Elementi di Probabilità e Statistica

Maurizio Pratelli

Anno Accademico 2017-18

Indice

Capitolo 1

Nozioni fondamentali di Calcolo delle Probabilità.

1.1 Prime definizioni.

Di fronte ad una situazione che suggerisce l'uso del Calcolo delle Probabilità, incontriamo alcune *affermazioni* legate tra loro dai connettivi logici "**o**", "**e**", "**non**": è facile convincersi che si può tradurre questo in una famiglia di sottinsiemi (chiamati *eventi*) di un opportuno insieme Ω, contenente l'insieme vuoto e tutto l'insieme, e stabile per le operazioni di *unione* (finita), *intersezione* e *complementazione*. Una tale famiglia di insiemi si chiama un'**algebra** di parti (il termine anglosassone è *field*).

L'insieme Ω, che usualmente rappresenta tutti i possibili esiti, è spesso chiamato *spazio fondamentale* o anche (soprattutto in Statistica) *spazio dei campioni*.

Il *grado di fiducia* che un sottinsieme si realizzi (chiamato *probabilità*), è rappresentato da un numero compreso tra 0 e 1; inoltre è intuitivo supporre che se due eventi sono incompatibili (cioè hanno intersezione vuota) la probabilità che si realizzi uno qualsiasi dei due debba essere la somma delle probabilità dei singoli eventi. Questo equivale a dire che la probabilità è una funzione d'insieme *(finitamente) additiva*.

Cominciamo a dare le prime definizioni (provvisorie):

Definizione 1.1.1 (**Algebra di parti).** Dato un insieme Ω, si chiama algebra di parti una famiglia $\mathcal{F}$ di sottinsiemi di Ω tale che:

a) l'insieme vuoto $\emptyset$ e l'intero insieme Ω sono elementi di $\mathcal{F}$;

b) se $A \in \mathcal{F}$, anche il suo complementare $A^c \in \mathcal{F}$;

c) se A e B sono elementi di $\mathcal{F}$, anche $A \cup B \in \mathcal{F}$.

Notiamo che automaticamente $\mathcal{F}$ è stabile anche per l'intersezione finita: questo segue dalle proprietà b) e c) e dal fatto che $\big(A \cap B\big)^c = A^c \cup B^c$. Inoltre le proprietà definite in a) sono ridondanti: è sufficiente ad esempio supporre che Ω sia un elemento di $\mathcal{F}$ ed automaticamente $\emptyset = \Omega^c$ è un elemento di $\mathcal{F}$.

Definizione 1.1.2 (**Probabilità finitamente additiva).** Data un'algebra $\mathcal{F}$ di parti di un insieme Ω , si chiama probabilità (finitamente additiva) una funzione $\mathbf{P} : \mathcal{F} \longrightarrow [0, 1]$ tale che

a) se $A, B \in \mathcal{F}$ e $A \cap B = \emptyset$, allora $\mathbf{P}\big(A \cup B\big) = \mathbf{P}\big(A\big) + \mathbf{P}\big(B\big)$;

b) $\mathbf{P}(\Omega) = 1$.

Gli elementi dell'algebra di parti $\mathcal{F}$ sono chiamati **eventi** , si chiama **trascurabile** un evento A tale che $\mathbf{P}(A) = 0$ e si chiama **quasi certo** un evento A tale che $\mathbf{P}(A) = 1$.

Vediamo alcune conseguenze immediate della definizione 1.1.2 che si possono provare facilmente per esercizio:

1. $\mathbf{P}(\emptyset) = 0$;
2. $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$;
3. se $B \subset A$, $\mathbf{P}\big(A \setminus B\big) = \mathbf{P}(A) - \mathbf{P}(B)$, dove si è posto $A \setminus B = A \cap B^c$;
4. $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$;
5. $\mathbf{P}(A \cup B \cup C) = \mathbf{P}(A) + \mathbf{P}(B) + \mathbf{P}(C) - \mathbf{P}(A \cap B) - \mathbf{P}(A \cap C) - \mathbf{P}(B \cap C) + \mathbf{P}(A \cap B \cap C)$, e così via ...

Le definizioni sopra riportate, oltre ad essere molto intuitive, sono supportate da valide argomentazioni logiche, tuttavia dal punto di vista matematico presentano una difficoltà: la **additività semplice** non consente di *andare al limite*, e di conseguenza di *calcolare degli integrali*. La buona proprietà per poter effettuare queste operazioni è la **additività numerabile**, detta anche **σ-additività**. Inoltre la famiglia di parti sulla quale possa essere definita una funzione σ-additiva è opportuno che sia stabile per *unione numerabile* e non *unione finita*.

Per questo motivo, seguendo quella che è ormai comunemente chiamata la *definizione assiomatica di Probabilità* secondo Kolmogorov, sostituiamo alle precedenti queste definizioni.

Definizione 1.1.3 (**σ-algebra di parti).** Dato un insieme Ω, si chiama σ-algebra di parti una famiglia $\mathcal{F}$ di sottinsiemi di Ω tale che:

a) l'insieme vuoto $\emptyset$ e l'intero insieme Ω sono elementi di $\mathcal{F}$;
b) se $A \in \mathcal{F}$, anche il suo complementare $A^c \in \mathcal{F}$;
c) se $(A_n)_{n\geq 1}$ è una successione di elementi di $\mathcal{F}$, anche $\bigcup_{n=1}^{+\infty} A_n \in \mathcal{F}$.

Naturalmente una σ-algebra è anche un'algebra di parti: infatti $A \cup B = A \cup B \cup \emptyset \cup \emptyset \ldots$.

Osservazione 1.1.4. La terminologia anglosassone per una famiglia di parti con tali proprietà è *σ-field*, che dovrebbe essere tradotto *σ-campo* (termine in realtà poco usato); la terminologia francese (introdotta dal Bourbaki) è *tribù*.

Definizione 1.1.5 (Probabilità). Assegnato un insieme Ω ed una σ-algebra $\mathcal{F}$ di parti di Ω, si chiama probabilità una funzione $\mathbf{P} : \mathcal{F} \longrightarrow [0,1]$ tale che
a) se $(A_n)_{n=1,2,\ldots}$ è una successione di elementi di $\mathcal{F}$ a due a due disgiunti, si ha $\mathbf{P}\big(\bigcup_{n=1}^{+\infty} A_n\big) = \sum_{n=1}^{+\infty} \mathbf{P}(A_n)$;
b) $\mathbf{P}(\Omega) = 1$.

Una funzione d'insieme che gode della proprietà a) della definizione 1.1.5 è detta *misura*; la probabilità è dunque una misura *normalizzata*. È facile constatare che una funzione σ-additiva è anche semplicemente additiva.

Una terna $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ formata da un insieme Ω, una σ-algebra $\mathcal{F}$ di parti di Ω ed una probabilità $\mathbf{P}$ definita su $\mathcal{F}$ viene chiamata *spazio probabilizzato* o anche *spazio di Probabilità*.

La proprietà seguente spiega perché la σ-additività può essere considerata una sorta di *continuità*.

Proposizione 1.1.6. *Sia $\mathcal{F}$ una σ-algebra di parti di un insieme Ω e sia $\mathbf{P} : \mathcal{F} \longrightarrow [0,1]$ semplicemente additiva (e tale che $\mathbf{P}(\Omega) = 1$). Sono equivalenti le seguenti proprietà:*
1) $\mathbf{P}$ è σ-additiva;
2) se $(A_n)_{n\geq 1}$ è una successione crescente di insiemi (cioè $A_n \subseteq A_{n+1}$), posto $A = \bigcup_{n\geq 1} A_n$, si ha $\lim_{n\to+\infty} \mathbf{P}(A_n) = \mathbf{P}(A)$;
3) se $(A_n)_{n\geq 1}$ è una successione decrescente di insiemi , posto $A = \bigcap_{n\geq 1} A_n$, si ha $\lim_{n\to+\infty} \mathbf{P}(A_n) = \mathbf{P}(A)$.

Dimostrazione. Mostriamo ad esempio l'equivalenza tra 1) e 2). Supponiamo che sia verificata 1), e poniamo $B_1 = A_1$, $B_n = A_n \setminus A_{n-1}$ per $n > 1$: gli insiemi $(B_n)_{n\geq 1}$ sono a due a due disgiunti e per l'additività finita si ha $\mathbf{P}(B_n) = \mathbf{P}(A_n) - \mathbf{P}(A_{n-1})$.

Poichè $\bigcup_{n\geq 1} A_n = \bigcup_{n\geq 1} B_n$, si ha $\mathbf{P}(A) = \sum_{n=1}^{+\infty} \mathbf{P}(B_n) =$
$= \lim_{n\to\infty} \sum_{h=1}^{n} \mathbf{P}(B_h) = \lim_{n\to\infty} \mathbf{P}(A_n)$.

Viceversa, supponendo che sia verificata la proprietà 2) , assegnata una successione $(B_n)_{n\geq 1}$ di eventi a due a due disgiunti, posto $A_n = B_1 \cup \ldots \cup B_n$, questa risulta essere una successione crescente di insiemi. Si ha allora $\mathbf{P}\big(\bigcup_{n\geq 1} B_n\big) = \mathbf{P}\big(\bigcup_{n\geq 1} A_n\big) = \lim_{n\to\infty} \mathbf{P}(A_n) = \lim_{n\to\infty} \sum_{h=1}^{n} \mathbf{P}(B_h) = \sum_{n=1}^{+\infty} \mathbf{P}(B_n)$

L'equivalenza tra 2) e 3) si dimostra facilmente passando al complementare. □

D'ora innanzi, le affermazioni 2) e 3) del precedente enunciato verranno anche scritte nella seguente maniera, telegrafica ma perfettamente chiara:

2) $A_n \uparrow A \quad \Longrightarrow \mathbf{P}(A_n) \to \mathbf{P}(A)$ (o anche $\mathbf{P}(A_n) \uparrow \mathbf{P}(A)$);

3) $A_n \downarrow A \quad \Longrightarrow \mathbf{P}(A_n) \to \mathbf{P}(A)$ (o anche $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$)

Inoltre le precedenti affermazioni sono anche equivalenti alle seguenti (lasciamo per esercizio la relativa facile dimostrazione):

2)bis: $A_n \uparrow \Omega \quad \Longrightarrow \mathbf{P}(A_n) \to 1$;

3)bis: $A_n \downarrow \emptyset \quad \Longrightarrow \mathbf{P}(A_n) \to 0$.

È naturale a questo punto chiedersi perchè la probabilità è assegnata solo su *alcuni* e non *tutti* i sottinsiemi di Ω : il motivo di questo è una difficoltà di ordine matematico, cioè non sempre è possibile estendere una funzione σ-additiva a tutti i sottinsiemi di un insieme Ω.

Esaminiamo in particolare un esempio concreto, immaginiamo di *scegliere a caso un numero compreso tra* 0 *e* 1 : lo spazio più naturale è $\Omega = [0, 1]$ e ad un intervallo $]a, b]$ (in verità non importa se questo intervallo è aperto, chiuso ..) sembra ragionevole attribuire come probabilità la sua lunghezza $(b - a)$. Inoltre è ovvio supporre che la probabilità attribuita sia *invariante per traslazioni (modulo 1)*, cioè $\mathbf{P}(A) = \mathbf{P}(A + c)$, dove con $A + c$ si intende il traslato di A (modulo 1).

Il famoso *controesempio di Vitali*, tradotto in questa situazione, può essere letto nel modo seguente:

Proposizione 1.1.7. *Non è possibile costruire una funzione* $\mathbf{P}$ *σ-additiva definita su* tutti *i sottinsiemi di* $[0, 1]$ *e tale che:*

1) $\mathbf{P}\big(]a, b]\big) = b - a \qquad$ *se* $0 \leq a \leq b \leq 1$ *;*

2) $\mathbf{P}$ *sia invariante per traslazioni (modulo 1).*

Osserviamo che quella enunciata sopra è una traduzione ai nostri scopi dell'esempio di Vitali, consistente nella costruzione di un sottinsieme della retta $\mathbb{R}$ non misurabile secondo Lebesgue. Torneremo su questo argomento nell'Appendice.

1.2 Il caso di uno spazio finito: elementi di calcolo combinatorio.

La difficoltà enunciata alla fine del paragrafo precedente (cioè l'impossibilità di estendere la probabilità a *tutti* i sottinsiemi di un insieme Ω) non si pone se Ω è un insieme finito (cioè $\Omega = \{\omega_1, \dots, \omega_n\}$). In tal caso è usuale (anche se non obbligatorio) considerare come σ-algebra degli eventi la famiglia $\mathcal{P}(\Omega)$ di tutte le parti di Ω ; inoltre la probabilità è univocamente determinata dai numeri $p_i = \mathbf{P}(\{\omega_i\})$, $(p_i \geq 0\,,\, p_1 + \cdots + p_n = 1)$. Per ogni evento $A \subset \Omega$ si ha infatti $\mathbf{P}(A) = \sum_{\omega_i \in A} p_i$. (D'ora innanzi scriveremo più brevemente $\mathbf{P}(\omega_i)$ anziché $\mathbf{P}(\{\omega_i\})$).

La stessa cosa vale se l'insieme Ω è numerabile ($\Omega = \{\omega_1, \omega_2, \dots\}$) : usualmente si considera come σ-algebra $\mathcal{F}$ la famiglia $\mathcal{P}(\Omega)$ di tutte le parti e vale la formula appena scritta, dove la somma finita diventa la somma di una serie se l'evento A è un insieme di cardinalità infinita.

Nel caso in cui Ω sia un insieme finito e gli *eventi elementari* ω_i siano equiprobabili, si parla di *distribuzione uniforme di probabilità su* Ω; naturalmente non esiste una distribuzione uniforme di probabilità su un insieme Ω numerabile ma infinito.

Tornando al caso di Ω finito e di distribuzione uniforme di probabilità, si ottiene la formula

$$\mathbf{P}(A) = \frac{\sharp A}{\sharp \Omega} = \frac{|A|}{|\Omega|}$$

dove con $\sharp A$ o con $|A|$ si indica la *cardinalità* (o numero degli elementi) dell'insieme A. La formula sopra scritta è anche chiamata *rapporto tra casi favorevoli e casi possibili* e talvolta ad essa ci si riferisce indicandola come la *definizione classica di Probabilità*.

In questo ambito, i problemi diventano molto spesso problemi di *calcolo combinatorio*: delle varie formule riportate dai libri (spesso con nomi diversi da un libro all'altro) bisogna, a mio avviso, conoscerne soltanto tre. Tutte le altre si possono dedurre da queste come esercizio. Prima di riportare queste formule premettiamo una comoda notazione: dato un intero n, anziché dire *un insieme di cardinalità* n, scriveremo più brevemente $\{1, \dots, n\}$.

Proposizione 1.2.1. *Siano k ed n due interi: il numero di applicazioni da $\{1, \dots, k\}$ a $\{1, \dots, n\}$ è n^k*

Proposizione 1.2.2 (Permutazioni). *Il numero di modi in cui si possono ordinare gli elementi di $\{1, \dots, n\}$ è $n!$*

Questa formula, così come la precedente, si dimostra per induzione.

Proposizione 1.2.3 (**Coefficiente binomiale**). *Siano* $0 \leq k \leq n$ *: il numero di sottinsiemi di* $\{1, \ldots, n\}$ *formati da* k *elementi è*

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Anche questa formula si dimostra per induzione, a scelta su k o su n.

Vediamo ora, a titolo d'esempio, due formule che si possono dedurre dalle presedenti: lasciamo la dimostrazione come esercizio.

Esercizio 1.2.4. Siano $0 \leq k \leq n$: il numero di sottinsiemi *ordinati* di $\{1, \ldots, n\}$ formati da k elementi è $\frac{n!}{(n-k)!}$

Notiamo che questo numero coincide anche con il numero delle *applicazioni iniettive* da $\{1, \ldots, k\}$ in $\{1, \ldots, n\}$.

Esercizio 1.2.5. Siano $k_1, \ldots, k_h$ interi con $k_1 + \cdots + k_h = n$: il numero di modi in cui si possono scegliere h sottinsiemi di $\{1, \ldots, n\}$ formati rispettivamente da $k_1, \ldots, k_h$ elementi è

$$\frac{n!}{k_1! \ldots k_h!}$$

1.3 Probabilità condizionata ed indipendenza.

Quando si à conoscenza della realizzazione di un evento, cambia la valutazione di probabilità di ogni altro evento: ad esempio se si sa che il numero uscito su un giro della roulette è un numero *pari*, la probabilità che sia uscito il numero 16 non è più $\frac{1}{37}$ ma $\frac{1}{18}$ (ricordiamo che la ruota della roulette contiene 37 caselle, numerate da 0 a 36, e che lo 0 non è considerato né pari né dispari). Se si è realizzato l'evento $B = \{2, 4, \ldots, 36\}$ (cioè è uscito un numero pari) sono rimasti 18 *casi possibili* dei quali uno è *favorevole*: se indichiamo con $A = \{16\}$, notiamo che la nuova probabilità che è stata attribuita ad A verifica dalla formula $\frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}$.

Si possono fornire diversi esempi simili che sempre verificano la formula sopra riportata: queste considerazioni sono all'origine della definizione che segue.

Definizione 1.3.1. Assegnato uno spazio di probabilità $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ ed un evento B *non trascurabile*, si chiama *probabilità condizionata* di A rispetto a B il numero

$$\mathbf{P}\big(A\big|B\big) \;=\; \frac{\mathbf{P}\big(A \cap B\big)}{\mathbf{P}\big(B\big)}$$

Essa indica la probabilità che viene associata all'evento A, coerentemente con la valutazione precedentemente assegnata, in seguito all'informazione che si è realizzato l'evento B.

Esercizio 1.3.2. Provare che, fissato B non trascurabile, la funzione $A \longrightarrow \mathbf{P}(A|B)$ è effettivamente una probabilità sulla σ-algebra $\mathcal{F}$.

Dati due eventi A e B non trascurabili, è immediato constatare che vale la formula $\mathbf{P}(A \cap B) = \mathbf{P}(A|B).\mathbf{P}(B) = \mathbf{P}(B|A).\mathbf{P}(A)$.

Proposizione 1.3.3. *Siano $A_1, \ldots, A_n$ eventi, e supponiamo che $A_1 \cap \ldots \cap A_{n-1}$ sia non trascurabile: vale la formula*

$$\mathbf{P}\big(A_1 \cap \ldots \cap A_n\big) \;=\; \mathbf{P}\big(A_1\big).\mathbf{P}\big(A_2\big|A_1\big) \ldots \mathbf{P}\big(A_n\big|A_1 \cap \ldots \cap A_{n-1}\big) \quad (1.3.1)$$

La dimostrazione si ottiene immediatamente scrivendo i vari termini; si noti che, se $1 \leq k < n-1$, anche $A_1 \cap \ldots \cap A_k$ è non trascurabile.

Definizione 1.3.4 (**Sistema di alternative).** Si chiama *sistema di alternative* una partizione di Ω in n eventi non trascurabili $B_1, \ldots, B_n$.

Ricordiamo che *partizione* significa che gli insiemi B_i sono a due a due disgiunti e che la loro unione è l'intero insieme Ω.

Proposizione 1.3.5 (**Formula di Bayes).** *Sia $B_1, \ldots, B_n$ un sistema di alternative: assegnato una qualunque evento A non trascurabile, valgono le formule*

$$\mathbf{P}(A) \;=\; \sum_{i=1}^{n} \mathbf{P}\big(A\big|B_i\big)\, \mathbf{P}\big(B_i\big) \quad (1.3.2)$$

$$\mathbf{P}\big(B_i\big|A\big) \;=\; \frac{\mathbf{P}\big(A\big|B_i\big)\, \mathbf{P}\big(B_i\big)}{\sum_{j=1}^{n} \mathbf{P}\big(A\big|B_j\big)\, \mathbf{P}\big(B_j\big)} \quad (1.3.3)$$

Dimostrazione. Per quanto riguarda la prima formula, si noti che $A = (A \cap B_1) \cup \ldots \cup (A \cap B_n)$ e questi eventi sono a due a due disgiunti: si ha pertanto

$$\mathbf{P}(A) = \sum_{i=1}^{n} \mathbf{P}\big(A \cap B_i\big) \;=\; \sum_{i=1}^{n} \mathbf{P}\big(A\big|B_i\big)\, \mathbf{P}\big(B_i\big)$$

La seconda formula ne è una conseguenza immediata. Usualmente si da il nome di *formula di Bayes* all'equazione 1.3.3, che è chiamata talvolta *formula delle probabilità delle cause.* □

Le formule della Proposizione 1.3.5 sono valide anche se il sistema di alternative anzichè essere finito è numerabile, naturalmente sostituendo alle somme finite le somme di una serie.

Esercizio 1.3.6. Qual è la probabilità che, in una estrazione del lotto, tutti e 5 i numeri estratti non siano superiori a 20? Provare a risolvere questo facile esercizio in due modi, utilizzando cioè il calcolo combinatorio e la formula 1.3.1.

Introduciamo ora il concetto di indipendenza (stocastica): vogliamo tradurre con una formula matematica l'idea che la conoscenza che si è realizzato l'evento A non modifica la valutazione di probabilità di B e viceversa. A tale scopo consideriamo due eventi A e B (non trascurabili) e proviamo a scrivere le eguaglianze $\mathbf{P}(A) = \mathbf{P}(A|B)$ e $\mathbf{P}(B) = \mathbf{P}(B|A)$: un esame immediato mostra che queste sono equivalenti tra loro ed equivalenti all'eguaglianza $\mathbf{P}(A \cap B) = \mathbf{P}(A).\mathbf{P}(B)$. A differenza delle due precedenti, quest'ultima è simmetrica rispetto ai due eventi ed ha senso anche se uno dei due (o anche tutti e due) sono trascurabili: ne segue che questa è la *buona* definizione di indipendenza.

Definizione 1.3.7 (**Indipendenza stocastica).** Due eventi A e B sono detti indipendenti se vale l'eguaglianza

$$\mathbf{P}(A \cap B) = \mathbf{P}(A).\mathbf{P}(B)$$

È un facile esercizio provare le seguenti affermazioni:

- Se A e B sono indipendenti, sono indipendenti anche A^c e B; A e B^c; A^c e B^c.
- Se $\mathbf{P}(A) = 0$ oppure $\mathbf{P}(A) = 1$, A è indipendente da qualsiasi altro evento.
- Due eventi *incompatibili* (cioè che hanno intersezione vuota) non possono essere indipendenti, a meno che uno dei due sia trascurabile.

Vediamo ora come si estende questa definizione al caso di n eventi (con $n \geq 3$).

Definizione 1.3.8 (**Indipendenza di più eventi).** Assegnati n eventi $A_1, \ldots, A_n$, questi si dicono *indipendenti* se per ogni intero k con $2 \leq k \leq n$ e per ogni scelta di interi $1 \leq i_1 < i_2 < \ldots < i_k \leq n$, vale l'eguaglianza

$$\mathbf{P}\big(A_{i_1} \cap \cdots \cap A_{i_k}\big) = \mathbf{P}\big(A_{i_1}\big) \ldots \mathbf{P}\big(A_{i_k}\big)$$

La definizione appena riportata è piuttosto misteriosa: risulterà più chiara quando verrà introdotta la nozione di indipendenza per variabili aleatorie. È istruttivo tuttavia provare per esercizio la proposizione seguente, che in qualche modo giustifica la definizione appena fornita.

Proposizione 1.3.9. *Gli eventi $A_1, \ldots, A_n$ sono indipendenti se e solo se, per ogni possibile scelta di $B_i = A_i$ oppure $B_i = A_i^c$, vale l'eguaglianza*

$$\mathbf{P}\big(B_1 \cap \ldots \cap B_n\big) \,=\, \mathbf{P}\big(B_1\big) \,.\, .\mathbf{P}\big(B_n\big)$$

Esercizio 1.3.10. Sull'insieme $\Omega = \{1, 2, 3, 4\}$ munito della distribuzione uniforme di probabilità, verificare che gli eventi $A = \{1, 2\}$, $B = \{1, 3\}$ e $C = \{2, 3\}$ sono *a due a due* indipendenti, ma non sono *globalmente* indipendenti

Osservazione 1.3.11. Un caso tipico di indipendenza si ha nelle *prove ripetute nelle medesime condizioni*: ad esempio sono indipendenti i risultati di successivi lanci di monete o successivi giri della ruota della roulette, ma *non sono indipendenti* i risultati delle 5 estrazioni nel lotto.

1.4 Appendice: alcuni complementi.

1.4.1 Il controesempio di Vitali.

Consideriamo l'intervallo [0,1]: Vitali ha provato che *non è possibile costruire una funzione* $\mathbf{m}$ *definita su tutti i sottinsiemi di [0,1] e tale che*

a) $\mathbf{m}$ è σ-additiva;

b) $\mathbf{m}$ è invariante per traslazioni (modulo 1);

c) $\mathbf{m}(\,[0,1]\,) \,=\, 1$.

Cominciamo ad osservare che se esiste una funzione d'insieme con le proprietà a), b) e c), necessariamente $\mathbf{m}\big(]a,b]\big) = (b-a)$, se $\;0 \le a < b \le 1$: è immediato verificare questa eguaglianza per a e b razionali e si estende al caso generale per continuità (vedi 1.1.6). Tuttavia questa eguaglianza in realtà non ci servirà nella costruzione dell'esempio.

Consideriamo su $[0,1]$ la *relazione d'equivalenza* : $x \,\mathcal{R}\, y$ se $x-y$ è razionale $\big((x-y) \in \mathbb{Q}\big)$. Sia A l'insieme delle *classi di equivalenza* e per ogni $a \in A$ consideriamo (utilizzando l'*assioma della scelta*) un elemento $x_a \in a$: chiamiamo poi E l'insieme formato da tutti questi punti, cioè $E = \big\{x_a \big| a \in A\big\}$.

Chiamiamo $\widetilde{\mathbb{Q}} = \mathbb{Q} \cap \big[0, 1\big[$ l'insieme dei razionali compresi tra 0 e 1, e per ogni $r \in \widetilde{\mathbb{Q}}$, sia E_r l'insieme ottenuto effettuando su E la traslazione di r modulo 1, più precisamente

$$E_r \,=\, \Big\{x \in [0,1] \,\Big|\quad (x-r) \in E, \quad \text{oppure} \quad (x-r+1) \in E\Big\}$$

Per ipotesi, $\mathbf{m}\big(E_r\big) = \mathbf{m}\big(E\big)$, qualunque sia r. Si provano facilmente queste due affermazioni:

1) se $r \neq s$, allora $E_r \cap E_s = \emptyset$;

2) $[0,1]$ è l'unione degli insiemi E_r, al variare di $r \in \widetilde{\mathbb{Q}}$.

A questo punto abbiamo costruito il controesempio: se $\mathbf{m}$ esiste, si deve avere infatti $1 = \mathbf{m}\big([0,1]\big) = \sum_{r\in\widetilde{\mathbb{Q}}} \mathbf{m}\big(E_r\big)$. Ma poiché questi numeri sono tutti eguali a $\mathbf{m}\big(E\big)$, la somma della serie non può che prendere il valore 0 (se $\mathbf{m}\big(E\big) = 0$), oppure $+\infty$ (se $\mathbf{m}\big(E\big) > 0$).

Notiamo che l'esistenza di questo insieme E non è data in modo costruttivo (detto intuitivamente *non si riesce a capire come sia fatto questo insieme*) ma è una conseguenza dell'*assioma della scelta*: se non si accetta l'assioma della scelta questa costruzione cade.

È interessante osservare che questa difficoltà non sussiste con le funzioni *finitamente* additive: è sempre possibile infatti prolungare (in modo però non unico) una funzione finitamente additiva definita su un'algebra di parti di un insieme a tutti i sottinsiemi. Ancora una volta però questo prolungamento non è costruttivo, ma una conseguenza dell'assioma della scelta.

Vedremo più avanti invece che è possibile prolungare (in modo unico) una funzione σ-additiva definita su un'algebra $\mathcal{A}$ di parti di un insieme Ω alla più piccola σ-algebra che la contiene, e questo sarà fatto con un procedimento effettivamente costruttivo.

1.4.2 Probabilità e teoria dei numeri.

Ci sono delle interessanti applicazioni della nozione di Probabilità alla Teoria dei numeri; in questo primo corso non c'è il tempo di addentrarci in questo capitolo, ma ci limitiamo ad un paio di esempi.

Esempio 1.4.1 (**La funzione di Eulero).** Si chiama *funzione di Eulero* la funzione $\phi(n)$ eguale (per $n \geq 2$) al numero di interi tra $1, \ldots, n$ primi con n: la *formula di Eulero* afferma che, se $p_1, \ldots, p_m$ sono i divisori primi di n, si ha

$$\phi(n) = n\Big(1 - \frac{1}{p_1}\Big) \ldots \Big(1 - \frac{1}{p_m}\Big)$$

Di questa formula di può dare una dimostrazione probabilistica: più precisamente si considerino sullo spazio $\Omega = \{1, \ldots, n\}$ la distribuzione di probabilità uniforme ed i sottinsiemi $A(p_i)$ costituiti dai multipli di p_i (compresi tra 1 e n).

1) Provare che gli eventi $A(p_i)$ sono indipendenti (e di conseguenza anche i loro complementari).

2) Osservare che l'intersezione dei complementari degli insiemi $A(p_i)$ coincide con l'insieme gli interi primi con n e dedurne la formula di Eulero.

Esempio 1.4.2 (**La densità di Dirichlet).** Sia A un sottinsieme dell'insieme dei numeri naturali $\mathbb{N}$, e definiamo (per i sottinsiemi A per il quali questo limite esiste)

$$\mathbf{d}(A) \;=\; \lim_{n\to\infty} \frac{|A \cap \{1,\ldots,n\}|}{n}$$

La funzione sopra definita è un tipico esempio di funzione *semplicemente additiva* ma non σ-additiva.

a) Verificare che la funzione $\mathbf{d}$ è additiva ma non σ-additiva ed esibire un sottinsieme $B \subset \mathbb{N}$ tale che $\mathbf{d}(B)$ non sia definita.

b) Assegnato un intero p , calcolare la densità dell'insieme G_p formato dai multipli di p e provare che, se p e q sono primi tra loro, gli insiemi G_p e G_q risultano indipendenti.

N.B. La famiglia dei sottinsiemi A per i quali è definita la densità in realtà *non è un'algebra*: tale famiglia infatti è stabile per passaggio al complementare (e la verifica di questo è immediata), *ma non è stabile per l'unione.*

Provare questo fatto (così come esibire un sottinsieme B che non ha densità) è un esercizio impegnativo.

Capitolo 2

Probabilità e variabili aleatorie su uno spazio numerabile

2.1 Richiami sulle serie numeriche.

Premettiamo alcuni richiami sulle serie numeriche. Data una successione di numeri reali $a_1, a_2, \ldots$, posto $s_n = a_1 + \cdots + a_n$, si chiama *somma della serie* il limite (se esiste) della successione $(s_n)_{n\geq 1}$, e si dice che *la serie converge* se questo limite esiste. Più precisamente, per definizione

$$\sum_{n=1}^{+\infty} a_n \;=\; \lim_{n\to\infty} \sum_{k=1}^{n} a_k \;=\; \lim_{n\to\infty} s_n$$

Se la serie converge, la successione $(a_n)_{n\geq 1}$ è infinitesima (infatti si ha $a_n = s_n - s_{n-1}$) , ma non è vero il viceversa (un esempio tipico è la successione $a_n = \frac{1}{n}$).

Vediamo ora alcune proprietà importanti delle serie a termini positivi (cioè $a_n \geq 0$, qualunque sia n): in tal caso la successione delle somme parziali $(s_n)_{n\geq 1}$ è *monotona crescente* e pertanto esiste comunque (finito o infinito) il limite. Ha sempre senso quindi scrivere $\sum_{n=1}^{+\infty} a_n \in [0, +\infty]$.

Le serie a termini di segno positivo hanno interessanti proprietà, in particolare si può *cambiare l'ordine della somma* e *sommare per pacchetti*: di seguito vediamo gli enunciati precisi nelle due seguenti proposizioni, nelle quali si suppone che la successione $(a_n)_{n\geq 1}$ sia formata da termini positivi.

Proposizione 2.1.1. *Sia* $v : \mathbb{N} \to \mathbb{N}$ *una applicazione biunivoca: allora*

$$\sum_{n=1}^{+\infty} a_n \;=\; \sum_{n=1}^{+\infty} a_{v(n)}$$

Proposizione 2.1.2. *Sia $A_1, A_2, \ldots$ una partizione di $\mathbb{N}$ (non importa se formata di insiemi finiti o infiniti): vale la formula*

$$\sum_{n=1}^{+\infty} a_n = \sum_{n=1}^{+\infty} \sum_{k \in A_n} a_k$$

Dimostrazione. Dimostriamo 2.1.1, lasciando per esercizio la analoga dimostrazione di 2.1.2. Chiamiamo $r(n) = \max\big(v(1), \ldots, v(n)\big)$ e sia $s'_n = a_{v(1)} + \cdots + a_{v(n)}$: per ogni n si ha

$$s'_n \leq a_1 + \cdots + a_{r(n)} \leq \sum_{n=1}^{+\infty} a_n$$

e quindi, al limite,

$$\sum_{n=1}^{+\infty} a_{v(n)} \leq \sum_{n=1}^{+\infty} a_n$$

In modo analogo si ottiene la diseguaglianza opposta e di conseguenza l'eguaglianza. □

Queste due proprietà si estendono immediatamente alle serie *assolutamente convergenti*: ricordiamo che una serie numerica è detta assolutamente convergente se si ha

$$\sum_{n=1}^{+\infty} \big|a_n\big| < +\infty$$

Senza scrivere una formalizzazione esplicita, notiamo che la serie è assolutamente convergente se (e solo se) convergono a un numero reale sia la serie dei termini positivi che quella dei termini negativi, e ad entrambe si possono applicare i risultati di 2.1.1 e 2.1.2.

Esercizio 2.1.3. Provare con dei controesempi che se la serie è convergente ma non assolutamente convergente gli enunciati precedenti sono falsi.

In particolare vale questo curioso risultato, del quale non diamo la dimostrazione (che non ci servirà più avanti) lasciandola come esercizio *impegnativo*.

Proposizione 2.1.4. *Supponiamo che la successione $(a_n)_{n\geq 1}$ sia tale che la seria ad essa associata* converga ma non converga assolutamente*: assegnato un qualsiasi $l \in [-\infty, +\infty]$, è possibile determinare una funzione biunivoca $v : \mathbb{N} \to \mathbb{N}$ tale che si abbia*

$$\lim_{n\to\infty} \sum_{k=1}^{n} a_{v(k)} = l$$

Come *suggerimento*, possiamo invitare a osservare che i termini della successione devono essere infinitesimi (poichè la serie converge) ed entrambe le serie dei termini positivi e di quelli negativi della successione divergono.

Abbiamo visto in sostanza che proprietà veramente buone di sommabilità si hanno solo con serie assolutamente convergenti.

2.2 Integrale rispetto ad una misura discreta.

Quando la misura è definita su insieme numerabile la costruzione dell'integrale è particolarmente semplice, sostanzialmente è una conseguenza delle proprietà delle somme di serie numeriche: cominciamo dunque ad esaminare questo caso semplificato, esplicitando le proprietà fondamentali dell'integrale.

Consideriamo un insieme numerabile $E = \{e_1, e_2, \ldots\}$ sul quale sia definita una misura $\mathbf{m}$: supponiamo che tutti i sottinsiemi di E siano misurabili (come abbiamo detto nel capitolo precedente, sugli insiemi numerabili non ci sono problemi di misurabilità) e supponiamo che, per ogni i , $\mathbf{m}(e_i) < +\infty$ (c'è un piccolo abuso di notazioni perchè avremmo dovuto scrivere $\mathbf{m}\big(\{e_i\}\big)$, ma usiamo questa notazione abbreviata). Per ogni insieme $A \subset E$ si ha

$$\mathbf{m}\big(A\big) \;=\; \sum_{e_i \in A} \mathbf{m}(e_i)$$

Consideriamo ora una funzione $f : E \to \mathbb{R}$; non ci poniamo problemi di *misurabilità* (sui quali invece saremo più accurati nei capitoli successivi) perché ogni sottinsieme di E è misurabile.

Definizione 2.2.1 (**Integrale).** Si dice che la funzione f è integrabile se

$$\sum_i \big|f(e_i)\big|\, \mathbf{m}(e_i) \;<\; +\infty$$

ed in tal caso chiamiamo *integrale* di f il numero

$$\int f \,\mathrm{d}\mathbf{m} \;=\; \sum_i f(e_i)\, \mathbf{m}(e_i)$$

Indichiamo con $\mathcal{L}^1$ lo spazio delle funzioni integrabili. Prima di procedere con le proprietà essenziali dell'integrale, osserviamo che dai risultati sulle serie numeriche che sono stati ricordati risulta evidente perché si richiede che la serie dei termini $f(e_i)\mathbf{m}(e_i)$ converga assolutamente : senza questa condizione infatti, se scegliessi di numerare i punti dell'insieme E secondo un altro ordinamento, potrei avere per l'integrale un risultato diverso.

Osserviamo ancora che, se f è a valori positivi, ha sempre senso parlare di integrale di f, cioè $\int f \, \mathrm{d}\mathbf{m} = \sum_{i\geq 1} f(e_i)\mathbf{m}(e_i) \in [0, +\infty]$.

Lasciamo per esercizio le seguenti facili proprietà:

1. se $f, g \in \mathcal{L}^1$, anche $(af + g) \in \mathcal{L}^1$ e $\int (af + g)\mathrm{d}\mathbf{m} = a \int f \, \mathrm{d}\mathbf{m} + \int g \, \mathrm{d}\mathbf{m}$;
2. se $0 \leq f \leq g$, allora $\int f \, \mathrm{d}\mathbf{m} \leq \int g \, \mathrm{d}\mathbf{m}$;
3. f è integrabile se e solo se $\int |f| \, \mathrm{d}\mathbf{m} < +\infty$, inoltre $\left| \int f \, \mathrm{d}\mathbf{m} \right| \leq \int |f| \, \mathrm{d}\mathbf{m}$;
4. se $0 \leq f$ e $\int f \, \mathrm{d}\mathbf{m} = 0$, allora f vale identicamente 0 eccetto eventualmente su un insieme *trascurabile*.

Ricordiamo che si chiama *trascurabile* un insieme che ha misura nulla; una proprietà verificata ovunque eccetto che su un insieme trascurabile è detta valere *quasi ovunque* (e si scrive q.o.), mentre in probabilità si preferisce dire *quasi certamente* (e si scrive q.c.).

I due enunciati che seguono sono le proprietà più importanti di *passaggio al limite sotto il segno d'integrale*.

Teorema 2.2.2 (Beppo Levi). *Sia* $(f_n)_{n\geq 1}$ *una successione* crescente *di funzioni positive, convergente ad* f *: la successione degli integrali* $\left(\int f_n \, \mathrm{d}\mathbf{m} \right)_{n\geq 1}$ *converge (crescendo) a* $\int f \, \mathrm{d}\mathbf{m}$.

In maniera più sintetica, scriveremo d'ora innanzi un enunciato come il precedente nella forma

$$0 \leq f_n \quad , \quad f_n \uparrow f \quad \Longrightarrow \int f_n \, \mathrm{d}\mathbf{m} \uparrow \int f \, \mathrm{d}\mathbf{m}$$

Dimostrazione. Innanzi tutto osserviamo che esiste $\lim_{n\to\infty} \int f_n \, \mathrm{d}\mathbf{m}$ (poiché si tratta di una successione monotona crescente) e che tale limite è inferiore o eguale a $\int f \, \mathrm{d}\mathbf{m}$: occorre poi distinguere i casi in cui l'integrale di f sia finito o infinito.

Consideriamo il primo caso, e sia $A = \int f \, \mathrm{d}\mathbf{m}$; per ogni $\varepsilon > 0$, esiste un k tale che la somma finita $\sum_{i=1,\ldots,k} f(e_i)\mathbf{m}(e_i) \geq A - \varepsilon$. Poiché per ogni punto (e_i), $f_n(e_i)\mathbf{m}(e_i)$ converge a $f(e_i)\mathbf{m}(e_i)$, convergono anche le somme finite e si trova che, per n abbastanza grande $\int f_n \, \mathrm{d}\mathbf{m} \geq \sum_{i=1,\ldots,k} f_n(e_i)\mathbf{m}(e_i) \geq A - 2\varepsilon$, e questo completa la dimostrazione.

Il caso in cui $\int f \, \mathrm{d}\mathbf{m} = +\infty$ è sostanzialmente identico: qualunque sia $B > 0$, esiste un k tale che $\sum_{i=1,\ldots,k} f(e_i)\mathbf{m}(e_i) \geq B$, e con gli stessi passaggi appena svolti si prova che, per n abbastanza grande, $\int f_n \, \mathrm{d}\mathbf{m} \geq \frac{B}{2}$. □

Teorema 2.2.3 (Convergenza dominata). *Sia $(f_n)_{n\geq 1}$ una successione di funzioni convergente* puntualmente *ad f e supponiamo che esista g positiva integrabile tale che si abbia $|f_n| \leq g$ qualunque sia n: vale allora la relazione*

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mathbf{m} \;=\; \int f \,\mathrm{d}\mathbf{m}$$

Dimostrazione. Cominciamo ad osservare che la condizione di *dominazione* $|f_n| \leq g$ (valida ovviamente anche per il limite f) implica che ogni f_n ed f siano integrabili. Notiamo poi che si ha la maggiorazione

$$\Big| \int f_n \,\mathrm{d}\mathbf{m} - \int f \,\mathrm{d}\mathbf{m}\Big| \leq \int \big|f_n - f\big| \,\mathrm{d}\mathbf{m} = \sum_{i\geq 1} \big|f_n(e_i) - f(e_i)\big| \,\mathbf{m}(e_i)$$

Dato $\varepsilon > 0$, esiste un intero k tale che $\sum_{i=k+1}^{+\infty} g(e_i)\mathbf{m}(e_i) < \varepsilon$, e di conseguenza (poiché $\big|f_n(e_i) - f(e_i)\big| \leq 2\, g(e_i)$) , qualunque sia n, $\sum_{i=k+1}^{+\infty} \big|f_n(e_i) - f(e_i)\big|\mathbf{m}(e_i) < 2\,\varepsilon$.

A questo punto, poiché le somme finite convergono, per n abbastanza grande, $\sum_{i=1}^{k} \big|f_n(e_i) - f(e_i)\big|\mathbf{m}(e_i) < \varepsilon$ e quindi $\big|\int f_n \,\mathrm{d}\mathbf{m} - \int f \,\mathrm{d}\mathbf{m}\big| < 3\varepsilon$ e questo conclude la dimostrazione. □

Proviamo ora con un controesempio che nell'enunciato precedente, se si toglie l'ipotesi di *dominazione*, il risultato di passaggio al limite sotto il segno d'integrale non è più vero.

Esercizio 2.2.4. Consideriamo sullo spazio $\mathbb{N}^*$ degli interi (strettamente positivi) la misura $\mathbf{m}$ tale che $\mathbf{m}(k) = 2^{-k}$ (notiamo che si tratta di una *probabilità*), e consideriamo la successione di funzioni così definite:

$$f_n(k) \;=\; \begin{cases} 2^n & \text{se } k = n \\ 0 & \text{se } k \neq n \end{cases}$$

Verificare che le funzioni così definite sono integrabili, che la successione non è *dominata*, che converge puntualmente a una funzione integrabile ma gli integrali non convergono.

Sarà importante il seguente risultato:

Teorema 2.2.5 (Diseguaglianza di Schwartz). *Siano f, g tali che $\int f^2 \,\mathrm{d}\mathbf{m} < +\infty$ e $\int g^2 \,\mathrm{d}\mathbf{m} < +\infty$: allora il prodotto fg è integrabile e vale la diseguaglianza*

$$\left|\int fg \,\mathrm{d}\mathbf{m}\right| \;\leq\; \sqrt{\int f^2 \,\mathrm{d}\mathbf{m}} \quad \sqrt{\int g^2 \,\mathrm{d}\mathbf{m}}$$

Inoltre, se la diseguaglianza sopra scritta è una eguaglianza, le funzioni f e g coincidono a meno di una costante moltiplicativa (cioè esiste t reale tale che $f(e_i) = t\, g(e_i)$ q.o.).

Dimostrazione. Cominciamo ad osservare che la funzione fg è integrabile: si ha infatti, per ogni punto e_i, $|f(e_i)g(e_i)| \leq \big(f^2(e_i) + g^2(e_i)\big)$. Per ogni t reale, si ha

$$0 \leq \int \big(tf + g\big)^2 \,\mathrm{d}\mathbf{m} = t^2 \int f^2 \,\mathrm{d}\mathbf{m} + \int g^2 \,\mathrm{d}\mathbf{m} + 2t\Big(\int fg \,\mathrm{d}\mathbf{m}\Big)$$

La funzione sopra scritta è un polinomio di II grado in t, e se è a valori positivi il relativo discriminante è negativo, cioè
$\big(\int fg \,\mathrm{d}\mathbf{m}\big)^2 - \big(\int f^2 \,\mathrm{d}\mathbf{m}\big).\big(\int g^2 \,\mathrm{d}\mathbf{m}\big) \leq 0.$

Inoltre se il discriminante è eguale a 0, il polinomio si annulla in un punto t, cioè esiste $t \in \mathbb{R}$ tale che si abbia $\int \big(tf + g\big)^2 \,\mathrm{d}\mathbf{m} = 0$ e questo equivale a dire che $(tf + g) = 0$ q.o □

Osservazione 2.2.6. La teoria esposta in questo paragrafo rimane valida se l'insieme E non è numerabile, ma la misura $\mathbf{m}$ è *concentrata* su un insieme numerabile, più precisamente se esiste una successione di punti $(e_1, e_2, \ldots)$ tale che, per ogni $A \subset E$, si abbia

$$\mathbf{m}\big(A\big) \;=\; \sum_{e_i \in A} \mathbf{m}(e_i)$$

Infatti in questo caso il complementare dell'unione dei punti che formano la successione è *trascurabile* e, nel calcolo degli integrali, interessa solo il valore di una funzione nei punti $(e_i)_{i \geq 1}$. Si usa dire in questo caso che la misura è *discreta*, o anche *atomica*.

2.3 Variabili aleatorie discrete.

Consideriamo ora, in questo e nel successivo capitolo, uno spazio di probabilità $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ nel quale l'insieme Ω è supposto numerabile. Alla definizione di *variabile aleatoria* premettiamo un esempio.

Supponiamo di aver puntato alla roulette 1 E sul numero 28 ed 1 E sul pari: possiamo domandarci qual è la probabilità di vincere più di 10 E, oppure la probabilità di perdere. Lo spazio naturale per descrivere l'esito di un giro della roulette è l'insieme $\Omega = \{0, 1, \ldots, 36\}$ munito della distribuzione uniforme di probabilità, ma le domande scritte sopra non corrispondono direttamente a sottinsiemi di Ω.

Siamo naturalmente portati a introdurre una funzione $X : \Omega \to \mathbb{R}$ (la funzione *vittoria netta*) che in questo esempio risulta essere così definita:

$$X(\omega) \;=\; \begin{cases} 36 & \omega = 28 \\ 0 & \omega \text{ pari, } \omega \neq 28 \\ -1 & \omega = 0 \\ -2 & \omega \text{ dispari} \end{cases}$$

La risposta alla prima domanda diventa $\mathbf{P}\big\{\omega_i \big|\, X(\omega_i) \geq 10\big\} = \mathbf{P}\big(X^{-1}\big([10, +\infty[\big)\big) = \frac{1}{37}$ e la risposta alla seconda è $\mathbf{P}\big\{\omega_i \big|\, X(\omega_i) < 0\big\} = \mathbf{P}\big(X^{-1}\big(]-\infty, 0[\big)\big) = \frac{19}{37}$.

In definitiva, abbiamo naturalmente introdotto una funzione $X : \Omega \to \mathbb{R}$ ed abbiamo *trasportato* la probabilità dai sottinsiemi di Ω ai sottinsiemi di $\mathbb{R}$.

Definizione 2.3.1 (**Variabile aleatoria).** Assegnato uno spazio di probabilità $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$ con Ω numerabile, si chiama *variabile aleatoria* reale (discreta) una funzione $X : \Omega \to \mathbb{R}$.

Definizione 2.3.2 (**Legge di Probabilità).** Si chiama *legge di probabilità* (o anche *distribuzione di probabilità*) della v.a. reale X la probabilità definita sui sottinsiemi di $\mathbb{R}$ dalla formula

$$\mathbf{P}_X\big(A\big) \;=\; \mathbf{P}\big(X^{-1}(A)\big)$$

La probabilità $\mathbf{P}_X$ viene anche chiamata la *probabilità immagine* (di $\mathbf{P}$ mediante X) e indicata $X\big(\mathbf{P}\big)$. Che si tratti effettivamente di una probabilità è immediato: se $(A_n)_{n\geq 1}$ è una successione di sottinsiemi di $\mathbb{R}$ a due a due disgiunti, anche le immagini inverse sono disgiunte e si ha

$$\mathbf{P}_X\big(\bigcup_n A_n\big) = \mathbf{P}\big(X^{-1}\big(\bigcup_n A_n\big)\big) = \sum_n \mathbf{P}\big(X^{-1}\big(A_n\big)\big) = \sum_n \mathbf{P}_X\big(A_n\big)$$

Si verifica inoltre immediatamente che $\mathbf{P}_X\big(\mathbb{R}\big) = 1$. È anche immediato constatare che l'immagine di una probabilità è *associativa* nel senso che, se $Y = g \circ X$, si ha $Y(\mathbf{P}) = \big(g \circ X\big)(\mathbf{P}) = g\big(\, X(\mathbf{P})\big)$.

Quando due variabili aleatorie hanno la stessa legge di probabilità sono dette **equidistribuite** (o anche *isonome*).

Vediamo più in dettaglio come è fatta la legge di probabilità di una v.a. discreta.

Poiché Ω è numerabile, anche l'immagine di X è un sottinsieme (finito o) numerabile della retta, cioè $(x_1, x_2, \ldots)$; per ogni punto x_i , si consideri il numero $p(x_i) = \mathbf{P}\big\{X = x_i\big\} = \mathbf{P}\big(X^{-1}(x_i)\big)$. Vale la formula:

$$\mathbf{P}_X\big(A\big) = \mathbf{P}\big(X^{-1}(A)\big) = \sum_{x_i \in A} p(x_i)$$

(infatti $X^{-1}(A) = \bigcup_{x_i \in A} \{X = x_i\}$). Naturalmente i numeri $p(x_i)$ sono positivi e $\sum_i p(x_i) = 1$; alla funzione $x \to p(x) = \mathbf{P}\{X = x\}$ viene dato il nome di *funzione di probabilità* (qualcuno usa anche il termine *densità discreta*).

Quanto alla scrittura $\{X = x\}$, è bene familiarizzarsi subito con la notazione (molto comoda) $\{X \in A\} = \{\omega_i \,|\, X(\omega_i) \in A\} = X^{-1}(A)$. Ad esempio $\{a < X \leq b\} = X^{-1}(]a, b])$.

Osservazione 2.3.3. Assegnata una *probabilità discreta* $\mathbf{Q}$ su $\mathbb{R}$ (cioè in pratica, come abbiamo visto, del valori $(x_1, x_2, \ldots)$ e dei numeri positivi $\big(p(x_1), p(x_2), \ldots\big)$ con $\sum_i p(x_i) = 1$) è naturale chiedersi se esiste una v.a. X la cui legge di probabilità sia $\mathbf{Q}$.

La risposta è affermativa e la costruzione è anche molto semplice: si può considerare come Ω l'insieme dei valori $\Omega = \{x_1, x_2, \ldots\}$, come probabilità $\mathbf{P}$ quella definita da $\mathbf{P}\big(\{x_i\}\big) = p(x_i)$ e come applicazione $X : \Omega \to \mathbb{R}$ l'applicazione identica (cioè $X(x_i) = x_i$). La verifica dell'eguaglianza $\mathbf{P}_X = \mathbf{Q}$ è immediata.

Questa osservazione sembra banale, ma dal punto di vista metodologico è invece importante: nella pratica spesso si incontra solo la *legge di probabilità* di una v.a., e questo ci dice che non dobbiamo porci domande sull'esistenza di uno spazio Ω e di una applicazione $X : \Omega \to \mathbb{R}$ perché la risposta è già data da questa costruzione canonica.

Vediamo ora rapidamente le principali variabili aleatorie discrete.

Esempio 2.3.4 (**Variabile Binomiale).** La variabile Binomiale (di parametri n e p, n intero positivo e $0 < p < 1$), considera n ripetizioni (in condizioni di indipendenza) di un esperimento che ha probabilità p di successo e conta il numero dei successi ottenuti. La legge binomiale viene indicata $B(n, p)$ e si scrive $X \sim B(n, p)$; quando $n = 1$ viene anche chiamata **legge di Bernoulli di parametro** p.

I valori della v.a. binomiale sono gli interi $\{0, 1, \ldots, n\}$ e vale, per $0 \leq k \leq n$, la formula

$$p(k) = \mathbf{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}$$

Esempio 2.3.5 (**Variabile di Poisson).** La variabile di Poisson (di parametro λ , $\lambda > 0$) è una variabile che assume tutti i valori interi positivi con probabilità

$$p(n) = \mathbf{P}\{X = n\} = \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!}$$

Esempio 2.3.6 (**Variabile Geometrica).** La variabile Geometrica (di parametro p , $0 < p < 1$) considera ripetizioni consecutive di un esperimento che ha probabilità p di successo e conta il numero di prove che è stato necessario effettuare per ottenere un successo.

I valori possibili sono gli interi strettamente positivi e si ha

$$p(n) = \mathbf{P}\big\{X = n\big\} = (1-p)^{n-1}p$$

Esercizio 2.3.7 (**Assenza di memoria della legge geometrica).** Provare che se X è una variabile geometrica, per n, h interi strettamente positivi, vale la formula

$$\mathbf{P}\big\{X = n + h \big| X > n\big\} = \mathbf{P}\big\{X = h\big\} \tag{2.3.1}$$

Provare viceversa che se X è una v.a. a valori interi strettamente positivi che soddisfa l'equazione 2.3.1, necessariamente è una variabile geometrica.

Esercizio 2.3.8 (**Variabile Binomiale negativa.).** La variabile Binomiale negativa può essere definita in questo modo: si ripete in condizioni di indipendenza un esperimento che ha probabilità p di successo fino a che questo si realizza k volte; la variabile conta il numero di tentativi che è stato necessario effettuare. Determinare la sua legge di probabilità.

Osservazione: il nome, un pò curioso, di binomiale negativa, deriva dall'eguaglianza

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

Ricordiamo che, se α è un numero reale qualsiasi e k un intero positivo, per definizione

$$\binom{\alpha}{k} = \frac{\alpha.(\alpha-1)\dots(\alpha-k+1)}{k!}$$

Esercizio 2.3.9 (**Variabile ipergeometrica).** Consideriamo un'urna contenente r sfere rosse e b sfere bianche, ed in essa compiamo n estrazioni *senza reimbussolamento* (ovviamente si deve avere $n \leq (r+b)$): consideriamo la v.a. X che conta il numero di sfere rosse che sono state estratte.

Di tale variabile determinare la distribuzione di probabilità, il valore atteso, la varianza.

2.4 Valori attesi e momenti.

Prima di dare la definizione di *valore atteso*, proviamo un teorema che si dimostra fondamentale in Calcolo delle Probabilità.

Teorema 2.4.1 (**Integrazione rispetto a una probabilità immagine).** *Siano X una v.a. discreta, $\mathbf{P}_X = X\big(\mathbf{P}\big)$ la sua* legge di probabilità *e $\varphi : \mathbb{R} \to \mathbb{R}$. φ è integrabile rispetto a $\mathbf{P}_X$ se e solo se $\varphi \circ X$ è integrabile rispetto a $\mathbf{P}$, e in tal caso vale l'eguaglianza*

$$\int_{\mathbb{R}} \varphi(x)\,\mathrm{d}\mathbf{P}_X(x) \;=\; \int_{\Omega} \varphi\big(X(\omega)\big)\,\mathrm{d}\mathbf{P}(\omega) \tag{2.4.1}$$

Dimostrazione. Cominciamo a supporre che φ sia a valori positivi. Poiché Ω è numerabile, la sua immagine mediante X è un sottinsieme (finito o) numerabile di $\mathbb{R}$ della forma $(x_1, x_2, \ldots)$. Consideriamo gli insiemi $A_i = \{X = x_i\} = \big\{\omega_j \,\big|\, X(\omega_j) = x_i\big\}$ e osserviamo che $p(x_i) = \sum_{\omega_j \in A_i} \mathbf{P}(\omega_j)$. Poichè quelle che seguono sono somme di serie a termini positivi, possiamo usare la proprietà associativa della somma: si ottiene pertanto

$$\int \varphi(x)\,\mathrm{d}\mathbf{P}_X(x) \;=\; \sum_i \varphi(x_i)p(x_i) \;=\; \sum_i \varphi(x_i)\Big(\sum_{\omega_j \in A_i} \mathbf{P}(\omega_j)\Big) =$$

$$\sum_i \Big(\sum_{\omega_j \in A_i} \varphi\big(X(\omega_j)\big)\mathbf{P}(\omega_j)\Big) \;=\; \sum_j \varphi\big(X(\omega_j)\big)\mathbf{P}(\omega_j) \;=\; \int_{\Omega} \varphi\big(X(\omega)\big)\,\mathrm{d}\mathbf{P}(\omega)$$

cioè l'eguaglianza desiderata. Il caso generale si ottiene scrivendo la funzione φ nella forma $\varphi = \varphi^+ - \varphi^-$ e sommando i due integrali. Ricordiamo che con $\varphi^+(x) = \max\big(\varphi(x), 0\big)$ e $\varphi^-(x) = -\min\big(\varphi(x), 0\big)$ intendiamo la *parte positiva* e *parte negativa* della funzione φ. □

Siamo ora in grado di dare la seguente definizione:

Definizione 2.4.2 (**Valore atteso).** Data una v.a. reale discreta X, si dice che essa ha *valore atteso* se è integrabile rispetto a $\mathbf{P}$, e in tal caso si chiama valore atteso l'integrale

$$\mathbf{E}\big[X\big] \;=\; \int_{\Omega} X(\omega)\,\mathrm{d}\mathbf{P}(\omega) \;=\; \sum_i X\big(\omega_i\big)\mathbf{P}\big(\omega_i\big)$$

Il valore atteso è anche chiamato *speranza matematica*; il termine anglosassone è *expectation* e quello francese *espérance*. Talvolte viene anche chiamato valor medio, ma è un termine improprio perché si potrebbe confondere con la media aritmetica dei valori della v.a. (quando questa prende un numero finito di valori).

In base al teorema 2.4.1 abbiamo la seguente regola pratica: data una v.a. discreta che prende i valori $(x_1, x_2, \ldots)$ con probabilità $\big(p(x_1), p(x_2), \ldots\big)$,

essa ammette valore atteso se e solo se $\sum_i |x_i|\, p(x_i) < +\infty$, ed in tal caso si ha $\mathbf{E}[X] = \sum_i x_i\, p(x_i)$.

Dalle proprietà dell'integrale derivano alcune proprietà immediate del valore atteso, ad esempio (se esiste) $\mathbf{E}[aX + b] = a\,\mathbf{E}[X] + b$.

Notiamo anche che se X è a valori positivi, ha sempre senso scrivere $\mathbf{E}[X] = \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) \in [0, +\infty]$.

Esercizio 2.4.3. Sia X una variabile aleatoria a valori interi positivi: provare che vale la formula

$$\mathbf{E}\big[X\big] \;=\; \sum_{n\geq 0} \mathbf{P}\big\{X > n\big\} \;=\; \sum_{n\geq 1} \mathbf{P}\big\{X \geq n\big\}$$

Definizione 2.4.4 (Momenti). Sia $1 \leq p < +\infty$ e X una v.a.: si chiama *momento assoluto* di ordine p il numero

$$\mathbf{E}\big[\big|X\big|^p\big] \;=\; \sum_i |x_i|^p\, p(x_i) \in [0, +\infty]$$

e se questo numero risulta finito, si dice che X ammette momento di ordine p. Dato un intero positivo n, se X ammette momento di ordine n, si chiama *momento di ordine n* il numero $\mathbf{E}\big[X^n\big]$.

Proposizione 2.4.5. *Siano $1 \leq p < q < +\infty$: se X ha momento di ordine q, ammette anche momento di ordine p.*

Dimostrazione. Per ogni numero reale x, vale la diseguaglianza $|x|^p \leq 1 + |x|^q$: si ha pertanto

$$\mathbf{E}\big[\big|X\big|^p\big] = \sum_i |x_i|^p\, p(x_i) \leq \sum_i \big(1 + |x_i|^q\big) p(x_i) = 1 + \mathbf{E}\big[\big|X\big|^q\big]$$

□

Osservazione 2.4.6. La dimostrazione sopra riportata (che è sufficiente per gli scopi di questo corso) è piuttosto rudimentale: il risultato in realtà è conseguenza di una diseguaglianza molto più precisa e importante (*diseguaglianza di Hölder*) che verrà presentata in corsi più avanzati.

Definizione 2.4.7 (Varianza). Sia X una variabile aleatoria dotata di momento secondo: si chiama *Varianza* di X il numero

$$Var\big(X\big) \;=\; \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)^2\big] \;=\; \mathbf{E}\big[X^2\big] - \mathbf{E}[X]^2$$

Esercizio 2.4.8. Provare che vale la relazione $Var\big(aX + b\big) \;=\; a^2\, Var\big(X\big)$.

Lemma 2.4.9 (**Diseguaglianza di Markov**). *Sia X una v.a. a valori positivi e t una costante positiva: vale la diseguaglianza*

$$t\,\mathbf{P}\big\{X \geq t\big\} \leq \mathbf{E}\big[X\big]$$

Dimostrazione. Introduciamo una notazione: se A è un insieme, si denota con I_A la *funzione indicatrice* dell'insieme A, più precisamente

$$I_A(\omega) \,=\, \begin{cases} 1 & \text{se } \omega \in A \\ 0 & \text{se } \omega \notin A \end{cases}$$

Si parte dunque dalla diseguaglianza tra variabili aleatorie $t\, I_{\{X \geq t\}} \leq X$, e passando alla conseguente diseguaglianza per gli integrali si ottiene il risultato. □

Conseguenza immediata della diseguaglianza di Markov è la seguente, che spiega perché la varianza è una *misura della dispersione* di una variabile aleatoria.

Proposizione 2.4.10 (**Diseguaglianza di Chebishev**). *Sia X una v.a. dotata di momento secondo: vale la diseguaglianza*

$$t^2\,\mathbf{P}\big\{\big|X - \mathbf{E}[X]\big| \geq t\big\} \leq Var\big(X\big)$$

Dimostrazione. Si applica la **diseguaglianza di Markov**, considerando come costante positiva t^2 e come variabile aleatoria $\big(X - \mathbf{E}[X]\big)^2$: si noti che

$$\big\{\big|X - \mathbf{E}[X]\big| \geq t\big\} \,=\, \big\{\big(X - \mathbf{E}[X]\big)^2 \geq t^2\big\}$$

□

Corollario 2.4.11. *La varianza di una v.a. X è eguale a 0 se e solo se X è* costante *q.c.*

Dimostrazione. Da una parte, se $X = c$ q.c., si ha $\mathbf{E}[X] = c$ e $\mathbf{E}[X^2] = c^2$ e quindi la varianza si annulla. Supponiamo viceversa che $Var\big(X\big) = 0$: poiché

$$\big\{\big|X - \mathbf{E}[X]\big| \neq 0\big\} \,=\, \bigcup_{n \geq 1} \big\{\big|X - \mathbf{E}[X]\big| \geq \frac{1}{n}\big\}$$

e ciascuno degli insiemi $\big\{\big|X - \mathbf{E}[X]\big| \geq \frac{1}{n}\big\}$ è trascurabile, anche $\big\{\big|X - \mathbf{E}[X]\big| \neq 0\big\}$ è trascurabile. □

2.5 Variabili aleatorie a più dimensioni, variabili aleatorie indipendenti.

Per semplicità di notazioni, trattiamo il caso di variabili aleatorie a valori in $\mathbb{R}^2$, ma identica è la trattazione di variabili aleatorie a valori in $\mathbb{R}^n$. Consideriamo dunque una variabile aleatoria *doppia* o *bidimensionale*, cioè una applicazione $\big(X,Y\big) : \Omega \longrightarrow \mathbb{R}^2$. La sua *legge di probabilità* (denotata $\mathbf{P}_{X,Y} = (X,Y)\big(\mathbf{P}\big)$ è una probabilità sui sottinsiemi di $\mathbb{R}^2$.

L'immagine di (X,Y) è un sottinsieme numerabile di $\mathbb{R}^2$ cioè un insieme di punti $\big\{(x_i,y_j)\big|\, i \geq 1, j \geq 1\big\}$ e la *funzione di probabilità* è definita da $p(x_i,y_j) = \mathbf{P}\big\{X = x_i, Y = y_j\big\}$. Per ogni sottinsieme $B \subset \mathbb{R}^2$ si ha

$$\mathbf{P}_{X,Y}\big(B\big) \;=\; \mathbf{P}\big\{(X,Y) \in B\big\} \;=\; \sum_{(x_i,y_j)\in B} p(x_i,y_j)$$

Teniamo presente che nelle formule la *virgola* sta per la *congiunzione*, che corrisponde insiemisticamente all'intersezione, cioè ad esempio

$$\big\{X = x_i\,, Y = y_j\big\} \;=\; (X,Y)^{-1}(x_i,y_j) \;=\; \big\{X = x_i\big\} \cap \big\{Y = y_j\big\}$$

Il *teorema di integrazione rispetto ad una misura immagine* 2.4.1 si traduce con minimi cambiamenti formali: valgono pertanto le eguaglianze

$$\mathbf{E}\big[\varphi(X,Y)\big] \;=\; \int_\Omega \varphi\big(X(\omega),Y(\omega)\big)\,\mathrm{d}\mathbf{P}(\omega) \;=\; \iint_{\mathbb{R}^2} \varphi(x,y)\,\mathrm{d}\mathbf{P}_{X,Y}(x,y) \;=$$

$$= \sum_{x_i,y_j} \varphi(x_i,y_j)\,p(x_i,y_j)$$

che si deve leggere: *$\varphi(X,Y)$ è integrabile rispetto a $\mathbf{P}$ se e solo se φ è integrabile rispetto a $\mathbf{P}_{X,Y}$, ed in tal caso è soddisfatta la formula scritta sopra.* Da questa formula e dalle proprietà dell'integrale seguono conseguenze immediate: ad esempio, se X e Y sono integrabili, vale l'eguaglianza $\mathbf{E}\big[X+Y\big] = \mathbf{E}[X] + \mathbf{E}[Y]$.

Definizione 2.5.1 (Covarianza). Supponiamo che X ed Y ammettano momento secondo: si chiama *covarianza* il numero

$$Cov\big(X,Y\big) \;=\; \mathbf{E}\big[\big(X - \mathbf{E}[X]\big)\big(Y - \mathbf{E}[Y]\big)\big] \;=\; \mathbf{E}\big[XY\big] - \mathbf{E}[X]\,\mathbf{E}[Y]$$

Notiamo che se X,Y ammettono momento secondo, per la diseguaglianza di Schwartz (teorema 2.2.5) il prodotto XY ammette momento primo. Notiamo ancora che $Var\big(X\big) = Cov\big(X,X\big)$; è immediato verificare che la

covarianza è bilineare $\big(\ Cov(aX+bY,Z) = a\,Cov(X,Z) + b\,Cov(Y,Z)\ \big)$ e che vale la formula

$$Var\big(X+Y\big) \;=\; Var\big(X\big) + Var\big(Y\big) + 2\,Cov\big(X,Y\big)$$

Se $Cov\big(X,Y\big) = 0$, le due variabili sono dette *incorrelate*.

Proposizione 2.5.2. *Siano $X\,, Y$ dotate di momento secondo: vale la diseguaglianza*

$$\Big|Cov\big(X,Y\big)\Big| \;\le \sqrt{Var\big(X\big)}\sqrt{Var\big(Y\big)}$$

Dimostrazione. È una conseguenza immediata della diseguaglianza di Schwartz 2.2.5, dove si è posto $f = \big(X - \mathbf{E}[X]\big)$ e $g = \big(Y - \mathbf{E}[Y]\big)$. Si ha dunque

$$\Big|Cov\big(X,Y\big)\Big| \;=\; \Big|\int (X-\mathbf{E}[X])\big(Y-\mathbf{E}[Y]\big)\mathrm{d}\mathbf{P}\Big| \le$$

$$\le \sqrt{\int \big(X-\mathbf{E}[X]\big)^2\mathrm{d}\mathbf{P}}\sqrt{\int \big(Y-\mathbf{E}[Y]\big)^2\mathrm{d}\mathbf{P}} \;=\; \sqrt{Var\big(X\big)}\sqrt{Var\big(Y\big)}$$

□

Si chiama *scarto quadratico medio* di X la radice della sua varianza (se esiste); e se $X\,, Y$ ammettono momento secondo e non sono costanti, si chiama *coefficiente di correlazione* il numero

$$\rho\big(X,Y\big) \;=\; \frac{Cov\big(X,Y\big)}{\sqrt{Var\big(X\big)}\sqrt{Var\big(Y\big)}}$$

Esempio 2.5.3 (**Retta di regressione).** Supponiamo che le due variabili X e Y siano dotate di momento secondo e con varianza strettamente positiva e cerchiamo

$$\min_{a,b} \mathbf{E}\Big[\big(Y-aX-b\big)^2\Big]$$

Verificare che la funzione $Q(a,b) = \mathbf{E}\big[(Y-aX-b)^2\big]$ tende a $+\infty$ per $|(a,b)| \to \infty$, che il gradiente di Q si annulla solo nel punto $(\overline{a},\overline{b})$ dove $\overline{a} = \dfrac{Cov(X,Y)}{Var(X)}$ e $\overline{b} = \mathbf{E}[Y] - \overline{a}\,\mathbf{E}[X]$ e che vale l'eguaglianza

$$Q(\overline{a}\,,\,\overline{b}) = \min_{a,b} \mathbf{E}\Big[\big(Y-aX-b\big)^2\Big] = Var(Y)\,\big(1-\rho(X,Y)^2\big)$$

Lasciamo per esercizio la dimostrazione della seguente proprietà della covarianza:

Proposizione 2.5.4 (**Matrice delle covarianze**). *Sia* $\big(X_1, \ldots, X_n\big)$ *una variabile aleatoria* n*–dimensionale, supponiamo che ogni componente* X_i *abbia momento secondo e indichiamo con* C *la* matrice delle covarianze *(cioè* $C_{ij} = Cov(X_i, X_j)$*).*

La matrice C *è* simmetrica, semidefinita positiva*; inoltre vale la formula*

$$Var\Big(\sum_{i=1}^{n} a_i X_i\Big) \;=\; \sum_{i,j=1}^{n} C_{ij}\, a_i a_j$$

Torniamo ad una variabile doppia (X, Y), la cui legge di probabilità è identificata dalla funzione di probabilità $p(x_i, y_j)$; ognuna delle due componenti X ed Y è una v.a. reale, e indichiamo con $p_X(x_i) = \mathbf{P}\{X = x_i\}$ (e analogamente per p_Y) le relative funzioni di probabilità.

Proposizione 2.5.5. *Valgono le formule*

$$p_X(x_i) = \sum_{y_j} p(x_i, y_j) \qquad\qquad p_Y(y_j) = \sum_{x_i} p(x_i, y_j)$$

Dimostrazione. L'insieme $\big\{X = x_i\big\}$ è unione numerabile degli insiemi (a due a due disgiunti) $\big\{X = x_i\,,\, Y = y_j\big\}$, $\big(j = 1, 2, \ldots\big)$; si ha pertanto

$$p_x(x_i) \;=\; \mathbf{P}\big\{X = x_i\big\} \;=\; \sum_{y_j} \mathbf{P}\big\{X = x_i\,,\, Y = y_j\big\} \;=\; \sum_{y_j} p(x_i, y_j)$$

□

Viceversa, conoscendo le distribuzioni di probabilità marginali delle componenti X ed Y, non si può ricostruire la distribuzione di probabilità globale del vettore aleatorio (X, Y). C'è tuttavia un caso nel quale questo si può fare, ed è quando le due variabili sono *indipendenti*.

Definizione 2.5.6. Due variabili aleatorie X ed Y si dicono *indipendenti* se, scelti comunque due sottinsiemi A e B di $\mathbb{R}$, gli eventi $X^{-1}(A)$ e $Y^{-1}(B)$ sono indipendenti, cioè se vale la formula

$$\mathbf{P}\big\{X \in A\,,\, Y \in B\big\} \;=\; \mathbf{P}\big\{X \in A\big\}\; \mathbf{P}\big\{Y \in B\big\}$$

Proposizione 2.5.7. *Due variabili discrete* X *ed* Y *sono indipendenti se e solo se le relative funzioni di probabilità sono legate dalla formula*

$$p(x_i, y_j) \;=\; p_X(x_i)\, p_Y(y_j) \tag{2.5.1}$$

Dimostrazione. Da una parte, se le variabili sono indipendenti, scegliendo $A = \{x_i\}$ e $B = \{y_j\}$, si verifica immediatamente che è soddisfatta la formula 2.5.1.

Supponiamo viceversa che la formula 2.5.1 sia soddisfatta, e scegliamo due sottinsiemi A e B di $\mathbb{R}$: si ha

$$\mathbf{P}\big\{X \in A\,,\, Y \in B\big\} \;=\; \sum_{x_i \in A\,,\, y_j \in B} p(x_i, y_j) \;=\; \sum_{x_i \in A} \sum_{y_j \in B} p_X(x_i) p_Y(y_j) \;=$$

$$= \Big(\sum_{x_i \in A} p_X(x_i)\Big)\Big(\sum_{y_j \in B} p_Y(y_j)\Big) \;=\; \mathbf{P}\big\{X \in A\big\}\; \mathbf{P}\big\{Y \in B\big\}$$

□

La nozione di indipendenza tra variabili aleatorie può essere formulata in un altro modo, più opportuno per successive dimostrazioni, ma dobbiamo premettere una definizione.

Definizione 2.5.8 (Probabilità prodotto). Siano $\mathbf{P}_1$ e $\mathbf{P}_2$ due probabilità sui sottinsiemi di $\mathbb{R}$: si chiama *probabilità prodotto* (e si indica $\mathbf{P}_1 \otimes \mathbf{P}_2$) la probabilità definita sui sottinsiemi di $\mathbb{R}^2$ tale che, se A, B sono sottinsiemi di $\mathbb{R}$, si abbia

$$\mathbf{P}_1 \otimes \mathbf{P}_2\big(A \times B\big) \;=\; \mathbf{P}_1\big(A\big)\; \mathbf{P}_2\big(B\big)$$

Naturalmente nella definizione appena data non è necessario che le due probabilità siano definite sui sottinsiemi di $\mathbb{R}$, ma si adatta senza modifiche a due probabilità discrete definite su due generici insiemi E_1 e E_2.

Nella definizione 2.5.8, occorre precisare quali sottinsiemi di $\mathbb{R}^2$ si considerano misurabili e come si costruisce effettivamente la probabilità prodotto (ci occuperemo di questi problemi nei successivi capitoli), ma se $\mathbf{P}_1$ e $\mathbf{P}_2$ sono probabilità *discrete* la costruzione è immediata. Più precisamente, se $\mathbf{P}_1$ (rispettivamente $\mathbf{P}_2$) è concentrata nei punti $(x_1, x_2, \ldots)$ (risp. $(y_1, y_2, \ldots)$) con funzione di probabilità $p_1(.)$ (risp. $p_2(.)$), la probabilità $\mathbf{P}_1 \otimes \mathbf{P}_2$ è la probabilità discreta concentrata nelle coppie di punti (x_i, y_j) con funzione di probabilità

$$p(x_i, y_j) = \mathbf{P}_1 \otimes \mathbf{P}_2\big(\{x_i, y_j\}\big) = p_1(x_i).p_2(y_j)$$

La verifica di questo fatto è sostanzialmente identica alla dimostrazione della proposizione 2.5.1, e una conseguenza immediata è la dimostrazione della seguente proprietà

Proposizione 2.5.9. *Due variabili aleatorie X_1, X_2 sono indipendenti se e solo se la legge di probabilità congiunta è il prodotto delle singole leggi, cioè se si ha*

$$\mathbf{P}_{X_1, X_2} \;=\; \mathbf{P}_{X_1} \otimes \mathbf{P}_{X_2}$$

La proprietà precedente (che potrebbe equivalentemente essere assunta come *definizione* di indipendenza) ammette una evidente estensione alla definizione di indipendenza per n variabili aleatorie $(X_1, \dots, X_n)$.

Cominciamo ad osservare che la definizione 2.5.8 si estende senza difficoltà al prodotto di 3 o più probabilità, purchè in numero finito: si constata inoltre facilmente che il prodotto è *associativo* nel senso che, ad esempio,

$$\mathbf{P}_1 \otimes \mathbf{P}_2 \otimes \mathbf{P}_3 = \big(\mathbf{P}_1 \otimes \mathbf{P}_2\big) \otimes \mathbf{P}_3 = \mathbf{P}_1 \otimes \big(\mathbf{P}_2 \otimes \mathbf{P}_3\big)$$

Di conseguenza si può dire, per **definizione**, che n v.a. $X_1, \dots, X_n$ sono indipendenti se la legge congiunta è il prodotto delle singole leggi, cioè se si ha

$$\mathbf{P}_{X_1,\dots,X_n} = \mathbf{P}_{X_1} \otimes \cdots \otimes \mathbf{P}_{X_n}$$

Osservazione 2.5.10. Vediamo come si può estendere la costruzione dell'osservazione 2.3.3 al caso n-dimensionale, cioè, assegnate n probabilità (discrete) $\mathbf{P}_1, \dots, \mathbf{P}_n$, come si possono costruire n v.a. indipendenti $X_1, \dots, X_n$ con legge rispettivamente $\mathbf{P}_1, \dots, \mathbf{P}_n$. Questa costruzione sarà molto usata nei modelli statistici.

Supponiamo che tutte le probabilità siano concentrate sullo stesso sottinsieme numerabile $C \subset \mathbb{R}$ (ci si può sempre ridurre a questa situazione), poniamo $\Omega = C^n$ (il prodotto cartesiano di C con sé stesso n volte) e su di esso mettiamo la probabilità prodotto $\mathbf{P}_1 \otimes \cdots \otimes \mathbf{P}_n$; sia poi X_i la *proiezione canonica* di indice i, cioè $X_i(x_1, \dots, x_n) = x_i$. È immediato constatare che $\mathbf{P}_{X_i} = X_i(\mathbf{P}) = \mathbf{P}_i$ e che (poichè la legge del vettore aleatorio $\mathbf{X} = (X_1, \dots, X_n)$ è il prodotto delle singole leggi) queste variabili sono indipendenti.

Proposizione 2.5.11. *Siano X, Y due v.a. indipendenti e f, g due funzioni reali: le variabili $f \circ X$ e $g \circ Y$ sono indipendenti.*

Dimostrazione. Dati due sottinsiemi A, B di $\mathbb{R}$, gli eventi $\big\{\big(f \circ X\big) \in A\big\} = \big\{X \in f^{-1}(A)\big\}$ e $\big\{\big(g \circ Y\big) \in B\big\} = \big\{Y \in g^{-1}(B)\big\}$ sono evidentemente indipendenti. □

Il risultato della Proposizione 2.5.11 si estende al caso di più variabili in questo modo: *funzioni di variabili aleatorie indipendenti che non coinvolgano la stessa variabile sono ancora indipendenti.* Per capirci meglio, se (X, Y, Z) sono indipendenti, anche $f(X, Y)$ e $g(Z)$ sono indipendenti, *ma non lo sono* $f(X, Y)$ e $g(Y, Z)$.

La prova di questa affermazione è una conseguenza dell'eguaglianza

$$\mathbf{P}_X \otimes \mathbf{P}_Y \otimes \mathbf{P}_Z \;=\; \big(\mathbf{P}_X \otimes \mathbf{P}_Y\big) \otimes \mathbf{P}_Z$$

che si può leggere nel modo seguente: *la coppia* (X, Y) *è indipendente dalla variabile* Z. Le estensioni di queste affermazioni a più variabili sono evidenti.

È istruttivo dimostrare il seguente risultato:

Proposizione 2.5.12. *Dati* n *eventi* $\big(A_1, \dots, A_n\big)$ *, questi sono indipendenti se e solo se le loro funzioni indicatrici* $\big(I_{A_1}, \dots, I_{A_n}\big)$ *sono indipendenti come variabili aleatorie.*

Definizione 2.5.13. Data una famiglia qualsiasi di variabili aleatorie $(X_i)_{i\in I}$, queste si dicono *indipendenti* se ogni sottofamiglia finita $\big(X_{i_1}, \dots, X_{i_n}\big)$ è formata da variabili indipendenti.

Abbiamo visto (diseguaglianza di Schwartz) che il prodotto di due v.a. di quadrato integrabile è integrabile, ma non è detto che il prodotto di due variabili integrabili sia integrabile (cercare un controesempio!). Tuttavia con le variabili indipendenti si ha il seguente risultato:

Teorema 2.5.14. *Siano* X, Y *due variabili indipendenti dotate di momento primo: anche* XY *ammette momento primo e vale la formula*

$$\mathbf{E}\big[\, XY \,\big] \;=\; \mathbf{E}\big[X\big]\, \mathbf{E}\big[Y\big]$$

Dimostrazione. Cominciamo a provare che XY è integrabile: si ha infatti

$$\mathbf{E}\big[\, \big|XY\big| \,\big] \;=\; \sum_{x_i, y_j} |x_i y_j|\, p(x_i, y_j) \;=\; \sum_{x_i} \sum_{y_j} |x_i||y_j| p_X(x_i) p_Y(y_j) \;=$$

$$=\; \Big(\sum_{x_i} |x_i|\, p_X(x_i)\Big)\Big(\sum_{y_j} |y_j|\, p_Y(y_j)\Big) \;=\; \mathbf{E}\big[\, |X| \,\big]\; \mathbf{E}\big[\, |Y| \,\big] \;< +\infty$$

A questo punto, essendo verificata la convergenza assoluta delle serie, si possono ripetere i passaggi sopra scritti senza i valori assoluti e si ottiene il risultato cercato. □

Una conseguenza evidente è il risultato seguente:

Corollario 2.5.15. *Due variabili indipendenti dotate di momento secondo sono* incorrelate

Naturalmente non è vero il viceversa (provare a costruire un esempio).

Proposizione 2.5.16 (**Formula della convoluzione discreta).** *Siano* X, Y *due v.a. indipendenti a valori interi (relativi) e sia* $Z = X + Y$ *: vale la formula*

$$p_Z(n) \;=\; \mathbf{P}\big\{\, Z = n \,\big\} \;=\; \sum_{h=-\infty}^{+\infty} p_X(h) p_Y(n-h)$$

Dimostrazione. La dimostrazione è una conseguenza della relazione

$$\{ X+Y=n \} = \bigcup_{h=-\infty}^{+\infty} \{ X=h\,,\, Y=n-h \}$$

e del fatto che gli insiemi scritti a destra sono a due a due disgiunti. Si noti che se X, Y sono a valori interi positivi, la formula diventa (per n positivo)

$$p_Z(n) = \sum_{h=0}^{n} p_X(h)\, p_Y(n-h)$$

□

Esercizio 2.5.17. Provare che, se $X \sim B(n,p)$, $Y \sim B(m,p)$ e sono indipendenti, allora $(X+Y) \sim B(n+m,p)$ (si noti che ci si può ridurre, per induzione, al caso in cui una delle due variabili sia di Bernoulli). Dedurne, per una variabile Binomiale X, le formule di $\mathbf{E}[X]$ e $Var(X)$.

2.6 La funzione generatrice delle Probabilità.

Premettiamo alcuni richiami sulle *serie di potenze*: data una successioni di numeri $(a_n)_{n\geq 0}$, si chiama serie di potenze ad essa associata la serie $\sum_{n=0}^{+\infty} a_n t^n$. Il *raggio di convergenza* R verifica l'equazione

$$R = \frac{1}{\limsup_{n\to\infty} \sqrt[n]{|a_n|}}$$

(con la convenzione $\frac{1}{0} = +\infty$ e $\frac{1}{+\infty} = 0$). La serie di potenze converge per $|t| < R$ e non converge per $|t| > R$; inoltre se $R > 0$, posto $\varphi(t) = \sum_{n=0}^{+\infty} a_n t^n$, si ha $a_n = \frac{1}{n!}\varphi^{(n)}(0)$ e di conseguenza due serie di potenze coincidono se e solo se tutti i coefficienti $(a_n)_{n\geq 0}$ sono eguali.

In questo paragrafo consideriamo solamente variabili aleatorie X, $Y, \ldots$ *a valori interi positivi.*

Definizione 2.6.1. Data una variabile aleatoria X a valori interi positivi, si chiama *funzione generatrice delle probabilità* la funzione $G_X(.)$ definita da

$$G_X(t) = \sum_{n=0}^{+\infty} t^n\, p(n) = \mathbf{E}\big[t^X\big]$$

Si noti che la funzione generatrice è sicuramente definita per $|t| \leq 1$ (infatti il raggio di convergenza è sicuramente maggiore o eguale a 1, e si verifica direttamente che la serie converge per $|t| = 1$).

Proposizione 2.6.2. *Valgono le seguenti proprietà:*

1. $G_X(t) = G_Y(t) \iff X$ *e* Y *sono equidistribuite;*
2. X *e* Y *indipendenti* $\Longrightarrow G_{X+Y}(t) = G_X(t).G_Y(t)$.

Dimostrazione. La prima proprietà è immediata. Per quanto riguarda la seconda, si noti che anche le variabili t^X e t^Y sono indipendenti; si ha pertanto (ricordando il Teorema 2.5.1)

$$G_{X+Y}(t) = \mathbf{E}\big[t^{(X+Y)}\big] = \mathbf{E}\big[t^X t^Y\big] = \mathbf{E}\big[t^X\big]\,\mathbf{E}\big[t^Y\big] = G_X(t).G_Y(t)$$

□

Il risultato seguente esprime una relazione tra i momenti di una v.a. e le derivate della sua funzione generatrice:

Proposizione 2.6.3. *Sia* X *una v.a. a valori interi positivi: valgono le seguenti eguaglianze*

1. $\mathbf{E}\big[X\big] = \lim_{t\to 1-} G'_X(t)$
2. $\mathbf{E}\big[\,X(X-1)\,\big] = \lim_{t\to 1-} G"_X(t)$

Dimostrazione. Ricordiamo che ha senso scrivere $\mathbf{E}[X] \in [0, +\infty]$; sia poi $0 < t < 1$.

Vale l'eguaglianza $G'_X(t) = \sum_{n\geq 1} p(n) n\, t^{n-1}$. Facendo convergere t a 1 da sinistra, questa serie converge (per convergenza monotona: può essere vista come conseguenza del Teorema di Beppo Levi) a $\sum_{n\geq 1} p(n)\, n = \mathbf{E}[X]$. La dimostrazione della seconda eguaglianza si fa sostanzialmente allo stesso modo, osservando preventivamente che la v.a. $X(X-1)$ è ancora a valori positivi. □

Riportiamo qua sotto una *tabella* delle funzioni generatrici delle più usuali variabili aleatorie a valori interi, che il lettore può facilmente verificare:

- $X \sim B(n,p) \qquad \Longrightarrow \qquad G_X(t) = \big[1 + p(t-1)\big]^n$;
- X Geometrica di parametro $p \qquad \Longrightarrow \qquad G_X(t) = \frac{tp}{1-t(1-p)}$;
- X di Poisson di parametro $\lambda \qquad \Longrightarrow \qquad G_X(t) = \mathrm{e}^{\lambda(t-1)}$.

Esercizio 2.6.4. Calcolare valore atteso e varianza delle variabili sopra scritte con un calcolo diretto e utilizzando il risultato della Proposizione 2.6.3.

Esercizio 2.6.5. Provare che la somma di due variabili di Poisson indipendenti è ancora una variabile di Poisson (specificando la relazione esistente tra i parametri).

2.7 Legge dei Grandi Numeri per variabili Binomiali.

In questa sezione ci occupiamo di un primo *teorema limite* che riguarda una successione di variabili di Bernoulli di parametro p $(0 < p < 1)$: indichiamo con $X_1, X_n, \ldots$ una successione di variabili indipendenti con tale distribuzione, e poniamo $S_n = X_1 + \ldots + X_n$, che sappiamo avere distribuzione Binomiale $B(n, p)$.

Teorema 2.7.1 (**Legge dei grandi numeri per variabili Binomiali).** *Con le notazioni sopra indicate, per ogni $\varepsilon > 0$, vale il seguente limite*

$$\lim_{n\to\infty} \mathbf{P}\Big\{\Big|\frac{S_n}{n} - p\Big| > \varepsilon\Big\} = 0$$

Dimostrazione. Un semplice calcolo prova che $\mathbf{E}\Big[\frac{S_n}{n}\Big] = p$, $Var\Big(\frac{S_n}{n}\Big) = \frac{p(1-p)}{n}$ e di conseguenza per la diseguaglianza di Chebishev 2.4.10

$$\mathbf{P}\Big\{\Big|\frac{S_n}{n} - p\Big| > \varepsilon\Big\} \leq \frac{Var\Big(\frac{S_n}{n}\Big)}{\varepsilon^2} = \frac{p(1-p)}{n\,\varepsilon^2}$$

□

Osservazione 2.7.2. La dimostrazione sopra riportata è molto semplice, e si estende quasi senza modifiche a situazioni più generali: ad esempio si può supporre che le variabili $X_1, X_2, \ldots$ siano *indipendenti, equidistribuite*, dotate di momento secondo e con varianza σ^2 strettamente positiva: se si pone $\mathbf{E}\big[X_i\big] = m$, la stessa dimostrazione prova che

$$\lim_{n\to\infty} \mathbf{P}\Big\{\Big|\frac{S_n}{n} - m\Big| > \varepsilon\Big\} = 0$$

Il risultato 2.7.1 è all'origine di diversi teoremi che vanno sotto il nome di *leggi dei grandi numeri*, e che saranno affrontati in corsi più avanzati.

Una famiglia (non necessariamente una successione) $\big(X_i\big)_{i\in I}$ di variabili aleatorie *indipendenti ed equidistribuite* verrà d'ora innanzi indicata con l'abbreviazione (largamente usata) *i.i.d.* (Independent Identically Distributed).

Diamo una definizione più precisa per il tipo di convergenza enunciato nel teorema 2.7.1.

Definizione 2.7.3 (**Convergenza in Probabilità).** Data una successione di v.a. $\big(X_n\big)_{n\geq 1}$ ed una v.a. X, si dice che la successione *converge in probabilità* verso X se, per ogni $\varepsilon > 0$

$$\lim_{n\to\infty} \mathbf{P}\Big\{\big|X_n - X\big| \geq \varepsilon\Big\} = 0$$

Questo tipo di convergenza verrà ripreso in un capitolo successivo, ma un esame più dettagliato sarà oggetto di un corso di Probabilità più avanzato.

Come abbiamo visto dalla dimostrazione, la *velocità di convergenza* a zero della *probabilità di deviazione* (cioè della probabilità $\mathbf{P}\big\{\big|\frac{S_n}{n}-p\big|>\varepsilon\big\}$) nella legge dei grandi numeri (come è enunciata nell' Osservazione 2.7.2) è dell'ordine di $\frac{1}{n}$; tuttavia nel caso delle Variabili Binomiali si può provare che tale velocità di convergenza è *esponenziale*.

Teorema 2.7.4. *Nelle ipotesi del Teorema 2.7.1, dato $\varepsilon>0$, esiste una costante positiva $H(p,\varepsilon)$ tale che si abbia*

$$\mathbf{P}\Big\{\Big|\frac{S_n}{n}-p\Big|>\varepsilon\Big\}\le 2\,\exp\big(-n\,H(p,\varepsilon)\big)$$

Dimostrazione. Poniamo $L(s)=\mathbf{E}\big[\exp(s\,X_1)\big]=1-p+p\,\mathrm{e}^s$, e di conseguenza $\mathbf{E}\big[\exp(s\,S_n)\big]=L(s)^n$; scegliamo poi a con $p<a<1$.

Valgono le seguenti diseguaglianze:

$$\mathbf{P}\Big\{\frac{S_n}{n}>a\Big\}=\mathbf{P}\Big\{\exp\Big(s\big(\frac{S_n}{n}-a\big)\Big)>1\Big\}\le$$
$$\mathbf{E}\Big[\exp\Big(s\big(\frac{S_n}{n}-a\big)\Big)\Big]=L\big(\frac{s}{n}\big)^n\,\mathrm{e}^{-as}$$

qualunque sia s positivo. Prendendo $t=\frac{s}{n}$, e nell'ultimo termine della precedente disequazione l'estremo inferiore sui valori possibili si ha

$$\mathbf{P}\Big\{\frac{S_n}{n}>a\Big\}\le\exp\Big[-n\Big(\sup_{t>0}\big(at-\log L(t)\big)\Big)\Big]$$

La funzione $t\longrightarrow at-\log\big(1-p+p\mathrm{e}^t\big)$ è concava, diverge a $-\infty$ per $t\to+\infty$, ed ha derivata in 0 strettamente positiva: ha pertanto un valore massimo finito e strettamente positivo per $0<t<+\infty$. Preso $\varepsilon>0$ con $p+\varepsilon<1$, e denotando $h(p,\varepsilon)$ il massimo della funzione sopra indicata dove si è posto $a=p+\varepsilon$, si ottiene

$$\mathbf{P}\Big\{\frac{S_n}{n}>p+\varepsilon\Big\}\le\exp\Big(-n\,h(p,\varepsilon)\Big)$$

Con passaggi analoghi, si ottiene

$$\mathbf{P}\Big\{\frac{S_n}{n}<p-\varepsilon\Big\}\le\exp\Big(-n\,h(p,-\varepsilon)\Big)$$

Ponendo $H(p,\varepsilon)=\min\big(h(p,\varepsilon),h(p,-\varepsilon)\big)$, poichè $\mathbf{P}\big\{\big|\frac{S_n}{n}-p\big|>\varepsilon\big\}=\mathbf{P}\big\{\frac{S_n}{n}-p>\varepsilon\big\}+\mathbf{P}\Big\{\frac{S_n}{n}-p<-\varepsilon\Big\}$, si ottiene finalmente il risultato voluto. □

2.8 Appendice: alcuni esercizi significativi.

Esercizio 2.8.1 (**Sul gioco del lotto).** Quando viene puntata una somma sul realizzarsi di un evento di probabilità p, se il gioco è equo nel caso che questo evento si realizzi la somma dovrebbe essere restituita moltiplicata per p^{-1}; in particolare nel gioco del lotto se si punta su un *numero secco* la probabilità che questo venga estratto è

$$\frac{\binom{89}{4}}{\binom{90}{5}} = \frac{1}{18}$$

e quindi il *moltiplicatore teorico* dovrebbe essere 18: invece il moltiplicatore effettivamente praticato è 11,2.

Ancora più vistose sono le discrepanze se si considerano ambi, terne, quaterne ecc.. Qui bisogna distinguere tra ambi ottenuti puntando due numeri oppure un insieme di numeri maggiore di due (per ogni estrazione, è possibile puntare fino a 10 numeri): limitiamoci per semplicità al caso di ambo ottenuto puntando due numeri, terna ottenuta puntando tre numeri, ecc..

Per l'*ambo* il moltiplicatore teorico equo è 400,5 e quello effettivamente praticato 250, per la *terna* il valore teorico 11.748 e quello praticato 4.500; per la *quaterna* rispettivamente 511.038 e 120.000 e infine per la *cinquina* 43.949.268 e 6.000.000.

Tra l'altro sulla somma eventualmente vinta viene praticato un prelievo fiscale forfettario del 6 %: ne segue che ogni persona che abbia un minimo di conoscenza di calcolo delle probabilità non dovrebbe assolutamente giocare al lotto.

Tuttavia alcune persone ritengono di poter aggirare la situazione evidentemente sfavorevole con *sistemi di puntate* che consentano di vincere a colpo sicuro, ma vediamo che cosa succede: supponiamo per semplicità che il moltiplicatore effettuato puntando su un numero secco sia 11 e consideriamo il caso di una persona che voglia assolutamente vincere 100 Euro al lotto, puntando su un numero (poniamo il 53 sulla ruota di Venezia).

La prima volta punterà 10 Euro: se vince ne incassa 110 di cui 10 risarciscono la somma puntata e 100 sono la vittoria netta (nel caso in cui il numero venga estratto). Se il numero non viene estratto, ne punta 11 all'estrazione successiva: dei 121 Euro incassati nel caso di estrazione favorevole, 21 risarciscono i soldi spesi nelle prime due puntate e 100 costituiscono il guadagno effettivo, e così di seguito.

Il ragionamento alla base di questo sistema è evidente: prima o poi il numero 53 uscirà ed a quel momento si avrà la vittoria netta di 100 Euro. Tuttavia il giocatore ha pur sempre un capitale limitato e potrebbe *andare in bancarotta* prima di aver ottenuto la vincita che desiderava.

1) Determinare quale deve essere, al passo n-mo, il valore $s(n)$ della puntata da effettuare per poter avere una vittoria netta di 100 Euro (recuperando le somme spese nelle puntate precedenti).

2) Supponiamo che il giocatore abbia un capitale iniziale di 200 Euro: qual è la probabilità che il giocatore debba fermarsi per insufficienza di fondi senza aver ottenuto la sua vittoria?

3) Supponiamo che il giocatore non abbia limitazioni di fondi, e indichiamo con X la variabile aleatoria che indica quale somma in totale il giocatore ha dovuto impiegare fino al momento nel quale riesce a vincere: qual è il valore atteso $\mathbf{E}[X]$?

Esercizio 2.8.2 (**I polinomi di Bernstein).** Consideriamo, per $0 \leq x \leq 1$ una v.a. X_n^x *binomiale* di parametri n ed x; sia poi f una funzione continua definita sull'intervallo $[0,1]$ e definiamo

$$B_n(x) \,=\, \mathbf{E}\Big[f\Big(\frac{X_n^x}{n}\Big)\Big]$$

Provare che, per ogni n, $B_n(x)$ è un polinomio di grado n, (chiamato *polinomio di Bernstein*) e che la successione $(B_n)_{n\geq 1}$ converge uniformemente alla funzione f.

Questo procedimento *probabilistico* fornisce (limitatamente al caso degli intervalli di $\mathbb{R}$) una dimostrazione alternativa di un importante teorema di Weierstrass.

Esercizio 2.8.3 (**Il paradosso di Borel).** Ogni evento, per quanto la sua probabilità sia piccola, prima o poi si realizza (verificare questa affermazione utilizzando la variabile Geometrica) e quindi, come si usa dire con linguaggio colorito, *la scimmia che batte a caso sui tasti di una macchina da scrivere prima o poi scrive la Divina Commedia*: questa affermazione va sotto il nome di *paradosso di Borel*, anche se in realtà non è affatto paradossale. Tuttavia il tempo necessario per ottenere questo può essere talmente lungo da rendere di fatto impossibile l'evento.

Esaminiamo una versione semplificata: una scimmia di nome *Lucilla* batte a caso 7 caratteri sui tasti di una macchina da scrivere che ha solo 26 tasti (corrispondenti alle lettere), al ritmo di un carattere al secondo. Qual è il valore atteso del tempo necessario per riuscire a scrivere il suo nome? (In realtà bisognerebbe esaminare una situazione un poco più generale, cioè che dopo aver battuto a caso un certo numero di caratteri -non necessariamente multiplo di 7- vengano scritte nell'ordine giusto le lettere *lucilla*; questa situazione è un poco più complicata da esaminare e ci accontentiamo della versione semplificata).

Una curiosità divertente: per riuscire a scrivere, battendo a caso sui tasti, il solo primo versetto della Divina Commedia, il valore atteso del tempo necessario è di *miliardi di volte superiore* all'età dell'Universo!

Capitolo 3

Probabilità e variabili aleatorie su uno spazio generale

3.1 Costruzione di una Probabilità

Cominciamo con una definizione:

Definizione 3.1.1. Sia $\mathcal{A}$ una famiglia di parti di un insieme E: si chiama *σ-algebra generata da $\mathcal{A}$* la più piccola σ-algebra contenente $\mathcal{A}$: essa coincide con l'intersezione di tutte le σ-algebre contenenti $\mathcal{A}$.

Notiamo che tale insieme non è vuoto, perché esiste almeno $\mathcal{P}(E)$ (cioè la famiglia di tutti i sottinsiemi di E) che contiene $\mathcal{A}$. È bene inoltre ribadire che non esiste un metodo *costruttivo* per caratterizzare la σ-algebra generata da $\mathcal{A}$.

Proposizione 3.1.2 (I boreliani). *Sulla retta reale $\mathbb{R}$ coincidono le σ-algebre generate, ad esempio, da queste famiglie di insiemi:*

1. *le semirette del tipo $]-\infty, x]$, al variare di $x \in \mathbb{R}$;*
2. *gli intervalli semiaperti $]a, b]$ (oppure $[a, b[$), con $-\infty < a < b < +\infty$;*
3. *gli* aperti *di $\mathbb{R}$;*
4. *i* chiusi *di $\mathbb{R}$.*

La σ-algebra da essi generata è chiamata σ-algebra di Borel su $\mathbb{R}$ (e indicata $\mathcal{B}(\mathbb{R})$) ed i relativi elementi sono detti *boreliani*.

Dimostrazione. Chiamiamo ad esempio $\mathcal{B}_1$ la σ-algebra generata dalle semirette e $\mathcal{B}_2$ quella generata dagli intervalli. Poiché $]a,b] =]-\infty,b]\backslash]-\infty,a]$ è un elemento di $\mathcal{B}_1$, ne segue che $\mathcal{B}_2 \subseteq \mathcal{B}_1$.

Viceversa, poiché $]-\infty,x] = \cup_{n\geq 1}]x-n,x]$, segue che le semirette sono elementi di $\mathcal{B}_2$ e di conseguenza $\mathcal{B}_1 \subseteq \mathcal{B}_2$: si ha quindi l'eguaglianza $\mathcal{B}_1 = \mathcal{B}_2$.

Le altre eguaglianze si dimostrano in maniera del tutto simile e comunque molto semplice. □

Sulla retta, se non sarà diversamente specificato, si considera la σ-algebra di Borel. Analoga è la definizione della σ-algebra $\mathcal{B}(\mathbb{R}^n)$ dei boreliani di $\mathbb{R}^n$ che è generata, ad esempio, dalle seguenti famiglie di insiemi:

1. gli aperti di $\mathbb{R}^n$;
2. i *prodotti cartesiani* $A_1 \times \ldots \times A_n$, dove ogni A_i è un boreliano di $\mathbb{R}$;
3. i *prodotti cartesiani* della forma $]-\infty,x_1] \times \ldots \times]-\infty,x_n]$.

Diamo per scontato che il lettore sia a conoscenza della teoria della *misura* e dell'*integrazione* secondo Lebesgue, e chiamiamo $\mathcal{L}$ la famiglia delle parti di $\mathbb{R}$ misurabili secondo Lebesgue: $\mathcal{L}$ è una σ-algebra e contiene gli intervalli, e di conseguenza si ha l'inclusione $\mathcal{B}(\mathbb{R}) \subseteq \mathcal{L}(\mathbb{R})$.

In realtà l'inclusione è stretta ma la dimostrazione di questo fatto non è affatto immediata. Questo può essere visto in diversi modi e forse il più naturale è passare attraverso la cardinalità: si prova infatti che la famiglia dei Boreliani ha la stessa cardinalità di $\mathbb{R}$ (risultato tutt'altro che elementare), mentre si può costruire un insieme C trascurabile secondo Lebesgue che ha la stessa cardinalità di $\mathbb{R}$ (l'esempio più noto è l'insieme di Cantor). Ogni sottinsieme di C è trascurabile e pertanto misurabile secondo Lebesgue e di conseguenza la famiglia $\mathcal{L}$ ha cardinalità *strettamente superiore* a quella dei boreliani.

Saranno fondamentali per quanto segue i due seguenti risultati:

Teorema 3.1.3 (**Unicità di Probabilità**). *Siano* $\mathbf{P}$ *e* $\mathbf{Q}$ *due probabilità definite su una* σ*-algebra* $\mathcal{F}$ *di parti di un insieme* E *e supponiamo che* $\mathbf{P}$ *e* $\mathbf{Q}$ *coincidano su una famiglia* $\mathcal{I}$ *di parti tale che:*
1) $\mathcal{I}$ *genera* $\mathcal{F}$ *;*
2) $\mathcal{I}$ *è stabile per l'intersezione (finita).*
Allora $\mathbf{P}$ *e* $\mathbf{Q}$ *coincidono su tutto* $\mathcal{F}$*.*

Teorema 3.1.4 (**Esistenza di Probabilità**). *Sia* $\mathcal{A}$ *un'algebra di parti di un insieme* E *e sia* $\mathbf{P} : \mathcal{A} \to [0,1]$ *una funzione* σ*-additiva (tale che* $\mathbf{P}(E) = 1$*):* $\mathbf{P}$ *si prolunga (in un sol modo) alla* σ*-algebra* $\mathcal{F}$ *generata da* $\mathcal{A}$*.*

È bene precisare che cosa significa affermare che una funzione $\mathbf{P}$ è σ–additiva su un'algebra $\mathcal{A}$ di parti: vuol dire che se $(A_n)_{n=1,2,\dots}$ è una successione di elementi di $\mathcal{A}$ a due a due disgiunti e *se anche* $\bigcup_{n=1}^{+\infty} A_n$ *è un elemento di* $\mathcal{A}$, allora $\mathbf{P}\big(\bigcup_{n=1}^{+\infty} A_n\big) = \sum_{n=1}^{+\infty} \mathbf{P}(A_n)$

La dimostrazione dei due teoremi precedenti è lasciata a un corso più avanzato, ma è opportuno qualche commento. Il primo risultato non è vero per misure in generale (se la misura di tutto lo spazio è infinita): provare ad esempio a costruire un controesempio di due misure su $\mathcal{B}(\mathbb{R})$ che coincidono su ogni semiretta $]-\infty, x]$ ma non coincidono. Il secondo risultato, viceversa, è vero per misure qualsiasi (e osserviamo che, nel caso delle probabilità, l'unicità del prolungamento è conseguenza del Teorema 3.1.3).

Applichiamo ora i due teoremi appena enunciati alla **costruzione delle probabilità su** $\mathbb{R}$.

Definizione 3.1.5 (**Funzione di ripartizione).** Sia $\mathbf{P}$ una probabilità definita su $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$: si chiama *funzione di ripartizione* la funzione $F : \mathbb{R} \to [0,1]$ definita da $F(x) = \mathbf{P}\big(]-\infty, x]\big)$.

Proposizione 3.1.6. *La funzione di ripartizione sopra definita gode delle seguenti proprietà:*

1. *è crescente;*
2. *è continua a destra;*
3. $F(+\infty) = \lim_{x\to+\infty} F(x) = 1 \quad e \quad F(-\infty) = \lim_{x\to-\infty} F(x) = 0\,.$

Dimostrazione. È evidente che F è crescente (in genere non strettamente crescente).

Delle proprietà successive proviamo ad esempio la continuità a destra: dato x, poiché F è monotona, è sufficiente considerare una successione $(x_n)_{n\geq 1}$ convergente ad x da destra (ad esempio $x_n = x + \frac{1}{n}$). A questo punto, usando le stesse notazioni del Capitolo 1,

$$]-\infty, x_n] \downarrow]-\infty, x] \implies F(x_n) = \mathbf{P}\big(]-\infty, x_n]\big) \downarrow \mathbf{P}\big(]-\infty, x]\big) = F(x)$$

Le altre proprietà si dimostrano in maniera praticamente identica.

□

Con facili passaggi si prova che $F(b) - F(a) = \mathbf{P}\big(]a, b]\big)$, che $F_-(x) = \lim_{y<x\,,\,y\to x} F(y) = \mathbf{P}\big(]-\infty, x[\big)$ e che $\Delta F(x) = F(x) - F_-(x) = \mathbf{P}\big(\{x\}\big)$.

Ma quello che è veramente importante è il risultato seguente, che è in un certo senso l'inverso della Proposizione 3.1.6.

Teorema 3.1.7 (Esistenza di una Probabilità su $\mathcal{B}(\mathbb{R})$). *Assegnata una funzione $F : \mathbb{R} \to [0,1]$ con le proprietà 1), 2) e 3) della Proposizione 3.1.6, esiste una ed una sola probabilità $\mathbf{P}$ su $\mathcal{B}(\mathbb{R})$ tale che, per ogni $x \in \mathbb{R}$, si abbia $F(x) = \mathbf{P}\big(]-\infty, x]\big)$.*

Dimostrazione. L'unicità di questa probabilità $\mathbf{P}$ è conseguenza del Teorema 3.1.3 (la famiglia delle semirette è stabile per intersezione e genera $\mathcal{B}(\mathbb{R})$): proviamo ora l'esistenza.

Chiamiamo $\mathcal{A}$ la famiglia dei *pluriintervalli*: più precisamente un elemento A di $\mathcal{A}$ è della forma

$$A =]x_1, y_1] \cup \ldots \cup]x_k, y_k] \quad \text{con} \quad -\infty \le x_1 < y_1 < \ldots < x_k < y_k \le +\infty$$

e, per A di quella forma, definiamo

$$\mathbf{P}(A) = \sum_{i=1}^{k} \big[F(y_i) - F(x_i)\big]$$

È piuttosto noioso ma elementare provare che $\mathcal{A}$ è un'algebra di parti di $\mathbb{R}$ (che genera $\mathcal{B}(\mathbb{R})$) e che $\mathbf{P}$ è una funzione semplicemente additiva definita su $\mathcal{A}$: notiamo tra l'altro che un elemento $A \in \mathcal{A}$ si può scrivere in modi diversi come unione finita e disgiunta di intervalli ma il numero $\mathbf{P}(A)$ che ne risulta non dipende dalla particolare rappresentazione scelta.

Il prolungamento di $\mathbf{P}$ a tutto $\mathcal{B}(\mathbb{R})$ è una conseguenza del Teorema 3.1.4 a patto di provare che $\mathbf{P}$ è σ-additiva su $\mathcal{A}$. È più comodo a questo scopo provare la proprietà seguente:

$$\text{se} \quad A_n \in \mathcal{A} \quad , \quad A_n \downarrow \emptyset \qquad \Longrightarrow \quad \mathbf{P}(A_n) \downarrow 0$$

Partiamo dal fatto seguente: dato $A \in \mathcal{A}$ ed $\varepsilon > 0$, esiste $B \in \mathcal{A}$ con $\overline{B}$ compatto e $\overline{B} \subset A$ ($\overline{B}$ è la *chiusura* di B) tale che $\mathbf{P}\big(A \setminus B\big) < \varepsilon$. L'esistenza di un tale B è più facile da capire che da scrivere formalmente: comunque per ognuno dei k intervalli $]x_i, y_i]$ che compongono A, si considera un intervallo a chiusura compatta $]z_i, w_i]$ tale che $\mathbf{P}\big(]x_i, y_i] \setminus]z_i, w_i]\big) < \frac{\varepsilon}{k}$ e poi si prende l'unione di questi intervalli.

Se x_i, y_i sono entrambi finiti, si prenderà $]x_i + \delta, y_i]$ con un opportuno δ sufficientemente piccolo, se il primo estremo è $-\infty$ (e l'altro finito), si prenderà $]-M, y_i]$ con M reale sufficientemente grande e così via ... le proprietà della funzione F permettono questa costruzione.

Consideriamo allora la successione $A_n \downarrow \emptyset$, $\varepsilon > 0$ e, per ogni n, un elemento $B_n \in \mathcal{A}$ con le proprietà sopra indicate e contenuto in A_n e tale che $\mathbf{P}\big(A_n \setminus B_n\big) < \frac{\varepsilon}{2^n}$.

Si ha $\bigcap_{n\geq 1} \overline{B}_n = \emptyset$ e, siccome questi insiemi sono compatti, ne esiste una sottofamiglia finita con intersezione vuota: scegliamo dunque $\bar{n}$ tale che $B_1 \cap \ldots \cap B_{\bar{n}} = \emptyset$. Si ha

$$A_{\bar{n}} = A_{\bar{n}} \cap \big(B_1 \cap \ldots \cap B_{\bar{n}}\big)^c = \bigcup_{j=1,\ldots,\bar{n}} \big(A_{\bar{n}} \cap B_j^c\big) \subseteq \bigcup_{j=1,\ldots,\bar{n}} \big(A_j \setminus B_j\big)$$

Ne segue che si ha $\mathbf{P}(A_{\bar{n}}) < \varepsilon$ e, poichè questo è vero per ogni ε, si ha $\lim_{n\to\infty} \mathbf{P}(A_n) = 0$. □

Vediamo i tipi più usuali di probabilità su $\mathbb{R}$ e le corrispondenti proprietà delle relative funzioni di ripartizione.

Esempio 3.1.8 (**Probabilità discrete).** Abbiamo già incontrato le probabilità *discrete* (dette anche *atomiche*) su $\mathcal{B}(\mathbb{R})$: $\mathbf{P}$ è concentrata sulla successione di punti $(x_1, x_2, \ldots)$ e, per ogni $A \in \mathcal{B}(\mathbb{R})$, vale l'eguaglianza $\mathbf{P}(A) = \sum_{x_i \in A} p(x_i)$ essendo $p(x_i) = \mathbf{P}\big(\{x_i\}\big)$.

In particolare la funzione di ripartizione soddisfa l'eguaglianza $F(x) = \sum_{x_i \leq x} p(x_i)$: disegnando in particolare la funzione di ripartizione delle leggi Binomiale, o di Poisson, o altre, si nota che ha un tipico *andamento a gradini*. Ma non tutte le funzioni di ripartizione delle probabilità discrete sono fatte così come mostra l'esempio che ora segue.

Esercizio 3.1.9. Sia $\mathbb{Q}$ l'insieme dei razionali e consideriamo una numerazione di $\mathbb{Q} = \{q_1, q_2, \ldots\}$; sia poi $\mathbf{P}$ concentrata su $\mathbb{Q}$ tale che $p(q_n) = \mathbf{P}\big(\{q_n\}\big) = 2^{-n}$ ed F la relativa funzione di ripartizione. Provare che F è *strettamente crescente*.

Esempio 3.1.10 (**Misura secondo Lebesgue).** La misura secondo Lebesgue non è limitata e quindi non può essere costruita come conseguenza del Teorema 3.1.4. Tuttavia si può costruire la misura di Lebesgue $\boldsymbol{\lambda}$ sui sottinsiemi boreliani di $[0, 1]$ considerando la funzione di ripartizione così definita:

$$F(x) = \begin{cases} 0 & \text{per } x < 0 \\ x & \text{per } 0 \leq x \leq 1 \\ 1 & \text{per } x > 1 \end{cases}$$

In modo analogo la si può costruire su ogni intervallo di $\mathbb{R}$ di lunghezza 1; si pone poi, per $A \in \mathcal{B}(\mathbb{R})$, $\boldsymbol{\lambda}(A) = \sum_{n=-\infty}^{+\infty} \boldsymbol{\lambda}\big(A \cap]n, n+1]\big)$.

Esempio 3.1.11 (**Probabilità diffusa).** Abbiamo visto che ogni punto è trascurabile per la probabilità $\mathbf{P}$ associata alla funzione di ripartizione F se e solo se F è continua: questo è una conseguenza della formula $\mathbf{P}\big(\{x\}\big) =$

$\Delta F(x)$. Le probabilità che godono di questa proprietà sono dette *diffuse*. Provare che in tal caso la funzione di ripartizione F è anche *uniformemente continua*.

In verità le probabilità diffuse non hanno particolari proprietà: sono molto più importanti e maneggevoli le *probabilità definite da una densità*, che verranno però introdotte nel successivo paragrafo.

3.2 Costruzione dell'integrale

Definizione 3.2.1 (**Spazio e applicazione misurabile).** Si chiama *spazio misurabile* una coppia $\big(E, \mathcal{E}\big)$ dove E è un insieme e $\mathcal{E}$ una σ-algebra di parti di E. Dati due spazi misurabili $\big(E, \mathcal{E}\big)$ e $\big(F, \mathcal{F}\big)$, una applicazione $f : E \to F$ è detta *misurabile* se, per ogni $A \in \mathcal{F}$, $f^{-1}(A) \in \mathcal{E}$.

Proposizione 3.2.2. *Con le notazioni della definizione precedente, se $\mathcal{A}$ è una famiglia di parti di F che genera la σ-algebra $\mathcal{F}$, affinché una funzione $f : E \to F$ sia misurabile, è sufficiente che, per ogni $A \in \mathcal{A}$, $f^{-1}(A) \in \mathcal{E}$.*

Dimostrazione. La dimostrazione è molto semplice: se noi chiamiamo $\mathcal{B}$ la famiglia dei sottisiemi $B \subseteq F$ tali che $f^{-1}(B) \in \mathcal{E}$, è una facile verifica provare che $\mathcal{B}$ è una σ-algebra. Poichè $\mathcal{B}$ contiene $\mathcal{A}$, contiene anche la σ-algebra generata cioè $\mathcal{F}$. □

Se non è specificato diversamente, dato uno spazio misurabile $\big(E, \mathcal{E}\big)$, una funzione $f : E \to \mathbb{R}$ è detta *misurabile* se è misurabile considerando su $\mathbb{R}$ la σ-algebra $\mathcal{B}(\mathbb{R})$.

Grazie al risultato 3.2.2, affinché f sia misurabile è sufficiente ad esempio che, $\forall x \in \mathbb{R}$, $\big\{f \leq x\big\} = f^{-1}\big(]-\infty, x]\big)$ (o, equivalentemente, $\forall a < b$, $\big\{a < f \leq b\big\} = f^{-1}\big(]a, b]\big)$) sia un elemento di $\mathcal{E}$).

Una funzione misurabile da $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ su $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$ è detta *boreliana*.

Definizione 3.2.3 (**Funzione semplice).** Dato uno spazio misurabile $\big(E, \mathcal{E}\big)$, si chiama *semplice* una funzione misurabile $\varphi : E \to \mathbb{R}$ *che prende un numero finito di valori* (cioè la cui immagine è un insieme finito).

Chiamati $a_1, \ldots, a_n$ i punti dell'immagine della funzione semplice φ e detti $A_i = \{\varphi = a_i\}$, è evidente che la funzione può essere scritta nella forma

$$\varphi \;=\; \sum_{i=1}^{n} a_i \, I_{A_i}$$

cioè φ è una combinazione lineare di indicatrici di insiemi misurabili, viceversa ogni combinazione lineare di indicatrici di insiemi misurabili (non necessariamente disgiunti) è evidentemente una funzione semplice. L'espressione di una funzione semplice in tale forma non è unica, tuttavia date due funzioni semplici φ e ψ è facile vedere che esistono $A_1, \dots, A_n$ disgiunti in modo tale che si possa scrivere

$$\varphi \,=\, \sum_{i=1}^{n} a_i \, I_{A_i} \quad ; \quad \psi \,=\, \sum_{i=1}^{n} b_i \, I_{A_i} \,,$$

cioè φ e ψ si possono scrivere come combinazione lineare delle funzioni indicatrici *degli stessi insiemi misurabili*.

Una conseguenza immediata di questa osservazione è che l'insieme delle funzioni semplici è uno *spazio vettoriale* ed un *reticolo* (l'ultima dizione significa che, se $\phi\,,\,\psi$ sono funzioni semplici, anche $\varphi \vee \psi \,=\, \max(\varphi, \psi)$ e $\varphi \wedge \psi = \min(\varphi, \psi)$ sono funzioni semplici).

Sopponiamo ora assegnato uno spazio misurabile $\big(E, \mathcal{E}\big)$ sul quale è definita una misura di *probabilità* $\mathbf{m}$.

Definizione 3.2.4 (**Integrale delle funzioni semplici).** Sia φ una funzione semplice della forma $\varphi \,=\, \sum_{i=1}^{n} a_i \, I_{A_i}$: definiamo *integrale* di φ il numero

$$\int_E \varphi(x) \, \mathrm{d}\mathbf{m}(x) \,=\, \sum_{i=1}^{n} a_i \, \mathbf{m}(A_i)$$

Se non c'è ambiguità, si può scrivere più semplicemente $\int \varphi \, \mathrm{d}\mathbf{m}$: è una verifica noiosa ma non difficile provare che questo numero non dipende dalla particolare rappresentazione di φ che si è scelta, mentre è facile provare che si ha

- $\int \big(a\varphi + \psi\big) \, \mathrm{d}\mathbf{m} \,=\, a \int \varphi \, \mathrm{d}\mathbf{m} \,+\, \int \psi \, \mathrm{d}\mathbf{m}$;
- se $\varphi \leq \psi$, allora $\int \varphi \, \mathrm{d}\mathbf{m} \leq \int \psi \, \mathrm{d}\mathbf{m}$.

Proposizione 3.2.5 (**Proprietà di Beppo Levi per funzioni semplici).** *Sia* $(\varphi_n)_{n\geq 1}$ *una successione di funzioni semplici e supponiamo che* $\varphi_n \uparrow \varphi$ *e che* φ sia ancora una funzione semplice*: allora*

$$\int \varphi_n \, \mathrm{d}\mathbf{m} \uparrow \int \varphi \, \mathrm{d}\mathbf{m}$$

Anche la dimostrazione di questo risultato è lasciata a un corso più avanzato, tuttavia è interessante osservare che se $\varphi_n = I_{A_n}$ dove $\big(A_n\big)_{n\geq 1}$ è una

successione *crescente* di insiemi, si ha che $I_{A_n} \uparrow I_A$ essendo $A = \cup_{n\geq 1} A_n$: la proprietà di Beppo Levi equivale alla *continuità* della probabilità, più precisamente

$$\int I_{A_n} \, \mathrm{d}\mathbf{m} \;=\; \mathbf{m}\big(A_n\big) \;\uparrow\; \mathbf{m}\big(A\big) \;=\; \int I_A \, \mathrm{d}\mathbf{m}$$

Allo scopo di estendere la definizione di integrale, sarà fondamentale il risultato seguente:

Teorema 3.2.6 (**Approssimazione con funzioni semplici**). *Sia f una funzione misurabile a valori positivi: esiste una successione di funzioni semplici $(\varphi_n)_{n\geq 1}$ tale che*

$$\varphi_n \uparrow f$$

Dimostrazione. Una possibile successione approssimante può essere definita in questo modo:

$$\varphi_n \;=\; n\, I_{\{f\geq n\}} + \sum_{h=0}^{n2^n-1} \frac{h}{2^n}\, I_{\left\{\frac{h}{2^n}\leq f< \frac{h+1}{2^n}\right\}}$$

È piuttosto noioso (ma per niente difficile) verificare che, qualunque sia x, $\varphi_n(x) \leq \varphi_{n+1}(x)$ e che $\lim_{n\to\infty} \varphi_n(x) = f(x)$.

□

La funzione f può anche prendere il valore $+\infty$ in qualche punto x; i boreliani su $\overline{\mathbb{R}} = [-\infty, +\infty]$ e le funzioni misurabili a valori in $\overline{\mathbb{R}}$ si definiscono in maniera identica a quanto si è fatto per la retta reale $\mathbb{R}$.

Osservazione 3.2.7 (**Sulla definizione di funzione misurabile**). Solitamente in analisi si chiama misurabile una funzione $f : \mathbb{R} \to \mathbb{R}$ tale che, per ogni $A \in \mathcal{B}(\mathbb{R})$, $f^{-1}(A) \in \mathcal{L}$ (sia cioè misurabile secondo Lebesgue): si considerano quindi due differenti σ-algebre su $\mathbb{R}$ come spazio di partenza e come spazio di arrivo. La ragione di questa apparente incongruenza va ricercata proprio nel Teorema 3.2.6: vedremo subito che quel risultato di approssimazione è fondamentale nella definizione di integrale, e per poter fare quella costruzione è necessario che gli insiemi $\{a \leq f < b\}$ siano misurabili (e questo equivale a dire che l'immagine inversa di ogni Boreliano è misurabile). Viceversa si ha interesse a disporre, sull'insieme su cui è definita la funzione, della famiglia di insiemi misurabili più grande possibile (la σ-algebra $\mathcal{L}$, quando si considera la misura di Lebesgue).

Una conseguenza di questa definizione è, ad esempio, che composizione di due funzioni misurabili non è necessariamente misurabile, però se $f : \mathbb{R} \to \mathbb{R}$ è misurabile e $g : \mathbb{R} \to \mathbb{R}$ è boreliana, allora $g \circ f$ è misurabile.

Inoltre, data una successione $(f_n)_{n\geq 1}$ di funzioni misurabili a valori reali, la funzione $\left(\sup_n f_n\right)$ è misurabile: si ha infatti $\left\{\sup_n f_n \leq a\right\} = \bigcap_n \left\{f_n \leq a\right\}$. In modo analogo sono misurabili $\left(\inf_n f_n\right)$, $\left(\limsup_n f_n\right)$, $\left(\liminf_n f_n\right)$ e, se esiste, $\left(\lim_n f_n\right)$.

Infine, come conseguenza del Teorema 3.2.6, ogni funzione misurabile a valori reali si può scrivere come limite puntuale di una successione di funzioni semplici: da qui segue facilmente che, se f e g sono misurabili, anche $(f + g)$, $(f \vee g)$ e $(f \wedge g)$ sono misurabili.

Definizione 3.2.8 (**Integrale delle funzioni a valori positivi).** Sia f una funzione misurabile a valori positivi e consideriamo una successione di funzioni semplici $(\varphi_n)_{n\geq 1}$ tale che $\varphi_n \uparrow f$: si definisce *integrale* di f il numero

$$\int f \,\mathrm{d}\mathbf{m} = \lim_{n\geq 1} \int \varphi_n \,\mathrm{d}\mathbf{m}$$

Il limite esiste poiché la successione di numeri $\left(\int \varphi_n \,\mathrm{d}\mathbf{m}\right)_{n\geq 1}$ è *crescente* (il limite eventualmente può essere $+\infty$); apparentemente però la definizione 3.2.8 è ambigua perché si possono prendere in considerazione diverse successioni approssimanti. In realtà questa ambiguità non sussiste come dimostra il risultato seguente:

Teorema 3.2.9 (**Proprietà di Beppo Levi).** *Se $(\varphi_n)_{n\geq 1}$ e $(\psi_n)_{n\geq 1}$ sono due successioni di funzioni semplici convergenti alla funzione f si ha*

$$\lim_{n\to\infty} \int \varphi_n \,\mathrm{d}\mathbf{m} = \lim_{n\to\infty} \int \psi_n \,\mathrm{d}\mathbf{m}$$

Inoltre se $(f_n)_{n\geq 1}$ è una successione di funzioni misurabili a valori positivi, si ha

$$f_n \uparrow f \quad \Longrightarrow \quad \int f_n \,\mathrm{d}\mathbf{m} \uparrow \int f \,\mathrm{d}\mathbf{m}$$

Dimostrazione. Fissiamo n e consideriamo la successione di funzioni semplici $(\varphi_n \wedge \psi_m)_{m\geq 1}$: questa è crescente e converge a φ_n. Per la Proposizione 3.2.5 si ha

$$\int \varphi_n \,\mathrm{d}\mathbf{m} = \lim_{m\to\infty} \int \left(\varphi_n \wedge \psi_m\right)\mathrm{d}\mathbf{m} \leq \lim_{m\to\infty} \int \psi_m \,\mathrm{d}\mathbf{m}$$

e, di conseguenza, $\lim_{n\to\infty} \int \varphi_n \,\mathrm{d}\mathbf{m} \leq \lim_{m\to\infty} \int \psi_m \,\mathrm{d}\mathbf{m}$. Scambiando le due successioni si ottiene la diseguaglianza opposta e quindi l'eguaglianza: questo dimostra la prima affermazione.

Per quanto riguarda la seconda, consideriamo per ogni n una successione di funzioni semplici $(\varphi_{n,m})_{m\geq 1}$ convergente crescendo ad f_n, e poniamo $\psi_n = \max_{i,j\leq n}(\varphi_{i,j})$.

È immediato constatare che $(\psi_n)_{n\geq 1}$ è una successione crescente di funzioni semplici, che per ogni n si ha $\psi_n \leq f_n$ e che $\psi_n \uparrow f$: si ha pertanto

$$\int f \,\mathrm{d}\mathbf{m} \,=\, \lim_{n\to\infty} \int \psi_n \,\mathrm{d}\mathbf{m} \leq \lim_{n\to\infty} \int f_n \,\mathrm{d}\mathbf{m}\,.$$

Ma, poiché per ogni n si ha $\int f_n \,\mathrm{d}\mathbf{m} \,\leq\, \int f \,\mathrm{d}\mathbf{m}$, si ottiene l'eguaglianza cercata. □

Si verifica facilmente che, se f, g sono misurabili positive ed $a > 0$, si ha $\int (af + g)\mathrm{d}\mathbf{m} = a \int f \mathrm{d}\mathbf{m} + \int g \mathrm{d}\mathbf{m}$; inoltre se $f \leq g$, allora $\int f \mathrm{d}\mathbf{m} \leq \int g \mathrm{d}\mathbf{m}$.

Consideriamo ora una generica funzione misurabile f , e poniamo $f^+ = f \vee 0 = \max(f, 0)$ e $f^- = -(f \wedge 0) = -\min(f, 0)$: entrambe sono funzioni misurabili (è una verifica immediata) e si ha $|f| = f^+ + f^-$ e $f = f^+ - f^-$.

Definizione 3.2.10 (**Funzione integrabile e integrale).** Si dice che la funzione misurabile f è *integrabile* se $\int |f| \mathrm{d}\mathbf{m} < +\infty$, e in tal caso si chiama *integrale* di f il numero

$$\int f \,\mathrm{d}\mathbf{m} \,=\, \int f^+ \mathrm{d}\mathbf{m} \,-\, \int f^- \mathrm{d}\mathbf{m}\,.$$

Lo spazio delle funzioni integrabili viene indicato $\mathcal{L}^1\big(E, \mathcal{E}, \mathbf{m}\big)$ (o più semplicemente $\mathcal{L}^1$ se non c'è ambiguità): se $f, g \in \mathcal{L}^1$ ed a è un numero qualsiasi, si ha $\int (af + g)\mathrm{d}\mathbf{m} = a \int f \mathrm{d}\mathbf{m} + \int g \mathrm{d}\mathbf{m}$. Mentre l'eguaglianza $\int af \,\mathrm{d}\mathbf{m} = a. \int f \,\mathrm{d}\mathbf{m}$ è immediata, l'eguaglianza $\int (f + g) \,\mathrm{d}\mathbf{m} = \int f \,\mathrm{d}\mathbf{m} + \int g \,\mathrm{d}\mathbf{m}$ è conseguenza di questo fatto che lasciamo provare come esercizio: se $f = g - h$ dove g, h sono misurabili, a valori positivi e integrabili, si ha $\int f \,\mathrm{d}\mathbf{m} = \int g \,\mathrm{d}\mathbf{m} - \int h \,\mathrm{d}\mathbf{m}$.

Teorema 3.2.11 (**Convergenza dominata).** *Sia $(f_n)_{n\geq 1}$ una successione di funzioni misurabili convergente puntualmente ad f e supponiamo che esista g integrabile a valori positivi tale che si abbia, per ogni $x \in E$, $|f_n(x)| \leq g(x)$: allora si ha*

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mathbf{m} \,=\, \int f \,\mathrm{d}\mathbf{m}\,.$$

Anche di questo risultato omettiamo la dimostrazione; ci limitiamo ad osservare che la condizione $|f_n(x)| \leq g(x)$ (valida ovviamente anche per il limite f) porta come conseguenza che ogni f_n (e così pure il limite f) è integrabile.

Osservazione 3.2.12. La costruzione esposta in questo paragrafo è valida (praticamente senza modifiche) per l'integrale rispetto ad una generica misura $\mathbf{m}$ non di probabilità (tale che si abbia $\mathbf{m}(E) = +\infty$). L'unica modifica sostanziale è nella definizione di *funzione semplice* : bisogna considerare delle funzioni φ della forma $\varphi = \sum_{i=1}^{n} a_i \, I_{A_i}$ con A_i tali che $\mathbf{m}(A_i) < +\infty$.

L'integrale della funzione f rispetto alla misura di Lebesgue (se esiste) è usualmente denotato $\int f(x)\,\mathrm{d}x$.

Sostanzialmente senza modifiche rispetto al Capitolo 2 si prova la *diseguaglianza di Schwartz*: se f^2 e g^2 sono integrabili, il prodotto $fg \in \mathcal{L}^1$ e si ha

$$\Big| \int fg \,\mathrm{d}\mathbf{m} \Big| \leq \sqrt{\int f^2 \,\mathrm{d}\mathbf{m}} \, \sqrt{\int g^2 \,\mathrm{d}\mathbf{m}} \,.$$

Osservazione 3.2.13 (Integrale rispetto ad una misura discreta). Quando l'insieme E è numerabile (o più in generale la misura è concentrata su un insieme numerabile), l'integrale come è stato definito in questo capitolo coincide con la definizione data nel Capitolo 2: basta verificare questo per le funzioni a valori positivi.

Data una tale funzione f, definiamo

$$\varphi_n(x_j) = \begin{cases} f(x_j) & \text{se } j \leq n \\ 0 & \text{se } j > n \end{cases}$$

La successione $(\varphi_n)_{n\geq 1}$ è una successione crescente di funzioni semplici convergente ad f: poiché per ogni n si ha $\int \varphi_n \,\mathrm{d}\mathbf{m} = \sum_{j\leq n} f(x_j)\mathbf{m}(x_j)$, al limite si ha la somma della serie, cioè la definizione data a suo tempo.

Possiamo ora introdurre una nuova categoria di probabilità su $\mathbb{R}$, molto importante nelle applicazioni.

Definizione 3.2.14 (Densità di probabilità). Si chiama *densità di probabilità* su $\mathbb{R}$ una funzione reale f definita su $\mathbb{R}$, misurabile e a valori positivi, integrabile (secondo Lebesgue) e tale che $\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x = 1$.

Ad una densità f è associata una probabilità $\mathbf{P}$ su $\mathcal{B}(\mathbb{R})$ mediante la formula

$$\mathbf{P}(A) = \int_A f(x)\,\mathrm{d}x$$

È immediato constatare che la funzione così definita è semplicemente additiva e che $\mathbf{P}(\mathbb{R}) = 1$; per provare che è anche σ-additiva viene più comodo mostrare la proprietà di continuità sulle successioni crescenti d'insiemi usando la proprietà di Beppo Levi.

Se $A_n \uparrow A$, si ha che $f.I_{A_n} \uparrow f.I_A$ e quindi

$$\mathbf{P}(A_n) \,=\, \int f.I_{A_n}\,\mathrm{d}x \uparrow \int f.I_A\,\mathrm{d}x \,=\, \mathbf{P}(A)\,.$$

Vale il seguente risultato

Teorema 3.2.15 (**Integrazione rispetto a una misura definita da una densità).** *Una funzione misurabile g definita su $\mathbb{R}$ è integrabile rispetto a $\mathbf{P}$ se e solo se il prodotto gf è integrabile rispetto alla misura di Lebesgue, e in tal caso si ha*

$$\int g(x)\,\mathrm{d}\mathbf{P}(x) \,=\, \int g(x)f(x)\,\mathrm{d}x\,.$$

Dimostrazione. Cominciamo a supporre che g sia l'indicatrice di un insieme misurabile A:

$$\int I_A\,\mathrm{d}\mathbf{P} \,=\, \mathbf{P}(A) \,=\, \int_A f\,\mathrm{d}x \,=\, \int f\,I_A\,\mathrm{d}x$$

Di conseguenza l'eguaglianza è vera per le funzioni semplici; data una generica g misurabile e positiva, e considerando una successione crescente approssimante $(\varphi_n)_{n\geq 1}$, applicando in entrambi gli integrali la proprietà di Beppo Levi, si ha

$$\int g\,\mathrm{d}\mathbf{P} = \lim_{n\to\infty}\int \varphi_n\,\mathrm{d}\mathbf{P} = \lim_{n\to\infty}\int \varphi_n\,f\,\mathrm{d}x = \int g\,f\,\mathrm{d}x$$

Considerata poi una funzione misurabile generica g, si considera la decomposizione $g = g^+ - g^-$ e si conclude facilmente. □

Analoga è la definizione di *probabilità definita da una densità* su $\big(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)\big)$, ed il relativo teorema di integrazione.

Esaminiamo ora la funzione di ripartizione di una probabilità definita da una densità, cioè $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$: naturalmente F è continua, ma non è vero il viceversa. Ci sono esempi di funzioni di ripartizione continue la cui probabilità associata non è definita da una densità: l'esempio più noto è quello della *misura di Cantor*, che sarà esposta in Appendice.

Vale il seguente risultato, che viene qui solo enunciato:

Proposizione 3.2.16 (**Funzioni assolutamente continue).** *La probabilità associata ad una funzione di ripartizione F è definita da una densità se e solo se F è* assolutamente continua*, cioè per ogni $\varepsilon > 0$, esiste $\delta > 0$ tale che, prese delle coppie di punti (x_i, y_i),*

$$\sum_{i\leq n} |x_i - y_i| < \delta \quad \Longrightarrow \sum_{i\leq n} \big|F(x_i) - F(y_i)\big| < \varepsilon$$

La Proposizione precedente fornisce una precisa caratterizzazione che però è poco *pratica*: di fatto si utilizza spesso questo criterio *sufficiente* (che lasciamo provare come esercizio). *Supponiamo che la funzione di ripartizione F sia continua e C^1 a tratti, cioè che sia derivabile con derivata continua eccetto che in un insieme finito di punti $a_1, \dots, a_n$: allora la probabilità associata ad F è definita da una densità e una versione della densità f è data (eccetto che nei punti $a_1, \dots, a_n$) dall'eguaglianza $f(x) = \frac{\mathrm{d}F(x)}{\mathrm{d}x}$.*

Notiamo che nei punti $a_1, \dots, a_n$ possiamo definire la densità in un modo qualsiasi, poiché si tratta di un insieme *trascurabile* (rispetto alla misura di Lebesgue) e la densità interviene solo attraverso integrali.

3.3 Variabili aleatorie reali e vettoriali su uno spazio di probabilità generale

Ora che disponiamo della teoria dell'integrazione rispetto ad una probabilità su uno spazio Ω generale, possiamo estendere senza difficoltà le definizioni date nel Capitolo 2 e riguardanti le variabili aleatorie (reali e vettoriali): c'è però una differenza sostanziale. Nel Capitolo 2 non avevamo menzionato problemi di misurabilità (perché in un insieme numerabile ogni sottinsieme è misurabile) mentre ora dobbiamo essere molto precisi proprio riguardo a questioni di misurabilità.

Definizione 3.3.1 (**Variabile aleatoria reale).** Assegnato uno spazio di Probabilità $\big(\Omega, \mathcal{F}, \mathbf{P}\big)$, si chiama *variabile aleatoria reale* una applicazione misurabile $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Quindi X deve essere tale che, ad esempio, per ogni $x \in \mathbb{R}$, $\big\{X \le x\big\} = X^{-1}\big(]-\infty, x]\big) \in \mathcal{F}$.

Allora, data una funzione boreliana $f : \mathbb{R} \to \mathbb{R}$, $f \circ X$ è ancora una variabile aleatoria (ma questo non è più vero con una generica funzione f).

Definizione 3.3.2 (**Legge di Probabilità).** Si chiama *legge di probabilità* (o anche *distribuzione di probabilità*) di una variabile aleatoria reale X l'immagine di $\mathbf{P}$ mediante X; si chiama *funzione di ripartizione* di X la funzione di ripartizione della sua legge di probabilità.

Si ha dunque, per ogni A *boreliano*, $\mathbf{P}_X(A) = \mathbf{P}\big(X^{-1}(A)\big)$.
Chiamata poi F_X la sua funzione di ripartizione, si ha

$$F_X(x) = \mathbf{P}_X\big(]-\infty, x]\big) = \mathbf{P}\big\{X \le x\big\}\,.$$

Osservazione 3.3.3. Assegnata comunque una probabilità $\mathbf{Q}$ su $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$, esiste una variabile aleatoria X la cui legge di probabilità sia eguale a $\mathbf{Q}$. La costruzione è simile a quella che è stata fatta per le leggi di probabilità discrete, ed è anche molto semplice (ma importante dal punto di vista metodologico): si può prendere $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ e $\mathbf{P} = \mathbf{Q}$. Si considera poi come applicazione $X : \mathbb{R} \to \mathbb{R}$ l'identità, cioè $X(x) = x$: è immediato constatare che $\mathbf{P}_X = \mathbf{Q}$. Una analoga costruzione (che non ripeteremo) si può fare per le variabili *vettoriali*.

Vediamo ora l'estensione al caso generale del Teorema 2.4.1.

Teorema 3.3.4 (Integrazione rispetto ad una probabilità immagine). *Sia $\varphi : \mathbb{R} \to \mathbb{R}$ boreliana: φ è integrabile rispetto a $\mathbf{P}_X$ se e solo se $\varphi \circ X$ è integrabile rispetto a $\mathbf{P}$ e in tal caso vale la formula*

$$\int_{\mathbb{R}} \varphi(x)\, \mathrm{d}\mathbf{P}_X(x) \;=\; \int_{\Omega} \varphi\big(X(\omega)\big)\, \mathrm{d}\mathbf{P}(\omega)\,.$$

Dimostrazione. La dimostrazione è simile a quella del teorema 3.2.15, ed è abbastanza semplice. Cominciamo a verificare la formula nel caso in cui $\varphi = I_A$, con A boreliano.

$$\int_{\mathbb{R}} I_A(x)\, \mathrm{d}\mathbf{P}_X(x) = \mathbf{P}_X(A) = \mathbf{P}\big(X^{-1}(A)\big) =$$

$$\int_{\Omega} I_{X^{-1}(A)}(\omega)\, \mathrm{d}\mathbf{P}(\omega) = \int_{\Omega} \big(I_A \circ X\big)(\omega)\, \mathrm{d}\mathbf{P}(\omega)$$

Di conseguenza la formula è vera per le combinazioni lineari di indicatrici di boreliani, cioè per le funzioni semplici. Data φ misurabile positiva, si prende una successione approssimante crescente $(\varphi_n)_{n\geq 1}$ di funzioni semplici: applicando Beppo Levi in entrambi gli integrali si ottiene

$$\int_{\mathbb{R}} \varphi(x)\, \mathrm{d}\mathbf{P}_X(x) = \lim_{n\to\infty} \int_{\mathbb{R}} \varphi_n(x)\, \mathrm{d}\mathbf{P}_X(x) =$$

$$\lim_{n\to\infty} \int_{\Omega} \varphi_n\big(X(\omega)\big)\, \mathrm{d}\mathbf{P}(\Omega) = \int_{\Omega} \varphi\big(X(\omega)\big)\, \mathrm{d}\mathbf{P}(\Omega)$$

Per passare poi al caso di φ di segno qualsiasi, si considera la decomposizione $\varphi = \varphi^+ - \varphi^-$ e si applica separatamente la formula a φ^+ e φ^-. □

Perfettamente analoghe a quanto si è visto per il caso delle variabili aleatorie discrete, sono le definizioni di valori attesi, momenti, varianza, ecc...

Ad esempio, il *valore atteso* di X (se esiste) è l'integrale

$$\mathbf{E}\big[X\big] \;=\; \int_\Omega X(\omega)\,\mathrm{d}\mathbf{P}(\omega) \;=\; \int_{\mathbb{R}} x\,\mathrm{d}\mathbf{P}_X(x)$$

La dimostrazione del fatto che, se $1 \le p < q < +\infty$ ed X ammette momento di ordine q, allora ammette anche momento di ordine p, è sostanzialmente identica a quanto fatto per le variabili discrete: provare per esercizio a *tradurre* questa dimostrazione. Allo stesso modo è identica la dimostrazione della *diseguaglianza di Chebishev*.

Passiamo ora al caso delle variabili aleatorie *vettoriali* $\mathbf{X} = \big(X_1, \ldots, X_n\big)$ limitando per semplicità di notazioni l'esposizione al caso delle variabili aleatorie *doppie* (X, Y) (l'estensione al caso n-dimensionale è del tutto immediata).

Per definizione, si chiama *variabile aleatoria doppia* una applicazione misurabile $(X, Y) : \big(\Omega, \mathcal{F}\big) \to \big(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)\big)$. Le componenti X e Y sono due funzioni definite su Ω a valori reali.

Proposizione 3.3.5. *La coppia (X, Y) è una variabile aleatoria (cioè è misurabile come applicazione a valori in $\mathbb{R}^2$) se e solo se entrambe le componenti X e Y sono variabili aleatorie reali (cioè misurabili come applicazioni a valori in $\mathbb{R}$).*

Dimostrazione. Ricordiamo che $\mathcal{B}(\mathbb{R}^2)$ è generata, ad esempio, dai prodotti cartesiani $]-\infty, x]\times]-\infty, y]$: pertanto, se X e Y sono misurabili,

$$\big(X, Y\big)^{-1}\big(]-\infty, x]\times]-\infty, y]\big) \;=\; X^{-1}\big(]-\infty, x]\big) \cap Y^{-1}\big(]-\infty, y]\big)$$

è un elemento di $\mathcal{F}$. Viceversa, supponendo che la coppia (X, Y) sia misurabile,

$$X^{-1}\big(]-\infty, x]\big) \;=\; (X, Y)^{-1}\big(]-\infty, x]\times]-\infty, +\infty[\big)$$

è un elemento di $\mathcal{F}$. □

La *legge di probabilità* della coppia (X, Y) è l'immagine di $\mathbf{P}$ mediante l'applicazione (X, Y): è quindi una probabilità su $\mathcal{B}(\mathbb{R}^2)$. Il Teorema 3.3.4 si estende senza difficoltà al caso vettoriale, in particolare presa $\varphi : \mathbb{R}^2 \to \mathbb{R}$ boreliana e limitata, vale la formula

$$\int_\Omega \varphi\big(X(\omega), Y(\omega)\big)\,\mathrm{d}\mathbf{P}(\omega) \;=\; \iint_{\mathbb{R}^2} \varphi(x, y)\,\mathrm{d}\mathbf{P}_{X,Y}(x, y)$$

Nella formula precedente, si è considerata una funzione *boreliana e limitata* perché in questo caso sicuramente è integrabile (rispetto ad una misura

di probabilità); un altro caso in cui sicuramente l'integrale esiste è quando φ è boreliana e a valori positivi.

La definizione di *indipendenza* di due variabili aleatorie X, Y è identica a quella data a suo tempo per variabili discrete (vedi Definizione 2.5.6) ed in maniera identica si prova il risultato seguente (vedi Corollario 2.5.11): *se X e Y sono indipendenti e f, g sono due funzioni* boreliane, *allora anche $f \circ X$ e $g \circ Y$ sono indipendenti.*

Per poter estendere al caso generale i risultati della Proposizione 2.5.9 e del Teorema 2.5.14, dobbiamo però insistere un poco sulla nozione di *probabilità prodotto.*

Definizione 3.3.6 (Probabilità prodotto). Siano $\mathbf{P}$ e $\mathbf{Q}$ due probabilità su $\big(\mathbb{R}, \mathcal{B}(\mathbb{R})\big)$: si chiama *probabilità prodotto* (e si indica $\mathbf{P}\otimes\mathbf{Q}$) la probabilità su $\big(\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2)\big)$ tale che, presi comunque due sottinsiemi boreliani A e B di $\mathbb{R}$, si abbia

$$\mathbf{P} \otimes \mathbf{Q}\big(A \times B\big) \;=\; \mathbf{P}\big(A\big).\mathbf{Q}\big(B\big)$$

L'unicità di una tale probabilità è una facile conseguenza del Teorema 3.1.3 : infatti i *rettangoli misurabili* $A \times B$ (con A, B boreliani) sono una famiglia di parti stabile per intersezione che genera la σ-algebra prodotto $\mathcal{B}(\mathbb{R}^2)$. L'esistenza invece è una conseguenza del Teorema 3.1.4 , ed è più impegnativa da dimostrare: si considera l'algebra $\mathcal{A}$ di parti di $\mathbb{R}^2$ formata da unioni disgiunte di *rettangoli misurabili* sulla quale è definita la naturale estensione della 3.3.6 e si dimostra che è σ-additiva. Non insistiamo su questa costruzione, cito soltanto il fatto (che ci servirà tra poco) che vale una estensione del *Teorema di Fubini-Tonelli.*

Più precisamente, se $\varphi : \mathbb{R}^2 \to \mathbb{R}$ è boreliana e limitata (oppure a valori positivi) vale la formula di integrazione

$$\iint_{\mathbb{R}^2} \varphi(x,y)\, \mathrm{d}\mathbf{P}\otimes\mathbf{Q}\,(x,y) \;=\; \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \varphi(x,y)\, \mathrm{d}\mathbf{Q}(y)\right] \mathrm{d}\mathbf{P}(x)$$

Nella parte destra della formula sopra scritta si può scambiare l'ordine di integrazione, inoltre quando vengono scritte delle integrazioni successive (se non vi sono parentesi) vengono svolte da destra verso sinistra: scriveremo così più semplicemente

$$\iint_{\mathbb{R}^2} \varphi(x,y)\, \mathrm{d}\mathbf{P}\otimes\mathbf{Q}\,(x,y) \;=\; \int_{\mathbb{R}} \mathrm{d}\mathbf{P}(x) \int_{\mathbb{R}} \varphi(x,y)\, \mathrm{d}\mathbf{Q}(y)$$

È immediata l'estensione al caso generale della caratterizzazione provata nel caso delle variabili discrete con la Proposizione 2.5.9: più precisamente X e Y sono indipendenti se e solo se $\mathbf{P}_{X,Y} = \mathbf{P}_X \otimes \mathbf{P}_Y$.

Ed in modo analogo, si estende facilmente il Teorema 2.5.14:

Teorema 3.3.7. *Supponiamo che X ed Y siano indipendenti e dotate di momento primo: anche XY ha valore atteso e vale la formula*

$$\mathbf{E}\big[XY\big] \,=\, \mathbf{E}\big[X\big]\,\mathbf{E}\big[Y\big]$$

Dimostrazione. Cominciamo ha provare che $\mathbf{E}[|XY|] < +\infty$ utilizzando il Teorema di Fubini-Tonelli:

$$\mathbf{E}\big[\,|XY|\,\big] \,=\, \iint_{\mathbb{R}^2} |xy|\,\mathrm{d}\mathbf{P}_X \otimes \mathbf{P}_Y(x,y) =$$

$$= \int_{\mathbb{R}} |x|\,\mathrm{d}\mathbf{P}_X(x) \int_{\mathbb{R}} |y|\,\mathrm{d}\mathbf{P}_Y(y) = \mathbf{E}\big[|X|\big]\,\mathbf{E}\big[|Y|\big] < +\infty$$

Ripetendo gli stessi passaggi senza i valori assoluti, si ottiene la tesi. □

3.4 Variabili aleatorie con densità

Definizione 3.4.1. Si dice che la v.a. reale X ha *densità* f se la sua legge di probabilià $\mathbf{P}_X$ ha densità f, cioè se per ogni boreliano A vale la formula

$$\mathbf{P}\big\{X \in A\big\} \,=\, \mathbf{P}_X\big(A\big) \,=\, \int_A f(x)\,\mathrm{d}x$$

Di conseguenza la *funzione di ripartizione* è data da $F(x) = \int_{-\infty}^{x} f(t)\,\mathrm{d}t$ ed è pertanto continua, ma come sappiamo non è vero il viceversa. Per questo motivo è piuttosto fuorviante la denominazione di *variabili aleatorie continue* che alcuni testi danno: bisognerebbe piuttosto dire *variabili aleatorie assolutamente continue*.

Se si modifica la densità f su un insieme *trascurabile* (per la misura di Lebesgue) il valore degli integrali $\int_A f(x)\,\mathrm{d}x$ non viene alterato: per questo la densità di probabilità, più che una funzione, è una *classe di equivalenza di funzioni* (intendendo per equivalenti due funzioni che differiscono su un insieme trascurabile).

Proposizione 3.4.2. *Sia X una variabile aleatoria reale. Sono equivalenti le due seguenti affermazioni:*

1. *X ha densità f;*
2. *per ogni funzione reale φ boreliana e limitata, vale la formula*

$$\mathbf{E}\big[\varphi(X)\big] \,=\, \int_{\mathbb{R}} \varphi(x)\,f(x)\,\mathrm{d}x$$

Dimostrazione. La dimostrazione è del tutto immediata, ma come vedremo il criterio fornito da questa Proposizione è molto utile.

Da una parte, supponendo che X abbia densità f, utilizzando i Teoremi 3.3.4 e 3.2.15, si ha

$$\mathbf{E}\big[\varphi(X)\big] \;=\; \int_{\mathbb{R}} \varphi(x)\,\mathrm{d}\mathbf{P}_X(x) \;=\; \int_{\mathbb{R}} \varphi(x)\,f(x)\,\mathrm{d}x$$

Viceversa, prendendo A boreliano e considerando $\varphi = I_A$, si ha

$$\mathbf{P}\big\{X \in A\big\} \;=\; \mathbf{E}\big[I_A \circ X\big] = \int_{\mathbb{R}} I_A(x) f(x)\,\mathrm{d}x = \int_A f(x)\,\mathrm{d}x$$

□

In maniera del tutto analoga viene data la definizione di variabile aleatoria *vettoriale* $\mathbf{X} = (X_1, \ldots, X_n)$ con densità, e l'estensione n-dimensionale della Proposizione 3.4.2.

Il risultato che viene ora enunciato è l'analogo per variabili con densità della Proposizione 2.5.5.

Proposizione 3.4.3. *Sia (X, Y) una variabile doppia con densità $f(x, y)$: anche le componenti X ed Y ammettono densità f_1 ed f_2 che soddisfano le formule*

$$f_1(x) = \int_{-\infty}^{+\infty} f(x,y)\mathrm{d}y \qquad\qquad f_2(y) = \int_{-\infty}^{+\infty} f(x,y)\,\mathrm{d}x$$

Dimostrazione. Si utilizza il criterio fornito dalla Proposizione 3.4.2. Sia $\varphi : \mathbb{R} \to \mathbb{R}$ boreliana limitata:

$$\mathbf{E}\big[\varphi(X)\big] \;=\; \iint \varphi(x) f(x,y)\,\mathrm{d}x\,\mathrm{d}y \;=\; \int \varphi(x)\Big[\int f(x,y)\,\mathrm{d}y\Big]\,\mathrm{d}x$$

Questo equivale a dire che la funzione $x \longrightarrow \int f(x,y)\,\mathrm{d}y$ è la densità di X. □

Osservazione 3.4.4. Viceversa, conoscendo le densità marginali delle componenti X e Y, non si può ricostruire la densità congiunta, anzi *non è neppure detto che la coppia (X, Y) abbia densità*! Per fornire un controesempio, consideriamo una variabile X con densità e la coppia (X, X) ; provare che quest'ultima non può avere densità.

Il risultato seguente è l'analogo per variabili con densità della Proposizione 2.5.9.

Proposizione 3.4.5. *Sia (X, Y) una variabile doppia con densità: le variabili X e Y sono indipendenti se e solo se tra le densità vale la seguente relazione (quasi ovunque)*

$$f(x, y) \;=\; f_1(x)\, f_2(y)$$

Dimostrazione. È un facile esercizio provare che, se $\mathbf{P}_1$ e $\mathbf{P}_2$ hanno densità rispettivamente f_1 ed f_2, la probabilità prodotto $\mathbf{P}_1 \otimes \mathbf{P}_2$ ha come densità la funzione $f_1(x)f_2(y)$ (che è talvolta chiamata il *prodotto tensore* delle due funzioni f_1 ed f_2).

Di conseguenza vale quella relazione tra le densità se e solo se la legge di probabilità congiunta è il prodotto delle singole leggi. □

Vediamo ora l'analogo per variabili con densità della Proposizione 2.5.16.

Proposizione 3.4.6 (**Formula della convoluzione**). *Siano X, Y due variabili indipendenti con densità rispettivamente f_1 ed f_2 : la somma $(X+Y)$ ha densità g data dalla formula*

$$g(x) \;=\; \int_{-\infty}^{+\infty} f_1(x-y) f_2(y)\,\mathrm{d}y$$

Dimostrazione. Di nuovo si usa la Proposizione 3.4.2. Sia $\varphi : \mathbb{R} \to \mathbb{R}$ boreliana limitata

$$\mathbf{E}\big[\varphi(X+Y)\big] = \iint \varphi(x+y) f_1(x) f_2(y) \mathrm{d}x\,\mathrm{d}y = \int f_2(y)\mathrm{d}y \int \varphi(x+y) f_1(x)\mathrm{d}x =$$

$$= \int f_2(y)\mathrm{d}y \int \varphi(t) f_1(t-y)\mathrm{d}t = \int \varphi(t)\Big[\int f_1(t-y) f_2(y)\mathrm{d}y\Big]\mathrm{d}t$$

□

Le formule che ora seguono esprimono come si trasforma la densità di una variabile aleatoria (reale o vettoriale) se si applica ad essa un *diffeomorfismo*: ricordiamo che si chiama diffeomorfismo un'applicazione biunivoca tra due aperti A e B di $\mathbb{R}^k$, che sia *differenziabile* con *inversa differenziabile*.

Proposizione 3.4.7. *Sia X una v.a. reale con densità f diversa da 0 su un aperto $A \subseteq \mathbb{R}$ e sia $h : A \to B$ un diffeomorfismo. Consideriamo la variabile $Y = h(X)$: essa ha densità g data da*

$$g(y) \;=\; \begin{cases} 0 & \text{se } y \notin B \\ f\big(h^{-1}(y)\big)\big|\frac{\mathrm{d}\,h^{-1}(y)}{\mathrm{d}\,y}\big| = f\big(x(y)\big)\big|\frac{\mathrm{d}x(y)}{\mathrm{d}y}\big| & \text{se } y \in B \end{cases}$$

Dimostrazione. È essenzialmente una conseguenza della formula del cambio di variabili per gli integrali. Data φ boreliana limitata, si ha

$$\mathbf{E}\big[\varphi(Y)\big] = \mathbf{E}\big[\varphi\big(h(X)\big)\big] = \int_A \varphi(h(x))\, f(x)\, \mathrm{d}x =$$

$$= \int_B \varphi(y)\, f\big(h^{-1}(y)\big) \Big| \frac{\mathrm{d}\, h^{-1}(y)}{\mathrm{d}\, y} \Big| \, \mathrm{d}y$$

□

Esempio 3.4.8. La densità più semplice che si possa immaginare è la densità *uniforme* sull'intervallo $[0, 1]$ così definita

$$f(x) \;=\; \begin{cases} 1 & \text{per } 0 < x < 1 \\ 0 & \text{altrimenti} \end{cases}$$

Sia X con tale densità e sia $Y = \log(X)$: la densità di Y è data da

$$g(y) \;=\; \begin{cases} \mathrm{e}^y & \text{per } y < 0 \\ 0 & \text{per } y \geq 0 \end{cases}$$

La formula per la trasformazione della densità di una v.a. vettoriale $\mathbf{X}$ mediante un diffeomorfismo è anch'essa conseguenza della formula del cambio di variabili per integrali (questa volta n-dimensionali) ed è del tutto analoga alla formula 3.4.7: il termine $\big|\frac{\mathrm{d}\, h^{-1}(y)}{\mathrm{d}\, y}\big|$ è sostituito col *valore assoluto del determinante della matrice Jacobiana* della funzione h^{-1}.

Vediamo come si usa in concreto questa formula, limitandoci per semplicità al caso di una variabile doppia (X, Y) con densità f diversa da 0 sull'aperto A di $\mathbb{R}^2$: consideriamo un diffeomorfismo h da A su B e sia $(U, V) = h(X, Y)$. La coppia (U, V) ha una densità g che si annulla fuori di B, mentre su B soddisfa la formula

$$g(u, v) \;=\; f\big(x(u, v), y(u, v)\big) \,.\, \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

dove con $\begin{vmatrix} a & b \\ c & d \end{vmatrix}$ si intende il *valore assoluto del determinante* della matrice $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$.

Esempio 3.4.9. Sia (X, Y) avente densità

$$f(x,y) \;=\; \begin{cases} 2\mathrm{e}^{-(x+y)} & \text{per } 0 < x < y \\ 0 & \text{altrimenti} \end{cases}$$

e sia $(U, V) = (X+Y, X-Y)$: vogliamo calcolare la densità di (U, V).

Innanzi tutto è facile verificare che la funzione sopra scritta è effettivamente una densità, cioè che si ha

$$\iint_{\mathbb{R}^2} f(x,y)\,\mathrm{d}x\,\mathrm{d}y \;=\; \iint_{\{0<x<y\}} 2\,\mathrm{e}^{-(x+y)}\,\mathrm{d}x\,\mathrm{d}y \;=\; 1$$

Inoltre è immediato constatare che l'applicazione $h(x, y) = (x+y, x-y)$ è un diffeomorfismo dall'aperto $A = \big\{(x,y) \in \mathbb{R}^2 \,\big|\, 0 < x < y\big\}$ sull'aperto $B = \big\{(u,v) \in \mathbb{R}^2 \,\big|\, u > 0\,,\; -u < v < 0\big\}$: l'inversa di h si calcola immediatamente, si ha infatti $x = \frac{u+v}{2}$ e $y = \frac{u-v}{2}$. È immediato anche il calcolo del modulo del determinante $\left|\begin{matrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{matrix}\right| \;=\; \frac{1}{2}$.

La densità g della coppia (U, V) risulta pertanto essere

$$g(u,v) \;=\; \begin{cases} \mathrm{e}^{-u} & \text{per } u > 0\,,\; -u < v < 0 \\ 0 & \text{altrove} \end{cases}$$

È sempre prudente verificare che si ha effettivamente, come in questo caso,

$$\iint_{\mathbb{R}^2} g(u,v)\,\mathrm{d}u\,\mathrm{d}v \;=\; \iint_{B} \mathrm{e}^{-u}\,\mathrm{d}u\,\mathrm{d}v \;=\; 1$$

3.5 Esempi di variabili aleatorie con densità

3.5.1 Densità uniforme

Si chiama densità uniforme sull'intervallo $]a, b[$ una densità che è costante su quell'intervallo e nulla fuori: si avrà quindi

$$f(x) \;=\; \begin{cases} \frac{1}{b-a} & \text{per } a < x < b \\ 0 & \text{altrimenti} \end{cases}$$

È un facile esercizio provare che, se X è una v.a. con tale densità, si ha $\mathbf{E}[X] = \frac{a+b}{2}$ e $Var(X) = \frac{(b-a)^2}{12}$.

3.5.2 Densità Gamma

Premettiamo la definizione della funzione *Gamma*: questa è definita, per $r > 0$, da $\Gamma(r) = \int_0^{+\infty} x^{r-1}\mathrm{e}^{-x}\,\mathrm{d}x$. Questa non si può calcolare esplicitamente, ma è immediato verificare (tramite una integrazione per parti) che, se $r > 1$, si ha $\Gamma(r) = \big(r-1\big)\Gamma(r-1)$. Inoltre $\Gamma(1) = 1$ e di conseguenza, per n intero, $\Gamma(n) = (n-1)!$

Definizione 3.5.1. Si chiama densità Gamma di parametri r e λ , $(r > 0\,,\, \lambda > 0)$, (e si indica $\Gamma(r,\lambda)$) la funzione definita da

$$f(x) \,=\, \begin{cases} \frac{1}{\Gamma(r)}\ \lambda^r x^{r-1}\mathrm{e}^{-\lambda x} & x > 0 \\ 0 & x \le 0 \end{cases}$$

È un facile calcolo provare che si tratta effettivamente di una densità di probabilità; quando $r = 1$, la densità $\Gamma(1,\lambda)$ si chiama più semplicemente *esponenziale di parametro* λ.

Se $X \sim \Gamma(r,\lambda)$ e $\beta > 0$, è facile provare che vale la seguente formula

$$\mathbf{E}\big[X^\beta\big] \,=\, \frac{\Gamma(r+\beta)}{\Gamma(r)\,\lambda^\beta}$$

e da questa si calcolano facilmente i momenti della variabile X : ad esempio $\mathbf{E}[X] = \frac{r}{\lambda}$.

Proposizione 3.5.2. *Se* $X \sim \Gamma(r_1,\lambda)$, $Y \sim \Gamma(r_2,\lambda)$ *e sono indipendenti, allora* $(X+Y) \sim \Gamma(r_1+r_2,\lambda)$

Dimostrazione. Si utilizza la formula della convoluzione (Proposizione 3.4.6): per semplificare i conti, limitiamoci al caso in cui X e Y sono esponenziali di parametro λ. La densità di $(X+Y)$ si annulla per $x \le 0$, e per $x > 0$ è eguale a

$$g(x) \,=\, \int_0^x \lambda^2\,\mathrm{e}^{-\lambda(x-y)}\mathrm{e}^{-\lambda y}\,\mathrm{d}y \,=\lambda^2\,x\,\mathrm{e}^{-\lambda x}$$

che è appunto la densità $\Gamma(2,\lambda)$. □

La densità esponenziale esibisce una sorta di *assenza di memoria* che è in un certo senso l'analogo per variabili con densità della proprietà delle variabili geometriche.

Esercizio 3.5.3. Sia X una variabile con densità esponenziale e siano x, y positivi: provare che si ha

$$\mathbf{P}\big\{X > x+y \,\big|\, X > x\big\} \,=\, \mathbf{P}\big\{X > y\big\} \tag{3.5.1}$$

Viceversa, sia X una variabile a valori positivi con legge di probabilità diffusa, e supponiamo che, presi comunque x e y positivi, valga l'eguaglianza (3.5.1): provare che X ha densità esponenziale.

3.5.3 Densità Gaussiana

Cominciamo ad osservare che la primitiva della funzione $\mathrm{e}^{-\frac{x^2}{2}}$ non si può scrivere in termine di funzioni elementari, e quindi l'integrale su un intervallo non si può calcolare esattamente: si può però calcolare l'integrale su tutta la retta grazie a un trucco geniale. L'idea brillante che segue è solitamente attribuita a Gauss, in realtà è stata introdotta da Laplace proprio nella sua generalizzazione di un precedente risultato di De Moivre, mentre Gauss ha estensivamente utilizzato la funzione che segue nella teoria degli errori (vedremo qualche cenno nell'ultimo capitolo).

Notiamo che vale l'eguaglianza $\left(\int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x\right)^2 = \iint_{\mathbb{R}^2} \mathrm{e}^{-\frac{x^2+y^2}{2}}\,\mathrm{d}x\,\mathrm{d}y$; passando a coordinate polari, questo integrale doppio diventa $\int_0^{2\pi} \mathrm{d}\theta \int_0^{+\infty} \mathrm{e}^{-\frac{\rho^2}{2}}\,\rho\,\mathrm{d}\rho = 2\pi$.

Ne segue che la funzione $f(x) = \frac{1}{\sqrt{2\pi}}\,\mathrm{e}^{-\frac{x^2}{2}}$ è una densità di probabilità, detta densità **Normale** o **Gaussiana** $N(0,1)$, e la funzione $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$ è la relativa *funzione di ripartizione.*

Gli integrali della funzione $\mathrm{e}^{-\frac{x^2}{2}}$ su un intervallo qualsiasi non possono venire calcolati esplicitamente ma solo approssimati numericamente; per venire incontro a questa difficoltà sono state compilate delle *tavole statistiche* della funzione $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$ (per x positivo).

Per una variabile $X \sim N(0,1)$ si ha $\mathbf{E}[X] = 0$ (non c'è bisogno di fare calcoli, poichè la funzione $x\,\mathrm{e}^{-\frac{x^2}{2}}$ è una funzione *dispari*, e quindi il suo integrale su tutto $\mathbb{R}$ è 0). Viceversa $Var\big(X\big) = \mathbf{E}\big[X^2\big] = 1$, come si verifica facilmente integrando per parti: si ha infatti

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} x^2\,\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x \;=\; \frac{-1}{\sqrt{2\pi}} x\,\mathrm{e}^{-\frac{x^2}{2}}\Big|_{-\infty}^{+\infty} \;+\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x \;=\; 1$$

Definizione 3.5.4 (**Variabile Gaussiana).** Si dice che la variabile X ha legge gaussiana $N(m,\sigma^2)$ $\big(m \in \mathbb{R}\,,\ \sigma > 0\big)$ se $\frac{X-m}{\sigma}$ ha legge $N(0,1)$

Si può pertanto rappresentare X nella forma $X = \sigma Y + m$, con $Y \sim N(0,1)$: ne segue immediatamente che $\mathbf{E}[X] = m$, $Var\big(X\big) = \sigma^2$. Inoltre, come conseguenza della Proposizione 3.4.7, la densità di X è la funzione g definita da

$$g(x) \;=\; \frac{1}{\sqrt{2\pi}\,\sigma}\;\mathrm{e}^{-\frac{(x-m)^2}{2\sigma^2}}$$

Proposizione 3.5.5. *Se* $X \sim N\big(m_1,\sigma_1^2\big)$, $Y \sim N\big(m_2,\sigma_2^2\big)$ *e sono indipendenti, allora* $(X+Y) \sim N\big(m_1+m_2,\sigma_1^2+\sigma_2^2\big)$.

Dimostrazione. Ci si può ridurre al caso in cui $m_1 = m_2 = 0$, e, per semplicità di conti, limitiamoci al caso in cui $\sigma_1 = \sigma_2 = 1$. Applicando la formula della convoluzione, la densità g di $(X + Y)$ è data da

$$g(x) \,=\, \frac{1}{2\pi} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{1}{2}\left(y^2+(x-y)^2\right)} \mathrm{d}y \,=\, \frac{1}{2\pi} \mathrm{e}^{-\frac{x^2}{4}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{1}{2}\left(\sqrt{2}y-\frac{x}{\sqrt{2}}\right)^2} \mathrm{d}y$$

Facendo il cambio di variabile $\sqrt{2}y - \frac{x}{\sqrt{2}} = t$, l'integrale sopra scritto risulta eguale a

$$\frac{1}{2\pi} \frac{\mathrm{e}^{-\frac{x^2}{4}}}{\sqrt{2}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}t \,=\, \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{2}} \mathrm{e}^{-\frac{x^2}{4}}$$

cioè $(X + Y) \sim N(0, 2)$. □

Esercizio 3.5.6. Se $X \sim N(0, 1)$, allora $X^2 \sim \Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$.

3.6 Appendice

3.6.1 Alcune leggi di probabilità di rilevante interesse in Statistica

Prima di illustrare alcune leggi di probabilità di rilevante interesse nell'inferenza statistica, introduciamo la definizione di *quantile*: data una funzione di ripartizione F ed un numero $0 < \alpha < 1$, intuitivamente lo α-quantile è il numero r_α tale che $F(r_\alpha) = \alpha$ (quindi, per una variabile aleatoria X con funzione di ripartizione F, si ha $\mathbf{P}\{X \leq r_\alpha\} = \alpha$).

Riserveremo in particolare la notazione q_α all'α-quantile della legge $N(0, 1)$, cioè al numero q_α tale che $\Phi(q_\alpha) = \alpha$.

La definizione sopra enunciata non presenta difficoltà se l'applicazione F è *biunivoca* da un intervallo $I \subseteq \mathbb{R}$ su $]0, 1[$, ma in generale si possono presentare due difficoltà: può darsi che F abbia una *discontinuità* intorno al valore α, in modo che non esista alcun numero r_α con la proprietà richiesta; e può darsi che sia *costante* su un intervallo in modo che esista tutto un intervallo di numeri r tali che $F(r) = \alpha$. La definizione deve allora essere modificata in questo modo:

Definizione 3.6.1 (Quantile). Data una funzione di ripartizione F ed un numero $0 < \alpha < 1$, si chiama *α-quantile* di F il numero così definito

$$r_\alpha \,=\, \inf \left\{x \in \mathbb{R} \,\middle|\, F(x) > \alpha\right\}.$$

Le leggi di probabilità che vengono ora esposte, sono state introdotte per l'applicazione a problemi di inferenza statistica.

Definizione 3.6.2 (**Legge chi-quadro).** Si chiama *legge chi-quadro a* n *gradi di libertà* (e si indica $\chi^2(n)$) la legge $\Gamma\left(\frac{n}{2}, \frac{1}{2}\right)$.

Il motivo per cui è stato dato un nome particolare a questa legge Gamma è il seguente: se $(X_1, \ldots, X_n)$ sono indipendenti gaussiane $N(0,1)$, allora $X_1^2 + \cdots + X_n^2$ ha legge $\chi^2(n)$ (la prova di questo fatto è una conseguenza immediata dell'Esercizio 3.5.6 e della Proposizione 3.5.2).

Per agevolare i conti con questa particolare legge di probabilità, sono state predisposte le *tavole* della legge Chi-quadro: più precisamente, in funzione dei gradi di libertà n e del numero α, queste tavole assegnano il valore $\chi^2_{(\alpha,n)}$ dello α–quantile della legge $\chi^2(n)$ (cioè, per una variabile X con densità $\chi^2(n)$ si ha $\mathbf{P}\{X \leq \chi^2_{(\alpha,n)}\} = \alpha$).

Definizione 3.6.3 (**Legge di Student).** Siano $X \sim N(0,1)$, $Y \sim \chi^2(n)$ indipendenti: si chiama *legge di Student a* n *gradi di libertà* (e si indica $T(n)$) la legge di

$$\frac{\sqrt{n}\,X}{\sqrt{Y}}$$

Prima di calcolare effettivamente la densità, osserviamo che se T è una variabile di Student, ha legge *simmetrica* (cioè T e $-T$ sono equidistribuite): infatti una variabile con densità è simmetrica se e solo se la sua densità è una funzione *pari*. Di conseguenza, poiché $X \sim N(0,1)$ è simmetrica, $\frac{\sqrt{n}\,X}{\sqrt{Y}}$ e $\frac{-\sqrt{n}\,X}{\sqrt{Y}}$ sono equidistribuite.

Il calcolo della densità (in verità piuttosto tedioso) è una conseguenza della Proposizione 3.4.2 : siano f_1 la densità di X ed f_2 la densità di Y, e sia φ boreliana limitata. Applicando il teorema di Fubini-Tonelli ed il cambio di variabili, si ha

$$\begin{aligned}
\mathbf{E}\Big[\varphi\Big(\tfrac{\sqrt{n}\,X}{\sqrt{Y}}\Big)\Big] &= \iint_{\{-\infty<x<+\infty\,,\,y>0\}} \varphi\Big(\tfrac{\sqrt{n}\,x}{\sqrt{y}}\Big)\, f_1(x) f_2(y)\,\mathrm{d}x\,\mathrm{d}y \\
&= \int_0^{+\infty} f_2(y)\,\mathrm{d}y \int_{-\infty}^{+\infty} \varphi\Big(\tfrac{\sqrt{n}\,x}{\sqrt{y}}\Big) f_1(x)\,\mathrm{d}x \\
&= \int_0^{+\infty} f_2(y)\,\mathrm{d}y \int_{\infty}^{+\infty} \varphi(t) f_1\Big(\tfrac{t\sqrt{y}}{\sqrt{n}}\Big)\,\tfrac{\sqrt{y}}{\sqrt{n}}\,\mathrm{d}t \\
&= \int_{-\infty}^{+\infty} \varphi(t)\Big[\int_0^{+\infty} f_1\Big(\tfrac{t\sqrt{y}}{\sqrt{n}}\Big)\, f_2(y)\,\tfrac{\sqrt{y}}{\sqrt{n}}\,\mathrm{d}y\Big]\,\mathrm{d}t
\end{aligned}$$

e ne segue che la densità di $\frac{\sqrt{n}\,X}{\sqrt{Y}}$ è la funzione

$$g(t) \;=\; \int_0^{+\infty} f_1\Big(\frac{t\sqrt{y}}{\sqrt{n}}\Big)\, f_2(y)\,\frac{\sqrt{y}}{\sqrt{n}}\,\mathrm{d}y$$

Inserendo al posto di f_1 ed f_2 i valori delle densità, e portando avanti conti faticosi anche se non difficili, si prova che la densità g è data da $g(x) = c_n \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}$ dove c_n è una opportuna costante.

Per poter fare dei conti effettivi, sono state predisposte le tavole della legge di Student: in funzione dei gradi di libertà n e di α, riportano il valore $t_{(\alpha,n)}$ dello α–quantile della legge $\tau(n)$).

Poiché T ha una legge simmetrica (cioè la sua densità è una *funzione pari*) si constata facilmente che vale l'eguaglianza $t_{(\alpha\, n)} = -t_{(1-\alpha\,,n)}$; ne segue che se serve individuare un numero t tale che si abbia $\mathbf{P}\big\{|T| > t\big\} = \alpha$, questo numero è dato da $t = t_{(1-\frac{\alpha}{2}\,,n)}$.

Definizione 3.6.4 (Legge di Fisher). Siano C_n e C_m due variabili indipendenti con legge rispettivamente $\chi^2(n)$ e $\chi^2(m)$: si chiama *legge di Fisher* $F_{n,m}$ la legge di

$$\frac{C_n/n}{C_m/m}$$

Il calcolo della densità di tale variabile può essere condotto con passaggi analoghi a quelli appena fatti: la densità risultante è evidentemente nulla sulla semiretta negativa, e per x positivo vale $c(n,m)\, \frac{x^{\frac{n}{2}-1}}{(m+nx)^{\frac{n+m}{2}}}$.

Anche per la legge di Fisher sono state compilate opportune tavole che danno, per alcuni valori di α, lo α-quantile della legge $F_{n,m}$.

Concludiamo osservando che l'uso delle tavole statistiche, nella pratica, è ora superato dall'uso di software statistici.

3.6.2 La misura di Cantor

L'insieme C di Cantor può essere definito come l'insieme dei numeri dell'intervallo $[0,1]$ che possono essere scritti, *in base 3*, utilizzando le sole cifre 0 e 2. Ricordiamo che ogni numero di quell'intervallo può essere scritto, in base 3, nella notazione $0, a_1a_2a_3\ldots$ intendendo con questa notazione $\sum_{n=1}^{+\infty} \frac{a_n}{3^n}$. La notazione è unica con una eccezione: ad esempio il numero $1/3$ si può scrivere $0,100\ldots = 0,1\overline{0}$ ma anche $0,0222\ldots = 0,0\overline{2}$. In questo caso scegliamo la seconda notazione (e quindi $1/3$ si può scrivere con le sole cifre 0 e 2 e pertanto appartiene a C).

L'insieme C si può costruire in questo modo: dall'intervallo $[0,1]$ cominciamo a togliere l'insieme A_1 dei numeri che hanno 1 come prima cifra decimale, cioè l'intervallo aperto $]\frac{1}{3}, \frac{2}{3}[$. Poi togliamo l'insieme A_2 dei numeri che non stanno in A_1 e che hanno 1 come seconda cifra decimale (l'unione dei due intervalli aperti $]\frac{1}{3^2}, \frac{2}{3^2}[$ e $]\frac{7}{3^2}, \frac{8}{3^2}[$) e così via ... Ogni insieme A_n è

formato da 2^{n-1} intervalli aperti di lunghezza 3^{-n} e quindi l'unione di questi insiemi $(A_n)_{n\geq 1}$ (che sono disgiunti) ha misura (secondo Lebesgue) eguale a $\sum_{n=1}^{+\infty} 2^{n-1}3^{-n} = 1$.

Di conseguenza l'insieme C di Cantor (che è il complementare in $[0,1]$ dell'unione di questi intervalli) è un insieme chiuso che ha misura 0 (cioè è trascurabile) secondo Lebesgue. Viceversa la cardinalità di C coincide con quella dell'intervallo $[0,1]$ (e quindi con quella di $\mathbb{R}$) : infatti C può essere rappresentato come $\{0,2\}^{\mathbb{N}}$ (cioè le successioni di cifre 0 e 2) , e la sua cordinalità coincide ovviamente con quella di $\{0,1\}^{\mathbb{N}}$ ed ogni numero tra 0 e 1 può essere rappresentato (in base 2) come successione infinita di cifre 0 e 1.

Costruiamo ora la funzione di ripartizione F della misura di Cantor (che è una probabilità) mediante limite di una successione $(F_n)_{\geq 1}$ di funzioni di ripartizione continue approssimanti (infatti F non può essere scritta con una espressione esplicita): ognuna delle $(F_n)_n$ (e quindi anche il limite) vale 0 per $x \leq 0$ e vale 1 per $x \geq 1$.

Poi F_1 è costante sull'insieme A_1 e lineare a tratti nel complementare: più precisamente vale $\frac{1}{2}$ nei punti $\frac{1}{3}$ e $\frac{2}{3}$ ed è lineare tra 0 e $\frac{1}{3}$ e tra $\frac{2}{3}$ e 1.

Invece F_2 coincide con F_1 su A_1 , è costante si ognuno degli intervalli che compongono A_2 e si raccorda negli altri punti in modo lineare a tratti: vale $\frac{1}{2^2}$ nei punti $\frac{1}{3^2}$ e $\frac{2}{3^2}$, vale $\frac{3}{2^2}$ nei punti $\frac{7}{3^2}$ e $\frac{8}{3^2}$ e così di seguito ...

È facile constatare che, dato $n < m$, si ha, per ogni x, $\left|F_n(x) - F_m(x)\right| \leq 2^{-n}$: di conseguenza la successione F_n è di Cauchy per la convergenza uniforme e pertanto converge uniformemente ad una funzione F che è crescente continua, vale 0 per $x \leq 0$ e 1 per $x \geq 1$, ed è costante su ognuno degli intervalli che compongono $\bigcup_{n\geq 1} A_n$. Pertanto la probabilità $\mathbf{m}$ associata ad F (la *misura di Cantor*) è una probabilità diffusa, concentrata sull'insieme C (nel senso che il complementare di C è trascurabile per $\mathbf{m}$).

Se $\mathbf{m}$ avesse una densità f, si dovrebbe avere

$$1 \,=\, \mathbf{m}\big(C\big) = \int_C f(x)\,\mathrm{d}x$$

ma questo è impossibile poiché l'integrale (secondo Lebesgue) di qualsiasi funzione sull'insieme trascurabile C è 0.

È interessante sapere che ogni probabilità $\mathbf{P}$ sulla retta $\mathbb{R}$ si può scrivere nella forma $\mathbf{P} = \mathbf{m}_1 + \mathbf{m}_2 + \mathbf{m}_3$ dove queste ultime sono *sottoprobabilità* (si ha infatti $\mathbf{m}_1(\mathbb{R}) + \mathbf{m}_2(\mathbb{R}) + \mathbf{m}_3(\mathbb{R}) = 1$) e sono tali che:

1) $\mathbf{m}_1$ è una misura *discreta*;

2) $\mathbf{m}_2$ è definita da una densità f ;

3) $\mathbf{m}_3$ è una misura *diffusa* concentrata su un insieme trascurabile secondo Lebesgue.

La costruzione si può fare in questo modo: si prende la funzione di ripartizione F associata a $\mathbf{P}$ e si considera l'insieme D (al più numerabile, eventualmente vuoto) dei punti di discontinuità di F. La misura $\mathbf{m}_1$ è concentrata nei punti di D e ad ogni punto $x \in D$ è tale che $\mathbf{m}_1(\{x\}) = \Delta F(x)$.

Si può dimostrare che la funzione F è derivabile quasi ovunque (secondo Lebesgue) e la sua derivata f risulta essere una funzione misurabile a valori positivi (e il suo integrale su $\mathbb{R}$ è ≤ 1) : la misura $\mathbf{m}_2$ è associata alla densità f.

La misura $\mathbf{m}_3$ si ottiene come differenza $\mathbf{P} - \mathbf{m}_1 - \mathbf{m}_2$ (cioè, per ogni $A \in \mathcal{B}(\mathbb{R})$, $\mathbf{m}_3(A) = \mathbf{P}(A) - \mathbf{m}_1(A) - \mathbf{m}_2(A)$), e si prova che $\mathbf{m}_3$ è diffusa e concentrata su un insieme trascurabile secondo Lebesgue.

Capitolo 4

Convergenza di variabili aleatorie e teoremi limite.

4.1 Convergenza in probabilità e in legge

Uno studio accurato della *convergenza di variabili aleatorie* sarà oggetto di un corso più avanzato; qui ci limitiamo a qualche elemento utile per i teoremi limite che sono impiegati nell'inferenza statistica.

Ricordiamo la definizione di *convergenza in probabilità*:

Definizione 4.1.1 (**Convergenza in probabilità).** Si dice che la successione di variabili aleatorie $\big(X_n\big)_{n\geq 1}$ converge in probabilità alla v.a. X se, per ogni $\varepsilon > 0$, si ha

$$\lim_{n\to\infty} \mathbf{P}\big\{\big|X_n - X\big| > \varepsilon\big\} \,=\, 0$$

La convergenza in probabilità ad una costante c è un caso particolare di quella definizione, poiché le costanti possono essere viste come delle variabili aleatorie. Notiamo ancora che nella definizione 4.1.1 imporre "$> \varepsilon$" oppure "$\geq \varepsilon$" porta allo stesso risultato, in quanto

$$\big\{\omega\big||X_n(\omega)-X(\omega)| > \varepsilon\big\} \subseteq \big\{\omega\big||X_n(\omega)-X(\omega)| \geq \varepsilon\big\} \subseteq \big\{\omega\big||X_n(\omega)-X(\omega)| > \frac{\varepsilon}{2}\big\}$$

e di conseguenza

$$\mathbf{P}\big\{|X_n - X| > \varepsilon\big\} \leq \mathbf{P}\big\{|X_n - X| \geq \varepsilon\big\} \leq \mathbf{P}\big\{|X_n - X| > \frac{\varepsilon}{2}\big\}$$

Vediamo la seguente leggera generalizzazione del Teorema 2.7.1:

Teorema 4.1.2 (**Legge dei grandi numeri**). *Sia $X_1, X_2, \dots$ una successione di variabili aleatorie dotate di momento secondo, incorrelate, e supponiamo che $\mathbf{E}[X_i] = m$ per ogni i (cioè hanno tutte lo stesso valore atteso) e che esista una costante K tale che si abbia $Var(X_i) \le K$ qualunque sia i (cioè le varianze sono equilimitate). Allora, posto $S_n = X_1 + \cdots + X_n$, la successione $\left(\frac{S_n}{n}\right)_{n\ge 1}$ converge in probabilità ad m.*

Dimostrazione. È sempre una conseguenza della diseguaglianza di Chebishev, osservando che $\mathbf{E}\left[\frac{S_n}{n}\right] = m$ e che $Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}\big(Var(X_1) + \cdots + Var(X_n)\big) \le \frac{K}{n}$. □

Soprattutto in statistica, è usuale indicare $\overline{X}_n = \frac{S_n}{n}$ (la *media empirica* delle variabili $X_1, \dots, X_n$).

A volte sono comodi i criteri seguenti, che vengono enunciati come esercizio:

Esercizio 4.1.3. Sia $(X_n)_{n\ge 1}$ una successione di variabili aleatorie dotate di momento secondo e supponiamo che

$$\lim_{n\to\infty} \mathbf{E}\big[X_n\big] = c \qquad \lim_{n\to\infty} Var\big(X_n\big) = 0$$

Provare che la successione converge in probabilità a c; provare con un controesempio che il criterio è soltanto sufficiente.

Esercizio 4.1.4. Sia $\big(X_n\big)_{n\ge}$ una successione di variabili aleatorie e siano $F_n(.)$ le relative funzioni di ripartizione. Sono equivalenti le affermazioni seguenti:

- $(X_n)_{n\ge 1}$ converge in probabilità a c;
- per $x < c$, $\lim_{n\to\infty} F_n(x) = 0$, e per $x > c$, $\lim_{n\to\infty} F_n(x) = 1$.

Tra le varie proprietà della convergenza in probabilità ci limitiamo alla seguente, che sarà utilizzata più avanti:

Proposizione 4.1.5. *Sia $\big(X_n\big)_{n\ge 1}$ una successione convergente in probabilità a c e sia g una funzione boreliana continua* nel punto *c: allora $Y_n = g(X_n)$ converge in probabilità a $g(c)$.*

Dimostrazione. Dato $\varepsilon > 0$, esiste $\delta > 0$ tale che: $|x - c| \le \delta \Rightarrow |g(x) - g(c)| \le \varepsilon$.

Di conseguenza vale la seguente inclusione di insiemi

$$\Big\{\big|g(X_n) - g(c)\big| > \varepsilon\Big\} \subseteq \Big\{\big|X_n - c\big| > \delta\Big\}$$

□

Per enunciare il Teorema limite di DeMoivre-Laplace è necessario introdurre un altro tipo di convergenza.

Definizione 4.1.6 (Convergenza in legge). Si dice che la successione di v.a. $\big(X_n\big)_{n\geq 1}$ converge *in legge* (o anche *in distribuzione*) alla v.a. X se per ogni $f : \mathbb{R} \to \mathbb{R}$ continua e limitata, si ha

$$\lim_{n\to\infty} \mathbf{E}\big[f\big(X_n\big)\big] \;=\; \mathbf{E}\big[f\big(X\big)\big]$$

Proposizione 4.1.7. *Siano X_n e X variabili aleatorie, F_n ed F le relative funzioni di ripartizione; supponiamo inoltre che F sia continua (cioè la legge di X sia diffusa). Allora sono equivalenti le seguenti affermazioni:*

a) la successione $(X_n)_{n\geq 1}$ converge a X in legge;

b) per ogni $x \in \mathbb{R}$, si ha $\lim_{n\to\infty} F_n(x) = F(x)$.

Dimostrazione. Supponiamo che sia verificato a): scegliamo $x \in \mathbb{R}$, $\delta > 0$ e consideriamo una funzione continua f tale che $f(t) = 1$ per $t \leq x$, $f(t) = 0$ per $t \geq (x + \delta)$, e decrescente tra x e $x + \delta$. Per ogni n, valgono le diseguaglianze

$$F_n(x) \leq \int f(t)\,\mathrm{d}F_n(t) = \mathbf{E}\big[f\big(X_n\big)\big] \leq F_n(x+\delta)$$

(la notazione $\int g(t)\,\mathrm{d}F(t)$ indica l'integrale di g rispetto alla probabilità associata alla funzione di ripartizione F) e le stesse diseguaglianze valgono per la variabile limite . Si ha pertanto

$$F(x+\delta) \geq \int f(t)\mathrm{d}F(t) = \lim_{n\to\infty} \int f(t)\mathrm{d}F_n(t) \geq \limsup_{n\to\infty} F_n(x)$$

In modo analogo si prova la diseguaglianza $F(x-\delta) \leq \liminf_{n\to\infty} F_n(x)$, e per la continuità di F si può concludere che $\lim_{n\to\infty} F_n(x) = F(x)$.

Supponiamo viceversa che sia soddisfatto b) , e consideriamo una funzione continua f uniformemente limitata in modulo dalla costante 1 (ci si può ridurre a questo caso).

Dato $\varepsilon > 0$, esiste $M > 0$ tale che si abbia $F(-M) \leq \varepsilon$ e $F(M) \geq 1 - \varepsilon$; esiste di conseguenza n_1 tale che, per $n \geq n_1$, si abbia $F_n(-M) \leq -2\varepsilon$ e $F_n(M) \geq 1 - 2\varepsilon$.

Consideriamo poi una funzione φ costante a tratti (più precisamente della forma $\varphi(x) = \sum_{i=1}^{n} a_i\, I_{]x_i,x_{i+1}]}(x)$) che sia nulla fuori di $]-M, M]$ e che su quell'intervallo differisca da f per meno di ε.

È evidente che si ha $\lim_{n\to\infty} \int \varphi \, \mathrm{d}F_n = \int \varphi \, \mathrm{d}F$, e dunque esiste n_2 tale che, per $n \geq n_2$, si abbia $\left| \int \varphi \, \mathrm{d}F_n - \int \varphi \, \mathrm{d}F \right| < \varepsilon$.

Sia ora $\overline{n} = \max(n_1, n_2)$ e consideriamo $n \geq \overline{n}$. Valgono le seguenti diseguaglianze

$$\int \left| f - \varphi \right| \mathrm{d}F \leq \int_{]-\infty,-M]} |f| \, \mathrm{d}F + \int_{]-M,M]} |f - \varphi| \, \mathrm{d}F + \int_{]M,+\infty[} |f| \, \mathrm{d}F \leq$$

$$\leq F(-M) + \varepsilon + \big(1 - F(M)\big) \leq 3\varepsilon$$

In modo analogo si prova che si ha $\int |f - \varphi| \, \mathrm{d}F_n \leq 5\varepsilon$.

Si ottengono allora le disuguaglianze:

$$\left| \int f \mathrm{d}F_n - \int f \mathrm{d}F \right| \leq \int |f - \varphi| \, \mathrm{d}F_n + \left| \int \varphi \, \mathrm{d}F_n - \int \varphi \, \mathrm{d}F \right| + \int |f - \varphi| \mathrm{d}F \leq 9\varepsilon$$

Poiché questo si verifica per ogni $\varepsilon > 0$, si ottiene così il risultato. □

Esercizio 4.1.8. Se il limite è una costante c, è equivalente affermare che la successione $(X_n)_{n\geq 1}$ converge in probabilità oppure in legge a c.

4.2 Il teorema di De Moivre-Laplace (e introduzione al teorema Limite Centrale)

Il teorema di De Moivre-Laplace che viene ora enunciato, è un caso particolare (limitato al caso delle variabili di Bernoulli) del *Teorema del Limite Centrale*: sia $X_1, X_2, \ldots$ una successione di variabili indipendenti di Bernoulli di parametro p con $0 < p < 1$, denotiamo $q = 1 - p$ e $S_n = X_1 + \cdots + X_n$.

Teorema 4.2.1 (**Limite Centrale per Variabili Binomiali**). *Presi due numeri a, b con $-\infty \leq a < b \leq +\infty$, si ha*

$$\lim_{n\to\infty} \mathbf{P}\Big\{ a \leq \frac{S_n - np}{\sqrt{npq}} \leq b \Big\} = \frac{1}{\sqrt{2\pi}} \int_a^b \mathrm{e}^{-\frac{x^2}{2}} \mathrm{d}x$$

Prima di affrontare la dimostrazione (elementare ma piuttosto tecnica) vediamo un esempio di applicazione del teorema 4.2.1.

Esempio 4.2.2. Sia $X \sim B(400\,;\, 0{,}05)$: vogliamo calcolare $\mathbf{P}\big\{X > 30\big\}$.

Il conto esplicito non è fattibile, tuttavia (essendo 400 *grande*) i conti che riguardano la variabile $\frac{X-20}{\sqrt{400\times 0{,}05\times 0{,}95}}$ si possono approssimare con la formula risultante dal teorema 4.2.1. Si ha pertanto

$$\mathbf{P}\big\{X > 30\big\} = \mathbf{P}\Big\{\frac{X-20}{\sqrt{400\times 0{,}05\times 0,95}} > \frac{30-20}{\sqrt{400\times 0{,}05\times 0{,}95}}\Big\} =$$

$$1 - \mathbf{P}\Big\{\frac{X-20}{\sqrt{400\times 0{,}05\times 0{,}95}} \le 2{,}29\Big\}$$

Questo numero si può approssimare con $\big(1-\Phi(2{,}29)\big) = 1 - 0{,}989 = 0{,}011$.

Per affrontare la dimostrazione del Teorema 4.2.1, è necessario stabilire prima alcuni risultati.

Lemma 4.2.3 (**Formula di Stirling**). *Esiste una costante positiva* c *tale che per ogni intero* n *si abbia*

$$n! \;=\; c\,\big(\frac{n}{e}\big)^n\sqrt{n}\,\exp(\theta_n) \;=\; c\,n^{n+\frac{1}{2}}\,e^{-n}\,\exp(\theta_n)$$

dove $\frac{1}{12n+1} \le \theta_n \le \frac{1}{12n}$

La dimostrazione di questo come del successivo lemma, entrambe elementari ma piuttosto tecniche, saranno riportate in Appendice.

Con le notazioni del Teorema 4.2.1, chiamiamo $Z_n = \frac{S_n - np}{\sqrt{npq}}$ e sia I_n l'insieme dei valori della variabile Z_n: notiamo che I_n è formato da $(n+1)$ punti che distano $\frac{1}{\sqrt{npq}}$ uno dall'altro, e che il minimo ed il massimo di questi punti convergono (quando $n \to +\infty$) rispettivamente a $-\infty$ ed a $+\infty$.

Lemma 4.2.4. *Presi* $-\infty < a < b < +\infty$, *il numero*

$$\max_{x\in I_n\cap[a,b]} \Big|c\,\sqrt{npq}\;\mathbf{P}\big\{Z_n = x\big\} - \exp\big(-\frac{x^2}{2}\big)\Big|$$

(dove c *è la stessa costante della formula di Stirling), converge a* 0 *se* n *tende a* $+\infty$.

Tenendo conto del fatto che il minimo della funzione $\exp\big(-\frac{x^2}{2}\big)$ sull'intervallo $[a,b]$ è strettamente positivo, si può riscrivere l'enunciato del lemma 4.2.4 nella forma seguente, che sarà più comoda per la successiva dimostrazione:

Fissati $-\infty < a < b < +\infty$ *e dato* $\varepsilon > 0$, *esiste* $\bar n = \bar n(\varepsilon, a, b)$ *tale che, per* $n \ge \bar n$ *ed* $x \in I_n \cap [a,b]$ *si abbia*:

$$\mathbf{P}\big\{Z_n = x\big\} \;=\; \frac{c^{-1}}{\sqrt{npq}}\exp\big(-\frac{x^2}{2}\big)\Big(1+\alpha(x)\Big) \quad \text{con} \quad \big|\alpha(x)\big| < \varepsilon.$$

Siamo ora in grado di affrontare la dimostrazione del Teorema 4.2.1.

Dimostrazione. Fissiamo $-\infty < a < b < +\infty$ (il caso $a = -\infty$ oppure $b = +\infty$ si riporta a questo con piccole modifiche) e, dato $\varepsilon > 0$, scegliamo $\bar{n} = \bar{n}(\varepsilon, a, b)$ come sopra. Si ha:

$$\mathbf{P}\big\{a \le Z_n \le b\big\} = \sum_{x \in I_n \cap [a,b]} \mathbf{P}\big\{Z_n = x\big\} = \frac{c^{-1}}{\sqrt{npq}} \sum_{x \in I_n \cap [a,b]} \exp\big(-\frac{x^2}{2}\big)\Big(1+\alpha(x)\Big)$$

La somma

$$\frac{c^{-1}}{\sqrt{npq}} \sum_{x \in I_n \cap [a,b]} \exp\big(-\frac{x^2}{2}\big)$$

è un'approssimazione dell'integrale (di Riemann) $c^{-1}\int_a^b \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x$ e pertanto converge (per $n \to \infty$) proprio a $c^{-1}\int_a^b \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x$.

Viceversa la somma

$$\frac{c^{-1}}{\sqrt{npq}} \sum_{x \in I_n \cap [a,b]} \exp\big(-\frac{x^2}{2}\big)|\alpha(x)|$$

è, per $n \ge \bar{n}$, inferiore a $K\varepsilon$, con K costante positiva indipendente da n, e pertanto converge a 0.

L'ultimo passo è provare che $c = \sqrt{2\pi}$. Partiamo dall'osservazione che ogni variabile Z_n ha valore atteso 0 e varianza 1 : di conseguenza, per la diseguaglianza di Chebishev,

$$\mathbf{P}\Big\{-a \le Z_n \le a\Big\} = 1 - \mathbf{P}\big\{\big|Z_n\big| > a\Big\} \ge 1 - \frac{1}{a^2}$$

è arbitrariamente vicino a 1 per a sufficientemente grande, e al limite, anche $c^{-1}\int_{-a}^{a} \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x$ è arbitrariamente vicino a 1.

Ricordando che $\int_{-\infty}^{+\infty} \exp\big(-\frac{x^2}{2}\big)\,\mathrm{d}x = \sqrt{2\pi}$, si ottiene l'eguaglianza cercata. □

Osservazione 4.2.5. Il Teorema 4.2.1 mostra quale è la velocità di convergenza di $\frac{S_n}{n}$ a p: dall'eguaglianza

$$\frac{S_n - np}{\sqrt{npq}} = \frac{\sqrt{n}}{\sqrt{pq}}\Big(\frac{S_n}{n} - p\Big)$$

segue che tale velocità è dell'ordine di $\frac{1}{\sqrt{n}}$.

Questa velocità, purtroppo piuttosto lenta, è la tipica velocità di convergenza dei teoremi limite della Statistica.

Alla luce paragrafo precedente, il Teorema 4.2.1 (teorema *Limite Centrale* per variabili Binomiali) è un risultato di *convergenza in Legge*. In verità quel risultato è valido in ipotesi molto più generali, e la dimostrazione è lasciata ad un corso più avanzato: tuttavia è comodo poter utilizzare subito il risultato generale. Quello che viene qui enunciato, senza dimostrazione, è il *Teorema Limite Centrale di Paul Lévy*:

Teorema 4.2.6. *Sia $X_1, X_2, \dots$ una successione di variabili indipendenti equidistribuite, dotate di momento primo μ e di varianza σ^2 (diversa da 0): posto $S_n = X_1 + \cdots + X_n$, la successione*

$$\frac{S_n - n\mu}{\sqrt{n}\,\sigma} = \sqrt{n}\Big(\frac{\overline{X}_n - \mu}{\sigma}\Big)$$

converge in legge alla variabile gaussiana $N(0,1)$.

Osservazione 4.2.7. Abbiamo visto come si possono costruire n v.a. $X_1, \dots, X_n$ indipendenti con leggi assegnate $\mathbf{P}_1, \dots, \mathbf{P}_n$, ma nei precedenti *teoremi limite* intervengono *successioni* di variabili aleatorie: in realtà si può costruire una sorta di *prodotto infinito* di probabilità, ma questo sarà l'oggetto di un corso più avanzato. Tuttavia questa costruzione non è necessaria per dare un senso sia alla legge dei Grandi Numeri che al teorema Limite Centrale. È sufficiente infatti costruire per ogni n, eventualmente su diversi spazi Ω_n , le variabili $X_1, \dots, X_n$: questo permette di dare un senso a quantità come $\mathbf{P}^n\big\{\big|\frac{S_n}{n} - m\big| > \varepsilon\big\}$ oppure $\mathbf{P}^n\big\{a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\big\}$, e solo queste intervengono negli enunciati dei teoremi limite sopra riportati.

4.3 Appendice

Quella che segue è la dimostrazione della *Formula di Stirling* (Lemma 4.2.3)

Dimostrazione. Partendo dalla diseguaglianza

$$\int_{k-1}^{k} \log(x)\,\mathrm{d}x < \log(k) < \int_{k}^{k+1} \log(x)\,\mathrm{d}x$$

si ottiene, per ogni intero n strettamente positivo,

$$\int_{0}^{n} \log(x)\,\mathrm{d}x < \log(n!) < \int_{1}^{n+1} \log(x)\,\mathrm{d}x$$

e, calcolando gli integrali,

$$n\log(n) - n < \log(n!) < (n+1)\log(n+1) - n$$

Consideriamo allora la differenza $d_n = \log(n!) - (n+\frac{1}{2})\log(n) + n$: notiamo che $d_n - d_{n+1} = (n+\frac{1}{2})\log\left(\frac{n+1}{n}\right) - 1$; inoltre

$$\frac{n+1}{n} = \frac{1+\frac{1}{2n+1}}{1-\frac{1}{2n+1}}$$

Ricordiamo ancora che vale lo sviluppo in serie (convergente per $|t|<1$):

$$\frac{1}{2}\log\left(\frac{1+t}{1-t}\right) = t + \frac{t^3}{3} + \frac{t^5}{5} + \cdots$$

Si ottiene pertanto

$$d_n - d_{n+1} = \frac{1}{3(2n+1)^2} + \frac{1}{5(2n+1)^4} + \cdots$$

Da quest'ultima eguaglianza (ricordando anche la somma di una serie di potenze) si ottiene:

$$\frac{1}{3(2n+1)^2} < d_n - d_{n+1} < \frac{1}{3\big((2n+1)^2-1\big)} = \frac{1}{12n} - \frac{1}{12(n+1)}$$

Un conto facile ma laborioso prova che

$$\frac{1}{12(n+1)} - \frac{1}{12(n+1)+1} < \frac{1}{3(2n+1)^2}$$

e da qui si ottengono le diseguaglianze:

$$d_n - \frac{1}{12n} < d_{n+1} - \frac{1}{12(n+1)} < d_{n+1} - \frac{1}{12(n+1)+1} < d_n - \frac{1}{12n+1}$$

Quindi la successione $\big(d_n - \frac{1}{12n}\big)_{n\geq 1}$ è crescente (rispettivamente $\big(d_n - \frac{1}{12n+1}\big)_{n\geq 1}$ è decrescente), e ponendo $c' = \lim_n d_n$ si ottiene

$$c' + \frac{1}{12n+1} < d_n < c' + \frac{1}{12n}$$

e, chiamato $c = \exp(c')$, si hanno finalmente le diseguaglianze:

$$c\, n^{n+\frac{1}{2}} \exp\big(-n + \frac{1}{12n+1}\big) < n! < c\, n^{n+\frac{1}{2}} \exp\big(-n + \frac{1}{12n}\big)$$

□

Segue la dimostrazione del Lemma 4.2.4

Dimostrazione. Sia $x \in I_n \cap [a,b\,] : \mathbf{P}\big\{Z_n = x\big\} = \mathbf{P}\big\{S_n = k\big\}$, essendo $k = np + x\sqrt{npq}$. Poniamo poi $j = n - k = nq - x\sqrt{npq}$.

Ricordando che $S_n \sim B(n,p)$ ed utilizzando la *formula di Stirling*, si ottiene

$$c\sqrt{npq}\ \mathbf{P}\big\{Z_n = x\big\} = \sqrt{npq}\ \sqrt{\frac{n}{kj}}\left(\frac{np}{k}\right)^k\left(\frac{nq}{j}\right)^j \exp\big(\theta_n - \theta_j - \theta_k\big)$$

Si osserva facilmente che, poichè $a \le x \le b$, j e k convergono a $+\infty$ (uniformemente rispetto a $x \in I_n$) quando n tende a $+\infty$; quindi (poiché $\big|\theta_n - \theta_j - \theta_k\big| \le \frac{1}{2n} + \frac{1}{2j} + \frac{1}{2k}$) il termine $\exp\big(\theta_n - \theta_j - \theta_k\big)$ converge uniformemente a 1.

Anche il termine

$$\sqrt{npq}\ \sqrt{\frac{n}{kj}} = \sqrt{\frac{n^2pq}{\big(np + x\sqrt{npq}\big)\big(nq - x\sqrt{npq}\big)}}$$

converge a 1, uniformemente rispetto a $x \in I_n$.

Proviamo ora che

$$\Big(\log\big(\frac{np}{k}\big)^k + x\sqrt{npq} + \frac{x^2q}{2}\Big)$$

converge uniformemente a 0 ; allo stesso modo si prova che

$$\Big(\log\big(\frac{nq}{j}\big)^j - x\sqrt{npq} + \frac{x^2p}{2}\Big)$$

converge uniformemente a 0, e questo completa la dimostrazione.

Esaminiamo dunque il termine

$$k\,\log\big(\frac{np}{k}\big) \;=\; \big(np + x\sqrt{npq}\big)\,\log\Big(1 - \frac{x\sqrt{npq}}{np + x\sqrt{npq}}\Big)\ ;$$

utilizzando lo sviluppo di Taylor $\log(1+t) = t - \frac{t^2}{2} + o(t^3)$, si ottiene che questo termine è eguale a

$$-x\sqrt{npq} \;-\; \frac{x^2npq}{2\big(np + x\sqrt{npq}\big)} \;+\; \big(np + x\sqrt{npq}\big)\,o\big(n^{-\frac{3}{2}}\big)$$

e questo è proprio il risultato cercato.

□

Capitolo 5

Introduzione all'inferenza statistica

5.1 Due parole sulla statistica descrittiva

Si parla di **statistica descrittiva** quando vengono analizzati i dati di una *indagine statistica* senza l'interpretazione di un modello probabilistico.

Possiamo rappresentare un'indagine statistica come una applicazione $\mathcal{X}$ da un insieme finito $\{1, 2, \ldots, n\}$ su un insieme C. Se C è un insieme di cardinalità piccola si parla di *indagine su un carattere qualitativo* (ad esempio un sondaggio sull'orientamento politico), mentre se $C = \mathbb{R}$ (o più generalmente $\mathbb{R}^d$) si parla di *indagine su un carattere quantitativo* (o su più caratteri quantitativi).

Limitiamoci all'indagine su un carattere quantitativo: l'indagine $\mathcal{X}$ corrisponde a una n-pla di numeri $\{x_1, \ldots, x_n\}$.

Assegnati questi numeri si chiama **media empirica** la quantità $\overline{x} = \dfrac{x_1 + \cdots + x_n}{n}$ e **varianza empirica** la quantità $\displaystyle\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$: si può osservare che questi possono essere interpretati come la speranza ed la varianza di una v.a. X che prende i valori $x_1, \ldots, x_n$ con distribuzione uniforme (cioè ciascuno con probabilità $1/n$).

Se invece abbiamo un'indagine su due caratteri quantitativi $(\mathcal{X}, \mathcal{Y})$ si chiama *covarianza empirica* la quantità

$$\sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n}$$

ed in modo analogo si può definire il *coefficiente di correlazione empirico*, la *retta di regressione*, ecc...

Non insistiamo ulteriormente su queste definizioni perché siamo interessati all'inferenza statistica: si parla di **inferenza statistica** quando si usano i risultati di una indagine statistica *per ricostruire un modello probabilistico che descriva opportunamente il fenomeno osservato.*

5.2 Modelli statistici

Introduciamo le idee fondamentali dell'*Inferenza Statistica* con un esempio, che d'ora innanzi chiameremo **Controllo di qualità**: è probabilmente il più semplice che si possa immaginare, ma sufficiente per presentare le idee fondamentali.

Vogliamo controllare la percentuale *sconosciuta* di pezzi difettosi in un insieme (ad esempio un grosso acquisto di certi componenti elettronici dall'estero), insieme che in statistica è usualmente denominato *popolazione*: per fare questo (non potendo verificare tutti i pezzi, per mancanza di tempo o altri motivi) estraiamo un *campione* di n pezzi che vengono verificati. I risultati di questa verifica saranno n variabili aleatorie $X_1, \ldots, X_n$ indipendenti, con legge di Bernoulli di parametro θ , $0 < \theta < 1$ (la variabile X_i prende il valore 1 se l'i-esimo pezzo risulta difettoso, altrimenti prende il valore 0): possiamo formalizzare la situazione in questo modo.

Consideriamo sullo spazio $\Omega = \{0,1\}^n$ (munito della σ-algebra di tutte le parti) la famiglia di probabilità $\Big(\mathbf{P}^\theta , \theta \in (0,1) \Big)$, definite da $\mathbf{P}^\theta(k_1, \ldots, k_n) = \theta^{k_1+\cdots+k_n}\big(1-\theta\big)^{n-(k_1+\cdots+k_n)}$; definiamo poi $X_i(k_1, \ldots, k_n) = k_i$ (cioè X_i è la proiezione coordinata di indice i). È immediato verificare che, se si considera su Ω la probabilità $\mathbf{P}^\theta$, (più avanti diremo sbrigativamente *sotto* $\mathbf{P}^\theta$) le variabili X_i risultano indipendenti, con legge di Bernoulli di parametro θ.

Possiamo cominciare a dare qualche definizione:

Definizione 5.2.1 (**Modello statistico).** Si chiama *modello statistico* una terna $\Big(\Omega, \mathcal{F}, \big(\mathbf{P}^\theta , \theta \in \Theta\big)\Big)$ dove Ω è un insieme, $\mathcal{F}$ una σ-algebra di parti di Ω e, per ogni $\theta \in \Theta$, $\mathbf{P}^\theta$ è una probabilità su $\big(\Omega, \mathcal{F}\big)$.

Supporremo sempre che a due parametri diversi θ_1 e θ_2 corrispondano due probabilità diverse (come si usa dire, il modello è **identificabile**).

In un modello statistico si chiama **trascurabile** un evento $A \in \mathcal{F}$ trascurabile per ogni probabilità $\mathbf{P}^\theta$.

Cominciamo a supporre che lo spazio Ω sia numerabile (e, se non ci sono ragioni per fare diversamente, diamo per sottinteso che $\mathcal{F}$ è la σ-algebra di tutte le parti di Ω).

Definizione 5.2.2 (**Verosimiglianza in un modello statistico discreto).** Assegnato un modello statistico $\Big(\Omega, \mathcal{F}, \big(\mathbf{P}^\theta\,, \theta \in \Theta\big)\Big)$ con Ω numerabile, si chiama *verosimiglianza* la funzione $L : \Theta \times \Omega \to \mathbb{R}^+$ definita da

$$L(\theta, \omega) \;=\; \mathbf{P}^\theta\big(\{\omega\}\big)$$

Naturalmente la verosimiglianza identifica la probabilità, poiché per ogni evento A vale la formula $\mathbf{P}^\theta\big(A\big) = \sum_{\omega_i \in A} L(\theta, \omega_i)$; la funzione L deve verificare la condizione $\sum_{\omega_i \in \Omega} L(\theta, \omega_i) = 1$. La notazione $L(.\,,.)$ deriva dall'inglese *Likelihood* e, nel caso discreto, in realtà L è a valori in $[0, 1]$; tuttavia nei casi che esamineremo più avanti sarà generalmente a valori in $\mathbb{R}^+$.

Abbiamo gli strumenti matematici per indagare il caso in cui Ω è uno spazio qualsiasi, tuttavia per evitare eccessive generalizzazioni e poter fare conti concreti, esamineremo come secondo esempio quello di un *modello con densità*.

Definizione 5.2.3 (**Modello con densità).** Il modello statistico è detto *con densità* se soddisfa le seguenti condizioni:

a) Ω è uno spazio euclideo $\mathbb{R}^n$ (o un sottinsieme misurabile di uno spazio euclideo);

b) $\mathcal{F}$ è la σ-algebra di Borel su Ω;

c) le probabilità $\mathbf{P}^\theta$ ammettono *densità* rispetto alla misura di Lebesgue n-dimensionale $\boldsymbol{\lambda}$.

Osservazione 5.2.4. La σ-algebra di Borel $\mathcal{B}(A)$ su un sottinsieme misurabile $A \subseteq \mathbb{R}^n$ è formata dalle intersezioni degli elementi di $\mathcal{B}(\mathbb{R}^n)$ con A, o (equivalentemente) è generata dagli aperti di A.

Definizione 5.2.5 (**Verosimiglianza in un modello con densità).** Si chiama *verosimiglianza* una funzione $L : \Theta \times \Omega \to \mathbb{R}^+$ tale che, fissato θ, $L(\theta, .)$ sia una versione della *densità* di $\mathbf{P}^\theta$ (rispetto alla misura di Lebesgue $\boldsymbol{\lambda}$).

Conoscere la verosimiglianza equivale a conoscere ogni probabilità $\mathbf{P}^\theta$, in quanto si ha per ogni $A \in \mathcal{F}$, $\mathbf{P}^\theta\big(A\big) \;=\; \int\!..\!\int_A L\big(\theta\,; x_1, \ldots, x_n\big)\,\mathrm{d}x_1 \ldots \mathrm{d}x_n$.

Osservazione 5.2.6. Apparentemente c'è una incongruenza tra le due definizioni, ma in realtà non è così : entrambe sono casi particolari della definizione *generale* di densità.

Date due misure $\mathbf{m}_1$ e $\mathbf{m}_2$ su $(E, \mathcal{E})$, si dice che $\mathbf{m}_2$ è definita dalla densità f rispetto a $\mathbf{m}_1$ se f è misurabile positiva e si ha, per ogni $A \in \mathcal{E}$,

$$\mathbf{m}_2(A) = \int_A f(e)\, \mathrm{d}\mathbf{m}_1(e)$$

Se si considera su un insieme numerabile Ω la misura $\mathbf{m}$ *che conta i punti* (cioè $\mathbf{m}(A) = \#A$ se A è in insieme finito, $\mathbf{m}(A) = +\infty$ se A è infinito), è facile verificare che la funzione $\omega \to \mathbf{P}^\theta(\{\omega\})$ è la *densità* di $\mathbf{P}^\theta$ rispetto a $\mathbf{m}$.

Scopo dell'inferenza statistica è partire dall'esperienza (l'osservazione del campione) per risalire a informazioni sulla legge di probabilità che meglio si adatta a descrivere il modello, e per ottenere questo i metodi dell'inferenza statistica sono essenzialmente tre:

- la stima statistica
- gli intervalli di fiducia
- i test statistici

Le definizioni precise verranno date nei prossimi paragrafi; cerchiamo ora di introdurre questi concetti a livello intuitivo, sempre riferendoci all'esempio del *controllo di qualità*. Indichiamo con $\overline{X}(\omega) = \frac{X_1(\omega)+\cdots+X_n(\omega)}{n}$ la media aritmetica (o meglio *media empirica*) delle variabili X_i (percentuale di pezzi difettosi riscontrati nell'indagine statistica), ed è importante ribadire che si tratta di una *variabile aleatoria*, cioè il risultato di questa indagine statistica dipende dal caso.

Non avendo per il momento risultati teorici più precisi, sembra opportuno considerare proprio $\overline{X}(\omega)$ come *stima* del parametro θ.

Quanto all'*intervallo di fiducia*, appare evidente che una maggiore ampiezza del campione permettere di rafforzare l'affidabilità dell'informazione: per spiegarci meglio, 2 pezzi difettosi su 10 oppure 200 su 1000 portano alla stessa stima (in entrambi i casi θ viene stimato 0,2), ma è evidente che il secondo risultato è molto più rassicurante. Come si può *misurare questa sicurezza*?
È interessante osservare che nella vita pratica si incontrano più volte gli intervalli di fiducia, senza rendersene conto, ad esempio quando vengono trasmesse le proiezioni sui risultati delle elezioni. Le prime proiezioni danno per il partito x una percentuale t con un'oscillazione ad esempio di 2 punti percentuali (in più o in meno), dopo due ore la percentuale è cambiata (magari di poco) ma l'oscillazione è stata ridotta a 0,5 punti, e così via ...

Effettuare un *test statistico* significa invece formulare un'ipotesi e pianificare un'esperienza per decidere se accettare o rifiutare l'ipotesi: ad esempio nel caso del controllo di qualità l'ipotesi potrebbe essere *"la ditta fornitrice garantisce che la percentuale di pezzi difettosi non supera il 5%"*(cioè $\theta \leq 0,05$). È evidente che l'ipotesi viene accettata se si osserva $\overline{X}(\omega) = 0,036$ e rifiutata se $\overline{X}(\omega) = 0,09$, ma che fare se $\overline{X}(\omega) = 0,049$ oppure 0,052 ?

A tutti questi problemi verrà data risposta nei paragrafi successivi.

Diamo ora una nuova definizione:

Definizione 5.2.7 (Campione). Sia $\big(\mathbf{m}^\theta\,,\,\theta\in\Theta\big)$ una famiglia parametrizzata di leggi di probabilità su $\mathbb{R}$: si chiama *campione di taglia n e legge* $\mathbf{m}^\theta$ una famiglia $(X_1,\ldots,X_n)$ di n variabili aleatorie indipendenti ciascuna con legge $\mathbf{m}^\theta$.

Notiamo che questa definizione è una generalizzazione dell'esempio del controllo di qualità: in questo caso $(X_1,\ldots,X_n)$ è un campione di legge di Bernoulli di parametro $\theta\,,\,0<\theta<1$.

Cominciamo col caso in cui ogni probabilità $\mathbf{m}^\theta$ è discreta: il modo canonico per rappresentare come modello statistico un campione di legge $\big(\mathbf{m}^\theta\,,\,\theta\in\Theta\big)$ è il seguente. Sia C l'insieme su cui sono concentrate le probabilità $\mathbf{m}^\theta$, e poniamo (per $\theta\in\Theta$ e $x_i\in C$), $p(\theta,x_i)=\mathbf{m}^\theta\big(\{x_i\}\big)$.

Poniamo poi $\Omega=C^n\ ,\ \mathcal{F}=\mathcal{P}(\Omega)$ e scegliamo come verosimiglianza

$$L(\theta\,;\,x_1,\ldots,x_n)\ =\ p(\theta,x_1)\cdots p(\theta,x_n)$$

(ricordiamo che assegnare una verosimiglianza equivale ad assegnare le probabilità $(\mathbf{P}^\theta\,,\,\theta\in\Theta)$). Consideriamo come X_i la proiezione canonica di indice i da Ω su C: le variabili $X_1,\ldots,X_n$ sono effettivamente indipendenti e ciascuna con legge $\mathbf{m}^\theta$ (se si considera su Ω la probabilità $\mathbf{P}^\theta$).

Vediamo ora il caso in cui le probabilità $\mathbf{m}^\theta$ sono definite da una densità. Sia $\big(f(\theta,.)\,,\theta\in\Theta\big)$ una famiglia parametrizzata di densità di probabilità su $\mathbb{R}$: si chiama *campione di taglia n e densità* $f(\theta,.)$ una famiglia di variabili aleatorie indipendenti, equidistribuite, aventi densità $f(\theta,.)$ (sotto $\mathbf{P}^\theta$).

La costruzione canonica del modello è la seguente: si prende $\Omega=\mathbb{R}^n$ e si considera come *verosimiglianza* la funzione

$$L\big(\theta\,;x_1,\ldots,x_n\big)\ =\ \prod_{i=1}^{n} f(\theta,x_i)$$

Si definiscono inoltre come variabili X_i le *proiezioni canoniche* di indice i: è immediato verificare che ponendo su Ω la probabilità $\mathbf{P}^\theta$ definita dalla densità $L(\theta,.)$ queste variabili risultano indipendenti ciascuna con densità $f(\theta,.)$.

Se ogni densità $f(\theta,.)$ si annulla fuori di un intervallo $\mathcal{I}\subseteq\mathbb{R}$, conviene considerare come spazio $\Omega=\mathcal{I}^n$ anzichè $\mathbb{R}^n$.

5.3 Teoria della Stima

Definizione 5.3.1 (Stima). Assegnato un modello statistico $\big(\Omega, \mathcal{F}, (\mathbf{P}^\theta, \theta \in \Theta)\big)$, si chiama *stima* una variabile aleatoria $U : \Omega \to \mathbb{R}$.

In genere una stima è accoppiata ad una funzione $g : \Theta \to \mathbb{R}$ e lo scopo di U è appunto valutare $g(\theta)$. Non si stima necessariamente direttamente θ per due motivi: non è detto che θ sia un numero e in ogni caso talvolta è più agevole stimare una funzione del parametro.

Definizione 5.3.2 (Stima corretta). Assegnata una funzione $g : \Theta \to \mathbb{R}$, la stima U di $g(\theta)$ è detta *corretta* se, per ogni θ, U è $\mathbf{P}^\theta$-integrabile e si ha $\mathbf{E}^\theta\big[U\big] = g(\theta)$.

Il termine anglosassone per stima corretta è *unbiased*, talvolta tradotto *non distorta*.

Esempio 5.3.3. In un campione di taglia n e legge *Geometrica* di parametro θ $(0 < \theta < 1)$, $\overline{X} = \frac{X_1 + \cdots + X_n}{n}$ è una stima corretta di θ^{-1}.

La definizione che viene ora presentata offre un criterio *asintotico* di bontà di una stima.

Definizione 5.3.4 (Stima consistente). Sia $\big(\mathbf{m}^\theta, \theta \in \Theta\big)$ una famiglia di leggi di probabilità discrete su $\mathbb{R}$ e consideriamo, per ogni n, un campione $X_1, \ldots, X_n$ di legge $\mathbf{m}^\theta$; sia poi $U_n = h_n(X_1, \ldots, X_n)$ una stima di $g(\theta)$ basata sulle osservazioni del campione n-simo. Si dice che la successione di stime $\big(U_n\big)_{n \geq 1}$ è *consistente* se, scelti comunque $\theta \in \Theta$ ed $\varepsilon > 0$, si ha

$$\lim_{n \to \infty} \mathbf{P}^\theta\big\{\big|U_n - g(\theta)\big| > \varepsilon\big\} = 0$$

Commentiamo la definizione appena data: la successione di stime è consistente se, qualunque sia la probabilità $\mathbf{P}^\theta$, U_n converge in probabilità a $g(\theta)$. La difficoltà che si pone però è poter costruire un modello statistico che contenga un *campione infinito*, cioè una estensione a una successione di variabili aleatorie della costruzione esposta alla fine della sezione precedente. Questo si può effettivamente fare, ma richiede risultati di teoria della misura più avanzati di quelli esposti in questo corso: con gli strumenti di cui disponiamo, però, si può costruire per ogni n un modello statistico $\Big(\Omega_n, \mathcal{F}_n, \big(\mathbf{P}_n^\theta, \theta \in \Theta\big)\Big)$ relativo al campione di taglia n. La definizione dovrebbe allora essere data nel modo seguente: scelti comunque $\theta \in \Theta$ ed $\varepsilon > 0$, si ha

$$\lim_{n \to \infty} \mathbf{P}_n^\theta\big\{\big|U_n - g(\theta)\big| > \varepsilon\big\} = 0$$

Il metodo più usuale per identificare stime consistenti consiste nell'utilizzare la legge dei grandi numeri, come si può verificare facilmente nell'esempio seguente:

Esempio 5.3.5. In un campione infinito di leggi di Poisson di parametro $\theta\,, (0 < \theta < \infty)$, la successione delle medie empiriche $\overline{X}_n = \frac{X_1 + \cdots + X_n}{n}$ è una stima consistente di θ.

5.4 Stime e riassunti esaustivi

Definizione 5.4.1 (Rischio). Sia U una stima della funzione $g(\theta)$: si chiama *Rischio* (quadratico) il numero

$$R\big(\theta, U\big) \;=\; \mathbf{E}^{\theta}\big[\big(U - g(\theta)\big)^2\big]$$

Notiamo che ha senso parlare di rischio anche se, per qualche θ, U non ha momento secondo: in tal caso il rischio è eguale a $+\infty$. Tuttavia, nel seguito di questo paragrafo, supponiamo tacitamente che tutte le stime considerate abbiano momento secondo qualunque sia la probabilità $\mathbf{P}^{\theta}$.

Osserviamo ancora che, se U è corretta, $R\big(\theta, U\big) = Var^{\theta}\big(U\big)$.

La definizione di rischio introduce un criterio di *ordinamento parziale* tra le stime, più precisamente diremo che

- U è *preferibile* a V se, per ogni θ, $R(\theta, U) \leq R(\theta, V)$;
- U è *strettamente preferibile* a V se è preferibile e, per almeno un parametro $\overline{\theta}$, $R(\overline{\theta}, U) < R(\overline{\theta}, V)$;
- U è *ammissibile* se non esistono stime strettamente preferibili a U;
- U è *ottimale* se è preferibile a ogni altra stima.

Naturalmente due stime non sono necessariamente confrontabili.

La nozione di rischio è strettamente legata alla nozione di riassunto esaustivo; prima di definire quest'ultima torniamo all'esempio del controllo di qualità. Negli esempi che abbiamo visto, non avevano importanza i singoli risultati delle varie prove, ma solo il numero totale di pezzi difettosi: trattenere questo unico dato costituisce evidentemente un notevole *risparmio di informazione*.

La definizione che segue ha proprio lo scopo di formalizzare questa idea di risparmio di informazione.

Definizione 5.4.2 (Riassunto esaustivo). Sia $T : \Omega \to E$ una variabile aleatoria: si dice che T è un *riassunto esaustivo* se si può scrivere la verosimiglianza nella forma

$$L\big(\theta, \omega\big) \;=\; h(\theta, T(\omega))\; k(\omega)$$

Quasi sempre T è a valori reali o più generalmente in uno spazio euclideo $\mathbb{R}^k$. Accanto alla terminologia di riassunto esaustivo, si usa anche quella di *statistica esaustiva* o *statistica sufficiente*.

Apparentemente la definizione 5.4.2 non ha nulla a che vedere con l'idea originale di risparmio di informazione; tutto sarà più chiaro dopo il risultato che segue.

Teorema 5.4.3. *Sia T un riassunto esaustivo, U una stima di $g(\theta)$ e supponiamo che U sia di quadrato integrabile per ogni probablità $\mathbf{P}^\theta$. Esiste una stima V della forma $V(\omega) = f\big(T(\omega)\big)$* preferibile *a U, inoltre V è* strettamente *preferibile a meno che U non sia già nella forma $f \circ T$. Infine, se U è corretta, anche V è corretta.*

Prima di affrontare la dimostrazione, commentiamo il risultato: se T è un riassunto esaustivo, le *buone stime* (in particolare le stime ammissibili) sono funzione di $T(\omega)$ e quindi $T(\omega)$ contiene tutte le informazioni rilevanti.

Vedremo la dimostrazione del Teorema 5.4.3 nel caso in cui lo spazio Ω è numerabile: in questo caso la dimostrazione è piuttosto lunga, ma del tutto elementare. Esiste naturalmente una dimostrazione analoga nel caso dei modelli con densità, che però richiede nozioni di integrazione più delicate (ed una dimostrazione generale che contiene come casi particolari entrambi i modelli, discreto e con densità).

Dimostrazione. (caso Ω numerabile) Cominciamo ad osservare che l'immagine dell'applicazione $T : \Omega \to E$ è un insieme numerabile $\big\{t_1, t_2, \dots\big\}$ e conseguentemente esiste una partizione numerabile $A_1, A_2, \dots$ di Ω, essendo $A_i = \big\{T = t_i\big\}$. È facile rendersi conto che una v.a. V si può scrivere nella forma $V = f \circ T$ se e solo se è *costante* su ogni insieme A_i.

Assegnata dunque U, costruiamo V nel modo seguente: V è costante su ogni insieme A_i dove prende il valore

$$\frac{1}{\mathbf{P}^\theta(A_i)} \int_{A_i} U \,\mathrm{d}\mathbf{P}^\theta \;=\; \frac{\sum_{\omega_j \in A_i} U(\omega_j) h\big(\theta, T(\omega_j)\big) k(\omega_j)}{\sum_{\omega_j \in A_i} h\big(\theta, T(\omega_j)\big) k(\omega_j)} \;=$$

$$= \frac{\sum_{\omega_j \in A_i} U(\omega_j) k(\omega_j)}{\sum_{\omega_j \in A_i} k(\omega_j)}$$

dove l'ultima eguaglianza è dovuta al fatto che $h\big(\theta, T(\omega)\big)$ è costante su ogni insieme A_i. Pertanto il numero $\frac{1}{\mathbf{P}^\theta(A_i)} \int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta$ non dipende da θ.

Sorge una difficoltà, nella definizione precedente, se $\mathbf{P}^\theta(A_i) = 0$. Se A_i è trascurabile per ogni probabilità $\mathbf{P}^\theta$ lo possiamo appunto *trascurare*, se invece è trascurabile solo per qualche valore del parametro θ, per definizione poniamo $\frac{1}{\mathbf{P}^\theta(A_i)} \int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta$ eguale al valore (costante) che si ottiene con i parametri θ per i quali A_i non è $\mathbf{P}^\theta$-trascurabile.

Cominciamo a verificare che, per ogni θ, si ha $\mathbf{E}^\theta[V] = \mathbf{E}^\theta[U]$ (e di conseguenza, se U è corretta, lo è pure V). Infatti

$$\mathbf{E}^\theta[U] = \int U \, \mathrm{d}\mathbf{P}^\theta = \sum_i \int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta = \sum_i \Big(\mathbf{P}^\theta(A_i) \frac{\int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta}{\mathbf{P}^\theta(A_i)} \Big)$$

Ora il numero $\frac{\int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta}{\mathbf{P}^\theta(A_i)}$ (che non dipende da θ) è eguale al valore di V sull'insieme A_i, quindi

$$\frac{\int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta}{\mathbf{P}^\theta(A_i)} = \frac{\int_{A_i} V \, \mathrm{d}\mathbf{P}^\theta}{\mathbf{P}^\theta(A_i)} \ :$$

ripetendo i passaggi precedenti nel verso opposto si ritrova quindi $\mathbf{E}^\theta[V]$, si ha cioè l'eguaglianza voluta.

Proviamo ora che si ha $\mathbf{E}^\theta\big[\big(V - g(\theta)\big)^2\big] \leq \mathbf{E}^\theta\big[\big(U - g(\theta)\big)^2\big]$ e osserviamo che ci si può ridurre al caso in cui $g(\theta) = 0$.

Poichè $\mathbf{E}^\theta\big[V^2\big] = \sum_i \int_{A_i} V^2 \, \mathrm{d}\mathbf{P}^\theta$, è sufficiente provare che, su ogni insieme A_i, si ha

$$\int_{A_i} V^2 \, \mathrm{d}\mathbf{P}^\theta \ \leq \ \int_{A_i} U^2 \, \mathrm{d}\mathbf{P}^\theta$$

e, poichè sull'insieme A_i la variabile aleatoria V assume costantemente il valore $\frac{1}{\mathbf{P}^\theta(A_i)} \int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta$, questo equivale a provare che si ha

$$\Big(\int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta \Big)^2 \leq \mathbf{P}^\theta\big(A_i\big) \Big(\int_{A_i} U^2 \, \mathrm{d}\mathbf{P}^\theta \Big)$$

L'ultima disuguaglianza è una conseguenza della *disuguaglianza di Schwartz*: infatti

$$\Big| \int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta \Big| = \Big| \int_{A_i} 1.U \, \mathrm{d}\mathbf{P}^\theta \Big| \leq$$

$$\leq \sqrt{\int_{A_i} 1 \, \mathrm{d}\mathbf{P}^\theta} \sqrt{\int_{A_i} U^2 \, \mathrm{d}\mathbf{P}^\theta} = \sqrt{\mathbf{P}^\theta(A_i)} \sqrt{\int_{A_i} U^2 \, \mathrm{d}\mathbf{P}^\theta}$$

Ricordiamo che la disuguaglianza di Schwartz è in realtà una eguaglianza se le due funzioni 1 e U sono proporzionali (sull'insieme A_i), cioè se U è costante sull'insieme A_i: di conseguenza si ha, per ogni θ, l'eguaglianza $\mathbf{E}^\theta\big[\big(V - g(\theta)\big)^2\big] = \mathbf{E}^\theta\big[\big(U - g(\theta)\big)^2\big]$ se e solo se U è costante su ogni insieme A_i, cioè se si può scrivere nella forma $f \circ T$. □

Osservazione 5.4.4. La dimostrazione precedente potrebbe essere fatta in una maniera molto più rapida, a patto di possedere qualche ulteriore nozione di misura e integrazione: essenzialmente il fatto che lo spazio delle variabili aleatorie U tali che $\int U^2 \, \mathrm{d}\mathbf{P}^\theta < +\infty$ è uno *spazio di Hilbert* $\mathcal{H}$ (munito del prodotto scalare $\langle U, V\rangle = \int UV \, \mathrm{d}\mathbf{P}^\theta$) e il sottospazio $\mathcal{V}$ delle v.a. costanti su ognuno degli insiemi A_i è un sottospazio *chiuso*. La costruzione che abbiamo fatto (di una variabile V che sull'insieme A_i coincide con $\frac{1}{\mathbf{P}^\theta(A_i)} \int_{A_i} U \, \mathrm{d}\mathbf{P}^\theta$) equivale alla costruzione della *proiezione ortogonale* di U sul sottospazio $\mathcal{V}$.

5.5 Stime di massima verosimiglianza

Diamo un'altra definizione:

Definizione 5.5.1 (Stima di massima verosimiglianza). Sia assegnato un modello statistico $\big(\Omega, \mathcal{F}, (\mathbf{P}^\theta, \theta \in \Theta)\big)$ tale che $\Theta \subset \mathbb{R}$: si dice che U è una stima di massima verosimiglianza del parametro θ se, per ogni $\omega \in \Omega$, si ha

$$L\big(U(\omega), \omega\big) \;=\; \sup_{\theta \in \Theta} L\big(\theta, \omega\big)$$

Di conseguenza il *"sup"* sopra scritto è in realtà un massimo. In verità non è necessario che l'eguaglianza sopra scritta sia verificata *esattamente* per ogni $\omega \in \Omega$, ma è sufficiente che sia soddisfatta *al di fuori di un insieme trascurabile* (si usa dire "per quasi ogni $\omega \in \Omega$").

Usualmente la stima di massima verosimiglianza, se esiste, viene indicata $\widehat{\theta}(\omega)$. Le stime di massima verosimiglianza sono facili da trovare, inoltre questo fornisce un criterio *costruttivo* per trovare una stima; viceversa è più difficile spiegare se e in quale senso una tale stima è una *buona stima*.

In un caso particolare si ha però il risultato seguente, che viene enunciato per ora limitatamente al caso di variabili aleatorie a valori interi positivi.

Teorema 5.5.2. *Sia* $\big(\mathbf{m}^\theta, \theta \in \Theta\big)$ *una famiglia di leggi di probabilità concentrate sugli interi positivi, e supponiamo che* Θ *sia un* intervallo *di* $\mathbb{R}$ *e che, ponendo* $p(\theta, k) = \mathbf{m}^\theta(\{k\})$*, questa si possa scrivere nella forma*

$$p(\theta, k) \;=\; c(\theta) \, \exp\big(\theta \, T(k)\big) \, g(k)$$

dove $T : \mathbb{N} \to \mathbb{R}$. *Consideriamo un campione infinito* $X_1, X_2, \ldots$ *di legge* $\mathbf{m}^\theta$ *e supponiamo che esista, per ogni* n*, la stima di massima verosimiglianza* $\widehat{\theta}_n$ *relativa al campione di taglia* n*: allora la successione di stime* $\big(\widehat{\theta}_n\big)_{n\geq 1}$ *è consistente.*

I modelli nei quali la funzione di probabilità ha la forma data dal Teorema 5.5.2 sono detti *modelli esponenziali*. A volte (come si vedrà anche negli esempi successivi) anziché l'espressione $\exp\big(\theta\, T(k)\big)$ compare un'espressione della forma $\exp\big(\, d(\theta)\, T(k)\big)$ dove l'applicazione $\theta \to d(\theta)$ è iniettiva: è sufficiente naturalmente considerare come nuovo parametro $\tilde{\theta} = d(\theta)$ per riportarsi alla situazione sopra enunciata.

Non riportiamo la dimostrazione del Teorema 5.5.2 , che è del tutto simile a quella dell'analogo risultato per modelli *con densità* che verrà esposta più avanti (per essere più precisi, entrambe le dimostrazioni sono riduzioni a casi particolari di un risultato più generale che in questo primo corso non abbiamo gli strumenti per dimostrare).

Limitiamoci ad osservare che la condizione del Teorema 5.5.2 è soddisfatta in molti esempi: nel caso delle leggi di Poisson si ha ad esempio $p(\theta, k) = \mathrm{e}^{-\theta}\theta^k(k!)^{-1} = \mathrm{e}^{-\theta}\exp\big(k\,\log(\theta)\big)(k!)^{-1}$ (è sufficiente considerare come parametro $\log(\theta)$ anziché θ).

Nel caso delle leggi geometriche si ha $p(\theta, k) = \theta\exp\big((k-1)\log(1-\theta)\big)$.

Esempio 5.5.3. Consideriamo il caso di un campione $(X_1, \ldots, X_n)$ di taglia n e legge Geometrica di parametro θ: sullo spazio $\Omega = (\mathbb{N}^*)^n$ la verosimiglianza è data da

$$L\big(\theta\,; k_1, \ldots, k_n\big) \;=\; \big(1-\theta\big)^{k_1+\cdots+k_n-n}\,\theta^n$$

Un facile calcolo prova che il massimo di questa funzione (al variare di θ) si ottiene nel punto $\frac{n}{k_1+\ldots+k_n}$, e questo identifica la stima di massima verosimiglianza. Ricordando che $X_1, \ldots, X_n$ sono le proiezioni coordinate, possiamo scrivere

$$\widehat{\theta}_n\big(k_1, \ldots, k_n\big) \;=\; \frac{n}{k_1 + \cdots + k_n}$$

oppure, indifferentemente,

$$\widehat{\theta}_n \;=\; \frac{n}{X_1 + \cdots + X_n}$$

mentre non è corretto scrivere $\widehat{\theta}_n = \frac{n}{k_1+\cdots+k_n}$ (in quest'ultimo caso, infatti, avrei a sinistra una variabile aleatoria, cioè una funzione, ed a destra un numero).

Considerando un campione infinito, il Teorema 5.5.2 afferma che la successione di stime $\big(\widehat{\theta}_n\big)_{n\geq 1}$ è consistente.

Vediamo ora l'analogo del Teorema 5.5.2 nel caso di *modelli con densità*, e di questo diamo una dimostrazione completa.

Teorema 5.5.4. *Supponiamo che* Θ *sia un* intervallo *di* $\mathbb{R}$ *e sia assegnata una famiglia di densità* $\big(f(\theta,x)\,,\theta\in\Theta\big)$ *che si possano scrivere nella forma*

$$f(\theta,x)\;=\;c(\theta)\,.\,\exp\big(\theta\,T(x)\big)\,.\,g(x)$$

con una opportuna applicazione $T:\mathbb{R}\to\mathbb{R}$*. Consideriamo un campione infinito* $X_1,X_2,\ldots$ *con densità* $f(\theta,.)$ *e supponiamo che esista, per ogni* n*, la stima di massima verosimiglianza* $\widehat{\theta}_n$ *relativa al campione di taglia* n*: allora la successione di stime* $\big(\widehat{\theta}_n\big)_{n\geq 1}$ *è consistente.*

Ricordiamo che quando le densità verificano la condizione del Teorema 5.5.4, si dice che si ha un *modello esponenziale*: la definizione può essere estesa al caso a dimensione maggiore di 1, supponendo $\Theta\subseteq\mathbb{R}^k$ e che esista una applicazione (boreliana) $T:\mathbb{R}\to\mathbb{R}^k$ in modo che si abbia

$$f(\theta,x)=c(\theta).\exp\big(\langle\theta,T(x)\rangle\big).g(x)$$

dove $\langle .,.\rangle$ è il prodotto scalare in $\mathbb{R}^k$. Con questa definizione più generale il Teorema 5.5.4 rimane vero ed il principio della dimostrazione non cambia, è solo un poco più complicato.

Vediamo ora la dimostrazione del Teorema 5.5.4.

Dimostrazione. Poichè si deve avere $\int f(\theta,x)\,\mathrm{d}x=1$, ne segue che

$$c(\theta)=\bigg[\int\exp\big(\theta\,T(x)\big)\,g(x)\,\mathrm{d}x\bigg]^{-1}=\exp\Big(-\psi(\theta)\Big)$$

essendo $\psi(\theta)\;=\;\log\big(\int\mathrm{e}^{\theta\,T(x)}g(x)\,\mathrm{d}x\big)$. Per calcolare $\psi'(\theta)$ si può derivare sotto il segno di integrale, e si ottiene

$$\psi'(\theta)\;=\;\frac{\int T(x)\,\mathrm{e}^{\theta\,T(x)}g(x)\,\mathrm{d}x}{\int\mathrm{e}^{\theta\,T(x)}g(x)\,\mathrm{d}x}\;=\;\mathbf{E}^{\theta}\Big[T\big(X_i\big)\Big]$$

Con conti analoghi, facili ma un poco più lunghi, si prova l'eguaglianza $\psi''(\theta)\;=\;Var^{\theta}\big(T(X_i)\big)$; poichè necessariamente $Var^{\theta}\big(T(X_i)\big)$ è strettamente positiva (vedi l'osservazione al termine della dimostrazione) ne segue che la funzione $\psi'(\theta)$ è strettamente crescente e quindi invertibile.

La verosimiglianza del campione n-simo assume la forma

$$L_n\big(\theta\,;x_1,\ldots,x_n\big)\;=\;\exp\Big(\theta\sum_{i\leq n}T(x_i)\;-n\,\psi(\theta)\Big)\prod_{i\leq n}g(x_i)$$

e per cercare il punto θ che rende massima questa espressione è sufficiente cercare il punto di massimo della funzione $\theta \longrightarrow \big(\theta \sum_{i\le n} T(x_i) - n\,\psi(\theta)\big)$. Questo si può fare risolvendo l'equazione (detta *equazione di massima verosimiglianza*)

$$\psi'\big(\theta\big)\Big|_{\theta=\widehat{\theta}_n} = \frac{\sum_{i\le n} T\big(X_i\big)}{n}$$

e di conseguenza la stima di massima verosimiglianza (che per ipotesi esiste) è data dall'espressione $\widehat{\theta}_n = \big(\psi'\big)^{-1}\Big(\frac{\sum_{i\le n} T(X_i)}{n}\Big)$.

Fissiamo una probabilità $\mathbf{P}^{\overline{\theta}}$: per la *Legge dei Grandi Numeri* (Teorema 4.1.2) la successione $\sum_{i\le n} \frac{T\left(X_i\right)}{n}$ converge in probabilità a $\mathbf{E}^{\overline{\theta}}\big[T(X_1)\big] = \psi'\big(\overline{\theta}\big)$ e quindi (poiché $\big(\psi'\big)^{-1}$ è una funzione continua) per la Proposizione 4.1.5, $\widehat{\theta}_n$ converge in probabilità a $\big(\psi'\big)^{-1}\big(\psi'(\overline{\theta})\big) = \overline{\theta}$. □

Osservazione 5.5.5. Vediamo perché (come è stato affermato nel corso della dimostrazione) necessariamente $Var^{\theta}\big(T(X_i)\big) > 0$: ricordo che solo le costanti hanno varianza 0, e se $T(x)$ fosse costante (quasi ovunque) la densità $f(\theta, x)$ sarebbe proporzionale alla funzione $g(x)$ e in definitiva queste densità sarebbero tutte eguali tra loro e questo contraddice l'ipotesi che a due parametri θ_1 e θ_2 diversi corrispondono due probabilità $\mathbf{P}^{\theta_1}$ e $\mathbf{P}^{\theta_2}$ diverse. Appare chiaro quindi che non si può avere $Var^{\theta}\big(T(X_i)\big) = 0$ per ogni parametro θ, ma si potrebbe obiettare che potrebbe essere eguale a 0 magari per un solo $\theta \in \Theta$.

In realtà non è così : la variabile $T(X_i)$ o è una costante *per ogni probabilità* $\mathbf{P}^{\theta}$ o non lo è per nessuna (e quindi $\psi''(\theta)$ o è sempre 0 oppure è sempre strettamente positivo). Infatti le probabilità definite dalle densità $f(\theta, x)$ ammettono gli stessi insiemi trascurabili (nel linguaggio della teoria della misura sono *equivalenti*), e ricordiamo che la densità $f(\theta, x)$ è la densità della variabile X_i sotto $\mathbf{P}^{\theta}$. Ricordando che una funzione a valori positivi ha integrale 0 se e solo se è nulla fuori di un insieme trascurabile, e poiché $\exp\big(\theta\, T(x)\big)$ è sempre strettamente positivo, un boreliano A è trascurabile per la densità $f(\theta, x)$ se e solo se $g(x)$ è nulla quasi ovunque sull'insieme A (rispetto alla misura di Lebesgue): questa condizione dunque non dipende dal parametro θ.

Osservazione 5.5.6. Nel Teorema precedente, abbiamo messo *per ipotesi* che esista la stima di massima verosimiglianza $\widehat{\theta}_n$: infatti siamo tentati di scrivere direttamente $\widehat{\theta}_n = \big(\psi'\big)^{-1}\Big(\frac{\sum_{i\le n} T\left(X_i\right)}{n}\Big)$, ma senza quella ipotesi non possiamo farlo perchè non siamo sicuri che, per ogni $\omega = (x_1, \ldots, x_n) \in \Omega$, $\frac{\sum_{i\le n} T\left(X_i(\omega)\right)}{n}$ sia un elemento di $\psi'\Big(\Theta\Big)$.

5.6 Intervalli di fiducia

Supponiamo assegnato un modello statistico, ed un numero α con $0 < \alpha < 1$; usualmente α è un numero vicino a 0, ed i valori tipici sono 0,1 ; 0,05 e 0,01.

Definizione 5.6.1 (Regione di Fiducia). Sia assegnato, per ogni $\omega \in \Omega$, un sottoinsieme dei parametri $C(\omega) \subset \Theta$: si dice che $C(\omega)$ è una *regione di fiducia* per il parametro θ *al livello* $(1-\alpha)$ se, qualunque sia θ, si ha

$$\mathbf{P}^\theta\big\{\omega\big|\, \theta \in C(\omega)\big\} \geq 1-\alpha$$

o (ciò che è lo stesso) $\mathbf{P}^\theta\big\{\omega\big|\, \theta \notin C(\omega)\big\} \leq \alpha$.

Se $\Theta \subseteq \mathbb{R}$ e $C(\omega)$ è un intervallo, si parla di *intervallo di fiducia*. Alcuni testi usano il termine *intervallo di confidenza*, ma è una cattiva traduzione dall'inglese: infatti la parola *confidence* vuole dire appunto fiducia (e non confidenza).

Naturalmente si ha interesse a individuare una regione di fiducia *più piccola possibile*, a patto che sia soddisfatta la condizione sul livello.

Non esistono veri risultati teorici per quanto riguarda le regioni di fiducia, esiste però un legame tra intervalli di fiducia e test statistici che esamineremo nel paragrafo successivo; vediamo piuttosto *alcuni esempi concreti*.

Esempio 5.6.2 (Intervallo di fiducia per il controllo di qualità). Consideriamo un campione $X_1, \ldots, X_n$ di legge di Bernoulli di parametro θ e vogliamo individuare un intervallo di fiducia per il parametro θ: partiamo dal fatto che $\overline{X} = \frac{X_1+\cdots+X_n}{n}$ è una stima corretta di θ e che $Var^\theta\big(\overline{X}\big) = \frac{\theta(1-\theta)}{n}$.

Ci aspettiamo un intervallo di fiducia per θ intorno alla sua stima, più precisamente della forma $I = \big[\overline{X}(\omega) - d\,,\, \overline{X}(\omega) + d\big]$ (con d da determinare).

Per determinare d (ricordiamo che abbiamo interesse che sia *più piccolo possibile*) partiamo dal fatto che si ha

$$\Big\{\theta \notin \big[\overline{X} - d\,,\, \overline{X} + d\big]\Big\} = \Big\{\big|\overline{X} - \theta\big| > d\Big\}$$

Dalla diseguaglianza di Chebishev di ottiene la maggiorazione $\mathbf{P}^\theta\big\{\big|\overline{X} - \theta\big| > d\big\} \leq \frac{\theta(1-\theta)}{nd^2}$; abbiamo bisogno di una maggiorazione indipendente da θ e poichè $\max_{0<\theta<1} \theta(1-\theta) = \frac{1}{4}$, si ottiene $\mathbf{P}^\theta\big\{\big|\overline{X} - \theta\big| > d\big\} \leq \alpha$ ponendo $d = \frac{1}{\sqrt{4n\alpha}}$, e di conseguenza

$$\mathbf{P}^\theta\Big\{ -\frac{1}{\sqrt{4n\alpha}} \leq \overline{X} - \theta \leq +\frac{1}{\sqrt{4n\alpha}}\Big\} \geq 1-\alpha$$

Si ottiene l'intervallo di fiducia $\big[\overline{X}(\omega) - \frac{1}{\sqrt{4n\alpha}}, \overline{X}(\omega) + \frac{1}{\sqrt{4n\alpha}}\big]$, o (come si scrive più sinteticamente) $\overline{X}(\omega) \pm \frac{1}{\sqrt{4n\alpha}}$.

L'intervallo di fiducia che abbiamo determinato sopra in realtà non è molto buono (cioè non è molto *stretto*) perché è basato sulla diseguaglianza di Chebishev, che in genere fa *perdere qualcosa* rispetto ai calcoli precisi; tuttavia quando n è grande i calcoli esatti sulla variabile $B(n,\theta)$ non sono praticabili. In questo caso però si può utilizzare il *Teorema Limite di De Moivre-Laplace*.

Esempio 5.6.3 (**Intervallo di fiducia approssimato mediante il teorema di De Moivre-Laplace).** Siamo nella stessa situazione dell'esercizio precedente, ma questa volta utilizziamo il fatto che

$$\mathbf{P}^\theta\Big\{\frac{X_1+\cdots+X_n-n\theta}{\sqrt{\theta(1-\theta)n}}\le x\Big\} = \mathbf{P}^\theta\Big\{\sqrt{n}\frac{\overline{X}-\theta}{\sqrt{\theta(1-\theta)}}\le x\Big\}\approx\Phi(x)$$

Il nostro scopo è trovare un numero d tale che valga la maggiorazione $\mathbf{P}^\theta\Big\{\sqrt{n}\dfrac{|\overline{X}-\theta|}{\sqrt{\theta(1-\theta)}}>d\Big\}\le\alpha\,.$

Ricordiamo che abbiamo usato la notazione q_β (dato $0<\beta<1$) per indicare il β-quantile della legge N(0,1) (vedi definizione 3.6.1), cioè il numero q_β tale che $\Phi(q_\beta)=\beta$: naturalmente questo numero non può essere calcolato esattamente, ma si può ricavare usando le tavole della funzione $\Phi(.)$.

Dunque $\mathbf{P}^\theta\big\{\sqrt{n}\frac{|\overline{X}-\theta|}{\sqrt{\theta(1-\theta)}}>q_{1-\frac{\alpha}{2}}\big\}\approx\alpha$: con passaggi analoghi a quelli fatti sopra, si ottiene l'intervallo di fiducia $\overline{X}(\omega)\pm\frac{q_{1-\frac{\alpha}{2}}}{2\sqrt{n}}$.

È interessante notare quanto l'intervallo così ottenuto si è ristretto rispetto al precedente: tenendo fisso n, sopra c'era un termine dell'ordine di $\frac{1}{\sqrt{\alpha}}$ (non dimentichiamo che α è un numero piccolo), mentre ora compare il numero $q_{1-\frac{\alpha}{2}}$ che è di solito vicino a 3.

Se noi consideriamo ad esempio $\alpha=0{,}01$, dalle tavole si ricava il valore approssimato $q_{0{,}995}=2{,}58$; gli intervalli di fiducia sono col primo metodo $\overline{X}(\omega)\pm\frac{5}{\sqrt{n}}$ e nel secondo caso $\overline{X}(\omega)\pm\frac{1{,}29}{\sqrt{n}}$.

Osservazione 5.6.4 (**Il metodo della quantità pivot).**

Si parla di *metodo della quantità pivot* quando si individua una funzione di una v.a. X e del parametro θ che sia

- invertibile rispetto al parametro θ;
- tale che la sua legge di probabilità non dipenda dal parametro θ.

Nei due esempi precedenti non abbiamo in realtà individuato una quantità pivot ma qualcosa di meno: nell'esempio 5.6.2 la variabile $\big(\overline{X}-\theta\big)$ non ha

legge indipendente da θ ma ha media 0 (indipendentemente dal parametro) ed una varianza che abbiamo potuto maggiorare uniformemente rispetto al parametro.

Useremo veramente il metodo della quantità pivot nell'ultimo capitolo.

5.7 Teoria dei test statistici

Il primo passo da compiere, di fronte a un test statistico, è *formulare un'ipotesi*: questo si ottiene effettuando una partizione dell'insieme Θ dei parametri in due sottinsiemi non vuoti Θ_0 e Θ_1 corrispondenti rispettivamente ai parametri dell'ipotesi e a quelli della sua negazione, detta *alternativa*.

Torniamo all'esempio del *controllo di qualità*, e consideriamo l'ipotesi *"la percentuale di pezzi difettosi non supera il 5%"*: in questo caso l'insieme dei parametri è $\Theta = \left]0\ ,\ 1\right[$, si ha $\Theta_0 = \left]0\ ,\ 0{,}05\right]$ e $\Theta_1 = \left]0{,}05\ ,\ 1\right[$.

L'ipotesi e l'alternativa sono indicate rispettivamente $\mathcal{H}_0\big)$ e $\mathcal{H}_1\big)$ e si usa dire, ad esempio nel caso precedente:

- consideriamo un test dell'ipotesi $\mathcal{H}_0\big)\ \theta \leq 0,05$ contro l'alternativa $\mathcal{H}_1\big)\ \theta > 0,05$.

Osserviamo che in linea di principio indicare l'alternativa è superfluo, in quanto Θ_1 è individuato dal fatto di essere il complementare di Θ_0 ; tuttavia nei fatti spesso è più chiaro indicare sia l'ipotesi che l'alternativa.

Il secondo passo è *pianificare un esperimento*, cioè stabilire una regola che, secondo il risultato dell'esperienza ω, permetta di decidere se accettare o rifiutare l'ipotesi. Questo equivale a scegliere un *evento* $D \in \mathcal{F}$ che consiste nell'insieme dei risultati ω che portano a rifiutare l'ipotesi: tale insieme D viene chiamato *regione di rifiuto* o più frequentemente *regione critica*.

Per capirci meglio, nell'esempio precedente, l'intuizione ci porta a rifiutare l'ipotesi se la percentuale di pezzi difettosi supera un certo numero a (da determinare secondo regole che vedremo): la regione critica sarà pertanto in questo caso

$$D = \Big\{ \omega \in \Omega \quad \Big| \quad \overline{X}(\omega) > a \Big\}$$

e diremo più sbrigativamente *"il test di regione critica* $D = \big\{\, \overline{X} > a \,\big\}$*"*.

Definizione 5.7.1 (**Livello e potenza).** Si chiama *taglia* di un test di regione critica D il numero

$$\sup_{\theta \in \Theta_0} \mathbf{P}^\theta\big(D\big)$$

Si dice che il test è di *livello* α se la sua taglia è minore o eguale ad α.

Si chiama *potenza* del test la funzione $\pi_D : \Theta_1 \to [0,1]$ definita da $\theta \longrightarrow \mathbf{P}^\theta\big(D\big)$.

Diremo che il test di regione critica D è *più potente* del test di regione critica D^* se, per ogni $\theta \in \Theta_1$, si ha $\mathbf{P}^\theta(D) \geq \mathbf{P}^\theta(D^*)$.

Scegliere un livello equivale a porre un confine superiore alle *probabilità dell'errore di prima specie* (cioè ai numeri $\mathbf{P}^\theta(D)$ per $\theta \in \Theta_0$) ; intuitivamente infatti errore di prima specie significa *"rifiutare l'ipotesi quando è vera"*). Invece la potenza è in un certo senso la *"capacità di accorgersi che l'ipotesi è falsa"* (ed errore di seconda specie è *"accettare l'ipotesi quando è falsa*).

Usualmente si procede in questo modo: si fissa un livello α (i valori tipici sono 0,1 ; 0,05 oppure 0,01) che fissi un limite superiore per l'errore di prima specie, e tra i test di livello α si cerca di ottenere la massima potenza possibile (cioè una regione critica *più grande possibile*).

Quando Θ_0 è ridotto a un solo punto (cioè $\Theta_0 = \{\theta_0\}$) si dice che *l'ipotesi è semplice*; perfettamente analoga naturalmente è la definizione di *alternativa semplice*. Come vediamo qua sotto, la ricerca della regione critica di un test a ipotesi semplice può essere ricondotta alla ricerca delle regioni di fiducia, e viceversa.

Osservazione 5.7.2 (**Legame tra test e regioni di fiducia).** Supponiamo di aver trovato, per ogni $\omega \in \Omega$, una *regione di fiducia* $C(\omega)$ al livello $(1-\alpha)$ e consideriamo il test dell'ipotesi $\mathcal{H}_0\big)\, \theta = \theta_0$ contro l'alternativa $\mathcal{H}_1\big)\, \theta \neq \theta_0$. Rifiutiamo l'ipotesi se $\theta_0 \notin C(\omega)$, consideriamo cioè come regione critica $D = \big\{\omega \,|\, \theta_0 \notin C(\omega)\big\}$: dalla definizione di regione critica segue che $\mathbf{P}^{\theta_0}\big(D\big) \leq \alpha$, cioè abbiamo ottenuto un test di livello α.

Quanto è stato fatto si può considerare nel senso inverso: cioè se per ogni $\bar\theta$ abbiamo la regione critica $D(\bar\theta)$ di livello α del test dell'ipotesi $\mathcal{H}_0\big)\, \theta = \bar\theta$, ponendo $C(\omega) = \big\{\theta \in \Theta \,\big|\, \omega \notin D(\theta)\big\}$, otteniamo una regione di fiducia al livello $(1-\alpha)$.

Esempio 5.7.3. Dato un campione $X_1, \ldots, X_n$ con *legge di Bernoulli*, pianifichiamo il test dell'ipotesi semplice $\mathcal{H}_0\big)\, \theta = \theta_0$ contro $\mathcal{H}_1\big)\, \theta \neq \theta_0$ al livello α.

Osservando che l'intervallo di fiducia si può equivalentemente scrivere nella forma $C(\omega) = \big\{\theta \,\big|\, -d \leq \overline{X}(\omega) - \theta \leq d\big\}$, si ottiene la regione critica della forma $D = \big\{\omega \,\big|\, \big|\overline{X}(\omega) - \theta_0\big| > d\big\}$, con un opportuno numero d da calcolare (questa forma della regione critica del resto si accorda con quello che suggerisce l'intuizione).

Per ottenere la regione critica *più grande possibile*, scegliamo il minimo d per il quale valga la maggiorazione

$$\mathbf{P}^{\theta_0}\big\{\, \big|\overline{X} - \theta_0\big| > d \,\big\} \leq \alpha$$

Utilizzando la diseguaglianza di Chebishev, si ottiene (omettiamo i facili conti, sostanzialmente identici a quelli svolti nel paragrafo precedente) per d il valore $\sqrt{\frac{\theta_0(1-\theta_0)}{n\alpha}}$.

Un valore più piccolo per il numero d si può ottenere utilizzando l'approssimazione suggerita dal Teorema di De Moivre-Laplace, cioè

$$\mathbf{P}^{\theta_0}\big\{\,|\overline{X}-\theta_0|>d\,\big\}=\mathbf{P}^{\theta_0}\Big\{\sqrt{n}\frac{|\overline{X}-\theta_0|}{\sqrt{\theta_0(1-\theta_0)}}>\frac{d\sqrt{n}}{\sqrt{\theta_0(1-\theta_0)}}\Big\}\approx$$

$$\approx 2\Big(1-\Phi\Big(\frac{d\sqrt{n}}{\sqrt{\theta_0(1-\theta_0)}}\Big)\Big)$$

Si ottiene in questo modo il valore $d=q_{1-\frac{\alpha}{2}}\sqrt{\frac{\theta_0(1-\theta_0)}{n}}$.

Nella stessa situazione del campione con legge di Bernoulli, cerchiamo di esaminare il test $\mathcal{H}_0\big)\,\theta\le\theta_0$ contro l'alternativa $\mathcal{H}_1\big)\,\theta>\theta_0$: facciamoci prima guidare dall'intuizione e poi arriveremo a dei risultati più precisi.

Ci aspettiamo una regione critica della forma $\big\{\,\overline{X}\ge d\,\big\}$ con un opportuno numero d da calcolare in funzione del livello scelto , ma sorgono delle difficoltà: cerchiamo *il più piccolo* numero d tale che valga la diseguaglianza seguente

$$\sup_{\theta\le\theta_0}\mathbf{P}^{\theta}\big\{\;\overline{X}\ge d\,\big\}\le\alpha$$

dove α è il livello scelto (cerchiamo il valore d più piccolo per avere la regione critica più grande possibile). Ci aspettiamo che la funzione $\theta\longrightarrow\mathbf{P}^{\theta}\big\{\overline{X}\ge d\big\}$ sia crescente (e questo semplificherebbe i conti) ma il calcolo diretto non è immediato: ci vengono però in aiuto dei risultati generali che ora esponiamo.

Nei risultati che seguono diamo per scontato che il modello sia dotato di verosimiglianza (e quindi, sulla base di quanto abbiamo svolto in questo corso, che il modello sia con spazio numerabile oppure con densità).

Lemma 5.7.4 (**Lemma di Neyman-Pearson).** *Supponiamo assegnato un modello statistico nel quale l'insieme* Θ *dei parametri è ridotto a due punti* ($\Theta=\big\{\theta_0,\theta_1\big\}$) *e sia dato il test dell'ipotesi* $\mathcal{H}_0\big)\,\theta=\theta_0$ *contro* $\mathcal{H}_1\big)\,\theta=\theta_1$. *Consideriamo l'insieme* D *così definito*

$$D\;=\;\big\{\omega\in\Omega\,\big|\,L(\theta_0,\omega)\le c\,L(\theta_1,\omega)\,\big\}$$

dove c *è una costante positiva. Allora*

1. *D è la regione critica di un test più potente di ogni altro test di livello $\mathbf{P}^{\theta_0}\big(D\big)$;*

2. *vale la diseguaglianza* $\mathbf{P}^{\theta_1}\big(D\big) \geq \mathbf{P}^{\theta_0}\big(D\big)$.

Dimostrazione. Consideriamo una generica funzione $\varphi : \Omega \to [0,1]$ e notiamo che per ogni $\omega \in \Omega$ vale la diseguaglianza

$$\Big(I_D(\omega) - \varphi(\omega)\Big)\,\Big(L(\theta_0,\omega) - c\,L(\theta_1,\omega)\Big) \leq 0$$

Infatti, se $\omega \in D$, $\big(I_D(\omega) - \varphi(\omega)\big) \geq 0$ e $\big(L(\theta_0,\omega) - c\,L(\theta_1,\omega)\big) \leq 0$ e dunque il prodotto è negativo; analoga è la verifica se $\omega \notin D$.

A questo punto la dimostrazione si diversifica nel caso di un modello discreto o di uno con densità: se Ω è numerabile, sommando su tutti i punti $\omega \in \Omega$ si ottiene

$$\mathbf{P}^{\theta_0}\big(D\big) - \int \varphi(\omega)\,\mathrm{d}\mathbf{P}^{\theta_0}(\omega) \leq c\Big(\mathbf{P}^{\theta_1}\big(D\big) - \int \varphi(\omega)\,\mathrm{d}\mathbf{P}^{\theta_1}(\omega)\Big)$$

Nel caso di un modello con densità, conviene indicare $\omega = (x_1,\dots,x_n)$ e scrivere la diseguaglianza nella forma

$$\Big(I_D(x_1,\dots,x_n) - \varphi(x_1,\dots,x_n)\Big)\,\Big(L(\theta_0\,;x_1,\dots,x_n) - c\,L(\theta_1\,;x_1,\dots,x_n)\Big) \leq 0$$

Integrando rispetto alla misura di Lebesgue, si ottiene

$$\mathbf{P}^{\theta_0}\big(D\big) - \int \varphi\ \mathrm{d}\mathbf{P}^{\theta_0}\ \leq c\Big(\mathbf{P}^{\theta_1}\big(D\big) - \int \varphi\ \ \mathrm{d}\mathbf{P}^{\theta_1}\Big)$$

Abbiamo quindi ottenuto lo stesso risultato e la dimostrazione prosegue identica in entrambi i casi: se D^* è la regione critica di un altro test, prendendo come funzione $\varphi = I_{D^*}$, si ottiene

$$\mathbf{P}^{\theta_0}\big(D\big) - \mathbf{P}^{\theta_0}\big(D^*\big) \leq c\Big(\mathbf{P}^{\theta_1}\big(D\big) - \mathbf{P}^{\theta_1}\big(D^*\big)\Big)$$

Se dunque D^* ha livello $\mathbf{P}^{\theta_0}\big(D\big)$ (cioè se $\mathbf{P}^{\theta_0}\big(D^*\big) \leq \mathbf{P}^{\theta_0}\big(D\big)$), ne segue che vale anche la diseguaglianza $\mathbf{P}^{\theta_1}\big(D^*\big) \leq \mathbf{P}^{\theta_1}\big(D\big)$ (cioè D è *più potente* di D^*).

Considerando poi come funzione φ la costante $\mathbf{P}^{\theta_0}\big(D\big)$, si ottiene $\mathbf{P}^{\theta_1}\big(D\big) - \mathbf{P}^{\theta_0}\big(D\big) \geq 0$, cioè il punto 2). □

Il lemma di Neyman-Pearson permette di identificare con precisione i *buoni test* nel caso in realtà poco significativo di un modello statistico nel quale i parametri siano solo due: il suo vero interesse consiste nel fatto che si può estendere a casi più generali, i cosiddetti *test unilateri*. Quando l'insieme dei parametri Θ è un intervallo di $\mathbb{R}$ (intervallo in senso lato, cioè anche una semiretta o tutta la retta) si parla di *test unilatero* se l'ipotesi è della forma $\mathcal{H}_0\big)\ \theta \leq \theta_0$ o della forma $\mathcal{H}_0\big)\ \theta \geq \theta_0$. Premettiamo una definizione.

Definizione 5.7.5 (**Rapporto di verosimiglianza crescente).** Supponiamo assegnato un modello statistico nel quale l'insieme dei parametri Θ è un intervallo di $\mathbb{R}$ e sia T una variabile aleatoria reale definita su Ω: si dice che il modello è *a rapporto di verosimiglianza crescente* rispetto a T se, scelti comunque $\theta_1 < \theta_2$, esiste una funzione reale (strettamente) crescente a valori positivi f_{θ_1,θ_2} tale che valga l'eguaglianza

$$\frac{L(\theta_2,\omega)}{L(\theta_1,\omega)} \;=\; f_{\theta_1,\theta_2}\big(T(\omega)\big)$$

Naturalmente quella definizione ha senso se le verosimiglianze sono sempre strettamente positive (o al più se si annullano tutte *sul medesimo sottinsieme* di Ω).

Teorema 5.7.6 (**Test unilatero).** *Supponiamo che il modello sia a rapporto di verosimiglianza crescente rispetto a T e consideriamo il test unilatero $\mathcal{H}_0\big)\,\theta \le \theta_0$ contro l'alternativa $\mathcal{H}_1\big)\,\theta > \theta_0$; consideriamo poi l'insieme $D = \big\{\omega \,\big|\, T(\omega) \ge d\big\}$ dove d è un opportuno numero. Il test di regione critica D è tale che:*

1. *vale l'eguaglianza* $\sup_{\theta\le\theta_0} \mathbf{P}^\theta\big(D\big) = \mathbf{P}^{\theta_0}\big(D\big)$;
2. *D è più potente di qualsiasi altro test D^* con* livello $\mathbf{P}^{\theta_0}\big(D\big)$.

Dimostrazione. Chiamiamo $c = f_{\theta_1,\theta_2}(d)$ (quindi c è un numero positivo): valgono le seguenti implicazioni

$$T(\omega) \ge d \iff f_{\theta_1,\theta_2}\big(T(\omega)\big) \ge c \iff L(\theta_2,\omega) \ge c\, L(\theta_1,\omega)$$

e da qui si ottiene $L(\theta_1,\omega) \le \frac{1}{c} L(\theta_2,\omega)$. A questo punto si può applicare il Lemma 5.7.4 e si trova (come conseguenza del punto 2)) $\mathbf{P}^{\theta_2}\big(D\big) \ge \mathbf{P}^{\theta_1}\big(D\big)$: poiché questo vale per ogni scelta di $\theta_1 < \theta_2$, ne segue che la funzione $\theta \to \mathbf{P}^\theta\big(D\big)$ è crescente e pertanto si ottiene la prova del punto 1) (tra l'altro questo semplifica notevolmente il calcolo della *taglia* del test, che risulta eguale a $\mathbf{P}^{\theta_0}\big(D\big)$).

Supponiamo inoltre che D^* abbia livello $\mathbf{P}^{\theta_0}\big(D\big)$, cioè che si abbia $\sup_{\theta\le\theta_0} \mathbf{P}^\theta\big(D^*\big) \le \mathbf{P}^{\theta_0}\big(D\big)$: prendendo un parametro $\theta > \theta_0$ si ha $\mathbf{P}^\theta\big(D^*\big) \le \mathbf{P}^\theta\big(D\big)$ (si applica di nuovo il Lemma 5.7.4, considerando θ al posto di θ_1).

Poichè questo vale per ogni $\theta > \theta_0$, ne segue che D è più potente di D^*. □

Osservazione 5.7.7. Naturalmente se l'ipotesi è della forma $\mathcal{H}_0\big)\,\theta \ge \theta_0$ (oppure se il modello è a rapporto di verosimiglianza *decrescente* rispetto a T) si "ribalta" la regione critica, più precisamente si sceglie della forma $D = \big\{T \le d\big\}$.

Esempio 5.7.8 (**Test unilatero per il controllo di qualità).** Riprendiamo l'esempio che abbiamo interrotto prima dell'enunciato del Lemma di Neyman-Pearson (test unilatero su un campione di Bernoulli): sullo spazio $\Omega = \{0,1\}^n$, il rapporto delle verosimiglianze è dato da

$$\frac{L(\theta_2; k_1, \dots, k_n)}{L(\theta_1; k_1, \dots, k_n)} = \Big(\frac{\theta_2}{\theta_1}\Big)^{k_1+\dots+k_n} \Big(\frac{1-\theta_2}{1-\theta_1}\Big)^{n-(k_1+\dots+k_n)}$$

e si verifica facilmente che è a rapporto di verosimiglianza crescente rispetto a $\overline{X}$. Si ha così una prova di quello che l'intuizione aveva suggerito, cioè che per il test unilatero $\mathcal{H}_0\big)\, \theta \leq \theta_0$ le *buone* regioni critiche siano della forma $\big\{\overline{X} \geq d\big\}$.

In funzione del livello α scelto, d deve essere il più piccolo numero tale che $\mathbf{P}^{\theta_0}\big\{\overline{X} \geq d\big\} \leq \alpha$ (questo per avere la regione critica più grande possibile): ancora una volta viene in aiuto l'approssimazione offerta dal Teorema di De Moivre-Laplace (purchè la numerosità n sia abbastanza grande). Si ha così

$$\mathbf{P}^{\theta_0}\big\{\overline{X} \geq d\big\} = \mathbf{P}^{\theta_0}\Big\{\sqrt{n}\frac{\overline{X}-\theta_0}{\sqrt{\theta_0(1-\theta_0)}} \geq \sqrt{n}\frac{d-\theta_0}{\sqrt{\theta_0(1-\theta_0)}}\Big\} \approx$$

$$\approx 1 - \Phi\Big(\sqrt{n}\frac{d-\theta_0}{\sqrt{\theta_0(1-\theta_0)}}\Big) = \alpha$$

Si prende allora $\sqrt{n}\dfrac{d-\theta_0}{\sqrt{\theta_0(1-\theta_0)}} = q_{1-\alpha}$ (si noti che $q_{1-\alpha}$ è un numero positivo perché α è tipicamente "piccolo", inferiore a $\frac{1}{2}$).

Si ottiene pertanto il valore $d = \theta_0 + \dfrac{q_{1-\alpha}\sqrt{\theta_0(1-\theta_0)}}{\sqrt{n}}$.

Osservazione 5.7.9 (**Soglia di accettazione).** Quando si pianifica un test statistico, per prima cosa si sceglie un *livello* α (solitamente vicino a 0) e in seguito si sceglie una *regione critica* D che abbia livello α.

Si deve cioè avere $\sup_{\theta\in\Theta_0} \mathbf{P}^{\theta}(D) \leq \alpha$: dunque più il livello diminuisce, più la regione critica tende ad essere piccola. Spesso ci si trova in questa situazione: per ogni numero $0 < \alpha < 1$, è assegnata una regione critica D_α di livello α in modo tale che, se $\alpha_1 \leq \alpha_2$, allora $D_{\alpha_1} \subseteq D_{\alpha_2}$. Inoltre $\cup_{0<\alpha<1} D_\alpha = \Omega$ e $\cap_{0<\alpha<1} D_\alpha = \emptyset$.

Allora, per ogni $\bar{\omega} \in \Omega$ (cioè per ogni *risultato dell'indagine statistica*) è assegnato un numero $\bar{\alpha}$ tale che, se $\alpha < \bar{\alpha}$, $\bar{\omega} \notin D_\alpha$ e se $\alpha > \bar{\alpha}$, $\bar{\omega} \in D_\alpha$. Tale numero $\bar{\alpha}$ sarà chiamato *soglia di accettazione.*

5.8 Due esempi di modelli con densità

Esempio 5.8.1 (Campione di legge esponenziale). Sia dato un campione $X_1, \dots, X_n$ con densità *esponenziale* di parametro $\theta, \theta > 0$.

Si considera $\Omega = (\mathbb{R}^+)^n$ e $L\big(\theta\,; x_1, \dots, x_n\big) = \theta^n\, \mathrm{e}^{-\theta(\sum x_i)}$.

La variabile $T = \sum_{i=1}^n X_i$ è un riassunto esaustivo. La ricerca della stima di massima verosimiglianza (per il campione di taglia n) porta a $\widehat{\theta}_n = \frac{n}{\sum_i X_i}$, ed in base al Teorema 5.5.4 la successione di stime $\widehat{\theta}_n$ è consistente.

Ci possiamo domandare se la stima $\widehat{\theta}_n$ è corretta: per effettuare tale calcolo ricordiamo che (sotto $\mathbf{P}^\theta$) , $\sum_{i\le n} X_i \sim \Gamma(n, \theta)$). Di conseguenza

$$\mathbf{E}^\theta\big[\widehat{\theta}_n\big] \;=\; \frac{n}{(n-1)!}\int_0^{+\infty} \theta^n\, x^{n-2}\mathrm{e}^{-\theta x}\,\mathrm{d}x = \frac{\theta\, n}{n-1}$$

Vogliamo esaminare ora un test *unilatero* dell'ipotesi $\mathcal{H}_0\big)\,\theta \le 1$ contro $\mathcal{H}_1\big)\,\theta > 1$ al livello α: notiamo che

$$\frac{L(\theta_2)}{L(\theta_1)} \;=\; \left(\frac{\theta_2}{\theta_1}\right)^n \mathrm{e}^{-(\theta_2-\theta_1)T}$$

cioè il modello è a *rapporto di verosimiglianza decrescente* rispetto a T.

Di conseguenza, conosciamo la forma della *buona* regione critica: deve essere $D = \big\{\sum_{i\le n} X_i \le c\big\}$ con c tale che $\mathbf{P}^1\big\{\sum_{i\le n} X_i \le c\big\} \le \alpha$, cioè $\mathbf{P}^1\big\{\sum_{i\le n} X_i > c\big\} \ge (1-\alpha)$. Per poter avere una regione critica più grande possibile (allo scopo di aumentare la *potenza* del test) imponiamo che la diseguaglianza appena scritta sia un'eguaglianza.

Si deve avere

$$(1-\alpha) = \frac{1}{(n-1)!}\int_c^{+\infty} x^{n-1}\,\mathrm{e}^{-x}\,\mathrm{d}x = \mathrm{e}^{-c}\Big[\frac{c^{n-1}}{(n-1)!} + \frac{c^{n-2}}{(n-2)!} + \cdots + c + 1\Big]$$

È evidente che, dato α, esiste uno ed un solo c positivo che soddisfa l'equazione sopra scritta, ma il calcolo esplicito deve essere fatto con approssimazioni numeriche.

Consideriamo il test dell'ipotesi $\mathcal{H}_0\big)\theta = 2$ contro l'alternativa $\mathcal{H}_1\big)\theta \ne 2$: partiamo dal fatto che, sotto $\mathbf{P}^2$, ogni variabile X_i ha valore atteso $1/2$ e varianza $1/4$.

Questo suggerisce una regione critica della forma $D = \big\{\big|\frac{\sum_i X_i}{n} - \frac{1}{2}\big| \ge c\big\}$ con $\mathbf{P}^2\big\{\big|\frac{\sum_i X_i}{n} - \frac{1}{2}\big| \ge c\big\} \le \alpha$. Il calcolo della probabilità sopra scritta può

essere fatto, con passaggi simili a quelli sopra indicati, ma i conti espliciti diventano complicati.

Possiamo allora accontentarci di una maggiorazione ottenuta con la diseguaglianza di Chebishev:

$$\mathbf{P}^2\left\{\left|\frac{\sum_i X_i}{n}-\frac{1}{2}\right|\geq c\right\}\leq\frac{Var^2\Big(\frac{\sum_i X_i}{n}\Big)}{c^2}=\frac{Var^2\big(X_i\big)}{n\,c^2}=\frac{1}{4nc^2}$$

Prendendo $c=\big(4n\alpha\big)^{-1/2}$ si ottiene la diseguaglianza voluta.

Esempio 5.8.2. Consideriamo la famiglia di densità (per $\theta>-1$)

$$f(\theta,x)\;=\;\begin{cases}(\theta+1)\,x^\theta & 0<x<1\\ 0 & \text{altrimenti}\end{cases}$$

e sia dato un campione di taglia n e densità $f(\theta,.)$.

Poichè la densità può essere scritta nella forma

$$f(\theta,x)=\big(\theta+1\big)\exp\big(\theta\,\log x\big)I_{]0,1[}(x)\,,$$

siamo in presenza di un *modello esponenziale*.

Considerando $\Omega=]0,1[^n$ e $\Theta=]-1,+\infty[$, si ottiene per la verosimiglianza l'espressione

$$L\big(\theta\,;x_1,\ldots,x_n\big)\;=\;\big(\theta+1\big)^n\Big(\prod_{i=1}^n x_i\Big)^\theta$$

e di conseguenza $T=\prod_i X_i$ è un riassunto esaustivo.

Il calcolo della stima di massima verosimiglianza (per il campione di taglia n) porta a

$$\widehat{\theta}_n\;=\;-1-\frac{n}{\sum_{i\leq n}\log\,X_i}$$

e la successione di stime $(\widehat{\theta}_n)_{n\geq1}$ è consistente.

Esaminiamo ora il test unilatero della forma $\mathcal{H}_0\big)\;\theta\geq0$ contro $\mathcal{H}_1\big)\;\theta<0$: il rapporto delle verosimiglianze

$$\frac{L\big(\theta_2\big)}{L\big(\theta_1\big)}\;=\;\left(\frac{\theta_2+1}{\theta_1+1}\right)^n\,\big(\prod_i X_i\big)^{\theta_2-\theta_1}$$

è *crescente* rispetto a $T=\prod_i X_i$ e si ottiene pertanto una regione critica della forma $D=\big\{\prod_i X_i\leq c\big\}$ con c tale che $\mathbf{P}^0\big\{\prod_i X_i\leq c\big\}=\alpha$, essendo α il livello desiderato.

I calcoli con *prodotti* di variabili indipendenti non sono agevoli, ma si può passare dai prodotti alle somme considerando i logaritmi: è immediato verificare che, sotto $\mathbf{P}^\theta$, $-\log X_i$ ha densità *esponenziale* di parametro $(\theta+1)$ e di conseguenza $-\log\left(\prod_i X_i\right) = -\sum_i \log X_i \sim \Gamma(n, \theta+1)$. Lasciamo completare i dettagli al lettore.

Capitolo 6

Inferenza statistica sui modelli gaussiani

6.1 Campioni statistici gaussiani

I modelli gaussiani sono largamente usati nell'inferenza statistica, sia perché sono molto maneggevoli dal punto di vista matematico, sia a causa del Teorema Limite Centrale: si pensa che un fenomeno casuale della realtà sia la combinazione di un numero elevato di *disturbi* casuali, e questo giustifica l'ipotesi che possa essere rappresentato con distribuzioni gaussiane.

Si pone però un problema metodologico: la densità $N(m, \sigma^2)$ (qualunque siano m e σ^2) è strettamente positiva su ogni intervallo. Ad esempio, che valore si può dare all'affermazione *"l'altezza media dei giovani che si presentano alla visita di leva a Pisa è gaussiana con media 180 (in cm) e varianza 100"*? Infatti risulta strettamente positiva la probabilità che l'altezza sia negativa, oppure superiore a 300 e questo appare assurdo.

Tuttavia le cose sono in realtà molto meno drastiche: abbiamo visto che i valori di una variabile con densità $N(0,1)$ sono di fatto compresi tra -3,5 e +3,5 (infatti $\Phi(3{,}5)$ differisce da 1 solo alla quarta cifra decimale) e di conseguenza i valori di una variabile $N(m, \sigma^2)$ sono compresi (a meno di eventi di probabilità inferiore a 10^{-3}) tra $m - 3{,}5\,\sigma$ e $m + 3{,}5\,\sigma$. Tornando all'esempio dei giovani alla visita di leva, questo si traduce nel considerare che l'altezza è compresa tra 145 e 215 cm, affermazione che appare perfettamente ragionevole.

Prima di addentrarci nell'esame di un campione di taglia n e densità gaussiana, vediamo alcuni risultati di probabilità preparatori.

Lemma 6.1.1. *Sia* $\mathbf{X} = (X_1, \ldots, X_n)$ *un vettore aleatorio formato da* n *v.a. indipendenti con densità* $N(0,1)$*, sia* A *una matrice* $n \times n$ ortogonale

(cioè la matrice di un cambio di base) e sia $\mathbf{Y} = A\,\mathbf{X}$. *Anche le componenti* $(Y_1, \dots, Y_n)$ *sono indipendenti con densità* $N(0,1)$.

Dimostrazione. La tesi equivale a dire che le variabili vettoriali $\mathbf{X}$ e $\mathbf{Y}$ sono equidistribuite.

La densità del vettore aleatorio $\mathbf{X}$ (scritta con notazione vettoriale) è $f(\mathbf{x}) = \big(2\pi\big)^{-\frac{n}{2}} \exp\big(-\frac{\|\mathbf{x}\|^2}{2}\big)$: se applichiamo la formula della Proposizione 3.4.7 (tenendo conto del fatto che la trasformazione $\mathbf{y} = A\,\mathbf{x}$ è un diffeomorfismo, con inversa $\mathbf{x} = A^{-1}\,\mathbf{y}$, e osservando che $\|A^{-1}\,\mathbf{y}\|^2 = \|\mathbf{y}\|^2$ poiché A è una matrice ortogonale), è immediato verificare che $\mathbf{Y}$ ha densità eguale a quella di $\mathbf{X}$.

□

Proposizione 6.1.2. *Siano* $(X_1, \dots, X_n)$ *indipendenti con densità* $N(0,1)$, *e definiamo* $\overline{X} = \frac{X_1 + \dots + X_n}{n}$. *Valgono i seguenti risultati:*

a) le variabili $\overline{X}$ *e* $\sum_{i \le n} \big(X_i - \overline{X}\big)^2$ *sono indipendenti;*

b) $\overline{X}$ *ha densità* $N(0, \frac{1}{n})$ *e* $\sum_{i \le n} \big(X_i - \overline{X}\big)^2$ *ha densità* $\chi^2(n-1)$ *;*

c) la variabile

$$\sqrt{n}\sqrt{n-1}\;\frac{\overline{X}}{\sqrt{\sum_{i \le n} \big(X_i - \overline{X}\big)^2}}$$

ha densità di Student $T(n-1)$.

Dimostrazione. Sia $\mathbf{e}_1$ il vettore $\mathbf{e}_1 = \big(\frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}}\big)$ e sia E_1 il sottospazio vettoriale di $\mathbb{R}^n$ generato da $\mathbf{e}_1$; sia poi E_2 l'ortogonale di E_1 e sia $\mathbf{e}_2, \dots, \mathbf{e}_n$ una base ortonormale di E_2. Sia poi A la matrice (ortogonale) di passaggio dalla base canonica di $\mathbb{R}^n$ alla base $\mathbf{e}_1, \dots, \mathbf{e}_n$.

Indichiamo con $\mathbf{X}$ il vettore aleatorio $(X_1, \dots, X_n)$ e sia $\mathbf{Y} = A\,\mathbf{X}$: in base al Lemma 6.1.1, le componenti $Y_1, \dots, Y_n$ sono ancora indipendenti con densità $N(0,1)$. Quindi Y_1 è indipendente da $(Y_2^2 + \dots + Y_n^2)$ che ha densità $\chi^2(n-1)$.

Notiamo che $Y_1 = \sqrt{n}\;\overline{X}$, inoltre $Y_2^2 + \dots + Y_n^2 = \sum_i Y_i^2 - Y_1^2 = \sum_i X_i^2 - n\overline{X}^2 = \sum_i \big(X_i - \overline{X}\big)^2$.

A questo punto le proprietà a) e b) sono immediate, e c) si ottiene come facile conseguenza tenendo conto della definizione della densità di Student.

□

La proposizione precedente era preparatoria del teorema che ora segue, che rappresenta il risultato preliminare fondamentale per l'inferenza statistica

su un campione gaussiano. Accanto alla notazione $\overline{X}$ che abbiamo appena definito, ne introduciamo un'altra che sarà usata fino alla fine di questo capitolo: se $(X_1, \dots, X_n)$ è un campione di n variabili aleatorie, indichiamo con

$$S^2 = \frac{\sum_i \big(X_i - \overline{X}\big)^2}{n-1}$$

(e naturalmente S ne è la radice quadrata). Se c'è pericolo di confusione (ad esempio se ci sono due campioni anche di taglia diversa $(X_1, \dots, X_n)$ e $(Y_1, \dots, Y_m)$) indicheremo $S^2(X)$ e $S^2(Y)$.

Teorema 6.1.3. *Siano $X_1, \dots, X_n$ indipendenti con densità $N(m, \sigma^2)$. Si hanno i seguenti risultati:*

a) le variabili $\overline{X}$ e S^2 sono indipendenti;

b) $\overline{X}$ ha densità $N(m, \frac{\sigma^2}{n})$ e $\frac{\sum_{i\le n}\big(X_i - \overline{X}\big)^2}{\sigma^2}$ ha densità $\chi^2(n-1)$;

c) la variabile

$$\frac{\sqrt{n}\,\big(\overline{X} - m\big)}{S}$$

ha densità di Student $T(n-1)$.

Dimostrazione. Possiamo scrivere $X_i = \sigma\, Y_i + m$, dove $Y_1, \dots, Y_n$ sono indipendenti con densità $N(0,1)$ e si applicano i risultati appena ottenuti nella Proposizione 6.1.2

Si hanno infatti le seguenti eguaglianze:

$$\overline{X} = \sigma\, \overline{Y} + m\,;$$

$$\frac{\sum_i (X_i - \overline{X})^2}{\sigma^2} = \sum_i (Y_i - \overline{Y})^2\,;$$

$$\frac{\sqrt{n}\,\big(\overline{X} - m\big)}{S} \;=\; \frac{\sqrt{n}\,\sigma\,\overline{Y}}{\sqrt{\frac{\sigma^2\,\sum_i (Y_i - \overline{Y})^2}{n-1}}} = \sqrt{n}\sqrt{n-1}\;\frac{\overline{Y}}{\sqrt{\sum_{i\le n}\big(Y_i - \overline{Y}\big)^2}}\,.$$

La facile conclusione è lasciata al lettore.

□

Consideriamo ora come modello statistico un **campione di taglia n e densità** $N(m, \sigma^2)$: sullo spazio $\Omega = \mathbb{R}^n$ consideriamo la verosimiglianza

$$L\big(m, \sigma^2\,;\, x_1, \dots, x_n\big) = \frac{1}{(2\pi)^{\frac{n}{2}}\,\sigma^n} \exp\Big(-\;\frac{\sum_i (x_i - m)^2}{2\,\sigma^2}\Big) =$$

$$= (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{\sum_i x_i^2}{2\,\sigma^2} + \frac{m}{\sigma^2}\,(\sum_i x_i) - \frac{n\,m^2}{2\,\sigma^2} - n\,\log\sigma \right)$$

L'insieme dei parametri Θ è $\mathbb{R}\times]0,+\infty[$ e come d'abitudine, indichiamo con $X_1,\dots,X_n$ le proiezioni coordinate.

Si dice che *la media è nota* se il parametro m è fisso (e di conseguenza come insieme dei parametri si considera $\Theta =]0,+\infty[$) ed analoga è naturalmente la definizione di modello con *varianza nota*.

Dalla formula della verosimiglianza, appare evidente che si ottiene un *riassunto esaustivo* con la variabile doppia $\left(\sum_i X_i, \sum_i X_i^2\right)$ (se la media è nota con $\sum_i \left(X_i - m\right)^2$, se la varianza è nota con $\sum_i X_i$).

Indaghiamo ora sull'esistenza delle *stime di massima verosimiglianza*: è sufficiente cercare i punti di massimo (rispetto a m ed a σ) dell'espressione

$$\left[-\frac{\sum_i x_i^2}{2\,\sigma^2} + \frac{m}{\sigma^2}\,(\sum_i x_i) - \frac{n\,m^2}{2\,\sigma^2} - n\,\log\sigma \right]$$

e per fare questo (dopo aver verificato le condizioni al limite, cioè l'andamento dell'espressione $\big[\dots\big]$ per $m \to \pm\infty$ e per $\sigma \to 0+$, $\sigma \to +\infty$) si annullano le derivate parziali, ottenendo le equazioni

$$\begin{cases} 0 = \dfrac{\partial}{\partial m}\big[\dots\big] = \dfrac{\sum_i x_i}{\sigma^2} - \dfrac{n\,m}{\sigma^2} \\ 0 = \dfrac{\partial}{\partial \sigma}\big[\dots\big] = \dfrac{\sum_i (x_i - m)^2}{\sigma^3} - \dfrac{n}{\sigma} \end{cases}$$

Facili conti provano che valgono le seguenti stime di massima verosimiglianza per i parametri:

1) $\widehat{m} = \overline{X}$ sempre;

2) $\widehat{\sigma}^2 = \dfrac{\sum_i \left(X_i - m\right)^2}{n}$ se m è nota;

3) $\widehat{\sigma}^2 = \dfrac{\sum_i \left(X_i - \overline{X}\right)^2}{n}$ se m è sconosciuta.

Notiamo ancora che la densità gaussiana $N(m,\sigma^2)$ si può scrivere nella forma $c(m,\sigma^2)\exp\left(-\frac{x^2}{2\sigma^2} + \frac{m}{\sigma^2}x \right)$ dove appare il prodotto scalare in $\mathbb{R}^2$ tra $T(x) = \left(x\,,\,x^2\right)$ ed il parametro bidimensionale $\left(\frac{m}{\sigma^2}\,,\,-\frac{1}{2\sigma^2}\right)$ (che è ovviamente in corrispondenza biunivoca col parametro *naturale* (m,σ^2)). Siamo dunque in presenza di un *modello esponenziale* e di conseguenza le stime di massima verosimiglianza sopra riportate sono consistenti.

È naturale chiedersi se queste stime siano *corrette*: è immediato constatare che $\overline{X}$ è una stima corretta del valore atteso, ma $\frac{\sum_i \left(X_i-\overline{X}\right)^2}{n}$ non è una stima corretta della varianza. Infatti $\frac{\sum_i \left(X_i-\overline{X}\right)^2}{\sigma^2}$ ha legge $\chi^2(n-1)$ e quindi valore atteso $(n-1)$.

Ne segue che una stima corretta della varianza è data da

$$S^2 = \frac{\sum_i \left(X_i - \overline{X}\right)^2}{n-1}$$

Osservazione 6.1.4. L'ultima proprietà non è specifica delle variabili gaussiane: infatti date n variabili $X_1, \ldots, X_n$ indipendenti equidistribuite, dotate di momento secondo, è sempre vero che

$$\mathbf{E}\Big[\frac{\sum_i \left(X_i - \overline{X}\right)^2}{n-1}\Big] \;=\; Var\big(X_1\big)$$

La prova di questo fatto è lasciata per esercizio.

6.2 Test sulla media di un campione gaussiano

In questo e nel successivo paragrafo supponiamo assegnato un campione $X_1, \ldots, X_n$ di taglia n e densità gaussiana.

Quando la varianza è nota, test e intervalli di fiducia sulla media m sono molto semplici e sono basati sul fatto che (sotto $\mathbf{P}^m$) $\overline{X}$ ha densità $N\big(m, \frac{\sigma^2}{n}\big)$ (o, equivalentemente, $\frac{\sqrt{n}\,\left(\overline{X}-m\right)}{\sigma}$ ha densità $N(0,1)$) : possiamo vedere un paio d'esempi come esercizi.

Esempio 6.2.1 (**Intervallo di fiducia per la media).** Trovare un intervallo di fiducia al livello 0,95 per la media di un campione gaussiano, con varianza nota.

Notiamo che abbiamo appena indicato una funzione del parametro e della variabie $\overline{X}$ la cui legge non dipende dal parametro m: possiamo dunque agevolmente utilizzare il *metodo della quantità pivot* cercando un intervallo di fiducia della forma $[\overline{X}(\omega) - d\,,\, \overline{X}(\omega) + d]$, con d tale che

$$\mathbf{P}^m\big\{\big|\overline{X} - m\big| > d\big\} = \mathbf{P}^m\Big\{\frac{\sqrt{n}}{\sigma}\big|\overline{X} - m\Big| > \frac{d\sqrt{n}}{\sigma}\Big\} \leq 0,05$$

Per avere un intervallo di fiducia più piccolo possibile, imponiamo che la diseguaglianza sopra scritta sia un'eguaglianza: ricordando che $\frac{\sqrt{n}}{\sigma}(\overline{X} - m)$ ha densità $N(0,1)$, scegliamo $\frac{d\sqrt{n}}{\sigma} = q_{0,975} = 1,96$ (dove q_α è lo α-*quantile* della legge $N(0,1)$).

Si ottiene così l'intervallo di fiducia $\overline{X}(\omega) \pm \frac{1,96\,\sigma}{\sqrt{n}}$.

Si noti la rassomiglianza con l'*intervallo di fiducia approssimato per il controllo di qualità* (Esempio 5.6.3)

Esempio 6.2.2 (**Test unilatero).** Individuare la regione critica di un test della forma $\mathcal{H}_0\big)\, m \le m_0$ contro $\mathcal{H}_1\big)\, m > m_0$, con varianza nota, al livello 0,02

Prendiamo $m_1 < m_2$ e scriviamo il rapporto delle verosimiglianze:

$$\frac{L\big(m_2\,; x_1, \ldots, x_n\big)}{L\big(m_1\,; x_1, \ldots, x_n\big)} = \exp\Big[\frac{m_2 - m_1}{\sigma^2}\big(\sum_i x_i\big) \;-\; \frac{n\big(m_2^2 - m_1^2\big)}{2\,\sigma^2}\Big]$$

Questo risulta crescente rispetto alla v.a. $\overline{X}$ e pertanto la regione critica sarà della forma $D = \big\{\overline{X} \ge c\big\}$ con c tale che $\mathbf{P}^{m_0}\big\{\overline{X} \ge c\big\} = 0,02$ (si pone l'eguale per avere la regione critica più grande possibile).

È più comodo scrivere la regione critica nella forma $\big\{\overline{X} - m_0 \ge d\big\}$, e ricordando che (sotto $\mathbf{P}^{m_0}$) $\frac{\sqrt{n}}{\sigma}\big(\overline{X} - m_0\big)$ ha densità $N(0,1)$, si pone

$$0,02 = \mathbf{P}^{m_0}\Big\{\overline{X} - m_0 \ge d\Big\} = \mathbf{P}^{m_0}\Big\{\frac{\sqrt{n}}{\sigma}\big(\overline{X} - m_0\big) \ge \frac{\sqrt{n}}{\sigma}d\Big\}$$

e di conseguenza si sceglie $\frac{\sqrt{n}}{\sigma}\,d = q_{0,98} = 2,055$. Si rifiuta quindi l'ipotesi se $\overline{X}(\omega)$ (cioè la media aritmetica dei dati osservati) supera $\big(m_0 + \frac{2,055\,\sigma}{\sqrt{n}}\big)$.

Esaminiamo ora il caso (molto più interessante e realistico) di test sulla media di un campione gaussiano *con varianza sconosciuta*, che è noto col nome di **test di Student**.

Nel caso in cui la varianza era nota, l'analisi era basata essenzialmente sulla variabile $\frac{\sqrt{n}\,\overline{X}}{\sigma}$, che ha densità $N\big(\frac{m\sqrt{n}}{\sigma}, 1\big)$: poichè ora la varianza σ non è nota, l'idea di Student è stata di sostituire a σ^2 la sua *stima corretta*, cioè S^2. L'analisi è ora concentrata sulla variabile

$$\frac{\sqrt{n}\,\overline{X}}{S} \;=\; \sqrt{n}\sqrt{n-1}\;\frac{\overline{X}}{\sqrt{\sum_{i\le n}\big(X_i - \overline{X}\big)^2}}$$

Cominciamo ad esaminare la sua distribuzione di probabilità.

Definizione 6.2.3 (**Legge di Student decentrata).** Si chiama legge di Student a n gradi di libertà decentrata di a (indicata anche $T(n)$ *decentrata di* a) la legge di

$$\frac{\sqrt{n}\ X}{\sqrt{Y}}$$

dove $X \sim N(a,1)$, $Y \sim \chi^2(n)$ e sono indipendenti.

La densità di questa legge di probabilità può essere calcolata, con conti molto tediosi, in modo analogo a quanto è stato fatto per la legge $T(n)$ non decentrata (vedi 3.6.3); in particolare è anche possibile verificare che le densità di Student decentrate di a , al variare di a , sono a rapporto di verosimiglianza crescente (rispetto all'identità, cioè alla variabile $T(x) = x$ su $\mathbb{R}$). Se questi conti sono molto pesanti, è invece un facile esercizio constatare che, se T sotto $\mathbf{P}^a$ ha legge di Student (n dimensionale) decentrata di a , la funzione $a \to \mathbf{P}^a\{T > c\}$ è crescente, ed è *questo solo che serve per il calcolo della taglia nel test unilatero.*

Osservazione 6.2.4. La variabile aleatoria $\frac{\sqrt{n}\,\overline{X}}{S}$ (sotto $\mathbf{P}^{m,\sigma^2}$) ha legge di Student $T(n-1)$ *decentrata* di $\frac{m\sqrt{n}}{\sigma}$.

Questa infatti è una conseguenza del fatto che si può scrivere

$$\frac{\sqrt{n}\ \overline{X}}{S} \;=\; \sqrt{n-1}\,\frac{\frac{\sqrt{n}}{\sigma}\ \overline{X}}{\sqrt{\frac{\sum_i \left(X_i-\overline{X}\right)^2}{\sigma^2}}}$$

In particolare, la legge di probabilità di questa variabile dipende dunque solo da $\frac{m}{\sigma}$.

Esempio 6.2.5 (**Test di Student unilatero).** Consideriamo, al livello α, la regione critica di un test dell'ipotesi $\mathcal{H}_0)\ m \le 0$, σ *qualsiasi* , contro l'alternativa $\mathcal{H}_1)\ m > 0$, σ *qualsiasi.*

Il test può essere scritto in questo modo:

$$\mathcal{H}_0)\ \frac{m}{\sigma} \le 0 \qquad \text{contro} \qquad \mathcal{H}_1)\ \frac{m}{\sigma} > 0$$

Poichè è stata individuata una variabile aleatoria (cioè $\frac{\sqrt{n}\,\overline{X}}{S}$) la cui distribuzione di probabilità dipende solo da $\frac{m}{\sigma}$ (ed è diversa per diversi valori di $\frac{m}{\sigma}$) restringiamo la nostra indagine a questa variabile aleatoria: la sua distribuzione di probabilità (cioè $T(n-1)$ decentrata di $\frac{m\sqrt{n}}{\sigma}$) è a rapporto

di verosimiglianza crescente rispetto a $\frac{m}{\sigma}$ e siamo pertanto condotti a una regione critica della forma

$$D = \{ \frac{\sqrt{n}\ \overline{X}}{S} \geq d \} \ = \ \Big\{ \omega \in \Omega \ \Big| \ \frac{\sqrt{n}\ \ \overline{X}(\omega)}{S(\omega)} \geq d \Big\}$$

con d tale che $\mathbf{P}^{0,\sigma^2} \big\{ \frac{\sqrt{n}\,\overline{X}}{S} \geq d \big\} = \alpha$ (ricordiamo che tale probabilità non dipende da σ se $m = 0$) : di conseguenza si prende $d = t_{(1-\alpha,\, n-1)}$ (vedi 3.6.3).

In base ai risultati teorici conseguenti al Lemma di Neyman-Pearson, sappiamo che questo test è *ottimale* tra tutti i test basati sull'osservazione della variabile $\frac{\sqrt{n}\,\overline{X}}{S}$ (vedi Teorema 5.7.6 per una formulazione più precisa di *"ottimale"*) ; in realtà si può dimostrare (facendo uso di nozioni più avanzate di quelle introdotte in questo corso) che è ottimale nella classe di tutti i possibili test sul modello.

Osservazione 6.2.6. Se il test è della forma

$$\mathcal{H}_0\big)\ m \leq m_0\,,\ \sigma \text{ qualsiasi} \qquad \text{contro} \qquad \mathcal{H}_1\big)\ m > m_0\,,\ \sigma \text{ qualsiasi}$$

non ci si può basare sul rapporto $\frac{m}{\sigma}$: allora (come spesso si fa in matematica) ci si riporta al caso precedente. Si considerano le variabili $\big(X_i - m_0\big)$ (che hanno legge $N\big(m - m_0, \sigma^2\big)$), e arriva di conseguenza a una regione critica della forma

$$D = \Big\{ \frac{\sqrt{n}\,\big(\overline{X} - m_0\big)}{S} \geq t_{(1-\alpha,\, n-1)} \Big\}$$

(lasciamo al lettore la verifica dei dettagli).

Esempio 6.2.7 (**Test di Student).** Consideriamo il test

$$\mathcal{H}_0\big) m = 0\,,\ \sigma \text{ qualsiasi} \qquad \mathcal{H}_1\big) m \neq 0\,,\ \sigma \text{ qualsiasi}$$

al livello α .

Il modo di procedere è simile a quello che è stato fatto precedentemente (non riportiamo i dettagli) ; si arriva ad una regione critica D della forma

$$D \ = \ \Big\{ \frac{\sqrt{n}\,\big|\overline{X}\big|}{S} \geq d \Big\}$$

con d tale che

$$\mathbf{P}^{0,\sigma^2} \Big\{ \frac{\sqrt{n}\,\big|\overline{X}\big|}{S} \geq d \Big\} = \alpha$$

Di conseguenza, si considera $d = t_{(1-\frac{\alpha}{2},\,n-1)}$ (vedi 3.6.3).

Il caso del test dell'ipotesi $\mathcal{H}_0\big)\, m = m_0\,,\ \sigma$ *qualsiasi* , viene trattato in modo analogo a quanto appena fatto: se α è il livello prescelto, si arriva alla regione critica

$$D = \left\{ \frac{\sqrt{n}\,\big|\overline{X} - m_o\big|}{S} \geq t_{(1-\frac{\alpha}{2},\,n-1)} \right\}$$

Esercizio 6.2.8. Il tempo medio di guarigione da una polmonite con i farmaci usuali è di 14 giorni: viene sperimentato su 17 pazienti un nuovo antibiotico (più costoso) e vengono rilevati i tempi di guarigione $x_1, \ldots, x_{17}$ che danno i risultati

$$\sum_{i=1}^{17} x_i = 197 \qquad \sum_{i=1}^{17} x_i^2 = 2596$$

Si può affermare che il nuovo farmaco in realtà non è più efficace?

Questi numeri $x_1, \ldots, x_{17}$ vengono interpretati come i valori osservati di un campione $X_1, \ldots, X_{17}$ con legge gaussiana $N\big(m, \sigma^2\big)$ sul quale viene effettuato il test dell'ipotesi

$$\mathcal{H}_0\big)\, m \geq 14\,,\ \sigma \text{ qualsiasi} \qquad \text{contro} \qquad \mathcal{H}_1\big)\, m < 14\,,\ \sigma \text{ qualsiasi}$$

ottenendo regione critica

$$\left\{ \frac{\sqrt{17}\,\big(\overline{X} - 14\big)}{S} \leq t_{(\alpha\,,16)} \right\}$$

dove α è il livello scelto. Ricordando che vale l'eguaglianza $t_{(\alpha\,,n)} = -t_{(1-\alpha\,,n)}$, dalle tavole della legge di Student si ricavano i valori $t_{(0{,}05\,;\,16)} = -1{,}746$ e $t_{(0{,}01\,;\,16)} = -2{,}58$.

I calcoli sui valori osservati portano a $\overline{x} = \frac{\sum_i x_i}{17} = 11,58$ e $s^2 = \frac{\sum_i (x_i - \overline{x})^2}{16} = 19,56$; e infine $\frac{\sqrt{17}(\overline{x}-14)}{s} = -2,25$. In conclusione, l'ipotesi viene rifiutata al livello 0,05 ed accettata al livello 0,01 .

In una situazione di incertezza come questa (cioè risultati diversi in corrispondenza di scelte diverse del livello) occorre essere cauti prima di arrivare a conclusioni pratiche.

Esempio 6.2.9 (Intervallo di fiducia per la media, con varianza sconosciuta).

Anche questa volta possiamo utilizzare il *metodo della quantità pivot* sfruttando il fatto che la variabile $\sqrt{n}\,\frac{\overline{X}-m}{S}$ ha legge di Student $T(n-1)$: lasciamo verificare per esercizio che un intervallo di fiducia per la media al livello $(1-\alpha)$, con varianza sconosciuta, è della forma

$$\overline{X}(\omega) \pm \frac{t_{(1-\frac{\alpha}{2}\,,\,n-1)}\, S(\omega)}{\sqrt{n}}\,.$$

6.3 Test sulla varianza di un campione gaussiano

Contrariamente a quanto si è visto per la media, l'indagine sulla varianza di un campione gaussiano è sostanzialmente identica nel caso in cui la media sia nota e in quello in cui sia sconosciuta, ed è basata su queste proprietà:

- se m è noto, $\frac{\sum_i \left(X_i - m\right)^2}{\sigma^2}$ ha densità $\chi^2(n)$;
- se m è sconosciuto, $\frac{\sum_i \left(X_i - \overline{X}\right)^2}{\sigma^2}$ ha densità $\chi^2(n-1)$.

Per essere precisi, le affermazioni sopra scritte sono vere sempre: si è detto *se* m *è noto* per evidenziare il fatto che la prima variabile va utilizzata solo nel primo caso. Per fissare le idee, concentriamoci sul secondo caso; come è stato fatto nel paragrafo precedente, limitiamo la nostra osservazione alla variabile $\sum_i \left(X_i - \overline{X}\right)^2$, la cui densità (sotto $\mathbf{P}^{m,\sigma^2}$) è, per x positivo, eguale a

$$f(x) = c(n)\sigma^{-(n+1)} x^{\frac{n-3}{2}} \mathrm{e}^{-\frac{x}{2\sigma^2}}$$

Lasciamo per esercizio la elementare verifica di questo, così come del fatto che queste densità siano a rapporto di verosimiglianza crescente.

Esempio 6.3.1 (**Test sulla varianza con media sconosciuta).** Consideriamo il test

$$\mathcal{H}_0)\ \sigma^2 \leq \sigma_0^2\ , m \text{ qualsiasi} \qquad \text{contro} \quad \mathcal{H}_1)\ \sigma^2 > \sigma_0^2\ , m \text{ qualsiasi}$$

al livello α .

Si tratta di un test *unilatero* sulla varianza, e si arriva alla regione critica

$$D = \left\{ \sum_i \left(X_i - \overline{X}\right)^2 \geq c \right\}$$

con c scelto in modo tale che si abbia

$$\mathbf{P}^{m,\sigma_0^2}\left\{ \frac{\sum_i \left(X_i - \overline{X}\right)^2}{\sigma_0^2} \geq \frac{c}{\sigma_0^2} \right\} = \alpha$$

e di conseguenza (poiché la distribuzione di $\dfrac{\sum_i \left(X_i - \overline{X}\right)^2}{\sigma_0^2}$ non dipende da m e, per $\sigma = \sigma_0$, è $\chi^2(n-1)$), si considera $\dfrac{c}{\sigma_0^2} = \chi^2_{(1-\alpha\,,\,n-1)}$ (vedi 3.6.2).

Quindi, osservati i dati $x_1, \ldots, x_n$, si rifiuta l'ipotesi se $\sum_i (x_i - \overline{x})^2 \geq \chi^2_{(1-\alpha\,,\,n-1)}\,\sigma_0^2$.

Osservazione 6.3.2. Il test dell'ipotesi $\mathcal{H}_0\big)\ \sigma^2 = \sigma_0^2$ (non importa se con m noto o sconosciuto) è meno agevole da trattare, ma per fortuna è anche meno importante nelle applicazioni. Sappiamo che la varianza è una misura della variabilità, di conseguenza applicato ad esempio a misurazioni su una produzione, l'ipotesi $\mathcal{H}_0\big)\ \sigma^2 \leq \sigma_0^2$ equivale a dire *la produzione è sufficientemente precisa* e quindi ha un evidente interesse pratico, mentre è meno importante indagare se *la variabilità corrisponde esattamente a un certo valore teorico.*

6.4 Confronto tra due campioni gaussiani indipendenti

In questo paragrafo ci occupiamo del caso in cui l'osservazione statistica sia formata da due campioni indipendenti $X_1, \ldots, X_n$ (di legge $N\big(m_1, \sigma_1^2\big)$) e $Y_1, \ldots, Y_k$ (di legge $N\big(m_2, \sigma_2^2\big)$).

Nel caso ad esempio in cui si abbiano dati su due siti archeologici diversi sarebbe un grave errore raggruppare tutti i dati in un unico campione: occorre tenere ben distinti i due campioni differenti. Quello che qui viene fatto con due, naturalmente può essere esteso a tre e più campioni ...

Il confronto tra i parametri di diversi campioni gaussiani indipendenti è un importante ed impegnativo capitolo dell'inferenza statistica che va sotto il nome di *analisi della varianza*: di esso ci limitiamo a dare qualche idea.

Volendo formalizzare come modello statistico il caso di due campioni indipendenti, si considera $\Omega = \mathbb{R}^{n+k}$, l'insieme dei parametri è $\Theta = \big(\mathbb{R}^2 \times]0, +\infty[^2\big)$ (si considera come parametro $(m_1, m_2, \sigma_1^2, \sigma_2^2)$) e la verosimiglianza è data da

$$L\big(m_1, m_2, \sigma_1^2, \sigma_2^2\,; x_1, \ldots, x_n, y_1, \ldots, y_k\big) = \prod_{i=1}^{n} f_{m_1,\sigma_2^2}(x_i) \prod_{j=1}^{k} f_{m_2,\sigma_2^2}(y_j)$$

essendo f_{m,σ^2} la densità $N(m, \sigma^2)$. Si considerano poi come X_i le proiezioni coordinate di indice i e come Y_j le proiezioni di indice $(n + j)$.

Esempio 6.4.1 (**Confronto tra due varianze).** Identifichiamo il test

$$\mathcal{H}_0\big)\ \sigma_1^2 \leq \sigma_2^2 \qquad \text{contro} \qquad \mathcal{H}_1\big)\ \sigma_1^2 > \sigma_2^2$$

al livello α prescelto.

Quando, come si è fatto sopra, non si scrive nulla sui parametri m_1 e m_2, si intende che questi sono qualsiasi.

Ricordiamo che la stima corretta di σ_1^2 è data da $S^2(X) = \sum_{i\le n}\big(X_i - \overline{X}\big)^2/(n-1)$ (e che $\sum_{i\le n}\big(X_i - \overline{X}\big)^2/\sigma_1^2$ ha densità $\chi^2(n-1)$), e analogamente per $S^2(Y)$: di conseguenza, se $\sigma_1^2 = \sigma_2^2$, la variabile

$$\frac{S^2(X)}{S^2(Y)} = \frac{\sum_i \big(X_i - \overline{X}\big)^2/(n-1)}{\sum_j \big(Y_j - \overline{Y}\big)^2/(k-1)}$$

ha legge di Fisher $F_{n-1\,,\,k-1}$ (vedi 3.6.4).

L'intuizione ci suggerisce di rifiutare l'ipotesi se il rapporto tra le stime delle due varianze è troppo grande (questa intuizione può essere sostenuta da un ragionamento più rigoroso, ma a prezzo di una certa fatica). Se chiamiamo $F_{(1-\alpha\,,\,n,\,k)}$ lo $(1-\alpha)$–quantile della legge $F_{n\,,k}$, la regione critica del test richiesto è data da

$$D = \left\{ \frac{\sum_{i\le n}\Big(X_i - \overline{X}\Big)^2/(n-1)}{\sum_{j\le k}\Big(Y_j - \overline{Y}\Big)^2/(k-1)} \ge F_{(1-\alpha\,,\,n-1,\,k-1)} \right\}$$

Esaminiamo ora il problema del confronto tra le medie, più impegnativo.

Definizione 6.4.2 (**Problema di Behrens-Fisher).** Si chiama *problema di Behrens-Fisher* l'individuazione della regione critica del test dell'ipotesi

$$\mathcal{H}_0\big)\; m_1 = m_2 \qquad \text{contro} \qquad \mathcal{H}_1\big)\; m_1 \neq m_2\,.$$

In questo problema non si pone alcuna condizione sulle varianze: questo problema ha ricevuto una soluzione completa (molto faticosa da ottenere) solo in tempi recenti. Noi ci limitiamo al caso più semplice nel quale si abbia $\sigma_1^2 = \sigma_2^2$ (cioè le varianze sono sconosciute, ma eguali).

Cominciamo con un facile risultato:

Lemma 6.4.3. *Se $m_1 = m_2$ e $\sigma_1^2 = \sigma_2^2$, la variabile*

$$Z_{n,k} = \frac{\overline{X} - \overline{Y}}{\sqrt{\sum_{i\le n}\big(X_i - \overline{X}\big)^2 + \sum_{j\le k}\big(Y_j - \overline{Y}\big)^2}}\; \frac{\sqrt{n+k-2}}{\sqrt{\frac{1}{n}+\frac{1}{k}}}$$

ha densità di Student $T(n+k-2)$.

Dimostrazione. Posto $\sigma^2 = \sigma_1^2 = \sigma_2^2$, la variabile $(\overline{X} - \overline{Y})/\sigma$ ha legge $N\big(0\,,\,\frac{1}{n} + \frac{1}{k}\big)$ e la variabile $\big[\sum_i \big(X_i - \overline{X}\big)^2 + \sum_j \big(Y_j - \overline{Y}\big)^2\big]/\sigma^2$ legge $\chi^2(n+k-2)$.

Inoltre le quattro variabili $\big(\overline{X}\,,\,\overline{Y}\,,\,\sum_i (X_i - \overline{X})^2\,,\,\sum_j (Y_j - \overline{Y})^2\big)$ sono indipendenti: la conclusione a questo punto è immediata.

□

La soluzione del problema di Behrens-Fisher (sotto l'ulteriore ipotesi $\sigma_1^2 = \sigma_2^2$) è a questo punto sostanzialmente un'estensione del test di Student: se consideriamo l'ipotesi $\mathcal{H}_0) m_1 = m_2$, si considera come regione critica (al livello α)

$$D = \left\{ \left| Z_{n,k} \right| \geq t_{(1-\frac{\alpha}{2}\,,\,n+k-2)} \right\}$$

mentre il test dell'ipotesi $\mathcal{H}_0) \, m_1 \leq m_2$ avrà regione critica

$$D = \left\{ Z_{n,k} \geq t_{(1-\alpha\,,\,n+k-2)} \right\}.$$

Esempio 6.4.4. Le misurazioni delle tibie da scheletri provenienti dalle tombe Etrusche di Cerveteri danno i seguenti risultati:

$$13 \text{ misurazioni} \qquad \overline{x} = 47,2 \qquad \frac{\sum (x_i - \overline{x})^2}{12} = 7,92\,,$$

mentre analoghe misurazioni dalle tombe di Ladispoli portano a

$$8 \text{ misurazioni} \qquad \overline{y} = 44,9 \qquad \frac{\sum (y_j - \overline{y})^2}{7} = 9,27\,.$$

Il risultato è casuale o si può affermare (al livello 0,05) che gli abitanti di Cerveteri erano effettivamente più alti?

Consideriamo i dati come risultati ottenuti su due campioni gaussiani indipendenti: per prima cosa ci poniamo il problema se possiamo considerare eguali le due varianze. Vogliamo più precisamente effettuare, al livelo 0,05, il test

$$\mathcal{H}_0) \, \sigma_2^2 = \sigma_1^2 \qquad \text{contro} \qquad \mathcal{H}_1) \, \sigma_2^2 > \sigma_1^2$$

(infatti, poiché la stima della varianza sul secondo campione risulta maggiore, non ci poniamo il problema che σ_2^2 possa essere minore: o è eguale, cioè il risultato è casuale, o è effettivamente maggiore).

Dalle tavole si ricava il valore $F_{(0{,}95\,;\,7,12)} = 2,91$, e poiché $\frac{9{,}27}{7{,}92} = 1{,}17$, accettiamo l'ipotesi dell'eguaglianza tra le due varianze.

A questo punto possiamo effettuare il test dell'ipotesi

$$\mathcal{H}_0) \, m_1 = m_2 \qquad \text{contro} \qquad \mathcal{H}_1) \, m_1 > m_2$$

I valori osservati per la variabile $Z_{13\,,\,8}$ portano a 1,761. Poiché $t_{(0{,}95\,;19)} = 1{,}729$, si rifiuta l'ipotesi e si conclude (al livello 0,05) che gli abitanti di Cerveteri erano effettivamente più alti.

6.5 Modelli statistici lineari: il teorema di Gauss-Markov

Definizione 6.5.1 (**Modelli lineari).** Si chiama *modello statistico lineare* un modello nel quale l'osservazione è data da n variabili aleatorie $X_1, \ldots, X_n$ che si possano scrivere nella forma

$$X_i = \sum_{j=1}^{k} a_{ij}\, \theta_j + \sigma W_i$$

con le seguenti proprietà:

a) $k < n$, $(\theta_1, \ldots, \theta_k) \in \mathbb{R}^k$ e $\sigma > 0$;

b) la matrice $n \times k$, $A = [a_{ij}]$ è di rango massimo (e quindi l'applicazione lineare ad essa associata $A : \mathbb{R}^k \to \mathbb{R}^n$ è iniettiva;

c) le variabili $W_1, \ldots, W_n$ sono gaussiane $N(0, 1)$ indipendenti.

Questa definizione è una generalizzazione della definizione che ora segue: i *modelli di regressione* sono all'origine dei modelli lineari.

Definizione 6.5.2 (**Modello di regressione).** Il modello è detto *di regressione* quando è della forma

$$X_i \;=\; \theta_1 + \theta_2 z_i + \cdots + \theta_k z_i^{k-1} + \sigma W_i$$

con $z_1 \neq z_2 \neq \cdots \neq z_n$ (e $k < n$).

In questo caso la matrice A corrispondente è della forma

$$A = \begin{bmatrix} 1 & z_1 \ldots z_1^{k-1} \\ & \cdots \\ 1 & z_n \ldots z_n^{k-1} \end{bmatrix}$$

ed è noto che una tale matrice (matrice di *Vandermonde*) è di rango massimo: i modelli di regressione sono dunque compresi nella Definizione 6.5.1.

Per i modelli lineari useremo anche la notazione vettoriale $\mathbf{X} = A\boldsymbol{\theta} + \sigma\mathbf{W}$.

Una prima osservazione è che le variabili aleatorie che costituiscono l'osservazione in un modello lineare *non formano un campione*: infatti non sono *equidistribuite*, sono tuttavia indipendenti, ed $X_i \sim N\big(\sum_j a_{ij}\theta_j, \sigma^2\big)$.

L'insieme dei parametri è $\Theta = \mathbb{R}^k \times]0, +\infty[$, e sullo spazio $\Omega = \mathbb{R}^n$ la verosimiglianza è data da

$$L\big(\boldsymbol{\theta}, \sigma^2 ; x_1, \dots, x_n\big) = (2\pi)^{-\frac{n}{2}} \exp\Big(- \frac{\sum_i \big(x_i - \sum_j a_{ij}\, \theta_j\big)^2}{2\,\sigma^2} - n\, \log \sigma \Big) =$$

$$= (2\pi)^{-\frac{n}{2}} \exp\Big(- \frac{\|\mathbf{x} - A\,\boldsymbol{\theta}\|^2}{2\,\sigma^2} - n \log\, \sigma \Big) .$$

Per essere precisi, non si dovrebbe dire nella Definizione 6.5.1 *"le variabili X_i ammettono la rappresentazione $X_i = \sum_{j=1}^k a_{ij}\, \theta_j + \sigma W_i$"*, bensì *"sotto la probabilità $\mathbf{P}^{\boldsymbol{\theta},\sigma^2}$, la legge di X_i è eguale alla legge di $\sum_{j=1}^k a_{ij}\, \theta_j + \sigma W_i$ "*.

Premettiamo un facile lemma:

Lemma 6.5.3. *Sia $A : \mathbb{R}^k \to \mathbb{R}^n$ una applicazione lineare* iniettiva. *Dato $\mathbf{x} \in \mathbb{R}^n$, il punto $\mathbf{y} \in \mathbb{R}^k$ che minimizza $\|\mathbf{x} - A\,\mathbf{y}\|^2$ è dato da $\mathbf{y} = U\mathbf{x}$, essendo $U = \big(A^t A\big)^{-1} A^t$.*

Dimostrazione. Cominciamo ad osservare che necessariamente $k \leq n$ (altrimenti A non potrebbe essere iniettiva); il caso $k = n$ è banale e quindi supponiamo $k < n$.

Proviamo che $A^t A$ (che è una matrice $k \times k$) è effettivamente invertibile: sia infatti $\mathbf{y} \in \mathbb{R}^k$ tale che $A^t A\,\mathbf{y} = 0$. Allora si ha

$$0 = \langle A^t A\,\mathbf{y}, \mathbf{y} \rangle = \langle A\,\mathbf{y}, A\,\mathbf{y} \rangle = \|A\,\mathbf{y}\|^2$$

e, poiché A è iniettiva, segue che $\mathbf{y} = 0$.

È facile constatare che la funzione $\mathbf{y} \to \|\mathbf{x} - A\mathbf{y}\|^2 = \sum_j \big(x_j - \sum_s a_{js} y_s\big)^2$ ammette minimo (è continua e tende a $+\infty$ per $\|\mathbf{y}\| \to +\infty$): per individuare il punto di minimo, annulliamo le derivate parziali. Si ottiene, per ogni i:

$$0 = -2 \sum_j a_{ji} \big(x_j - \sum_s a_{js} y_s\big)$$

cioè

$$\sum_j a^t_{ij}\, x_j \;=\; \sum_j \sum_s a^t_{ij}\, a_{js}\, y_s$$

che, scritta in notazione vettoriale, equivale a $A^t\mathbf{x} = A^t A\,\mathbf{y}$. La conclusione è immediata.

□

Osservazione 6.5.4. Nelle ipotesi del Lemma precedente, si ha $AU = P$, dove P è la proiezione ortogonale da $\mathbb{R}^n$ sul sottospazio $A\big(\mathbb{R}^k\big)$.

Torniamo all'espressione della verosimiglianza del modello nella forma vettoriale

$$L\big(\boldsymbol{\theta}, \sigma^2\,; \mathbf{x}\big) = (2\pi)^{-\frac{n}{2}} \exp\Big(- \frac{\|\mathbf{x} - A\,\boldsymbol{\theta}\|^2}{2\,\sigma^2} - n \log\,\sigma\Big)$$

per individuare le *stime di massima verosimiglianza*: in base al Lemma 6.5.3 la stima di $\boldsymbol{\theta}$ è $\widehat{\boldsymbol{\theta}}(\mathbf{x}) = U\mathbf{x}$ (o, scritta come variabile aleatoria, $\widehat{\boldsymbol{\theta}} = U\,\mathbf{X}$), e la stima di σ^2 è

$$\widehat{\sigma}^2 = \frac{\|\mathbf{X} - A\widehat{\boldsymbol{\theta}}\|^2}{n} = \frac{\|\mathbf{X} - A\,U\,\mathbf{X}\|^2}{n}\,.$$

Le *buone proprietà* di queste stime di massima verosimiglianza sono messe in luce dal risultato che viene ora enunciato.

Teorema 6.5.5 (**Teorema di Gauss Markov**). *$U\,\mathbf{X}$ è una stima corretta di $\boldsymbol{\theta}$, di rischio minimo tra tutte le stime lineari corrette. Inoltre*

$$\frac{\|\mathbf{X} - A\,U\,\mathbf{X}\|^2}{n-k}$$

è una stima corretta di σ^2.

Dimostrazione. Sia $V\,\mathbf{X}$ una stima *lineare* di $\boldsymbol{\theta}$: più precisamente V è una matrice $k \times n$ e $\big(V\mathbf{X}\big)_i = \sum_{j\le n} v_{ij}\,X_j$ è una stima di θ_i.

Poiché $\big(V\mathbf{X}\big)_i = \sum_{j,s} v_{ij}\,a_{js}\,\theta_s + \sigma \sum_j v_{ij} W_j$ ed ogni variabile W_j ha valore atteso 0, affinché valga l'eguaglianza $\mathbf{E}^{\boldsymbol{\theta},\sigma^2}\big[\big(V\mathbf{X}\big)_i\big] = \theta_i$, deve valere l'equazione $VA = I_k$, intendendo con I_k la matrice identità su $\mathbb{R}^k$.

È immediato constatare che la matrice U soddisfa questo requisito. Consideriamo viceversa una matrice V che soddisfa questa condizione, e calcoliamo il *rischio* della stima $\big(V\mathbf{X}\big)_i$:

$$\mathbf{E}^{\theta,\sigma^2}\Big[\Big(\theta_i - \sum_{j\le n} v_{ij}\,X_j\Big)^2\Big] = \sigma^2\mathbf{E}\Big[\Big(\sum_{j\le n} v_{ij}\,W_j\Big)^2\Big] = \sigma^2 \sum_{j\le n} v_{ij}^2 = \sigma^2 \sum_{j\le n} (v^t_{ji})^2$$

cioè è la norma della colonna i-ma della matrice V^t.

Sia P la proiezione ortogonale di $\mathbb{R}^n$ sul sottospazio $A(\mathbb{R}^k)$ e ricordiamo che $P = AU$ (vedi 6.5.4) : $VP = VAU$ e di conseguenza $U^t = PV^t$ (cioè la colonna i-ma della matrice U^t è la proiezione della colonna i-ma della matrice V^t). Poiché la proiezione diminuisce la norma, segue che il rischio di $U\mathbf{X}$ è inferiore a quello di $V\mathbf{X}$.

La seconda parte del teorema è una conseguenza del fatto che

$$\mathbf{X} - AU\mathbf{X} = \sigma\Big(\mathbf{W} - AU\mathbf{W}\Big) = \sigma\Big(\mathbf{W} - P\mathbf{W}\Big)$$

coincide con $\mathbf{W}$ proiettato sull'ortogonale del sottospazio $A(\mathbb{R}^k)$ (che è $(n-k)$-dimensionale).

Se questo fosse costituito dal sottospazio delle prime $(n-k)$ coordinate, sarebbe immediato verificare che $\mathbf{E}\big[\|\mathbf{X} - A\,U\,\mathbf{X}\|^2\big] = \sigma^2(n-k)$; in generale, si applica prima un cambio di base ortonormale in modo che i primi $(n-k)$ vettori della nuova base siano una base dell'ortogonale di $A(\mathbb{R}^k)$ e si tiene conto del Lemma 6.1.1.

□

Osservazione 6.5.6. Nella pratica, se non si dispone di un idoneo software statistico, non si calcola la matrice $\left(A^t A\right)^{-1} A^t$, ma, osservati i valori $x_1, \ldots, x_n$, i parametri $(\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ si stimano cercando

$$\min_{(\theta_1,\ldots,\theta_k)\in\mathbb{R}^k} \quad \sum_{i=1}^{n} \Big(x_i - \sum_{j=1}^{k} a_{ij}\,\theta_j\Big)^2$$

cioè, come si usa dire, si stimano i parametri col **metodo dei minimi quadrati**.

Osservazione 6.5.7 (Una curiosità storica). È facile verificare che Gauss è morto un anno prima che nascesse Markov, e viene dunque naturale chiedersi come possano aver trovato un teorema insieme: in realtà la formulazione del Teorema 6.5.5 come è enunciata sopra è una rielaborazione dovuta a Markov del *metodo dei minimi quadrati* ideato da Gauss.

Il primo utilizzo di questo metodo è stata fatto per risolvere un problema di astronomia: nel 1801 l'astronomo Piazzi aveva scoperto Cerere (il più grande degli asteroidi del sistema solare interno) e ne aveva seguito la traiettoria per qualche giorno, poi Cerere era diventato invisibile.

Le misurazioni effettuate vennero pubblicate e ne nacque una specie di *sfida scientifica* per ricostruire la traiettoria del pianetino: Gauss (che aveva solo 24 anni) a partire dalle misurazioni effettuate da Piazzi e ideando il metodo dei minimi quadrati, ricostruì la traiettoria di Cerere e previde quando e dove sarebbe riapparso. Dopo alcuni mesi Cerere venne nuovamente osservato proprio dove Gauss aveva previsto.