo ricrederci sulla sua probabilità oppure sul fatto che le prove fossero indipendenti.*

Per calcolare il valore atteso e la varianza di una variabile aleatoria X con legge geometrica, usare direttamente le definizioni richiederebbe di sommare delle serie, ad esempio

$$\mathbb{E}\left[X|\mathrm{Geom}(p)\right] = \sum_{k=1}^{\infty} k(1-p)^{k-1}p.$$

Anche se si tratta di un semplice esercizio sulle serie, preferiamo evitare. Possiamo invece ragionare nel seguente modo, usando il sistema di alternative $R_1 =$ "successo alla prima estrazione", $B_1 =$ "fallimento alla prima estrazione". Se I è l'informazione iniziale (che al solito descrive l'urna) ed X indica il primo successo in una successione (potenzialmente infinita) di estrazioni, troviamo

$$\begin{aligned}\mathbb{E}\left[X|I\right] &= \mathbb{E}\left[X|R_1 \cap I\right] P(R_1|I) + \mathbb{E}\left[X|B_1 \cap I\right] P(B_1|I)\\ &= \mathbb{E}\left[1|R_1 \cap I\right] p + \mathbb{E}\left[X|B_1 \cap I\right](1-p)\\ &= p + \mathbb{E}\left[X|B_1 \cap I\right](1-p)\end{aligned}$$

D'altra parte, se sappiamo che la prima estrazione è stata un fallimento, dalla seconda estrazione è come se ricominciassimo da capo (per via

dell'indipendenza), però aggiungendo 1 al conto dei tentativi:

$$\mathbb{E}[X|B_1 \cap I] = \mathbb{E}[(1+X)|I] = 1 + \mathbb{E}[X|I].$$

Se trattiamo il valore atteso $v = \mathbb{E}[X|I]$ come una incognita, abbiamo trovato una equazione per v,

$$v = p + (1+v)(1-p) \quad \text{la cui soluzione è} v = \frac{1}{p}$$

Concludiamo quindi che

$$\mathbb{E}[X|\text{Geom}(p)] = \frac{1}{p}.$$

Osserviamo che questo risultato è convincente, perché se p è molto piccola il tempo di attesa sarà grande, mentre se p tende ad 1 avremo che il valore atteso è vicino ad 1.

Per la varianza, al solito ragioniamo prima calcolando il valore atteso del quadrato, decomponendo allo stesso modo:

$$\begin{aligned}
\mathbb{E}\left[X^2|I\right] &= \mathbb{E}\left[X^2|R_1 \cap I\right] p + \mathbb{E}\left[X^2|B_1 \cap I\right](1-p) \\
&= p + \mathbb{E}\left[(1+X)^2|I\right](1-p) \\
&= p + \mathbb{E}\left[1 + 2X + X^2|I\right](1-p) \\
&= p + (1-p) + 2\mathbb{E}[X|I](1-p) + \mathbb{E}\left[X^2|I\right](1-p) \\
&= 1 + 2(1-p)/p + \mathbb{E}\left[X^2|I\right](1-p)
\end{aligned}$$

e troviamo un'equazione per $\mathbb{E}\left[X^2|I\right]$, da cui

$$\mathbb{E}\left[X^2|I\right] = (1 + 2(1-p)/p)/p = \frac{2-p}{p^2}.$$

Infine

$$\text{Var}(X|\text{Geom}(p)) = \frac{2-p}{p^2} - \left(\frac{1}{p}\right)^2 = \frac{1-p}{p^2}.$$

Anche in questo caso, notiamo che per $p \to 0$ la varianza tende a infinito, con la deviazione standard $\sigma = \sqrt{(1-p)/p^2} \approx 1/p$ che è confrontabile con il valore atteso. All'opposto, per $p \to 1$, la varianza tende a zero, la variabile tende ad essere costante (e uguale ad 1).

Esercizio 61 (funzione di sopravvivenza). Mostrare che la funzione di sopravvivenza di una variabile aleatoria X con legge geometrica soddisfa, per ogni $k \in \{1, 2, 3, \ldots\}$,

$$P(X > k|\text{Geom}(p)) = (1-p)^k.$$

(Suggerimento: come si scrive $\{X > k\}$ in termini di successi/fallimenti?)

Esercizio 62 (assenza di memoria). Supponiamo che $X \in \{1, 2, \ldots\}$ sia una variabile aleatoria con legge geometrica di un dato parametro $p \in [0, 1]$, rispetto ad una certa informazione I. Mostrare che per ogni numero naturale $n \geq 1$, la variabile $X - n$ è geometrica (dello stesso parametro p), rispetto all'informazione $I \cap \{X > n\}$. (Suggerimento: scrivere l'evento $\{X > k\}$ in termini di successi/fallimenti ed usare l'indipendenza tra le varie estrazioni, oppure usare la formula di Bayes).

10. Indipendenza tra variabili aleatorie

Abbiamo introdotto le variabili aleatorie come un linguaggio molto comodo per trattare i sistemi di alternative $\{X = e\}$ associati ad una quantità X a valori in un insieme E. In questa sezione estendiamo il concetto di *indipendenza* tra eventi a variabili aleatorie. Ricordiamo che l'idea di fondo dell'indipendenza tra A e B (sapendo I) è che non riusciamo ad utilizzare l'informazione dell'uno (A) per cambiare il grado di fiducia sull'altro (B),

$$P(A|B \cap I) = P(A|I).$$

Consideriamo ora due variabili aleatorie $X \in E$, $Y \in F$. Per definire che esse sono indipendenti (sapendo I), possiamo dire che *qualunque informazione* otteniamo da una di esse (ad esempio X), questa non cambia il grado di fiducia, ossia la *legge* che attribuiamo all'altra (Y), ossia la probabilità degli eventi $\{Y \in F'\}$ per ogni $F' \subseteq F$. È bene però specificare che cosa intendiamo per *informazione ottenibile* da una variabile aleatoria. Ad esempio, possiamo supporre di sapere che $\{X = e\}$, ma più in generale, possiamo anche supporre di sapere solamente che $\{X \in E'\}$ per un qualche $E' \subseteq E$. Stiamo quindi richiedendo che gli eventi $\{X \in E'\}$ $\{Y \in F'\}$ siano indipendenti.

Formuliamo quindi la seguente definizione precisa di due variabili aleatorie indipendenti:

Definizione 63 (variabili aleatorie indipendenti)**.** Siano $X \in E$, $Y \in F$ variabili aleatorie (discrete). Diciamo che X ed Y sono indipendenti (sapendo I) se, per ogni $E' \subseteq E$, $F' \subseteq F$, si ha che gli eventi $\{X \in E'\}$, $\{Y \in F'\}$ sono indipendenti (sapendo I), ossia

$$P(Y \in F' | \{X \in E'\} \cap I) = P(Y \in F'|I). \tag{14}$$

Nel caso $F' = \{f\}$, $E' = \{e\}$, l'identità sopra diventa

$$P(Y = f | \{X = e\} \cap I) = P(Y = f|I). \tag{15}$$

Osservazione 64 (equivalenza tra (15) e (14))*.* In realtà, per verificare (o definire) l'indipendenza di due variabili aleatorie discrete, basta controllare che valga l'identità (15) (questo semplifica notevolmente i calcoli, perché i possibili sottoinsiemi di E, F sono molti di più dei singoli elementi). Verifichiamolo: supponendo che valga (15), siano $F' \subseteq F$ ed $E' \subseteq E$. Allora, decomponendo con il sistema di alternative $\{X = e\}$,

$$\begin{aligned} P(Y \in F' | \{X \in E'\} \cap I) &= \sum_{e \in E} P(Y \in F' | \{X = e\} \cap \{X \in E'\} \cap I) P(X = e | \{X \in E'\}) \\ &= \sum_{e \in E'} P(Y \in F' | \{X = e\} \cap I) P(X = e | \{X \in E'\} \cap I) \end{aligned}$$

Perché le alternative in cui $\{X = e\}$ ma $e \in (E')^c$ hanno probabilità nulla, rispetto $\{X \in E'\} \cap I$. Decomponendo anche $\{Y \in F'\} = \bigcup_{f \in F'} \{Y = f\}$,

abbiamo

$$\begin{aligned} P(Y \in F' | \left\{X \in E'\right\} \cap I) &= \sum_{e \in E'} \sum_{f \in F'} P(Y = f | \left\{X = e\right\} \cap I) P(X = e | \left\{X \in E'\right\} \cap I) \\ &= \sum_{e \in E'} \sum_{f \in F'} P(Y = f | I) P(X = e | \left\{X \in E'\right\} \cap I) \\ &= \sum_{e \in E'} P(Y \in F' | I) P(X = e | \left\{X \in E'\right\} \cap I) \\ &= P(Y \in F' | I) P(X \in E' | \left\{X \in E'\right\} \cap I) = P(Y \in F' | I). \end{aligned}$$

Come nel caso dell'indipendenza tra eventi, la condizione di indipendenza tra X ed Y (sapendo I) è completamente simmetrica nei ruoli di X ed Y.

Osservazione 65 (indipendenza e composizione con funzioni). Un'osservazione importante è che, se $X \in E$ e $Y \in F$ sono variabili indipendenti (sapendo I) e introduciamo due funzioni $u : E \to U$, $v : F \to V$, anche le variabili aleatorie $u(X)$, $v(Y)$ sono indipendenti (sapendo I). Infatti, l'informazione ottenibile da $u(X)$ è una *parte* dell'informazione ottenibile da X. In formule,

$$\left\{u(X) \in U'\right\} = \left\{X \in u^{-1}(U')\right\}.$$

Osservazione 66 (indipendenza tra più variabili aleatorie). Nel caso di più di due variabili aleatorie $X_1 \in E_1$, $X_2 \in E_2$, $\dots X_n \in E_n$, per definirne l'indipendenza (sapendo I) basterà ricondursi all'indipendenza tra più di due eventi. Diremo quindi che, per ogni possibile scelta di sottoinsiemi $E_1' \subseteq E_1$, $E_2' \subseteq E_2$, $\dots E_n' \subseteq E_n$, gli eventi

$$\left\{X_1 \in E_1'\right\}, \quad \left\{X_2 \in E_2'\right\}, \dots, \left\{X_n \in E_n'\right\}$$

sono indipendenti. Ricordando la definizione, in pratica questo significa che, ogni volta che stiamo calcolando una probabilità che riguarda alcune variabili X_i con $i \in F$ (per un qualche insieme $F \subseteq \{1, 2, \dots, n\}$) e, oltre ad I, abbiamo dell'informazione che riguarda solo le rimanenti variabili X_j con $j \in F^c$, possiamo sempre trascurarla e ritornare all'informazione I.

Concludiamo questa sezione con una osservazione molto utile nei problemi.

Esercizio 67. Siano $X = 1_A$, $Y = 1_B$ variabili aleatorie indicatrici degli eventi A, B. Mostrare che X e Y sono indipendenti (sapendo I) se e solo se gli eventi A e B sono indipendenti (sapendo I).

In particolare, due eventi A, B sono indipendenti se e solo se A^c e B sono indipendenti: basta infatti scrivere le variabili indicatrici 1_A, 1_B e notare che $1_{A^c} = 1 - 1_A = u(1_A)$ è funzione della variabile indicatrice di A (dove $u(x) = 1 - x$). Negli esercizi a volte può essere quindi per dimostrare che due eventi sono (o non sono) indipendenti, porsi la stessa domanda per l'evento complementare (negazione) di uno oppure entrambi.

10.1. **Varianza della somma.** Se $X \in E$, $Y \in F$ sono variabili aleatorie discrete e indipendenti, allora, qualunque informazione otteniamo da X,

non soltanto la legge di Y non cambia ma, se $F \subseteq \mathbb{R}$ il valore atteso (e la varianza) di Y non cambiano. Infatti, dato $E' \subseteq E$, abbiamo

$$\begin{aligned}\mathbb{E}\left[Y|\left\{X \in E'\right\} \cap I\right] &= \sum_{f \in F} fP(Y = f|\left\{X \in E'\right\} \cap I) \\ &= \sum_{f \in F} fP(Y = f|I) \\ &= \mathbb{E}\left[Y|I\right].\end{aligned}$$

Esercizio 68. Mostrare che anche

$$\mathrm{Var}\left(Y|\left\{X \in E'\right\} \cap I\right) = \mathrm{Var}\left(Y|I\right).$$

Una importante risultato che segue da questo fatto è la formula

$$\mathbb{E}\left[XY|I\right] = \mathbb{E}\left[X|I\right]\mathbb{E}\left[Y|I\right], \tag{16}$$

se X, $Y \in E \subseteq \mathbb{R}$ sono variabili aleatorie (discrete) indipendenti. Per mostrarla, basta decomporre con il sistema di alternative $\{X = e\}$,

$$\begin{aligned}\mathbb{E}\left[XY|I\right] &= \sum_{e \in E} \mathbb{E}\left[XY|\left\{X = e\right\} \cap I\right] P(X = e|I) \\ &= \sum_{e \in E} \mathbb{E}\left[eY|\left\{X = e\right\} \cap I\right] P(X = e|I) \\ &= \sum_{e \in E} e\mathbb{E}\left[Y|\left\{X = e\right\} \cap I\right] P(X = e|I) \quad \text{perché } e \text{ è una costante} \\ &= \sum_{e \in E} e\mathbb{E}\left[Y|I\right] P(X = e|I) \quad \text{per quanto abbiamo visto sopra} \\ &= \mathbb{E}\left[Y|I\right] \sum_{e \in E} eP(X = e|I) = \mathbb{E}\left[Y|I\right]\mathbb{E}\left[X|I\right].\end{aligned}$$

Dalla (16), segue un altro fatto molto rilevante (anche negli esercizi!).

Proposizione 69. *Se X, $Y \in E \subseteq \mathbb{R}$ sono variabili aleatorie (discrete) indipendenti (sapendo I), allora*

$$\mathrm{Var}\left(X + Y|I\right) = \mathrm{Var}\left(X|I\right) + \mathrm{Var}\left(Y|I\right).$$

Dimostrazione. Intanto osserviamo che

$$\left(\mathbb{E}\left[X + Y|I\right]\right)^2 = \left(\mathbb{E}\left[X|I\right]\right)^2 + \left(\mathbb{E}\left[Y|I\right]\right)^2 + 2\mathbb{E}\left[X|I\right]\mathbb{E}\left[Y|I\right].$$

Poi, calcoliamo

$$\begin{aligned}\mathbb{E}\left[(X + Y)^2|I\right] &= \mathbb{E}\left[X^2 + Y^2 + 2XY|I\right] \\ &= \mathbb{E}\left[X^2|I\right] + \mathbb{E}\left[Y^2|I\right] + 2\mathbb{E}\left[XY|I\right] \\ &== \mathbb{E}\left[X^2|I\right] + \mathbb{E}\left[Y^2|I\right] + 2\mathbb{E}\left[X|I\right]\mathbb{E}\left[Y|I\right]\end{aligned}$$

avendo usato la formula (16). Usando l'espressione alternativa per la varianza, si conclude che

$$\begin{aligned}\mathrm{Var}\left(X + Y|I\right) &= \mathbb{E}\left[(X + Y)^2|I\right] - \left(\mathbb{E}\left[X + Y|I\right]\right)^2 \\ &= \mathbb{E}\left[X^2|I\right] + \mathbb{E}\left[Y^2|I\right] + 2\mathbb{E}\left[XY|I\right] - \left(\mathbb{E}\left[X|I\right]\right)^2 - \left(\mathbb{E}\left[Y|I\right]\right)^2 - 2\mathbb{E}\left[X|I\right]\mathbb{E}\left[Y|I\right] \\ &= \mathbb{E}\left[X^2|I\right] - \left(\mathbb{E}\left[X|I\right]\right)^2 + \mathbb{E}\left[Y^2|I\right] - \left(\mathbb{E}\left[Y|I\right]\right)^2 \\ &= \mathrm{Var}\left(X|I\right) + \mathrm{Var}\left(Y|I\right).\end{aligned}$$

□

Esercizio 70. Date $X, Y \in E \subseteq \mathbb{R}$ variabili aleatorie indipendenti, mostrare che $\mathrm{Cov}(X, Y|I) = 0$.

Più in generale, con una dimostrazione simile, si mostra che, se rispetto all'informazione I le variabili aleatorie $X_1, X_2, \dots X_n \in E \subseteq \mathbb{R}$ sono (a due a due) indipendenti, allora *la varianza della somma è uguale alla la somma delle varianze*,

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i|I\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i|I\right). \tag{17}$$

10.2. **Legge dei grandi numeri.** L'identità (17) è notevole perché, essendo la varianza *quadratica*, ci si aspetterebbe che sommando n termini si ottenga come risultato una somma che coinvolga n^2 addendi (si pensi al caso in cui $X_1 = X_2 = \ldots = X_n$), quindi è molto più "piccolo". Se pensiamo alle differenze $X_i - \mathbb{E}[X_i|I]$ come a degli errori che possiamo commettere sostituendo al vero valore di X_i il suo valore atteso $\mathbb{E}[X_i|I]$, il fatto che le variabili siano indipendenti implica che molti di questi, con alta probabilità, si "cancellano" tra loro.

Una conseguenza di questa proprietà di "cancellazione" è il seguente risultato.

Teorema 71 (legge (debole) dei grandi numeri)**.** *Siano $X_1, \dots X_n \in E \subseteq \mathbb{R}$ variabili aleatorie discrete indipendenti (rispetto ad I), tutte con lo stesso valore atteso*

$$m = \mathbb{E}[X_i|I] \quad \textit{per ogni } i \in \{1, \dots, n\},$$

e deviazione standard

$$\sigma = \sqrt{\mathrm{Var}\left(X_i|I\right)} \quad \textit{per ogni } i \in \{1, \dots, n\}.$$

Allora, per ogni $\varepsilon > 0$, si ha

$$P\left(\left|m - \frac{1}{n}\sum_{i=1}^{n} X_i\right| < \varepsilon\sigma\middle|I\right) \geq 1 - \frac{1}{n\varepsilon^2}.$$

In termini meno matematici, possiamo pensare alle variabili aleatorie $X_1, \dots X_n$ come a delle "copie" (ma indipendenti) di una certa variabile aleatoria X che vogliamo studiare – pensiamo ad esempio ad un laboratorio di fisica, in cui si ripetono tante volte le misure di un fenomeno che si vuole studiare. Allora, la *media empirica* dei risultati che otterremo,

$$\frac{1}{n}\sum_{i=1}^{n} X_i$$

per n molto grande grande dovrebbe avvicinarsi al valore atteso $m = \mathbb{E}[X|I]$, con grande probabilità. Precisamente, se vogliamo che la differenza sia minore di ε volte la deviazione standard σ, $(\varepsilon\sigma)$ dovremo effettuare un numero di tentativi "proporzionale" ad $1/\varepsilon^2$, ad esempio con $n = 100/\varepsilon^2$, si trova che la probabilità è maggiore del 99%.

Come nel caso di tanti altri risultati della teoria della probabilità, il condizionale "dovrebbe" è obbligatorio: infatti forniamo delle previsioni circa

la realtà basate sul nostro grado di fiducia e della informazione parziale, che quindi semplicemente andrà aggiornata, se vediamo che la previsione della legge dei grandi numeri non si realizza.

Osservazione 72 (probabilità come frequenza). Tuttavia, in tante situazioni (pensiamo ad esempio ai lanci di monete ripetute, alle statistiche sul sesso delle nascite) la legge dei grandi numeri si verifica in modo così evidente che alcuni scienziati in passato hanno proposto una definizione di probabilità di un evento proprio come *frequenza relativa* con cui questo si realizza, in un numero idealmente infinito di esperimenti indipendenti. In un certo senso, in questo modo, la legge dei grandi numeri assume il valore di una "legge" del mondo reale. Al di là del problema pratico di realizzare questi infiniti esperimenti, c'è in questa definizione un problema strutturale, perché non tiene conto che la probabilità può cambiare a seconda dell'informazione di cui uno dispone: se la probabilità è verificabile con "esperimenti" in modo così inequivocabile, non potrebbe dipendere dall'osservatore.

Dimostrazione della legge dei grandi numeri. Consideriamo la variabile aleatoria

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Il suo valore atteso (rispetto ad I) è dato da

$$\begin{aligned}\mathbb{E}[Y|I] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i|I\right] = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i|I\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i|I] = \frac{1}{n}\sum_{i=1}^{n} m \\ &= m.\end{aligned}$$

Per la varianza, troviamo

$$\begin{aligned}\operatorname{Var}(Y|I) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i|I\right) = \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i|I\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i|I) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Per concludere, applichiamo la disuguaglianza di Chebychev nella forma (13) con Y invece di X. □

Esempio 73. Si effettuano n estrazioni con rimpiazzo da un'urna contenente R palline rosse su N totali, e poniamo $p = R/N$. Usare la legge dei grandi numeri per stimare la probabilità che il la frequenza delle palline rosse estratte

$$X = \frac{\sharp\{\text{palline rosse estratte}\}}{n}$$

sia compreso in un intervallo

$$X \in (p - \varepsilon\sqrt{p(1-p)}, p + \varepsilon\sqrt{p(1-p)})$$

per un dato $\varepsilon > 0$.

Scrivendo $X = \frac{1}{n}\sum_{i=1}^{n} X_i$, dove $X_i \in \{0,1\}$ è la variabile indicatrice del successo (estrazione rossa) all'estrazione i, abbiamo che le X_i sono indipendenti (perché le estrazioni sono con reimmissione) e Bernoulli di parametro $p = R/N$. Quindi

$$\mathbb{E}[X_i|I] = p \quad \mathrm{Var}(X_i|I) = \sigma^2 = p(1-p)$$

e la legge dei grandi numeri ci dice

$$P(X \in (p - \varepsilon\sqrt{p(1-p)}, p + \varepsilon\sqrt{p(1-p)}|I) = P(|X-p| < \varepsilon\sigma|I) \geq 1 - \frac{1}{\varepsilon^2 n}.$$

10.3. **Operazioni tra variabili aleatorie indipendenti.** Alcune leggi discrete si comportano bene rispetto a delle operazioni naturali tra variabili aleatorie, ad esempio la somma.

Proposizione 74 (somma di binomiali indipendenti)**.** *Sia $p \in [0,1]$, n, $m \geq 1$ e siano (rispetto ad una informazione I)*

(1) $X \in \{0,1,\ldots,n\}$ una variabile aleatoria con legge $B(n,p)$,
(2) $Y \in \{0,1,\ldots,m\}$ una variabile aleatoria con legge $B(m,p)$,
(3) X e Y indipendenti.

Allora la somma $X + Y \in \{0,1,\ldots,m+n\}$ è una variabile aleatoria con legge $B(n+m,p)$.

Osserviamo che il parametro $p \in [0,1]$ deve essere lo stesso per entrambe le variabili X e Y.

Dimostrazione. Per dimostrare questo fatto, ricordiamo che le variabili binomiali $B(n,p)$ indicano il numero di successi in n esperimenti indipendenti (ad esempio, estrazioni con rimpiazzo) dove p è la probabilità del singolo successo. Allora, se immaginiamo di fare $n+m$ estrazioni, possiamo rappresentare X come il numero di successi nelle prime n estrazioni e Y come il numero di successi nelle estrazioni che vanno dalla $n+1$ alla $n+m$ (siccome le estrazioni sono diverse, X e Y sono indipendenti). D'altra parte, il numero totale di successi è $X+Y$, che quindi ha legge $B(n+m,p)$. □

Un risultato simile vale per le leggi Poisson. Questo non ci dovrebbe stupire, perché possiamo pensare di approssimare una variabile X avente legge Poisson di parametro λ con una legge $B(n,p)$, dove $n = \lambda/p$

Proposizione 75 (somma di Poisson indipendenti)**.** *Siano λ_1, $\lambda_2 > 0$ e (rispetto ad una informazione I)*

(1) $X \in \{0,1,\ldots,\}$ una variabile aleatoria con legge Poisson(λ_1),
(2) $Y \in \{0,1,\ldots\}$ una variabile aleatoria con legge Poisson(λ_2),
(3) X e Y indipendenti.

Allora la somma $X + Y \in \{0,1,\ldots\}$ è una variabile aleatoria con legge Poisson($\lambda_1 + \lambda_2$).

Più in generale, ci possiamo chiedere: date variabili aleatorie $X \in E$, $Y \in F$ con $E, F \subseteq \mathbb{R}$, come calcolare la legge di $X+Y$? Una formula, detta

di convoluzione, si ottiene decomponendo rispetto al sistema di alternative $\{X = k\}$, per cui dato $z \in \mathbb{R}$, abbiamo

$$\begin{aligned} P(X+Y=z|I) &= \sum_{k\in E} P(X+Y=z|\,\{X=k\}\cap I)P(X=k|I) \\ &= \sum_{k\in E} P(e+Y=z|\,\{X=k\}\cap I)P(X=k|I) \\ &= \sum_{k\in E} P(Y=z-e|\,\{X=k\}\cap I)P(X=k|I), \\ &= \sum_{\substack{k\in E\\(z-k)\in F}} P(Y=z-k|\,\{X=k\}\cap I)P(X=k|I), \end{aligned}$$

dove nell'ultimo passaggio abbiamo notato che basta sommare solo quando $z-k$ è un possibile valore di Y (altrimenti la probabilità è nulla).

Se aggiungiamo l'ipotesi che Y sia indipendente da X, sapendo I, possiamo scrivere

$$P(Y=z-k|\,\{X=k\}\cap I) = P(Y=z-k|I)$$

e quindi

$$P(X+Y=z|I) = \sum_{\substack{k\in E\\(z-k)\in F}} P(Y=z-k|I)P(X=k|I). \tag{18}$$

Esempio 76. Usando (18), mostriamo che la somma di variabili Poisson indipendenti è Poisson. Dato $z \in \{0,1,\dots\}$, dobbiamo sommare su tutti i $k \in \{0,1,\dots\}$ tali che $z-k \in \{0,1,2,\dots\}$. Notiamo che se $k > z$, allora $z-k < 0$ e quindi dovremo sommare solamente per $k \in \{0,1,2,\dots,z\}$. Otteniamo

$$\begin{aligned} P(X+Y=z|I) &= \sum_{k=0}^{z} e^{-\lambda_2}e^{-\lambda_1}\frac{\lambda_2^{z-k}\lambda_1^k}{(z-k)!k!} \\ &= e^{-\lambda_2}e^{-\lambda_1}\frac{1}{z!}\sum_{k=0}^{z}\frac{z!}{(z-k)!k!}\lambda_2^{z-k}\lambda_1^k \\ &= e^{-\lambda_2}e^{-\lambda_1}\frac{1}{z!}\sum_{k=0}^{z}\binom{z}{k}\lambda_2^{z-k}\lambda_1^k \\ &= e^{-(\lambda_1+\lambda_2)}\frac{1}{z!}(\lambda_1+\lambda_2)^z, \end{aligned}$$

dove nell'ultimo passaggio abbiamo usato la formula di Newton per la potenza di un binomio.

Esercizio 77. Usare la formula (18) per calcolare la legge della variabile $Z = X + Y$ che indica la somma degli esiti del lancio di due dadi (a sei facce). Calcolare valore atteso e varianza di Z.

Un'altra operazione utile tra variabili aleatorie è considerarne il *minimo* oppure il *massimo*. Precisamente, se X, $Y \in E \subseteq \mathbb{R}$, il minimo tra X e Y, $\min\{X,Y\}$ è il più piccolo tra i due valori, mentre il massimo $\max\{X,Y\}$

è il più grande tra i due. Allo stesso modo si può estendere le definizione a più di due variabili aleatorie.

Esercizio 78. Date $X = 1_A$, $Y = 1_B$ variabili indicatrici, la variabili $\min\{X, Y\}$, $\max\{X, Y\}$ sono pure variabili indicatrici. Di quali eventi?

Per calcolare la legge di $\min\{X, Y\}$, un'osservazione importante è che, per ogni $t \in \mathbb{R}$,

$$\{\min\{X, Y\} > t\} = \{X > t \quad \text{e} \quad Y > t\}.$$

Perciò, passando alle probabilità, possiamo calcolare la funzione di sopravvivenza della variabile $\min\{X, Y\}$,

$$P(\min\{X, Y\} > t|I) = P(X > t \quad \text{e} \quad Y > t|I) = P(X > t|I)P(X > t|\{Y > t\} \cap I)$$

e se X e Y sono indipendenti, concludiamo che

$$P(\min\{X, Y\} > t|I) = P(X > t|I)P(Y > t|I),$$

ossia *la funzione di sopravvivenza del minimo tra due (o più) variabili indipendenti è il prodotto delle funzioni di sopravvivenza delle singole variabili.*

Analogamente, per il massimo $\max X, Y$, usando l'identità

$$\{\min\{X, Y\} \leq t\} = \{X \leq t \quad \text{e} \quad Y \leq t\},$$

otteniamo che *la funzione di ripartizione del massimo tra due (o più) variabili indipendenti è il prodotto delle funzioni di ripartizione delle singole variabili.*

Esercizio 79. Date due variabili X, Y, Bernoulli indipendenti di parametri p, $q \in [0, 1]$, calcolare la legge di $\min X, Y$ e $\max X, Y$.

Proposizione 80 (minimo tra due geometriche indipendenti)**.** *Siano p_1, $p_2 \in [0, 1]$ e (rispetto ad una informazione I)*

(1) $X \in \{1, 2, \dots, \}$ una variabile aleatoria con legge $Geom(p_1)$,
(2) $Y \in \{1, 2, \dots\}$ una variabile aleatoria con legge $Geom(p_2)$,
(3) X e Y indipendenti.

Allora la variabile $\min\{X, Y\} \in \{1, 2, \dots\}$ *è una variabile aleatoria con legge* $Geom(p_1 + p_2 - p_1p_2)$.

Dimostrazione. Si può dimostrare questo risultato usando la funzione di sopravvivenza per una variabile geometrica di parametro $z \in [0, 1]$ che sappiamo essere $(1 - z)^k$ (per $k \in \{1, 2, \dots\}$). Siccome prendendo il minimo la funzione di sopravvivenza si moltiplica, abbiamo che $\min X, Y$ ha funzione di sopravvivenza $(1-p_1)^k(1-p_2)^k = (1-p_1-p_2+p_1p_2)^k$, che è la funzione di sopravvivenza di una variabile geometrica con parametro $p_1 + p_2 - p_1p_2$.

Un'altra dimostrazione, più intuitiva, è la seguente. Immaginiamo che le due variabili X e Y indichino rispettivamente il numero dell'estrazione relativo al primo successo in una successione di estrazioni da due urne diverse, una in cui la probabilità di successo è p_1 e l'altra p_2. Immaginiamo inoltre che le estrazioni avvengano contemporaneamente, e indichiamo con $X_i \in \{0, 1\}$ le variabili indicatrici del successo nella estrazione $i \geq 1$ dalla prima urna e $Y_i \in \{0, 1\}$ le variabili che indicano il successo nella estrazione $i \geq 1$ dalla seconda urna. A questo punto $\min\{X, Y\}$ indica il primo successo nella successione di esperimenti in cui l'esperimento $i \geq 1$ consiste

nel vedere se abbiamo avuto successo estraendo dalla prima oppure dalla seconda urna, quindi la probabilità di successo è

$$\begin{aligned} &P(X_i = 1 \,\mathrm{o}\, Y_i = 1|I) \\ &\quad = P(X_i = 1|I) + P(Y_i = 1|I) - P(X_i = 1 \,\mathrm{e}\, Y_i = 1|I) \\ &\quad = p_1 + p_2 - p_1 p_2. \end{aligned}$$

□

Appendice A. Regole di calcolo (eventi)

Riassumiamo le regole del calcolo della probabilità di eventi più utili per gli esercizi. Diamo un elenco in "ordine di utilità", per quanto possibile, per risolvere i problemi. Indichiamo con A, B, A_1, $\ldots$, A_n, I, eventi (o proposizioni). Usiamo la notazione insiemistica per le operazioni tra eventi.

i) $P(A|I) \in [0,1]$.

ii) Se A è certamente vero sapendo I, allora $P(A|I) = 1$; se è falso $P(A|I) = 0$.

iii) (Somma) $P(A^c|I) = 1 - P(A|I)$.

iv) (Prodotto) $P(A \cap B|I) = P(A|I)P(B|A \cap I)$

v) (Decomposizione in alternative) Se A_1, $\ldots A_n$ sono un sistema di alternative, allora

$$P(B|I) = \sum_{i=1}^{n} P(B|A_i \cap I)P(A_i|I).$$

vi) (Formula di Bayes) $P(B|A \cap I) = P(A|B \cap I)P(B|I)/P(A|I)$.

vii) (Eventi indipendenti) $P(B|A \cap I) = P(B|I)$, oppure

$$P(A \cap B|I) = P(A|I)P(B|I).$$

viii) (Additività, due eventi qualunque) $P(A \cup B|I) = P(A|I) + P(B|I) - P(A \cap B|I)$.

ix) (Additività, n eventi incompatibili) Se $P(A_i \cap A_j|I) = 0$ per ogni $i \neq j$, allora

$$P(\bigcup_{i=1}^{n} A_i|I) = \sum_{i=1}^{n} P(A_i|I).$$

x) (Probabilità delle cause) Se $A_1, \ldots, A_n$ sono un sistema di alternative, allora

$$P(A_i|B \cap I) = P(A_i|I) \cdot \frac{P(B|A_i \cap I)}{\sum_{j=1}^{n} P(B|A_j \cap I)P(A_j|I)}.$$

xi) (Intersezione tra n eventi)

$$P(\bigcap_{i=1}^{n} A_i|I) = P(A_1|I)P(A_2|A_1 \cap I) \ldots P(A_n|A_{n-1} \cap \ldots A_1 \cap I).$$

xii) (Monotonia) se B è vero ogni volta che A è vero, ossia $A = A \cap B$, o $A \subseteq B$ (o ancora A implica B) allora

$$P(A|I) \leq P(B|I).$$

xiii) Caso particolare della monotonia:

$$P(A \cap B|I) \leq P(A|I) \quad \text{e} \quad P(A \cap B|I) \leq P(B|I).$$

Appendice B. Regole di calcolo (variabili aleatorie)

Similmente, ricordiamo le definizioni e regole di calcolo con variabili aleatorie (viste finora) più utili negli esercizi. Indichiamo con X, Y, Z variabili aleatorie (discrete, a valori numerici) $A_1, \dots, A_n$, I, eventi, $c, \lambda \in \mathbb{R}$ costanti.

i) (legge o densità discreta) di $X \in E$ (sapendo I): $e \mapsto P(X = e|I)$.

ii) (valore atteso e funzione composta) Se $X \in E$, $f : E \to \mathbb{R}$, allora
$$\mathbb{E}\left[f(X)|I\right] = \sum_{e \in E} f(e) \cdot P(X = e|I).$$

iii) (linearità) $\mathbb{E}\left[cX + Y|I\right] = c\mathbb{E}\left[X|I\right] + \mathbb{E}\left[Y|I\right]$. Se $X = c$ è costante, $\mathbb{E}\left[c|I\right] = c$.

iv) (decomposizione) Se $A_1, \dots, A_n$ sono un sistema di alternative,
$$\mathbb{E}\left[X|I\right] = \sum_{i=1}^{n} \mathbb{E}\left[X|A_i \cap I\right] P(A_i|I).$$

v) (monotonia) Se $X \geq 0$, allora $\mathbb{E}\left[X|I\right] \geq 0$.

vi) (varianza) $\operatorname{Var}(X|I) = \mathbb{E}\left[(X - \mathbb{E}\left[X|I\right])^2|I\right] = \mathbb{E}\left[X^2|I\right] - (\mathbb{E}\left[X|I\right])^2$. $\operatorname{Var}(X|I) \geq 0$

vii) (quadraticità) $\operatorname{Var}(X|I) = 0$ se e solo se $X = \mathbb{E}\left[X|I\right]$.
$$\operatorname{Var}(\lambda X + c|I) = \lambda^2 \operatorname{Var}(X|I).$$

viii) (funzione di ripartizione) di $X \in E \subseteq \mathbb{R}$:
$$t \mapsto P(X \leq t|I) = \sum_{\substack{e \in E \\ e \leq t}} P(X = e|I)$$

ix) (funzione di sopravvivenza) di $X \in E \subseteq \mathbb{R}$:
$$t \mapsto P(X > t|I) = 1 - P(X \leq t|I)$$

Appendice C. Estrazioni dall'urna

Raccogliamo i principali risultati visti riguardanti il modello dell'urna contenente N palline, di cui R rosse e B blu (scriviamo $I(N,R,B)$).

Estrazioni senza reimmissione.

i) Le varie estrazioni *non sono* indipendenti. Tenere conto dello stato dell'urna (l'ordine non è importante), ad esempio

$$P(R_1|R_2 \cap I(N,R,B)) = \frac{R-1}{N-2}, \quad P(R_1|B_2 \cap I(N,R,B)) = \frac{R}{N-1}.$$

ii) $P(Ri|I(N,R,B)) = R/N$ per ogni $i \in \{1,\dots,N\}$.

iii) Probabilità di estrarre una fissata sequenza ordinata σ, lunga $n \leq N$ contenente $r \leq R$ rosse e $b \leq B$ blu:

$$P(\sigma|I(N,R,B)) = \frac{R(R-1)\dots(R-r+1)\cdot B(B-1)\dots(B-b+1)}{N(N-1)\dots(N-n+1)}.$$

iv) In $n \leq N$ estrazioni, poniamo $X \in \{0,1,\dots,R\}$ il numero di palline rosse estratte. Allora X ha legge ipergeometrica (dove $b = n-r$)

$$P(X = r|I(N,R,B)) = \frac{\binom{R}{r}\binom{B}{b}}{\binom{N}{n}}.$$

Estrazioni con reimmissione.

i) Le varie estrazioni *sono* indipendenti. Ad esempio,

$$P(R_1|R_2 \cap I(N,R,B)) = \frac{R}{N} \quad P(R_1|B_2 \cap I(N,R,B)) = \frac{R}{N}.$$

ii) $P(Ri|I(N,R,B)) = R/N$ per ogni $i \in \{1,\dots,N\}$.

iii) Probabilità di estrarre una fissata sequenza ordinata σ, lunga $n \leq N$ contenente $r \leq R$ rosse e $b \leq B$ blu:

$$P(\sigma|I(N,R,B)) = \frac{R^r B^b}{N^n} = p^r(1-p)^{n-r} \quad \text{con } p = R/N.$$

iv) In $n \leq N$ estrazioni, poniamo $X \in \{0,1,\dots,R\}$ il numero di palline rosse estratte. Allora X ha legge binomiale $B(n,p)$,

$$P(X = r|I(N,R,B)) = \binom{n}{r} p^r(1-p)^{n-r} \quad \text{con } p = R/N..$$

Appendice D. Riassunto delle principali leggi discrete

Bernoulli (indicatrice). Parametro $p \in [0,1]$. Possibili valori $X \in \{0,1\}$. Densità

$$P(X=1|\text{Bernoulli}(p)) = p \quad P(X=0|\text{Bernoulli}(p)) = 1-p.$$

$$\mathbb{E}\left[X|\text{Bernoulli}(p)\right] = p \quad \text{Var}\left(X|\text{Bernoulli}(p)\right) = p(1-p).$$

Dato un evento A, $X = 1_A$ è Bernoulli di parametro $p = P(A|I)$.

Uniforme (intervallo discreto). Parametro $n \in \{1,2,\ldots\}$. Possibili valori $X \in \{1,2,\ldots,n\}$. Densità

$$P(X=k|\text{Unif}\,\{1,\ldots,n\}) = 1/n \quad \text{per } k \in \{1,\ldots,n\}.$$

$$\mathbb{E}\left[X|\text{Unif}\,\{1,\ldots,n\}\right] = \frac{n+1}{2} \quad \text{Var}\left(X|\text{Unif}\,\{1,\ldots,n\}\right) = \frac{n^2-1}{12}.$$

Binomiale. Parametri $n \in \{1,2,\ldots\}$, $p \in [0,1]$. Possibili valori $X \in \{0,1,\ldots,n\}$. Densità

$$P(X=k|B(n,p)) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{per } k \in \{0,\ldots,n\}.$$

$$\mathbb{E}\left[X|B(n,p)\right] = np \quad \text{Var}\left(X|B(n,p)\right) = np(1-p).$$

Indica il numero di successi in n esperimenti indipendenti, ciascuno con probabilità di successo p. Se $n = 1 \Rightarrow X \in \{0,1\}$ Bernoulli(p).

Poisson. Parametro $\lambda > 0$ (reale). Possibili valori $X \in \mathbb{N} = \{0,1,2,\ldots\}$. Densità

$$P(X=k|\text{Poisson}(\lambda)) = e^{-\lambda}\lambda^k/(k!) \quad \text{per } k \in \{0,1,2,\ldots\}.$$

$$\mathbb{E}\left[X|\text{Poisson}(\lambda)\right] = \lambda \quad \text{Var}\left(X|\text{Poisson}(\lambda)\right) = \lambda.$$

Approssima $B(n,p)$ con p piccolo, n grande e $np \approx \lambda$.

Geometrica. Parametro $p \in [0,1]$. Possibili valori $X \in \{1,2,3,\ldots\}$. Densità

$$P(X=k|\text{Geom}(p)) = (1-p)^{k-1}p \quad \text{per } k \in \{1,2,3,\ldots\}.$$

$$\mathbb{E}\left[X|\text{Geom}(p)\right] = \frac{1}{p} \quad \text{Var}\left(X|\text{Geom}(p)\right) = \frac{1-p}{p^2}.$$

Indica il numero del tentativo del primo successo in una successione (infinita) di esperimenti indipendenti, ciascuno con probabilità p di successo.