Cenni sull'impostazione assiomatica del calcolo delle probabilità

Marco Bramanti

3 novembre 2020

Nell'impostazione assiomatica moderna del calcolo delle probabilità, dovuta a N. Kolmogorov, negli anni 1930[1], uno *spazio di probabilità* è definito come un particolare spazio di misura (nel senso della teoria della misura astratta "alla Lebesgue"), precisamente come una terna

$$(\Omega, \mathcal{M}, P)$$

dove:

Ω, detto *spazio campionario*, è un insieme, che rappresenta l'insieme dei possibili esiti elementari di un esperimento aleatorio; gli elementi di Ω si dicono *eventi elementari*;

$\mathcal{M}$, detta *sigma algebra degli eventi*, è una σ-algebra di sottoinsiemi di Ω, e rappresenta la famiglia dei possibili eventi (non solo elementari) di cui potremmo voler calcolare la probabilità; gli elementi di $\mathcal{M}$ si dicono *eventi*, e sono insiemi di eventi elementari.

P, detta *misura di probabilità*, è una misura su $(\Omega, \mathcal{M})$, tale che $P(\Omega) = 1$. Perciò P è una funzione d'insieme,

$$P : \mathcal{M} \to [0, 1],$$

numerabilmente additiva. Per ogni $E \in \mathcal{M}$, $P(E)$ è la probabilità dell'evento E, ed è espressa da un numero fra 0 e 1.

Esempio 1 *Se consideriamo l'esperimento aleatorio "lancio di due dadi", lo spazio campionario Ω è l'insieme delle coppie ordinate (i, j) con $i, j = 1, 2, ..., 6$, dove l'evento elementare (i, j) significa "il primo dado ha fatto i, il secondo ha fatto j". La σ-algebra degli eventi in questo caso è semplicemente*[2] $\mathcal{M} = \mathcal{P}(\Omega)$,

[1] Nel 1933 fu stampata la prima edizione dell'opera di Kolmogorov "Fondamenti di teoria della probabilità", in tedesco; fu tradotta in russo nel 1936 e in inglese nel 1950. Si noti la coincidenza temporale tra la ricerca della sistemazione assiomatica della probabilità e i lavori fondamentali di analisi funzionale: nel 1932 vennero pubblicati tre trattati fondamentali sulla teoria degli operatori lineari tra spazi astratti, da parte di Stefan Banach, John von Neumann, Marshall Stone.

[2] il simbolo $\mathcal{P}(\Omega)$ indica l'insieme delle parti di Ω, cioè l'insieme di tutti i possibili sottoinsiemi di Ω, compreso l'insieme vuoto e Ω stesso.

perché nel caso di uno spazio campionario finito possiamo facilmente assegnare una probabilità a ogni sottoinsieme di Ω*. Infine, la misura* P *è quella uniforme, cioè* $P(E)$ *è il rapporto tra il numero di elementi di* E *e il numero totale di elementi di* Ω *(cioè* 36*).*

Una *variabile aleatoria* (d'ora in poi abbreviata v.a.) è una quantità che assume un valore (reale) in dipendenza dall'evento elementare che si è realizzato, perciò è una funzione

$$X : \Omega \to \mathbb{R}.$$

Nell'esempio del lancio dei due dadi, una v.a. potrebbe essere

$$X = \text{somma dei punti dei due dadi,}$$

che assume valori interi da 2 a 12.

Ci interessa calcolare la probabilità che X assuma certi valori, o assuma valori in certi intervalli di $\mathbb{R}$, ossia ci interessa calcolare, ad esempio

$$P\left(\{\omega \in \Omega : X(\omega) \in (a,b)\}\right),$$

espressione solitamente abbreviata in

$$P\left(X \in (a,b)\right).$$

La funzione

$$I \mapsto P\left(X \in I\right)$$

che ad ogni intervallo assegna la probabilità che X assuma valori in quell'intervallo si dice *legge della v.a.* X. Se conosciamo la legge di X, siamo in grado di fare tutti i calcoli che ci interessano sulla v.a. X. Affinché sia concettualmente possibile calcolare $P\left(X \in (a,b)\right)$ occorre che l'insieme $(X \in (a,b))$ sia un evento, cioè appartenga alla σ-algebra $\mathcal{M}$. Col linguaggio della teoria della misura, questo significa che *una variabile aleatoria su* Ω *è una (qualsiasi) funzione* $X : \Omega \to \mathbb{R}$ *misurabile*. Quest'ultima è la definizione precisa di variabile aleatoria.

Si presti attenzione alla differenza tra il concetto di *variabile aleatoria* e quello di *legge di una variabile aleatoria*. Ad esempio, supponiamo di lanciare due dadi, uno rosso e uno blu. Siano X, Y le variabili aleatorie "punteggio dato dal dado rosso" e "punteggio dato dal dado blu". Ovviamente sono due variabili aleatorie diverse: in ogni lancio, i due dadi potrebbero dare punteggio diverso. Tuttavia è altrettanto evidente che le due variabili hanno la stessa legge: ognuna assume i valori $1, 2, 3, 4, 5, 6$ con probabilità $1/6$ per ciascun esito.

Le due classi più comuni di variabili aleatorie sono le *v.a. discrete* e le *v.a. continue*.

Una v.a. X è *discreta* se i valori che può assumere sono un insieme finito o numerabile; ad esempio se X assume solo valori interi, o comunque solo una

successione $\lambda_1, \lambda_2, \lambda_3...$ di valori possibili[3]. Una v.a. X invece si dice *continua* se può assumere qualsiasi valore reale in un certo intervallo (a, b). La v.a. "somma dei punti dei due dadi" è discreta, mentre la v.a. "tempo di vita di una lampadina, espresso in secondi" è continua, perché il suo valore, potenzialmente, è qualsiasi numero reale in un intervallo (a, b) ragionevole.

Per una v.a. discreta che può assumere solo i valori $\lambda_1, \lambda_2, \lambda_3...$, una volta che conosciamo i numeri

$$c_n = P(X = \lambda_n)$$

possiamo calcolare la probabilità di qualsiasi evento legato a X, calcolando

$$P(X \in (a, b)) = \sum_{n:\lambda_n \in E} c_n,$$

dove ovviamente, per il loro significato,

$$0 \leq c_n \leq 1 \text{ per ogni } n, \text{ e}$$

$$\sum_{n=1}^{\infty} c_n = 1.$$

La successione

$$p_X(n) = c_n = P(X = \lambda_n)$$

si chiama talvolta *densità discreta* della v.a. X, e la sua conoscenza, come detto, è sufficiente a calcolare le probabilità degli eventi legati a X, mediante una somma finita o una serie numerica.

Esempio 2 *Si dice che una v.a. discreta X ha* legge di Poisson di parametro $\lambda > 0$*, e si scrive $X \sim P_0(\lambda)$, se assume i valori $k = 0, 1, 2, 3, ...$ con probabilità*

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

Notiamo che $e^{-\lambda} \frac{\lambda^k}{k!} > 0$ per ogni $k = 0, 1, 2, ...$ e

$$\sum_{k=0}^{\infty} p_X(k) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1$$

(sfruttando la serie di Taylor di e^x). Ad esempio,

$$P(X \geq 5) = e^{-\lambda} \sum_{k=5}^{\infty} \frac{\lambda^k}{k!}.$$

[3] Questa definizione, anche se è quella solitamente scritta nei testi, non è del tutto soddisfacente. Ciò che rende una v.a. "discreta" non è semplicemente la cardinalità numerabile dell'insieme dei valori assunti, ma il fatto che questi valori siano "separati uno dall'altro". Una definizione più soddisfacente potrebbe essere: una v.a. è discreta se l'insieme dei valori che può assumere *ha chiusura numerabile.* Ad esempio $\mathbb{Q}$ è un insieme numerabile, ma non è un insieme discreto, perché la sua chiusura è il continuo $\mathbb{R}$. Invece, l'insieme $\{1/n : n = 1, 2, 3, ...\}$ ha chiusura numerabile. Finezze a parte, le v.a. discrete più comuni hanno semplicemente come possibili valori *gli interi.*

Per una v.a. continua, la situazione è un po' diversa. Esisterà in generale una funzione densità, detta *densità continua*[4],

$$f_X(t) \geq 0, \text{ tale che}$$
$$\int_{\mathbb{R}} f_X(t)\, dt = 1$$

e

$$P(X \in (a,b)) = \int_a^b f_X(t)\, dt.$$

Di nuovo, quando conosciamo la funzione densità di X, possiamo calcolare le probabilità di tutti gli eventi che ci interessano.

Esempio 3 *Si dice che una v.a. continua X ha* legge normale standard *se ha densità*

$$f_X(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Si noti che $f_X(t) > 0$ per ogni t e $\int_{\mathbb{R}} f_X(t)\, dt = 1$. Si avrà

$$P(X \in (a,b)) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2} dt.$$

Esempio 4 *Più in generale, una v.a. continua X ha* legge normale $N(\mu, \sigma^2)$ *(e si scrive $X \sim N(\mu, \sigma^2)$) se X ha densità*

$$f_X(t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(t-\mu)^2/2\sigma}.$$

Si noti che $f_X(t) > 0$ per ogni t e $\int_{\mathbb{R}} f_X(t)\, dt = 1$. Si avrà

$$P(X \in (a,b)) = \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-(t-\mu)^2/2\sigma} dt.$$

Spesso, nella pratica, dopo aver definito "a parole" il significato di una v.a. X, dal punto di vista analitico ci si limita a esplicitare la sua funzione densità (discreta o continua), senza stare neppure a precisare lo spazio campionario Ω e la definizione di X come funzione da Ω a $\mathbb{R}$. In sostanza, si specifica la *legge* della v.a. senza definire accuratamente la v.a. stessa.

Nell'impostazione assiomatica della meccanica quantistica data da von Neumann, le grandezze fisiche "osservabili" sono variabili aleatorie.

Una v.a., anziché avere valori scalari, potrebbe avere valori vettoriali. Ad esempio, la posizione aleatoria di una particella quantistica nello spazio tridimensionale è un *vettore aleatorio* $\underline{X}$, ovvero un vettore di v.a.; la sua densità continua è una funzione di tre variabili $f_{\underline{X}}(x_1, x_2, x_3)$ e avrà il significato

[4] Questo nome non deve trarre in inganno: è semplicemente un'abbreviazione di "funzione densità di una v.a. continua", ma non significa che la funzione f_X sia continua. Ad esempio, potrebbe essere $f_X(t) = \chi_{(0,1)}(t)$.

espresso dalla formula:

$$P\left(\underline{X} \in E\right) = \int_E f_{\underline{X}}\left(x_1, x_2, x_3\right) dx_1 dx_2 dx_3 \text{ per } E \subset \mathbb{R}^3.$$

Due quantità fondamentali legati alle v.a. sono il *valore atteso* e la *varianza*, che ora definiamo.

Il *valore atteso* di una v.a. X (che indica in un certo senso la sua "media"), indicato solitamente nei testi matematici col simbolo $\mathrm{E}\left(X\right)$ e nei testi fisici col simbolo $\langle X \rangle$, è definito formalmente come

$$\mathrm{E}\left(X\right) = \langle X \rangle = \int_\Omega X\left(\omega\right) dP\left(\omega\right). \tag{1}$$

Si presti attenzione al fatto che questo è un integrale astratto, fatto non (ad esempio) su $\mathbb{R}$, ma sull'insieme degli eventi elementari; è un integrale rispetto alla misura di probabilità. Si noti come il calcolo delle probabilità moderno utilizzi a fondo l'astrazione e la generalità della teoria della misura e dell'integrazione "alla Lebesgue". Questa definizione di valore atteso come integrale della v.a. aiuta a capire e dimostrare facilmente certe proprietà del valore atteso, come la sua linearità, cioè il fatto che risulti

$$\mathrm{E}\left(\lambda X + \mu Y\right) = \lambda \,\mathrm{E}\, X + \mu \,\mathrm{E}\, Y,$$

dal momento che questa relazione non è altro che la linearità dell'integrale. Raramente però la (1) è di aiuto al calcolo effettivo del valore atteso. Per il calcolo si usa normalmente un'altra caratterizzazione del valore atteso, che si può fare, però, separatamente per le v.a. discrete o continue.

Se X è una v.a. discreta, che assume i possibili valori $\{\lambda_n\}_{n=1}^{\infty}$ con probabilità

$$P\left(X = \lambda_n\right) = p_X\left(n\right),$$

allora

$$\mathrm{E}\, X = \sum_{n=1}^{\infty} \lambda_n p_X\left(n\right).$$

Si noti che questa è una serie numerica (a termini non necessariamente positivi), di cui non è garantita a priori la convergenza: il valore atteso potrebbe non esistere. Del resto, anche dalla definizione astratta (1) questo è apparente: una v.a. è per definizione una funzione $X : \Omega \to \mathbb{R}$ *misurabile*, non necessariamente *integrabile*.

Esempio 5 *Calcoliamo il valore atteso di una v.a. X avente legge di Poisson $P_0\left(\lambda\right)$ (v. Esempio 2) cioè tale che $p_X\left(k\right) = e^{-\lambda}\frac{\lambda^k}{k!}$ per $k = 0, 1, 2...$ Si ha:*

$$\begin{aligned}\mathrm{E}\, X &= \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{(k-1)}}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.\end{aligned}$$

In particolare, il valore atteso è finito.

Esempio 6 *Una v.a. (discreta) si dice avere* legge geometrica di parametro $p \in (0,1)$ *se assume i valori possibili*

$$n = 1, 2, 3, ...$$

con probabilità

$$p_X(n) = p(1-p)^{n-1}.$$

Se $p = \frac{1}{2}$*, ad esempio, questa v.a. può rappresentare il "numero di volte in cui devo lanciare una moneta per ottenere la prima volta testa". Il suo valore atteso è*

$$\mathrm{E}\, X = \sum_{n=1}^{\infty} np(1-p)^{n-1} = \frac{1}{p}$$

(come si può dimostrare). In particolare, è finito.

Esempio 7 *Se* X *è una v.a. che assume i valori*

$$n = 1, 2, 3, ...$$

con probabilità[5]

$$p_X(n) = \frac{6}{\pi^2}\frac{1}{n^2},$$

allora

$$\mathrm{E}(X) = \frac{6}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n} = +\infty$$

(serie armonica divergente).

Se invece X è una v.a. continua, con densità $f_X(t)$, allora

$$\mathrm{E}(X) = \int_{\mathbb{R}} t f_X(t)\, dt.$$

Di nuovo, il valore atteso potrebbe anche non esiste finito.

Esempio 8 *Una v.a.* X *di legge normale standard, cioè con*

$$f_X(t) = \frac{1}{\sqrt{2\pi}}e^{-t^2/2}$$

avrà

$$\mathrm{E}(X) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} te^{-t^2/2}dt = 0.$$

[5] E' noto che

$$\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{\pi^2}{6}, \text{ quindi } \sum_{n=1}^{\infty} p_X(n) = 1.$$

Esempio 9 *Una v.a. X con densità continua*[6]

$$f_X(t) = \frac{1}{\pi}\frac{1}{1+t^2}$$

avrà

$$\mathrm{E}(X) = \frac{1}{\pi}\int_{\mathbb{R}} \frac{t}{1+t^2}dt \text{ non convergente}$$

(almeno nel senso di Lebesgue), perché tende a zero all'infinito come $1/t$, non integrabile.

Se $\underline{X} = (X_1, X_2, X_3)$ è un *vettore aleatorio* in $\mathbb{R}^3$, di densità continua $f_X(x_1, x_2, x_3)$, potremo calcolare il valore atteso di ognuna delle sue componenti,

$$\mathrm{E}(X_i) = \int_{\mathbb{R}^3} x_i f_X(x_1, x_2, x_3)\, dx_1 dx_2 dx_3.$$

Introduciamo ora la *varianza* di una v.a. Si tratta di un "indice di dispersione" che misura quanto i valori assunti dalla variabile sono probabilmente lontani dal suo valore atteso (quindi, quanto sono "dispersi"). La definizione è ricondotta a quella di valore atteso:

$$\mathrm{Var}(X) = \sigma_X^2 = \mathrm{E}\left((X - \mathrm{E}\,X)^2\right) = \int_\Omega (X(\omega) - \mathrm{E}(X))^2\, dP(\omega)$$

che, fatti i conti in base alla linearità del valore atteso, si può riscrivere anche così:

$$\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - (\mathrm{E}\,X)^2. \tag{2}$$

Per il calcolo effettivo è bene di nuovo distinguere il caso discreto da quello continuo. Si usa la formula (2), in cui $(\mathrm{E}(X))^2$ sappiamo già calcolarla, mentre $\mathrm{E}\left(X^2\right)$ si calcola nel caso discreto così:

$$\mathrm{E}\left(X^2\right) = \sum_{n=1}^{\infty} \lambda_n^2 p_X(n)$$

e nel caso continuo così:

$$\mathrm{E}\left(X^2\right) = \int_{\mathbb{R}} t^2 f_X(t)\, dt.$$

Come per il valore atteso, a maggior ragione per la varianza non è garantita la sua finitezza.

[6] Si noti che

$$\int_{\mathbb{R}} \frac{1}{\pi}\frac{1}{1+t^2}dt = 1.$$

Dalla prima definizione di varianza si legge in particolare che una *v.a. con varianza nulla è costante*:

$$\operatorname{Var}(X) = 0 = \int_{\Omega} (X(\omega) - \mathrm{E}(X))^2 \, dP(\omega) \Rightarrow X(\omega) - \mathrm{E}(X) = 0,$$
$$X = E(X)$$

(per la precisione, dovremmo dire che è costante quasi ovunque, o come si dice in linguaggio probabilistico, "costante quasi certamente").

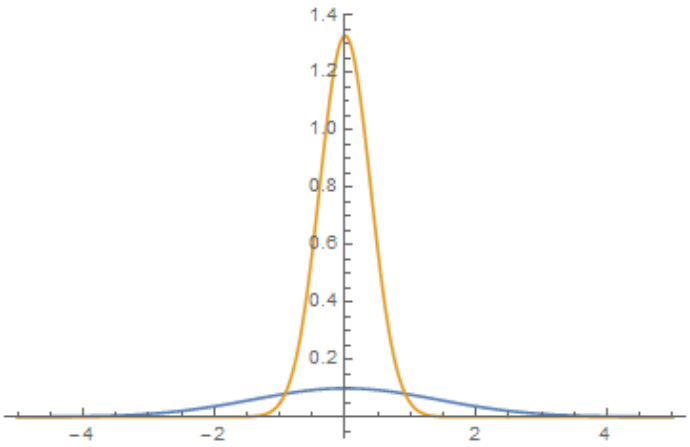
Si osservi che, dimensionalmente, la varianza è una grandezza quadratica rispetto alla v.a. corrisponente. Per avere una grandezza che rappresenti ancora un indice di dispersione, e sia dimensionalmente omogenea a X, si introduce la *deviazione standard* σ_X, che è per definizione la radice quadrata della varianza:

$$\sigma_X = \sqrt{\operatorname{Var} X}.$$

Esempio 10 *Per una v.a. continua di legge* $N(\mu, \sigma^2)$ *si ha*

$$E(X) = \mu$$
$$\operatorname{Var}(X) = \sigma^2.$$

Quindi i due parametri μ, σ *della normale hanno il significato di media e deviazione standard. Si capisce bene il significato della varianza (o della deviazione standard) come indice di dispersione osservando i grafici di densità normali per diversi valori di* σ*. Valori di* σ *piccoli indicano una legge poco dispersa, cioè molto concentrata, con un grafico della densità a "campana alta e stretta", mentre valori di* σ *grandi corrispondono a una legge molto dispersa, con un grafico della densità a "campana bassa e larga":*



Grafici di densità $N(0, \sigma^2)$ *con* $\sigma = 4$ *(curva più bassa) e* $\sigma = 0.3$ *(curva più alta)*

Un teorema, semplice ma fondamentale, che illustra bene il significato della varianza, è il seguente:

Teorema 11 (Disuguaglianza di Cebicev) *Sia X una v.a. qualsiasi avente valore atteso μ_X e varianza σ_X^2 finiti. Allora per ogni $\delta > 1$ si ha*

$$P\left(|X - \mu_X| < \delta\sigma_X\right) \geq 1 - \frac{1}{\delta^2}.$$

Per δ grande la quantità $1 - \frac{1}{\delta^2}$ si avvicina a 1. Dunque la disuguaglianza precedente esprime il fatto che, con grande probabilità, una variabile aleatoria si discosta dal proprio valor medio per meno di un certo multiplo della propria deviazione standard. Qualche esempio numerico spiegherà meglio quest'idea. Scegliendo $\delta = 2, 3, 5, 10$ otteniamo:

$$\begin{aligned} P\left(|X - \mu_X| < 2\sigma_X\right) &\geq 0.75 \\ P\left(|X - \mu_X| < 3\sigma_X\right) &\geq 0.88 \\ P\left(|X - \mu_X| < 5\sigma_X\right) &\geq 0.96 \\ P\left(|X - \mu_X| < 10\sigma_X\right) &\geq 0.99. \end{aligned}$$

Ad esempio, con probabilità di almeno il 99% una qualsiasi variabile aleatoria non si discosta dalla propria media per più di 10 deviazioni standard. Si rifletta sull'interesse di questo tipo di affermazioni quando X rappresenta la misura di una grandezza fisica e σ_X l'incertezza di questa misura.

La cosa notevole della precedente disuguaglianza è che vale per *qualsiasi* v.a., *comunque* sia distribuita (purché μ_X e σ_X siano finite). Questo pregio suggerisce anche il difetto di questo risultato: dovendo valere così in generale, questa stima non sarà molto precisa. Ad esempio, se sappiamo che X ha legge Gaussiana $N\left(\mu, \sigma^2\right)$ allora si può calcolare che è:

$$\begin{aligned} P\left(|X - \mu| < 2\sigma\right) &\geq 0.9545 \\ P\left(|X - \mu| < 3\sigma\right) &\geq 0.9973 \\ P\left(|X - \mu| < 5\sigma\right) &\geq 0.999999 \\ P\left(|X - \mu| < 10\sigma\right) &\simeq 1. \end{aligned}$$

Queste ultime stime sono *compatibili* con quelle del teorema di Cebicev, ma molto più precise. Ad esempio, su molti testi di statistica si dice che una v.a. normale assume "in pratica" valori compresi nell'intervallo $(\mu - 3\sigma, \mu + 3\sigma)$. Quel che si può dire rigorosamente è che una v.a. normale assume valori in quest'intervallo *con probabilità maggiore di* 0.99.

Media e varianza di X richiedono di calcolare $\mathrm{E}(X)$ e $\mathrm{E}\left(X^2\right)$. Più in generale, si dice *momento* n-esimo di una v.a. X il valore atteso $\mathrm{E}(X^n)$. I momenti successivi al secondo rappresentano, o servono a costruire, opportuni *indici di forma* legati alla v.a. Inoltre vari risultati trattano il problema di ricostruire la legge di una v.a. dai suoi (infiniti) momenti, nell'ipotesi che siano tutti finiti.

Concludiamo con quest'osservazione. In tutte le presentazioni didattiche "elementari" del calcolo delle probabilità ci si limita a considerare le due classi

di v.a. discrete e continue, e le si considerano separatamente, stabilendo due serie distinte di risultati, analoghi, ma comunque differenti. Uno dei punti di forza della teoria di Kolmogorov è che v.a. discrete e continue sono trattate simultanemanete come casi particolari di un concetto più generale. Ad esempio valore atteso e varianza di una v.a., non sono definite da formule distinte nel caso discreto o continuo, ma da un'unica formula generale, che però fa uso di un integrale astratto, di cui le serie numeriche e gli integrali su $\mathbb{R}$ sono casi particolari.