Note di Probabilità e Statistica

per il corso di Analisi Matematica II — a.a. 2011-12

A. Visintin — Facoltà di Ingegneria di Trento

Indice

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

Premessa. Questo scritto raccoglie alcune osservazioni, perlopiù riferite al testo *Introduzione alla Probabilità* di Baldi del 2003 (nel seguito citato semplicemente come [B]), volte a facilitare la comprensione dell'argomento, e ad offrire un approfondimento di alcuni aspetti.

Il [B] è sufficiente ai fini del corso per la parte di Statistica Descrittiva e di Probabilità, mentre manca quasi del tutto della parte di Statistica Inferenziale, che ai fini ingegneristici rappresenta il principale punto di arrivo della trattazione. A ciò si cerca qui di sopperire con i due paragrafi dedicati alla stima di parametri ed ai test di ipotesi.

I due paragrafi aggiuntivi su tolleranza ed affidabilità forniscono degli esempi di applicazioni dei metodi statistici a problemi ingegneristici. Si sono poi aggiunti alcuni esercizi con risoluzione.

Gli argomenti più inportanti sono contrassegnati da un pallino nero (•). I punti che quest'anno non sono stati trattati a lezione sono contrassegnati da un asterisco (∗).

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

Introdurre alcuni elementi di statistica inferenziale: la stima di parametri puntuale ed intervallare, ed i test per la verifica di ipotesi. Presentare alcune applicazioni ingegneristiche dei metodi probabilistici.

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 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]

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 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 [3] 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:

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.

---

[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.

[3] Per *campione* qui intendiamo un insieme di dati.

È facile fornire un campione, ma non è facile individuare un buon campione: quest'ultimo deve soddisfare due requisiti basilari:

(i) essere rappresentativo dell'intera popolazione, ovvero essere ben assortito (questo richiede cura nella selezione);

(ii) essere abbastanza numeroso (questo può essere costoso).

• **Due Modi di Sommare.** Sia $\{x_i\}_{i=1,...,N}$ un campione di dati numerici, ovvero un insieme di numeri $x_i \in \mathbf{R}$ con $i = 1, ..., N$. 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. [4] $x_2 = x_5$. 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.* [5] 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],

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

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: $\sigma_{xx} = \sigma_x^2$.

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

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

$*$ **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 [6] 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 linearmente 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 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 $\{\mathrm{Cov}(X_i, X_j)\}_{i,j=1,...,N}$; 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.

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

(i) La teoria classica avviata alla metà del '600 dai pionieri del calcolo delle probabilità (Pierre Fermat, Blaise Pascal, Christian 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.* [7] 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 1689). 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 inquadrano 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 in epoca successiva 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

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

l'evento *impossibile;* applicando un altro assioma della teoria della misura, abbiamo $\mathbf{P}(\emptyset) = \mathbf{P}(\Omega \backslash \Omega) = 1 - \mathbf{P}(\Omega) = 0$. Possono esservi eventi non impossibili di probabilità nulla, che saranno detti *trascurabili* o anche *quasi impossibili.* [8] 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 (Stocastica) di Eventi.** [9] 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 variabili aleatorie 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

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

• **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 (richiamata più avanti), 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?).

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

[9] L'indipendenza stocastica non ha nulla a che vedere con l'indipendenza lineare. Salvo avviso contrario, quando useremo il termine indipendenza ci riferiremo sempre a quella stocastica.

• **La Formula di Bayes.** Questa poggia sulla nozione di probabilità condizionata: per ogni coppia di eventi A, B

$$\mathbf{P}(A|B) := \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)} \qquad \text{se } \mathbf{P}(B) \neq 0. \tag{2.4}$$

Pertanto, supponendo che $\mathbf{P}(A), \mathbf{P}(B) \neq 0$,

$$\mathbf{P}(A|B) \cdot \mathbf{P}(B) := \mathbf{P}(A \cap B) = \mathbf{P}(B|A) \cdot \mathbf{P}(A),$$

da cui banalmente consegue che

$$\mathbf{P}(A|B) = \frac{\mathbf{P}(B|A)}{\mathbf{P}(B)} \cdot \mathbf{P}(A). \tag{2.5}$$

Più in generale, data una partizione $\{A_i\}_{i=1,2,...}$ di Ω costituita da eventi non trascurabili, sempre supponendo $\mathbf{P}(B) \neq 0$, abbiamo

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(B|A_i)}{\mathbf{P}(B)} \cdot \mathbf{P}(A_i) \qquad \text{per } i = 1, 2, ...$$

ed applicando la (2.3) otteniamo la formula di Bayes:

$$\mathbf{P}(A_i|B) = \frac{\mathbf{P}(B|A_i)}{\sum_j \mathbf{P}(B|A_j)\mathbf{P}(A_j)} \cdot \mathbf{P}(A_i) \qquad \text{per } i = 1, 2, ... \tag{2.6}$$

Questa formula si presta a diversi usi ed interpretazioni. Ad esempio, se le A_i sono interpretate come possibili cause di B, è naturale assegnare $P(B|A_i)$, ovvero la probabilità dell'effetto B corrispondente ad ogni possibile causa A_i. La formula di Bayes fornisce allora $\mathbf{P}(A_i|B)$, ovvero la probabilità che l'effetto B possa essere attribuito alla causa A_i. La formula di Bayes è quindi anche detta la *formula delle probabilità delle cause.*

Si possono anche interpretare le $\mathbf{P}(A_i)$ come le probabilità attribuite alle alternative A_i a priori di un certo esperimento. Le $\mathbf{P}(A_i|B)$ sono quindi le probabilità a posteriori dell'esperimento, ovvero conseguenti all'esito B dello stesso. In base alla formula di Bayes, il rapporto tra la probabilità a posteriori e quella a priori è pari a $\mathbf{P}(B|A_i)/\mathbf{P}(B)$. Se questo rapporto è maggiore (minore, rispett.) di 1, allora l'esperimento tende a confermare (a smentire, rispett.) A_i. La formula di Bayes può quindi rappresentare il progredire della nostra conoscenza in seguito all'eperienza.

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 al più m sottoinsiemi (ovvero m sottoinsiemi, uno o più dei quali eventualmente vuoti): 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. In effetti, 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 rosse, e sia k un intero tale che $0 \leq k \leq r$. Ci chiediamo quanti diversi sottoinsiemi di n elementi contenenti esattamente k biglie rosse (e quindi $n-k$ biglie bianche) 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 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 variabile aleatoria è

$$\begin{aligned} &\mathbf{P}(X=k) = \frac{\binom{r}{k}\binom{b}{n-k}}{\binom{r+b}{n}} && \text{per } k=0,...,r, \\ &\mathbf{P}(X=k) = 0 && \text{per ogni altro } k \in \mathbf{R}. \end{aligned} \tag{2.7}$$

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.** [10] La speranza di una variabile aleatoria 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 variabile aleatoria (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 consista in una successione di punti (questo esclude il caso in cui Ω contiene un intervallo). Sia allora $\Omega = \{\omega_1, ..., \omega_n, ...\}$, e per ogni $X : \Omega \to \mathbf{R}$ poniamo

$$\mathbf{E}(X) = \sum_i X(\omega_i)\mathbf{P}(\{\omega_i\}) \left(= \sum_{\omega\in\Omega} X(\omega)\mathbf{P}(\{\omega\})\right) \tag{3.1}$$

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:

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

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

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

[10] Per la speranza (o meglio la *speranza matematica)* tradizionalmente si usa il simbolo **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.

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

$$\mathbf{P}_X(y_j) := \mathbf{P}(X^{-1}(y_j)) = \sum_{i \in \alpha_j} \mathbf{P}(\{\omega_i\}) \qquad \forall j. \tag{3.5}$$

Pertanto, sempre assumendo la convergenza assoluta,

$$\mathbf{E}(X) \overset{(3.1),(3.4)}{=} \sum_j \sum_{i \in \alpha_j} X(\omega_i)\mathbf{P}(\{\omega_i\}) = \sum_j y_j \sum_{i \in \alpha_j} \mathbf{P}(\{\omega_i\}) = \sum_j y_j \mathbf{P}_X(y_j). \tag{3.6}$$

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

$$\mathbf{E}(f) = \mathbf{E}\Big(\sum_i a_i 1_{A_i}\Big) = \sum_i a_i \mathbf{E}(1_{A_i}) = \sum_i a_i \mathbf{P}(A_i), \tag{3.7}$$

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 $X : \Omega \to \mathbf{R}$ fa corrispondere un numero ad ogni evento elementare $\omega \in \Omega$; questo dipersè non sembrerebbe determinare una probabilità. Ma P (la misura di probabilità su Ω) e X determinano un'altra misura di probabilità, questa volta su $\mathbf{R}$:

$$\mathbf{P}_X(A) := \mathbf{P}(X^{-1}(A)) = \mathbf{P}(X \in A) \qquad \forall A \subset \mathbf{R}. \tag{3.8}$$

Questa nuova misura di probabilità, la $\mathbf{P}_X$, è detta la *distribuzione di probabilità* o la *legge* di X. Essa 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.** Diversi concetti della teoria delle variabili aleatorie 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 collezione di dati e la densità di probabilità di una variabile aleatoria. Queste analogie non sono ... casuali: ogni variabile aleatoria $X = X(\omega)$ si incarna in un valore numerico una volta che il caso [11] 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 [12] tanto più le frequenze relative approssimano la distribuzione di probabilità della variabile aleatoria in esame.

[11] o la fortuna, 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 presenza pervasiva della realtà in cui viviamo... 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.

[12] ovvero *numeroso* (il termine *ampiezza del campione* appartiene al gergo statistico).

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. (Simili incongruenze terminologiche non si presentano per varianza e covarianza.)

Variabili Aleatorie e loro Leggi. Si consideri la definizione (3.8): la *legge* (o *distribuzione di probabilità)* $\mathbf{P}_X$ è definita in termini della probabilità $\mathbf{P}$ e della variabile aleatoria X, ma in generale né $\mathbf{P}$ né X possono essere ricostruite a partire da $\mathbf{P}_X$. In altri termini, passando da una variabile aleatoria alla sua legge ci può essere una perdita di informazione. [13]

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

Ogni variabile aleatoria 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 variabili aleatorie *reali* (ovvero scalari) che per variabili aleatorie *multidimensionali* (ovvero vettoriali), e sia per variabili aleatorie discrete che per variabili aleatorie 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:

(i) La media, la varianza e gli altri momenti dipendono solo dalla legge delle variabili aleatorie interessate. Lo stesso vale per l'integrale di ogni funzione di una variabile aleatoria. La covarianza dipende solo dalla legge congiunta delle variabili aleatorie interessate. La speranza di una funzione di più variabili aleatorie è determinato dalla legge congiunta (ovvero, la legge della variabile aleatoria congiunta); la conoscenza delle leggi marginali (ovvero, le leggi delle variabili aleatorie marginali) non è invece sufficiente a determinarlo, a meno che le variabili aleatorie non siano (stocasticamente) indipendenti. Si noti che non ha senso parlare di indipendenza di leggi.

(ii) Una N-pla $\{X_1, .., X_N\}$ di variabili aleatorie determina la variabile aleatoria congiunta $X = (X_1, .., X_N)$, in quanto conoscere le componenti di un vettore ovviamente equivale a conoscere il vettore stesso. La legge congiunta $\mathbf{P}_X$ determina le leggi marginali $\mathbf{P}_{X_1}, ..., \mathbf{P}_{X_N}$. [14] 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 variabili aleatorie è determinata dalla legge congiunta, ma non dalle leggi delle variabili aleatorie della famiglia (ovvero dalle leggi marginali). 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. [15]

(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):

$$\mathbf{P}_{f(X)}(A) = \sum_{x \in f^{-1}(A)} \mathbf{P}_X(\{x\}) = \sum_{x \in f^{-1}(A)} p_X(x) \qquad \forall A \subset \mathbf{R}. \tag{3.9}$$

[13] In effetti una stessa legge si può incarnare (termine non tecnico!) in diverse variabili aleatorie, 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 variabile aleatoria 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 variabile aleatoria, il numero di vittorie e la sua legge.

[14] Le variabili aleatorie X_i sono dette *marginali* della $X = (X_1, .., X_N)$ perchè, per $N = 2$, tipicamente sono rappresentate ai margini della tabella matriciale che rappresenta la variabili aleatoria congiunta (X_1, X_2). Questa terminologia è usata anche per $N > 2$.

[15] A proposito di indipendenza, il [B] manca di sottolineare quanto segue. Sia $\{X_1, .., X_N\}$ una famiglia di variabili aleatorie reali (discrete o continue), sia X la variabile aleatoria 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 variabili aleatorie $\{X_1, .., X_N\}$ 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, $p_X = p_{X_1} \cdots p_{X_N}$.

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

$$p_{f(X)}(y) = \mathbf{P}_{f(X)}(\{y\}) = \sum_{x \in f^{-1}(y)} p_X(x) \qquad \forall y \in f(X(\Omega)). \tag{3.10}$$

La legge $\mathbf{P}_{f(X)}$ è quindi determinata da $\mathbf{P}_X$. In altri termini, date due variabili aleatorie X_1 e X_2, se $\mathbf{P}_{X_1} = \mathbf{P}_{X_2}$ (ovvero se X_1 e X_2 hanno la stessa legge, ovvero se 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 variabili aleatorie congiunte $X = (X_1, ..., X_N)$ che forniscono variabili aleatorie multidimensionali.

Quadro Riassuntivo. Qui $A \to B$ significa "A determina B". Le due frecce "destra verso sinistra" escluse non sono mai valide: una legge non può determinare una variabile aleatoria. Le due frecce "basso verso l'alto" escluse divengono valide se le variabili marginali sono stocasticamente indipendenti.

$$\begin{array}{ccccc} \text{v.a. congiunta} & \to\ (\not\leftarrow) & \text{legge congiunta} & \leftrightarrow & \text{densità congiunta,} \\ \updownarrow & & \downarrow\not\uparrow & & \downarrow\not\uparrow \\ \text{v.a. marginali} & \to\ (\not\leftarrow) & \text{leggi marginali} & \leftrightarrow & \text{densità marginali.} \end{array}$$

Esempi.

— (i) Siano X_1, X_2, Y_1, Y_2 quattro variabili aleatorie, tali che le variabili aleatorie 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 variabili aleatorie, tali che X_1 e Y_1 siano equidistribuite, e lo stesso valga per X_2 e Y_2. Questo non implica che le variabili aleatorie 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 variabili aleatorie sono equidistribuite.) Tuttavia le variabili aleatorie 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 variabili aleatorie 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.

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

Ecco un altro controesempio, in cui le X e Y sono appunto indipendenti. 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 variabili aleatorie 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 variabili aleatorie 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 di due diversi colori. Se le estrazioni sono con reimmissione (o "rimpiazzo"), allora esse sono indipendenti, e ad ogni estrazione ciascuna biglia ha probabilità $1/N$ di essere estratta. In questo caso il procedimento di estrazione è detto uno *schema successo-insuccesso,* ed il numero di successi (ovvero di biglie di uno dei due colori) ha distribuzione binomiale.

Se invece le estrazioni sono senza reimmissione, allora esse sono dipendenti, ed il numero di biglie di un colore estratte ha distribuzione ipergeometrica. Nondimeno, se si suppone di non sapere quali biglie sono state estratte precedentemente, allora ad ogni estrazione ciascuna biglia ha ancora probabilità $1/N$ di essere estratta, cf. [B, p. 54].

Ad esempio, siano b biglie bianche e r biglie rosse (quindi $N = b + r$), e si effettuino n estrazioni. Sia X il numero di biglie bianche estratte. Allora:

— nel caso con reimmissione, ad ogni estrazione la probabilità di estrarre una biglia bianca è $q = b/N$, e quindi

$$\mathbf{P}(X = k) = \binom{n}{k} q^k (1-q)^{n-k} \qquad \text{per } k = 0, ..., n; \tag{3.11}$$

— nel caso senza reimmissione, il rapporto "# casi favorevoli"/"# casi pssibili" fornisce

$$\mathbf{P}(X = k) = \binom{b}{k}\binom{r}{n-k} \Big/ \binom{b+r}{n} \qquad \text{per } k = 0, ..., n. \tag{3.12}$$

La Matrice di Covarianza. Sia N un qualsiasi intero ≥ 1, e $\{X_i\}_{i=1,...,N}$ una famiglia di variabili aleatorie. 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

$$\begin{aligned} &\mathrm{Var}\,(\textstyle\sum_i X_i) = \mathrm{Cov}\,(\sum_i X_i, \sum_j X_j) = \sum_i \sum_j \mathrm{Cov}\,(X_i, X_j) \\ &= \textstyle\sum_{i=j} \mathrm{Cov}\,(X_i, X_j) + \sum_{i\neq j} \mathrm{Cov}\,(X_i, X_j) = \sum_i \mathrm{Var}\,(X_i) + 2\sum_{i<j} \mathrm{Cov}\,(X_i, X_j). \end{aligned} \tag{3.13}$$

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

$$\mathrm{Cov}(X \cdot X^\tau) := \{\mathrm{Cov}(X_i, X_j)\}_{i,j=1,...,N}\ , \tag{3.14}$$

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 variabili aleatorie $\{X_i\}_{i=1,...,N}$ è linearmente indipendente. Essa è diagonale se e solo se queste variabili aleatorie sono non correlate. [16] In questo caso allora la varianza della somma è uguale alla somma delle varianze:

$$\mathrm{Var}\Big(\sum_i X_i\Big) = \sum_i \mathrm{Var}\,(X_i) \quad \text{ per variabili aleatorie non correlate.} \tag{3.15}$$

Come noto, in particolare questo si verifica se le variabili aleatorie sono indipendenti.

A prima vista questo risultato può apparire un po' sorprendente, poiché una formula del genere vale per funzionali lineari (e.g. la speranza), mentre la varianza è quadratica! Come è evidente dalla

[16] Si osservi che la non correlazione è quella che si potrebbe definire *indipendenza (stocastica) 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 variabili aleatorie misura l'eventuale esistenza di una relazione lineare tra le due variabili aleatorie. Tuttavia la covarianza potrebbe essere nulla anche in presenza di un legame non lineare tra le variabili aleatorie.

dimostrazione, questo poggia sull'ipotesi di non correlazione. [17] Questi risultati si estendono al caso di una successione di variabili aleatorie, sotto la condizione che tutte le quantità che compaiono in questa formula (le speranze, le covarianze, le serie) convergano.

• **Il Coefficiente di Correlazione.** Date due variabili aleatorie X, Y, si definisce il loro coefficiente di correlazione

$$\rho_{X,Y} := \frac{\mathrm{Cov}(X,Y)}{\sigma_X\ \sigma_Y} \qquad \text{se } \sigma_X, \sigma_Y \neq 0. \tag{3.16}$$

Valgono le seguenti proprietà:

$$\begin{aligned} &\rho_{X,Y} = 1 \quad \Rightarrow \quad \exists a > 0, \exists b \in \mathbf{R} : Y = aX + b, \\ &\rho_{X,Y} = -1 \quad \Rightarrow \quad \exists a < 0, \exists b \in \mathbf{R} : Y = aX + b, \\ &\rho_{X,Y} = 0 \quad \Leftarrow \quad (\not\Rightarrow) \quad \text{se } X, Y \text{ sono indipendenti}, \\ &\forall a, b, c, d \in \mathbf{R} \text{ con } a \cdot c > 0,\ U := aX + b, V := cY + d \ \Rightarrow\ \rho_{U,V} = \rho_{X,Y}. \end{aligned} \tag{3.17}$$

• **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 variabili aleatorie 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

$$\mathrm{Var}\left(\frac{1}{N}\sum_i X_i\right) = \frac{1}{N^2}\,\mathrm{Var}\left(\sum_i X_i\right) \overset{(3.15)}{=} \frac{1}{N^2}\sum_i \mathrm{Var}\,(X_i) = \frac{1}{N}\,\mathrm{Var}\,(X_1). \tag{3.18}$$

Più in generale, si definisce *campione aleatorio di ampiezza* N una famiglia di N variabili aleatorie indipendenti equidistribuite. [18] Quindi la *media campionaria* (o *media empirica)* di un campione aleatorio, ovvero la variabile aleatoria $\frac{1}{N}\sum_i X_i$, ha varianza ridotta 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. [19] Il fatto che una popolazione [20] 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.

---

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

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

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.

[18] Una tale situazione si presenta in svariati ambiti applicativi. Ad esempio la conferma sperimentale propria del metodo scientifico presuppone la riproducibilià dei risultati sperimentali; ovvero la possibilità di effettuare prove sperimentali il cui esito sia rappresentato da una successione di variabili aleatorie indipendenti ed equidistribuite.

[19] 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 aleatorio, allora si possono effettuare due tipi di medie. Si può mediare rispetto a i, ottenendo la *media campionaria;* questa è la variabile aleatoria $\bar{X}_n := \frac{1}{n}\sum_i X_i$ (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.

[20] 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,* ma possono ben essere inanimati.

Nella pratica l'uso di un campione aleatorio 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. [21] La (3.15) 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 variabili aleatorie, 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]). [22] Questa stima è alquanto grossolana, dovendo valere per ogni campione aleatorio. 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 aleatori infiniti — il che ovviamente richiede un passaggio al limite. [23]

Illustriamo brevemente la dimostrazione del teorema. Preliminarmente ricordiamo che per ogni evento $A \subset \Omega$, 1_A si definisce 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 variabile aleatoria, la funzione indicatrice di quell'evento appunto.

Per ogni variabile aleatoria $Y \geq 0$ ed ogni $c > 0$, ovviamente $c1_{\{Y \geq c\}} \leq Y$; quindi

$$c\mathbf{P}(Y \geq c) = \mathbf{E}(c1_{\{Y \geq c\}}) \leq \mathbf{E}(Y) \qquad \text{(disuguaglianza di Markov).} \tag{3.19}$$

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

$$\mathbf{P}(|X - \mu| \geq c) = \mathbf{P}(|X - \mu|^2 \geq c^2) \leq \frac{1}{c^2}\mathbf{E}(|X - \mu|^2) = \frac{1}{c^2}\operatorname{Var}(X). \tag{3.20}$$

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

$$\mathbf{P}(|\bar{X}_n - \mu| \geq c) \leq \frac{1}{c^2}\operatorname{Var}(\bar{X}_n) \overset{(3.18)}{=} \frac{1}{nc^2}\operatorname{Var}(X_1) \to 0 \qquad \text{per } n \to \infty, \forall c > 0; \tag{3.21}$$

oppure equivalentemente

$$\mathbf{P}(|\bar{X}_n - \mu| \leq c) = 1 - \mathbf{P}(|\bar{X}_n - \mu| \geq c) \geq 1 - \frac{1}{c^2}\operatorname{Var}(\bar{X}_n) \to 1 \qquad \text{per } n \to \infty, \forall c > 0. \tag{3.22}$$

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

[22] 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.

[23] 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.

Questo teorema permette di approssimare non solo la speranza delle funzioni del campione aleatorio $\{X_i\}$, ma anche la speranza di ogni funzione del campione aleatorio $\{X_i\}$. Sia $f : \mathbf{R} \to \mathbf{R}$ una funzione continua tale che $Y_i := f(X_i)$ ha speranza finita $\widetilde{\mu}$ (essendo il campione equidistribuito, questa speranza non dipende da i) e varianza pure finita. Denotando con $\bar{Y}_n$ la media campionaria delle Y_i, allora la (3.21) fornisce

$$\mathbf{P}(|\bar{Y}_n - \widetilde{\mu}| \geq c) \leq \frac{1}{c^2} \operatorname{Var}(\bar{Y}_n) \stackrel{(3.18)}{=} \frac{1}{nc^2} \operatorname{Var}(Y_1) \to 0 \qquad \text{per } n \to \infty, \forall c > 0. \tag{3.23}$$

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

Bernoullizzazione. [24] Il teorema dei grandi numeri permette di approssimare non solo la speranza di legge discreta, ma anche di simulare una qualsiasi legge discreta, nel senso che ora illustriamo; cf. [B, p. 72]. Sia $\{Y_i\}_{i \in \mathbf{N}}$ un qualsiasi campione aleatorio avente quella legge, e si fissi un $r \in Y_1(\Omega)$ (questo insieme immagine è comune a tutte le Y_i, che sono equidistribuite). Si ponga poi $X_i = 1_{\{Y_i = r\}}$ per ogni $i \in \mathbf{N}$, [25] si noti che anche questo è un campione aleatorio (ovvero è costituito da variabili aleatorie equidistribuite indipendenti). Pertanto

$$\mathbf{E}(X_i) = \mathbf{E}(1_{\{Y_i = r\}}) = \mathbf{P}(\{Y_i = r\}) = \mathbf{P}(\{Y_1 = r\}) =: q \qquad \text{per ogni } i.$$

Definiamo poi al solito la media campionaria $\bar{X}_n := \frac{1}{n} \sum_{i \leq n} X_i$ per ogni n, e notiamo che $\mathbf{E}(\bar{X}_n) = \mathbf{E}(X_1) = q$ per ogni n. Allora il teorema dei grandi numeri fornisce

$$\bar{X}_n \to q \quad \text{(nel senso della convergenza in probabilità).}$$

Applicando questo procedimento ad ogni $r \in Y_1(\Omega)$, si può approssimare l'intera densità di probabilità della legge discreta prefissata.

La variabile aleatoria $\bar{X}_n$ rappresenta la frequenza empirica dell'esito $Y_i = r$:

$$\bar{X}_n(\omega) = \frac{1}{n} \sum_{i \leq n} 1_{\{Y_i = r\}}(\omega) = \frac{\#\{i \leq n : Y_i(\omega) = r\}}{n} \qquad \text{per } \omega \in \Omega. \tag{3.24}$$

In questo modo si vede come il teorema dei grandi numeri giustifichi l'interpretazione frequentista del calcolo delle probabilità.

Ad esempio la bernoullizzazione di un dado equilibrato consiste nel simulare la legge uniforme sull'insieme $\{1, ..., 6\}$ mediante la sequente legge di Bernoulli:

$$p(k) = 1/6 \quad \text{per } k = 1 \text{ (ovvero testa)}, \qquad p(k) = 0 \quad \text{per } k = 0 \text{ (ovvero croce)}.$$

In questo modo, per ogni $r \in \{1, ..., 6\}$, la probabilità che un lancio del dado dia esito r è pari alla probabilità che un lancio della moneta dia esito testa.

Qui è bastata una sola legge di Bernoulli poiché le facce erano equiprobabili. Per un dado non equilibrato occorrono sei leggi di Bernoulli: per ogni $r \in \{1, ..., 6\}$, indicato con q_r la probabilità che un lancio del dado dia esito r, in questo caso si pone

$$p_r(k) = q_r \quad \text{per } k = 1 \text{ (ovvero testa)}, \qquad p_r(k) = 0 \quad \text{per } k = 0 \text{ (ovvero croce)}.$$

$*$ **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 variabile aleatoria X_i di

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

[25] Questa è ovviamente una variabile aleatoria di Bernoulli, una *bernoullizzata* della X_i (il termine non è standard!).

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 aleatorio.

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

$$\mathbf{E}(\bar{X}_n) = \frac{1}{n}\sum_i \mathbf{E}(X_i) = 0. \tag{3.25}$$

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

$$\begin{aligned} &\mathbf{E}[(\bar{X}_n)^2] = \tfrac{1}{n^2}\mathbf{E}(\textstyle\sum_i X_i \cdot \sum_j X_j) = \tfrac{1}{n^2}\sum_{i,j}\mathbf{E}(X_i \cdot X_j) \\ &= \tfrac{1}{n^2}\textstyle\sum_{i=j}\mathbf{E}(X_i \cdot X_j) + \tfrac{1}{n^2}\sum_{i\neq j}\mathbf{E}(X_i \cdot X_j) = \tfrac{n}{n^2} + 0 = \tfrac{1}{n}\ . \end{aligned} \tag{3.26}$$

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

$$\mathbf{E}[(\bar{X}_n)^3] = \tfrac{1}{n^3}\mathbf{E}(\textstyle\sum_i X_i \cdot \sum_j X_j \cdot \sum_k X_k) = \tfrac{1}{n^3}\sum_{i,j,k}\mathbf{E}(X_i \cdot X_j \cdot X_k) = 0, \tag{3.27}$$

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

$$\begin{aligned} &\mathbf{E}[(\bar{X}_n)^4] = \tfrac{1}{n^4}\mathbf{E}(\textstyle\sum_i X_i \cdot \sum_j X_j \cdot \sum_k X_k \cdot \sum_\ell X_\ell) \\ &= \tfrac{1}{n^4}\textstyle\sum_{i,j,k,\ell}\mathbf{E}(X_i \cdot X_j \cdot X_k \cdot X_\ell) = \tfrac{3}{n^4}\sum_{i,k}\mathbf{E}(X_i^2 \cdot X_k^2) = \tfrac{3n^2}{n^4} = \tfrac{3}{n^2}\ . \end{aligned} \tag{3.28}$$

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

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

$$\mathbf{E}(\bar{X}_n) \overset{(3.25)}{=} 0, \qquad \mathrm{Var}(\bar{X}_n) \overset{(3.25)}{=} \mathbf{E}[(\bar{X}_n)^2] \overset{(3.26)}{=} 1/n\ . \tag{3.29}$$

Confrontando quest'ultima uguaglianza con la (3.18), vediamo che questa è una manifestazione del principio di mutua compensazione — d'altra parte la (3.26) è stata dimostrata proprio 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 $\delta(t)^2$ rappresenta la *distanza quadratica media* dall'origine all'istante t, ovvero la varianza della distanza (che è una variabile aleatoria centrata). Abbiamo

$$\delta(t)^2 = \mathbf{E}[(nh\bar{X}_n)^2] \overset{(3.25)}{=} (nh)^2\,\mathrm{Var}(\bar{X}_n) \overset{(3.29)}{=} nh^2 = Dt \quad (\text{ponendo } D := h^2/\tau); \tag{3.30}$$

D è il *coefficiente di diffusione.* Nei fenomeni di diffusione quindi la *distanza quadratica media* è proporzionale al tempo: $\delta(t)^2 = Dt$; nei fenomeni di trasporto invece la distanza media è proporzionale al tempo: $\delta(t) = Ct$ (per un $C > 0$). Questo avviene 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 variabile aleatoria $X : \Omega \to \mathbf{R}$ definiamo la *funzione di ripartizione* [26]

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

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, a meno di un insieme H_X formato al più da una successione di punti. [27] 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,

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

da cui (un po' disinvoltamente)

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

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 variabile aleatoria 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 variabile aleatoria 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 variabile aleatoria X è definita dal [B] in termini della legge della variabile aleatoria 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

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

Quindi h è la densità della misura di probabilità $\mathbf{P}$ su Ω. [28] Allora la speranza di una variabile aleatoria (discreta o continua) $X : \Omega \to \mathbf{R}$ è

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

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

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

[26] La funzione $S_X : \mathbf{R} \to [0,1]$ definita da $S_X(y) = \mathbf{P}(X > y) = 1 - F_X(y)$ per ogni $y \in \mathbf{R}$ è detta *funzione di sopravvivenza* di X.

[27] 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.

[28] 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 variabile aleatoria 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.

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 variabili aleatorie discrete; si vedano gli sviluppi (3.1), ..., (3.6).

La definizione (4.7) è posta solo per variabili aleatorie continue; è noto che per variabili aleatorie discrete la speranza va scritta diversamente, con una serie invece di un integrale. Per tali variabili aleatorie 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 variabili aleatorie $X, Y : \Omega \to \mathbf{R}$, ciascuna discreta o continua. La loro covarianza vale

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

Se X e Y sono variabili aleatorie 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:

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

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

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

e h dovrà soddisfare la condizione di normalizzazione [29]

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

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

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

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

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

La funzione $\widetilde{h}$ 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

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

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 variabile aleatoria. 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

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

allora il guadagno atteso è

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

La varianza della variabile aleatoria g è

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

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 $\{g \le t\} := \{\omega \in \Omega : g(\omega) \le t\}$ 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 aleatorio avente la 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.

• **Osservazioni sulle Leggi Simmetriche.** Sia X una qualsiasi legge continua simmetrica, ovvero la cui densità $p_X : \mathbf{R} \to [0,1]$ è una funzione pari. Siano F_X e $F_{|X|}$ le funzioni di ripartizione di X e $|X|$. Allora

$$F_X(\lambda) + F_X(-\lambda) = 1 \qquad \forall \lambda \in [0,1], \tag{4.14}$$

$$F_{|X|}(\lambda) = F_X(\lambda) - F_X(-\lambda) \overset{(4.14)}{=} 2F_X(\lambda) - 1 \qquad \forall \lambda \in [0,1]. \tag{4.15}$$

Se la densità è ovunque non nulla, allora F_X è invertibile; poniamo quindi $\phi := F_X^{-1}$. Questa è la funzione dei quantili della legge prescritta; in altri termini, ϕ_α è il quantile di ordine α, ovvero $F_X(\phi_\alpha) = \alpha$ per ogni $\alpha \in [0,1]$. Grazie alle formule precedenti, si verifica facilmente che

$$F_{|X|}(\phi_{(\alpha+1)/2}) = \alpha, \qquad \phi_\alpha + \phi_{1-\alpha} = 0 \qquad \forall \alpha \in [0,1]. \tag{4.16}$$

Pertanto la funzione F_X individua la $F_{|X|}$, ed anche le rispettive funzioni dei quantili.

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 variabili aleatorie.

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

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

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

(Si noti che $\mathbf{P}(|X_n - X| \geq c) \to 0$ equivale a $\mathbf{P}(|X_n - X| < c) \to 1$.) Ad esempio la convergenza in probabilità compare nel teorema dei grandi numeri, e quella in legge nel teorema limite centrale. Si noti che

$$\begin{gathered} \text{la convergenza in probabilità coinvolge le variabili aleatorie,} \\ \text{mentre quella in legge dipende solo dalle loro leggi.} \end{gathered} \tag{5.3}$$

Sussiste un importante legame tra questi due concetti:

$$\text{la convergenza in probabilità implica quella in legge, ma non viceversa.} \tag{5.4}$$

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

(i) *Legge binomiale ⇒ 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]. [30] 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 $\lambda = np$. [31]

[30] Poiché si parla della convergenza di distribuzioni (piuttosto che di variabile aleatoria), si tratta necessariamente di una convergenza in legge.

[31] 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.

(ii) *Legge geometrica ⇒ 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 aleatorio di variabili aleatorie 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 variabili aleatorie 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) [32] Se $\{X_i\}_{i\in\mathbf{N}}$ è un campione aleatorio di variabili aleatorie con speranza μ e varianza finita $\sigma^2 > 0$, [33] allora la distribuzione della media campionaria standardizzata converge alla distribuzione normale standard [B, p. 124], ovvero

$$S_n^* := \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \to Y \quad \text{in legge,} \qquad \text{con } Y \sim N(0,1). \tag{5.5}$$

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

Tabella Riassuntiva. Lo schema di Bernoulli rappresenta un serie di accadimenti di due tipi (e.g., successi e insuccessi). Il numero di successi in n avvenimenti ha distribuzione binomiale. Il numero di avvenimenti prima del primo successo ha distribuzione geometrica.

Lo schema di Poisson rappresenta un serie di accadimenti indipendenti che avvengono ad istanti casuali. Il numero di accadimenti in un intervallo di tempo ha distribuzione di Poisson. Il tempo che intercorre tra due accadimenti successivi ha distribuzione esponenziale.

La freccia rappresenta il passaggio al limite sopra illustrato.

Schema di Bernoulli		Schema di Poisson
legge binomiale	$\to$	legge di Poisson
legge geometrica	$\to$	legge esponenziale

Osservazioni sul TLC. (i) Per il principio di mutua compensazione, $\mathrm{Var}(\bar{X}_n) = \mathrm{Var}(X_i)/n$ per ogni n ed 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 variabile aleatoria standardizzata della media campionaria, ed è diversa dalla media campionaria delle variabili aleatorie X_i standardizzate:

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

[32] De Moivre dimostrò questo risultato per variabili aleatorie 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 rilevante fu quello di De Moivre, già nel 1733!)

Perché il teorema si chiama così? Su questo non sono tutti d'accordo: i più lo intendono come *teorema centrale del limite,* ed 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 del limite centrale.* Tra l'altro, la denominazione inglese *central limit theorem* si presta ad entrambe le interpretazioni.

[33] 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.

(ii) La (5.5) può essere riscritta

$$\begin{aligned}
&\sqrt{n}(\bar{X}_n - \mu) \simeq \sigma Y \sim N(0, \sigma^2), \quad \text{oppure} \\
&\bar{X}_n \simeq \mu + \frac{\sigma}{\sqrt{n}} Y \sim N(\mu, \sigma^2/n), \quad \text{oppure} \\
&\sum_{i=1}^{n} X_i \simeq n\mu + \sqrt{n}\sigma Y \sim N(n\mu, n\sigma^2) \quad \text{per } n \to \infty.
\end{aligned} \tag{5.6}$$

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 variabile aleatoria Z ha distribuzione normale di speranza μ e varianza σ^2. Quindi quanto più grande è n, tanto più la legge della variabile aleatoria $\bar{X}_n$ 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 fondamentali proprietà:

(a) la speranza di una somma di variabili aleatorie è la somma delle speranze;

(b) la varianza di una somma di variabili aleatorie 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 variabili aleatorie.

(iii) La seconda formula della (5.6) può anche essere interpretata come segue: per n "abbastanza" grande, la legge della variabile aleatoria $\bar{X}_n$ finisce con dipendere dalle variabili X_i "essenzialmente" solo attraverso la loro media μ e la loro varianza σ^2/n. [34] Più precisamente, quanto più n è grande, tanto più questo è vero.

(iv) Se ciascuna variabile aleatoria X_i del campione aleatorio ha legge normale $N(\mu, \sigma^2)$, allora si può dimostrare che la (5.6) 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. [35] Sia $\{X_i\}_{i \in \mathbf{N}}$ un campione aleatorio di variabili aleatorie con speranza μ e varianza finita $\sigma^2 > 0$, e sia Y una variabile aleatoria normale standard. Fissiamo una qualsiasi costante $c > 0$, ed osserviamo che, definendo S_n^* come in (5.5) e denotando con $F_{S_n^*}$ la sua funzione di ripartizione,

$$\begin{aligned}
&\mathbf{P}(|\bar{X}_n - \mu| \le c) = \mathbf{P}\Big(|S_n^*| \le \frac{\sqrt{n}\, c}{\sigma}\Big) = \mathbf{P}\Big(-\frac{\sqrt{n}\, c}{\sigma} \le S_n^* \le \frac{\sqrt{n}\, c}{\sigma}\Big) \\
&= F_{S_n^*}\Big(\frac{\sqrt{n}\, c}{\sigma}\Big) - F_{S_n^*}\Big(-\frac{\sqrt{n}\, c}{\sigma}\Big).
\end{aligned} \tag{5.7}$$

Per via della (5.5),

$$F_{S_n^*}(t) \to F_Y(t) \qquad \text{per } n \to \infty, \forall t \in \mathbf{R}. \tag{5.8}$$

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 $F_{S_n^*}(-\frac{\sqrt{n}\, c}{\sigma}) \to F_Y(-\infty)$. La (5.7) quindi implica

$$\mathbf{P}(|\bar{X}_n - \mu| \le c) \to F_Y(+\infty) - F_Y(-\infty) = 1. \tag{5.9}$$

Quindi

$$\mathbf{P}(|\bar{X}_n - \mu| \ge c) = 1 - \mathbf{P}(|\bar{X}_n - \mu| \le c) \to 0 \qquad \forall c > 0, \text{ per } n \to \infty, \tag{5.10}$$

ovvero $\bar{X}_n \to \mu$ in probabilità, come affermato dal TGN, cf. (3.21).

(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 (5.5)): per $n \to \infty$, in base al TGN

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

[35] 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 (5.4) abbiamo visto che la convergenza in legge non implica quella in probabilità.

questa è una forma indeterminata del tipo $\frac{0}{0}$ (nel senso della convergenza in probabilità). Il TLC sostanzialmente afferma che

$$\bar{X}_n - \mu \simeq \frac{\sigma}{\sqrt{n}} Y \quad \text{con } Y \sim N(0,1), \text{ per } n \to \infty, \tag{5.11}$$

nel senso della convergenza in legge. Grazie alla (5.5), per n abbastanza grande quindi abbiamo

$$\mathbf{P}(|\bar{X}_n - \mu| \le c) = F_{|S_n^*|}\Big(\frac{\sqrt{n}\,c}{\sigma}\Big) \simeq F_{|Y|}\Big(\frac{\sqrt{n}\,c}{\sigma}\Big) = 2F_Y\Big(\frac{\sqrt{n}\,c}{\sigma}\Big) - 1 \qquad \forall c > 0; \tag{5.12}$$

ovvero

$$\mathbf{P}(|\bar{X}_n - \mu| \ge c) = 1 - F_{|S_n^*|}\Big(\frac{\sqrt{n}\,c}{\sigma}\Big) \simeq 2 - 2F_Y\Big(\frac{\sqrt{n}\,c}{\sigma}\Big) \qquad \forall c > 0, \tag{5.13}$$

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

Quest'ultima formula può essere confrontata con la (3.21):

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

$*$ **Il Teorema di Berry-Esseen.** Assodato che $F_{S_n^*} \to F_Y$ (vedi (5.8)) resta da valutare la velocità di questa convergenza. Sotto le ipotesi del TLC, si può dimostrare (teorema di Berry-Esseen) che, se $\mathbf{E}(|X_i|^3) < +\infty$, allora esiste una costante $C > 0$ tale che

$$\sup_{y \in \mathbf{R}} \Big|\mathbf{P}\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le y\Big) - \mathbf{P}(Y \le y)\Big| \le C\frac{\mathbf{E}(|X_i|^3)}{\sigma^3\sqrt{n}}. \tag{5.14}$$

Si noti che, dato un campione aleatorio, 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 (3.21) 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 (5.6), che va intesa nel senso della convergenza 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 una certa analogia con lo sviluppo di Taylor.

Sull'Uso del TLC. Fissata la legge del campione aleatorio, se $\mathbf{E}(|X_i|^3) < +\infty$ allora per ogni $\epsilon > 0$ esiste un N tale che, per ogni $Y \sim N(0,1)$,

$$\sup_{y \in \mathbf{R}} \Big|\mathbf{P}\Big(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \le y\Big) - \mathbf{P}(Y \le y)\Big| \le \epsilon \qquad \forall n \ge N. \tag{5.15}$$

Infatti basta scegliere $N = C^2\mathbf{E}(|X_i|^3)^2/\sigma^6\epsilon^2$ e poi applicare la (5.14). Poiché $\mathbf{E}(|X_i|^3)^2$ e σ sono determinati dalla legge del campione aleatorio, per ogni legge e per ogni tolleranza $\epsilon > 0$, resta così individuato un N che soddisfa la (5.15). Sottolineiamo che non è possibile indicare un N che valga per ogni campione e per ogni $\epsilon > 0$. [36]

$*$ **Legge Geometrica $\Rightarrow$ Legge Esponenziale.** Dimostriamo questo teorema limite. Una variabile aleatoria X a valori interi abbia distribuzione geometrica $G(p)$ con $0 < p \ll 1$; [37] quindi $\mathbf{P}(X > k) = (1-p)^k$ per ogni $k \in \mathbf{N}$. Fissiamo un intervallo temporale unitario $0 < \delta \ll 1$, cosicché al passo k corrisponde l'istante $t = k\delta$; alla variabile aleatoria X è quindi associata la variabile aleatoria riscalata

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

[37] $0 < p \ll 1$ 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).

$T_\delta := X\delta$. 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

$$\begin{aligned} &\mathbf{P}(T_\delta > t) = \mathbf{P}(X\delta > k\delta) = \mathbf{P}(X > k) = (1-p)^k = (1-\lambda\delta)^{t/\delta} \\ &= [(1-\lambda\delta)^{-1/(\lambda\delta)}]^{-\lambda t} \to e^{-\lambda t} \quad \text{per } \delta \to 0, \forall t > 0; \end{aligned} \tag{5.16}$$

d'altra parte per una variabile aleatoria 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

$$X_n \sim G(\lambda/n) \quad \Rightarrow \quad \frac{X_n}{n} \to T \quad \text{in legge,} \quad \text{con} \quad T \sim Exp(\lambda). \tag{5.17}$$

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 prive di memoria.

There are three kind of lies: lies, damned lies, and statistics. (Disraeli)

Statistical thinking will one day be as necessary
for efficient citizenship as the ability to read and write. (H.G. Wells)

6 • Stima di Parametri

Parole chiave: Stimatori e stime puntuali. Metodo dei momenti. Legge di Student. Stimatori di massima verosimiglianza. Stime intervallari. Confidenza.

Consideriamo uno spazio di probabilità che dipende da un parametro incognito θ appartenente ad un insieme Θ. (In alternativa, una famiglia di leggi di probabilità parametrizzate da $\theta \in \Theta$, che comunque definisce uno spazio di probabilità su $\mathbf{R}$.) Un tipico problema statistico consiste nella valutazione di θ (oppure di una funzione di θ) tramite la ripetizione di un esperimento aleatorio. Questa procedura è rappresentata da un campione aleatorio di ampiezza $n \in \mathbf{N}$, ovvero da una famiglia di n variabili aleatorie indipendenti equidistribuite: $X_1, ..., X_n$. L'esperimento può dar luogo ad esiti diversi; ciascun risultato è rappresentato dalla scelta di un valore di $\omega \in \Omega$, e quindi da una delle n-ple di valori che possono essere assunte da questa famiglia di n variabili aleatorie.

Stimatori. Si distinguono *stime puntuali* e *stime intervallari;* le prime forniscono degli scalari o dei vettori, le seconde degli intervalli o dei prodotti cartesiani di intervalli. Incominciamo con considerare le stime puntuali. Poiché la misura di probabilità $\mathbf{P}$ dipende dal parametro incognito $\theta \in \Theta$, si tratta di una probabilità condizionata. La speranza e per la varianza dipendono dalla misura di probabilità, e quindi dal parametro incognito $\theta \in \Theta$. Metteremo in evidenza questa dipendenza scrivendo $\mathbf{P}_\theta$, $\mathbf{E}_\theta$ e Var_θ invece di $\mathbf{P}$, $\mathbf{E}$ e Var.

Si dice *statistica campionaria* (o più semplicemente *statistica)* una funzione delle variabili aleatorie $X_1, ..., X_n$ che costituiscono il campione, che non dipende da alcun ulteriore parametro; ovvero una funzione $T(X_1, ..., X_n)$ ove $T : \mathbf{R}^n \to \mathbf{R}$, se il campione ha ampiezza n. [38] Quindi $T(X_1, ..., X_n)$ è una variabile aleatoria; a volte per brevità scriveremo T al posto di $T(X_1, ..., X_n)$.

Si dice *stimatore* (puntuale) di un parametro scalare θ incognito una statistica (ovvero una funzione di $X_1, ..., X_n$) che serve a stimare (ovvero, a congetturare) θ. Più in generale, se f è una funzione

[38] In effetti si tratta di una famiglia di funzioni: una funzione per ciascun n.

$\Theta \to \mathbf{R}$, si dice stimatore di $f(\theta)$ una statistica che serve a stimare $f(\theta)$. Per indicare uno stimatore di un parametro incognito θ si usa anche la notazione $\widehat{\theta}$. Quando alle variabili aleatorie che costituiscono il campione sostituiamo i valori osservati (ovvero effettuiamo un *campionamento,* che è rappresentato dalla scelta di un $\omega \in \Omega$), lo stimatore di un parametro incognito θ diventa una *stima* (di quel parametro incognito). Quindi uno stimatore è una variabile aleatoria, mentre le stime sono i valori che esso assume in seguito ai diversi campionamenti.

Introduciamo ora alcune delle principali proprietà che uno stimatore può avere o meno. Se $T = T(X_1, ..., X_n)$ è uno stimatore di un parametro incognito θ,

$$\begin{aligned} &\mathbf{E}_\theta(T(X_1, ..., X_n)) - \theta \ \text{ è detta } \textit{distorsione} \text{ di } T, \\ &\mathbf{E}_\theta[(T(X_1, ..., X_n) - \theta)^2] \ \text{ è detto} \\ &\textit{errore quadratico medio} \text{ o } \textit{rischio quadratico} \text{ di } T. \end{aligned} \tag{6.1}$$

Si noti che queste due quantità sono funzioni di $\theta \in \Theta$; il campione aleatorio invece è fissato (sono fissate le variabili $X_1, ..., X_n$, non i loro valori!) .

Uno stimatore T di θ è detto

$$\begin{aligned} &\textit{corretto} \text{ (o } \textit{non distorto)} \ \text{ se } \ \mathbf{E}_\theta[T(X_1, ..., X_n)] = \theta \quad \forall n, \forall \theta \in \Theta, \\ &\textit{asintoticamente corretto} \ \text{ se } \ \mathbf{E}_\theta[T(X_1, ..., X_n)] \to \theta \text{ per } n \to \infty, \ \ \forall \theta \in \Theta, \\ &\textit{consistente in media quadratica} \ \text{ se } \ \mathbf{E}_\theta[(T(X_1, ..., X_n) - \theta)^2] \to 0 \text{ per } n \to \infty, \forall \theta \in \Theta. \end{aligned} \tag{6.2}$$

Si intende che ciascuna di queste condizioni è richiesta per ogni campione $(X_1, ..., X_n, ...)$.

Errore Quadratico Medio = Varianza + Distorsione^2. Ponendo $\bar{T} := \mathbf{E}_\theta(T)$ abbiamo

$$\begin{aligned} \mathbf{E}_\theta[(T-\theta)^2] &= \mathbf{E}_\theta[(\{T - \bar{T}\} + \{\bar{T} - \theta\})^2] \\ &= \mathbf{E}_\theta[(T - \bar{T})^2] + \mathbf{E}_\theta[(\bar{T} - \theta)^2] + 2\mathbf{E}_\theta[(T - \bar{T})(\bar{T} - \theta)] \\ &= \mathrm{Var}_\theta(T) + [\mathbf{E}_\theta(T - \theta)]^2. \end{aligned} \tag{6.3}$$

(La speranza del doppio prodotto si annulla perchè, dal momento che $\bar{T} - \theta$ non dipende da ω,

$$\mathbf{E}_\theta[(T - \bar{T})(\bar{T} - \theta)] = \mathbf{E}_\theta[(T - \bar{T})]\ (\bar{T} - \theta) = 0.)$$

Quindi

$$\begin{aligned} &\text{errore quadratico medio} = \text{varianza} + \text{distorsione}^2, \\ &\text{errore quadratico medio} = \text{varianza}, \quad \text{se lo stimatore è corretto.} \end{aligned} \tag{6.4}$$

In generale uno stimatore corretto è preferibile ad uno distorto; tuttavia vi sono anche degli utili stimatori distorti. La minimizzazione dell'errore quadratico medio è un buon criterio di scelta tra gli stimatori. Ad esempio, un'importante classe di stimatori è costituita da quelli che minimizzano l'errore quadratico medio (ovvero la varianza) tra tutti gli stimatori corretti. In altri termini, uno stimatore corretto T di θ appartiene a questa classe se solo se la sua varianza è minore di quella di ogni altro stimatore corretto di θ.

Metodo dei Momenti. Questo semplicemente consiste nell'utilizzare i momenti campionari per stimare certe quantità che dipendono dal parametro θ. Tipicamente questo si applica quando tali quantità sono proprio i momenti di una legge, che dipendono dal parametro θ. Più in generale, se θ è un vettore N dimensionale, si calcolano N momenti campionari, e li si pone uguali ai corrispondenti momenti della legge, che appunto dipendono da θ. Si ottiene così un sistema di N equazioni in N incognite; la sua soluzione (se esiste) fornisce i parametri incogniti.

Ad esempio, se X è una variabile aleatoria, la sua speranza $E_\theta(X)$ (supposta finita) potrà essere convenientemente stimata dalla media campionaria $\bar{X}_n := \frac{1}{n}\sum_i X_i$, essendo $\{X_1, .., X_n\}$ un campione aleatorio avente la stessa distribuzione di X. Poiché $E_\theta(\bar{X}_n) := \frac{1}{n}\sum_i E_\theta(X_i) = E_\theta(X)$, questo stimatore è corretto.

La varianza (supposta finita) di una variabile aleatoria X di cui sia nota la speranza $\mu := E_\theta(X)$ (necessariamente finita) potrà essere stimata dalla *varianza campionaria*

$$T_2(X_1, .., X_n) := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \mu^2. \tag{6.5}$$

Si verifica immediatamente che anche questo stimatore è corretto.

Se invece la speranza $E_\theta(X)$ non è nota, non si può stimare la varianza di X mediante l'analoga variabile aleatoria $\frac{1}{n}\sum_i[X_i - E_\theta(\bar{X}_n)]^2$ $(= \frac{1}{n}\sum_i[X_i - E_\theta(X)]^2)$, per il semplice motivo che questa dipende anche da θ, e quindi non è una statistica. Sembra allora naturale sostituire nell'ultima formula $E_\theta(X)$ con il suo stimatore $\bar{X}_n$, e quindi stimare la varianza di X mediante lo stimatore

$$\tilde{T}_2(X_1, .., X_n) := \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \ \left(= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2\right). \tag{6.6}$$

Questa è una statistica, quindi è uno stimatore della varianza di X; tuttavia non è il migliore stimatore. Un semplice conto [B, p. 127-8] mostra infatti che

$$E_\theta(\tilde{T}_2) = \frac{n-1}{n}\mathrm{Var}_\theta(X); \tag{6.7}$$

lo stimatore $\tilde{T}_2$ quindi non è corretto (comunque è asintoticamente corretto). Lo stimatore

$$\bar{s}_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 \ \left(\neq \frac{1}{n-1}\sum_{i=1}^{n} X_i^2 - (\bar{X}_n)^2\right) \tag{6.8}$$

invece è uno stimatore corretto della varianza di X, poiché la (6.7) fornisce $E_\theta(\bar{s}_n^2) = \mathrm{Var}_\theta(X)$.

La notazione $\bar{s}_n^2$ è standard. Questo stimatore è solitamente denominato *varianza campionaria;* col rischio di creare dei fraintendimenti poiché lo stesso termine è usato per la T_2 (che si può usare solo se è nota la speranza), e ahimè anche per la varianza campionaria introdotta in statistica descrittiva (che si scrive come la $\tilde{T}_2$ [B, p. 7]).

Si può mostrare che

$$\begin{aligned}&\text{gli stimatori } \bar{X}_n \text{ (della speranza) e } \bar{s}_n^2 \text{ (della varianza)}\\ &\text{sono di varianza minima tra gli stimatori corretti,}\\ &\text{e sono consistenti in media quadratica.}\end{aligned} \tag{6.9}$$

La verifica che $\bar{X}_n$ è uno stimatore (della speranza) consistente in media quadratica è immediata.

Legge di Student. [39] La statistica $\bar{s}_n^2$ ha un'importante applicazione, che ora illustriamo.

Se $\{X_1, .., X_n\}$ è un campione aleatorio avente legge normale $N(\mu, \sigma^2)$, allora si verifica immediatamente che

$$Z_n := \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \text{(ovvero } Z_n \text{ è la standardizzata di } \bar{X}_n\text{)}. \tag{6.10}$$

In seguito al teorema limite centrale, questo è approssimativamente valido anche per una vasta classe di altre leggi, se n è abbastanza grande. Come mostrato da [B, p. 126-7], questo permette di stimare la media μ, se la deviazione standard σ è nota.

[39] Questa legge fu proposta nel 1908 da W.S. Gosset, in un articolo che firmò con lo pseudonimo di Student (non si capisce bene cosa c'entrino gli studenti...). Infatti a quei tempi Gosset lavorava come statistico per la nota marca di birra Guinness, e per motivi contrattuali non poteva pubblicare a suo nome sui giornali scientifici.

Se invece σ non è nota, sostituendola con s_n nella (6.10) si ottiene una nuova legge di probabilità, che dipende da n. Gosset studiò le proprietà di questa legge, che denotò $t(n-1)$: la t di Student. Egli dimostrò che, per un campione aleatorio $\{X_1, .., X_n\}$ avente distribuzione normale $N(\mu, \sigma^2)$,

$$\frac{\bar{X}_n - \mu}{s_n/\sqrt{n}} \sim t(n-1) \qquad \forall n \in \mathbf{N}. \tag{6.11}$$

Si noti che qui compare $t(n-1)$ (non $t(n)$); $n-1$ è detto il *numero di gradi di libertà* (per motivi che qui non chiariremo). In proposito si veda [B, p. 128-9].

La legge di Student t (o meglio, la famiglia $\{t(n)\}_{n\in\mathbf{N}}$) è una delle più importanti della statistica. Per questo motivo la sua funzione dei quantili (o equivalentemente, la funzione di ripartizione) è tabulata da [B, p. 134-5], unitamente alla legge normale standard. [40]

Stimatori di Massima Verosimiglianza. Questo metodo ha portata più generale di quello dei momenti. Sia ad esempio X una variabile aleatoria discreta [41] di densità $p_X(\cdot;\theta)$, con $\theta \in \Theta$ parametro incognito. Un campione aleatorio $(X_1, ..., X_n)$ per definizione è costituito da variabili indipendenti equidistribuite, ed ha quindi densità congiunta

$$p_{(X_1,...,X_n)}(x_1, ..., x_n; \theta) = p_X(x_1;\theta) \cdot .. \cdot p_X(x_n;\theta) = \prod_{i=1,...,n} p_X(x_i;\theta). \tag{6.12}$$

Si dice *funzione di verosimiglianza* L (relativa alla densità p_X) questa stessa quantità, considerata come funzione di θ:

$$L(\theta; x_1, ..., x_n) = \prod_{i=1,...,n} p_X(x_i;\theta) \qquad \forall \theta \in \Theta.$$

(Si noti che, per la funzione L, θ è la variabile; $x_1, ..., x_n$ invece sono dei parametri, e a volte sono sottintesi.) [42]

Una statistica T è detta uno *stimatore di massima verosimiglianza* se e solo se $T = T(x_1, ..., x_n)$ massimizza la funzione $\theta \mapsto L(\theta, x_1, ..., x_n)$ per ogni $(x_1, ..., x_n)$ e per ogni n; ovvero se e solo se

$$L(T(x_1, ..., x_n), x_1, ..., x_n) \geq L(\theta, x_1, ..., x_n) \qquad \forall \theta \in \Theta, \forall (x_1, ..., x_n) \in \mathbf{R}^n, \forall n. \tag{6.13}$$

Poiché la funzione logaritmo è strettamente crescente, massimizzare la funzione di verosimiglianza L è equivalente a massimizzarne il logaritmo $\log L$, cosa che spesso è più comoda.

Se la funzione di verosimiglianza $L(\theta, x_1, ..., x_n)$ è differenziabile rispetto a θ $(\in \Theta)$ per ogni $(x_1, ..., x_n) \in \mathbf{R}^n$, allora ogni stimatore di massima verosimiglianza T soddisfa l'*equazione di verosimiglianza:* [43]

$$[\nabla_\theta L(\theta, x_1, ..., x_n)]_{\theta=T(x_1,...,x_n)} = 0 \qquad \forall (x_1, ..., x_n) \in \mathbf{R}^n, \forall n. \tag{6.14}$$

Sostituendo L con $\log L$, è chiaro che si perviene alla stessa equazione. Questa implicazione non è invertibile, poiché un punto stazionario non è necessariamente di massimo.

Si può dimostrare che ogni stimatore di massima verosimiglianza è asintoticamente corretto.

Una Lotteria. Si pensi ad una lotteria con 10 partecipanti $A_1, ..., A_{10}$: 9 prendono un biglietto, uno (che denomineremo il "superacquirente") prende tutti gli altri biglietti — ad esempio 11 se il totale dei biglietti disponibili era 20. Non sappiamo chi sia il superacquirente, quindi dobbiamo formulare 10

[40] La conoscenza di queste tavole non è da ritenersi parte del programma di esame, almeno per i prossimi appelli ...

[41] Questo conto si estende facilmente al caso di variabili aleatorie continue, semplicemente sostituendo la densità discreta p_X con una densità continua f_X.

[42] La funzione di verosimiglianza è spesso indicata con L, poiché in Inglese il termine *verosimiglianza* è tradotto con *likelihood* — o meglio viceversa, poiché questa nozione è stata introdotta dal celebre biologo e statistisco inglese Ronald Fischer. In Inglese likely e likelihood sono rispettivamente sinonimi di probabile e probabilità. Tuttavia quella che in statistica si denomina likelihood non è una misura di probabilità. (Quindi in questo caso è meglio l'Italiano, poiché il termine verosimiglianza è sinonimo di plausibilità, non di probabilità.)

[43] Se θ è uno scalare, allora il gradiente $\nabla_\theta L$ va sostituito con la derivata ordinaria $dL/d\theta$.

modelli statistici alternativi, ciascuno con la sua distribuzione di probabilità. Sappiamo però che A_5 ha poi vinto la lotteria. Si chiede di individuare la distribuzione di probabilità più verosimile, ovvero di congetturare chi possa essere stato il superacquirente. Mostreremo che è ragionevole rispondere che costui sia A_5, ovvero il vincitore.

Si noti la differenza tra:

(i) il problema probabilistico di calcolare le probabilità di vittoria di ciascun partecipante, sapendo quale è il superacquirente,

(ii) il problema statistico di individuare il superacquirente, sapendo chi ha vinto la lotteria.

Entrambi i problemi ammettono una soluzione rigorosa, che però fornisce solo un risultato probabilistico, ovvero fornisce un risultato espresso in termini di probabilità.

Ciascuno dei 10 partecipanti $A_1, ..., A_{10}$ può essere il superacquirente. In modo più formale, posto $\Theta := \{1, ..., 10\}$, ad ogni $\theta \in \Theta$ associamo il modello statistico M_θ: "A_θ ha acquistato 11 biglietti, gli altri ne hanno acquistati 9". Ciascun M_θ individua la distribuzione di probabilità definita come segue: posto $p(i|\theta) :=$ "probabilità di vittoria di A_i condizionata dal fatto che A_θ sia il superacquirente",

$$p(i|\theta) = 11/20 \quad \text{per } i = \theta, \qquad p(i|\theta) = 1/20 \quad \text{per } i \neq \theta. \tag{6.15}$$

(Si noti che la dipendenza dal parametro θ è rappresentata usando la notazione tipica del condizionamento probabilistico.) La funzione di verosimiglianza $L(\theta, i) := p(i|\theta)$ è quindi

$$L(\theta, i) = 11/20 \quad \text{per } \theta = i, \qquad L(\theta, i) = 1/20 \quad \text{per } \theta \neq i. \tag{6.16}$$

Pertanto $L(\cdot, i)$ è massimizzata per $\theta = i$, ovvero attribuendo l'acquisto di 11 biglietti al vincitore. [44]

Quanto è plausibile questa conclusione? Si noti che essa non dipende dal numero di biglietti acquistati dall'ignoto superacquirente: possiamo sostituire 11 con 100 o con 2, il risultato non cambia, anche se la conclusione è molto più plausibile nel primo caso che nel secondo. Infatti, se il superacquirente ha acquistato 100 biglietti, è alquanto verosimile che sia lui vincitore. Se invece ne ha acquistati solo 2, è più verosimile che il vincitore sia uno degli altri 9. Tuttavia, dovendo indicare il superacquirente, non si può rispondere "uno qualsiasi dei non vincitori": dobbiamo sceglierne uno! Ed allora la scelta più ragionevole cade proprio sul vincitore (anche se è proprio quello che avremmo voluto escludere!)

Un Altro Esempio. Si consideri una moneta che ignoriamo se sia equilibrata o meno; al fine di stabilirlo, [45] lanciamo la moneta n volte. L'esito del lancio i-esimo è una variabile aleatoria X avente distribuzione di Bernoulli $\text{Ber}(\theta)$, con $\theta \in \Theta := [0, 1]$:

$$X = 1 \quad \text{se viene testa}, \qquad X = 0 \quad \text{se viene croce}.$$

L'esito degli n lanci è quindi rappresentato da un campione aleatorio $(X_1, ..., X_n)$.

Possiamo stimare θ mediante l'esito del primo lancio, X_1, o più astutamente mediante la media campionaria $\bar{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$. Sia $T_1(X_1, ..., X_n) := X_1$ che $T_2(X_1, ..., X_n) := \bar{X}_n$ sono statistiche (che indicheremo brevemente con T_1 e T_2), poichè sono variabili aleatorie che dipendono solo da $X_1, ..., X_n$ (e non da θ); possono essere considerate degli stimatori di θ, ovvero della speranza della variabili aleatorie X ($\sim\text{Ber}(\theta)$). Effettuare i lanci corrisponde a fissare un $\omega \in \Omega$. I valori ottenuti $X_1(\omega)$ e $\bar{X}_n(\omega)$ sono quindi stime di θ.

Gli stimatori T_1 che T_2 sono non distorti. Gli errori quadratici medi coincidono quindi con le rispettive varianze, che sono diverse:

$$\text{Var}_\theta(T_1) = \theta(1-\theta), \qquad \text{Var}_\theta(T_2) \overset{(3.18)}{=} \frac{1}{n}\text{Var}_\theta(X_1) = \frac{1}{n}\theta(1-\theta).$$

[44] Si noti la presenza di due diversi punti di vista. Dal punto di vista probabilistico, si suppone θ nota e I variabile; si usa quindi la distribuzione di probabilità data dalla (6.15). Dal punto di vista statistico, invece si suppone I nota e θ incognita; entra quindi in gioco la funzione di verosimiglianza (6.16).

[45] o meglio, al fine di congetturarlo. In statistica non si perviene ad alcuna certezza circa i modelli probabilistici, pur essendo le conclusioni precise e rigorose.

Come si vede, lo stimatore T_2 è consistente in media quadratica. Per contro T_1 non è consistente in media quadratica. Lo stimatore T_2 è pertanto da preferirsi a T_1, come peraltro si poteva ben intuire.

Ci chiediamo ora se vi sia uno stimatore di massima verosimiglianza per θ. La densità discreta di $X \sim \text{Ber}(\theta)$ vale

$$p_X(x|\theta) = \begin{cases} \theta & \text{se } x = 1 \\ 1-\theta & \text{se } x = 0 \end{cases} \qquad \forall \theta \in [0,1], \tag{6.17}$$

ovvero

$$p_X(x|\theta) = \theta^x (1-\theta)^{1-x} \qquad \text{per } x \in \{0,1\}, \forall \theta \in [0,1]. \tag{6.18}$$

Pertanto, osservando che $x_1 + ... + x_n = n\bar{X}_n$,

$$L(\theta; x_1, ..., x_n) = \prod_{i=1,...,n} p_X(x_i|\theta) = \theta^{x_1}(1-\theta)^{1-x_1} \cdots \theta^{x_n}(1-\theta)^{1-x_n} = \theta^{n\bar{X}_n}(1-\theta)^{n-n\bar{X}_n},$$

ovvero, passando ai logaritmi,

$$\log L(\theta; x_1, ..., x_n) = n\bar{X}_n \log\theta + n(1-\bar{X}_n)\log(1-\theta) \qquad \forall \theta \in [0,1].$$

Lo stimatore di massima verosimiglianza $T = T(X_1, ..., X_n)$, se esiste, deve soddisfare l'equazione (6.14):

$$\frac{n\bar{X}_n}{T} - \frac{n(1-\bar{X}_n)}{1-T} = 0,$$

e questa ha soluzione $T = \bar{X}_n$. Possiamo concludere che questo è uno stimatore di massima verosimiglianza.

Stime Intervallari. Data una distribuzione di probabilità dipendente da un parametro scalare $\theta \in \mathbf{R}$ incognito, invece di attribuire a θ un valore, possiamo cercare un intervallo $[a,b] \subset \mathbf{R}$ per il quale si possa confidare che contenga θ. In alternativa all'intervallo $]a,b[$, possiamo determinare una semiretta $]-\infty, a[$ oppure $]b, +\infty[$: si parla allora rispettivamente di stima bilatera, stima unilatera sinistra, stima unilatera destra. [46] (Potremmo riunire questi tre casi in uno solo della forma $]a,b[$, se consentissimo ad a, b di assumere anche i valori $\pm\infty$.)

Fissiamo un $\alpha \in]0,1[$ (tipicamente $\alpha = 0.05$, o anche $\alpha = 0.01$, o più raramente $\alpha = 0.001$), [47] ed un intero n, che rappresenterà l'ampiezza del campione. Nel caso della stima bilatera, si dice che due statistiche $\theta_1, \theta_2 : \mathbf{R}^n \to \mathbf{R}$ definiscono una *stima intervallare* $]\theta_1(X_1, ..., X_n), \theta_2(X_1, ..., X_n)[$ *per* θ *al livello di confidenza* $1-\alpha$ se

$$\mathbf{P}_\theta(\theta_1(X_1, ..., X_n) < \theta < \theta_2(X_1, ..., X_n)) = 1 - \alpha. \tag{6.19}$$

Questo corrisponde al taglio di due code, ovvero di $]-\infty, \theta_1(X_1, ..., X_n)] \cup [\theta_2(X_1, ..., X_n), +\infty[$.

Questo significa che, se si potesse ripetere un gran numero di volte il campionamento $(x_1, ..., x_n)$ che ha dato luogo alla stima intervallare, la percentuale dei casi in cui $\theta_1(x_1, ..., x_n) < \theta < \theta_2(x_1, ..., x_n)$ dovrebbe essere vicina a $1-\alpha$. Si usa allora dire che [48]

$$\theta_1(X_1, ..., X_n) < \theta < \theta_2(X_1, ..., X_n) \quad \text{con livello di confidenza } 1-\alpha.$$

Si vede facilmente che, per ogni livello di confidenza $1-\alpha$, esistono infinite statistiche θ_1, θ_2 tali che $\mathbf{P}_\theta(\theta_1 < \theta < \theta_2) = 1-\alpha$. Infatti, per ogni $\theta_1 \in \mathbf{R}$ abbastanza piccolo, esiste un $\theta_2 \in \mathbf{R}$ tale che

[46] Le stime bilatere sono anche dette *a due code,* e quelle unilatere *ad una coda,* per ovvi motivi.

[47] Tradizionalmente, si preferisce usare la notazione $1-\alpha$ con α piccolo, piuttosto che l'equivalente $\beta = 1-\alpha$ con β vicino a 1.

[48] Perché usiamo il termine "livello di confidenza" piuttosto che quello di probabilità? Perché, θ non è una variabile aleatoria, in quanto Θ non è stato dotato di una misura di probabilità. Una volta eseguito il campionamento (ovvero fissato un $\omega \in \Omega$ e trovato $x_1 = X_1(\omega), ..., x_n = X_n(\omega)$), risulta determinato $\bar{\theta}_i(\omega) := \theta_i(x_1, ..., x_n)$ per $i = 1, 2$; quindi $\{\bar{\theta}_1(\omega) < \theta < \bar{\theta}_2(\omega)\}$ non è un evento. Pertanto non ha senso scrivere $\mathbf{P}_\theta(\bar{\theta}_1(\omega) < \theta < \bar{\theta}_2(\omega))$ per un ω fissato.

valga tale uguaglianza. Di solito si scelgono θ_1, θ_2 in modo tale che le due code abbiano area uguale, ovvero

$$\mathbf{P}_\theta(\theta \leq \theta_1) = \mathbf{P}_\theta(\theta_2 \leq \theta) = \alpha/2;$$

quindi $\theta_1 = -\theta_2$ se la distribuzione è simmetrica.

Nel caso di una stima unilatera (ad esempio destra) si individua una statistica $\theta_1 : \mathbf{R}^n \to \mathbf{R}$ tale che

$$\mathbf{P}_\theta(\theta_1(X_1, ..., X_n) < \theta) = 1 - \alpha. \tag{6.20}$$

Questo corrisponde al taglio di una sola coda, ovvero $]-\infty, \theta_1(X_1, ..., X_n)]$. L'intervallo di confidenza unilatero sinistro (o destro) è ovviamente unico.

Più è piccolo l'intervallo di confidenza, maggiore è la *precisione* della stima. Pertanto precisione e confidenza sono esigenze contrapposte. In linea di massima, si possono migliorare entrambe aumentando l'ampiezza del campione — un'operazione che in generale ha un costo.

Distribuzione Pivotale, Distribuzione del χ^2 e Teorema di Cochran. Per la costruzione della stima intervallare di solito si usa una variabile aleatoria la cui distribuzione sia nota (tabulata oppure calcolabile al computer); questa è detta la *distribuzione pivotale* della stima. Nel seguente esempio questo ruolo sarà svolto dalla distribuzione normale standard $N(0, 1)$.

Nel successivo esempio utilizzeremo invece la *legge del* $\chi^2(n)$ [49] per n intero ≥ 1. Questa distribuzione è definita come segue:

$$\text{se } \{Z_1, ..., Z_n\} \text{ è un campione aleatorio avente legge } N(0,1), \text{ allora } \quad \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n). \tag{6.21}$$

Quindi, se $\{X_1, ..., X_n\}$ è un campione aleatorio avente legge $N(\mu, \sigma^2)$, standardizzando le variabili X_i ed applicando la (6.21) si ottiene

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu)^2 \sim \chi^2(n). \tag{6.22}$$

Se la media μ non è nota, la si può sostituire con la media campionaria $\bar{X}_n$, a prezzo di perdere un grado di libertà. Il Teorema di Cochran infatti afferma che, per ogni campione aleatorio $\{X_1, ..., X_n\}$ avente legge normale di varianza σ^2, ponendo come al solito $\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i$,

$$\frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \sim \chi^2(n-1) \qquad \forall n. \tag{6.23}$$

Stime della Media. Sia $(X_1, ..., X_n)$ un campione aleatorio avente legge normale con varianza σ^2 nota; vogliamo stimarne la media μ. La variabile aleatoria $Y_n := (\bar{X}_n - \mu)\sqrt{n}/\sigma$ (che, detto tra parentesi, non è una statistica) ha allora distribuzione normale standard. Fissiamo un livello di confidenza $1 - \alpha$; valori tipici sono $\alpha = 0.05$, oppure $\alpha = 0.01$, o anche $\alpha = 0.001$. Denotando con z la funzione dei quantili della distribuzione normale standard, abbiamo

$$\mathbf{P}(-z_{1-\alpha/2} < Y_n < z_{1-\alpha/2}) = 1 - \alpha,$$

ovvero

$$\mathbf{P}\Big(\bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\Big) = 1 - \alpha. \tag{6.24}$$

[49] Si legge: *chi quadro ad n gradi di libertà.* I quantili di questa importante distribuzione sono tabulati da diversi testi, ma non dal [B].

Questo significa che, osservato un campione $(x_1, ..., x_n) = (X_1(\omega), ..., X_n(\omega))$ e posto $\bar{x}_n := \frac{1}{n}\sum_{i=1}^n x_i$, [50] siamo confidenti allo $1-\alpha$ (ovvero al $100(1-\alpha)\%$...) che valga la seguente stima bilatera per la media μ: [51]

$$\bar{x}_n - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} < \mu < \bar{x}_n + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \qquad \text{ovvero} \qquad |\mu - \bar{x}_n| < z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}. \tag{6.25}$$

Ad esempio, per $\alpha = 0.05$ dalle tavole dei quantili della distribuzione normale standard [B, p. 134] ricaviamo che il quantile $1-\alpha/2 = 0.975$ vale circa 1.96, ovvero $z_{0.975} \simeq 1.96$.

In seguito al Teorema Limite Centrale, questo risultato si applica anche se la variabile aleatoria X non ha distribuzione normale, purché il campione sia sufficientemente ampio: più ampio è il campione, migliore è l'approssimazione.

Se la varianza di X non fosse stata nota, avremmo sostituito la deviazione standard σ con la deviazione standard *campionaria* s_n, e la distribuzione $t(n-1)$ di Student sarebbe stata la nostra distribuzione pivotale. Avremmo quindi dovuto sostituire il quantile normale standard $z_{1-\alpha/2}$ con il corrispondente quantile $t_{1-\alpha/2,n-1}$ di Student (pure tabulato in [B, p. 135]):

$$\mathbf{P}\Big(\bar{X}_n - t_{1-\alpha/2,n-1}\,\frac{s_n}{\sqrt{n}} < \mu < \bar{X}_n + t_{1-\alpha/2,n-1}\,\frac{s_n}{\sqrt{n}}\Big) = 1-\alpha. \tag{6.26}$$

Stime della Varianza. Sia $(X_1, ..., X_n)$ un campione aleatorio avente distribuzione normale di media μ nota; vogliamo stimarne la varianza. Posto ancora $Z_i := (X_i - \mu)/\sigma \sim N(0,1)$, otteniamo

$$\frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n Z_i^2 \overset{(6.21)}{\sim} \chi^2(n). \tag{6.27}$$

Fissiamo un livello di confidenza $1-\alpha$. Denotando $\rho_{\alpha,n}$ (≥ 0) il quantile di ordine α della distribuzione $\chi^2(n)$, abbiamo allora

$$\mathbf{P}\Big(\rho_{\alpha/2,n} < \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \mu)^2 < \rho_{1-\alpha/2,n}\Big) = 1-\alpha,$$

ovvero

$$\mathbf{P}\Big(\frac{\sum_{i=1}^n (X_i-\mu)^2}{\rho_{1-\alpha/2,n}} < \sigma^2 < \frac{\sum_{i=1}^n (X_i-\mu)^2}{\rho_{\alpha/2,n}}\Big) = 1-\alpha. \tag{6.28}$$

Ovviamente questa stima può anche essere espressa in termini della varianza campionaria (con media nota), ovvero di $\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$.

In seguito al Teorema Limite Centrale, questo risultato si applica anche se la variabile aleatoria X non ha distribuzione normale, purché il campione sia sufficientemente ampio.

Se la media μ di X non fosse stata nota, l'avremmo sostituita con la media campionaria $\bar{X}_n$ nella (6.27), perdendo un grado di libertà. Il teorema di Cochran avrebbe infatti fornito

$$\Big(\frac{(n-1)s_n^2}{\sigma^2} =\Big)\ \frac{1}{\sigma^2}\sum_{i=1}^n (X_i - \bar{X}_n)^2 \overset{(6.23)}{\sim} \chi^2(n-1), \tag{6.29}$$

e quindi

$$\mathbf{P}\Big(\frac{\sum_{i=1}^n (X_i-\bar{X}_n)^2}{\rho_{1-\alpha/2,n-1}} < \sigma^2 < \frac{\sum_{i=1}^n (X_i-\bar{X}_n)^2}{\rho_{\alpha/2,n-1}}\Big) = 1-\alpha. \tag{6.30}$$

Sintesi. Riassumendo, queste sono le distribuzioni pivotali che abbiamo utilizzato per stimare la media o la varianza di un campione normale, oppure di un campione tanto ampio da consentire l'uso del Teorema Limite Centrale:

[50] Solitamente si usano lettere maiuscole per indicare le variabili aleatorie, e quelle minuscole per rappresentare i corrispondenti campionamenti. Quindi ad esempio i valori assunti dalla variabile aleatoria X sono indicati con x.

[51] Con scrittura più ingegneristica che matematica, qualcuno scriverebbe $\mu = \bar{x}_n \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$.

Stima della media con varianza nota: distribuzioni pivotale $N(0,1)$, cf. (6.24).
Stima della media con varianza incognita: distribuzioni pivotale $t(n-1)$, cf. (6.26).
Stima della varianza con media nota: distribuzioni pivotale $\chi(n)$, cf. (6.28).
Stima della varianza con media incognita: distribuzioni pivotale $\chi(n-1)$, cf. (6.30).

7 • Verifica di Ipotesi

Parole chiave: Ipotesi nulla ed alternativa. Tipi di errore. Significatività. p-value.

Ipotesi e Regione Critica. La verifica di ipotesi è un altro classico problema di statistica inferenziale. Tale verifica consiste

$$\begin{array}{l}\text{nello stabilire non se una data ipotesi è vera,}\\ \text{ma se è compatibile con i dati a disposizione.}\end{array} \tag{7.1}$$

La corrispondente procedura è denominata un *test;* questo sfocia nella decisione di accettare o meno l'ipotesi data; Questo rientra in quella che è anche detta *teoria della decisione.*

La progettazione di un test si fonda sui seguenti passi:

(i) la formulazione di un'ipotesi di partenza, detta *ipotesi nulla,* riguardante la distribuzione di probabilità di una variabile aleatoria (scalare o vettoriale). La sua negazione è denominata *ipotesi alternativa.* Indicheremo le due ipotesi rispettivamente con H_0 e H_A; [52]

(ii) l'individuazione di un'opportuna statistica campionaria U, detta la *statistica del test;* (o *statistica-test);*

(iii) la determinazione di una *regione critica* C, ovvero di un insieme di valori che può assumere T, in base al quale si perviene alla seguente *regola di decisione:*

$$\begin{array}{l}\text{si rifiuta } H_0 \text{ se } U(X_1,...,X_n)\in C,\\ \text{si accetta } H_0 \text{ se } U(X_1,...,X_n)\notin C.\end{array} \tag{7.2}$$

(La regione di accettazione è quindi il complementare in Ω della regione critica C.)

Questo significa che, effettuato un campionamento, ovvero scelto un $\omega\in\Omega$,

$$\begin{array}{l}\text{si rifiuta } H_0 \text{ se } U(X_1(\omega),...,X_n(\omega))\in C,\\ \text{si accetta } H_0 \text{ se } U(X_1(\omega),...,X_n(\omega))\notin C.\end{array} \tag{7.3}$$

Anche se dal punto di vista formale le ipotesi H_0 e H_A giocano ruoli simmetrici, in realtà tra di esse sussiste una fondamentale differenza concettuale. Si effettua un test solo se si ha motivo di ritenere che l'ipotesi H_0 possa essere contraddetta dai fatti. Infatti, se il campione non smentirà nettamente H_0, non si giungerà ad alcuna conclusione statisticamente significativa. [53] Questo ovviamente richiede una certa cura nella formulazione dell'ipotesi H_0 da testare.

Ad esempio, in un sistema penale garantista, si processa qualcuno solo se ci sono dei forti motivi per ritenere che possa essere colpevole (non si pesca uno a caso e gli si dice "adesso dimostrami di non essere colpevole"). Almeno in linea di principio, si condanna un imputato solo se le prove a lui avverse sono schiaccianti — aldilà di ogni ragionevole dubbio, come si usa dire. Quindi l'ipotesi H_0 sarà "l'imputato è innocente", e di conseguenza H_A sarà "l'imputato è colpevole" — e non viceversa.

[52] Con riferimento alle ipotesi del test, si distingue tra *test parametrici* e *test non parametrici.* Nei primi l'ipotesi H_0 riguarda uno o più valori della distribuzione di probabilità; ad esempio la sua media e/o la sua varianza. Nei secondi H_0 riguarda altre proprietà della distribuzione; ad esempio, la distribuzione è normale? (nel caso di una distribuzione vettoriale) sono le sue componenti indipendenti? ecc..

[53] In tal caso sarebbe pertanto più preciso dire che "non si rifiuta H_0", e non che "si accetta H_A". Questa è solo una sfumatura, ma coglie il senso del discorso.

[54] Analogamente, si mette un farmaco in commercio solo se si è convinti che esso sia efficace e non dannoso, in modo da evitare danni al portafogli o alla salute. Una commissione ministeriale effettua quindi un test in cui si assume come ipotesi H_0 "il farmaco è non efficace o è dannoso" (o anche entrambe, naturalmente); di conseguenza H_A è "il farmaco è efficace e non è dannoso" — e non viceversa. Si potrebbe trattare analogamente il problema della validità di un progetto edilizio: le varianti sono infinite.

Un test è considerato significativo solo se porta al rifiuto di H_0; in caso contrario è considerato poco più che inutile. In questo senso, ad esempio, uscire assolti da un processo penale non significa che si è ritenuti innocenti, ma semplicemente che le prove non sono state ritenute schiaccianti.

Assumeremo che l'ipotesi H_0 determini completamente la distribuzione di probabilità; in tal caso si dice che l'ipotesi H_0 è *semplice*. Ad esempio, questo vale per H_0: "$\theta = 1/2$", ma non per H_A: "$\theta < 1/2$". Questa limitazione ci permetterà di semplificare la trattazione.

Errori e Livello di Significatività. Ogni test può comportare due tipi di errore:

$$\begin{array}{ll} \text{errore del I tipo:} & \text{rifiutare } H_0 \text{ quando } H_0 \text{ è vera} \\ & \text{(ovvero, accettare } H_A \text{ quando } H_A \text{ è falsa),} \\ \text{errore del II tipo:} & \text{accettare } H_0 \text{ quando } H_0 \text{ è falsa} \\ & \text{(ovvero, rifiutare } H_A \text{ quando } H_A \text{ è vera).} \end{array} \tag{7.4}$$

Per quanto detto, un errore del I tipo (ad esempio, condannare un innocente) è considerato più grave di uno del II tipo, e la progettazione del test tiene conto dell'esigenza di ridurre per quanto possibile tale rischio. A parità di ampiezza del campione, si può ridurre l'errore di un tipo solo aumentando quello dell'altro tipo: occorre quindi giungere ad un punto di equilibrio tra queste due esigenze. Si possono comunque ridurre entrambi gli errori aumentando l'ampiezza del campione; ma questo ha un costo.

Occorre ora tradurre questi intenti in precisi concetti quantitativi. Allo scopo di contenere il rischio di errore del I tipo, si fissa un *livello di significatività* $\alpha \in]0,1[$, e si definisce la regione critica del test in modo che la probabilità di commettere un errore del I tipo sia pari ad α, ovvero (usando la probabilità condizionata)

$$\begin{aligned} \alpha = \mathbf{P}(\text{errore del I tipo}) &\overset{(7.4)}{=} \mathbf{P}(\text{rifiutare } H_0 \mid H_0) \\ &\overset{(7.2)}{=} \mathbf{P}(U(X_1, ..., X_n) \in C \mid H_0). \end{aligned} \tag{7.5}$$

Più α è piccolo, più il test risulta significativo, poiché minore è la probabilità che H_0 venga respinta, e quindi è più alto il contenuto di informazione dei dati quando ciò avviene. Si noti il paradosso terminologico:

più il livello di significatività di un test è basso, più il test è ritenuto significativo.

(Questa è la terminologia corrente ... sorry.) [55] In letteratura si definisce anche il seguente parametro:

$$\begin{aligned} \beta = \mathbf{P}(\text{errore del II tipo}) &\overset{(7.4)}{=} \mathbf{P}(\text{accettare } H_0 \mid H_A) \\ &\overset{(7.2)}{=} \mathbf{P}(U(X_1, ..., X_n) \notin C \mid H_A). \end{aligned} \tag{7.6}$$

[54] Un saldo principio giuridico prescrive appunto che ogni imputato sia da considerare innocente fino a prova contraria. Nelle aule dei tribunali il giudizio su un imputato non è mai affidato ad un test. Tuttavia l'analogia è senz'altro calzante, ed aiuta a comprendere l'asimmetria delle due ipotesi alla base dei test.

[55] Per le stime intervallari si parla di *livello di confidenza* $1 - \alpha$, mentre per i test si usa il *livello di significatività* α. Si noti che in entrambi i casi α è scelto il più piccolo possibile. Malgrado le analogie, i due concetti vanno comunque ben distinti.

Come già osservato per il livello di confidenza, il livello di significatività non è una probabilità. L'ipotesi H_0 è vera o falsa, ad essa quindi non può essere associata alcuna probabilità.

La quantità $1-\beta$ è poi detta la *potenza del test;* questa ovviamente è la probabilità di rifiutare H_0 quando questa ipotesi è falsa. Nella scelta di un test, si cerca di minimizzare la significatività α e di massimizzarne la potenza; queste due esigenze sono contrastanti, pertanto occorre individuare un punto di equilibrio. Dal momento che l'errore del I tipo è ritenuto più grave di quello del II tipo, spesso prima si fissa α, e successivamente si cerca di individuare un test che massimizzi la potenza (ovvero minimizzi β).

Un Esempio. Diversi test sono costruiti ricollegandosi alle stime intervallari; la regione critica è spesso rappresentata da una o due code. Riprendiamo l'esempio già visto di una variabile aleatoria X avente distribuzione normale di media μ incognita e varianza σ^2 nota. Poniamo $Y_n(\omega,\mu) := (\bar{X}_n(\omega) - \mu)\sqrt{n}/\sigma$ [56] mettendo in evidenza anche la dipendenza dall'incognita $\mu \in \mathbf{R}$. Come sappiamo, $Y_n(\cdot,\mu)$ ha distribuzione normale standard. Fissato un valore μ_0, consideriamo le ipotesi

$$H_0: \ \mu = \mu_0, \qquad H_A: \ \mu \neq \mu_0, \tag{7.7}$$

e fissiamo un livello di significatività $\alpha \in]0,1[$. Per ogni $\beta \in]0,1[$, sia z_β il quantile di ordine β della distribuzione normale standard. Si noti che

$$\mathbf{P}(|Y_n(\cdot,\mu)| \leq z_{1-\alpha/2} \mid H_0) = \mathbf{P}(|Y_n(\cdot,\mu_0)| \leq z_{1-\alpha/2});$$

inoltre, in seguito a [B, p. 98], per ogni variabile aleatoria Z,

$$\mathbf{P}(|Z| \leq z_{1-\alpha/2}) = \mathbf{P}(Z \leq z_{1-\alpha}) = 1-\alpha.$$

Pertanto

$$\mathbf{P}(|Y_n(\cdot,\mu)| > z_{1-\alpha/2} \mid H_0) = 1 - \mathbf{P}(|Y_n(\cdot,\mu_0)| \leq z_{1-\alpha/2}) = 1-(1-\alpha) = \alpha.$$

Effettuato un campionamento [57] si respinge quindi l'ipotesi H_0 (ovvero si accetta l'ipotesi H_A) se e solo se $|Y_n(\omega,\mu_0)| > z_{1-\alpha/2}$; ovvero, posto $\bar{x}_n = \bar{X}_n(\omega)$,

$$\begin{array}{c} \text{si rifiuta } H_0 \text{ al livello di significatività } \alpha \quad \Leftrightarrow \\ \mu_0 < \bar{x}_n - z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \quad \text{oppure} \quad \mu_0 > \bar{x}_n + z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}\,. \end{array} \tag{7.8}$$

Ad esempio per $\alpha = 0.05$, $z_{1-\alpha/2} = z_{0.975}$ ($\simeq 1.96$ [B, p. 134]); pertanto

$$\begin{array}{c} \text{si rifiuta } H_0 \text{ al livello di significatività } 0.05 \quad \Leftrightarrow \\ \mu_0 < \bar{x}_n - z_{0.975}\,\frac{\sigma}{\sqrt{n}} \quad \text{oppure} \quad \mu_0 > \bar{x}_n + z_{0.975}\,\frac{\sigma}{\sqrt{n}}\,. \end{array} \tag{7.9}$$

La regione critica è quindi l'unione di due semirette:

$$C = \Big]-\infty, \bar{x}_n - z_{0.975}\,\frac{\sigma}{\sqrt{n}}\Big[\cup \Big]\bar{x}_n + z_{0.975}\,\frac{\sigma}{\sqrt{n}}, +\infty\Big[.$$

Nel caso in cui la varianza non fosse stata nota, allora avremmo proceduto come abbiamo visto alla fine del paragrafo precedente. Ovvero, avremmo sostituito la deviazione standard σ con la deviazione standard *campionaria* s_n, ed il quantile normale $z_{1-\alpha/2}$ con il corrispondente quantile della distribuzione $t(n-1)$ di Student (pure tabulato in [B, p. 135]). Quest'ultima quindi sarebbe stata la nostra distribuzione pivotale.

Se la legge campionata non fosse stata normale, ma il campione fosse stato comunque abbastanza ampio, grazie al teorema limite centrale avremmo potuto approssimare quella legge con una normale, e quindi utilizzare la procedura testé descritta.

[56] Questa non è una statistica. Tuttavia sostituiremo μ con un valore $\mu_0 \in \mathbf{R}$, ottenendo la statistica $Y_n(\omega,\mu_0)$.

[57] Come sappiamo, nel formalismo probabilistico-statistico questo è rappresentato dalla scelta di un $\omega \in \Omega$.

Infine, ricordiamo che per α fissato c'è un'ampia scelta degli intervalli relativi alle stime bilatere, anche se ovviamente l'intervallo simmetrico è unico. Pertanto la scelta della regione critica C del test non è necessariamente unica, nemmeno se la statistica del test ed il livello di significatività sono fissati. Si giunge quindi alla seguente conclusione un po' paradossale: si possono progettare due test diversi che hanno la stessa statistica-test e lo stesso livello di significatività, e che nondimeno, a fronte dello stesso campionamento, forniscono due decisioni opposte (!) [58]

p-Value. La procedura che abbiamo appena descritta consiste nel prescrivere a priori il livello di significatività α, e poi respingere o meno l'ipotesi H_0 in base ai valori assunti dal campione — o meglio dalla statistica del test valutata sui valori campionati. Si può impostare la significatività di un test anche ribaltando questo punto di vista: in alternativa si può prima valutare la statistica sul campione selezionato, e su questa base individuare per quali livelli di significatività l'ipotesi H_0 verrebbe respinta.

Si definisce allora il concetto di *p-value* relativo ad un campionamento della statistica del test $U(X_1, ..., X_n)$. [59] (Come sappiamo, il campionamento è individuato dalla scelta di un $\omega \in \Omega$.) Per ogni livello di significatività α, indichiamo con C_α una regione critica che soddisfi la relazione (7.9), e poniamo

$$\begin{aligned} p\text{-value} &= \sup \{\alpha \in]0,1[\colon U(X_1(\omega), ..., X_n(\omega)) \notin C_\alpha\} \\ &(= \inf \{\alpha \in]0,1[\colon U(X_1(\omega), ..., X_n(\omega)) \in C_\alpha\}). \end{aligned} \tag{7.10}$$

In altri termini, il p-value è l'estremo superiore dei livelli di significatività α per cui si accetta l'ipotesi H_0, e questo coincide con l'estremo inferiore degli α per cui si rifiuta H_0.

Più piccolo è il p-value, più grande è l'evidenza sperimentale che l'ipotesi H_0 sia falsa. Il p-value può essere quindi interpretato come una misura delle prove avverse all'ipotesi H_0, relative al risultato del campionamento.

Ad esempio, torniamo al problema della stima della media di una variabile aleatoria X avente distribuzione normale di media μ incognita e varianza σ^2 nota. Vogliamo definire un test per verificare se $\mu = \mu_0$, cf. (7.7). Precedentemente abbiamo scelto un livello di significatività α, abbiamo effettuato un campionamento, e ne abbiamo tratte le conseguenze. Adesso in base alla (7.10) calcoliamo il p-value relativo al campionamento; in questo modo possiamo delineare un quadro più preciso della situazione.

Consideriamo ad esempio i tre casi seguenti:

(i) p-value $= 0.049$. Allora al livello di significatività $\alpha = 0.05$ rifiutiamo H_0, mentre al livello $\alpha = 0.01$ non rifiutiamo tale ipotesi. Dal momento che 0.049 è prossimo a 0.05 e distante da 0.01, per $\alpha = 0.05$ il rifiuto di H_0 dovrebbe essere (per così dire...) poco convinto, mentre per $\alpha = 0.01$ l'accettazione di H_0 (o, a voler essere precisi, il non rifiuto di H_0) dovrebbe essere più convinta.

(ii) p-value $= 0.011$. Come nel caso (i), al livello di significatività $\alpha = 0.05$ rifiutiamo H_0, mentre al livello $\alpha = 0.01$ non rifiutiamo tale ipotesi. Dal momento che 0.011 è distante da 0.05 e prossimo a 0.01, per $\alpha = 0.05$ il rifiuto di H_0 dovrebbe essere convinto, mentre per $\alpha = 0.01$ l'accettazione di H_0 dovrebbe essere poco convinta.

(iii) p-value $= 0.009$. Rifiutiamo H_0 sia al livello di significatività $\alpha = 0.05$ che a quello $\alpha = 0.01$. Dal momento che 0.009 è distante da 0.05 e prossimo a 0.01, il rifiuto di H_0 dovrebbe essere convinto per $\alpha = 0.05$, mentre per $\alpha = 0.01$ dovrebbe essere meno convinto.

Queste differenze qualitative sono l'essenza del significato del p-value. Anche qui, onestà vorrebbe che si scegliesse il livello di significatività prima di aver visto il p-value.

Nella pratica, calcolare il p-value non è più oneroso che decidere l'esito del test per un prescritto livello di significatività α, e l'uso del p-value permette di fornire una rappresentazione più precisa del quadro statistico. Tuttavia stabilire a priori il livello di significatività può presentare dei vantaggi;

[58] Ci possono essere dei motivi statistici per scegliere un test piuttosto che l'altro. Onestà vorrebbe che uno facesse la scelta del test prima di aver visto il risultato...

[59] Il p-value è anche detto *livello di significatività osservato;* per distinguerlo da questo, l'α precedentemente introdotto è allora anche detto *livello di significatività prescritto.*

[60] ad esempio, permette di giungere ad una decisione circa l'esito del test, evitando le esitazioni che potrebbero presentarsi in seguito all'uso del p-value. Quest'ultimo strumento è appunto meno decisionale e più conoscitivo, e lo scopo di un test vuole essere soprattuto decisionale.

Per quel che possono valere gli aggettivi, diversi sperimentatori usano questa terminologia: [61]

$$\begin{aligned} &\text{gli indizi sono molto fortemente a sfavore di } H_0 && \text{se } p\text{-value} < 0.01, \\ &\text{gli indizi sono fortemente a sfavore di } H_0 && \text{se } p\text{-value} \in]0.01, 0.05[, \\ &\text{gli indizi sono debolmente a sfavore di } H_0 && \text{se } p\text{-value} \in]0.05, 0.1[. \end{aligned} \tag{7.11}$$

Si noti che un alto p-value può essere un debole indizio contro H_0 (però se è troppo alto non è nemmeno un indizio!), ma non può mai essere un *forte* indizio in favore di H_0. Infatti nessun test può concludersi *nettamente* in favore di H_0, ovvero nettamente contro H_A.

Test o Stime? Test e stime rispondono ad esigenze diverse:

$$\begin{aligned} &\text{le stime sono descrittive: forniscono una conoscenza;} \\ &\text{i test sono decisionali: danno luogo ad una decisione.} \end{aligned} \tag{7.12}$$

In diversi casi si può scegliere tra una stima intervallare ed un test; l'uso del p-value rappresenta una sorta di compromesso tra queste due esigenze.

I test possono anche presentare il seguente vantaggio rispetto alle stime intervallari: essi offrono la possibilità di utilizzare l'ipotesi H_0 nei calcoli. Come è mostrato dal seguente esempio, questo può essere utile specialmente quando l'ipotesi H_0 è semplice (ovvero non composta).

Alla fine della precedente sezione, abbiamo illustrato dei procedimenti per la stima intervallare della media e della varianza di una distribuzione normale. Per quanto già osservato, queste stime danno luogo a corrispondenti test sul valore della media e della varianza.

8 Tolleranza

Le dimensioni dei manufatti prodotti industrialmente si discostano necessariamente da un prescritto valore nominale. Per *tolleranza* si intende la massima deviazione consentita dalla norma. Nella produzione di manufatti composti da più elementi, si presenta la necessità di valutare la variazione del manufatto assemblato M, sulla base delle variazioni dei componenti $M_1, ..., M_N$ *(analisi dell'errore).*

Consideriamo il caso in cui una quantità X relativa ad M sia la somma delle corrispondenti quantità dei componenti: $X = \sum_{i=1}^{N} X_i$. Ad esempio, gli X_i potrebbero rappresentare delle lunghezze, dei pesi, delle concentrazioni, delle resistenze elettriche, ecc.. [62] Poiché ciascun X_i è affetto da un errore δ_i (positivo o negativo), pure X sarà affetto dall'errore $\delta = \sum_{i=1}^{N} \delta_i$. Supponiamo che $|\delta_i| \leq \varepsilon_i$, quest'ultima essendo la tolleranza della componente i-esima. Si pone l'esigenza di maggiorare $|\delta| = |\sum_{i=1}^{N} \delta_i|$. Vi sono due modi principali di valutare δ:

(i) un metodo deterministico, in cui si presuppone il peggior caso possibile:

$$|\delta| = \left| \sum_{i=1}^{N} \delta_i \right| \leq \sum_{i=1}^{N} |\delta_i| \leq \sum_{i=1}^{N} \varepsilon_i =: \varepsilon'; \tag{8.1}$$

(ii) un metodo probabilistico, che qui illustriamo brevemente.

[60] Accontentarsi della grossolana distinzione tra *"buono e non buono"* può essere più comodo che formulare un giudizio qualitativo più raffinato. La comodità comunque non dovrebbe essere il solo criterio di scelta.

[61] Questo richiede una precisazione. Il termine *indizio* è una maldestra traduzione dell'inglese *evidence,* che viene usato in statistica inferenziale. In senso giuridico evidence significa prova o testimonianza, e quindi è sostanzialmente diverso da indizio, oltre ad essere più forte. D'altra parte non sarebbe corretto tradurre evidence con *evidenza,* che in Italiano ha un altro significato.

[62] Il problema dell'analisi dell'errore e della tolleranza si presenta in ogni produzione industriale.

Questo procedimento è basato sulla rappresentazione degli errori $\{\delta_i : i = 1, ..., N\}$ come variabili aleatorie reali, centrate e dotate di media finita $\{\mu_i : i = 1, ..., N\}$ e varianza finita $\{\sigma_i^2 : i = 1, ..., N\}$. Lo stesso quindi vale per δ, che avrà media finita $\mu = \sum_{i=1}^N \mu_i$ e varianza incognita ma finita σ^2.

Ciascuna varianza σ_i^2 è poi messa in relazione con la rispettiva tolleranza ε_i; si può porre ad esempio $\varepsilon_i = K\sigma_i$, per un'opportuna costante $K > 0$. [63] Infine si suppone che

$$\text{gli errori } \{\delta_i : i = 1, ..., N\} \text{ siano stocasticamente indipendenti. } [64] \tag{8.2}$$

Sotto queste ipotesi, il *principio di mutua compensazione* fornisce

$$\sigma^2 \le \sum_{i=1}^N \sigma_i^2 = \frac{1}{K^2} \sum_{i=1}^N \varepsilon_i^2 \quad \text{ovvero} \quad \varepsilon'' := K\sigma \le \Big(\sum_{i=1}^N \varepsilon_i^2 \Big)^{1/2}. \tag{8.3}$$

Ad esempio, se gli ε_i sono uniformi, ovvero $\varepsilon_i = \varepsilon_1$ per $i = 1, ..., N$,

$$\varepsilon' = \sum_{i=1}^N \varepsilon_i = N\varepsilon_1, \qquad \varepsilon'' \le \Big(\sum_{i=1}^N \varepsilon_i^2 \Big)^{1/2} = \sqrt{N}\, \varepsilon_1. \tag{8.4}$$

Pertanto $\varepsilon' = \sqrt{N}\, \varepsilon''$, indipendentemente da K e dalle distribuzioni assunta per gli errori $\{\delta_i : i = 1, ..., N\}$. Quindi [65]

$$\varepsilon'' < \varepsilon' \quad \text{per ogni } N, \qquad \varepsilon'' << \varepsilon' \quad \text{per } N \text{ grande.} \tag{8.5}$$

Quando applicabile, l'approccio probabilistico pertanto fornisce una stima più stringente della tolleranza. Questo metodo può offrire diversi vantaggi, anche se richiede un'analisi più sofisticata.

Quanto sopra illustrato si inquadra nel tema dell'*analisi dell'errore,* che va ben aldilà del problema della tolleranza.

Un'Applicazione al Calcolo Numerico. Un computer calcola in parallelo con grande precisione un gran numero di quantità reali $\{A_i\}_{i=1,...,N}$. Supponiamo quindi che esista un piccolo $\delta > 0$ tale che, denotando ε_i l'errore commesso nel calcolo di A_i, sia $|\varepsilon_i| \le \delta$ per ogni i. Ci si chiede come si possa stimare l'errore complessivo ε commesso nel calcolo di $A := \sum_{i=1}^N A_i$.

Per $i = 1, ..., N$, interpreteremo gli errori ε_i come realizzazioni di variabili aleatorie X_i (ovvero, $\varepsilon_i = X_i(\omega)$ per ogni $\omega \in \Omega$), aventi deviazione standard $\sigma_i \le \delta$. L'errore complessivo ε risulta quindi la realizzazione della variabile aleatoria $X := \sum_{i=1}^N X_i$ avente deviazione standard σ da valutare. Dal momento che le operazioni sono eseguite in parallelo, supporremo che le variabili X_i siano stocasticamente indipendenti. Allora, per il principio di compensazione, $\sigma^2 = \sum_{i=1}^N \sigma_i^2$ $(\le N\delta^2)$.

Mentre l'approccio deterministico conduce a considerare l'errore massimo possibile, ovvero $|\varepsilon| \le N\delta =: e_{max}$, l'impostazione probabilistica fornisce $\sigma \le \sqrt{N}\delta$. Quindi

$$\sigma \le \frac{e_{max}}{\sqrt{N}} << e_{max} \qquad \text{per } N \text{ grande.} \tag{8.6}$$

[63] Ad esempio, se si suppone che ogni errore δ_i sia distribuito normalmente, ovvero $\delta_i \sim N(0, \sigma_i^2)$ per $i = 1, ..., N$, ponendo ad esempio $\varepsilon_i = 3\sigma_i$, si tollerano erroneamente meno dello 0.3% dei manufatti. Se questa percentuale è ritenuta eccessiva, si può ovviamente scegliere un K più grande.

Si noti che l'ipotesi di distribuzione normale dei δ_i non è necessaria per il presente discorso sulla tolleranza. Tra l'altro, tipicamente l'errore δ_i è soggetto alla maggiorazione $|\delta_i| \le \varepsilon_i$, che sembra sottintendere una distribuzione uniforme.

[64] Questa ipotesi è cruciale. Ad esempio, i pezzi prodotti da una macchina possono essere affetti da due tipi di errore:

(a) un errore deterministico sistematico, ovvero uniforme per tutti i pezzi; questo corrisponde al caso limite di una variabile aleatoria avente varianza nulla e media non nulla (ovvero una variabile non aleatoria);

(b) un errore casuale a media nulla, che varia da pezzo a pezzo.

Gli errori del primo tipo non sono stocasticamente indipendenti, e quindi vanno trattati con il metodo deterministico (i).

[65] Nella pratica, a fronte di un prescritta tolleranza del manufatto assemblato, ha interesse calcolare la tolleranza che è sufficiente richiedere per i componenti.

Se poi supponiamo che l'errore X abbia distribuzione gaussiana, allora ad esempio in oltre il 99% dei casi $|\varepsilon| \leq 3\sigma$, quindi

$$|\varepsilon| \leq 3\sigma \leq \frac{3}{\sqrt{N}} e_{max} << e_{max} \qquad \text{per } N \text{ grande.} \tag{8.7}$$

9 Affidabilità

Quanto segue sviluppa quanto esposto da [B, pp. 96, 101-103]; si veda anche [B, p. 59]. Si tratta di nozioni di ovvio interesse ingegneristico.

Affidabilità ed Inaffidabilità. Il funzionamento di un'apparato D (in un certo periodo di tempo) può essere rappresentato da una variabile aleatoria X definita sullo spazio Ω dei possibili stati del sistema:

$$X(\omega) = 1 \quad \text{se } D \text{ funziona in } \omega, \qquad X(\omega) = 0 \quad \text{se } D \text{ non funziona in } \omega.$$

Si definisce *affidabilità (reliability* in Inglese) di un dispositivo funzionante D (o di una sua componente) la probabilità R che esso funzioni (come richiesto e sotto specifiche condizioni ambientali, almeno per il periodo di tempo prescritto). Si definisce *inaffidabilità (unreliability* in Inglese) di un dispositivo D la probabilità $U = 1 - R$ che esso non funzioni. Essendo X la funzione indicatrice degli stati del sistema in cui D funziona, abbiamo

$$R = \mathbf{P}(X = 1) = \mathbf{E}(X), \qquad U = \mathbf{P}(X = 0) = 1 - \mathbf{E}(X).$$

Struttura di Affidabilità. Si dice che un sistema ha *struttura di affidabilità in serie* se il guasto (ovvero il mancato funzionamento) di una qualsiasi sua componente comporta il non funzionamento dell'intero sistema. Si dice invece che ha *struttura di affidabilità in parallelo* se il mancato funzionamento dell'intero sistema richiede il guasto di tutte le sue componenti. [66]

Sia D un dispositivo costituito da un insieme di componenti $D_1, ..., D_N$ indipendenti; con questo si intende che le rispettive funzioni di funzionamento X_i costituiscono una famiglia di variabili aleatorie stocasticamente indipendenti. Siano R l'affidabilità di D, e R_i quella di D_i per $i = 1, ..., N$; le corrispondenti *inaffidabilità* allora sono $U = 1 - R$ e $U_i = 1 - R_i$ per ogni i. L'indipendenza delle componenti fornisce le seguenti leggi fondamentali:

$$\begin{array}{ll} \text{legge dell'affidabilità per sistemi in serie:} & R = R_1 \cdot ... \cdot R_N, \\ \text{legge dell'inaffidabilità per sistemi in parallelo:} & U = U_1 \cdot ... \cdot U_N. \end{array} \tag{9.1}$$

Definiamo *durata di vita* di un dispositivo l'istante $V = V(\omega)$ in cui esso cessa di funzionare; queste è una variabile aleatoria. Se un sistema consiste di n componenti aventi durata di vita $V_1, ..., V_N$, allora per la durata di vita V dell'intero sistema abbiamo

$$\begin{array}{ll} \text{per sistemi in serie:} & V(\omega) = \min\{V_1(\omega), ..., V_N(\omega)\}, \\ \text{per sistemi in parallelo:} & V(\omega) = \max\{V_1(\omega), ..., V_N(\omega)\}, \end{array} \qquad \text{per } \omega \in \Omega. \tag{9.2}$$

Funzione di Affidabilità. Sopra abbiamo definito l'affidabilità per un tempo di funzionamento prescritto. Se invece tale tempo è variabile, allora si definisce la *funzione di sopravvivenza* R di V, ovvero $R(t) = \mathbf{P}(V \geq t)$ per ogni $t > 0$. Questa è detta *funzione di affidabilità,* e rappresenta la probabilità che un dispositivo (o una sua componente) funzioni almeno fino al tempo $t > 0$, supposto che sia funzionante all'istante 0. Questa ovviamente è una funzione non crescente di t. Si noti che la

[66] Questi concetti non hanno nulla a che vedere con le combinazioni in serie ed in parallelo proprie dei modelli reologici e circuitali, che pure sono ampiamente usate in ingegneria.

funzione di inaffidabilità $U(t)$ $(= 1 - R(t)) = \mathbf{P}(V < t)$ coincide con la funzione di ripartizione della durata di vita V.

Se supponiamo che i dispositivi siano stocasticamente indipendenti, allora le leggi (9.1) e (9.2) sono ovviamente generalizzate come segue:

$$\begin{array}{ll} \text{per sistemi in serie:} & R(t) = R_1(t) \cdot ... \cdot R_N(t) \\ \text{per sistemi in parallelo:} & U(t) = U_1(t) \cdot ... \cdot U_N(t) \end{array} \qquad \forall t > 0. \tag{9.3}$$

$$\begin{array}{ll} \text{per sistemi in serie:} & V(t,\omega) = \min\{V_1(t,\omega), ..., V_N(t,\omega)\} \\ \text{per sistemi in parallelo:} & V(t,\omega) = \max\{V_1(t,\omega), ..., V_N(t,\omega)\} \end{array} \qquad \forall t > 0, \text{per } \omega \in \Omega. \tag{9.4}$$

L'applicazione ripetuta di queste semplici regole permette di estenderle a strutture ben più generali delle combinazioni in serie e in parallelo: ad esempio combinazioni in serie di più dispositivi, ciascuno consistente in una combinazione in parallelo di componenti, ecc..

*** Tasso di Guasto.** Per la definizione di probabilità condizionata, la probabilità che un dispositivo funzionante all'istante $t > 0$ subisca un guasto in un intervallo di tempo $[t, t + \Delta t]$ è pari a

$$\frac{\mathbf{P}(t < V \leq t + \Delta t)}{\mathbf{P}(t < V)} = \frac{\mathbf{P}(t < V) - \mathbf{P}(t + \Delta t < V)}{\mathbf{P}(t < V)} = \frac{R(t) - R(t + \Delta t)}{R(t)}.$$

Il tasso medio di guasto in tale intervallo temporale è quindi $[R(t) - R(t + \Delta t)]/[\Delta t\, R(t)]$. Passando al limite per $\Delta t \to 0$, si perviene quindi al

$$\textit{tasso istantaneo di guasto:} \quad Z(t) = -\frac{R'(t)}{R(t)} \ (\geq 0) \quad \forall t > 0, \tag{9.5}$$

a cui corrisponde, integrando questa semplice equazione differenziale e ponendo $R(0) = 1$, la

$$\textit{funzione di affidabilità:} \quad R(t) = e^{-\int_0^t Z(\tau)\, d\tau} \quad \forall t > 0. \tag{9.6}$$

Come già osservato, $U(t)$ $(= 1 - R(t)) = \mathbf{P}(V < t)$ coincide con la funzione di ripartizione della durata di vita V; la corrispondente densità di probabilità è il

$$\textit{tempo di guasto:} \quad f(t) = -R'(t) = Z(t) e^{-\int_0^t Z(\tau)\, d\tau} \quad \forall t > 0. \tag{9.7}$$

$*$ **Legge di Weibull.** A volte la funzione Z viene assunta costante, il che corrisponde ad un tempo di guasto $f(t)$ esponenziale:

$$f(t) = Z e^{-Zt} \quad \forall t > 0. \tag{9.8}$$

In tal caso i tempi di attesa tra due diversi guasti hanno una distribuzione di Poisson. In questo modo però si escludono effetti di memoria, in particolare si trascura l'usura del dispositivo (quindi l'usato sarebbe come il nuovo...). In diversi casi è più realistico supporre che inizialmente Z decresca (un dispositivo rodato è più affidabile di uno nuovo...), resti poi pressoché costante per un lungo tratto, ed infine cresca In una certa misura, questo si applica anche al tasso di mortalità umana. Questo è rappresentato dalla curva detta *a vasca da bagno* si veda la figura di [B, p. 103]. Un tempo di guasto spesso utilizzato è la

$$\textit{legge di Weibull:} \quad f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta} \quad \text{per } t > 0, \text{ con } \alpha, \beta \text{ parametri } > 0. \tag{9.9}$$

Questa corrisponde al tasso istantaneo di guasto $Z(t) = \alpha\beta t^{\beta-1}$, che è decrescente per $0 < \beta < 1$ e crescente per $\beta > 1$.

Si definisce anche *durata media* di un dispositivo il tempo medio di guasto:

$$\mathbf{E}(T) = \int_0^{+\infty} t f(t)\, dt = \int_0^{+\infty} R(t)\, dt. \tag{9.10}$$

10 * Un Problema di Controllo Stocastico

Supponiamo di voler produrre pezzi di lunghezza nominale L^* mediante una macchina che produce pezzi di lunghezza effettiva

$$X_n = L + \delta_n \qquad \forall n \geq 0, \tag{10.1}$$

con L inaccessibile, ma con δ_n che possiamo controllare (ad esempio regolando una vite micrometrica). Distingueremo il caso deterministico da quello stocastico.

Caso Deterministico. Per un qualsiasi δ_0, produciamo un primo pezzo di lunghezza $X_0 = L + \delta_0$. Misuriamo X_0, imponiamo

$$\delta_1 = \delta_0 + L^* - X_0 \ \ (= L^* - L\ !), \tag{10.2}$$

e produciamo un secondo pezzo di lunghezza

$$X_1 = L + \delta_1 = L + \delta_0 + L^* - X_0 = L^*. \tag{10.3}$$

Proseguiamo ponendo

$$\delta_n = \delta_{n-1} + L^* - X_{n-1}\ (= \delta_1) \qquad \forall n \geq 1. \tag{10.4}$$

e quindi ottendendo pezzi di lunghezza $X_n = L^*$ per ogni $n \geq 1$.

Caso Stocastico. Supponiamo ora che la produzione di ogni pezzo sia soggetto ad un errore, che rappresentiamo mediante una variabile aleatoria W_n:

$$X_n = L + \delta_n + W_n \qquad \forall n \geq 0. \tag{10.5}$$

Supponiamo anche che le W_n siano indipendenti ed equidistribuite, con

$$\mathbf{E}(W_n) = 0, \qquad \mathrm{Var}(W_n) = \sigma^2 \qquad \forall n \geq 0. \tag{10.6}$$

Se il valore di L fosse accessibile, potremmo porre $\delta_n = L^* - L$, ottenendo $\delta_n = L^* + W_n$, e quindi

$$\mathbf{E}(X_n) = L^*, \qquad \mathrm{Var}(X_n) = \sigma^2 \qquad \forall n \geq 0. \tag{10.7}$$

Questo risultato sarebbe ottimale, poiché non si può scendere sotto la varianza del "rumore" W.

Ma, essendo il valore di L inaccessibile, possiamo procedere in tre modi diversi:

(i) Possiamo ancora definire

$$\delta_1 = \delta_0 + L^* - X_0 \ \ (= L^* - L - W_0)$$

come in (10.2), e $\delta_n = \delta_1$ per ogni $n \geq 1$. Otteniamo

$$X_n = L + \delta_n + W_n = L + \delta_0 + L^* - X_0 + W_n = L^* - W_0 + W_n \qquad \forall n \geq 1. \tag{10.8}$$

Quindi

$$\mathbf{E}(X_n) = L^*, \qquad \mathrm{Var}(X_n) = 2\sigma^2 \qquad \forall n \geq 0.$$

(iii) Per ogni pezzo possiamo cercare di compensare l'errore del pezzo precedente. In analogia con la (10.4), poniamo allora

$$\delta_n = \delta_{n-1} + L^* - X_{n-1} \qquad \forall n \geq 1, \tag{10.9}$$

ottenendo

$$X_n = L + \delta_{n-1} + L^* - X_{n-1} + W_n \overset{(10.1)}{=} L^* - W_{n-1} + W_n \qquad \forall n \geq 0. \tag{10.10}$$

Quindi anche in questo caso

$$\mathbf{E}(X_n) = L^*, \qquad \mathrm{Var}(X_n) = 2\sigma^2 \qquad \forall n \geq 0.$$

(iii) Ad ogni pezzo possiamo cercare di compensare l'errore utilizzando l'informazione fornita da tutti i pezzi precedenti.

$$\delta_n = \delta_{n-1} + L^* - \frac{1}{n}\sum_{k=1}^{n} X_{k-1} \qquad \forall n \geq 1. \tag{10.11}$$

Questo fornisce

$$\begin{aligned} X_n &= L + \delta_n = L + \delta_{n-1} + L^* - \frac{1}{n}\sum_{k=1}^{n} X_{k-1} + W_n \\ &\overset{(10.1)}{=} L^* - \frac{1}{n}\sum_{k=1}^{n} W_{k-1} + W_n \qquad \forall n \geq 0. \end{aligned} \tag{10.12}$$

Per il principio di compensazione e per la (10.1), $\mathrm{Var}\,(\frac{1}{n}\sum_{k=1}^{n} W_{k-1}) = \sigma^2/n$; quindi

$$\mathbf{E}(X_n) = L^*, \qquad \mathrm{Var}(X_n) = (1 + \tfrac{1}{n})\sigma^2 \qquad \forall n \geq 0. \tag{10.13}$$

Conclusioni.

La procedura (i) è la meno onerosa, poiché richiede solo una misurazione ed una regolazione.

La procedura (ii) è ben più onerosa, poiché richiede una misurazione ed una regolazione per ogni pezzo prodotto, ma non presenta vantaggi rispetto alla (i).

La procedura (iii) è tanto onerosa quanto la (ii), ma al limite per $n \to \infty$ riduce la varianza al minimo irriducibile σ^2.

La procedura (ii) è quindi svantaggiosa rispetto alle altre due; si dice che presenta un *eccesso di controllo.*

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

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

Stima di parametri. Stimatori e stime puntuali. Stimatori di massima verosimiglianza. Stime intervallari. Confidenza.

Test di ipotesi. Ipotesi nulla ed alternativa. Tipi di errore. Significatività. p-value.

Principali Leggi Esaminate.

Leggi discrete finite: legge uniforme, di Bernoulli, binomiale, ipergeometrica:

$$\mathrm{Unif}(n), \quad \mathrm{Ber}(p)\ (= \mathrm{Bin}(1,p)), \quad \mathrm{Bin}(n,p), \quad \mathrm{Iper}(r,b,n).$$

Leggi discrete infinite: legge geometrica, di Poisson:

$$\mathrm{Geo}(p), \quad \mathrm{Poi}(\lambda).$$

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

$$\mathrm{Unif}([a,b]), \quad \mathrm{Exp}(\lambda), \quad \mathrm{N}(\mu,\sigma^2).$$

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 variabile aleatoria X, ciascuna di queste probabilità individua poi una legge.

12 Esercizi

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

(a) 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?

(b) 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 $2\varphi(X)+2$.

Risoluzione. Definiamo i seguenti eventi:

F: si è scelto il primo cassetto, S: si è scelto il secondo cassetto,

B: la biglia estratta è bianca, N: la biglia estratta è nera.

Quindi $\mathbf{P}(F) = \mathbf{P}(S) = 1/2$, $\mathbf{P}(N|F) = 2/5$, $\mathbf{P}(N|S) = 1/3$.

(a) Grazie alla formula della probabilità totale,

$\mathbf{P}(N) = \mathbf{P}(N|F)\cdot\mathbf{P}(F) + \mathbf{P}(N|S)\cdot\mathbf{P}(S) = (2/5)\cdot 1/2 + (1/3)\cdot 1/2 = 11/30$.

(Pertanto $\mathbf{P}(B) = 1 - 11/30 = 19/30$.)

(b) $\mathbf{E}(\varphi(X)) = 2\mathbf{P}(N) + 3\mathbf{P}(B) = 79/11$; $\mathbf{E}(2\varphi(X)+2) = 2\mathbf{E}(\varphi(X)) + 2 = 158/11 + 2$.

$\mathrm{Var}(\varphi(X)) = 4\mathbf{P}(N) + 9\mathbf{P}(B) - \mathbf{E}(\varphi(X))^2 = ...$; $\mathrm{Var}(2\varphi(X)+2) = 4\,\mathrm{Var}(\varphi(X)) = ...$

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

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

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

Risoluzione. (a) La probabilità che la terza biglia sia rossa è la stessa che la prima biglia sia rossa, ovvero 2/5.

(b) La probabilità la prima e la terza biglia estratta siano dello stesso colore è la stessa che lo siano le prime due biglie estratte. Definiamo i seguenti eventi:

BB: la prima e la seconda biglia estratte sono bianche,

RR: la prima e la seconda biglia estratte sono rosse.

Allora $\mathbf{P}(BB) = (4/10)\cdot(3/9) = 4/30$, $\mathbf{P}(RR) = (6/10)\cdot(5/9) = 1/3$. Pertanto

$$\mathbf{P}(BB) + \mathbf{P}(RR) = (4/30) + (1/3) = 7/15.$$

— **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.

Risoluzione. (i) L'evento indicato corrisponde a otto successi (testa) e due insuccessi (croce) in dieci esperimenti bernoulliani di parametro p, quindi ha probabilità $p^8(1-p)^2$.

(ii) L'evento indicato corrisponde a otto successi in uno schema bernoulliano di dieci prove di parametro p, quindi ha probabilità

$$C_8^{10}p^8(1-p)^2 = C_2^{10}p^8(1-p)^2 = \frac{10\cdot 9}{2}\cdot p^8(1-p)^2.$$

(iii) Definiamo i seguenti eventi:

PT: le monete pari danno testa, PC: le monete pari danno croce,

DT: le monete dispari danno testa, DC: le monete dispari danno croce.

Abbiamo $\mathbf{P}(PT) = \mathbf{P}(DT) = p^5$, $\mathbf{P}(PC) = \mathbf{P}(DC) = (1-p)^5$.

L'evento indicato ha probabilità

$$\begin{aligned}&\mathbf{P}(PT)\cdot\mathbf{P}(DT) + \mathbf{P}(PC)\cdot\mathbf{P}(DC) + \mathbf{P}(PT)\cdot\mathbf{P}(DC) + \mathbf{P}(PC)\cdot\mathbf{P}(DT)\\ &= p^{10} + (1-p)^{10} + 2p^5\cdot(1-p)^5.\end{aligned} \tag{12.1}$$

— **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?

Risoluzione. (i) Poiché ad ogni lancio la probabilità di ottenere un numero pari è $1/2$, per un dado equilibrato la probabilità di ottenere tre numeri pari è $1/2^3$.

(ii) Ad ogni lancio, la probabilità di ottenere un numero pari è $q_2 + q_4 + q_6$. Quindi la probabilità di ottenere tre numeri pari è $(q_2 + q_4 + q_6)^3$.

(iii) La probabilità di ottenere tre numeri pari diversi nel caso di un dado equilibrato è $(3/6)\cdot(2/6)\cdot(1/6) = 1/36$.

— **Esercizio 5.** Estraggo una dopo l'altra 6 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 3 estrazioni vanno bene, qual è la probabilità di vittoria?

Si risponda distinguendo tra questi due casi:

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

Risoluzione. Definiamo i seguenti eventi, per $n = 1, 2, ..., 6$:

N_n: la n-sima carta estratta è nera, R_n: la n-sima carta estratta è rossa.

(a) Poiché le estrazioni sono indipendenti e ad ogni estrazione la probabilità di estrarre la carta del colore giusto è pari a $20/40 = 1/2$, la probabilità di vittoria è $1/2^3$.

(b) La probabilità di vittoria partendo con una carta nera è

$$\mathbf{P}(R_4 \cap N_5 \cap R_6 \mid N_1 \cap R_2 \cap N_3) = (19/37)\cdot(18/36)\cdot(17/35).$$

Lo si vede facilmente anche senza bisogno di calcolare la probabilità condizionata, poiché

dopo $N_1 \cap R_2 \cap N_3$ restano 19 carte rosse su 37,

dopo $N_1 \cap R_2 \cap N_3 \cap R_4$ restano 18 carte nere su 36,

dopo $N_1 \cap R_2 \cap N_3 \cap R_4 \cap N_5$ restano 17 carte rosse su 35.

La probabilità di vittoria partendo con una carta rossa è la stessa.

La probabilità di vittoria è quindi $2 \cdot (19/37) \cdot (18/36) \cdot (18/35)$.

— **Esercizio 6.** Si estraggano successivamente (senza reimmissione) due biglie da un'urna contenente 5 biglie bianche e 3 nere. Sia definita una "funzione premio" nel seguente modo: l'estrazione di ogni biglia bianca viene premiata con 2 E, e quella di ogni biglia nera con 1 E. Sia X_i il premio riscosso in base alla i-esima estrazione.

(a) Qual è la probabilità che le due biglie estratte abbiano colore diverso?

(b) Si calcoli la legge congiunta di X_1 e X_2.

(c) Si calcoli il valore atteso di $X_1 + X_2$.

Risoluzione. Definiamo i seguenti eventi:

BN: la prima biglia estratta è bianca, la seconda è nera,

BB: la prima biglia estratta è bianca, la seconda è bianca,

NN: la prima biglia estratta è nera, la seconda è nera,

NB: la prima biglia estratta è nera, la seconda è bianca.

(a) $\mathbf{P}(BN) = 15/56$, $\mathbf{P}(NB) = 15/56$. Pertanto $\mathbf{P}(BN) + \mathbf{P}(NB) = 15/28$.

(b) $\mathbf{P}(BN) = 15/56$, $\mathbf{P}(NB) = 15/56$, $\mathbf{P}(BB) = 10/28$, $\mathbf{P}(NN) = 3/28$.

(c) $\mathbf{E}(X_1+X_2) = 3\mathbf{P}(BN)+3\mathbf{P}(NB)+4\mathbf{P}(BB)+2\mathbf{P}(NN) = 3\cdot 15/56+3\cdot 15/56+4\cdot 10/28+2\cdot 3/28$.

— **Esercizio 7.** Ho probabilità 2/5 di incontrare Aldo (da solo o con altri), e probabilità 1/5 di incontrare sia Aldo che Bruno. Assumendo che l'incontro con Aldo sia indipendente da quello con Bruno,

(a) che probabilità ho di incontrare Bruno?

(b) incontrando Aldo, che probabilità ho di incontrare anche Bruno?

(c) se incontro Aldo, questi mi restituisce 1 Euro (che mi doveva); se incontro Bruno, sono io che gli restituisco 2 Euro (che gli dovevo). Se sono uscito con 10 Euro, alla fine qual è il valore atteso che mi dovrei trovare in tasca?

Risoluzione. Definiamo i seguenti eventi:

A: incontro Aldo, B: incontro Bruno.

Abbiamo $\mathbf{P}(A) = 2/5$, $\mathbf{P}(A \cap B) = 1/5$.

(a) In seguito all'ipotesi di indipendenza, $\mathbf{P}(A \cap B) = \mathbf{P}(A) \cdot \mathbf{P}(B)$. Quindi $\mathbf{P}(B) = 1/2$.

(b) Inoltre $\mathbf{P}(B|A) = \mathbf{P}(A \cap B)/\mathbf{P}(A) = (1/5)/(1/2) = 2/5$.

(c) "Incasso" atteso: $1\mathbf{P}(A) + (-2)\mathbf{P}(B) = 2/5 - 1 = -3/5$.

Valore atteso residuo $10 - 3/5 = 47/5$.

— **Esercizio 8.** Da un cassetto contenente 4 monete d'oro e 3 d'argento si estraggono successivamente tutte le monete (senza reimmissione!).

(a) Qual è la probabilità che la sesta moneta estratta sia d'oro?

(b) L'esito della quinta estrazione è indipendente dall'esito della prima? Vale il viceversa?

(c) Sapendo che la prime due monete estratte sono di metalli diversi, qual è la probabilità che la prima e la quinta siano dello stesso metallo?

(d) Sapendo che la moneta d'oro vale 10 E e quella d'argento 2 E, qual è il valore atteso corrispondente all'estrazione delle ultime due monete?

Risoluzione. (a) La probabilità che la sesta moneta estratta sia d'oro è la stessa che la prima moneta estratta sia d'oro (anche se non cè reimmissione!). Quindi vale 4/7.

(b) L'esito della quinta estrazione non è indipendente dall'esito della prima. Lo stesso vale per il viceversa, poiché la relazione di indipendenza è simmetrica.

(c) Definiamo i seguenti eventi:

OO: la prima e la seconda moneta estratte sono d'oro, $\mathbf{P}(OO) = (4/7) \cdot (3/6) = 12/42$,

AA: la prima e la seconda moneta estratte sono d'argento, $\mathbf{P}(AA) = (3/7) \cdot (2/6) = 1/7$,

OA: la prima moneta estratta è d'oro, la seconda è d'argento, $\mathbf{P}(OA) = (4/7) \cdot (3/6) = 12/42$,

AO: la prima moneta estratta è d'argento, la seconda è d'oro, $\mathbf{P}(AO) = (3/7) \cdot (4/6) = 12/42$,

OAO: la prima e la quinta moneta estratte sono d'oro, la seconda d'argento,
AOA: la prima e la quinta moneta estratte sono d'argento, la seconda d'oro.

$$P(OAO) = (4/7) \cdot (3/6) \cdot (3/5) = 6/35, \quad \mathbf{P}(AOA) = (3/7) \cdot (4/6) \cdot (2/5) = 12/105.$$

Abbiamo

$$\begin{aligned} \mathbf{P}(OAO|OA \cup AO) &= \frac{\mathbf{P}(OAO \cap (OA \cup AO))}{\mathbf{P}(OA \cup AO)} = \frac{\mathbf{P}(OAO)}{\mathbf{P}(OA) + \mathbf{P}(AO)} \\ \mathbf{P}(AOA|OA \cup AO) &= \frac{\mathbf{P}(AOA \cap (OA \cup AO))}{\mathbf{P}(OA \cup AO)} = \frac{\mathbf{P}(AOA)}{\mathbf{P}(OA) + \mathbf{P}(AO)} \end{aligned} \tag{12.2}$$

Pertanto

$$\mathbf{P}(OAO \cup AOA|OA \cup AO) = \frac{\mathbf{P}(AOA)}{\mathbf{P}(OA) + \mathbf{P}(AO)} + \frac{\mathbf{P}(AOA)}{\mathbf{P}(OA) + \mathbf{P}(AO)} = ... \tag{12.3}$$

(d) È equivalente calcolare il valore atteso relativo alle prime due estrazioni piuttosto che alle ultime due. Il valore atteso quindi vale

$$20 \cdot \mathbf{P}(OO) + 4 \cdot \mathbf{P}(AA) + 12 \cdot \mathbf{P}(OA) + 12 \cdot \mathbf{P}(AO) = ...$$

— **Esercizio 9.** (i) Un test d'esame consta di dieci domande, per ciascuna delle quali si deve scegliere l'unica risposta corretta fra quattro alternative. Quale è la probabilità che, rispondendo a caso alle dieci domande, almeno due risposte risultino corrette? [Maturità scientifica PNI del 2011]

(ii) Se almeno quattro risposte risultano corrette, si riceve un premio di 8 Euro. Rispondendo a caso alle dieci domande, quanto vale il "guadagno atteso?

(iii) Se invece si ricevono 2 Euro per ogni risposta corretta e si perde 1 Euro per ogni risposta scorretta, rispondendo a caso alle dieci domande, quanto vale il "guadagno atteso?

Risoluzione. (i) L'evento C: "almeno due risposte risultano corrette" è il complementare dell'evento S: "9 o 10 risposte risultano scorrette". Quindi

$$\mathbf{P}(C) = 1 - \mathbf{P}(S) = 1 - \sum_{k=0,1} C_k^{10} \cdot (3/4)^k \cdot (1/4)^{10-k} = 1 - (3/4)^{10} - 10 \cdot (3/4) \cdot (1/4)^9 = ...$$

(ii) Analogamente a quanto appena calcolato, la probabilità che almeno quattro risposte risultino corrette vale $1 - \sum_{k=0}^{3} C_k^{10} \cdot (3/4)^k \cdot (1/4)^{10-k}$. Il guadagno atteso è quindi pari a

$$8 \cdot \Big(1 - \sum_{k=0}^{3} C_k^{10} \cdot (3/4)^k \cdot (1/4)^{10-k}\Big).$$

(iii) Il guadagno atteso per ogni domanda è pari a $2 \cdot (1/4) - 1 \cdot (3/4) = -1/4$ E.
II guadagno atteso totale è quindi $10 \cdot (-1/4) = -5/2$ E.

— **Esercizio 10.** Dalle statistiche risulta che solo un terzo degli studenti di medicina civile ed ambientale passa l'esame di Anatomia II (An. II) entro il secondo anno, solo un quarto degli studenti si laurea restando in corso, e che il 70 % degli studenti che si laureano restando in corso hanno passato An. II entro il secondo anno.

(a) Che probabilità ha di laurearsi restando in corso uno studente che passa l'esame di An. II entro il secondo anno?

(b) Ci sono dei premi: 100 E per passare l'esame di An. II entro il secondo anno, 500 E per laurearsi restando in corso. Si calcoli il valore atteso di "guadagno per ogni matricola", distinguendo il caso in cui i premi sono cumulabili da quello in cui non lo sono.

Risoluzione. (1) *Definizioni e Dati.*
Fissiamo un generico studente S, e definiamo i seguenti eventi:

A: S passa l'esame di An. II entro il secondo anno,
L: S si laurea restando in corso.
Quindi $\mathbf{P}(A) = 1/3$, $\mathbf{P}(L) = 1/4$, $\mathbf{P}(A|L) = 7/10$,
(2) *Risoluzione.*
(a) $\mathbf{P}(L|A) = \mathbf{P}(A|L) \cdot \mathbf{P}(L)/\mathbf{P}(A) = (7/10) \cdot (1/4)/(1/3) = 21/40$.
(b) Il guadagno atteso con cumulo è pari a

$$500 \cdot \mathbf{P}(L) + 100 \cdot \mathbf{P}(A) = 500/4 + 100/3 = 1900/12\ E.$$

Poiché $\mathbf{P}(A \cap L) = \mathbf{P}(A|L) \cdot \mathbf{P}(L) = (7/10) \cdot (1/4) = 7/40$, il guadagno atteso senza cumulo è pari a

$$\begin{aligned} 500 \cdot \mathbf{P}(L) + 100 \cdot [\mathbf{P}(A \setminus L)] &= 500 \cdot \mathbf{P}(L) + 100 \cdot [\mathbf{P}(A) - \mathbf{P}(A \cap L)] \\ &= 500/4 + 100/3 - 100 \cdot (7/40) = 1690/12\ E. \end{aligned} \tag{12.4}$$

(3) *Conclusioni.*
(a) $\mathbf{P}(L|A) = 21/40$.
(b) Guadagno atteso con cumulo: 1900/12 E. Guadagno atteso senza cumulo: 1690/12 E.

— **Esercizio 11.** Nel gioco del lotto vengono estratti successivamente 5 numeri (tra 90) senza reimmissione.
(a) Qual è la probabilità che 7 sia il secondo numero estratto?
(b) Sapendo che 7 non è stato il primo numero estratto, qual è la probabilità che venga estratto come secondo numero?
(c) La probabilità di estrarre il numero 7 aumenta dopo ogni estrazione mancata?
(d) Sapendo che il primo numero estratto è stato pari, qual è la probabilità che 7 sia estratto come secondo numero?
(e) Sapendo che il primo numero estratto è stato dispari, qual è la probabilità che 7 sia estratto come secondo numero?
(f) Sapendo che il primo numero estratto è risultato ≤ 30, qual è la probabilità che 7 sia estratto come secondo numero?

Risoluzione. (a) Per ciascun numero, la probabilità di essere estratto come primo numero è pari a quella di essere estratto come settimo numero, ovvero 1/90.
(b) Dal momento che dopo la prima estrazione restano 89 numeri, la probabilità è 1/89.
(c) Sì: dopo l'n-esima estrazione mancata vale $1/(90 - n)$.
(d) Sapendo che il primo estratto è stato un pari, la probabilità è 1/89.
(e) Definiamo gli eventi
D: il primo estratto è stato un dispari; E: il primo estratto è stato un dispari diverso da 7;
S: il secondo estratto è il 7.
Osserviamo che $S \cap D = S\cap$ e che E ed S sono indipendenti. Quindi

$$\begin{aligned} P(S|D) = P(S \cap D)/P(D) &= P(S \cap E)/P(D) = P(S) \cdot P(E)/P(D) \\ &= (1/89) \cdot (44/90)/(1/2) = 44/(89 \cdot 45). \end{aligned} \tag{12.5}$$

(f) Procedendo in modo analogo al caso (e), si ottiene $P = (29/30) \cdot (1/89)$.

— **Esercizio 12.** Da una scatola contenente 6 biglie rosse e 4 bianche si estraggono successivamente tre biglie (senza reimmissione). Sapendo che la prime due biglie estratte sono dello stesso colore, qual è la probabilità che anche la terza sia dello stesso colore?

Risoluzione. La probabilità richiesta vale

$$\begin{aligned}
&P((R,R,R)\cup(B,B,B)|(R,R)\cup(B,B))\\
&=P((R,R,R)|(R,R)\cup(B,B))+P((B,B,B)|(R,R)\cup(B,B))\\
&=P((R,R,R))/P((R,R)\cup(B,B))+P((B,B,B))/P((R,R)\cup(B,B))\\
&=P((R,R,R)/[P((R,R)+P(B,B))]+P((B,B,B))/[P((R,R)+P(B,B))]\\
&=\Big(\tfrac{6}{10}\cdot\tfrac{5}{9}\cdot\tfrac{4}{8}+\tfrac{4}{10}\cdot\tfrac{3}{9}\cdot\tfrac{2}{8}\Big)\Big/\Big(\tfrac{6}{10}\cdot\tfrac{5}{9}+\tfrac{4}{10}\cdot\tfrac{3}{9}\Big).
\end{aligned}\tag{12.6}$$

In alternativa si poteva usare la formula di Bayes:

$$\begin{aligned}
&P(3\text{ uguali}|(2\text{ uguali})=P(2\text{ uguali}|3\text{ uguali})\cdot P(3\text{ uguali})/P(2\text{ uguali})\\
&=1\cdot P(3\text{ uguali})/P(2\text{ uguali})=P((R,R,R)\cup(B,B,B))/P((R,R)\cup(B,B)),
\end{aligned}\tag{12.7}$$

e poi procedere come sopra.

— **Esercizio 13.** Un libro di n pagine presenta m refusi. Sia X il numero di refusi a pag. 10. È più probabile $X=2$ o $X=1$?

Risoluzione. X è una variabile aleatoria con distribuzione Bernoulliana: $X\sim Ber(m,1/n)$. Quindi

$$P(X=1)=m(1/n)\cdot[1-(1/n)]^{m-1},\qquad P(X=2)=[m(m-1)/2]\cdot(1/n)^2\cdot[1-(1/n)]^{m-2}.$$

Pertanto $P(X=1)/P(X=2)=2(n-1)/[(m-1)]\simeq 2n/m$ per n,m grandi.

— Per n,m grandi, posto $\lambda:=m/n$, si può approssimare X con una variabile aleatoria Y avente distribuzione di Poisson: $Y\sim Poi(\lambda)$. Quindi

$$P(X=1)=e^{-\lambda}\lambda^1/1,\qquad P(X=2)=e^{-\lambda}\lambda^2/2.$$

Pertanto $P(X=1)/P(X=2)=2/\lambda=2n/m$.

Osservazione per $m=2n$: per la distribuzione Bernoulliana prevale di poco $X=2$; la distribuzione di Poisson invece li dà alla pari (!).

— **Esercizio 14.** Metà dei gemelli eterozigoti hanno lo stesso sesso, mentre tutti i gemelli omozigoti hanno lo stesso sesso. Due terzi dei gemelli hanno lo stesso sesso.
(i) Si calcoli la percentuale dei gemelli omozigoti.
(i) Si calcoli la probabilità che due gemelli dello stesso sesso siano omozigoti.

Risoluzione. (a) Si definiscano i seguenti eventi:
O: i gemelli sono omozigoti;
E: i gemelli sono eterozigoti;
E: i gemelli hanno lo stesso sesso.
Allora $P(S|O)=1$, $P(S|E)=1/2$, $P(S)=2/3$. Si noti che

$$P(S)=P(S|O)\cdot P(O)+P(S|E)\cdot P(E)=P(O)+(1/2)\cdot[1-P(O)];$$

quindi $P(O)=1/3$.
(b) La formula di Bayes fornisce $P(O|S)=P(S|O)\cdot P(O)/P(S)=(1/3)/(2/3)=1/2$.

— **Esercizio 15.** Supponiamo che un campione radioattivo contenga $M=2\cdot 10^{10}$ nuclidi, ciascuno dei quali ha la probabilità $p=10^{-10}$ di decadere in un secondo. Calcolare:
(i) il numero atteso di decadimenti in un secondo,
(ii) la probabilità di osservare più di 4 decadimenti in un secondo.

Risoluzione. (i) Il numero N di nuclidi che decade in un secondo può essere rappresentato mediante una variabile aleatoria binomiale $N \sim Bin(M, p)$. Come noto, la variabile N può essere approssimata da una variabile aleatoria X di Poisson: $N \simeq X \sim Poi(\lambda)$, con $\lambda = Mp = 2$.

Il numero atteso di decadimenti in un secondo quindi vale $E(N) \simeq E(X) = \lambda = 2$.

(ii) La probabilità di osservare più di 4 decadimenti in un secondo è

$$P(N > 4) \simeq P(X > 4) = e^{-2} \sum_{k=5}^{\infty} \frac{2^k}{k!} = 1 - e^{-2} \sum_{k=0}^{4} \frac{2^k}{k!} = ...$$

I prossimi quesiti riguardano l'indipendenza delle variabili aleatorie; in proposito si vedano anche gli esempi discussi nella sezione 3.

— **Esercizio 16.** (a) Siano $X, Y : \Omega \to \mathbf{R}$ due variabili aleatorie. Si può calcolare la legge di $Z_1 = X - Y$? e quella di $Z_2 = XY$?

(b) Supponiamo ora che siano note soltanto le leggi $\mathbf{P}_X, \mathbf{P}_Y$. Si pongono allora le stesse due domande della parte (a).

(c) Ancora supponendo che siano note soltanto le leggi $\mathbf{P}_X, \mathbf{P}_Y$, e supponendo anche che X e Y siano limitate [in modo che non ci siano problemi di convergenza delle speranze] si può calcolare la speranza di Z_1? e quella di Z_2?

Risoluzione. (a) Sì per entrambe le domande, poiché è nota la variabile aleatoria congiunta (X, Y).

(b) No per entrambe le domande, poiché non è nota la legge congiunta $\mathbf{P}_{(X,Y)}$, ovvero la legge della variabile aleatoria congiunta (X, Y).

(c) Si può calcolare la speranza di $Z_1 = X - Y$, poichè $E(X - Y) = E(X) - E(Y)$. Invece per calcolare la speranza di Z_2 occorrerebbe conoscere la legge congiunta $\mathbf{P}_{(X,Y)}$.

— **Esercizio 17.** (a) Siano $X, Y : \Omega \to \mathbf{R}$ due variabili aleatorie limitate [in modo che non ci siano problemi di convergenza delle speranze]. Si può calcolare la varianza di $Z_1 = X - Y$? e quella di $Z_2 = XY$? e la covarianza di Z_1 e Z_2?

(b) Supponiamo ora che siano note soltanto le leggi $\mathbf{P}_X, \mathbf{P}_Y$. Si può calcolare la varianza di $Z_1 = X - Y$? e quella di $Z_2 = XY$?

(c) Stessa domanda di (b), supponendo che X e Y siano indipendenti.

(d) Nelle ipotesi di (b), si può calcolare la speranza di $Z_2 = e^{XY}$?

Risoluzione. (a) Sì per tutte e tre le domande. Poiché, essendo nota la variabile aleatoria congiunta (X, Y), si ricava facilmente la legge congiunta di Z_1 e Z_2.

(b) No per entrambe le domande, poiché non è nota la legge congiunta $\mathbf{P}_{(X,Y)}$.

(c) Si può calcolare la varianza di Z_1, poiché per via dell'indipendenza di X e Y

$$\mathrm{Var}(Z_1) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X, Y) = \mathrm{Var}(X) + \mathrm{Var}(Y),$$

Si può calcolare la varianza di Z_2. Infatti l'indipendenza di X e Y implica quella di X^2 e Y^2, [67] e quindi

$$\mathrm{Var}(Z_2) = \mathbf{E}(X^2 \cdot Y^2) - \mathbf{E}(X \cdot Y)^2 = \mathbf{E}(X^2) \cdot \mathbf{E}(Y^2) - \mathbf{E}(X)^2 \cdot \mathbf{E}(Y)^2;$$

per calcolare queste medie bastano le leggi di X e Y (senza bisogno della legge congiunta).

Per lo stesso motivo si può calcolare anche la covarianza di Z_1 e Z_2:

$$\mathrm{Cov}(Z_1, Z_2) = \mathbf{E}(Z_1^2 \cdot Z_2^2) - \mathbf{E}(Z_1 \cdot Z_2)^2 = \mathbf{E}(X^2 \cdot Y^2) - \mathbf{E}(X^2 \cdot Y - X \cdot Y^2)^2$$

$$= \mathbf{E}(X^2) \cdot \mathbf{E}(Y^2) - [\mathbf{E}(X^2) \cdot \mathbf{E}(Y) - \mathbf{E}(X) \cdot \mathbf{E}(Y^2)]^2.$$

(d) Si può calcolare la speranza di $Z_2 = e^{XY}$, poiché $Z_2 = e^X \cdot e^Y$, e le variabili aleatorie e^X e e^Y sono indipendenti, dal momento che lo sono X e Y.

[67] L'importante Proposizione 3.19 del [B; p. 56] si estende anche alle variabili aleatorie continue.

— **Esercizio 18.** (a) Siano $X, Y : \Omega \to \mathbf{R}$ due variabili aleatorie discrete limitate [in modo che non ci siano problemi di convergenza delle speranze] e $g : \mathbf{R} \to \mathbf{R}$ una funzione continua.

(a) Si può calcolare la speranza di $Z_1 = g(X) + Y$? e quella di $Z_2 = g(X)Y$?

(b) Si può calcolare la varianza di Z_1? e quella di Z_2?

(c) Stesse domanda di (a), supponendo che siano note soltanto le leggi $\mathbf{P}_X, \mathbf{P}_Y$.

(d) Stesse domanda di (b), supponendo che siano note soltanto le leggi $\mathbf{P}_X, \mathbf{P}_Y$.

(e) Stessa domanda di (a), supponendo che X e Y siano indipendenti.

(f) Stessa domanda di (b), supponendo che $g(X)$ e Y siano indipendenti.

(g) Supponendo che siano note soltanto le leggi $\mathbf{P}_X, \mathbf{P}_Y$ e che X e Y siano indipendenti, si può calcolare la speranza di $Z_3 = g(e^{XY})$?

Risoluzione. (a) Sì per entrambe le domande, poiché è nota la variabile aleatoria congiunta (X, Y).

(b) Sì per entrambe le domande, poiché è nota la variabile aleatoria congiunta (X, Y).

(c) La legge (ovvero la densità) di X determina quella di $g(X)$:

$$p_{g(X)}(z) = \textstyle\sum_{\{x \in \mathbf{R} : g(x) = z\}} p_X(x) \qquad \forall z \in \mathbf{R}.$$

Si può allora calcolare la speranza di $g(X)$, quindi anche quella di $Z_1 = g(X) + Y$.

Non si può calcolare la speranza di $Z_2 = g(X)Y$, poiché non è nota la legge congiunta $\mathbf{P}_{(g(X),Y)}$.

(d) No per entrambe le domande, poiché non è nota la legge della variabile aleatoria congiunta $(g(X), Y)$.

(e) L'indipendenza di X e Y implica quella di $g(X)$ e Y. Quindi si può calcolare anche la speranza di $Z_2 = g(X)Y$.

(f) Sì per entrambe le domande, poiché la legge congiunta $\mathbf{P}_{(g(X),Y)}$ è individuata dalle leggi marginali di $g(X)$ e Y, grazie all'indipendenza di $g(X)$ e Y.

(g) Sì per entrambe le domande, per i motivi della risposta (f).

— **Esercizio 19.** Siano $X, Y : \Omega \to \mathbf{R}$ due variabili aleatorie discrete limitate [in modo che non ci siano problemi di convergenza delle speranze] e $g : \mathbf{R} \to \mathbf{R}$ una funzione continua.

(a) L'indipendenza di $g(X)$ e Y implica quella di X e Y?

(b) Le leggi di X^3 e $\arctan Y$ determinano quelle di X e Y?

(c) L'indipendenza di $g(X)$ e e^Y implica quella di X e Y?

(d) L'indipendenza di X^3 e e^Y implica quella di X e Y^2?

Risoluzione. (a) No. Per un controesempio, basta scegliere g uguale alla funzione nulla (poiché allora la trasformazione $X \mapsto g(X)$ comporta una perdita di informazione).

(b) Sì, poiché le funzioni cubo ed arco-tangente sono invertibili.

(c) No. Per un controesempio, basta scegliere g uguale alla funzione nulla.

(d) Sì, per la Proposizione 3.19 del [B; p. 56].

— **Esercizio 20.** Sia X una variabile aleatoria con distribuzione $N(\mu, \sigma^2)$.

(a) È più probabile $X = \mu$ oppure $X = \mu + \sigma$?

(b) È più probabile l'evento $\{X \in [\mu, \mu + \sigma]\}$ oppure $\{X \in [\mu + \sigma, \mu + 3\sigma]\}$? I due eventi sono indipendenti?

Risoluzione. (a) Entrambi gli eventi hanno probabilità nulla.

(b) $\mathbf{P}(X \in [\mu, \mu + \sigma]) > 0.5$, quindi questo evento è più probabile. I due eventi hanno intersezione vuota e ciascuno ha probabilità non nulla, quindi non sono indipendenti.

13 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. Il [Verri] è un valido eserciziario (con risoluzioni). Esistono 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

— M. Verri: Probabilità e statistica: 400 esercizi d'esame risolti. Progetto Leonardo, Bologna 2010

I was at the mathematical school, where a master taught his pupils after a method scarcely imaginable to us in Europe. The proposition and demonstration was fairly written on a thin wafer, with ink composed of a cephalic tincture. This the student was to swallow upon a fasting stomach, and for three days following eat nothing but bread and water. As the wafer digested the tincture mounted to the brain, bearing the proposition along with it.

from Jonathan Swift's *Gulliver's Travels*