Divagazioni in margine all'*Introduzione alla Probabilità* di P. Baldi

A. Visintin — Facoltà di Ingegneria di Trento — a.a. 2010-11

Indice

Le secret d'ennuyer est celui de tout dire. (Voltaire)

Premessa. Questo scritto raccoglie alcune osservazioni, perlopiù riferite al testo di Baldi del 2003 (nel seguito citato semplicemente come [B]), e diverse divagazioni, volte a facilitare la comprensione dell'argomento. Sono poca cosa, poiché il [B] è ben fatto e sufficiente ai fini del corso.

Queste pagine sono scaricabili in pdf dal sito http://www.science.unitn.it/∼visintin/ ove si trovano anche altre informazioni circa il corso.

Gli Obiettivi del Minicorso di Probabilità. Introdurre alcune nozioni fondamentali di statistica descrittiva, utili persino per la vita di tutti i giorni: istogrammi, media, mediana, quantili, boxplots, varianza, regressione lineare, ecc..

Introdurre gli spazi di probabilità, e stabilirne alcune proprietà fondamentali mediante operazioni insiemistiche elementari. Presentare il calcolo combinatorio.

Definire la probabilità condizionata, da questa derivare la classica formula di Bayes, ed introdurre l'indipendenza di eventi.

Introdurre la teoria delle variabili aleatorie discrete e continue, e le principali distribuzioni di probabilità. Derivare il teorema dei grandi numeri, ed enunciare il teorema limite centrale.

Svolgere diversi esercizi, usando strumenti di analisi e di calcolo delle probabilità.

1 Statistica descrittiva

Cosa Sono la Statistica ed il Calcolo delle Probabilità? La statistica è la disciplina che studia la raccolta, l'organizzazione, l'elaborazione e l'interpretazione dei dati. [1]

Comunemente si distingue tra *statistica descrittiva,* che tratta la raccolta, l'organizzazione e (appunto) la descrizione sintetica dei dati, e *statistica inferenziale (o deduttiva o induttiva o matematica:* sono tutti sinonimi), che trae conclusioni probabilistiche dai dati usando massicciamente concetti e metodi del *calcolo delle probabilità.* Con quest'ultimo si intende l'apparato matematico che tratta la nozione di probabilità per la *modellizzazione* (ovvero la rappresentazione matematica) di fenomeni aleatori. [2]

Gran parte del calcolo delle probabilità è volto a determinare la probabilità di eventi complessi a partire dalla probabilità di eventi elementari. Compito fondamentale della statistica inferenziale è

[1] Siamo sempre più sommersi da dati, dai quali diventa sempre più importante saper estrarre informazioni — un'operazione meno banale di quanto possa sembrare.

[2] La denominazione di *calcolo* è tradizionale, e ci permette di riunire i corsi matematici di base sotto un comune denominatore: accanto al calcolo integro-differenziale (ovvero la più blasonata analisi matematica), abbiamo il calcolo algebrico (ovvero l'algebra lineare), il calcolo numerico (ovvero la più paludata analisi numerica), ed appunto il calcolo delle probabilità. Nella consuetudine solo quest'ultimo resta privo di una denominazione ... nobiliare.

Oltre ci sarebbero l'analisi delle equazioni differenziali, l'analisi di Fourier (una pallida idea è fornita dagli ultimi due capitoli del [Bramanti, Pagani, Salsa]), l'analisi complessa (ovvero in $\mathbf{C}$ piuttosto che in $\mathbf{R}$), ecc. E naturalmente la fisica-matematica, disciplina di cerniera tra ingegneria, fisica e matematica. E questo conclude questa piccola *Weltanschaung* matematica per l'ingegneria, senz'altro incompleta.

il problema inverso, ovvero risalire alle probabilità di eventi elementari partendo dalla probabilità di eventi complessi.

La statistica descrittiva può usare strumenti matematici, ad esempio di algebra lineare. Comunque solitamente la componente matematica della statistica inferenziale è ben più ampia. In ogni caso la distinzione tra statistica inferenziale e calcolo delle probabilità è a volte sfumata.

Ad esempio, se precedentemente ad un'elezione si effettua un sondaggio su un campione necessariamente ridotto di votanti, la raccolta e l'organizzazione dei dati del campione rientra nella statistica descrittiva. Ma l'estrapolazione di tali risultati allo scopo di desumere informazioni circa l'orientamento dell'intero corpo elettorale, unitamente ad una stima dei margini di errore, è compito della statistica inferenziale. In seguito all'elezione, l'organizzazione dei dati e la sintesi dei risultati è ancora affidata alla statistica descrittiva.

Pensiamo anche al classico esempio delle estrazioni, con o senza reimmissioni, da un'urna contenente biglie di diversi colori. Mediante il calcolo delle probabilità, note le probabilità delle estrazioni elementari (ovvero le percentuali delle biglie dei diversi colori) si potrà determinare la probabilità di eventi più complessi (ad esempio, la probabilità di estrarre biglie tutte dello stesso colore). Se però non si ha accesso all'urna, si dovranno effettuare alcune estrazioni per cercare di stimare il numero delle biglie dei diversi colori, ovvero la probabilità delle estrazioni elementari. Quindi:

(i) prima facciamo delle estrazioni volte ad identificare la composizione dell'urna, ed rappresentiamo i dati raccolti mediante la statistica descrittiva;

(ii) sulla base di quei risultati e mediante la statistica inferenziale, stimiamo le probabilità degli eventi elementari;

(iii) infine, mediante il calcolo delle probabilità, possiamo determinare le probabilità di eventi più complessi.

In questa presentazione (invero alquanto schematica) mancano dei protagonisti importanti: a monte la *modellistica,* uno dei punti di contatto tra l'elaborazione matematica e la realtà, ed il *calcolo numerico,* che ovviamente si avvale dei moderni calcolatori elettronici.

Due Modi di Sommare. Sia $\{x_i\}_{i=1,...,N}$ un campione di dati numerici, ovvero con $x_i \in \mathbf{R}$ per ogni i. Siano $\{z_j\}_{j=1,...,M}$ le corrispondenti *modalità,* ovvero i valori assunti dal complesso delle x_i. Si noti che $M \leq N$, poiché diversi elementi del campione possono assumere la stessa modalità: e.g. [3] $x_2 = x_5$. Sia f una funzione $\mathbf{R} \to \mathbf{R}$. Si possono sommare gli $f(x_i)$ rispetto agli elementi del campione (ovvero gli i), oppure rispetto alle modalità, dopo aver sostituito gli $f(x_i)$ con gli $\{f(z_j)$, pesati con i rispettivi *effettivi.* [4] Più esplicitamente, posto

$$\alpha_j = \{i : x_i = z_j\}, \qquad N_j = \#\alpha_j \qquad \text{per } j = 1, ..., M \tag{1.1}$$

(N_j è il numero di elementi che costituiscono α_j, ovvero l'effettivo della modalità z_j), abbiamo

$$\sum_{i=1}^{N} f(x_i) = \sum_{j=1}^{M} \sum_{i \in \alpha_j} f(x_i) = \sum_{j=1}^{M} \sum_{i \in \alpha_j} f(z_j) = \sum_{j=1}^{M} N_j f(z_j). \tag{1.2}$$

In particolare, definita la proporzione (o *frequenza relativa)* $q_j := N_j/N$ per $j = 1, ..., M$,

$$\bar{x} := \tfrac{1}{N} \textstyle\sum_{i=1}^{N} x_i = \sum_{j=1}^{M} q_j z_j, \tag{1.3}$$

$$\sigma_x^2 := \tfrac{1}{N} \textstyle\sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{j=1}^{M} q_j (z_j - \bar{x})^2; \tag{1.4}$$

ed anche, si veda il calcolo di [B, p. 7],

$$\sigma_x^2 = \left(\frac{1}{N} \sum_{i=1}^{N} x_i^2 \right) - \bar{x}^2 = \left(\sum_{j=1}^{M} q_j z_j^2 \right) - \bar{x}^2. \tag{1.5}$$

[3] e.g.: exempli gratia = ad esempio.

[4] Qui ed altrove si usa la terminologia del [B].

Si noti che, poiché le proporzioni q_j sono ≥ 0 ed hanno somma 1, esse sono la densità di una distribuzione di probabilità.

Consideriamo ora il caso di due campioni numerici $X = \{x_i\}_{i=1,...,N}$ e $Y = \{y_\ell\}_{j=\ell,...,P}$, con rispettive modalità $\{z_j\}_{j=1,...,M}$ e $\{w_m\}_{m=1,...,Q}$. Abbiamo anche le *modalità congiunte*

$$\{(z_j, w_m) : j = 1, ..., M,\ m = 1, ..., Q\},$$

che hanno effettivi L_{jm} e frequenze relative congiunte q_{jm}:

$$L_{jm} = \#\{(i, \ell) : (x_i, y_\ell) = (z_j, w_m)\}, \quad q_{jm} = \frac{L_{jm}}{NP} \qquad \text{per } j = 1, ..., M,\ m = 1, ..., Q.$$

In modo analogo a sopra si può rappresentare la covarianza del campione

$$\sigma_{xy} := \frac{1}{NP} \sum_{i=1}^{N} \sum_{\ell=1}^{P} (x_i - \bar{x})\,(y_\ell - \bar{y}) = \sum_{j=1}^{M} \sum_{m=1}^{Q} q_{jm} (z_j - \bar{x})\,(w_m - \bar{y}); \tag{1.6}$$

o anche, sviluppando il prodotto,

$$\sigma_{xy} = \left(\frac{1}{NP} \sum_{i=1}^{N} \sum_{\ell=1}^{P} x_i y_\ell \right) - \bar{x}\bar{y} = \left(\sum_{j=1}^{M} \sum_{m=1}^{Q} q_{jm} z_j w_m \right) - \bar{x}\bar{y}. \tag{1.7}$$

Anche in questo caso le proporzioni q_{jm} sono la densità di una distribuzione di probabilità.

Se i due campioni coincidono (ovvero $Y = X$), ritroviamo la varianza del campione: $\sigma_{xx} = \sigma_x^2$.

Componenti Principali in Analisi Multivariata. Per *analisi statistica multivariata* s'intende l'analisi statistica di dati multidimensionali (ovvero vettoriali). L'individuazione delle cosiddette *componenti principali* è un'importante tecnica di rappresentazione sintetica di dati vettoriali. Supponiamo di avere un campione [5] di ampiezza M di dati N-dimensionali (ovvero, abbiamo M vettori di $\mathbf{R}^N$). Cerchiamo di determinare una rotazione del sistema di riferimento ortogonale di $\mathbf{R}^N$, in modo tale che, denotando con $Y_1, ..., Y_N$ gli assi del nuovo sistema di riferimento, per ogni $m \in \{1, ..., N\}$ la varianza di $Y_1, ..., Y_m$ sia massima. Più specificamente, gli assi sono individuati uno dopo l'altro, in modo tale che l'asse m-esimo massimizzi la varianza tra tutte le direzioni che sono indipendenti da $Y_1, ..., Y_{m-1}$ (le quali sono stati individuate ai passi precedenti).

Ad esempio, sia $\{X_i = (X_{i1}, X_{i2}, X_{i3})\}_{1 \leq i \leq M}$ un campione di *ampiezza* M di dati tridimensionali, ciascuno decomposto nelle sue componenti su tre assi ortogonali prefissati. Per via dell'orgonalità degli assi, la varianza totale del campione vale

$$\sum_{i=1}^{M} |X_i|^2 = \sum_{i=1}^{M} |X_{i1}|^2 + \sum_{i=1}^{M} |X_{i2}|^2 + \sum_{i=1}^{M} |X_{i3}|^2 \geq \sum_{i=1}^{M} |X_{i1}|^2.$$

Ovvero, la varianza del campione è non minore della varianza della prima componente del campione. Questo traduce il fatto che i componenti del campione lungo il primo asse, ovvero le proiezioni dei dati lungo quella direzione, contengono meno informazioni del campione stesso — poiché ogni componente di un vettore contiene meno informazione dell'intero vettore. (Ovviamente, invece del primo asse avremmo potuto sceglierne un altro.)

Tra tutte le possibili direzioni dello spazio ve n'è una, chiamiamola ξ, che massimizza la varianza del campione lungo quella direzione. Questa viene selezionata come primo asse principale di quel campione. Il secondo asse principale principale verrà individuato nel piano ortogonale a ξ come segue. Innanzi tutto ruotiamo gli assi in modo che ξ sia la direzione del primo asse nel nuovo riferimento. Poi spogliamo il campione della sua componente nella direzione ξ, riducendolo ad un campione ancora di ampiezza M ma di dati bidimensionali. Quindi, tra tutte le direzioni ortogonali a ξ, individuiamo

[5] Per *campione* qui possiamo intendere semplicemente un insieme strutturato di dati.

il secondo asse principale massimizzando la varianza di questo campione bidimensionale. E così poi procediamo per individuare i restanti assi principali — anzi nel nostro esempio ci fermiamo, poiché la scelta del terzo asse ortogonale è obbligata (insomma, abbiamo finito gli assi).

Sottolineiamo che gli assi principali dipendono dal campione. In diversi casi, bastano i primissimi assi principali per esprimere le caratteristiche più salienti anche di campioni con tante componenti (cioè con N grande). Inoltre campioni diversi estratti da una popolazione abbastanza omogenea possono avere assi principali che differiscono di poco tra i diversi campioni, cosicché basta individuare gli assi principali una volta per tutte.

Si può dimostrare che il primo asse principale ha la direzione dell'autovettore associato al più grande autovalore della matrice di covarianza $\{\mathrm{Cov}(X_i, X_j)\}_{i,j=1,\dots,N}$; analogamente, il secondo asse principale ha la direzione dell'autovettore associato al secondo autovalore della stessa matrice, e così via. L'effettiva implementazione di questo metodo quindi richiede il calcolo di autovalori ed autovettori della matrice di covarianza.

Questo metodo è stato ampiamente impiegato nell'ultimo secolo ad esempio in biologia. La statistica descrittiva offre anche altri metodi per estrarre informazioni dai dati. Ad esempio la *cluster analysis* (ovvero, analisi dei raggruppamenti) cerca di individuare i modi più significativi di ripartire in gruppi un campione di grande ampiezza. Non presenteremo queste procedure, che possono essere reperite ad esempio sul [Baldi 1996].

2 Spazi di probabilità e calcolo combinatorio

Interpretazioni del Concetto di Probabilità. In estrema sintesi, possiamo distinguere le seguenti impostazioni emerse storicamente.

(i) La teoria classica avviata alla metà del '600 dai pionieri del calcolo delle probabilità (Pierre Fermat, Blaise Pascal, Christiaan Huygens, ecc.), originata da alcuni quesiti circa i giochi di azzardo, e centrata sulla nozione di equiprobabilità — ovvero distribuzione uniforme di probabilità per insiemi finiti. Qui la probabilità di un evento è tipicamente vista come rapporto tra il numero dei casi favorevoli e quello dei casi possibili, e quindi si basa sul *calcolo combinatorio.* [6] Rientrano comunque in questa stagione pionieristica anche la derivazione della Formula di Bayes e dei primi teoremi limite.

(ii) L'interpretazione *frequentista* del concetto di probabilità, sviluppata nel '700 soprattutto in Inghilterra, che definisce la probabilità di un evento come il limite a cui tende il rapporto tra il numero dei casi in cui si è verificato l'evento ed il numero di esperimenti, al tendere di quest'ultimo all'infinito. In tal modo si attribuisce alla probabilità una base del tutto empirica, coerentemente con la tradizione filosofica anglosassone.

Questa definizione di probabilità trova fondamento nel Teorema dei Grandi Numeri (dimostrata da Jakob Bernoulli già nel suo trattato del 1713). L'applicazione di questo punto di vista è necessariamente ristretto agli eventi indefinitamente ripetibili.

(iii) L'approccio *assiomatico,* codificato nel 1933 dal grande matematico russo A.N. Kolmogorov, in cui la probabilità è intepretata come una *misura* non negativa e di massa totale 1, ed èè quindi trattata nell'ambito della teoria matematica della misura, sviluppata all'inizio del '900. Al giorno d'oggi questo è l'approccio più comunemente adottato.

(iv) L'interpretazione *soggettivista,* introdotto dal matematico italiano De Finetti nella prima metà del '900, secondo cui la probabilità di un evento esprime il grado di fiducia nel suo verificarsi, e quindi è frutto di una valutazione soggettiva (piuttosto che oggettiva come nell'approccio frequentista). Questo può permettere di attribuire una probabilità ad eventi irripetibili.

Questa schematizzazione è alquanto rudimentale, ed incompleta; ad esempio trascura la statistica, che ha diversi punti di contatto con il calcolo delle probabilità. Ad esempio, la classica formula di Bayes (del 1763) ha dato luogo a notevolissimi sviluppi di statistica inferenziale, che ben si inquadra-

[6] Con quest'ultimo di intende lo studio della cardinalità (ovvero la *numerosità)* degli insiemi costituiti da un numero finito di elementi.

no nell'impostazione soggettivista. Un'altra scuola di pensiero è invece più incline ad un approccio frequentista alla statistica.

Una conciliazione di queste opinioni non sembra in vista. Un autorevole commentatore ha osservato che raramente è stato registrato un simile disaccordo, perlomeno successivamente alla costruzione della Torre di Babele.

La Nozione di σ-Algebra. Il Baldi, dovendo mettere tutto nero su bianco, non se l'è sentita di omettere questo concetto; questo l'ha costretto a varie precisazioni e verifiche in diversi punti del testo. A noi, che in questo corso siamo meno interessati agli aspetti teorici, questo più che difficile può forse apparire un po' sterile. In effetti nei casi più tipici la σ-algebra è costituita da $\mathcal{P}(\Omega)$ — l'insieme delle parti di Ω, ovvero la famiglia di tutti i sottoinsiemi di Ω. Ora ogni buon matematico sa che, sviluppando la teoria delle variabili aleatorie continue, non è corretto assumere $\mathcal{P}(\Omega)$ come σ-algebra, perché in tal modo si includono degli insiemi estremamente patologici. Verissimo, ma nella vita ed anche nella matematica di ogni giorno affrontiamo rischi ben maggiori senza curarcene minimamente: il rischio che ha un ingegnere di imbattersi in un insieme patologico del tipo qui paventato è ben minore di quello che gli cada in testa un meteorite.

Quindi possiamo ben semplificarci la vita prendendo $\mathcal{P}(\Omega)$ come σ-algebra. E possiamo farlo senza suscitare la disapprovazione di Baldi, che pure (immagino) lo avrebbe fatto, se solo non avesse dovuto affidarlo per iscritto all'eternità...

Il σ compare anche nella locuzione di σ-additività, ovvero l'estensione ad una successione di eventi della proprietà di additività della misura di probabilità. Questa nozione è invece ineludibile.

L'Evento Certo e l'Evento Impossibile. L'evento *certo* è Ω, ovvero l'intero spazio dei possibili risultati. Uno degli assiomi della teoria della misura prescrive $\mathbf{P}(\Omega) = 1$, comunque possono esservi eventi non certi di probabilità 1; questi eventi sono detti *quasi certi.* Analogamente, $\emptyset$ rappresenta l'evento *impossibile;* applicando un altro assioma della teoria della misura, abbiamo $\mathbf{P}(\emptyset) = \mathbf{P}(\Omega \setminus \Omega) = 1 - \mathbf{P}(\Omega) = 0$. Possono esservi eventi non impossibili di probabilità nulla, che saranno detti *quasi impossibili.* [7] Ad esempio, $\Omega =]0, 1[$ sia dotato della misura euclidea, che ovviamente è una misura di probabilità (ovvero, non negativa, di massa totale 1, oltre che σ-additiva). L'insieme $\{0.5\}$ ha misura nulla, e lo stesso vale per ogni unione finita ed ogni successione di punti.

Indipendenza di Eventi. Per ogni coppia di interi $m, n \geq 2$ l'indipendenza per n-ple non implica quella per m-ple. Ecco un controesempio per $m = 3, n = 2$. Si doti l'insieme $\Omega = \{1, 2, 3, 4\}$ della densità di probabilità uniforme. Si verifica immediatamente che la famiglia $\{1, 2\}, \{1, 3\}, \{2, 3\}$ è indipendente per coppie ma non per terne.

Per le v.a. invece, se $n \leq m$, l'indipendenza per m-ple implica quella per n-ple (ma non viceversa). Questo segue dalla definizione 3.16 [B, p. 53]: basta scegliere $m - n$ degli A_i uguali a Ω...

Eventi Indipendenti e loro Complementari. Una famiglia $A_1, ..., A_N \in \mathcal{P}(\Omega)$ di eventi indipendenti resta tale, se uno o più di essi sono sostituiti dal rispettivo complementare in Ω.

Lo verifichiamo per $N = 2$. Ovvero, supponiamo che $A_1, A_2 \in \mathcal{P}(\Omega)$ siano indipendenti, e mostriamo che ad esempio lo stesso vale per A_1, A_2' (quest'ultimo indica il complementare di A_2 in Ω). Infatti, poiché

$$A_1 \cap A_2' = A_1 \setminus (A_1 \cap A_2), \quad A_1 \cap A_2 \subset A_1 \quad \text{(per insiemistica elementare)},$$

$$\mathbf{P}(A_1 \cap A_2) = \mathbf{P}(A_1) \cdot \mathbf{P}(A_2) \quad \text{(per ipotesi)},$$

abbiamo

$$\begin{aligned} &\mathbf{P}(A_1 \cap A_2') = \mathbf{P}(A_1 \setminus (A_1 \cap A_2)) = \mathbf{P}(A_1) - \mathbf{P}(A_1 \cap A_2) \\ &= \mathbf{P}(A_1) - \mathbf{P}(A_1) \cdot \mathbf{P}(A_2) = \mathbf{P}(A_1)[1 - \mathbf{P}(A_2)] = \mathbf{P}(A_1) \cdot \mathbf{P}(A_2'). \end{aligned} \tag{2.1}$$

[7] Nel calcolo delle probabilità, in generale il termine *quasi* è riferito ad eventi o proprietà che valgono a meno di insiemi di misura nulla.

Formula della Probabilità Totale. Il [B] riporta questa formula ((2.7) a p. 31), senza sottolinearne l'importanza. Sia $\{A_i\}_{i=1,2,...}$ una partizione di Ω (l'evento certo), ovvero una famiglia di sottoinsiemi disgiunti di Ω (al più una successione in questo caso) la cui unione è tutto Ω. Allora, per ogni $B \subset \Omega$,

$$B = B \cap \Omega = B \cap \bigcup_i A_i = \bigcup_i (B \cap A_i);$$

essendo questa un'unione di insiemi disgiunti, otteniamo quindi

$$\mathbf{P}(B) = \mathbf{P}\Big(\bigcup_i (B \cap A_i)\Big) = \sum_i \mathbf{P}(B \cap A_i) \qquad \forall B \subset \Omega. \tag{2.2}$$

Possiamo fare un ulteriore passo: usando la definizione di probabilità condizionata, perveniamo alla *formula della probabilità totale* (detta anche *formula della partizione dell'evento certo* o *formula delle alternative):*

$$\mathbf{P}(B) = \sum_i \mathbf{P}(B \cap A_i) = \sum_i \mathbf{P}(B|A_i)\mathbf{P}(A_i) \qquad \forall B \subset \Omega. \tag{2.3}$$

Si noti che possiamo scrivere la seconda somma solo se $\mathbf{P}(A_i) \neq 0$ per ogni i (perché?).

Formulario di Calcolo Combinatorio.

Numero di sottoinsiemi di un insieme di n elementi: 2^n.

Numero di disposizioni di un insieme di n elementi: $\#D_k^n = \dfrac{n!}{(n-k)!}$.

Numero di combinazioni di un insieme di n elementi =
numero di sottoinsiemi di k elementi di un insieme di n elementi:

$$\#C_k^n = \binom{n}{k} = \frac{n!}{k!(n-k)!} \quad \text{(detto coefficiente binomiale).}$$

Numero di permutazioni di un insieme di n elementi: $\#P_n = n!$.

Numero di partizioni di un insieme di n elementi in m sottoinsiemi: m^n.

Numero di partizioni di un insieme di n elementi in m sottoinsiemi, rispettivamente di cardinalià $k_1, ..., k_m$, con $k_1 + ... + k_m = n$:

$$\#C^n_{k_1,...,k_m} = \binom{n}{k_1 \, ... \, k_m} = \frac{n!}{k_1! \, ... \, k_m!} \quad \text{(detto coefficiente multinomiale).}$$

Si noti che $\#C_k^n = \#C^n_{k,n-k}$ $(= \#C^n_{n-k})$.

La Distribuzione Ipergeometrica. Il [B] (al pari di altri testi) introduce questa distribuzione di probabilità nell'ambito del calcolo combinatorio. Comunque anche questa distribuzione discende da un risultato combinatorio.

Fissiamo tre numeri interi r, b, n, con $n \leq r + b$. Consideriamo un insieme I di due tipi di elementi, diciamo b biglie bianche e r biglie rossee sia k un intero tale che $0 \leq k \leq r$. Ci chiediamo quanti diversi sottoinsiemi di n elementi contenenti esattamente k biglie rosse si possono estrarre da I. La risposta è $\binom{r}{k}\binom{b}{n-k}$. Se vogliamo individuare la probabilità di un tale tipo di estrazione, dobbiamo dividere il risultato per il numero t delle possibili estrazioni di n biglie da I, ovvero $\binom{r+b}{n}$. Pertanto, indicando con X il numero di biglie rosse estratte (senza reimmissione), la distribuzione di questa v.a. è

$$\begin{aligned} &\mathbf{P}(X = k) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{r+b}{n}} && \text{per } k = 0, ..., r, \\ &\mathbf{P}(X = k) = 0 && \text{per ogni altro } k \in \mathbf{R}. \end{aligned} \tag{2.4}$$

Diremo che X ha distribuzione ipergeometrica di parametri r, b, n, ovvero $X \sim \text{Iper}(r, b, n)$.

3 Variabili aleatorie discrete

La Speranza. [8] La speranza di una v.a. discreta $X : \Omega \to \mathbf{R}$ è definita dal [B] mediante la legge di X. Tuttavia esiste una definizione equivalente che fa riferimento direttamente alla v.a. (senza coinvolgere la sua legge), e che può meglio chiarire certe proprietà della speranza.

Cominciamo con assumere che l'insieme Ω sia finito o al più è una successione di punti. Sia allora $\Omega = \{\omega_1, ..., \omega_n, ...\}$, e per ogni $X : \Omega \to \mathbf{R}$ poniamo

$$\mathbf{E}(X) = \sum_{i=1}^{\infty} X(\omega_i)\mathbf{P}(\{\omega_i\}) \ \Big(= \sum_{\omega\in\Omega} X(\omega)\mathbf{P}(\{\omega\})\Big) \tag{3.1}$$

se questa serie converge assolutamente. Si confronti questa definizione con quella che impiega la legge $\mathbf{P}_X$ di X, [B, p. 60], che qui riscriviamo in modo del tutto equivalente:

$$\mathbf{E}(X) = \sum_j y_j \mathbf{P}_X(y_j) \ \Big(= \sum_{y\in X(\Omega)} y\mathbf{P}_X(\{y\})\Big), \tag{3.2}$$

sempre assumendo la convergenza assoluta.

Possiamo dimostrare l'equivalenza tra queste due definizioni raggruppando gli indici che corrispondono ad uno stesso valore di X (ovvero alla stessa *modalità),* mediante un procedimento analogo a quello di (1.2). Con notazione standard poniamo

$$X(\Omega) = \{X(\omega) : \omega \in \Omega\}, \quad \text{ovvero in questo caso} \quad X(\Omega) = \{X(\omega_i) : i = 1, 2, ...\}. \tag{3.3}$$

Allora $X(\Omega)$ è un insieme finito o al più una successione di numeri: $X(\Omega) = \{y_1, y_2, ...\}$, che ovviamente non è più numeroso di Ω. Definiamo

$$A_j = X^{-1}(y_j), \qquad \alpha_j = \{i : X(\omega_i) = y_j\} \qquad \forall j; \tag{3.4}$$

quindi $X(\omega_i) = y_j$, ovvero $\omega_i \in X^{-1}(y_j)$, per ogni $i \in \alpha_j$. Osserviamo che

$$\mathbf{P}_X(y_j) := \mathbf{P}(X^{-1}(y_j)) = \sum_{i\in\alpha_j} \mathbf{P}(\{\omega_i\}) \qquad \forall j.$$

Pertanto, sempre assumendo la convergenza assoluta,

$$\mathbf{E}(X) \overset{(3.1),(3.4)}{=} \sum_j \sum_{i\in\alpha_j} X(\omega_i)\mathbf{P}(\{\omega_i\}) = \sum_j y_j \sum_{i\in\alpha_j} \mathbf{P}(\{\omega_i\}) = \sum_j y_j \mathbf{P}_X(y_j). \tag{3.5}$$

Abbiamo quindi ritrovato la (3.2).

Assumiamo ora che Ω sia qualsiasi, ma f sia *costante a tratti:* con questo intendiamo che f è della forma $f = \sum_i a_i 1_{A_i}$, con $a_i \in \mathbf{R}$ per $i = 1, 2, ...$, ed $\{A_i\}_{i=1,2,...}$ è una partizione di Ω, ovvero una famiglia di sottoinsiemi disgiunti di Ω (al più una successione in questo caso) la cui unione è tutto Ω. Allora abbiamo

$$\mathbf{E}(f) = \mathbf{E}\Big(\sum_i a_i 1_{A_i}\Big) = \sum_i a_i \mathbf{E}(1_{A_i}) = \sum_i a_i \mathbf{P}(A_i), \tag{3.6}$$

se questa serie converge assolutamente.

Distribuzione di Probabilità e Legge. In Calcolo delle Probabilità questi sono due sinonimi (qualcuno usa anche il termine di *probabilità immagine).* Una variabile aleatoria ("v.a.", per brevità) $X : \Omega \to \mathbf{R}$ fa corrispondere un numero ad ogni evento elementare $\omega \in \Omega$; questo dipersè non

[8] Per la speranza tradizionalmente si usa il simbolo $\mathbf{E}$, che può andare bene per Inglesi, Francesi e Tedeschi, che rispettivamente usano i termini Expected value, Espérance, Erwartungswert. Gli Italiani invece parlano di Speranza, Media, Valore atteso, ... niente E.

sembrerebbe determinare una probabilità. Ma, poiché Ω è munito di una misura di probabilità, in effetti X determina un'altra misura di probabilità, questa volta su $\mathbf{R}$:

$$\mathbf{P}_X(A) := \mathbf{P}(X^{-1}(A)) = \mathbf{P}(X \in A) \qquad \forall A \subset \mathbf{R}. \tag{3.7}$$

Questo ci dice come la probabilità di essere assunto da X si distribuisce tra i diversi sottoinsiemi di $\mathbf{R}$. Questo giustifica pienamente la denominazione di *distribuzione di probabilità;* l'origine del termine *legge* invece sembra meno chiara.

Si parla di leggi fisiche, come se ci fosse una prescrizione nel comportamento fisico. Questo punto di vista è applicabile alla probabilità? La (cosiddetta) *Legge dei Grandi Numeri* sembra prescrivere la convergenza delle medie empiriche (ovvero quelle rilevate ripetendo un esperimento) al valor medio (ovvero la speranza matematica) [B, p. 42]. Sembra quasi una legge fisica, e forse è per questo che tradizionalmente si parla di Legge dei Grandi Numeri. Comunque, sia chiaro che la Legge dei Grandi Numeri è in effetti un teorema, e soprattutto che qui il termine "legge" non sta per "distribuzione di probabilità".

Variabili Aleatorie e Statistiche. Chiameremo *statistica* una collezione di dati. [9] Diversi concetti della teoria della v.a. sono analoghi ad altri concetti che abbiamo incontrato in statistica descrittiva, e ne condividono sia la denominazione che diverse proprietà; tra questi figurano la media, la varianza, la deviazione standard, la covarianza, il coefficiente di correlazione, la funzione di ripartizione, ecc.. Si noti anche l'ovvia analogia tra la frequenza relativa delle diverse modalità di una statistica e la densità di probabilità di una v.a.. Queste analogie non sono ... casuali: ogni v.a. $X = X(\omega)$ diventa ovviamente un valore numerico una volta che il caso [10] ha scelto l'*evento elementare* rappresentato dal punto $\omega \in \Omega$. Viceversa, per via del Teorema dei Grandi Numeri, quanto più il campione è ampio, tanto più le frequenze relative approssimano la distribuzione di probabilità della v.a. in esame.

Nondimeno certe sfumature lessicali riflettono i differenti punti di vista della statistica descrittiva e del calcolo delle probabilità. Ad esempio, una volta che i dati sono acquisiti, non ha molto senso usare i termini di speranza o valore atteso, ed è meglio parlare di media o di valor medio.

Variabili Aleatorie e loro Leggi. Si consideri la definizione (3.7): la *legge* (o *distribuzione di probabilità)* $\mathbf{P}_X$ è definita in termini della probabilità $\mathbf{P}$ e della v.a. X, ma né $\mathbf{P}$ né X possono essere ricostruite a partire da $\mathbf{P}_X$. In altri termini, passando da una v.a. alla sua legge ci può essere una perdita di informazione. [11] Ad esempio, se X è la v.a. che rappresenta le estrazioni successive delle biglie da un'urna senza reimmissione, le disposizioni degli elementi estratti dall'urna rappresentano una v.a. X, mentre le combinazioni degli stessi elementi rappresentano la sua legge $\mathbf{P}_X$. Ogni disposizione individua un'unica combinazione, ma non viceversa.

Riassumiamo alcuni punti che il Baldi espone più o meno esplicitamente.

Ogni v.a. determina la sua legge, ma non viceversa. Comunque ogni legge determina ed è determinata dalla sua densità, o equivalentemente dalla sua funzione di ripartizione. Legge, densità e funzione di ripartizione quindi contengono la stessa informazione. Questo vale sia per v.a. *reali* (ovvero scalari) che per v.a. *multidimensionali* (ovvero vettoriali), e sia per v.a. discrete che per v.a. continue.

La conoscenza della legge $\mathbf{P}_X$ può surrogare quella della X stessa solo in alcuni casi; è cruciale comprendere quando questo vale. Ad esempio:

[9] A dire il vero, questa non coincide con la definizione standard, ma a noi può bastare.

[10] o la dea bendata, o la iella, o il destino cinico e baro: non mancano immagini più o meno pittoresche del *caso,* che d'altra parte è una delle presenze più pervasive della realtà... Un famoso libro del 1970 del celebre biologo Jacques Monod si intitolava *il Caso e la Necessità,* ed interpretava i fenomeni biologici come (co-)stretti tra quello che succede per forza (la necessità delle leggi fisiche) e quello che non riusciamo a ridurre a necessità, e quindi attribuiamo al caso. Comunque, sia ben chiaro, questa è filosofia e non calcolo.

[11] In effetti una stessa legge si può incarnare (termine non tecnico!) in diverse v.a., che possono essere considerate come diverse *realizzazioni* (termine tecnico questo!) della stessa legge.

In diversi casi è più naturale pensare alla legge piuttosto che ad una v.a. ad essa associata. Ad esempio se gioco a testa o croce con una certa posta in gioco ad ogni lancio, mi interesserà sapere quante volte ho vinto, piuttosto che avere il dettaglio dell'esito dei singoli lanci. L'esito dettagliato dei lanci è una v.a. aleatoria, il numero di vittorie e la sua legge.

(i) La media, la varianza e gli altri momenti dipendono solo dalla legge delle v.a. interessate. Lo stesso vale per l'integrale di ogni funzione di una v.a.. La covarianza dipende solo dalla legge congiunta delle v.a. interessate. L'integrale di una funzione di più v.a. è determinato dalla legge congiunta (ovvero, la legge della v.a. congiunte); la conoscenza delle leggi marginali (ovvero, le leggi delle v.a. marginali) non è invece sufficiente a determinarlo, a meno che le v.a. non siano (stocasticamente) indipendenti.

(ii) Una famiglia $\{X_1, .., X_N\}$ di v.a. ovviamente determina la v.a. congiunta $X = (X_1, .., X_N)$, e quindi la legge congiunta $\mathbf{P}_X$. Quest'ultima determina le leggi marginali $\mathbf{P}_{X_1}, ..., \mathbf{P}_{X_N}$. Il viceversa in generale non vale, a meno che la famiglia $\{X_1, .., X_N\}$ non sia indipendente [B, p. 53].

(iii) L'indipendenza di una famiglia di v.a. è determinata dalla legge congiunta, ma non dalle leggi delle v.a. della famiglia. Infatti per stabilire l'eventuale indipendenza occorrono sia la legge congiunta che quelle marginali, e la legge congiunta determina le leggi marginali ma non viceversa. [12]

(iv) Sia $f : \mathbf{R} \to \mathbf{R}$ una funzione continua. La legge di $Y = f(X)$ può essere espressa in termini della legge $\mathbf{P}_X$ di X (o equivalentemente della sua densità p_X):

$$\mathbf{P}_{f(X)}(A) = \sum_{x \in f^{-1}(A)} \mathbf{P}_X(\{x\}) = \sum_{x \in f^{-1}(A)} p_X(x) \qquad \forall A \subset \mathbf{R}. \tag{3.8}$$

Quindi per la densità $p_{f(X)}$ abbiamo (banalmente)

$$p_{f(X)}(y) = \mathbf{P}_{f(X)}(\{y\}) = \sum_{x \in f^{-1}(y)} p_X(x) \qquad \forall y \in f(X(\Omega)) = X(f(\Omega)). \tag{3.9}$$

La legge $\mathbf{P}_{f(X)}$ è quindi determinata da $\mathbf{P}_X$. In altri termini, date due v.a. X_1 e X_2, se $\mathbf{P}_{X_1} = \mathbf{P}_{X_2}$ (ovvero se X_1 e X_2 sono *equidistribuite),* allora $\mathbf{P}_{f(X_1)} = \mathbf{P}_{f(X_2)}$. Questo è facilmente esteso a funzioni $f : \mathbf{R}^N \to \mathbf{R}^M$, ovvero funzioni di v.a. congiunte $X = (X_1, ..., X_N)$ che forniscono v.a. multidimensionali.

Esempi.

— (i) Siano X_1, X_2, Y_1, Y_2 quattro v.a., tali che le v.a. congiunte (X_1, X_2) e (Y_1, Y_2) siano equidistribuite. Se X_1 e X_2 sono indipendenti, allora pure Y_1 e Y_2 sono indipendenti. Questo consegue direttamente dalla definizione di indipendenza.

— (ii) Siano X_1, X_2, Y_1, Y_2 quattro v.a., tali che X_1 e Y_1 siano equidistribuite, e lo stesso valga per X_2 e Y_2. Questo non implica che le v.a. congiunte $X = (X_1, X_2)$ e $Y = (Y_1, Y_2)$ siano equidistribuite.

Ecco un controesempio. Si lanci due volte una moneta (equilibrata o meno), e si ponga:

$X_1 = 1$ se il primo lancio ha dato Testa, $X_1 = 0$ altrimenti,

$X_2 = 1$ se il secondo lancio ha dato Croce, $X_2 = 0$ altrimenti,

$Y_1 = X_1$,

$Y_2 = 1$ se il primo lancio ha dato Croce, $Y_2 = 0$ altrimenti.

Allora X_1 e Y_1 sono equidistribuite, e che lo stesso vale per X_2 e Y_2. (Se la moneta è equilibrata, addirittura tutte e quattro le v.a. sono equidistribuite.) Tuttavia le v.a. congiunte $X = (X_1, X_2)$ e $Y = (Y_1, Y_2)$ non sono equidistribuite. Ad esempio, se la moneta è equilibrata, la probabilità di avere $(X_1, X_2) = (1, 1)$ è 0.25, mentre l'evento $(Y_1, Y_2) = (1, 1)$ è impossibile.

Si osservi che le v.a. congiunte X e Y sono dipendenti. Ad esempio se $(X_1, X_2) = (1, 1)$ allora $(Y_1, Y_2) = (1, 0)$. È questo il motivo per cui X e Y non sono equidistribuite? No. Adesso lo vediamo.

[12] A proposito di indipendenza, il [B] manca di sottolineare quanto segue. Sia $\{X_1, .., X_N\}$ una famiglia di v.a. reali (discrete o continue), sia X la v.a. congiunta, e si definisca la *funzione di ripartizione multivariata* F_X:

$$F_X(x_1, ..., x_N) := \mathbf{P}(X_1 \leq x_1, ..., X_N \leq x_N) \qquad \forall (x_1, ..., x_N) \in \mathbf{R}^N.$$

La famiglia di v.a. $\{X_1, .., X_N\}$ è indipendente se e solo se, denotando con F_{X_i} le rispettive funzioni di ripartizione,

$$F_X(x_1, ..., x_N) = F_{X_1}(x_1) \cdots F_{X_N}(x_N) \qquad \forall (x_1, ..., x_N) \in \mathbf{R}^N.$$

Inoltre questo vale se e solo se, denotando con p_X e p_{X_i} le rispettive densità, $p_X = p_{X_1} \cdots p_{X_N}$.

— (iii) L'implicazione (ii) non vale nemmeno se X e Y sono indipendenti.

Ecco un controesempio. Si lanci tre volte una moneta (equilibrata o meno), e si ponga:

$X_1 = 1$ se il primo lancio ha dato Testa, $X_1 = 0$ altrimenti,

$X_2 = 1$ se il secondo lancio ha dato Croce, $X_2 = 0$ altrimenti,

$Y_1 = 1$ se il terzo lancio ha dato Testa, $Y_1 = 0$ altrimenti,

$Y_2 = 1$ se il terzo lancio ha dato Croce, $Y_2 = 0$ altrimenti.

Allora X_1 e Y_1 sono equidistribuite, e che lo stesso vale per X_2 e Y_2. (Se la moneta è equilibrata, addirittura tutte e quattro sono equidistribuite.) Tuttavia le v.a. congiunte $X = (X_1, X_2)$ e $Y = (Y_1, Y_2)$ non sono equidistribuite. Ad esempio, se la moneta è equilibrata, la probabilità di avere $(X_1, X_2) = (1, 1)$ è 0.25, mentre l'evento $(Y_1, Y_2) = (1, 1)$ è impossibile.

Qui le v.a. congiunte X e Y sono indipendenti, poiché X è determinata dai primi due lanci, Y dal terzo. (Invece Y_1 e Y_2 sono dipendenti, ma questo è irrilevante.)

Estrazioni Con o Senza Reimmissione. Si effettuino successive estrazioni da un'urna contenente N biglie distinte. Se le estrazioni sono con reimmissione, allora queste sono indipendenti, e ciascuna biglia ha probabilità $1/N$ di essere estratta. Se invece le estrazioni sono senza reimmissione, allora esse sono dipendenti; nondimeno ciascuna biglia ha ancora probabilità $1/N$ di essere estratta, cf. [B, p. 54].

Ricordiamo che, se le biglie sono di due colori, il numero di biglie di un colore estratte è rappresentato dalla distribuzione binomiale nel caso con reimmissione, da quella ipergeometrica nel caso senza reimmissione.

La Matrice di Covarianza. Sia N un qualsiasi intero ≥ 1, e $\{X_i\}_{i=1,...,N}$ una famiglia di v.a.. Sottintendendo che le somme sono da 1 a N, ed osservando che $\mathrm{Cov}(X_i, X_j) = \mathrm{Cov}(X_j, X_i)$ per ogni i, j, abbiamo

$$\begin{aligned} &\mathrm{Var}\,(\textstyle\sum_i X_i) = \mathrm{Cov}\,(\sum_i X_i, \sum_j X_j) = \sum_i \sum_j \mathrm{Cov}\,(X_i, X_j) \\ &= \textstyle\sum_{i=j} \mathrm{Cov}\,(X_i, X_j) + \sum_{i\neq j} \mathrm{Cov}\,(X_i, X_j) = \sum_i \mathrm{Var}\,(X_i) + 2\sum_{i<j} \mathrm{Cov}\,(X_i, X_j). \end{aligned} \tag{3.10}$$

Si definisce la *matrice di covarianza* (o matrice di varianza-covarianza, o matrice di dispersione) della famiglia di v.a. $\{X_i\}_{i=1,...,N}$

$$\mathrm{Cov}(X \cdot X^\tau) := \{\mathrm{Cov}(X_i, X_j)\}_{i,j=1,...,N}\,, \tag{3.11}$$

ove X^τ denota il vettore riga ottenuto per trasposizione del vettore colonna $X = (X_1, ..., X_N)$. Si può dimostrare che questa matrice è semidefinita positiva, ed è definita positiva se la famiglia di v.a. $\{X_i\}_{i=1,...,N}$ è linearmente indipendente. Essa è diagonale se e solo se queste v.a. sono non correlate. [13] In questo caso allora la varianza della somma è uguale alla somma delle varianze:

$$\mathrm{Var}\Big(\sum_i X_i\Big) = \sum_i \mathrm{Var}\,(X_i) \quad \text{per v.a. non correlate.} \tag{3.12}$$

Come noto, in particolare questo è verificato se le v.a. sono indipendenti.

Questo risultato è sorprendente, poiché una formula del genere vale per funzionali lineari (e.g. la speranza), mentre la varianza è quadratica! Come è evidente dalla dimostrazione, questo poggia sull'ipotesi di non correlazione. [14] Questi risultati si estendono al caso di una successione di v.a.,

[13] Si osservi che la non correlazione è quella che si potrebbe definire *indipendenza per coppie.* L'indipendenza vera è propria invece deve valere per coppie, per terne, ecc. [B, p. 53].

Si noti anche che la covarianza di due v.a. misura l'eventuale esistenza di una relazione lineare tra le due v.a.. Tuttavia essa potrebbe essere nulla anche in presenza di un legame non lineare tra le v.a..

[14] e ovviamente anche dalla definizione di varianza. Esaminiamo un momento questa definizione. Come noto, la varianza è una misura della dispersione intorno alla media; lo stesso si può dire della quantità $\mathbf{E}(|X - \mathbf{E}(X)|)$, ma la

sotto la condizione che tutte le quantità che compaiono in questa formula (le speranze, le covarianze, le serie) convergano.

Il Principio di Mutua Compensazione. Il Teorema dei Grandi Numeri poggia sul teorema di Chebyshev e su un risultato tanto semplice quanto potente, che potremmo definire come *il principio di mutua compensazione,* che ora illustriamo.

Si consideri una successione $\{X_i\}_{i\in\mathbf{N}}$ di v.a. equidistribuite (cioé aventi la stessa distribuzione di probabilità) e non correlate. Poiché $\mathrm{Var}(cY) = c^2\,\mathrm{Var}(Y)$ per ogni Y ed ogni $c \in \mathbf{R}$, e $\mathrm{Var}(X_i) = \mathrm{Var}(X_1)$ (per via della equidistribuzione), abbiamo

$$\mathrm{Var}\left(\frac{1}{N}\sum_i X_i\right) = \frac{1}{N^2}\,\mathrm{Var}\left(\sum_i X_i\right) \overset{(3.12)}{=} \frac{1}{N^2}\sum_i \mathrm{Var}\,(X_i) = \frac{1}{N}\,\mathrm{Var}\,(X_1). \tag{3.13}$$

Più in generale, si definisce *campione statistico (o aleatorio) di ampiezza N* una famiglia di N v.a. indipendenti equidistribuite. Quindi la *media campionaria* (o *media empirica)* di un campione statistico, ovvero la v.a. $\frac{1}{N}\sum_i X_i$, ha varianza attenuata di un fattore $1/N$ rispetto alle X_i. Questo si può interpretare osservando che mediando su più individui le oscillazioni (ovvero, le deviazioni dal valore atteso) delle diverse X_i in parte si compensano. [15] Il fatto che una popolazione [16] possa presentare una dispersione statistica minore di quella dei singoli è anche esperienza della vita di tutti i giorni. La teoria qui sviluppata ne fornisce una rappresentazione matematica, evidenziando le ipotesi essenziali: l'equidistribuzione e l'assenza di correlazione.

Nella pratica l'uso di un campione statistico più ampio richiede la raccolta di un maggior numero di dati, il che ha un costo in generale. Occorre quindi trovare un punto di equilibrio tra il beneficio della minore dispersione ed il prezzo dell'ulteriore acquisizione di dati. [17] La (3.12) fornisce una valutazione quantitativa della riduzione della dispersione che può essere utile per calcolare tale punto di equilibrio.

Il Teorema dei Grandi Numeri. Questo è uno dei principali risultati del corso; la sua prima formulazione è dovuta a Jakob Bernoulli, e risale al 1689, ovvero ai primordi del calcolo delle probabilità. Dimostrando che la media campionaria converge alla comune speranza delle v.a., questo teorema stabilisce un legame fondamentale tra l'impostazione assiomatica del calcolo delle probabilità e l'approccio frequentista. Si noti pure che esso fornisce una *stima dell'errore,* tramite la disuguaglianza di Chebyshev (si veda l'ultima riga di [B, p. 71]). [18] Questa stima è alquanto grossolana, dovendo valere

presenza del valore assoluto rende quest'ultima non derivabile, e quindi poco maneggevole. E' vero che il quadrato distorce un po' la dispersione, poiché

$$|X-\mathbf{E}(X)|^2 < |X-\mathbf{E}(X)| \ \text{ per } \ |X-\mathbf{E}(X)|<1, \qquad |X-\mathbf{E}(X)|^2 > |X-\mathbf{E}(X)| \ \text{ per } \ |X-\mathbf{E}(X)|>1;$$

però la funzione quadrato è derivabile. La definizione della varianza presenta anche altri vantaggi, tra i quali l'utile formula $\mathrm{Var}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2$. E comunque diversi risultati che vedremo dipendono in modo essenziale da questa definizione.

[15] Occorre prestare attenzione all'uso del termine *media,* (o *valor medio),* che è leggermente ambiguo. Se $\{X_i\}_{i=1,\dots,n}$ è un campione statistico, allora si possono effettuare due tipi di medie. Si può mediare rispetto a i, ottenendo la *media campionaria;* questa è la v.a. $\bar{X}_n := \frac{1}{n}\sum_i X_i$ (per ogni n), come l'abbiamo definita in statistica descrittiva. Oppure si può mediare rispetto a $\omega \in \Omega$, ottenendo il valor medio $\mathbf{E}(X_i)$ (lo stesso per ogni i, essendo le X_i equidistribuite). Per quest'ultimo parleremo piuttosto di speranza, e cercheremo di evitare i sinonimi valor medio o media.

[16] In statistica il termine *popolazione* è usato in senso molto esteso, e può anche essere riferito ad un insieme di oggetti simili tra di loro. I componenti di una popolazione sono allora detti *individui.*

[17] L'ampiezza del campione è un rilevante elemento per valutare la qualità di un'indagine statistica, ad esempio un sondaggio d'opinione.

[18] In analisi può essere utile sapere che una successione $\{x_n\}$ converge ad un certo valore x. Nelle applicazioni, in particolare per l'analisi numerica, spesso è molto utile maggiorare l'errore commesso sostituendo x con x_n, mediante una quantità che ovviamente dipenderà da n. Una simile maggiorazione è detta una *stima dell'errore.*

L'errore effettivamente commesso ovviamente dipende dai dati del problema. È impossibile darne una valutazione esatta (ove fosse possibile, mediante un'ovvia correzione del risultato si potrebbe eliminare l'errore!). Occorre quindi accontentarsi di una maggiorazione, che corrisponde all'ipotesi più pessimistica. Ovviamente si cerca di fornire una stima il più possibile stringente.

per ogni campione statistico. Per specifiche classi di leggi essa può essere raffinata: un esempio sarà fornito dal Teorema Limite Centrale. Entrambi sono *teoremi limite,* in quanto esprimono proprietà che valgono per campioni statistici infiniti — il che ovviamente richiede un passaggio al limite. [19]

Illustriamo brevemente la dimostrazione del teorema. Per ogni v.a. $Y \geq 0$ ed ogni $c > 0$, $c1_{\{Y \geq c\}} \leq Y$; [20] quindi

$$c\mathbf{P}(Y \geq c) = \mathbf{E}(c1_{\{Y \geq c\}}) \leq \mathbf{E}(Y) \qquad \text{(disuguaglianza di Markov).} \tag{3.14}$$

Per ogni v.a. X con varianza finita e speranza μ, applicando questa disuguaglianza a $Y = |X - \mu|^2$ otteniamo la disuguaglianza di Chebyshev:

$$\mathbf{P}(|X - \mu| \geq c) = \mathbf{P}(|X - \mu|^2 \geq c^2) \leq \frac{1}{c^2}\mathbf{E}(|X - \mu|^2) = \frac{1}{c^2}\operatorname{Var}(X). \tag{3.15}$$

Se $\{X_i\}_{i \in \mathbf{N}}$ è un campione statistico (di ampiezza infinita) di varianza finita e speranza μ, applicando quest'ultima disuguaglianza alla successione $\{\bar{X}_n\}$ e ricordando il principio di mutua compensazione, perveniamo alla *convergenza in probabilità* di $\bar{X}_n$ a μ:

$$\mathbf{P}(|\bar{X}_n - \mu| \geq c) \leq \frac{1}{c^2}\operatorname{Var}(\bar{X}_n) \overset{(3.13)}{=} \frac{1}{nc^2}\operatorname{Var}(X_1) \to 0 \qquad \text{per } n \to \infty, \forall c > 0; \tag{3.16}$$

oppure equivalentemente

$$\mathbf{P}(|\bar{X}_n - \mu| \leq c) = 1 - \mathbf{P}(|\bar{X}_n - \mu| \geq c) \geq 1 - \frac{1}{c^2}\operatorname{Var}(\bar{X}_n) \to 1 \qquad \text{per } n \to \infty, \forall c > 0. \tag{3.17}$$

Questo teorema permette di approssimare anche le funzioni del campione statistico $\{X_i\}$. Sia $f : \mathbf{R} \to \mathbf{R}$ una funzione continua tale che $Y_i := f(X_i)$ ha speranza finita $\tilde{\mu}$ (essendo il campione equidistribuito, questa speranza non dipende da i). Denotando con $\bar{Y}_n$ la media campionaria delle Y_i, allora la (3.16) fornisce

$$\mathbf{P}(|\bar{Y}_n - \tilde{\mu}| \geq c) \leq \frac{1}{c^2}\operatorname{Var}(\bar{Y}_n) \overset{(3.13)}{=} \frac{1}{nc^2}\operatorname{Var}(Y_1) \to 0 \qquad \text{per } n \to \infty, \forall c > 0. \tag{3.18}$$

In questo modo si possono approssimare ad esempio i momenti del campione statistico $\{X_i\}$.

Bernoullizzazione. Il teorema dei grandi numeri permette di approssimare non solo la speranza di una v.a., ma anche la legge di una qualsiasi v.a. discreta Y. A questo scopo basta *bernoullizzare* (il termine non è standard!) la v.a. ovvero trasformarla in una v.a. di Bernoulli, più precisamente in una funzione indicatrice; [B, p. 72]. [21] Si fissi un $y \in Y(\Omega)$, e si ponga $A = \{Y = y\}$. Sia $\{X_i\}_{i \in \mathbf{N}}$ un campione statistico avente la distribuzione della funzione indicatrice 1_A. Allora il teorema dei grandi numeri fornisce

$$p(y) = \mathbf{P}(\{Y = y\}) = \mathbf{E}(1_{\{Y=y\}}) = \lim_{n \to \infty} \bar{X}_n \quad \text{(nel senso della convergenza in probabilità).}$$

[19] La legge (o teorema) dei grandi numeri deve il suo nome al fatto che si basa su un passaggio al limite per $n \to \infty$: questi sono i grandi numeri. Lo stesso si può dire per gli altri teoremi limite, ma questo è stato il primo ad essere scoperto e in certo senso si è accaparrato il nome.

[20] Per ogni evento $A \subset \Omega$, 1_A è la *funzione indicatrice* di A, ovvero

$$1_A(\omega) = 1 \quad \forall \omega \in A, \qquad 1_A(\omega) = 0 \quad \forall \omega \in \Omega \setminus A;$$

la densità p della legge di 1_A quindi vale

$$p(1) = \mathbf{P}(A), \qquad p(0) = 1 - \mathbf{P}(A), \qquad p(y) = 0 \quad \forall y \in \mathbf{R} \setminus \{0, 1\}.$$

Si noti che $\mathbf{E}(1_A) = \mathbf{P}(A)$. In altri termini, si può rappresentare la probabilità di ogni evento come la speranza di una v.a., la funzione indicatrice di quell'evento appunto.

[21] Questo risponde alla domanda: come si può trasformare un dado in una moneta? (senza venderla naturalmente...)

Applicando questo procedimento ad ogni $y \in Y(\Omega)$, si può approssimare l'intera densità di probabilità di Y, ovvero la sua legge.

Random Walk. Ovvero passeggiata aleatoria, chiamata anche passeggiata dell'ubriaco. Per semplicità, supponiamo che il moto avvenga lungo una retta. Simuliamo questo comportamento mediante il ripetuto lancio di una moneta equilibrata: un passo in avanti se viene testa, un passo indietro se viene croce. Il lancio i-esimo può essere pertanto rappresentato con una v.a. X_i di Bernoulli: per $X_i = 1$ si va avanti, per $X_i = -1$ si va indietro, ciascun esito con probabilità $1/2$. I lanci sono supposti indipendenti, cosicché le X_i costituiscono un campione statistico.

Calcoliamo alcuni momenti della media campionaria $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$ a titolo di esercizio. Ovviamente

$$\mathbf{E}(\bar{X}_n) = \tfrac{1}{n}\textstyle\sum_i \mathbf{E}(X_i) = 0. \tag{3.19}$$

Inoltre, essendo $\mathbf{E}(X_i^2) = \mathbf{E}(1) = 1$ per ogni i, e per via dell'indipendenza (qui basterebbe la non correlazione)

$$\mathbf{E}(X_i \cdot X_j) = \mathbf{E}(X_i) \cdot \mathbf{E}(X_j) = 0 \qquad \text{se } i \neq j,$$

abbiamo

$$\begin{aligned} \mathbf{E}[(\bar{X}_n)^2] &= \tfrac{1}{n^2}\mathbf{E}(\textstyle\sum_i X_i \cdot \sum_j X_j) = \tfrac{1}{n^2}\sum_{i,j}\mathbf{E}(X_i \cdot X_j) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=j}\mathbf{E}(X_i \cdot X_j) + \tfrac{1}{n^2}\sum_{i\neq j}\mathbf{E}(X_i \cdot X_j) = \tfrac{n}{n^2} + 0 = \tfrac{1}{n}\ . \end{aligned} \tag{3.20}$$

Si verifica facilmente che, per via dell'indipendenza (qui la non correlazione non basta), $\mathbf{E}(X_i \cdot X_j \cdot X_k) = 0$ per ogni i, j, k. Pertanto

$$\mathbf{E}[(\bar{X}_n)^3] = \tfrac{1}{n^3}\mathbf{E}(\textstyle\sum_i X_i \cdot \sum_j X_j \cdot \sum_k X_k) = \tfrac{1}{n^3}\sum_{i,j,k}\mathbf{E}(X_i \cdot X_j \cdot X_k) = 0, \tag{3.21}$$

e lo stesso vale per ogni momento dispari. Avendo ormai compreso quali termini sono nulli, poi abbiamo

$$\begin{aligned} \mathbf{E}[(\bar{X}_n)^4] &= \tfrac{1}{n^4}\mathbf{E}(\textstyle\sum_i X_i \cdot \sum_j X_j \cdot \sum_k X_k \cdot \sum_\ell X_\ell) \\ &= \tfrac{1}{n^4}\textstyle\sum_{i,j,k,\ell}\mathbf{E}(X_i \cdot X_j \cdot X_k \cdot X_\ell) = \tfrac{3}{n^4}\sum_{i,k}\mathbf{E}(X_i^2 \cdot X_k^2) \\ &= \tfrac{3}{n^4}\textstyle\sum_{i=k}\mathbf{E}(X_i^2 \cdot X_k^2) + \tfrac{3}{n^4}\sum_{i\neq k}\mathbf{E}(X_i^2 \cdot X_k^2) = \tfrac{3n}{n^4} + \tfrac{3n(n-1)}{n^4} = \tfrac{3}{n^2}\ . \end{aligned} \tag{3.22}$$

Ci arrestiamo qui con il calcolo dei momenti si $\bar{X}_n$.

I più rilevanti risultati ottenuti riguardano la speranza e la varianza:

$$\mathbf{E}(\bar{X}_n) \overset{(3.19)}{=} 0, \qquad \mathrm{Var}(\bar{X}_n) \overset{(3.19)}{=} \mathbf{E}[(\bar{X}_n)^2] \overset{(3.12)}{=} 1/n\ . \tag{3.23}$$

Confrontando quest'ultima uguaglianza con la (3.13), vediamo che questa è una manifestazione del principio di mutua compensazione — d'altra parte la (3.13) è stata dimostrata riproducendo la procedura usata per derivare quel principio.

Moto Browniano. Il random walk è alla base di un importante modello fisico, noto come *moto Browniano,* che rappresenta fenomeni di diffusione, ad esempio la dispersione di una sostanza in un fluido, o la propagazione del calore. Se denotiamo con h la lunghezza di un passo e con τ l'unità di tempo, allora $h\sum_{i=1}^n X_i = nh\bar{X}_n$ rappresenta l'ascissa raggiunta dopo l'n-esimo lancio, ovvero all'istante $t = n\tau$. Poniamo

$$\delta(t)^2 = \mathbf{E}[(h\textstyle\sum_{i=1}^n X_i)^2]\text{: media del quadrato della distanza dall'origine all'istante } t,$$

cosicché $\delta(t)$ rappresenta la *distanza quadratica media* dall'origine all'istante t, ovvero la deviazione standard della distanza (che è una v.a. centrata). Abbiamo

$$\delta(t)^2 = \mathbf{E}[(nh\bar{X}_n)^2] \overset{(3.19)}{=} (nh)^2\,\mathrm{Var}(\bar{X}_n) \overset{(3.23)}{=} nh^2 = Dt \quad (\text{ponendo } D := h^2/\tau); \tag{3.24}$$

D è il *coefficiente di diffusione.* Nei fenomeni di diffusione quindi la *distanza quadratica media* è proporzionale alla radice quadrata del tempo: $\delta(t) = \sqrt{Dt}$, mentre per i fenomeni di trasporto la distanza media è proporzionale al tempo. Questo in virtù del principio di mutua compensazione, come già osservato.

4 Variabili aleatorie continue

Legge e Densità di Probabilità. Supponiamo che $\mathbf{P}$ sia una misura di probabilità su uno spazio campionario Ω. Per ogni v.a. $X : \Omega \to \mathbf{R}$ definiamo la funzione di ripartizione

$$F_X : \mathbf{R} \to [0,1], \qquad F_X(y) = \mathbf{P}(X \leq y) \quad \forall y \in \mathbf{R}. \tag{4.1}$$

Si verifica facilmente che $F_X(-\infty) = 0$ e, essendo $\mathbf{P}$ una misura di probabilità, $F_X(+\infty) = 1$ e F è non decrescente. Distinguiamo a seconda che sia X continua o discontinua.

(a) Se F_X è continua, possiamo supporre che sia derivabile ovunque, salvo in un insieme H_X formato al più da una successione di punti. [22] In tal caso $f_X := F_X'$ (detta *densità di probabilità)* è definita solo in $\mathbf{R} \setminus H_X$; ma questo non causa particolari difficoltà, e comunque non ci impedisce di scrivere

$$F_X(y) = \int_{-\infty}^{y} f_X(s)\, ds \qquad \forall s \in \mathbf{R}. \tag{4.2}$$

Quindi, ad esempio,

$$\mathbf{P}(a < X \leq b) = F_X(b) - F_X(a) = \int_a^b f_X(s)\, ds \qquad \forall]a,b[\subset \mathbf{R}, \tag{4.3}$$

da cui (un po' disinvoltamente)

$$\mathbf{P}(x < X \leq x + dx) \simeq f_X(x)\, dx \qquad \forall x \in \mathbf{R}.$$

Si noti che $\mathbf{P}(X = x) = 0$ per ogni $x \in \mathbf{R}$. Quindi $f_X(x)$ non è la probabilità dell'evento $\{X = x\}$, a differenza di quanto visto per la funzione di densità di una v.a. discreta.

(b) Ben diversa è la situazione in cui F_X sia discontinua, ovvero abbia dei salti. Il caso limite in cui la F_X cresce solo a salti corrisponde esattamente ad una v.a. X discreta, che è stato già trattato. Il caso intermedio in cui F_X cresca sia a salti che aldifuori dei salti non rientra nella trattazione del [B], in quanto richiede degli strumenti analitici più raffinati.

Un'altra Rappresentazione della Speranza. La speranza di una v.a. X è definita dal [B] in termini della legge della v.a. stessa. Se $\Omega \subset \mathbf{R}^N$ possiamo dare una definizione equivalente che sfrutta la teoria dell'integrazione su $\mathbf{R}^N$. Per semplicità qui ci limitiamo a $N = 2$, ma l'estensione ad un generico N non presenta difficoltà. In questo caso il generico $\omega \in \Omega$ è della forma $\omega = (\omega_1, \omega_2)$, e possiamo usare l'integrale bidimensionale, che indichiamo con $\iint ...d\omega_1 d\omega_2$. Si assuma che esista una funzione $h : \Omega \to \mathbf{R}$ tale che

$$h \geq 0, \qquad \iint_\Omega h(\omega)\, d\omega_1 d\omega_2 = 1, \qquad \mathbf{P}(A) = \iint_A h(\omega)\, d\omega_1 d\omega_2 \qquad \forall A \subset \Omega. \tag{4.4}$$

Quindi h è la densità della misura di probabilità $\mathbf{P}$ su Ω. [23] Allora la speranza di una v.a. (discreta o continua) $X : \Omega \to \mathbf{R}$ è

$$\mathbf{E}(X) = \iint_\Omega X(\omega) h(\omega)\, d\omega_1 d\omega_2, \quad \text{se} \quad \iint_\Omega |X(\omega)| h(\omega)\, d\omega_1 d\omega_2 < +\infty. \tag{4.5}$$

[22] Questo non è del tutto vero: ci sono dei controesempi. Tuttavia possiamo ignorare questi casi, che sono estremamente patologici. Il [B] invece si pone qualche scrupolo in più.

[23] Occorre prestare attenzione a distinguere la densità della misura di probabilità su Ω, dalla densità della misura di probabilità su $\mathbf{R}$ indotta da una v.a. X. Il termine è lo stesso, ma la prima è riferita alla probabilità $\mathbf{P}$ su Ω; la seconda alla probabilità $\mathbf{P}_X$ su $\mathbf{R}$, ovvero alla legge di X.

Questo rende conto del fatto che diverse proprietà della speranza riproducono quelle dell'integrale. Ad esempio, sotto opportune restrizioni,

$$\mathbf{E}(X_1 + X_2) = \mathbf{E}(X_1) + \mathbf{E}(X_2), \qquad \mathbf{E}(cX) = c\mathbf{E}(X) \qquad \forall c \in \mathbf{R}.$$

Se X è discreta e p_X è la sua densità, possiamo confrontare la (4.5) con la nota definizione

$$\mathbf{E}(X) = \sum_{x_i \in X(\Omega)} x_i p_X(x_i), \quad \text{se} \quad \sum_{x_i \in X(\Omega)} |x_i| p_X(x_i) < +\infty. \tag{4.6}$$

D'altra parte, denotando con F_X la funzione di ripartizione di X e supponendo che questa sia continua (e derivabile salvo al più in una successione di punti),

$$\mathbf{E}(X) = \int_{\mathbf{R}} x\, F_X'(x)\, dx, \quad \text{se} \quad \int_{\mathbf{R}} |x|\, F_X'(x)\, dx < +\infty. \tag{4.7}$$

Confrontando la (4.5) con la (4.7) ritroviamo i due diversi modi di sommare che abbiamo già incontrato in statistica descrittiva e nell'integrazione delle v.a. discrete; si vedano gli sviluppi di (3.1),...,(3.5).

La definizione (4.7) è posta solo per v.a. continue; è noto che per v.a. discrete la speranza va scritta diversamente, con una serie invece di un integrale. Per tali v.a. la (4.7) è più generale della (4.5), che si applica solo se $\Omega \subset \mathbf{R}^N$ (qui abbiamo posto $N = 2$).

Supponiamo ora di avere due v.a. $X, Y : \Omega \to \mathbf{R}$, ciascuna discreta o continua. La loro covarianza vale

$$\begin{aligned} &\mathrm{Cov}(X,Y) = \iint_\Omega [X(\omega) - \mathbf{E}(X)]\, [Y(\omega) - \mathbf{E}(X)]\, h(\omega)\, d\omega_1 d\omega_2, \\ &\text{se} \quad E(X), \mathbf{E}(Y), \iint_\Omega |[X(\omega) - \mathbf{E}(X)]\, [Y(\omega) - \mathbf{E}(X)]|\, h(\omega)\, d\omega_1 d\omega_2 < +\infty. \end{aligned} \tag{4.8}$$

Se X e Y sono v.a. discrete e $p_{X,Y}$ è la loro densità congiunta, allora

$$\begin{aligned} &\mathrm{Cov}(X,Y) = \sum_{x_i \in X(\Omega)} \sum_{y_j \in Y(\Omega)} [x_i - \mathbf{E}(X)]\, [y_j - \mathbf{E}(Y)]\, p_{X,Y}(x_i, y_j) \\ &\text{se} \quad E(X), \mathbf{E}(Y), \sum_{x_i \in X(\Omega)} \sum_{y_j \in Y(\Omega)} |[x_i - \mathbf{E}(X)]\, [y_j - \mathbf{E}(Y)]|\, p_{X,Y}(x_i, y_j) < +\infty. \end{aligned} \tag{4.9}$$

Se X e Y sono indipendenti e le loro densità sono rispettivamente p_X e p_Y, allora $p_{X,Y}(x_i, y_j) = p_X(x_i) p_Y(y_j)$, e ritroviamo la nota proprietà che $\mathrm{Cov}(X,Y) = 0$.

Tiro al Bersaglio. Il bersaglio sia un tabellone circolare Ω di raggio R. Se si tira a casaccio, la densità di probabilità $h(\omega)$ sarà uniforme:

$$h(\omega) = \tilde{h}(\rho) = (\pi R^2)^{-1} \qquad \forall \rho = \sqrt{\omega_1^2 + \omega_2^2}.$$

Se invece si mira al centro e si ha una buona mira, allora la densità di probabilità sarà una funzione decrescente della ρ:

$$h(\omega) = \tilde{h}(\rho) \qquad \text{con } \tilde{h} : [0, R] \to \mathbf{R}^+ \text{ decrescente},$$

e h dovrà soddisfare la condizione di normalizzazione [24]

$$\iint_\Omega h(\omega)\, d\omega_1 d\omega_2 = 2\pi \int_0^R \tilde{h}(\rho) \rho\, d\rho = 1. \tag{4.10}$$

[24] Per un bersaglio di raggio infinito (ovvero $\Omega = \mathbf{R}^2$), per ogni $\sigma > 0$ si può usare la densità di probabilità

$$\tilde{h}(\rho) = \frac{1}{2\pi\sigma} \exp\big(\frac{-\rho^2}{2\sigma}\big) \qquad \forall \rho > 0,$$

che è parente stretta della nota densità gaussiana. Si verifichi che la condizione (4.10) è soddisfatta anche in questo caso.

Si noti che $\tilde{h}$ potrebbe divergere per $\rho \to 0$.

La funzione $\tilde{h}$ può essere definita la *funzione di mira,* e dipenderà dal tiratore. La probabilità di colpire un punto di un insieme $A \subset \Omega$ è allora pari a

$$\mathbf{P}(A) = \mathbf{E}(1_A) = \iint_\Omega 1_A(\omega) h(\omega)\, d\omega_1 d\omega_2 = \iint_A h(\omega)\, d\omega_1 d\omega_2. \tag{4.11}$$

Se A è radiale, ovvero è un cerchio di centro $(0,0)$ e raggio $0 \le r \le 1/R$, allora

$$\mathbf{P}(A) = \mathbf{E}(1_A) = \iint_A h(\omega)\, d\omega_1 d\omega_2 = 2\pi \int_0^r \tilde{h}(\rho)\rho\, d\rho.$$

Supponiamo che il tiro venga ricompensato con un premio $g(\omega)$ (≥ 0), che dipende dal punto colpito (oppure dal centro della regione circolare colpita, se non si vuole pensare la freccetta puntiforme); g è quindi una v.a.. Allora il guadagno atteso dal tiratore con funzione di mira h è

$$\mathbf{E}(g) = \iint_\Omega g(\omega) h(\omega)\, d\omega_1 d\omega_2, \tag{4.12}$$

se questo integrale è finito. Se la funzione premio g dipende dalla distanza dal centro, ovvero se

$$g(\omega) = \tilde{g}(\rho) \qquad \text{con } \tilde{g} : [0, R] \to \mathbf{R}^+,$$

allora il guadagno atteso è

$$\mathbf{E}(g) = \iint_\Omega g(\omega) h(\omega)\, d\omega_1 d\omega_2 = 2\pi \int_0^{1/R} \tilde{g}(\rho)\tilde{h}(\rho)\rho\, d\rho.$$

La varianza della v.a. g è

$$\mathrm{Var}(g) = \iint_\Omega g(\omega)^2 h(\omega)\, d\omega_1 d\omega_2 - \mathbf{E}(g)^2 = 2\pi \int_0^{1/R} \tilde{g}(\rho)^2 \tilde{h}(\rho)\rho\, d\rho - \mathbf{E}(g)^2. \tag{4.13}$$

La funzione di ripartizione di g è

$$F_g(t) = \mathbf{P}(g \le t) = \iint_{\{g \le t\}} h(\omega)\, d\omega_1 d\omega_2 \qquad \forall t \ge 0,$$

ed il suo calcolo richiede la determinazione dell'insieme $\{\omega \in \Omega : g(\omega) \le t\}$ per ogni $t \ge 0$.

Si noti che le funzioni h e g hanno ruoli del tutto diversi; inoltre h ha media 1, a differenza di g.

Il Metodo di Montecarlo. Il modello del tiro al bersaglio può anche essere usato ... all'inverso. La determinazione della probabilità di un evento A (ovvero di colpire un punto di un insieme A) richiede il calcolo dell'integrale della funzione densità. D'altra parte $\mathbf{P}(A) = \mathbf{E}(1_A)$, e questa speranza può essere approssimata mediante la legge dei grandi numeri. Se si individua un campione statistico con legge di 1_A, allora effettuando un gran numero di volte quell'esperimento si può calcolare in modo approssimato $\mathbf{P}(A) = \mathbf{E}(1_A)$; questo permette quindi di approssimare l'integrale della densità. In diversi casi questi esperimenti possono essere effettuati al calcolatore.

Questa procedura per il calcolo approssimato degli integrali è detto *metodo di Montecarlo* (per via del noto casinò).

Il Paradosso del Tiro al Segno. Dalla (4.11) consegue che

$$\mathbf{P}(\{(x,y)\}) = 0 \qquad \forall (x,y) \in \Omega, \forall \text{ tiratore}.$$

Quindi tutti i tiratori hanno la stessa probabilità di colpire il centro (!), poiché per tutti la probabilità è nulla. Per lo stesso motivo, ciascun tiratore ha la stessa probabilità di colpire il centro o qualsiasi altro punto prefissato. Se invece di considerare punti si considerano piccoli cerchi, allora le cose stanno diversamente — sempre che, come finora supposto, la densità di probabilità (la densità di probabilità, non la probabilità!) dipenda dal punto e dal tiratore. Questo risolve quello che poteva sembrare un paradosso.

5 Teoremi limite

Diverse Nozioni di Convergenza. Per le successione di numeri esiste un solo concetto di convergenza; non è così per le successioni di funzioni, in particolare per quelle di v.a..

Sia $\mathbf{P}$ una misura di probabilità su un insieme Ω, e sia $\{X_n\}$ una successione di v.a. $\Omega \to \mathbf{R}$, discrete o continue. Tra le altre, si definiscono le seguenti nozioni di convergenza:

$$X_n \to X \text{ in probabilità} \quad \Leftrightarrow \quad \mathbf{P}(|X_n - X| \geq c) \to 0 \quad \text{per } n \to \infty, \forall c > 0, \tag{5.1}$$

$$X_n \to X \text{ in legge} \quad \Leftrightarrow \quad \begin{cases} F_{X_n}(t) \to F_X(t) \quad \text{per } n \to \infty, \\ \forall t \in \mathbf{R} \text{ in cui } F_X \text{ è continua.} \end{cases} \tag{5.2}$$

Ad esempio la convergenza in probabilità compare nel teorema dei grandi numeri, e quella in legge nel teorema limite centrale. Si noti che la convergenza in probabilità coinvolge le v.a., mentre quella in legge dipende solo dalle leggi delle X_n e di X. Sussiste un importante legame tra questi due concetti:

$$\text{la convergenza in probabilità implica quella in legge, ma non viceversa.} \tag{5.3}$$

Teoremi Limite. Nel corso abbiamo considerati quattro teoremi limite, ovvero teoremi che esprimono la convergenza di una successione di v.a. o leggi, secondo una delle nozioni di convergenza appena definite:

(i) *Legge binomiale $\Rightarrow$ legge di Poisson.* Sia λ una costante > 0. Per $n \to \infty$, la distribuzione binomiale $B(n, \lambda/n)$ converge alla distribuzione di Poisson $Poi(\lambda)$ [B, p. 48]. [25] Questo significa che uno schema di Bernoulli in cui un evento di probabilità p molto piccola viene ripetuto un gran numero n di volte può essere approssimato da un processo di Poisson con parametro $\lambda = np$. [26]

(ii) *Legge geometrica $\Rightarrow$ legge esponenziale.* Sia λ una costante > 0. Per $n \to \infty$, la distribuzione geometrica $G(\lambda/n)$ converge alla distribuzione esponenziale $Exp(\lambda)$, riscalando opportunamente il tempo — si veda più avanti.

(iii) *Teorema dei Grandi Numeri (o TGN).* (Jakob Bernoulli, 1689) Se $\{X_i\}_{i \in \mathbf{N}}$ è un campione statistico di v.a. con speranza μ e varianza finita, allora la media campionaria $\bar{X}_n$ converge in probabilità a μ [B, p. 71]. Lo stesso risultato vale anche per le v.a. continue, con la stessa dimostrazione. Come abbiamo visto, questo giustifica il punto di vista frequentista.

(iv) *Teorema Limite Centrale (o TLC).* (De Moivre, 1733; Lindeberg 1922) [27] Se $\{X_i\}_{i \in \mathbf{N}}$ è un campione statistico di v.a. con speranza μ e varianza finita $\sigma^2 > 0$, [28] allora la distribuzione della media campionaria standardizzata converge alla distribuzione normale standard [B, p. 124], ovvero

$$S_n^* := \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to Y \quad \text{in legge,} \quad \text{con } Y \sim N(0,1). \tag{5.4}$$

Pertanto la distribuzione standardizzata di un campione statistico di sufficiente ampiezza può essere approssimata in legge da una distribuzione gaussiana; e questo avviene indipendentemente dalla distribuzione del campione statistico (!).

---

[25] Poiché si parla della convergenza di distribuzioni (piuttosto che di v.a.), si tratta necessariamente di una convergenza in legge.

[26] Per questo motivo la legge di Poisson è detta la *legge degli eventi rari,* o anche la *legge dei piccoli numeri* — qui intendendo il termine *legge* in un senso del tutto diverso da quello della "legge" dei grandi numeri.

[27] De Moivre dimostrò questo risultato per v.a. binomiali $X_i \sim B(1,p)$. Lindeberg lo estese poi alla forma riportata da [B, p. 124]. (Pur se apparentemente più modesto, il passo più importante fu quello di De Moivre, già nel 1733!)

Perché il teorema si chiama così? Su questo non sono tutti d'accordo: i più attribuiscono l'aggettivo *centrale* a *teorema,* per il suo ruolo centrale nel calcolo delle probabilità ed in statistica; per altri è il limite ad essere centrale, e parlano di *teorema centrale del limite.* Tra l'altro, la denominazione inglese *central limit theorem* si presta ad entrambe le interpretazioni.

[28] Se invece $\sigma^2 = 0$ allora ... ancora meglio: in tal caso $X_i = \mu$ in tutto Ω, quindi $\bar{X}_n = \mu$ e per ogni n.

Osservazioni sul TLC. (i) Per il principio di mutua compensazione, $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_i)/n$ per ogni i. Quindi $\sigma_{\bar{X}_n} = \sigma/\sqrt{n}$, ovvero il denominatore della S_n^* è la deviazione standard del numeratore:

$$S_n^* = \frac{\bar{X}_n - \mu}{\sigma_{\bar{X}_n}} \, .$$

Questa è la v.a. standardizzata della media campionaria, ed è diversa dalla media campionaria delle v.a. X_i standardizzate (perché?):

$$\text{standardizzata della media } = \frac{\frac{1}{n}\sum_{i=1}^n X_i - \mu}{\sigma/\sqrt{n}} \neq \frac{1}{n}\sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \text{ media delle standardizzate.}$$

(ii) La (5.4) può essere riscritta

$$\sqrt{n}(\bar{X}_n - \mu) \simeq \sigma Y \sim N(0, \sigma^2) \qquad \text{oppure} \qquad \bar{X}_n \simeq \mu + \frac{\sigma}{\sqrt{n}} Y \sim N(\mu, \sigma^2/n), \tag{5.5}$$

per $n \to \infty$. Qui "$\simeq$" corrisponde all'approssimazione nel senso della convergenza in legge per $n \to \infty$, mentre "$Z \sim N(\mu, \sigma^2)$" significa che la v.a. ha distribuzione normale di speranza μ e varianza σ^2. Quindi quanto più grande è n, tanto più la legge della v.a. $\bar{X}_n$ è vicina alla legge normale di media μ e varianza σ^2/n, e tanto più la varianza di quest'ultima legge è piccola. Questo è coerente con due note proprietà:

(a) la speranza di una somma di v.a. è la somma delle speranze;

(b) la varianza di una somma di v.a. non correlate è la somma delle varianze; quindi, essendo la varianza quadratica, la varianza della loro media è la media delle varianze divisa per il numero delle v.a..

(iii) La seconda formula della (5.5) può anche essere interpretata come segue: per n "abbastanza" grande, la legge della v.a. $\bar{X}_n$ finisce con dipendere dalle variabili X_i "essenzialmente" solo attraverso la loro media μ e la loro varianza σ^2/n. [29] Più precisamente, quanto più n è grande, tanto più questo è vero.

(iv) Se ciascuna v.a. X_i del campione statistico ha legge normale $N(\mu, \sigma^2)$, allora si può dimostrare che la (5.5) è verificata esattamente (senza bisogno di approssimare), ovvero che $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

Relazione tra TLC e TGN. Il TLC spiega il ruolo fondamentale della distribuzione normale in statistica.

(i) TLC $\Rightarrow$ TGN. [30] Sia $\{X_i\}_{i \in \mathbf{N}}$ un campione statistico di v.a. con speranza μ e varianza finita $\sigma^2 > 0$, e sia Y una v.a. normale standard. Fissiamo una qualsiasi costante $c > 0$, ed osserviamo che, definendo S_n^* come in (5.4),

$$\begin{aligned} &\mathbf{P}(|\bar{X}_n - \mu| \le c) = \mathbf{P}\Big(|S_n^*| \le \frac{\sqrt{n}\, c}{\sigma}\Big) = \mathbf{P}\Big(-\frac{\sqrt{n}\, c}{\sigma} \le S_n^* \le \frac{\sqrt{n}\, c}{\sigma}\Big) \\ &= F_{S_n^*}\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) - F_{S_n^*}\Big(-\frac{\sqrt{n}\, c}{\sigma}\Big). \end{aligned} \tag{5.6}$$

Per via della (5.4), $F_{S_n^*}(t) \to F_Y(t)$ per $n \to \infty$, per ogni $t \in \mathbf{R}$. Poiché $\frac{\sqrt{n}\, c}{\sigma} \to +\infty$ per $n \to \infty$, è allora facile rendersi conto che $F_{S_n^*}(\frac{\sqrt{n}\, c}{\sigma}) \to F_Y(+\infty)$. Analogamente $F_{S_n^*}(-\frac{\sqrt{n}\, c}{\sigma}) \to F_Y(-\infty)$. Pertanto

$$\mathbf{P}(|\bar{X}_n - \mu| \le c) = F_{S_n^*}\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) - F_{S_n^*}\Big(-\frac{\sqrt{n}\, c}{\sigma}\Big) \to F_Y(+\infty) - F_Y(-\infty) = 1. \tag{5.7}$$

[29] Queste virgolette indicano delle espressioni che andrebbero precisate.

[30] A prima vista questo può apparire un po' sorprendente, poiché il TLC fornisce una convergenza in legge, il TGN una convergenza in probabilità; e in (5.3) abbiamo visto che la convergenza in legge non implica quella in probabilità.

Quindi

$$\mathbf{P}(|\bar{X}_n - \mu| \geq c) = 1 - \mathbf{P}(|\bar{X}_n - \mu| \leq c) \to 0 \qquad \forall c > 0, \text{ per } n \to \infty, \tag{5.8}$$

ovvero $\bar{X}_n \to \mu$ in probabilità, come affermato dal TGN, cf. (3.16).

(ii) Il TLC fornisce anche un'informazione più precisa del TGN circa la velocità con cui $\bar{X}_n \to \mu$. Si consideri la definizione di S_n^* (ovvero l'uguaglianza della (5.4)): per $n \to \infty$, questa è una forma indeterminata del tipo $\frac{0}{0}$. Il TLC sostanzialmente afferma che

$$\bar{X}_n - \mu \simeq \frac{\sigma}{\sqrt{n}} Y \quad \text{ con } Y \sim N(0,1), \text{ per } n \to \infty, \tag{5.9}$$

nel senso della convergenza in legge. Grazie alla (5.4), per n abbastanza grande quindi abbiamo

$$\mathbf{P}(|\bar{X}_n - \mu| \leq c) = F_{|S_n^*|}\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) \simeq F_{|Y|}\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) = 2F_Y\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) - 1 \qquad \forall c > 0; \tag{5.10}$$

ovvero

$$\mathbf{P}(|\bar{X}_n - \mu| \geq c) = 1 - F_{|S_n^*|}\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) \simeq 2 - 2F_Y\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) \qquad \forall c > 0, \tag{5.11}$$

F_Y è la funzione di ripartizione della distribuzione normale standard, che si trova tabulata.

Quest'ultima formula può essere confrontata con la (3.16):

$$\mathbf{P}(|\bar{X}_n - \mu| \geq c) \leq \frac{n\sigma^2}{c^2} \qquad \forall c > 0.$$

(iii) Resta da valutare la velocità con cui $F_{S_n^*} \to F_Y$. Si può dimostrare (teorema di Berry-Esseen) che, se $\mathbf{E}(|X_i|^3) < +\infty$, allora esiste una costante $C > 0$ tale che

$$\sup_{y \in \mathbf{R}} \Big|\mathbf{P}\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq y\Big) - \mathbf{P}(Y \leq y)\Big| \leq C \frac{\mathbf{E}(|X_i|^3)}{\sigma^3 \sqrt{n}}. \tag{5.12}$$

Si noti che, dato un campione statistico, il TGN fornisce una successione (la $\bar{X}_n$) che converge in probabilità alla speranza μ; questo può essere considerato come uno sviluppo arrestato al momento del primo ordine: la speranza, appunto. La (3.16) maggiora l'errore di quest'ultimo sviluppo mediante il momento del secondo ordine: la varianza σ^2.

Analogamente, il TLC esibisce uno sviluppo fino al momento del secondo ordine, che compare attraverso la σ; si veda la (5.5) che converge in legge. Il teorema di Berry-Esseen maggiora poi l'errore di quest'ultimo sviluppo mediante il momento del terzo ordine, $\mathbf{E}(|X_i|^3)$. Si può notare l'analogia con lo sviluppo di Taylor.

Sull'Uso del TLC. Fissata la legge del campione statistico, se $\mathbf{E}(|X_i|^3) < +\infty$ allora per ogni $\epsilon > 0$ esiste un N tale che

$$\sup_{y \in \mathbf{R}} \Big|\mathbf{P}\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \leq y\Big) - \mathbf{P}(Y \leq y)\Big| \leq \epsilon \qquad \forall n \geq N. \tag{5.13}$$

Infatti basta scegliere $N = C^2\mathbf{E}(|X_i|^3)^2/\sigma^6\epsilon^2$ e poi applicare la (5.12). Poiché $\mathbf{E}(|X_i|^3)^2$ e σ sono determinati dalla legge del campione statistico, per ogni legge e per ogni tolleranza $\epsilon > 0$, resta così individuato un N che soddisfa la (5.13). Sottolineiamo che non è possibile indicare un N che valga per ogni campione e per ogni $\epsilon > 0$. [31]

Legge Geometrica ⇒ Legge Esponenziale. Dimostriamo questo teorema limite. Una v.a. X abbia distribuzione geometrica $G(p)$ con $0 < p \ll 1$; [32] quindi $\mathbf{P}(X > k) = (1-p)^k$ per ogni $k \in \mathbf{N}$. Fissiamo

[31] Questo lo osserva anche il [B, p. 125], che poi però aggiunge che *tradizionalmente* si assume $N = 30$ o 50. In effetti la statistica ha anche una rilevante componente empirica.

[32] $0 < p \ll 1$ significa: p positivo ma molto piccolo. Difatti poi passeremo al limite per $p \to 0$ (o meglio, per $\delta = p/\lambda \to 0$, il che poi è la stessa cosa).

un intervallo temporale unitario $0 < \delta \ll 1$, cosicché al passo k corrisponde l'istante $t = k\delta$; alla v.a. X è quindi associata la v.a. riscalata $T_\delta := X\delta$. Facciamo ora tendere sia p che δ a zero, tenendo fisso il loro rapporto: $\lambda := p/\delta =$ costante. Ricordando il limite notevole $\lim_{\delta\to 0}(1-\lambda\delta)^{-1/(\lambda\delta)} = e$, otteniamo

$$\begin{aligned} &\mathbf{P}(T_\delta > t) = \mathbf{P}(X\delta > k\delta) = \mathbf{P}(X > k) = (1-p)^k = (1-\lambda\delta)^{t/\delta} \\ &= [(1-\lambda\delta)^{-1/(\lambda\delta)}]^{-\lambda t} \to e^{-\lambda t} \quad \text{per } \delta \to 0, \forall t > 0; \end{aligned} \tag{5.14}$$

d'altra parte per una v.a. T_0 avente distribuzione esponenziale $Exp(\lambda)$, $\mathbf{P}(T_0 > t) = e^{-\lambda t}$. Abbiamo così visto che $T_\delta \to T$ in legge, con $T \sim Exp(\lambda)$. Ponendo $\delta = 1/n$ e $X_n = T_\delta/\delta$ per ogni $n \in \mathbf{N}$ e quindi passando al limite per $n \to \infty$, possiamo concludere che

$$X_n \sim G(\lambda/n) \quad \Rightarrow \quad \frac{X_n}{n} \to T \quad \text{in legge,} \quad \text{con} \quad T \sim Exp(\lambda). \tag{5.15}$$

Questo significa che, per uno schema di Bernoulli in cui un evento di probabilità p molto piccola viene ripetuto ad intervalli temporali δ molto brevi, il tempo di attesa del primo successo (che ha distribuzione geometrica) può essere approssimato da una distribuzione esponenziale con parametro $\lambda = p/\delta$. Questo risultato non è sorprendente, essendo entrambe le distribuzioni sono prive di memoria.

6 Sintesi

Principali Temi Trattati.

Statistica descrittiva. Istogrammi. Mediana, funzione di ripartizione e quantili, boxplots. Media, varianza, covarianza, momenti. Regressione lineare.

Probabilità. Spazi di probabilità Ω e misura di probabilità $\mathbf{P}$. Probabilità condizionata. Formula della probabilità totale (2.3). Formula di Bayes. Indipendenza di eventi.

Calcolo combinatorio. Disposizioni, permutazioni, combinazioni e partizioni.

Variabili aleatorie discrete. Variabili aleatorie X e loro leggi $\mathbf{P}_X$. V.a. discrete e continue. Densità di probabilità discreta p_X. Leggi di Bernoulli, binomiale, ipergeometrica, geometrica e di Poisson. Funzione di ripartizione per v.a. discrete. V.a. congiunte, loro marginali e rispettive leggi. Indipendenza di v.a. discrete. Calcoli con densità. Speranza, varianza, covarianza di v.a. discrete. Teorema di Chebyshev e teorema dei grandi numeri.

Variabili aleatorie continue. Funzione di ripartizione F_X e densità $f_X = F_X'$ di v.a. continue. Quantili. Legge uniforme e legge esponenziale. Indipendenza di v.a. continue. Legge normale. Speranza, varianza, covarianza di v.a. continue. Convergenza in legge e teorema limite centrale.

Principali Leggi Esaminate.

Leggi discrete: legge di Bernoulli, binomiale, ipergeometrica, geometrica, di Poisson:

$$\text{Ber}(p)\ (= \text{Bin}(1,p)), \quad \text{Bin}(n,p), \quad \text{Iper}(r,b,n), \quad \text{Geo}(p), \quad \text{Poi}(\lambda).$$

Leggi continue: legge uniforme, esponenziale, normale (o gaussiana):

$$\text{Unif}, \quad \text{Exp}(\lambda), \quad \text{N}(\mu,\sigma^2).$$

Le rispettive densità, funzioni di ripartizione, speranze e varianze sono riassunte in una tabella [B, p. 136]. Diverse altre leggi notevoli sono studiate in letteratura. Nondimeno un ventaglio alquanto limitato di esse è sufficiente per rappresentare un gran numero di fenomeni fisici ed ingegneristici.

Si noti che si possono costruire infinite misure di probabilità, usando una densità discreta oppure, se $\Omega \subset \mathbf{R}^N$, una densità continua come in (4.4). Per ogni v.a. X, ciascuna di queste probabilità individua poi una legge.

7 Alcuni esercizi

Esercizio 1. Un mobile ha 2 cassetti. Uno contiene 2 biglie nere e 3 bianche, un'altro 1 biglia nera e 2 bianche.

(1) Prima si sceglie a caso un cassetto, poi al suo interno si estrae a sorte una biglia.

Qual è la probabilità di estrarre una biglia nera?

(2) Associamo ora il numero 2 alle biglie nere, ed il numero 3 a quelle bianche. In questo modo all'estrazione con esito X è associato un numero $\varphi(X)$.

Si calcolino speranza e varianza di $\varphi(X)$ e di $2\varphi(X) + 2$.

Esercizio 2. (1) Si rappresenti la densità del prodotto di due variabili aleatorie discrete X, Y in termini della loro densità congiunta.

(2) Da una scatola contenente 6 biglie rosse e 4 bianche si estraggono successivamente tre biglie (senza reimmissione).

(2a) Qual è la probabilità che la terza biglia sia rossa?

(2b) Qual è la probabilità che la prima e la terza biglia estratta siano dello stesso colore?

Esercizio 3. Si lancino 10 monete numerate da 1 a 10, tutte con la stessa probabilità $p \in]0, 1[$ di dare testa. Si calcolino le probabilità dei seguenti eventi:

(i) Le monete numero 3 e numero 7 danno croce e le altre danno testa,

(ii) Due monete (non specificate) danno croce e le altre danno testa,

(iii) Le monete pari danno uno stesso risultato, e pure tutte le monete dispari danno uno stesso risultato.

Esercizio 4. Si lanci un dado 3 volte.

(i) Che probabilità c'è di ottenere tre numeri pari (eventualmente ripetuti) nel caso di un dado equilibrato?

(ii) E se il dado non è equilibrato? (Si indichi con q_i la probabilità di ottenere la faccia i, con $i = 1, 2, ..., 6$.)

(iii) Che probabilità c'è di ottenere tre numeri pari diversi nel caso di un dado equilibrato?

Esercizio 5. Estraggo una dopo l'altra 10 carte da un mazzo di 40 (che consiste di 20 carte rosse e 20 nere). Scommetto che la prima carta estratta sia nera e che nell'estrazione le carte rosse si alterneranno a quelle nere. Se le prime 6 estrazioni vanno bene, qual è la probabilità di vittoria?

Si risponda distinguendo tra questi due casi:

(a) estraggo con rimpiazzo, (b) estraggo senza rimpiazzo.

8 Per saperne di più

[Baldi, 2003] è il testo di riferimento del corso; questa è una versione ridotta del [Baldi, 1992], che è più ampio soprattutto nella parte di statistica. Da segnalare i capitoli di probabilità e statistica dell'ottimo [Prodi], ed anche il [Baldi 1996] ed il [Bramanti] pure loro soprattutto per la trattazione della statistica. Il [Johnson] è la traduzione di un classico testo anglosassone ed ha carattere più applicativo. Esistono anche molti altri manuali di probabilità e statistica, anche in lingua italiana.

— P. Baldi: Calcolo delle probabilità e statistica. McGraw-Hill, Milano 1992

— P. Baldi: Appunti di metodi matematici e statistici. CLUEB, Bologna 1996

— P. Baldi: Introduzione alla probabilità con elementi di statistica. McGraw-Hill, Milano 2003

— M. Bramanti: Calcolo delle probabilità e statistica. Esculapio, Bologna 1997

— R.A. Johnson: Probabilità e statistica per ingegneria e scienze. Pearson, Paravia, Bruno Mondadori, Milano 2007

— G. Prodi: Metodi matematici e statistici per le scienze applicate. McGraw-Hill, Milano 1992