Università degli Studi di Firenze

Dipartimento di Matematica "Ulisse Dini"

Appunti del corso

Applicazioni di Matematiche e Statistica

Luigi Barletti

Dipartimento di Matematica "Ulisse Dini"
Università degli Studi di Firenze

Anno Accademico 2007/2008

Disponibile on-line all'indirizzo: `www.math.unifi.it/~barletti/`

Avvertenza agli studenti

Queste dispense sono leggermente *sovrabbondanti* rispetto al contenuto del corso e al programma di esame. Per la preparazione di quest'ultimo si possono *saltare* le seguenti parti:

- osservazione 1.9, esercizio 1.10 e oss. 1.11;
- nel paragrafo 1.4 le *permutazioni con ripetizioni* e la dimostrazione della formula per le *combinazioni con ripetizioni*;
- dimostrazione della prop. 2.7 e della prop. 2.8;
- esempio 2.11;
- dimostrazione del teorema 2.19;
- dimostrazione delle proposizioni 3.14 e 3.16;
- tutto il paragrafo 3.7;
- osservazione 4.5;
- tutto il paragrafo 4.5 tranne la prop. 4.23 (senza dimostrazione) e l'esercizio 4.24;
- paragrafi 5.3 e 5.4 (contengono prerequisiti di algebra lineare che *non* chiedo all'esame ma che sono indispensabili per capire gli argomenti successivi);
- i paragrafi 6.2.2 e 6.2.3;
- le stime di σ^2 (nei due casi μ nota e μ incognita) nel paragrafo 6.4;
- dimostrazione della prop. 6.11;
- nel paragrafo 8.1 i passaggi per la soluzione del sistema (8.6);
- tutto il paragrafo 8.2;
- tutto il paragrafo 8.4.

Avverto comunque che il programma può subire delle piccole variazioni da un anno all'altro e che per l'esame si fa riferimento al programma svolto in aula. In caso di dubbio siete invitati a contattarmi per tempo.

Indice

Parte I

Calcolo delle Probabilità

Capitolo 1

Il concetto matematico di probabilità

1.1 Alcuni esempi

Esempio 1.1.: Testa o croce (1 lancio)
Lanciamo una moneta; il risultato del lancio è il verificarsi di uno dei due eventi:

$$T\ (=\ \textit{esce testa}) \qquad \text{oppure} \qquad C\ (=\ \textit{esce croce}).$$

Intuitivamente, assegnamo a questi eventi le probabilità

$$P(T) = 1/2 \qquad P(C) = 1/2.$$

Il motivo per cui si assegna la probabilità $1/2$ a entrambi questi eventi è completamente "soggettivo", nel senso che non c'è alcuna regola che imponga la scelta della probabilità . D'altra parte, in questo caso la scelta è dettata dal buon senso. Se abbiamo detto "$P(T) = 1/2$" è perché abbiamo ragionato così: se provo a lanciare la moneta tante volte, la metà circa delle volte uscirà testa e l'altra metà croce. Oppure potremmo aver ragionato "per simmetria": non c'è motivo per cui una faccia della moneta sia "privilegiata" rispetto all'altra, dunque assegno lo stesso valore ad entrambe le facce.
In particolari situazioni (perché la moneta è marcatamente asimmetrica? perché sto giocando contro un noto truffatore? ...) nessuno mi vieta di scegliere una probabilità diversa

$$P(T) = p$$

dove $0 \leq p \leq 1$ è un numero reale compreso fra 0 e 1. A questo punto però sono obbligato a porre

$$P(C) = 1 - p,$$

poiché la probabilità che esca "o testa o croce" è 1 (certezza). □

Esempio 1.2.: Lancio di un dado
Giocando a testa o croce possiamo scommettere solo su due possibili eventi. Se giochiamo a lanciare un dado abbiamo molti più eventi a disposizione: gli eventi "elementari"

esce 1, *esce 2*, ... *esce 6*

e gli eventi "composti", come

esce un numero pari, *esce un numero maggiore di 4*, ...

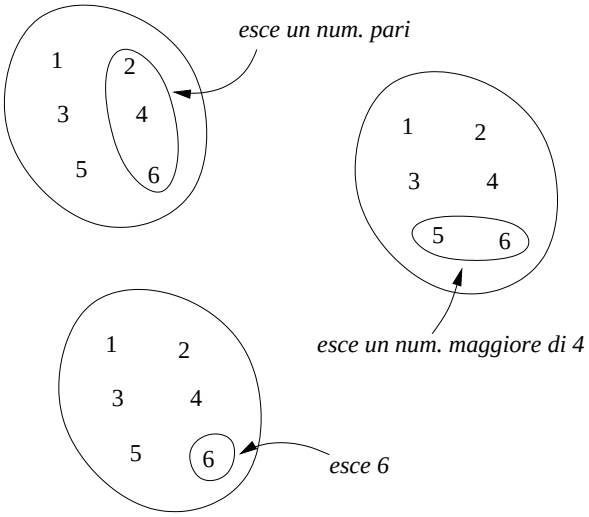


Figura 1.1: Alcuni eventi nel lancio del dado, visti come sottoinsiemi dell'insieme degli eventi elementari.

Senza bisogno di ripetere le considerazioni dell'esempio precedente, pare giusto scegliere le seguenti probabilità per gli eventi elementari:

$$P(\textit{esce 1}) = 1/6, \quad P(\textit{esce 2}) = 1/6, \quad \ldots \quad P(\textit{esce 6}) = 1/6.$$

A questo punto le probabilità degli eventi composti sono obbligate; ad esempio,

"esce un numero maggiore di 4" = *"esce 5 oppure esce 6"*

e dunque

$$P(\textit{esce un numero maggiore di 4}) = 1/6 + 1/6 = 1/3.$$

Infatti, se crediamo che il 5 e il 6 escano "mediamente" una volta ogni sei, dobbiamo anche ammettere che, sempre mediamente, due volte su sei esca un numero maggiore di 4.
Notiamo che gli eventi composti, e anche gli eventi elementari stessi, sono sottoinsiemi dell'insieme degli eventi elementari (Fig. 1.1). □

Esempio 1.3.: Lancio di due dadi
Supponiamo di lanciare due dadi, uno rosso e uno blu: quali sono gli eventi elementari? Chiaramente, sono le trentasei coppie

$$\begin{array}{cccccc} (1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6) \\ (2,1) & (2,2) & (2,3) & (2,4) & (2,5) & (2,6) \\ (3,1) & (3,2) & (3,3) & (3,4) & (3,5) & (3,6) \\ (4,1) & (4,2) & (4,3) & (4,4) & (4,5) & (4,6) \\ (5,1) & (5,2) & (5,3) & (5,4) & (5,5) & (5,6) \\ (6,1) & (6,2) & (6,3) & (6,4) & (6,5) & (6,6) \end{array}$$

a

$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$
$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(5,5)$	$(5,6)$
$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$

b

$(1,1)$	$(1,2)$	$(1,3)$	$(1,4)$	$(1,5)$	$(1,6)$
$(2,1)$	$(2,2)$	$(2,3)$	$(2,4)$	$(2,5)$	$(2,6)$
$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(3,5)$	$(3,6)$
$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$	$(4,5)$	$(4,6)$
$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(5,5)$	$(5,6)$
$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$	$(6,6)$

Figura 1.2: Due eventi nel lancio di due dadi, visti come sottoinsiemi dell'insieme degli eventi elementari. a) *sul dado rosso esce 5 e sul blu un numero pari* (la probabilità di questo evento è 1/12); b) T_7, ovvero *il punteggio totale dei due dadi è 7* (la probabilità di questo evento è 1/6).

dove la prima cifra indica il numero uscito sul dado rosso e la seconda il numero uscito sul dado blu. Se abbiamo fiducia nel fatto che ogni faccia del primo dado abbia probabilità 1/6 e ogni faccia del secondo dado abbia probabilità 1/6, pare giusto attribuire probabilità 1/36 a ciascuno degli eventi elementari sopraelencati. La scelta dei possibili eventi su cui scommettere è in questo caso molto più ricca che nei due esempi precedenti. Ad esempio, possiamo considerare l'evento

sul dado rosso esce 5 e sul blu un numero pari

(vedi figura 1.2a) oppure gli eventi

$$T_n = \textit{il punteggio totale dei due dadi è } n,$$

dove n è un numero che va da 2 a 12 (vedi figura 1.2b). □

Esercizio 1.4. *Con riferimento all'esempio precedente, calcolare $P(T_n)$ per ogni n compreso tra 2 e 12.* □

In questi primi, semplicissimi esempi abbiamo già fatto la conoscenza con tutti i concetti fondamentali del calcolo delle probabilità:

1. la *probabilità* è un numero, compreso tra 0 e 1, con il quale esprimiamo il nostro "grado di fiducia" nel verificarsi di un *evento*;

2. i possibili eventi sono sottoinsiemi di un *insieme di eventi elementari*, che dev'essere opportunamente individuato a seconda dello specifico sistema che si sta analizzando (lancio di una moneta, lancio di un dado, lancio di due dadi, etc.).

3. la discilplina matematica del "calcolo delle probabilità" *non serve a scegliere la probabilità ma piuttosto a dedurre ulteriori probabilità a partire da altre assegnate*, mediante deduzioni logico-matematiche basate su alcuni assiomi (paragrafo 1.2).

Quest'ultimo concetto è particolarmente importante anche perché la gente tende a ad avere le idee piuttosto confuse a riguardo: *non è il calcolo delle probabilità che mi dice che la probabilità di ottenere testa lanciando una moneta è 1/2.* Questa, per quanto ragionevole e di buon senso, è in realtà una scelta soggettiva. Il calcolo delle probabilità può servire invece per fare un ragionamento del tipo:

Se la probabilità di ottenere testa a ogni lancio è 1/2,
allora la probabilità di ottenere 4 teste in 10 lanci è 105/512

(questo calcolo, in particolare, impareremo a farlo nel paragrafo 3.1). La scienza che, invece, si occupa di fornire dei criteri per la scelta delle probabilità è semmai la statistica, che tratteremo nella seconda parte del corso.
C'è da dire che questa visione "soggettivista" della probabilità e "assiomatico-deduttiva" del calcolo delle probabilità non è affatto scontata e in passato (specialmente nelle epoche del positivismo) si intendeva fondare la teoria su una "definizione" di probabilità. In base a questa definizione la probabilità di ottenere testa "deve" essere $1/2$ perché se faccio un numero N molto grande di prove, il rapporto tra il numero di teste e il numero di lanci tende a $1/2$, in formule:

$$P(T) = \lim_{N\to\infty} \frac{\text{numero di teste}}{N} = \frac{1}{2}.$$

Questa impostazione, cosiddetta "frequentista", ha certamente la sua validità in determinate situazioni ma in generale è chiaramente troppo restrittiva. Invece, vorremmo che ricadessero sotto la disciplina del calcolo delle probabilità anche situazioni irripetibili, come ad esempio una corsa di cavalli, una partita di calcio o la quotazione di un titolo in borsa. Meglio quindi non pretendere di definire la probabilità, lasciarla come concetto primitivo e intuitivo, e costruirci sopra una teoria assiomatica rigorosa.

Esempio 1.5.: Il gioco del Totocalcio I
Un buon esempio per illustrare il significato soggettivo della probabilità è il gioco del Totocalcio. Come sappiamo, si tratta di indovinare l'esito (1 = vince la squadra di casa, X = pareggio, 2 = vince la squadra ospite) di 13 partite di calcio. Un evento elementare è quindi un possibile esito delle 13 gare (quella che si chiama "la colonna vincente"). L'insieme degli eventi elementari è dunque formato da tutte le possibili colonne di 13 simboli che possono essere 1, X o 2:

$$\Omega = \left\{ \begin{array}{|c|}\hline \omega_1 \\ \hline \omega_2 \\ \hline \omega_3 \\ \hline \vdots \\ \hline \omega_{13} \\ \hline \end{array} \;\middle|\; \omega_i \in \{1, \mathrm{X}, 2\} \right\}.$$

Qualè la probabilità di ogni singolo evento elementare, ovvero, qual'è la probabilità che una data colonna sia vincente? Se supponiamo che tutti gli eventi elementari siano equiprobabili (come negli esempi visti finora) la probabilità di ognuno di essi sarà $1/N$, dove N è il numero dei possibili eventi elementari, cioè di tutti gli elementi di Ω. È facile convincersi che $N = 3^{13} = 1594323$ (si veda anche il paragrafo 1.4) per cui la probabilità è 3^{-13} (che è circa 0.0000006: sei su dieci milioni), proprio come se giocassimo a lanciare un dado con 3^{13} facce.

Tuttavia non possiamo davvero credere che tutti gli eventi elementari siano equiprobabili! Sappiamo ad esempio che i 2 sono ben più rari degli 1 e delle X, sappiamo che ci sono squadre più forti e altre più deboli, sappiamo che il rendimento delle squadre può dipendere dalla posizione in classifica e così via. Anche la regola "frequentista" ha poco senso, poiché ogni giornata di campionato fa storia a sé, è unica e irripetibile. Perciò, a meno di giocare completamente a caso, quando compiliamo la schedina abbiamo in mente risultati più o meno probabili, secondo il nostro giudizio soggettivo. □

Terminiamo questa breve carrellata con un esempio, tratto sempre dal mondo dello sport, in cui si ha a che fare con eventi "continui".

Esempio 1.6.: Una gara fra due centometristi
Due atleti, A e B, corrono una gara di 100 metri piani della quale vogliamo indagare probabilisticamente l'esito. Come nell'esempio precedente, la scelta degli eventi elementari dipende da csa ci interessa prevedere. Facciamo due casi:

1. ci interessa solo chi arriva primo;
2. ci interessano i tempi.

Se ci interessa solo chi arriva primo, i possibili eventi elementari sono solo due:

$$\Omega_1 = \big\{\text{"vince } A\text{"},\ \ \text{"vince } B\text{"}\big\}.$$

Nel caso in cui ci interessano anche i tempi dei due atleti, ogni evento elementare è una coppia di numeri reali (t_A, t_B), dove

$$t_A = \text{tempo dell'atleta } A, \qquad t_B = \text{tempo dell'atleta } B.$$

Dunque avremo il seguente insieme di eventi elementari:

$$\Omega_2 = \big\{(t_A, t_B) \;\big|\; 0 \le t_A, t_B, \le T\big\} = [0,T] \times [0,T],$$

dove T è un tempo massimo oltre i quali siamo sicuri che gli atleti non vanno (p. es. $T = 20\,\text{sec}$). Notiamo che stavolta abbiamo a che fare con eventi elementari "continui" (i possibili tempi possono variare con continuità sull'intervallo $[0,T]$, mentre finora avevamo visto solo eventi "discreti" (testa o croce, facce del dado, colonne di una schedina, ecc.) L'insieme Ω_2 è rappresentato da un quadrato di lato T nel piano Cartesiano $\mathbb{R}^2$ ogni punto del quale, di coordinate (t_A, t_B), rappresenta una possible coppia di tempi ottenuti dagli atleti nella gara. Questa descrizione "contiene la precedente", nel senso che gli eventi elementari di Ω_1 sono eventi composti di Ω_2 e precisamente:

$$\text{"vince } A\text{"} = \big\{(t_A, t_B) \in \Omega_2 \;\big|\; t_A < t_B\big\}$$

$$\text{"vince } B\text{"} = \big\{(t_A, t_B) \in \Omega_2 \;\big|\; t_A > t_B\big\}$$

(vedi figura 1.3). □

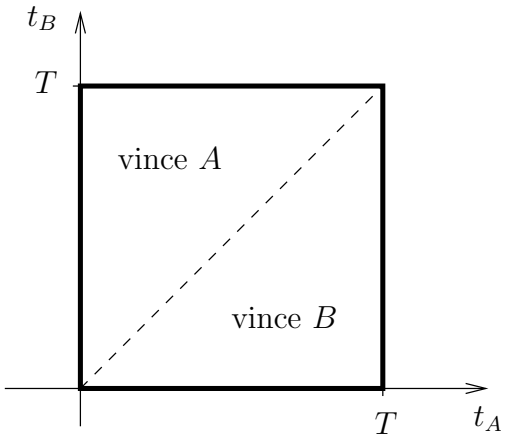


Figura 1.3: Gara fra due centometristi: l'insieme Ω_2 e gli eventi "vince A" e "vince B".

1.2 Assiomi del calcolo delle probabilità e prime conseguenze

Come già sottolineato nel paragrafo precedente, il moderno *calcolo delle probabilità* è una teoria assiomatico-deduttiva, che parte da definizioni e proprietà "primitive" e indimostrate, e da queste deduce conseguenze logicamente rigorose, né più né meno di quanto accade per la Geometria Euclidea. L'assiomatizzazione del calcolo delle probabilità è dovuta soprattutto al matematico russo Andrei Kolmogorov (1903-1987) che individuò la struttura matematica fondamentale comune a *ogni* sistema probabilistico, come ad esempio quelli che abbiamo visto nel paragrafo precedente. In questo paragrafo cercheremo ora di presentare e di capire tale struttura.

Ogni analisi probabilistica si basa innanzitutto sull'individuazione di un insieme Ω, i cui elementi sono interpretati come *eventi elementari*. I sottoinsiemi di Ω saranno detti *eventi*. L'insieme di tutti i sottoinsiemi di Ω, cioè di tutti gli eventi, si indica con $S(\Omega)$. Notiamo che ogni elemento di Ω può essere visto come un sottoinsieme (dunque gli eventi elementari sono eventi) e lo stesso Ω può essere visto come un sottoinsieme, (dunque anche Ω è un evento).
Se A_1 e A_2 sono due sottoinsiemi di Ω, cioè due eventi, l'evento $A_1 \cup A_2 \in S(\Omega)$ sarà interpretato come "si verifica A_1 oppure A_2". In particolare, poiché ogni sottoinsieme è un unione di elementi di Ω, ogni evento A si interpreta come *"si verifica almeno uno degli eventi elementari contenuti in A"*. Per lo stesso motivo, se A è un evento, il suo complementare A^c si interpreta come l'evento "A non si verifica" e, se A_1 e A_2 sono due eventi, l'intersezione $A_1 \cap A_2$ si interpreta come "si verifica sia A_1 che A_2. Riassumendo:

A^c		"non A"
$A_1 \cup A_2$	corrispondono a:	"A_1 o A_2"
$A_1 \cap A_2$		"A_1 e A_2"

Una volta definiti gli eventi, possiamo definire che cosa si intende per probabilità.

Definizione 1.7. Sia Ω un insieme fissato. Una *misura di probabilità* è una funzione $P : S(\Omega) \to [0,1]$, con le due seguenti proprietà:

(i) $P(\Omega) = 1$;

(ii) se $A_1, A_2 \in S(\Omega)$ e $A_1 \cap A_2 = \emptyset$, allora $P(A_1 \cup A_2) = P(A_1) + P(A_2)$.

□

Dunque, la misura di probabilità è una funzione che ad ogni evento $A \in S(\Omega)$ assegna un numero reale $P(A)$ compreso fra 0 e 1, che sarà detto *probabilità* di A. Le condizioni (i) e (ii) sono assiomi intuitivamente ben comprensibili: (i) ci dice che la probabilità dell'evento Ω (interpretabile come l'evento *Si verifica almeno uno degli eventi elementari*) è 1 (si tratta quindi di un *evento certo*); (ii) ci dice che se due sottoinsiemi A_1 e A_2 sono disgiunti (diremo anche che A_1 e A_2 sono *eventi incompatibili*), allora la probabilità che si verifichi o l'uno o l'altro è la somma delle probabilità di ciascuno.
Dalle proprietà (i) e (ii) possiamo dedurre le seguenti semplici coseguenze.

Proposizione 1.8. *Sia $P : S(\Omega) \to [0,1]$ una misura di probabilità. Valgono allora le seguenti proprietà:*

(1) $P(A^c) = 1 - P(A)$;

(2) $P(\emptyset) = 0$;

(3) $B \subset A \Rightarrow P(B) \leq P(A)$;

(4) se $A_1, A_2, \dots A_n$, sono tali che $A_i \cap A_j = \emptyset$ per $i \neq j$, allora

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) \qquad (1.1)$$

(5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Dimostrazione (1) Poiché $A \cup A^c = \Omega$ si ha $P(A \cup A^c) = 1$ (per la (i)) e, poiché $A \cap A^c = \emptyset$ si ha $P(A \cup A^c) = P(A) + P(A^c)$ (per la (ii)); dunque $P(A) + P(A^c) = 1$.
(2) Osserviamo innanzitutto che $\emptyset \in S(\Omega)$ (si tratta dell'evento "impossibile": *nessun evento elementare si verifica*); poiché $\emptyset = \Omega^c$, per il punto (1) e per la (i) si ha $P(\emptyset) = 1 - P(\Omega) = 0$.
(3) Scriviamo A come $A = B \cup (A \setminus B) = B \cup (A \cap B^c)$; poiché $B \cap (A \setminus B) = \emptyset$, per la (ii) si ha $P(A) = P(B) + P(A \setminus B)$. Dunque, essendo $P(A \setminus B) \geq 0$, si ottiene $P(A) - P(B) \geq 0$.
(4) Basta applicare ripetutamente la (ii):

$$P(A_1 \cup \cdots \cup A_n) = P(A_1 \cup \cdots \cup A_{n-1}) + P(A_n)$$

$$= P(A_1 \cup \cdots \cup A_{n-2}) + P(A_{n-1}) + P(A_n) = \cdots$$

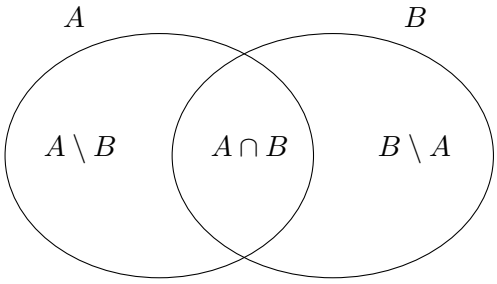


Figura 1.4:

(5) Scriviamo $A \cup B$ come $A \cup B = (A \setminus B) \cup (A \cap B) \cup (B \setminus A)$ (figura 1.4), da cui, usando la (4),

$$P(A \cup B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A)$$

$$= P(A) - P(A \cap B) + P(A \cap B) + P(B) - P(A \cap B)$$

$$= P(A) + P(B) - P(A \cap B)\,.$$

□

Abbiamo di fatto già utilizzato le proprietà (i) e (ii), o le loro conseguenze (1)-(4), nel paragrafo precedente. Nell'esempio 1.1 si ha $\Omega = \{T, C\}$ e, fissata $P(T) = p$, abbiamo concluso che $P(C) = 1 - p$, coerentemente con la (2). Negli esempi 1.2 e 1.3 abbiamo calcolato le probabilità degli eventi composti come somma delle probabilità degli eventi elementari che li compongono (ovviamente tutti gli eventi elementari sono disgiunti fra loro), il che segue dalla (4).

Osservazione 1.9. In molte circostanze la richiesta di poter assegnare una probabilità a *tutti* i sottoinsiemi di Ω è troppo forte. Pensiamo a insiemi Ω "continui", come Ω_2 dell'esempio 1.6: $S(\Omega)$ contiene anche un'infinità di sottoinsiemi "singolari" (si pensi a insiemi ancora più strani dei frattali) per i quali la definizione stessa di "misurabilità" può essere problematica. È più prudente indebolire la richiesta che P sia definita su tutto $S(\Omega)$ e richiedere invece che la probabilità sia definita su una certa classe di sottoinsiemi. Perciò, più in generale di quanto richiesto dalla definizione 1.7, la misura di probabilità è una funzione $P : \sigma(\Omega) \to [0, 1]$, dove $\sigma(\Omega) \subset S(\Omega)$ è un insieme più ristretto di sottoinsiemi, che diremo *misurabili*. Solamente i sottoinsiemi misurabili saranno detti *eventi*. Naturalmente l'insieme degli eventi $\sigma(\Omega)$ deve avere certe proprietà minime. In particolare richiediamo che:

(a) $\Omega \in \sigma(\Omega)$

(b) $A \in \Omega \Rightarrow A^c \in \sigma(\Omega)$

(c) se $A_1, A_2, \ldots$ sono un numero finito o un'infinità numerabile di elementi di $\sigma(\Omega)$ allora $A_1 \cup A_2 \cup \cdots$ è un elemento di $\sigma(\Omega)$.

Un sottoinsieme $\sigma(\Omega) \subset S(\Omega)$ con le proprietà (a), (b) e (c) viene detto *σ-algebra*. Notiamo che $S(\Omega)$ stesso è una σ-algebra. Nel caso di insiemi Ω con un numero finito di elementi, prenderemo sempre $\sigma(\Omega) = S(\Omega)$. □

Esercizio 1.10. *Sia Ω un insieme qualsiasi e $A \in S(\Omega)$ un qualsiasi sottoinsieme. Dimostrare che $\{\emptyset, A, A^c, \Omega\}$ è una σ-algebra. Dimostrare anche che $S(\Omega)$ stesso è una σ-algebra.* □

Osservazione 1.11. Avendo a che fare con infinità di eventi, può essere utile rafforzare l'assioma (ii) (e la sua conseguenza (4)) aggiungendo il seguente assioma:

(iii) se $A_1, A_2, \ldots$ sono un'infinità numerabile di eventi incompatibili cioè tali che $A_i \cap A_j = \emptyset$ per $i \neq j$, allora

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Notiamo che *non* si richiede un'analoga proprietà per un'infinità non-numerabile (ad esempio un continuo) di eventi. □

Concludiamo il paragrafo ricordando le seguenti proprietà degli insiemi: siano A, B e C sottoinsiemi di un insieme Ω, allora

1. $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$;
2. $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$;
3. $(A \cup B)^c = A^c \cap B^c$;
4. $(A \cap B)^c = A^c \cup B^c$.

Esercizio 1.12. *Dimostrare le precedenti identità 1-4 (utilizzare diagrammi del tipo di quello in figura 1.4).* □

1.3 Eventi elementari equiprobabili I

Casi particolarmente semplici di sistemi probabilistici sono quelli in cui gli eventi elementari sono in numero *finito* N,

$$\Omega = \{\omega_1, \omega_2, \ldots \omega_N\}$$

e inoltre gli eventi elementari sono *equiprobabili*, ovvero

$$P(\omega_i) = p, \qquad \text{per ogni } i,$$

per un opportuno $0 < p \leq 1$. Sono sistemi di questo tipo quelli trattati negli esempi 1.1, 1.2 e 1.3. Dalla proprietà (1.1), si deduce subito che la probabilità p di ogni evento elementare è data dall'inverso del numero di eventi elementari:

$$p = 1/N$$

(infatti deve risultare $P(\omega_1) + \cdots + P(\omega_N) = Np = 1$). Più in generale, la probabilità di un evento $A \in S(\Omega)$ è data da

$$P(A) = \frac{\textit{numero di elementi di } A}{N}. \tag{1.2}$$

Spesso la formula precedente si trova scritta nella forma

$$P(A) = \frac{\textit{numero dei casi favorevoli}}{\textit{numero dei casi possibili}},$$

dove per "casi possibili" si intendono tutti gli eventi elementari e per "casi favorevoli" si intendono gli eventi elementari contenuti in A. Ricordiamo che queste semplici formule valgono *solo* nel caso di eventi elementari equiprobabili! Abbiamo di fatto già usato la formula (1.2) negli esempi 1.2 e 1.3. Infatti, riprendendo l'esempio del lancio di due dadi, per calcolare la probabilità dell'evento T_7 (cioè la probabilità di totalizzare 7 sui due dadi) devo dividere per il numero dei casi favorevoli, ovvero il numero di eventi elementari contenuti in T_7, che è 6 (figura 1.2b) per il numero dei casi possibili, cioè di tutti i possibili eventi elementari, che è 36. Si ottiene così $P(T_7) = 6/36 = 1/6$.

È evidente che i sistemi probabilistici di questo tipo riconducono il calcolo delle probabilità a problemi di *enumerazione*, cioè al problema di contare il numero di elementi in certi insiemi.
La disciplina matematica che si occupa dei problemi di enumerazione è il *calcolo combinatorio*. Nel prossimo paragrafo vedremo alcuni semplici concetti e risultati del calcolo combinatorio che ci forniranno gli strumenti per risolvere problemi del tipo di quello dell'esempio seguente.

Esempio 1.13.: Il gioco del Poker
Il Poker "all'italiana" in quattro giocatori si gioca con un mazzo di 32 carte (dai "sette" ai "dieci" più fanti, regine, re e assi). Pescando a caso cinque carte dal mazzo, che probabilità ho di avere un poker (4 carte dello stesso valore)?
Per rispondere a questa domanda si deve innanzitutto individuare il corretto insieme di eventi elementari, il che ci fornisce il numero di casi possibili N, contare il numero di casi favorevoli e applicare la formula (1.2). Un evento elementare è un gruppo di cinque carte estratto da un mazzo di 32. Dunque N, il numero di eventi elementari, lo trovo calcolando in quanti modi diversi posso prendere cinque carte da un mazzo di 32. Il numero di casi favorevoli N_f è il numero di tali gruppi che contengono 4 carte dello stesso valore. Poiché gli eventi elementari sono tutti equiprobabili (salvo casi di baro), la probabilità di avere poker la calcolo come N_f/N.
Come si vede, il problema probabilistico si riduce a calcolare i numeri N e N_f ossia, in definitiva, a un problema di enumerazione. Questa semplicità concettuale del problema non deve trarre in inganno: i problemi di enumerazione possono essere anche molto difficili (e anche quello di questo esempio non è proprio banalissimo). Riprenderemo questo esempio nel paragrafo 1.5, dopo aver dato un po' di basi di calcolo combinatorio. □

1.4 Elementi di calcolo combinatorio

Come abbiamo visto nel precedente paragrafo, calcolare la probabilità degli eventi, nel caso di eventi elementari equiprobabili, equivale a risolvere problemi di calcolo combinatorio. Qui di seguito enunceremo alcuni problemi tipici del calcolo combinatorio e dedurremo le formule che li risolvono.

Permutazioni di n oggetti distinti

Se ho un certo numero n di oggetti distinti, una loro *permutazione* è un possibile modo di ordinarli. Ad esempio, le possibili permutazioni dei numeri 1, 2 e 3 sono

$$(1,2,3),\ (1,3,2),\ (2,1,3),\ (2,3,1),\ (3,1,2),\ (3,2,1).$$

Il numero delle possibili permutazioni di n oggetti distinti si indica con $\pi(n)$ ed è dato da

$$\pi(n) = n! \tag{1.3}$$

dove ricordiamo che $n! = n(n-1)(n-2)\cdots 1$. Infatti possiamo contare le possibili permutazioni nel seguente modo. Immaginiamo di avere n oggetti e di doverli ordinare: abbiamo n possibili scelte dell'oggetto da mettere per primo; scelto il primo, restano $n-1$ oggetti tra cui scegliere il secondo e così via, finché non rimane che un unico oggetto da mettere per ultimo. Dunque i pssibili modi per ordinare gli oggetti sono $n(n-1)(n-2)\cdots 1$, ovvero $n!$.

Permutazioni di n oggetti con ripetizioni

Supponiamo adesso che nel gruppo di n oggetti non tutti gli oggetti siano distinti ma che invece ci possano essere più copie indistinguibili dello stesso oggetto (figura 1.5). Supponiamo che il numero di oggetti distinti sia s e che l' i-esimo

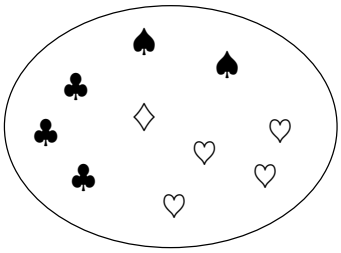


Figura 1.5: Gruppo di oggetti con ripetizioni.

oggetto sia ripetuto n_i volte (dunque $n_1 + n_2 + \cdots + n_s = n$). Ad esempio, nella figura 1.5 ci sono $n = 10$ oggetti, gli oggetti distinti sono $s = 4$ e ci sono $n_1 = 3$ fiori, $n_2 = 2$ picche, $n_3 = 1$ quadri e $n_4 = 4$ cuori.
Sia $\pi^*(n_1, n_2, \ldots, n_s)$ il numero di possibili ordinamenti degli n oggetti. Per calcolare questo numero supponiamo dapprima che gli n oggetti siano tutti distinti: allora le combinazioni sono $n!$, come visto nel caso precedente. Ma ora dobbiamo tener conto che ci sono degli oggetti indistinguibili, permutando i quali si ottiene sempre lo stesso ordinamento: ogni permutazione con ripetizioni ha $n_1!$ repliche indistinguibili ottenute per permutazione degli oggetti 1, $n_2!$ repliche indistinguibili ottenute per permutazione degli oggetti 2 e così via.

Pertanto, ogni permutazione ha $n_1!\,n_2!\cdots n_s!$ copie indistinguibili. Si ottiene dunque

$$\pi^*(n_1, n_2, \ldots, n_s) = \frac{n!}{n_1!\,n_2!\cdots n_s!}. \tag{1.4}$$

Disposizioni di n oggetti presi k alla volta

Una *disposizione* di n oggetti distinti presi k alla volta (con $k \leq n$, ovviamente), è un possibile modo di scegliere k degli n oggetti e di ordinarli. Ad esempio, tutte le disposizioni di 1, 2, 3 e 4 presi due alla volta sono:

$$\begin{array}{cccccc} (1,2) & (2,1) & (1,3) & (3,1) & (1,4) & (4,1) \\ (2,3) & (3,2) & (2,4) & (4,2) & (3,4) & (4,3). \end{array}$$

Se indichiamo con $D(n,k)$ il numero delle possibili disposizioni di n oggetti presi k alla volta otteniamo

$$D(n,k) = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}. \tag{1.5}$$

Infatti, analogamente a quanto fatto per le permutazioni, possiamo contare le disposizioni dicendo che ci sono n possibilità per scegliere il primo oggetto, $n-1$ per scegliere il secondo e così via, solo che stavolta ci dobbiamo fermare alla k-esima posizione, per la quale rimangono $n-k+1$ possibilità.
Notiamo infine che $D(n,n) = \pi(n) = n!$, il che è consistente con la formula (1.5) se si pone, com'è consuetudine, $0! = 1$.

Disposizioni con ripetizioni di n oggetti presi k alla volta

Una *disposizione con ripetizioni* di n oggetti distinti presi k alla volta è un possibile modo di sceglere k oggetti *eventualmente ripetuti* dagli n e ordinarli. Stavolta k può essere anche maggiore di n in quanto ogni oggetto lo posso ripetere quante volte voglio. Ad esempio, le disposizioni con ripetizioni degli oggetti $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ presi tre alla volta, sono tutti i numeri interi da 0 a 999:

$$000,\ 001,\ 002,\ 003,\ 004,\ \ldots\ldots\ 995,\ 996,\ 997,\ 998,\ 999.$$

Se $D^*(n,k)$ è il numero di tutte le disposizioni con ripetizioni di n oggetti presi k alla volta, si ottiene

$$D^*(n,k) = n^k. \tag{1.6}$$

Infatti le posso conteggiare così: ho n possibili scelte per il primo oggetto ed ho ancora n possibili scelte per il secondo (perché stavolta posso ripetere lo stesso oggetto in seconda posizione) e così via fino alla k-esima posizione. Dunque ho n^k possibilità in tutto.

Combinazioni di n oggetti presi k alla volta

Una *combinazione* di n oggetti distinti presi k alla volta (con $k \leq n$) è un possibile modo di sceglere k degli n oggetti (senza ordinarli). Dunque, a differenza delle disposizioni, stavolta l'ordine in cui prendo gli oggetti non conta. Ad esempio, tutte le possibili combinazioni di 1, 2, 3 e 4 presi due alla volta sono:[1]

$$\{1,2\}\quad \{1,3\}\quad \{1,4\}\quad \{2,3\}\quad \{2,4\}\quad \{3,4\}$$

[1] L'uso delle parentesi graffe al posto delle tonde serve a sottolineare il fatto che si tratta di insiemi e non di coppie ordinate: convenzionalmente, infatti, $(1,2)$ indica una coppia ordinata (per cui $(1,2) \neq (2,1)$) mentre $\{1,2\}$ indica un insieme (per cui $\{1,2\} = \{2,1\}$).

Il numero delle possibili combinazioni di n oggetti presi k alla volta si indica con $C(n,k)$ oppure con $\binom{n}{k}$ (che prende il nome di *coefficiente binomiale*) e si ha

$$C(n,k) = \binom{n}{k} = \frac{n(n-1)(n-2)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!\,k!}. \tag{1.7}$$

Infatti, per contare tutte le combinazioni di n oggetti presi k alla volta posso dapprima contare tutte le possibili disposizioni e poi identificare le disposizioni che differiscono solo per l'ordine degli oggetti. Ad esempio, supponiamo di voler contare le combinazioni degli oggetti $1,2,3,4$ presi tre alla volta; allora disposizioni differenti, come

$$(1,2,4),\ (1,4,2),\ (2,1,4),\ (2,4,1),\ (4,1,2),\ (4,2,1),$$

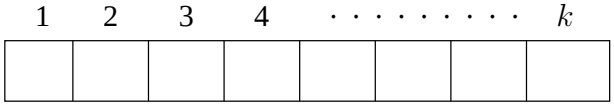
si identificano con l'unica combinazione $\{1,2,4\}$. È chiaro quindi che, contando tutte le disposizioni (e sappiamo che ce ne sono $D(n,k) = n(n-1)(n-2)\cdots(n-k+1)$), abbiamo contato $\pi(k) = k!$ volte ogni diversa combinazione. Dobbiamo perciò dividere $D(n,k)$ per $\pi(k)$, ottenendo esattamente la (1.7).

Combinazioni con ripetizioni di n oggetti presi k alla volta

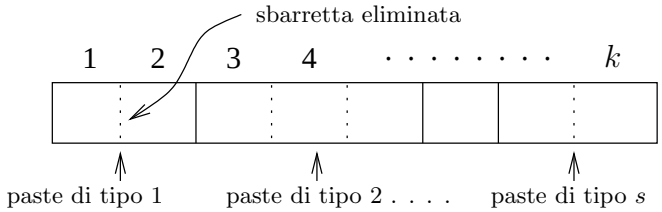
Una *combinazione con ripetizioni* di n oggetti distinti presi k alla volta è un possibile modo di sceglere k degli n oggetti (senza ordinarli) potendo ripetere ogni oggetto quante volte vogliamo. Stavolta k può essere più grande di n. Quant'è il numero, che indicheremo con $C^*(n,k)$, di tutte le possibili combinazioni con ripetizioni? C'è un modo tradizionale di enunciare questo problema noto come "problema della pasticceria": nel banco della pasticceria ci sono n tipi diversi di paste e io voglio riempire un vassoio con k paste (con eventuali ripetizioni, cioè posso prendere due o più o, al limite, tutte e k le paste dello stesso tipo). In quanti modi diversi posso riempire il vassoio? La risposta è:

$$C^*(n,k) = C(n+k-1,k) = \binom{n+k-1}{k} = \frac{(n+k-1)!}{(n-1)!\,k!}. \tag{1.8}$$

La dimostrazione di questa formula non è proprio immediata. Restando nel linguaggio paste-vassoio, per riempire il vassoio con k paste posso cominciare a scegliere quanti tipi di paste diverse voglio. Sia s questo numero: voglio riempire il vassoio con k paste di s tipi diversi (naturalmente deve essere $s \le k$). Fissato s, devo decidere quante paste metterò di ciascun tipo. Numeriamo gli s tipi di paste prescelti da 1 a s: comincio a mettere le paste di tipo 1 nel vassoio, a un certo punto mi fermo e comincio a mettere le paste di tipo 2, e così via fino alle paste di tipo s. Se vedo il vassoio come k caselle da riempire con gli s tipi di paste:



una possibile configurazione è univocamente determinata eliminando $k-s$ delle $k-1$ sbarrette verticali e lasciando $s-1$ sbarrette a delimitare le caselle occupate dagli s tipi di paste:



Dunque, fare una combinazione con ripetizioni di k oggetti scelti da n equivale a scegliere s oggetti su n e $k-s$ numeri tra 1 e $k-1$ (le sbarrette da togliere): si tratta perciò di scegliere k elementi (senza ripetizioni e senza ordine) da un insieme di $n+k-1$ oggetti, il che può essere fatto in $C(n+k-1,k)$ modi diversi.

1.5 Eventi elementari equiprobabili II

Vediamo adesso alcuni esempi di utilizzo delle formule del calcolo combinatorio per calcolare probabilità di eventi nel caso di eventi elementari equiprobabili, come discusso nel paragrafo 1.3.

Riprendiamo innanzitutto l'esempio 1.13. Abbiamo visto i "casi possibili", cioè gli eventi elementari, sono tutti i modi in cui possiamo estrarre 5 carte dal mazzo di 32. È evidente che si tratta di tutte le possibili combinazioni (senza ripetizioni) di 32 oggetti presi 5 alla volta e dunque i casi possibili sono $C(32,5)$. I "casi favorevoli" sono tutti gli eventi elementari (ovvero tutte le suddette combinazioni) che che contengono un poker (ovvero quattro carte dello stesso valore). Li possiamo contare direttamente: i possibili poker sono 8 (da quello di "sette" a quello di assi) e, per ciascuno di essi, ci sono $32-4=28$ possibilità per la rimanente carta (quattro sono quelle che formano il poker); dunque i casi favorevoli sono $8\times 28=224$. La probabilità di avere un poker scegliendo a caso 5 carte da un mazzo di 32 è perciò[2]

$$\frac{\textit{casi favorevoli}}{\textit{casi possibili}}=\frac{224}{C(32,5)}=\frac{224}{201376}\approx 0.0011.$$

Nel caso del gioco "all'americana" (mazzo di 52 carte), i casi possibili sono $C(52,5)$ mentre i casi favorevoli sono $13\times(52-4)=624$ (tredici tipi di poker per $52-4$ possibilità per la carta rimanente). Dunque la probabilità di avere un poker scegliendo a caso 5 carte da un mazzo all'americana è

$$\frac{\textit{casi favorevoli}}{\textit{casi possibili}}=\frac{624}{C(52,5)}=\frac{624}{2598960}\approx 0.00024$$

(circa cinque volte più piccola).

Esempio 1.14.
Ho quattro scatole chiuse, numerate da 1 a 4, e so che dentro ci sono complessivamente tre palline (ma non so come queste siano distribuite nelle scatole).

[2]Nel fornire le soluzioni numeriche degli esercizi scriveremo il risultato sotto forma di frazione esatta (quando possibile) o in forma di numero decimale approssimato a due cifre significative.

Che probabilità ho che le tre palline siano tutte nella scatola n. 1?

Gli eventi elementari, che possono essere considerati equiprobabili, sono le possibili distribuzioni delle tre palline nelle quattro scatole. Detto ciò, risolviamo l'esercizio in due modi diversi.

1º metodo ("diretto"). Contiamo i casi possibili e i casi favorevoli. I casi favorevoli sono solamente uno (la configurazione in cui le tre palline sono nella scatola 1). Contare i casi possibili è facile: ci sono 4 configurazioni in cui le tre palline sono tutte in una sola scatola e 4 in cui le tre palline sono tutte in scatole diverse. Restano i casi di tipo 2+1 (due palline nella stessa scatola, una in una scatola diversa e due scatole vuote): le due scatole vuote possono essere sceltre in 6 modi diversi e per ogni scelta delle due scatole vuote restano 2 possibilità per mettere la pallina singola. Dunque le configurazioni di tipo 2+1 sono $6 \times 2 = 12$. Complessivamente abbiamo contato $4 + 4 + 12 = 20$ casi possibili e pertanto la probabilità cercata è $1/20$.

2º metodo (applicazione delle formule del c. combinatorio). Osserviamo che posso interpretare il"mettere una pallina in una scatola" come lo "scegliere" quella scatola e il "mettere due palline" in una scatola come lo "scegliere due volte" quella scatola. Ma allora una distribuzione delle tre palline nelle quattro scatole la posso vedere come una scelta, con possibili ripetizioni, di tre scatole dalle quattro disponibili. In altre parole, è una combinazione con ripetizioni di 4 oggetti presi tre alla volta (nel linguaggio paste-vassoio: è come se dovessi riempire un vassoio con tre paste scegliendole da quattro tipi diversi e per scegliere una pasta ci mettessi una pallina accanto). Dunque i casi possibili sono $C^*(4,3)$ mentre il caso favorevole è, come già detto, uno solo. (tutte e tre le palline nella prima scatola, ovvero la prima scatola "scelta tre volte"). Pertanto, ricordando la formula (1.8), la probabilità è

$$\frac{\textit{casi favorevoli}}{\textit{casi possibili}} = \frac{1}{C^*(4,3)} = \frac{1}{C(6,3)} = \frac{1}{20}.$$

Notiamo che questo secondo metodo è migliore quando si ha a che fare con numeri grandi. □

Esempio 1.15.: Paradosso dei compleanni
Si tratta di un classico problema didattico di calcolo delle probabilità. Possiamo formularlo così: *quanti studenti ci devono essere in una classe affinché la probabilità che almeno due siano nati lo stesso giorno dell'anno sia maggiore del 50%?*.
Supponiamo per semplicità che l'anno sia fatto di 365 giorni, tutti equiprobabili. Sia s il numero di studenti nella classe e sia A l'evento di cui vogliamo calcolare la probabilità:

$$A = \textit{"in classe ci sono almeno due studenti nati lo stesso giorno"}.$$

Come spesso succede negli esercizi di calcolo delle probabilità, in questo caso non conviene andare dritti al problema ma piuttosto studiare l'evento complementare:

$$A^c = \textit{"in classe } \underline{\textit{non}} \textit{ ci sono due studenti nati lo stesso giorno"}.$$

Identifichiamo ogni studente con il giorno dell'anno in cui è nato: dunque ad ogni studente è associato un numero da 1 a 365 e, se due o più studenti sono nati lo stesso giorno, questi avranno tutti lo stesso numero. Un evento elementare è quindi una possibile associazione (con eventuali ripetizioni) dei numeri da 1 a 365 agli s studenti; in altre parole è una sequenza ordinata, con eventuali ripetizioni, di s oggetti scelti fra $n = 365$. Pertanto gli eventi elementari, cioè i casi possibili, sono $D^*(n, s) = n^s$. Se vogliamo calcolare la probabilità dell'evento A^c, i casi favorevoli (ovvero sfavorevoli per A) sono le disposizioni *senza* ripetizioni che sono $D(n, s) = n(n-1)\cdots(n-s+1)$. Pertanto

$$P(A^c) = \frac{D(n,s)}{D^*(n,s)} == \frac{n(n-1)\cdots(n-s+1)}{n^s}$$

e dunque (ricordando la (1) della proposizione (1.8))

$$P(A) = 1 - \frac{n(n-1)\cdots(n-s+1)}{n^s}.$$

Notiamo che la probabilità di A è funzione $p(s)$ del numero degli studenti. Calcolando questa funzione per alcuni valori di s (in figura 1.6 riportiamo i valori di $P(A)$ per s che va da 1 a 40) ci accorgiamo che $p(s)$ supera il 50% (cioè 1/2) non appena $s = 23$. Dunque, basta avere una classe di 23 studenti

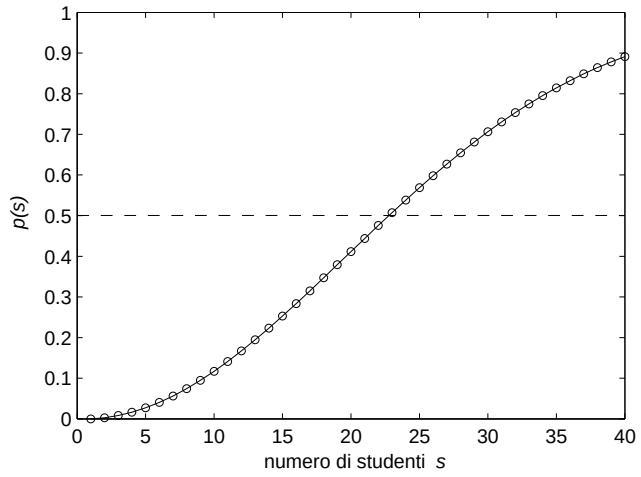


Figura 1.6: La probabilità $p(s)$ che almeno due studenti siano nati lo stesso giorno in funzione del numero degli studenti s. Notiamo che $p(22) < 0.5$ mentre $p(23) > 0.5$. Notiamo anche che $p(1) = 0$ e che $p(s) \to 1$ al crescere di s (per $s > 365$ si avrà, chiaramente, $p(s) = 1$).

affinché la probabilità che almeno due festeggino insieme il compleanno sia più del 50%. Il nome "paradosso dei compleanni" è dovuto al fatto che questo è un risultato abbastanza anti-intuitivo (infatti ci si aspetterebbe un numero più grande di 23). □

1.6 Indipendenza e probabilità condizionata

Definizione 1.16. Sia $P : S(\Omega) \to [0,1]$ una misura di probabilità. Due eventi $A_1, A_2 \in S(\Omega)$ si dicono *indipendenti* se

$$P(A_1 \cap A_2) = P(A_1)\,P(A_2) \tag{1.9}$$

cioè se la probabilità che si verifichino entrambi è il prodotto delle probabilità che ciascuno si verifichi separatamente. □

Sottolineiamo il fatto che l'indipendenza non è una proprietà "insiemistica" degli eventi (come ad esempio l'incompatibilità) ma è una proprietà legata alla misura di probabilità. Riprendendo l'esempio 1.2 del lancio di un dado, consideriamo gli eventi

$$\begin{aligned} A_1 &= \textit{esce un numero pari}, & P(A_1) &= 1/2, \\ A_2 &= \textit{esce un numero maggiore di 3}, & P(A_2) &= 1/2. \end{aligned}$$

Questi *non* sono indipendenti; infatti

$$P(A_1 \cap A_2) = P(\{4,6\}) = \frac{1}{3} \neq \frac{1}{4} = P(A_1)\,P(A_2)\,.$$

Riprendendo invece l'esempio 1.3 del lancio di due dadi, consideriamo gli eventi

$$\begin{aligned} B_1 &= \textit{esce un numero pari sul dado rosso}, & P(B_1) &= 1/2, \\ B_2 &= \textit{esce un numero maggiore di 3 sul dado blu}, & P(B_2) &= 1/2, \end{aligned}$$

Stavolta si ha

$$P(B_1 \cap B_2) = P\left(\left\{ \begin{array}{ccc} (2,4) & (2,5) & (2,6) \\ (4,4) & (4,5) & (4,6) \\ (6,4) & (6,5) & (6,6) \end{array} \right\}\right) = \frac{1}{4} = P(B_1)\,P(B_2)$$

e quindi gli eventi sono indipendenti. Almeno in questi esempi, la definizione matematica di indipendenza conferma la nozione intuitiva che ne abbiamo: gli eventi A_1 e A_2 non sono indipendenti poiché fra le facce maggiori di 3 ci sono più facce pari (4 e 6) che non sulle facce minori o uguali a 3 (solo il 2); invece B_1 e B_2 sono indipendenti poiché quello che succede sul dado rosso non dipende (è, appunto, indipendente) da quello che succede sul dado blu.

Osservazione 1.17. Due eventi incompatibili, cioè disgiunti, non sono mai indipendenti, a meno che uno dei due non abbia probabilità nulla: il fatto stesso di escludersi vicendevolmente fa sì che il verificarsi dell'uno dipenda dal verificarsi o meno dell'altro. Ciò è confermato dalla definizione 1.16: se A_1 e A_2 sono incompatibili, cioè $A_1 \cap A_2 = \emptyset$, si ha $P(A_1 \cap A_2) = P(\emptyset) = 0$, che è uguale a $P(A_1)\,P(A_2)$ se e solo se una delle due probabilità è nulla. □

Strettamente legato al concetto di indipendenza è il concetto di probabilità condizionata.

Definizione 1.18. Sia $P : S(\Omega) \to [0,1]$ una misura di probabilità e sia $B \in S(\Omega)$ un evento tale che $P(B) \neq 0$. Per ogni $A \in S(\Omega)$ si definisce *probabilità di A condizionata a B* il numero

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}. \tag{1.10}$$

□

Intuitivamente, $P(A \mid B)$ è *la probabilità che si verifichi A sapendo che si è verificato B*. Lasciamo per esercizio la dimostrazione della seguente proposizione.

Proposizione 1.19. *Sia $P : S(\Omega) \to [0,1]$ una misura di probabilità e sia $B \in S(\Omega)$ un evento tale che $P(B) \neq 0$. La probabilità condizionata $P(A \mid B)$ ha le seguenti proprietà:*

(i) la funzione $A \mapsto P(A \mid B)$, che ad ogni $A \in S(\Omega)$ associa il numero $P(A \mid B)$, è una misura di probabilità su $S(\Omega)$ (secondo la definizione 1.7);

(ii) $P(A \mid B) = P(A)$ se e solo se A e B sono indipendenti.

Osservazione 1.20. Se gli eventi elementari sono finiti ed equiprobabili (cfr. par. 1.3) allora $P(A \mid B)$ si può calcolare come

$$P(A \mid B) = \frac{\textit{casi favorevoli ad } A \textit{ e a } B}{\textit{casi favorevoli a } B},$$

ovvero, più precisamente,

$$P(A \mid B) = \frac{\textit{numero di elementi di } A \cap B}{\textit{numero di elementi di } B}.$$

□

Nel paragrafo 1.1 abbiamo accennato al fatto che in certi casi è più naturale attribuire una probabilità a eventi elementari a partire dalla probabilità di eventi composti, piuttosto che il contrario. Spesso per far ciò si utilizza un'ipotesi di indipendenza, come illustreremo nell'esempio seguente.

Esempio 1.21.
Ogni mattina il Sig. Rossi prende il bus alle 8 in punto. Sfortunatamente la fermata non ha pensilina per cui il Sig. Rossi arriva puntuale e, in caso di pioggia, spera nella puntualità del bus. Dopo un po' di anni di esperienza, il Sig. Rossi sa che il bus arriva puntuale il 75% delle volte. Inoltre quel giorno le previsioni indicano probabilità di pioggia al 20%. Domanda: qual'è la probabilità che il Sig. Rossi si bagni aspettando il bus in ritardo, ovvero la probabilità dell'evento *piove e il bus è in ritardo*?
L'insieme degli eventi elementari è il seguente:

bus in orario *non piove*	*bus in orario* *piove*
bus in ritardo *non piove*	*bus in ritardo* *piove*

Notiamo che gli eventi *"bus in orario"* e *"non piove"*, così come i loro complementari *"bus in ritardo"* e *"piove"*, sono eventi composti e sono rappresentati, rispettivamente, dalla prima riga, dalla prima colonna, dalla seconda riga e dalla seconda colonna della precedente tabella. Il Sig. Rossi può dunque attribuire una probabilità a eventi *composti*, esattamente

$$P(\textit{bus in orario}) = 3/4 \qquad P(\textit{non piove}) = 4/5$$
$$P(\textit{bus in ritardo}) = 1/4 \qquad P(\textit{piove}) = 1/5$$

e ha il problema di calcolarsi la probabilità dell'evento elementare

il bus è in ritardo e piove

intersezione di *"bus in ritardo"* e *"piove"* (rappresentato della casella in basso a destra). Tuttavia le informazioni che abbiamo finora *non sono sufficienti*: non c'è nessuna prescrizione per calcolare $P(A \cap B)$ conoscendo $P(A)$ e $P(B)$.
Una prima strada percorribile la si vede subito ed è quella di fare l'ipotesi di *indipendenza*. Se assumiamo che la pioggia e la puntualità del bus siano due fatti che non hanno alcuna influenza reciproca, possiamo usare la (1.9) e scrivere

$$P(\textit{bus in ritardo} \cap \textit{piove}) = P(\textit{bus in ritardo})\, P(\textit{piove}) = \frac{1}{4} \times \frac{1}{5} = \frac{1}{20}.$$

Nell'ipotesi di indipendenza possiamo calcolare allo stesso modo la probabilità di ciascun evento elementare e il risultato è:

3/5	3/20
1/5	1/20

D'altra parte, assumere l'indipendenza può non essere molto corretto. Infatti sappiamo benissimo che le condizioni metereologiche e la puntualità degli autobus non sono affatto indipendenti! Il Sig. Rossi non dovrebbe avere difficoltà ad accorgersi che la pioggia fa aumentare la probabilità di ritardo del bus. Con un po' di osservazioni si accorge che *quando piove, la probabilità di ritardo del bus è del 50%*; questa è precisamente una probabilità condizionata:

$$P(\textit{bus in ritardo} \mid \textit{piove}) = 1/2.$$

Pertanto, usando la definizione (1.10), si ottiene

$$P(\textit{bus in ritardo} \cap \textit{piove}) = P(\textit{bus in ritardo} \mid \textit{piove})\, P(\textit{piove})$$
$$= \frac{1}{2} \times \frac{1}{5} = \frac{1}{10}$$

Abbandonando l'ipotesi irrealistica di indipendenza la probabilità dell'evento più sfortunato è aumentata (Legge di Murphy!). Le probabilità degli altri eventi elementari le posso calcolare così: indicando con X, Y e Z le restanti probabilità incognite

X	Y
Z	1/10

otterremo il sistema in tre incognite

$$\begin{cases} Z + \dfrac{1}{10} = P(\textit{bus in ritardo}) = \dfrac{1}{4} \\ X + Z = P(\textit{non piove}) = \dfrac{4}{5} \\ Y + \dfrac{1}{10} = P(\textit{piove}) = \dfrac{1}{5} \end{cases}$$

che, risolto, ci dà

13/20	1/10
3/20	1/10

.

□

Esempio 1.22.: Estrazioni con reimbussolamento
Un'urna contiene tre palline bianche e una pallina rossa. Estraiamo a caso una pallina dall'urna, poi rimettiamola dentro ("reimbussolamento") e facciamo una seconda estrazione. Qual'è la probabilità di estrarre una pallina bianca alla prima e rossa alla seconda estrazione?
Lo spazio degli eventi elementari è dato da tutti i possibili esiti delle due estrazioni:

$$\begin{array}{cc} (\mathrm{B},\mathrm{B}) & (\mathrm{B},\mathrm{R}) \\ (\mathrm{R},\mathrm{B}) & (\mathrm{R},\mathrm{R}) \end{array}$$

dove, ad esempio, la coppia ordinata (B,R) sta per "pallina bianca alla prima estrazione e rossa alla seconda estrazione". Lo spazio degli eventi elementari si può anche rappresentare con il diagramma "ad albero" di fig. 1.7 dove, ad esempio, l'evento (B,R) corrisponde al percorso evidenziato in neretto. Sarà utile anche introdurre la notazione $(\mathrm{B},*)$ per indicare l'evento composto *pallina bianca alla prima estrazione*, ovvero

$$(\mathrm{B},*) = \big\{(\mathrm{B},\mathrm{B}),\ (\mathrm{B},\mathrm{R})\big\}.$$

La rappresentazione di questo evento nel diagramma ad albero è data dall'unione di tutti i percorsi che hanno una pallina bianca come prima estrazione ed è perciò quella evidenziata in fig. 1.8. Similmente, $(*,\mathrm{R})$ indicherà l'evento *pallina rossa alla seconda estrazione*, e così via. Calcoliamo ora la probabilità dell'evento

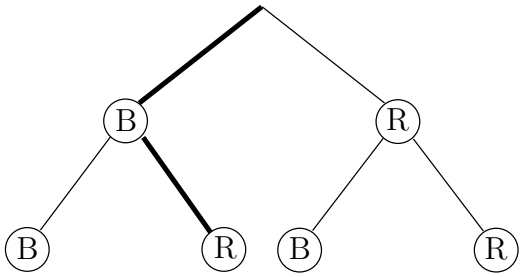


Figura 1.7: Diagramma degli eventi elementari relativi a due estrazioni di palline dall'urna. In neretto il percorso corrispondente all'evento (B,R).

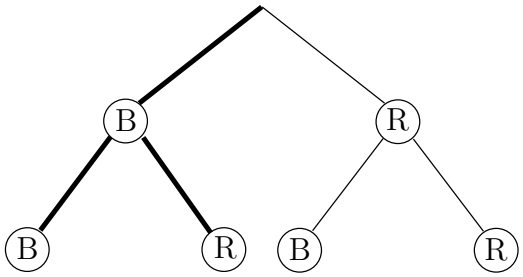


Figura 1.8: L'evento $(\mathrm{B}, *)$.

(B,R). La probabilità di estrarre una pallina bianca alla prima estrazione è, ovviamente, 3/4:

$$P((\mathrm{B}, *)) = 3/4.$$

Per via del reimbussolamento la seconda estrazione avviene esattamente nelle medesime condizioni della prima e quindi la probabilità di estrarre una pallina rossa alla seconda estrazione è 1/4:

$$P((*, \mathrm{R})) = 1/4.$$

Le due estrazioni sono completamente indipendenti (la prima non ha influenza sulla seconda e tantomeno il viceversa) e quindi

$$P((\mathrm{B}, \mathrm{R})) = P((\mathrm{B}, *))\, P((*, \mathrm{R})) = 3/16.$$

Possiamo schematizzare questa procedura come in figura 1.9: l'indipendenza delle due estrazioni implica che è sufficiente moltiplicare i due numeri che si incontrano su un percorso per ottenere la probabilità di quel percorso. È facile così verificare che le probabilità dei quattro eventi elementari sono le seguenti:

$$P((\mathrm{B}, \mathrm{B})) = 9/16 \qquad P((\mathrm{B}, \mathrm{R})) = 3/16$$
$$P((\mathrm{R}, \mathrm{B})) = 3/16 \qquad P((\mathrm{R}, \mathrm{R})) = 1/16.$$

□

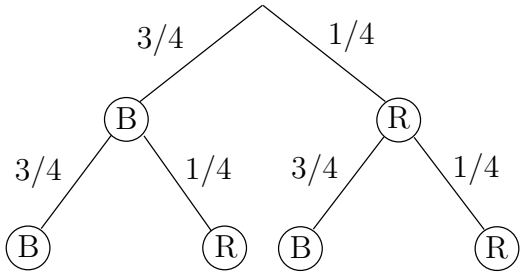


Figura 1.9:

Esempio 1.23.: Estrazioni senza reimbussolamento
Rimettiamoci nella situazione dell'esempio precedente ma stavolta supponiamo di *non* riporre la pallina nell'urna dopo la prima estrazione. Allora questa volta le due estrazioni non sono più indipendenti perché le probabilità della seconda estrazione dipendono da cosa è avvenuto nella prima. Se ad esempio abbiamo estratto una pallina bianca alla prima estrazione, per la seconda estrazione avremo probabilità $2/3$ di estarre una B e $1/3$ di estarre una R. Si tratta esattamente di *probabilità condizionate*:

$$P((*,\mathrm{B}) \mid (\mathrm{B},*)) = 2/3, \quad P((*,\mathrm{R}) \mid (\mathrm{B},*)) = 1/3.$$

Per calcolare $P((\mathrm{B},\mathrm{R}))$ basterà quindi usare la (1.10):

$$P((\mathrm{B},\mathrm{R})) = P((\mathrm{B},*))\,P((*,\mathrm{R}) \mid (\mathrm{B},*)) = \frac{3}{4} \times \frac{1}{3} = \frac{1}{4}.$$

Schematizziamo questa procedura in fig. (1.10): i numeri (e il loro significato) sono cambiati ma per calcolare la probabilità di un percorso si deve sempre moltiplicare i due numeri che si incontrano sul percorso stesso. In questo modo

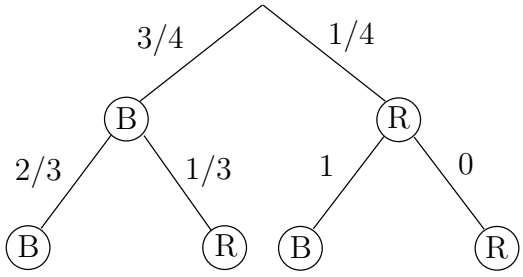


Figura 1.10:

si ottengono le seguenti probabilità per i quattro eventi elementari:

$$P((\mathrm{B},\mathrm{B})) = 1/2 \qquad P((\mathrm{B},\mathrm{R})) = 1/4$$
$$P((\mathrm{R},\mathrm{B})) = 1/4 \qquad P((\mathrm{R},\mathrm{R})) = 0.$$

□

Esercizio 1.24. *In un'urna ci sono 3 palline bianche, 2 rosse e 4 gialle. Effettuando 2 estrazioni senza reimbussolamento calcolare la probabilità che alla seconda estrazione si ottenga una pallina rossa. Rispondere poi alla stessa domanda nel caso con reimbussolamento.* □

1.7 Il teorema di Bayes

I concetti di indipendenza e di probabilità condizionata trovano un'interessante applicazione nel cosiddetto *Teorema di Bayes* (dovuto al matematico inglese Thomas Bayes, 1702-1761).

Definizione 1.25. Sia Ω un insieme qualsiasi. Una *partizione* (finita) di Ω è un insieme di sottoinsiemi $A_1, A_2 \dots A_n$ di Ω tali che

(i) $A_i \cap A_j = \emptyset$ per $i \neq j$

(ii) $A_1 \cup A_2 \cup \cdots \cup A_n = \Omega$

(figura 1.11). □

Ad esempio, se A è un qualunque sottoinsieme di Ω, A e A^c formano una partizione di Ω.

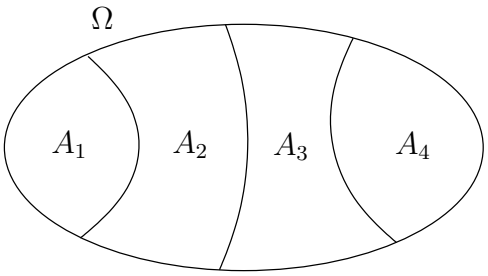


Figura 1.11: Una (quadri)partizione di un insieme Ω

Teorema 1.26. (di Bayes)
Sia $P : S(\Omega) \to [0,1]$ una misura di probabilità e sia $A_1, A_2, \dots A_n \in S(\Omega)$ una partizione di Ω tale che $P(A_i) \neq 0$ per ogni i. Allora, per ogni $B \in S(\Omega)$ tale che $P(B) \neq 0$ si ha

$$P(A_k \mid B) = \frac{P(B \mid A_k)\, P(A_k)}{\sum_{i=1}^{n} P(B \mid A_i)\, P(A_i)}, \tag{1.11}$$

per ogni $k = 1, 2, \cdots n$.

Prima di scrivere la (semplice) dimostrazione di questo teorema, osserviamo che l'interesse della formula (1.11) ("formula di Bayes") consiste nell'esprimere le probabilità condizionate $P(A_k \mid B)$ in termini delle probabilità condizionate

$P(B \mid A_k)$. Questo può essere utile in molte circostanze, come cercheremo di spegare nei successivi esempi.

Dimostrazione Poiché $\bigcup_{i=1}^{n} A_i = \Omega$ possiamo scrivere $B = \bigcup_{i=1}^{n} B \cap A_i$ e quindi, poiché gli A_i sono disgiunti, si ha $P(B) = \sum_{i=1}^{n} P(B \cap A_i)$. Dunque, usando ripetutemante la (1.10), si ottiene

$$P(A_k \mid B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(B \mid A_k)\, P(A_k)}{\sum_{i=1}^{n} P(B \cap A_i)} = \frac{P(B \mid A_k)\, P(A_k)}{\sum_{i=1}^{n} P(B \mid A_i)\, P(A_i)}.$$

□

Vediamo adesso alcuni di esempi di utilizzo della formula di Bayes che ci aiuteranno a capirne il significato.

Esempio 1.27.: Falsi positivi nei test medici
Si tratta di un classico esempio di applicazione della formula di Bayes. Supponiamo di sottoporre un paziente a un test per determinare se ha una certa malattia. Supponiamo di sapere, da prove sperimentali, che il test dà un risultato positivo corretto (cioè in effettiva presenza della malattia) nel 99% dei casi e che dà un risultato negativo corretto (cioè in assenza della malattia) nel 93% dei casi. Supponiamo anche di sapere che la malattia ha un'incidenza dello 0.2% nella popolazione (possiamo interpretare questo dato come la probabilità che il paziente abbia la malattia). Siamo interessati a determinare *la probabilità che il paziente sia sano anche se il il test ha dato un risutato positivo*, ovvero di quello che si chiama un "falso positivo".
Sia A l'evento *"il paziente ha la malattia"* e sia B l'evento *"il test è positivo"*. I nostri dati sono:

$$P(A) = 0.002, \qquad P(B \mid A) = 0.99, \qquad P(B^c \mid A^c) = 0.93,$$

mentre la probabilità che vogliamo calcolare è $P(A^c \mid B)$. Poiché A e A^c formano una (bi)partizione[3] di Ω si può scrivere la formula di Bayes nella seguente forma

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B \mid A)\, P(A) + P(B \mid A^c)\, P(A^c)}. \tag{1.12}$$

Osservando che $P(B \mid A^c) = 1 - P(B^c \mid A^c) = 0.07$, abbiamo tutti i dati che ci servono per poter usare la formula precedente. Si ottiene così

$$P(A \mid B) = \frac{0.99 \times 0.002}{0.99 \times 0.002 + 0.07 \times 0.998} \approx 0.028$$

e dunque

$$P(A^c \mid B) = 1 - P(A \mid B) \approx 0.97.$$

[3]Non abbiamo descritto esplicitamente Ω ma è chiaro che si tratta di un insieme dello stesso tipo di quelli incontrati finora:

paziente malato *test positivo*	*paziente sano* *test positivo*
paziente malato *test negativo*	*paziente sano* *test negativo*

Notiamo tuttavia che stavolta non è degli eventi elementari che ci interessa la probabilità.

La morale è che, nonostante la buona affidabilità del test, se la malattia è rara il falso positivo è probabile. Questo risultato non significa che il test sia inutile: rispetto alla probabilità "a priori" di avere la malattia, cioè $P(A) = 0.2\%$, dopo il test la probabilità è diventata "a posteriori" $P(A \mid B) \approx 2.8\%$, ovvero 14 volte maggiore.[4] Semmai, questo risultato incoraggia a ripetere i test medici più di una volta. Se ad esempio immaginiamo di ripetere il test e di trovarlo ancora positivo, la nuova probabilità a posteriori può essere valutata prendendo come nuova probabilità a priori la precedente probabilità a posteriori (cioè usando $P(A) = 0.028$) e ottenendo così:

$$P(A \mid B) = \frac{0.99 \times 0.028}{0.99 \times 0.028 + 0.07 \times 0.972} \approx 0.29.$$

Se iterassimo ancora una volta il procedimento, e il test risultasse ancora positivo, otterremmo $P(A \mid B) \approx 0.85$. □

Come secondo esempio di utilizzo della formula di Bayes risolveremo il famosissimo *paradosso dei prigionieri*.

Esempio 1.28.: Paradosso dei tre prigionieri
Tre prigionieri, che chiameremo A, B e C, sono condannati a morte e attendono la loro sorte rinchiusi in tre celle separate. Ma la notte prcedente l'esecuzione il Re decide che uno dei tre prigionieri sarà graziato. I prigionieri, che sono all'oscuro del nome del graziato, sanno quindi di avere 1/3 di probabilità ciascuno di essere il fortunato che avrà salva la vita. Quella notte il secondino, che non può rivelare a nessun prigioniero la sua sorte, parla col prigioniero A e gli rivela che B *è condannato*. Allora il prigioniero A si sente un po' sollevato perché pensa che la sua probabilità di salvezza sia aumentata al 50%. Domanda: *ha ragione il prigioniero A a credere che le sue possibilità di salvezza siano aumentate?*
Il modo corretto di impostare la soluzione del dilemma è mediante l'uso della formula di Bayes. Consideriamo gli eventi

$$\begin{aligned} A &:= \textit{il prigioniero } A \textit{ è graziato}, \\ B &:= \textit{il prigioniero } B \textit{ è graziato}, \\ C &:= \textit{il prigioniero } C \textit{ è graziato}, \\ S_B &:= \textit{il secondino rivela che } B \textit{ è condannato}. \end{aligned}$$

Notiamo che, a priori, tutti e tre i prigionieri hanno la stessa probabilità di essere graziati:

$$P(A) = P(B) = P(C) = 1/3,$$

ma quello che veramente interessa al prigioniero A è la probabilità a posteriori

$$P(A \mid S_B)$$

[4] La scelta che abbimo fatto della probabilità "a priori" $P(A)$ come percentuale di incidenza della malattia in tutta la popolazione, può essere discutibile. In molti casi $P(A)$ sarà valutata soggettivamente dal medico in base ad altri fattori e alla sua esperienza personale. Resta comunque il fatto che, una volta scelta la probabilità $P(A)$, la formula di Bayes può essere usata "per correggere il tiro", come visto nell'esempio.

cioè la probabilità di essere graziato sapendo che il secondino gli ha rivelato che B sarà condannato. La formula di Bayes ci dice che

$$P(A \mid S_B) = \frac{P(S_B \mid A)\,P(A)}{P(S_B \mid A)\,P(A) + P(S_B \mid B)\,P(B) + P(S_B \mid C)\,P(C)}$$
$$= \frac{P(S_B \mid A)}{P(S_B \mid A) + P(S_B \mid B) + P(S_B \mid C)},$$

dove abbiamo tenuto conto del fatto che $P(A) = P(B) = P(C)$. È quindi evidente che il problema si riduce a valutare le probabilità condizionate $P(S_B \mid A)$, $P(S_B \mid B)$, $P(S_B \mid C)$. Poiché il secondino *non può far conoscere ad A la sua sorte*, allora ha senso porre

$$P(S_B \mid A) = 1/2, \qquad P(S_B \mid B) = 0, \qquad P(S_B \mid C) = 1,$$

e dalla formula di Bayes si ottiene:

$$P(A \mid S_B) = 1/3.$$

Dunque, la probabilità di grazia per A, contrariamente a quello che si può a prima vista pensare, *non è aumentata*. Se usiamo di nuovo la formula di Bayes, stavolta per calcolare la probabilità a posteriori di C, ottieniamo invece

$$P(C \mid S_B) = 2/3.$$

Perciò la probabilità di grazia del prigioniero C è davvero aumentata dopo la rivelazione del secondino. Il contenuto intuitivo di questi risultati è il seguente: la rivelazione del secondino, in realtà, non aggiunge nessuna informazione significativa per il prigioniero A, infatti *chiaramente* uno dei due prigionieri rimanenti è condannato e sapere se si tratti di B o di C non è rilevante. Invece, dal punto di vista di C, l'informazione ottenuta da A è rilevante poiché, non essendo il secondino vincolato a non rivelare la sorte di C, il fatto che il secondino abbia detto B e non C "rafforza" la posizione di quest'ultimo.

Questo paradosso ha un'altra versione, nota come *dilemma di Monty Hall*. Monty Hall era il conduttore di "Let's make a deal", un gioco a premi televisivo americano: dietro una di tre porte chiuse, A, B e C, c'è un'auto che il concorrente vince se indovina la porta giusta. Il concorrente sceglie la porta A ma prima di aprirla il conduttore lo ferma, gli apre la porta B, che è vuota, e gli propone una scelta: continuare a puntare sulla porta A o cambiare e scegliere la porta C? Voi cosa avreste fatto? □

Esercizio 1.29. *Ho tre scatole identiche: una contiene una pallina bianca e una nera, le altre due contengono due palline nere ciascuna. Lo scopo del gioco è individuare la scatola con la pallina bianca. Supponiamo di aver scelto una scatola e di aver estratto da questa una pallina nera. Se adesso mi vengono prospettate due possibilità:*

- *guardare l'altra pallina (puntare sulla scatola già scelta);*
- *scegliere un'altra scatola (puntare su un'altra scatola);*

cosa mi conviene fare? □

Capitolo 2

Variabili aleatorie semplici

2.1 Il concetto di variabile aleatoria

Nel capitolo precedente abbiamo incontrato molti eventi probabilistici di tipo numerico. Ci sono diverse situazioni in cui si ha a che fare con *numeri aleatori* (ossia "casuali"):

- gli eventi elementari possono essere essi stessi dei numeri, come nel caso del lancio di un dado (esempio 1.2), oppure coppie o n-uple di numeri, come nel caso del lancio di due dadi (esempio 1.3, della corsa dei 100 m (esempio 1.6) e del paradosso dei compleanni (esempio 1.15);

- in altri casi può convenire identificare con numeri gli eventi elementari; ad esempio, giocando a testa o croce come nell'esempio 1.1, possiamo associare 1 all'evento T e 0 all'evento C;

- in generale capita spesso di considerare numeri che dipendono da eventi; l'esempio più tipico è quallo di una vincita in denaro: un numero (la cifra che guadagno/perdo) che dipende dagli eventi di un sistema probabilistico (una schedina del totocalcio, una corsa di cavalli, una partita a dadi o a poker).

D'altra parte il lavorare coi numeri permette di spingersi più a fondo nell'analisi matematica dei sistemi probabilistici.
Per tutte queste ragioni si introduce il concetto di *variabile aleatoria*.

Definizione 2.1. Sia $P : S(\mathbb{R}) \to [0, 1]$ una misura di probabilità sull'insieme dei numeri reali $\mathbb{R}$. Allora in questo caso l'evento elementare $X \in \mathbb{R}$ si chiama *variabile aleatoria* ("semplice" o "scalare") e la misura di probabilità P si chiama *legge* o *distribuzione* della variabile aleatoria X.[1] □

Intuitivamente, una variabile aleatoria X è un numero che non è conosciuto con certezza ma che ha una certa probabilità di avere un certo valore (un sinonimo di "variabile aleatoria" è, difatti, "numero casuale").

[1] Nella letteratura più rigorosa le definizioni di variabile aleatoria e di distribuzione sono un po' diverse ma, per i nostri scopi, la definizione 2.1 sarà più che soddisfacente.

D'ora in poi abbrevieremo spesso il termine "variabile aleatoria" con "v.a.". Conviene inoltre introdurre una notazione snella per la probabilità degli eventi composti: se A è un sottoinsieme generico di $\mathbb{R}$ scriveremo $P(X \in A)$ anziché $P(A)$ e, nello stesso spirito, scriveremo semplicemente

$$P(X = a)\,, \quad P(X \le a)\,, \quad P(a \le X < b)\,, \quad \text{etc.}$$

al posto di

$$P(\,\{a\}\,)\,, \quad P(\,(-\infty, a]\,)\,, \quad P(\,[a, b)\,)\,, \quad \text{etc.}$$

Esempio 2.2.
Ricordiamo le misure di probabilità introdotte negli esempi 1.1 e 1.2. Nel primo caso

$$\Omega_2 = \{T, C\}\,, \qquad P(T) = P(C) = 1/2$$

e nel secondo

$$\Omega_2 = \{1, 2, 3, 4, 5, 6\}\,, \qquad P(1) = P(2) = \cdots P(6) = 1/6.$$

Costruiamo due v.a., che indicheremo con X e Y, ponendo

$$X = 1, \text{ se sulla moneta esce } T, \qquad X = 0, \text{ sesulla moneta esce } C$$

$$Y = 1 \text{ se sul dado esce 1, 3 o 5}, \qquad Y = 0 \text{ se sul dado esce 2, 4 o 6}$$

(si veda la figura 2.1).

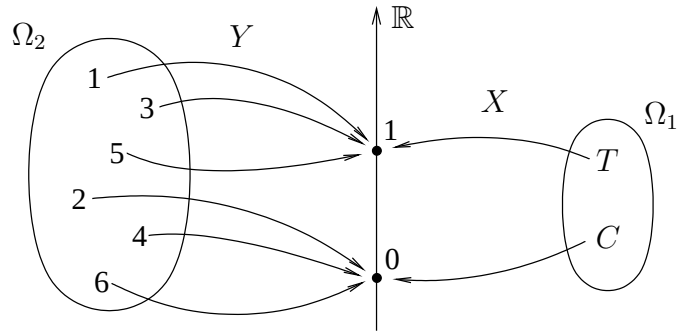


Figura 2.1: Le variabili aleatorie X e Y.

Dunque le leggi di X è di Y sono date da

$$P(X = 0) = P(X = 1) = 1/2, \qquad P(Y = 0) = P(Y = 1) = 1/2.$$

Notiamo che X e Y hanno la stessa legge. In questo caso si dice che le due v.a. sono *identicamente distribuite*. □

L'esempio precedente ci mostra due v.a. X e Y nate in contesti diversi ma con la stessa distrubuzione: giocare a testa o croce con una moneta oppure a pari o dispari con un dado è sostanzialmente la stessa cosa. Il concetto di distribuzione di una v.a. ci permette, in questo caso come in molti altri, di astrarre dal contesto specifico e concentrarci su una struttura matematica, che è racchiusa nella *distribuzione*.

2.2 Funzione di ripartizione e quantili

Un modo conveniente di visualizzare la distribuzione di una variabile aleatoria semplice X è quello di ricorrere alla sua *funzione di ripartizione*, detta anche *distribuzione cumulativa*. Questa è una funzione $F_X : \mathbb{R} \to [0,1]$ così definita:

$$F_X(x) := P(X \leq x)\,. \tag{2.1}$$

Costruiamo ad esempio la funzione di ripartizione della v.a. X dell'esempio 2.2 (che sarà identica a quella della Y, per quanto appena osservato). Se $x < 0$ si ha chiaramente $F_X(x) = 0$ in quanto la retroimmagine della semiretta $(-\infty, x)$ è l'insieme vuoto. Se $0 \leq x < 1$ si ha $F_x(0) = 1/2$ poiché in questo caso la semiretta $(-\infty, x)$ contiene sempre lo 0 e dunque la retroimmagine è C (croce) che ha probabilità $1/2$. Infine, se $x \geq 1$ si ha $F_x(0) = 1$ poiché adesso la semiretta $(-\infty, x)$ contiene sempre sia 0 che 1 e dunque la sua retroimmagine è $\{T, C\}$ che ha probabilità 1. Il grafico di F_X è riportato in figura 2.2; notiamo che la funzione è continua da destra ovvero $\lim_{x\to 0^+} F_X(x) = 1/2 = F_X(1/2)$ e $\lim_{x\to 1^+} F_X(x) = 1 = F_X(1)$.

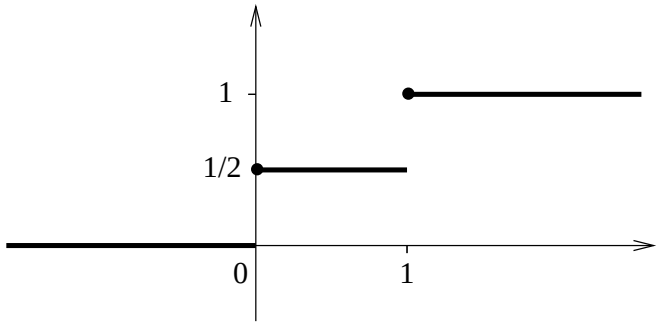


Figura 2.2: Funzione di ripartizione delle v.a. X e Y dell'esempio 2.2. La grafica vuole evidenziare il fatto che la funzione è continua da destra.

Come ulteriore esempio possiamo considerare la variabile aleatoria "completa" del lancio del dado,[2] in cui $\Omega = \{1,2,3,4,5,6\}$ e $X(k) = k$ per $k = 1, 2, \ldots, 6$; il grafico della sua funzione di ripartizione è disegnato in figura 2.3.

Proposizione 2.3. *Ogni funzione di ripartizione F_X ha le seguenti proprietà:*

(i) *è non-decrescente;*

(ii) $\lim_{x\to-\infty} F_X(x) = 0$*;*

(iii) $\lim_{x\to+\infty} F_X(x) = 1$*;*

(iv) *è continua da destra, ovvero* $\lim_{x\to x_0^+} F_X(x) = F_X(x_0)$.

Omettiamo la dimostrazione di queste proprietà, peraltro abbastanza intuitive. Il lettore interessato può consultare il libro di Dall'Aglio [3].

[2] Notiamo per inciso che il termine "aleatorio" viene dal latino *alea* che significa proprio "dadi" (*Alea iacta est...*).

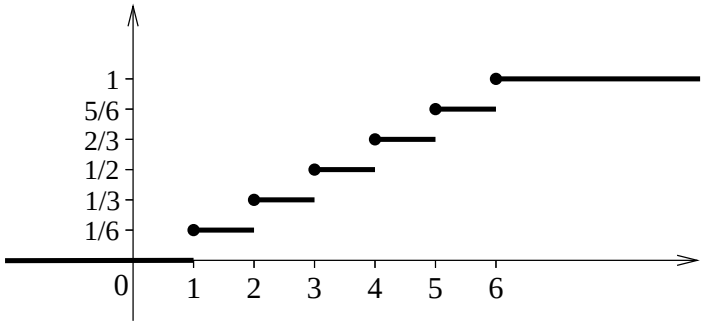


Figura 2.3: Funzione di ripartizione della v.a. del lancio del dado.

Se la funzione di ripartizione è *invertibile*, ovvero se per ogni $p \in [0,1]$ esiste $x \in \mathbb{R}$ tale che $F_X(x) = p$, tale x è detta il *quantile di ordine p della v.a. X* e viene indicata con $x = q_p^X$. Dunque il quantile di ordine p è quella x tale che la probabilità che X sia minore di x è proprio p; in formule:

$$F_X(q_p^X) = p, \quad \text{ovvero} \quad P\left(X \leq q_p^X\right) = p \tag{2.2}$$

(si veda la figura 2.4).

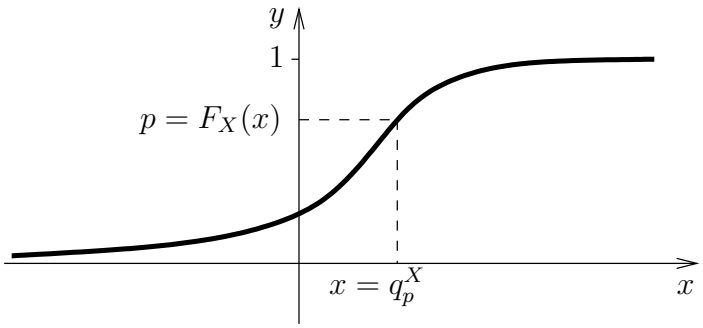


Figura 2.4: Corrisponenza fra la probabilità $p = F_X(x)$ e il quantile $x = q_p^X$.

Naturalmente, non tutte le funzioni di ripartizione sono invertibili (per esempio quelle delle figure 2.2 e 2.3 non lo sono), poiché ci possono essere valori di p cui non corrisponde nessuna x o valori di p cui non corrisponde una sola x. Ad esempio, nel grafico di figura 2.2, non c'è nessuna x per cui $F_X(x) = 1/3$ e a $p = 1/2$ corrispondono tutte le x nell'intervallo $[0,1)$. Tuttavia, come vedremo nel prossimo paragrafo, per distribuzioni *continue* con densità strettamente positiva la funzione di ripartizione è sempre invertibile e quindi i quantili sono sempre ben definiti

2.3 Variabili aleatorie discrete e continue

D'ora in avanti ci concentreremo su due importanti categorie di variabili aleatorie semplici: quelle *discrete* e quelle *continue*.

Definizione 2.4. Una variabile aleatoria X si dice *discreta* se assume un insieme numerabile di valori $\{x_1, x_2, x_3, \ldots\}$ con probabilità $\{p_1, p_2, p_3, \ldots\}$, ovvero

$$P(X = x_k) = p_k, \qquad k = 1, 2, 3, \ldots,$$

dove $p_k \in [0, 1]$ e $\sum_k p_k = 1$. □

Notiamo quindi che la probabilità di una v.a. discreta è concentrata su un insieme discreto di valori. Questo può essere finito, $\{x_1, x_2, x_3, \ldots, x_n\}$, o infinito $\{x_1, x_2, x_3, \ldots\}$. Nel primo caso la v.a. si dice *discreta finita* e la condizione sulle probabilità p_k è espressa da una semplice sommatoria:

$$\sum_{k=1}^{n} p_k = 1. \tag{2.3}$$

Nel secondo caso la v.a. si dice *discreta infinita* e la condizione sulle probabilità p_k è espressa da una *serie*:

$$\sum_{k=1}^{\infty} p_k = 1. \tag{2.4}$$

Quello discreto è un caso limite (ancorché molto importante) di v.a. in cui la probabilità è tutta concentrata in punti isolati della retta reale. All'estremo opposto si trova l'altro caso limite (anch'esso della massima importanza) di v.a. continue, che sono "spalmate" su un continuo di valori della retta reale. Per esse ogni singolo punto isolato ha probabilità nulla mentre ha più senso chiedersi qual'è la probabilità che la v.a. stia in un intervallo di valori. La definizione rigorosa è la seguente.

Definizione 2.5. Una variabile aleatoria X si dice *(assolutamente) continua* se esiste una funzione integrabile $f_X : \mathbb{R} \to [0, +\infty)$ tale che

$$P(a \le X \le b) = \int_a^b f_X(x)\, dx, \tag{2.5}$$

per ogni $-\infty < a \le b < +\infty$. La funzione f_X si chiama *densità* della variabile aleatoria X. □

Graficamente, la densità di una v.a. continua X è una funzione f_X tale che la probabilità che X cada tra a e b è pari all'area sotto il grafico di f_X (figura 2.5) È chiaro che la funzione di densità f_X deve soddisfare la condizione

$$\int_{-\infty}^{+\infty} f_X(x)\, dx = 1, \tag{2.6}$$

che è l'analogo continuo delle condizioni (2.3) e (2.4). Notiamo anche che la funzione di ripartizione di una v.a. continua X con densità f_X si può esprimere come

$$F_X(x) = \int_{-\infty}^{x} f_X(y)\, dy. \tag{2.7}$$

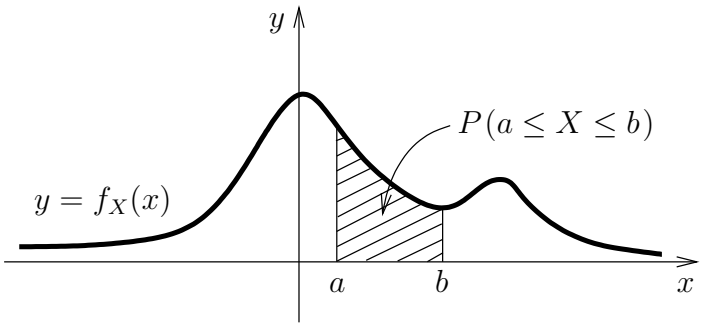


Figura 2.5:

Esempio 2.6.: Quando arriva la telefonata?
Il Sig. Rossi aspetta una telefonata dal Sig. Bianchi il quale ha preannunciato che chiamerà, in un istante non meglio precisato, fra le 16 e le 18. Il Sig. Rossi si deve assentare dalle 16:45 alle 17. Qualè la probabilità che la telefonata arrivi mentre il Sig. Rossi è assente?
L'istante della telefonata è una variabile aleatoria X. Per quanto ne sa il Sig. Rossi, tutti i momenti dalle 16 alle 18 sono equiprobabili mentre fuori da questo intervallo la probabilità è zero. Dunque è intuitivo considerare X come una v.a. continua la cui densità f_X ha valore costante α sull'intervallo $[16, 18]$ e sia zero fuori di esso. Quanto vale la costante α? Dev'essere tale da soddisfare la condizione di normalizzazione (2.6), ovvero tale che l'area del retatngolo con base $[16, 18]$ e altezza α sia 1, ed è perciò $\alpha = 1/2$. La densità di X è pertanto

$$f_X(t) = \begin{cases} 0, & \text{se } t < 16, \\ \frac{1}{2}, & \text{se } 16 \le t \le 18, \\ 0, & \text{se } t > 18, \end{cases}$$

(figura 2.6). Questa è la tipica distribuzione *uniforme* su un intervallo (si veda la sezione 3.5) che assicura che la probabilità che X stia in un certo intervallo di tempo $[t_1, t_2]$ è proporzionale all'ampiezza dell'intervallo stesso. Notiamo, in particolare, che la probabilità che la telefonata arrivi fra le 16:45 e le 17 è 1/8. La funzione di ripartizione è data da

$$F_X(t) = \int_{-\infty}^{t} f_X(s)\, ds = \begin{cases} 0, & \text{se } t < 16, \\ \frac{t-16}{2}, & \text{se } 16 \le t \le 18, \\ 1, & \text{se } t > 18 \end{cases}$$

(figura 2.6). □

Proposizione 2.7. *Una variabile aleatoria continua X ha le seguenti proprietà:*

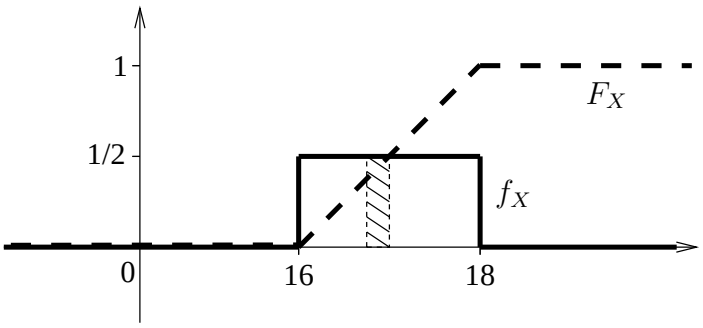


Figura 2.6: Densità (linea continua) e funzione di ripartizione (linea tratteggiata) della v.a. continua descritta nell'esempio 2.6. L'area tratteggiata corrisponde alla probabilità $P(16:45 \leq X \leq 17)$.

(i) $P(X = x) = 0$ *(la probabilità di un singolo punto è nulla);*[3]

(ii) la funzione di ripartizione F_X *è continua;*

inoltre, se la funzione di densità f_X *è continua in un punto* x*, si ha*

(iii) $f_X(x) = F'_X(x)$ *(la densità è la derivata della funzione di ripartizione).*

Dimostrazione La proprietà (i) segue immediatamente dalla definizione: infatti in questo caso l'intervallo di integrazione si riduce a un punto e perciò si ha

$$P(X = x) = \int_x^x f_X(y)\, dy = 0.$$

Le proprietà (ii) e (iii) seguono da teoremi noti sugli integrali (in particolare, la (iii) è il cosiddetto teorema fondamentale del calcolo integrale). □

Notiamo che la proprietà (ii) è strettamente legata alla (i); infatti, come si vede anche dalle figure 2.2 e 2.3, i salti della funzione di ripartizione sono in corrispondenza dei punti "pesanti", ovvero dei punti con probabilità non nulla.

Proposizione 2.8. *Se una v.a. continua* X *ha densità strettamente positiva, allora la sua funzione di ripartizione è invertibile e di conseguenza i quantili sono sempre ben definiti.*

Dimostrazione Da risultati elementari di analisi, [6], sappiamo che F_X è strettamente crescente se f_X è strettamente positiva e che una funzione strettamente crescente è sempre invertibile. □

[3] Il fatto che la probabilità di ogni singolo punto sia nulla potrebbe apparire contraddittorio con la proprietà $P(\cup_n A_n) = \sum_n P(A_n)$; tuttavia questa proprietà vale solo per unioni *numerabili*, mentre gli intervalli della retta reale sono unioni *non-numerabili* dei propri punti. Per lo stesso motivo non c'è contraddizione tra il fatto che i punti hanno misura nulla e gli intervalli, che sono unione di punti, hanno misura positiva: un intervallo è un'unione non-numerabile di punti!

Le v.a. discrete e quelle continue sono senza dubbio importanti e molto utili sia per la teoria che per le applicazioni, e nel capitolo 3 ne vedremo molti esempi. Questo non significa che non ci siano v.a. molto importanti che non rientrano completamente in nessuna delle due categorie (e che potremmo chiamare "variabili aleatorie miste"). Un esempio lampante che si trova in natura è dato dai livelli energetici degli atomi. Secondo le leggi della meccanica quantistica l'energia di un elettrone in un atomo non è una quantità determinata (né determinabile) a priori ma è piuttosto una variabile aleatoria. Risulta dalla teoria (e viene confermato dagli esperimenti) che questa v.a. ha una distribuzione in parte discreta e in parte continua (fig.2.7). La parte discreta corrisponde a li-

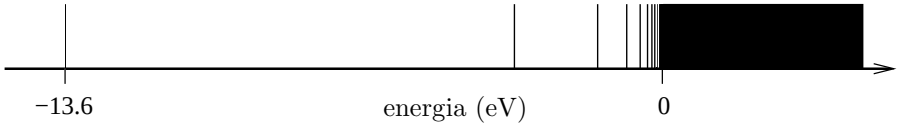


Figura 2.7: Lo "spettro" dell'atomo di idrogeno, ovvero i possibili valori che può assumere la v.a. "energia" dell'elettrone. Notare gli infiniti valori discreti che si accumulano nello 0 e il successivo continuo.

velli energetici E_n in cui l'elettrone è legato al nucleo. Tali livelli formano una successione infinita che si accumula nello 0 e precisamente

$$E_n = -\alpha/n^2, \qquad n = 1, 2, \ldots,$$

dove α vale circa 13.6 eV. La parte continua comprende tutte le energie da 0 a $+\infty$ e corrisponde all'elettrone che ha abbastanza energia da slegarsi e ionizzare l'atomo.

Nonostante che, come abbiamo appena visto, ci siano importanti v.a. che non sono né discrete né continue, in queste note ci occuperemo d'ora in avanti solo di v.a. discrete o continue.

2.4 Valore atteso di una variabile aleatoria

Definizione 2.9. Sia X una v.a. discreta e finita che assume i valori $x_1, \ldots, x_n$ con probabilità $p_1, \ldots, p_n$. Si definisce *valore atteso* (o *media* o *aspettazione* o *speranza matematica*) di X il numero

$$E[X] := x_1 p_1 + x_2 p_2 + \cdots + x_n p_n. \tag{2.8}$$

□

Come si vede, la definizione di $E[X]$ corrisponde all'idea intuitiva di "media pesata" dei possibili valori che può assumere la v.a. X, dove i "pesi" sono proprio le probabilità con cui si presentano tali valori.

Esempio 2.10.: Lotteria Italia
Quanto "vale", veramente, il biglietto di una lotteria? La risposta più naturale è senz'altro: *vale tanto quanto è il valore atteso della vincita*. Prendiamo ad esempio i dati relativi alla Lotteria Italia 2002/2003 (reperibili sul sito `www.lotteriaitalia.net`), riportati nella tabella (2.1). Essendo tutti i biglietti

Biglietti venduti: 17 949 331	Costo del biglietto: 3 €
Premi (in €)	
s_1	5 mln
s_2	2 mln
s_3	1 mln
s_4	800 000
s_5	700 000
s_6	600 000
s_7	500 000
s_8	400 000
s_9	300 000
s_{10}	200 000
da s_{11} a s_{40}	100 000
da s_{41} a s_{90}	50 000
Montepremi totale: $s = s_1 + s_2 + \cdots + s_{90} = 17$ mln	

Tabella 2.1: Dati relativi alla Lotteria Italia 2002/2003

equiprobabili, la probabilità di estrazione di un singolo biglietto è

$$p = \frac{1}{17\,949\,331} \approx 5.57 \times 10^{-8}.$$

Se ho acquistato un biglietto, la mia vincita è una variabile aleatoria discreta X che assume i valori $s_1, s_2, \ldots s_{90}$, ciascuno con probabilità p. Dunque, il valore atteso della vincita sarà

$$E[X] = \sum_{k=1}^{90} s_k p = sp = \frac{s}{\textit{biglietti venduti}} \approx 0.95\,€.$$

È evidente che il biglietto costa molto di più del suo vero "valore" (dal punto di vista strettamente probabilistico). Tolte le spese di organizzazione, resterà un certo guadagno netto per lo Stato (che è l'organizzatore della lotteria). Notiamo che, essendo il montepremi fissato fin da prima della vendita dei biglietti, il guadagno dello Stato è incerto e dipende da quanti biglietti sono venduti. Se ad esempio i biglietti venduti fossero stati 25 milioni (come l'anno precedente) la vincita attesa da un singolo biglietto sarebbe stata pari a $E[X] \approx 0.68\,€$ (per un ricavo da parte dello Stato di ben 2.32 € a bigletto). È proprio vera la citazione anonima[4]

[4]Anche se questa frase è assolutamente condivisibile nella sostanza, il nostro Anonimo mostra di aver egli stesso poca familiarità con la "statistica", infatti qui la statistica non c'entra proprio niente e si dovrebbe piuttosto parlare di "probabilità".

"Le lotterie sono una tassa su chi non conosce la statistica"

□

Esempio 2.11.: Scommesse sui cavalli
Le scommesse sulle corse dei cavalli (e anche su altri eventi sportivi) funzionano così: scommetto una somma s_k su un certo cavallo k (posso scommettere cifre diverse su cavalli diversi) e ricevo $s_k + q_k s_k$ se il cavallo k vince (altrimenti perdo s_k). Il numero q_k è la "quota" del cavallo k e viene fissata dall'allibratore.
Ci si domanda: qual'è la quota "equa"? Per rispondere a questa domanda bisogna prima mettersi d'accordo su cosa significhi "equa". Analogamente a quanto visto per le lotteria, il concetto di valore atteso, o media, può fornirci un criterio accettabile: la scommessa è equa se *in media* ricavo quanto ho speso (cioè se il "guadagno atteso" è nullo). Per calcolarci le quote, guidati da questo concetto di equità, formalizziamo la scommessa. Se i cavalli in gara sono n possiamo numerarli da 1 a n e dire che l'esito della corsa è descritto dall'insieme di eventi elementari

$$\Omega = \{1, 2, \dots, n\}$$

dove l'evento elementare k significa *vince il cavallo* k. Se p_k è la probabilità che vinca il cavallo k, il mio possibile ricavo è descritto dalla variabile aleatoria $X : \Omega \to \mathbb{R}$ così definita:

$$X(k) = s_k + q_k s_k, \qquad k = 1, 2, \dots, n.$$

Perciò il mio ricavo medio è

$$E[X] = \sum_{k=1}^{n} p_k (s_k + q_k s_k).$$

Come si è detto, la scommessa è equa se il ricavo medio è uguale alla spesa, ovvero se

$$\sum_{k=1}^{n} p_k (s_k + q_k s_k) = s := \sum_{k=1}^{n} s_k.$$

Notiamo allora che se

$$q_k = \frac{1 - p_k}{p_k}, \qquad \text{ovvero} \quad p_k = \frac{1}{1 + q_k},$$

la scommessa è equa mentre, se $q_k < \frac{1-p_k}{p_k}$, ovvero $p_k < \frac{1}{1+q_k}$, l'allibratore in media ci guadagna. Posto $p'_k = \frac{1}{1+q_k}$ valutiamo il quadagno atteso dall'allibratore in funzione delle p'_k:

$$s - E[X] = s - \sum_{k=1}^{n} p_k (s_k + q_k s_k) = s - \sum_{k=1}^{n} p_k \Big(s_k + \frac{1 - p'_k}{p'_k} s_k \Big) = s - \sum_{k=1}^{n} \frac{p_k}{p'_k} s_k.$$

Se ad esempio $p'_k = c p_k$ si ha

$$\text{guadagno atteso dall'allibratore} = s\Big(1 - \frac{1}{c}\Big).$$

k	*Nome*	q_k	p'_k
1	Axum	10	0.0909
2	Kimmelstiel	15	0.0625
3	Lorenz de Bergerac	6	0.1429
4	Niki Dance	4	0.2000
5	Odilla	5	0.1667
6	Prevail	3	0.2500
7	San Leo	3/2	0.4000
8	Show must go on	3	0.2500
9	Sopran Fitas	5	0.1667
10	Sopran Poldin	12	0.0769
			$c = \mathbf{1.8065}$

Tabella 2.2: Quote e coefficienti p'_k e c relative alla corsa I del 17.10.2003 all'Ippodromo di Roma (dati SNAI).

Perciò l'allibratore di solito valuta le probabilità p_k in base alla sua esperienza e poi applica un coefficiente $c > 1$ per ottenere le p'_k e quindi le quote q_k. Ci si può accorgere del coefficiente c applicato calcolando

$$\sum_{k=1}^{n} \frac{1}{1+q_k} = \sum_{k=1}^{n} p'_k = \sum_{k=1}^{n} cp_k = c.$$

Nella tabella 2.2 è riportato una situazione tipica di scommesse sui cavalli. Le quote q_k sono state prese dal sito `www.snai.it`, dopodiché si sono calcolati i corrispondenti p'_k e se ne è fatta la somma per ottenere il valore di c che è risultato essere circa 1.8. Calcolando poi $1 - \frac{1}{c}$ si ottiene che il guadagno atteso dall'allibratore è circa il 45% del totale delle puntate ricevute. A questo modo, certamente, la scommessa non è proprio equa ma è chiaro che l'allibratore deve mantenere un margine di sicurezza, ripagarsi le spese e, infine, essendo questo il suo lavoro, deve assicurarsi un introito più o meno stabile. Dal punto di vista dello scommettitore invece, c'è una perdita media netta (nell'impostazione frequentista si potrebbe anche dire che "alla lunga" le perdite sarrano più dei ricavi). Scommettere sui cavalli è quindi da considerarsi un divertimento che si paga e non certamente un modo per guadagnare. □

Una prima naturale estensione della definizione 2.9 è quella del valore atteso per una v.a. discreta infinita.

Definizione 2.12. Sia X una v.a. discreta che può assumere gli infiniti valori $x_1, x_2, \ldots$ con probabilità $p_1, p_2, \ldots$. Si definisce *valore atteso* di X il numero

$$E[X] := \sum_{k=1}^{\infty} x_k p_k, \tag{2.9}$$

purché la serie che definisce $E[X]$ converga. □

Notiamo che estendere la definizione di media al caso di una v.a. che può assumere infiniti valori implica fare ricorso al concetto di "serie". Pertanto in

questo caso *non tutte le v.a. hanno valore atteso*, poiché non sempre la serie che lo definisce è convergente. Se la serie che definisce $E[X]$ converge, diremo che X ha "media finita".
In queste note non ci interessa di soffermarci sui valori attesi di v.a. discrete infinite ma piuttosto di servircene per capire la definizione di valore atteso nel caso di v.a. continue. Sia X una v.a. continua con densità f_X. Suddividiamo $\mathbb{R}$ in tanti intervallini di ampiezza Δx e centro x_k (con $k = 0, \pm 1, \pm 2, \ldots$). Se la suddivisione è abbastanza fine e f_X è sufficientemente regolare possiamo approssimare la v.a. continua X con la v.a. discreta $\tilde{X}$ che assume il valore x_k con probabilità

$$p_k = f_X(x_k)\,\Delta x.$$

Approssimiamo cioè la curva di densità con un *istogramma* (figura 2.8). Per tale

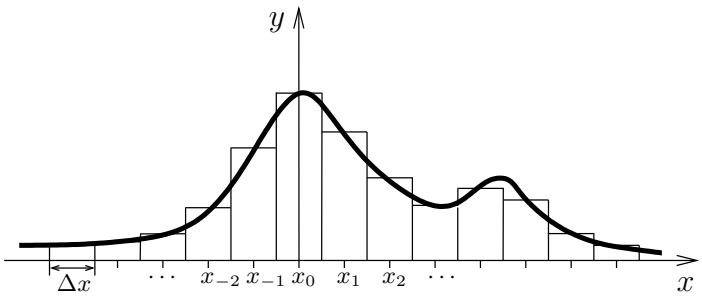


Figura 2.8: Aprossimazione discreta di una v.a. continua: notiamo che la curva è approssimata da un istogramma.

v.a., in base alla definizione 2.12, la media è data da

$$E\left[\tilde{X}\right] = \sum_{k=-\infty}^{+\infty} x_k p_k = \sum_{k=-\infty}^{+\infty} x_k f_X(x_k)\,\Delta x.$$

Per $\Delta x \to 0$, $E\left[\tilde{X}\right]$ tende all'integrale (se questo esiste) su $\mathbb{R}$ della funzione $x f_X(x)$:

$$E\left[\tilde{X}\right] \xrightarrow{\Delta x \to 0} \int_{-\infty}^{+\infty} x f_X(x)\,dx.$$

Questa discussione intuitiva ci porta alla seguente definizione rigorosa.

Definizione 2.13. Sia X una v.a. continua con densità f_X. Si definisce *valore atteso* di X il numero

$$E[X] := \int_{-\infty}^{+\infty} x f_X(x)\,dx, \tag{2.10}$$

purché l'integrale esista. □

Se l'integrale che definisce $E[X]$ esiste, diremo che X ha media finita.

Esercizio 2.14. *Sia X una v.a. continua con media finita; dimostrare che*

(1) se f_X è simmetrica rispetto a $x = 0$ (ovvero è "pari"), allora $E[X] = 0$,

(2) più in generale, se f_X è simmetrica rispetto a $x = m$, allora $E[X] = m$

(per la dimostrazione utilizzare le note proprietà degli integrali). □

Esercizio 2.15. *Calcolare la media della v.a. dell'esempio 2.6.* □

2.5 Varianza di una variabile aleatoria

Definizione 2.16. Sia X una v.a. (discreta o continua) con media finita $\mu = E[X]$. Si definisce *varianza*, o *scarto quadratico medio* di X il numero

$$\mathrm{Var}[X] = E\left[(X-\mu)^2\right], \tag{2.11}$$

purché questo numero esista, ovvero purché la variabile aleatoria $(X-\mu)^2$ abbia media finita. □

Supponiamo che X sia discreta e che assuma i valori $x_1, x_2, \ldots$ con probabilità $p_1, p_2, \ldots$. Allora la v.a. $(X-\mu)^2$ assumerà i valori $(x_1-\mu)^2, (x_2-\mu)^2, \ldots$ con probabilità $p_1, p_2, \ldots$. Dunque, nel caso discreto e finito si avrà

$$\mathrm{Var}[X] = \sum_{k=1}^{n} (x_k-\mu)^2 p_k, \tag{2.12}$$

nel caso discreto e infinito

$$\mathrm{Var}[X] = \sum_{k=1}^{\infty} (x_k-\mu)^2 p_k, \tag{2.13}$$

Con ragionamento analogo, nel caso continuo si avrà.

$$\mathrm{Var}[X] = \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)\,dx. \tag{2.14}$$

La varianza è dunque *la media dei quadrati degli scarti dalla media* e ci dà pertanto un'idea di quanto una v.a. è "raccolta" o "dispersa" attorno al suo valore atteso. Se la varianza esiste, si dirà che X ha varianza finita; notiamo che una v.a. X potrebbe non avere varianza finita pur avendo media finita. Chiaramente, ogni v.a. discreta e finita ha media e varianza finite.

Se X una v.a. con varianza finita possiamo scrivere

$$E\left[(X-\mu)^2\right] = E\left[X^2 - 2\mu X + \mu^2\right] = E\left[X^2\right] - 2\mu E[X] + E\left[\mu^2\right]$$

dove si sono usate le cosiddette *proprietà di linearità* della media:

(i) $E[f(X)+g(X)] = E[f(X)] + E[g(X)]$, dove $f(X)$ e $g(X)$ sono due funzioni della v.a. X.

(ii) $E[cX] = cE[X]$, dove c è una costante;

(proprietà la cui dimostrazione, nei casi discreto e continuo, è lasciata per esercizio). Ricordando che $\mu = E[X]$, e usando l'ulteriore proprietà

(iii) $E[c] = c$, dove c è una costante;

si ha dunque la seguente espressione alternativa per la varianza:

$$\mathrm{Var}[X] = E\left[X^2\right] - E[X]^2. \tag{2.15}$$

Una quantità collegata alla varianza, che risulta spesso utile in quanto dimensionalmente omogenea a X, è la deviazione standard.

Definizione 2.17. Sia X una v.a. con varianza finita. Si definisce *deviazione standard* di X la radice quadrata della varianza:

$$\mathrm{Std}[X] = \sqrt{\mathrm{Var}[X]}. \tag{2.16}$$

□

Esempio 2.18. Sia X una v.a. discreta con probabilità costante, ovvero che assume i valori $x_1, x_2, \ldots x_n$, ciascuno con probabilità $1/n$. Allora il valore atteso di X è

$$E[X] = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

e la varianza è data da

$$\mathrm{Var}[X] = \frac{x_1^2 + x_2^2 + \cdots + x_n^2}{n} - \Big(\frac{x_1 + x_2 + \cdots + x_n}{n}\Big)^2.$$

Utilizzando questa formula per calcolare la varianza della v.a. X dell'esempio 2.10 si ottiene $\mathrm{Var}[X] \approx 1.81 \times 10^6$ e $\mathrm{Std}[X] \approx 1.34 \times 10^3$. □

Teorema 2.19. (Disuguaglianza di Cebicev)
Sia X una v.a. con varianza finita e valore atteso μ. Per ogni $\epsilon > 0$ si ha

$$P(|X - \mu| \geq \epsilon) \leq \frac{\mathrm{Var}[X]}{\epsilon^2}. \tag{2.17}$$

Dimostrazione Facciamo la dimostrazione solo nel caso di X continua. Per una tale X si ha

$$\begin{aligned}
\mathrm{Var}[X] &= \int_{-\infty}^{+\infty} (x-\mu)^2 f_X(x)\,dx \\
&\geq \int_{-\infty}^{\mu-\epsilon} (x-\mu)^2 f_X(x)\,dx + \int_{\mu+\epsilon}^{+\infty} (x-\mu)^2 f_X(x)\,dx \\
&\geq \epsilon^2 \int_{-\infty}^{\mu-\epsilon} f_X(x)\,dx + \epsilon^2 \int_{\mu+\epsilon}^{+\infty} f_X(x)\,dx = \epsilon^2 P(|X-\mu| \geq \epsilon).
\end{aligned}$$

□

Definizione 2.20. Sia X una v.a. con media e varianza finite. Posto $\mu = E[X]$ e $\sigma = \mathrm{Std}[X]$, la variabile aleatoria

$$\tilde{X} = \frac{X - \mu}{\sigma} \tag{2.18}$$

è chiamata *standardizzata* della X. □

Esercizio 2.21. *Verificare che la standardizzata $\tilde{X}$ di X ha media 0 e varianza 1.* □

Capitolo 3

Alcune distribuzioni notevoli

Come abbiamo sottolineato nel capitolo precedente ci sono variabili aleatorie, anche molto diverse tra loro, che sono però accomunate dall'avere la medesima distribuzione, ovvero dal fatto di indurre la stessa misura di probabilità su $\mathbb{R}$ (cfr. paragrafo 2.1). Talune ditribuzioni ricorrono molto frequentemente nel calcolo delle probabilità e in statistica, sia per motivi "fondamentali" sia anche per la particolare semplicità del loro impiego e della loro analisi. È di alcune di tali distribuzioni "notevoli", sia discrete che continue, che ci vogliamo occupare in questo capitolo.

3.1 La distribuzione binomiale

Consideriamo un esperimento probabilistico consistente in un'unica prova i cui possibili esiti possono essere solamente *successo* o *insuccesso*, il primo con probabilità $p \in [0,1]$ e il secondo, ovviamente, con probabilità $1-p$. Tale prova viene detta *Bernoulliana*. Il lancio di una moneta è una prova di questo tipo, con $p = 1/2$; giocare a "fare 6" lanciando un dado è un'altra prova di questo tipo, con $p = 1/6$, e così via: gli esempi possono essere numerosissimi. Possiamo costruire una v.a. associata ponendo

$$X = \begin{cases} 1, & \text{se il risultato è } \textit{successo}; \\ 0, & \text{se il risultato è } \textit{insuccesso}. \end{cases}$$

La v.a. così costruita è una v.a. discreta semplicissima: assume il valore 1 con probabilità p e il valore 0 con probabilità $1-p$:

$$P(X=1) = p, \qquad P(X=0) = 1-p. \tag{3.1}$$

Definizione 3.1. La distribuzione definita dalla (3.1) è detta *Bernoulliana* ed è indicata con $\mathcal{B}(p)$. Se X è una v.a. con tale distribuzione, scriveremo

$$X \sim \mathcal{B}(p),$$

che si legge "X ha distribuzione $\mathcal{B}(p)$". □

Notiamo che la distribuzione Bernoulliana dipende dal parametro $p \in [0,1]$. Lasciamo per esercizio la verifica del fatto che, se $X \sim \mathcal{B}(p)$, allora

$$E[X] = p, \qquad \text{Var}[X] = p(1-p). \tag{3.2}$$

Supponiamo ora di eseguire n prove Bernoulliane *identiche e indipendenti* come, ad esempio, n estrazioni da un'urna con reimbussolamento, n lanci di una moneta o di un dado, ma anche n test del palloncino eseguiti da una pattuglia della Polizia Stradale. Come sopra, ad ogni prova ci interessa solo l'esito *successo* o *insuccesso* ("successo", ad esempio, potrebbe essere l'estrazione di una pallina rossa da un'urna contenente palline bianche, rosse e verdi, l'ottenere T nel lancio della moneta, l'ottenere 6 nel lancio di un dado, il risultato positivo di un test del palloncino, e così via).
Poiché assumiamo che tutte le prove siano identiche e indipendenti l'una dall'altra, cioè che ciascuna prova avviene sempre nelle medesime condizioni, la probabilità di successo o di insuccesso è sempre la stessa ad ogni prova. Ogni singola prova, considerata a sé, è quindi una v.a. Bernoulliana: se p è la probabilità di successo, con $0 \le p \le 1$, potremo allora scrivere:

$$\begin{cases} P(X_i = 1) = p, \\ P(X_i = 0) = 1 - p, \end{cases}$$

per ogni $i = 1, 2, \ldots, n$ (indice che numera le prove). Consideriamo la v.a. X che conta il *numero di successi nelle n prove*:

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

(notiamo che $X = k$ se ci sono k successi, comunque posizionati). Vogliamo trovare la distribuzione di X. Chiaramente, X è una v.a. discreta che può assumere i valori $k = 0, 1, 2, \ldots, n$. La distribuzione di X si determina perciò calcolando $P(X = k)$, cioè calcolando la probabilità di avere k successi in n prove (indipendentemente dalla posizione dei successi).
Per far ciò, conviene tuttavia cominciare col calcolare la probabilità di avere k successi in posizioni fissate. A questo scopo consideriamo gli eventi

$$S_i = \textit{successo alla i-esima prova} \quad \textit{(cioè } X_i = 1\textit{)},$$

fissiamo k indici $i_1, i_2, \ldots, i_k$ (indicando con $j_1, j_2, \ldots, j_{n-k}$ gli indici rimanenti) e calcoliamo la probabilità di avere k successi nelle prove $i_1, i_2, \ldots, i_k$, ovvero

$$P\left(S_{i_1} \cap \cdots \cap S_{i_k} \cap S^c_{j_1} \cap \cdots \cap S^c_{j_{n-k}}\right).$$

Per l'ipotesi di indipendenza, questa è data da

$$P(S_{i_1}) \cdots P(S_{i_k})\, P\left(S^c_{j_1}\right) \cdots P\left(S^c_{j_{n-k}}\right) = p^k (1-p)^{n-k}.$$

Ma allora la probabilità di avere k successi in qualunque posizione la si ottiene come probabilità dell'unione, per tutte le possibili scelte dei k indici $i_1, i_2, \ldots, i_k$

in $1, 2, \ldots, n$, degli eventi dei quali sopra abbiamo calcolato la probabilità: $S_{i_1} \cap \cdots \cap S_{i_k} \cap S^c_{j_1} \cap \cdots \cap S^c_{j_{n-k}}$. Poiché questi eventi sono tutti disgiunti si ottiene

$$P(X = k) = P\left(\bigcup_{i_1, i_2, \ldots, i_k} S_{i_1} \cap \cdots \cap S_{i_k} \cap S^c_{j_1} \cap \cdots \cap S^c_{j_{n-k}} \right)$$

$$= \sum_{i_1, i_2, \ldots, i_k} P(S_{i_1}) \cdots P(S_{i_k}) \, P\left(S^c_{j_1}\right) \cdots P\left(S^c_{j_{n-k}}\right) = \sum_{i_1, i_2, \ldots, i_k} p^k (1-p)^{n-k}.$$

I termini della somma (costanti) sono tanti quanti i possibili modi di scegliere i k indici $i_1, i_2, \ldots, i_k$ in $1, 2, \ldots, n$, ovvero $C(n, k)$ (cfr. paragrafo 1.4). Si ha perciò

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \ldots, n. \tag{3.3}$$

Definizione 3.2. La distribuzione descritta dalla (3.3) è detta *binomiale* ed è indicata con $\mathcal{B}(n, p)$. Se X è una v.a. con distribuzione $\mathcal{B}(n, p)$ scriveremo

$$X \sim \mathcal{B}(n, p).$$

□

Notiamo che la distribuzione binomiale dipende da due parametri: $n \in \mathbb{N}$ e $p \in [0, 1]$. Notiamo anche che per $n = 1$ si ritrova la distribuzione Bernoulliana, dunque

$$\mathcal{B}(1, p) = \mathcal{B}(p). \tag{3.4}$$

Esempio 3.3.: Il gioco del Totocalcio II
Nell'esempio 1.5 abbiamo formalizzato il gioco del Totocalcio, calcolando la probabilità di "fare 13" giocando una colonna. Supponendo per semplicità che tutti i risultati siano equiprobabili (cosa in realtà ben lungi dall'essere vera), proviamo adesso a calcolare la probabilità di "fare k", con $k = 0, 1, 2, \ldots, 13$. Se tutti i risultati sono equiprobabili, allora le partite in schedina possono essere considerate come $n = 13$ prove Bernoulliane, ciascuna con probabilità di successo $p = 1/3$. Dunque, la v.a. X del punteggio totalizzato ha distribuzione binomiale

$$X \sim \mathcal{B}\,(13, 1/3)$$

e perciò la probabilità di "fare k" è data da

$$P(X = k) = \binom{13}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{13-k}, \qquad k = 0, 1, \ldots, 13.$$

In figura 3.1 è mostrato il grafico della distribuzione (le probabilità sono espresse in percentuale). Notiamo che la probabilità di fare 13 è, come già calcolato nell'esempio 1.5, $P(X = 13) \approx 6.3 \times 10^{-7}$, quella di non indovinare neanche una

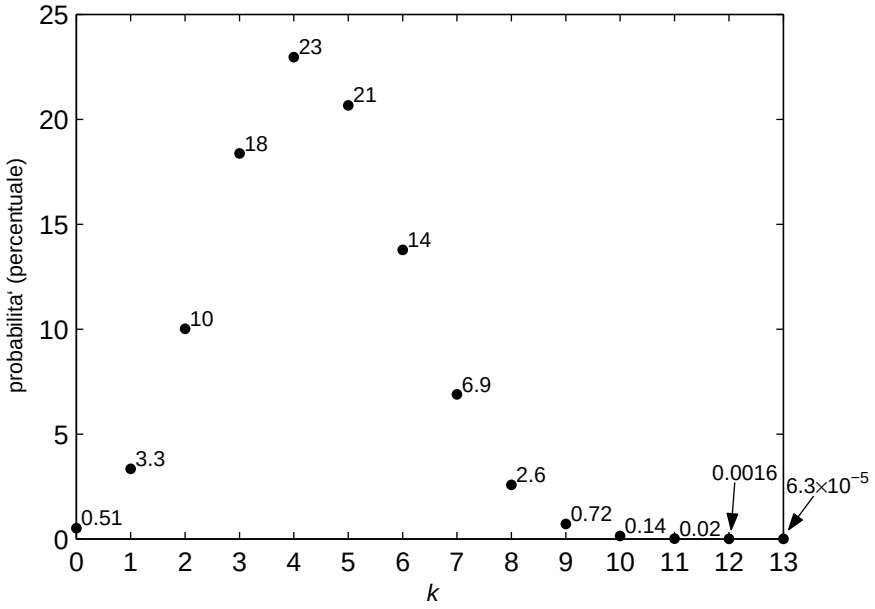


Figura 3.1: La distribuzione binomiale del Totocalcio $\mathcal{B}(13, 1/3)$.

partita è $P(X = 0) \approx 0.0051$ mentre il risultato più probabile è $P(X = 4) \approx 0.23$.[1] □

Calcoliamo adesso il valore atteso di una v.a. binomiale $X \sim \mathcal{B}(n, p)$:

$$E[X] = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} \frac{(n-1)!}{(n-k)!\,(k-1)!}\, p^k (1-p)^{n-k}$$

$$= np \sum_{r=0}^{n} \frac{(n-1)!}{(n-1-r)!\,r!}\, p^r (1-p)^{n-1-r}.$$

Poiché dalla formula del binomio di Newton segue

$$\sum_{r=0}^{n} \frac{(n-1)!}{(n-1-r)!\,r!}\, p^r (1-p)^{n-1-r} = 1,$$

si ottiene in definitiva

$$E[X] = np. \tag{3.5}$$

Nel paragrafo 4.4 vedremo come ottenere questo risultato in modo molto più semplice nel contesto delle variabili aleatorie multiple.

[1] Nella pratica il risultato che si presenta con maggior frequenza è superiore a 4, poiché in realtà p è maggiore di 1/3 (si ha 1/3 solo se si fanno i pronostici tirando a caso). Aver posto $p = 1/3$ è un "modello" della realtà, che qui adottiamo per semplicità. Un procedura migliore sarebbe quella di *stimare* il paramtero p con le tecniche della statistica (si vedano i capitoli 6 e 7).

Per calcolare la varianza di $X \sim \mathcal{B}(n,p)$, secondo la definizione (2.12), si può usare la formula

$$\text{Var}\,[X] = \sum_{k=0}^{n} (k-np)^2 \binom{n}{k} p^k (1-p)^{n-k}.$$

I calcoli sono simili ai precedenti, ma un po' più complicati e per questo li omettiamo. Tuttavia, anche in questo caso, è più semplice ricorrere alla teoria delle variabili aleatorie multiple (esercizio 4.18). In ogni caso, il risultato è

$$\text{Var}\,[X] = np(1-p). \tag{3.6}$$

Se calcoliamo ad esempio media e varianza della v.a. binomiale del Totocalcio, $X \sim \mathcal{B}\,(13, 1/3)$ (esempio 3.3), otteniamo $E\,[X] = 13/3$ e $\text{Var}\,[X] = 26/9$.

Esercizio 3.4.[2] *Calcolare la probabilità di*

A) ottenere almeno un 6 lanciando un dado 4 volte;

B) ottenere almeno un doppio 6 lanciando due dadi 24 volte.

3.2 La distribuzione geometrica

Come nel paragrafo precedente consideriamo una successione di prove Bernoulliane, cioè identiche e indipendenti i cui possibili esiti siano "successo", con probabilità p, o "insuccesso", con probabilità $1-p$. Stavolta però non mettiamo un limite fissato al numero di prove, che potrebbero essere potenzialmente infinite. Consideriamo la v.a. X definita come il numero di insuccessi prima del primo successo. Tale v.a. è detta anche *ritardo* del primo successo. Chiaramente X può assumere i valori $r = 0, 1, 2, \ldots$ e si tratta perciò di una VA discreta infinita. Utilizzando le notazioni introdotte nel paragrafo precedente la distribuzione di X si calcola così:

$$P(X = r) = P(S_1^c \cap \cdots S_r^c \cap S_{r+1}) = P(S_1^c) \cdots P(S_r^c)\, P(S_{r+1}) = (1-p)^r p,$$

dove si è usata l'ipotesi di indipendenza.

Definizione 3.5. La distribuzione di una v.a. discreta infinita X che può assumere i valori $r = 0, 1, 2, \ldots$ con probabilità

$$P(X = r) = p(1-p)^r \tag{3.7}$$

è detta *geometrica* e si indica con $\mathcal{G}(p)$, per cui scriveremo

$$X \sim \mathcal{G}(p).$$

□

[2] Questo esercizio è un celebre problema di probabilità, celebre perché fu posto dal Cavaliere de Méré a Pascal agli albori del calcolo delle probabilità; il Cavaliere scommetteva intuitivamente "alla pari" i due eventi (ovvero come se entrambi avessero probabilità 1/2) mentre, usando la distribuzione binomiale, ci si accorge che il primo ha probabilità leggermente maggiore di 1/2 e il secondo leggermente minore.

Notiamo che la distribuzione geometrica dipende da un solo parametro $p \in [0,1]$. Senza eseguire i calcoli, che richiedono di maneggiare un po' di serie di potenze, riportiamo media e varianza di una v.a. $X \sim \mathcal{G}(p)$:

$$\begin{aligned} E[X] &= \sum_{r=0}^{\infty} rp(1-p)^r = \frac{1-p}{p}\,, \\ \mathrm{Var}[X] &= \sum_{r=0}^{\infty} \Big(r - \frac{1-p}{p}\Big)^2 p(1-p)^r = \frac{1-p}{p^2}\,. \end{aligned} \tag{3.8}$$

Esempio 3.6.: Ritardi nel gioco del Lotto
Su ogni "ruota" del Lotto vengono estratti 5 numeri da 1 a 90. La probabilità di estrazione di un dato numero compreso tra 1 e 90 è perciò $p = 5/90 = 1/18$. Fissato un certo numero fra 1 e 90, ciascuna estrazione la si può vedere come una prova Bernoulliana in cui il "successo" è l'uscita di quel numero e ha probabilità 1/18. Il ritardo di quel numero, ovvero il numero di estrazioni prima della sua uscita, è dunque una v.a. di tipo geometrico e precisamente

$$X \sim \mathcal{G}\Big(\frac{1}{18}\Big).$$

Pertanto, la probabilità che un dato numero ritardi di esattamente r estrazioni è

$$P(X = r) = \frac{1}{18}\Big(\frac{17}{18}\Big)^r$$

e la probabilità che ritardi di *almeno* r estrazioni è

$$P(X \geq r) = P(S_1^c \cap \cdots \cap S_r^c) = \Big(\frac{17}{18}\Big)^r.$$

Notiamo che quest'ultima coincide con la probabilità che una v.a. con distribuzione binomiale $\mathcal{B}(r, 1/18)$ valga 0 (nessun successo in r prove).
I giornali "specializzati" pubblicizzano i numeri con i maggiori ritardi sulle varie ruote e i giocatori tendono a puntare su di essi, come se l'accumularsi del ritardo aumentasse la probabilità di uscita la volta succssiva. Tuttavia questo è chiaramente falso, dal momento che ogni prova è indipendente dalle altre! Anche se un ritardo lungo è *complessivamente* poco probabile, l'accumularsi o meno del ritardo settimana per settimana ha sempre la stessa probabilità, $1-p$ e p, indipendentemente da quanto ritardo si è già accumulato.[3] Questa proprietà di "mancanza di memoria" della distribuzione geometrica sarà formulata e dimostrata più in generale qui di seguito. □

Proposizione 3.7. (Mancanza di memoria della distribuzione geometrica) *Se X ha una distribuzione geometrica, $X \sim \mathcal{G}(p)$, allora*

$$P(X = r + s \mid X \geq r) = P(X = s)\,. \tag{3.9}$$

[3] Semmai, ragionando dal punto di vista statistico, se un numero non esce tante volte di seguito si dovrebbe piuttosto essere portati a pensare che la sua uscita sia *meno probabile* di quella degli altri numeri! A questo proposito segnaliamo un'utile fonte di dati statistici e storici sul gioco del Lotto: il sito web `www.giocodellotto.com`.

Dimostrazione La dimostrazione consiste in un semplice calcolo:

$$P(X = r + s \mid X \geq r) = \frac{P((X = r + s) \cap (X \geq r))}{P(X \geq r)}$$

$$= \frac{P(X = r + s)}{P(X \geq r)} = \frac{p(1-p)^{r+s}}{(1-p)^r} = p(1-p)^s = P(X = s).$$

□

Ad esempio, nei ritardi del Lotto la probabilità che un numero non uscito per r settimane esca la settimana successiva è

$$P(X = r \mid X \geq r) = P(X = 0) = p$$

(con $p = 1/18$): com'è ovvio la probabilità di estrazione di quel numero resta sempre p, indipendentemente dal ritardo già accumulato.

3.3 La distribuzione di Poisson

La distribuzione di Poisson è il limite per un "continuo di prove" della distribuzione binomiale. Vediamo di precisare questo concetto. Supponiamo di voler registrare nel tempo le occorrenze di eventi casuali e indipendenti tra loro come, ad esempio, l'osservazione di stelle cadenti o, più prosaicamente, l'arrivo di chiamate a un centralino telefonico. Fissiamo un intervallo di osservazione $[0, t]$ che dividiamo in n intervallini di campionamento di uguale ampiezza

$$\Delta t = t/n.$$

Se gli eventi osservati sono sufficientemente "rari" (rispetto al tempo di osservazione t) possiamo supporre che, per n sufficientemente grande, la probabilità che in un singolo intervallino si verifichino due o più occorrenze dell'evento sia completamente trascurabile. Pertanto, in ogni intervallino registreremo o "successo" (una occorrenza in quell'intervallino) o "insuccesso" (zero occorrenze in quell'intervallino). Se i vari istanti di tempo sono tutti equivalenti ai fini del verificarsi dell'evento, in ciascun intervallino la probabilità di successo p sarà la stessa. Gli n intervallini di campionamento costituiscono perciò n prove Bernoulliane e la variabile aleatoria X_n che conta il numero di successi (occorrenze) fra 0 e t ha una distribuzione binomale

$$X_n \sim \mathcal{B}(n, p).$$

Per poter passare al limite per $n \to \infty$ dobbiamo essere più precisi riguardo alla probabilità p. È chiaro che questa deve "scalare con l'ampiezza dell'intervallino di campionamento", ovvero, quanto più piccolo è quest'ultimo tanto più piccola è p. Per Δt piccoli è lecito supporre una dipendenza lineare di p da Δt:

$$p = \nu \, \Delta t = \nu t/n,$$

dove $\nu > 0$ è detta *frequenza media di occorrenza*. Si ha dunque

$$X_n \sim \mathcal{B}(n, \nu t/n).$$

Che succede allora se infittisco sempre più il campionamento, ovvero se $n \to \infty$? Dalla (3.3) segue

$$P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \Big(\frac{\nu t}{n}\Big)^k \Big(1 - \frac{\nu t}{n}\Big)^{n-k}$$

$$= \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \frac{(\nu t)^k}{k!} \Big(1 - \frac{\nu t}{n}\Big)^n \Big(1 - \frac{\nu t}{n}\Big)^{-k}$$

$$\xrightarrow{n\to\infty} \quad \frac{(\nu t)^k}{k!} \mathrm{e}^{-\nu t}.$$

Perciò, al limite per $n \to \infty$, la distribuzione della v.a. X_n tende alla distribuzione di una v.a. X, discreta infinita, che può assumere i valori $k = 0, 1, \ldots$ con probabilità

$$P(X = k) = \frac{(\nu t)^k}{k!} \mathrm{e}^{-\nu t}, \qquad k = 0, 1, \ldots. \tag{3.10}$$

Definizione 3.8. La distribuzione discreta e infinita descritta dalla (3.3) è detta *di Poisson*[4] ed è indicata con $\mathcal{P}(\nu t)$, per cui scriveremo

$$X \sim \mathcal{P}(\nu t).$$

□

Notiamo che la distribuzione di Poisson dipende dai parametri reali $\nu > 0$ e $t \geq 0$; tuttavia questi si presentano sempre sotto forma di prodotto νt e quindi, in realtà, la distribuzione di Poisson dipende dell'unico parametro reale νt. La distribuzione di poisson (detta anche "degli eventi rari") fa la sua apparizione in una miriade di contesti, i più diversi fra loro:

- chiamate a un centralino;
- processi in arrivo a una stampante di rete;
- clienti che entrano in un negozio;
- batteri in una capsula di Petri;
- tracce di animali in un'area di osservazione;

e molti altri. Negli ultimi due esempi dei questo elenco t non avrà un significato temporale ma spaziale (lunghezza, area, volume) e ν sarà, conseguentemente, una frequenza media di occorrenza per unità di lunghezza, area, volume.

Calcoliamo adesso media e varianza di una variabile aleatoria di Poisson $X \sim \mathcal{P}(\nu t)$. Per la media si ha:

$$E[X] = \sum_{k=0}^{\infty} k \frac{(\nu t)^k}{k!} \mathrm{e}^{-\nu t} = \nu t \, \mathrm{e}^{-\nu t} \sum_{k=1}^{\infty} \frac{(\nu t)^{k-1}}{(k-1)!} = \nu t \tag{3.11}$$

[4] Dal nome del grande matematico francese Siméon Deins Poisson, 1781-1840.

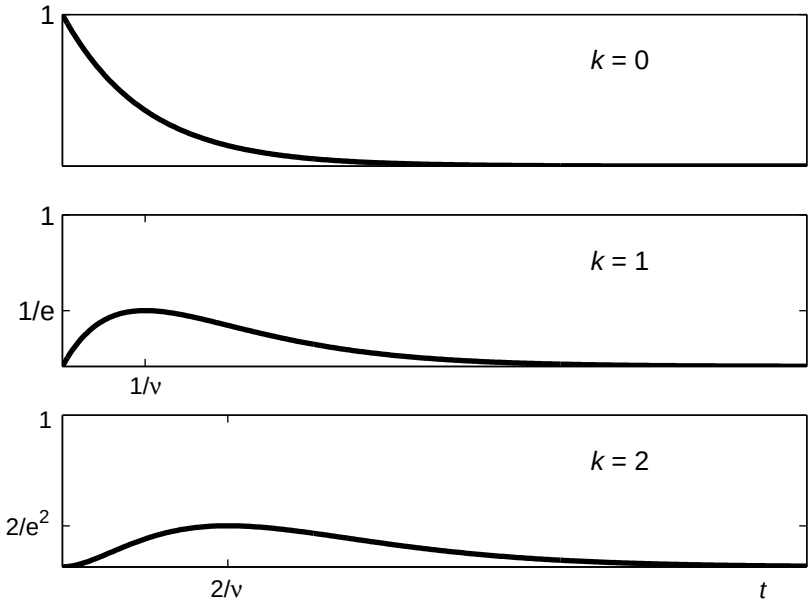


Figura 3.2: Grafici della distribuzione di Poisson al variare di t per tre diversi valori di k fissati. Ciascun grafico risponde alla domanda: "qualè la probabilità di avere k occorrenze ($k = 0, 1, 2$) fra 0 e t?".

(in accordo col significato di ν, che è una frequenza media di successi). Per la varianza, usando la (2.15), si ha:

$$\mathrm{Var}\,[X] = \sum_{k=0}^{\infty} k^2 \frac{(\nu t)^k}{k!}\,\mathrm{e}^{-\nu t} - (\nu t)^2 = \nu t\,\mathrm{e}^{-\nu t} \sum_{k=1}^{\infty} k \frac{(\nu t)^{k-1}}{(k-1)!} - (\nu t)^2$$

$$= \nu t\,\mathrm{e}^{-\nu t}\Big[\sum_{k=1}^{\infty}(k-1)\frac{(\nu t)^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty}\frac{(\nu t)^{k-1}}{(k-1)!}\Big] - (\nu t)^2.$$

Nella prima sommatoria si riconosce la media νt già calcolata sopra e nella seconda lo sviluppo in serie di un esponenziale; si ottiene quindi

$$\mathrm{Var}\,[X] = (\nu t)^2 + \nu t - (\nu t)^2 = \nu t. \tag{3.12}$$

3.4 La distribuzione esponenziale

Mettiamoci nella stessa situazione in cui nasce la distribuzione di Poisson: la registrazione delle occorrenze nel tempo di eventi casuali e indipendenti, sufficientemente "rari", con frequenza media ν. Adesso però siamo interessati alla variabile aleatoria T definita come il *tempo di attesa della prima occorrenza*. Qual'e la disribuzione di T? Fissato $t \geq 0$, chiaramente si ha

$$P(T \leq t) = 1 - P(T > t) = 1 - P(X_t = 0) = 1 - \mathrm{e}^{-\nu t},$$

dove $X_t \sim \mathcal{P}(\nu t)$ è la v.a. , con distribuzione di Poisson, che ci dà il numero di occorrenze al tempo t (l'indice t serve per sottolineare la dipendenza dal tempo della v.a. di Poisson). Inoltre, ovviamente, possiamo porre

$$P(T \leq t) = 0, \qquad \text{per } t < 0$$

in quanto l'attesa del primo evento comincia da $t = 0$. Abbiamo dunque trovato la funzione di ripartizione della variabile aleatoria T:

$$F_T(t) = \begin{cases} 1 - \mathrm{e}^{-\nu t}, & \text{se } t \geq 0, \\ 0, & \text{se } t < 0. \end{cases} \tag{3.13}$$

Per la proprietà (iii) della proposizione 2.7 si ha che la densità di T è data da[5] $f_T(t) = \frac{d}{dt} F_T(t)$ e dunque

$$f_T(t) = \begin{cases} \nu\, \mathrm{e}^{-\nu t}, & \text{se } t \geq 0, \\ 0, & \text{se } t < 0, \end{cases} \tag{3.14}$$

(figura 3.3).

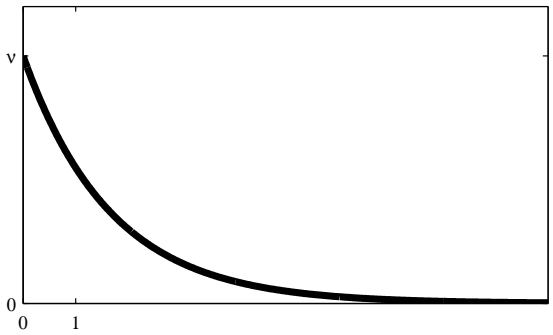


Figura 3.3: Grafico della parte $x \geq 0$ della densità esponenziale (3.14)

Definizione 3.9. La distribuzione della v.a. continua T descritta dalla (3.14) è detta *esponenziale* ed è indicata con $\mathcal{E}(\nu)$, per cui scriveremo

$$T \sim \mathcal{E}(\nu).$$

□

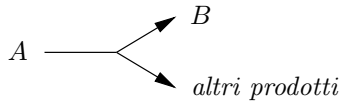
Notiamo che la distribuzione esponenziale dipende da un solo parametro reale $\nu > 0$. Risulta inoltre che

$$E[T] = \int_0^{+\infty} t\, \nu\, \mathrm{e}^{-\nu t}\, dt = \frac{1}{\nu}, \tag{3.15}$$

[5] Questo, a rigore, è vero solo per $t \neq 0$, dove F_T è derivabile. In ogni caso $f_T(0)$ può sempre definita a piacimento, in quanto gli integrali di f_T non cambiano per la sua alterazione in un singolo punto. Nella (3.14) si è scelto $f_T(0) = \nu$.

$$\mathrm{Var}\,[T] = \int_0^{+\infty} \left(t - \frac{1}{\nu}\right) \nu\, \mathrm{e}^{-\nu t}\, dt = \frac{1}{\nu^2}\,. \tag{3.16}$$

Esempio 3.10.: Decadimento radioattivo
Il decadimento radioattivo di una specie atomica A in una specie atomica B avviene secondo lo schema



Gli esperimenti mostrano che il tempo di attesa del decadimento è di tipo esponenziale: $T \sim \mathcal{E}(\nu)$ (dove ν è un valore tipico del particolare decadimento). Se abbiamo N atomi della specie A, poiché ciascuno di essi ha una probabilità $P(T \le t) = 1 - \mathrm{e}^{-\nu t}$ di essere decaduto prima del tempo t, allora al tempo t saranno decaduti $(1 - \mathrm{e}^{-\nu t})N$ atomi[6] e dunque restano in media

$$\mathrm{e}^{-\nu t} N$$

atomi della specie A. Quanto tempo t_0 deve passare affinché mi ritrovi con la metà degli atomi della specie A che avevo inizialmente? La risposta è

$$t_0 = \frac{1}{\nu}\, \ln 2.$$

Tale t_0 è detto *tempo di dimezzamento*. Notiamo che $\ln 2 \approx 0.69 \approx 2/3$. Inutile sottolineare l'importanza di questi concetti ai fini applicativi, come ad esempio la datazione col C_{14}. □

La distribuzione esponenziale è per quella di Poisson l'analogo della distribuzione geometrica per quella binomiale. Si potrebbe anzi dedurre la funzione di ripartizione (3.13) discretizzando il tempo in n intervallini (come nel paragrafo precedente), considerando la v.a. discreta "ritardo del primo successo" (che ha distribuzione geometrica) e passando al limite per $n \to \infty$. Possiamo inoltre dimostrare una proprietà di "mancanza di memoria" della distribuzione esponenziale, analoga a quella già dimostrata per la distribuzione geometrica (proposizione 3.7).

Proposizione 3.11. (Mancanza di memoria della distribuzione esponenziale) *Se T ha una distribuzione esponenziale, $T \sim \mathcal{E}(\nu)$, allora*

$$P(T > t + s \mid T > t) = P(T > s)\,. \tag{3.17}$$

Dimostrazione La dimostrazione è analoga al caso della distribuzione geometrica:

$$P(T > t + s \mid T > t) = \frac{P((T > t + s) \cap (T > t))}{P(T > t)}$$

[6] Questa affermazione si può dimostrare rigorosamente utilizzando la legge dei grandi numeri (esempio 4.27).

$$= \frac{P(T > t+s)}{P(T > t)} = \frac{\mathrm{e}^{-\nu(t+s)}}{\mathrm{e}^{-\nu t}} = \mathrm{e}^{-\nu s} = P(T > s)\,.$$

□

Nell'esempio del decadimento radioattivo si può conculdere che se l'atomo non è decaduto dopo un tempo t, la probabilità di dover attendere ancora un tempo s non dipende dal tempo t già passato ma è la stessa probabilità di dover attendere un tempo s che si sarebbe avuta all'inizio.

3.5 La distribuzione uniforme

La distribuzione uniforme su un intervallo rale $[a, b]$ è già stata introdotta nell'esempio della telefonata (esempio 2.6). È una distribuzione continua in cui la densità ha un valore costante su $[a, b]$ ed è nulla su tutti gli altri punti della retta. Il valore della costante è determinato dalla condizione (2.6) ed è necessariamente uguale all'inverso dell'ampiezza dell'intervallo, $\frac{1}{b-a}$.

Definizione 3.12. La distibuzione *uniforme* su un intervallo $[a, b]$ è la distribuzione di una v.a. continua X con densità

$$f_X(x) = \begin{cases} 0, & \text{se } x < a, \\ \frac{1}{b-a}, & \text{se } a \le t \le b, \\ 0, & \text{se } t > b, \end{cases} \tag{3.18}$$

(figura 3.4). Scriveremo in questo caso

$$X \sim \mathcal{U}(a, b).$$

□

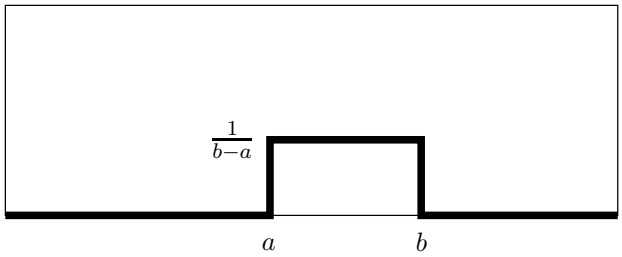


Figura 3.4: Densità della distribuzione uniforme sull'intervallo $[a, b]$.

Notiamo che la probabilità che X sta in un qualunque intervallo $[x_1, x_2]$ della retta reale, è proporzionale all'ampiezza dell'intervallo $[x_1, x_2] \cap [a, b]$.
Calcoliamo ora valore atteso e varianza di una v.a. $X \sim \mathcal{U}(a, b)$. A questo scopo calcoliamo i seguenti integrali:

$$\int_{-\infty}^{+\infty} x\, f_X(x)\, dx = \frac{1}{b-a} \int_a^b x\, dx = \frac{1}{b-a} \left(\frac{b^2}{2} - \frac{a^2}{2} \right) = \frac{a+b}{2},$$

$$\int_{-\infty}^{+\infty} x^2 f_X(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\left(\frac{b^3}{3} - \frac{a^3}{3}\right) = \frac{(b^2+ab+a^2)}{3}$$

Usando le formule (2.10) e (2.15) si ottiene quindi

$$E[X] = \frac{b-a}{2}, \tag{3.19}$$

$$\mathrm{Var}[X] = \frac{(b^2+ab+a^2)}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12} \tag{3.20}$$

3.6 La distribuzione normale

La *distribuzione normale* è la distribuzione notevole per eccellenza. Oltre alla sua importanza teorica (si veda ad esempio il suo ruolo nel teorema centrale di convergenza, paragrafo 4.6, che in parte ne spiega l'onnipresenza), essa è di primaria importanza nelle applicazioni; sono infatti distribuite normalmemte moltissime variabili aleatorie nei più svariati contesti come ad esempio

- errori nelle misurazioni di grandezze fisiche;
- velocità delle molecole in un gas all'equilibrio;
- grandezze bio-ecologiche;
- grandezze econometriche;
- grandezze antropometriche.

Definizione 3.13. La distibuzione *normale* o *Gaussiana*[7] è la distribuzione di una v.a. continua X con densità

$$f_X(x) = g_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.21}$$

dove μ e $\sigma > 0$ sono due parametri reali. per tale v.a. scriveremo

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

□

La distribuzione normale dipende da due parametri reali, $\mu \in \mathbb{R}$ e $\sigma > 0$, il cui significato è chiarito dalla seguente proposizione.

Proposizione 3.14. *La densità gaussiana g_{μ,σ^2}, definita dalla* (3.21)*, ha le seguenti proprietà:*

(i) $$\int_{-\infty}^{+\infty} g_{\mu,\sigma^2}(x)\,dx = 1,$$

(ii) $$\int_{-\infty}^{+\infty} x\, g_{\mu,\sigma^2}(x)\,dx = \mu,$$

(iii) $$\int_{-\infty}^{+\infty} (x-\mu)^2\, g_{\mu,\sigma^2}(x)\,dx = \sigma^2.$$

[7] Dal nome del grandissimo matematico e fisico tedesco Carl Friedrich Gauss, 1777-1855.

Dimostrazione (i) Segue dalla nota formula $\int_{-\infty}^{+\infty} \mathrm{e}^{-x^2}\,dx = \sqrt{\pi}$ (si veda ad esempio [7]), facendo l'opportuno cambio di variabile.
(ii) Notiamo che g_{μ,σ^2} è simmetrica rispetto a $x = \mu$, dunque il risultato segue dall'esercizio 2.14.
(iii) Utilizzando l'integrazione per parti e il punto (i) si ottiene

$$\begin{aligned}\int_{-\infty}^{+\infty} (x-\mu)^2\, g_{\mu,\sigma^2}(x)\,dx &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (x-\mu)^2\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ &= -\frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} (x-\mu)\Big(-\frac{x-\mu}{\sigma^2}\Big)\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx \\ = (x-\mu)\, \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\Big|_{x=\infty}^{x=+\infty} &+ \frac{\sigma^2}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{+\infty} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}\,dx = \sigma^2.\end{aligned}$$

□

La precedente proposizione ci dice che g_{μ,σ^2} è effettivamente una densità (verifica la proprietà (2.6)) e che, se la variabile aleatoria X ha distribuzione $\mathcal{N}(\mu,\sigma^2)$, allora

$$E[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2, \qquad \mathrm{Std}[X] = \sigma. \tag{3.22}$$

Dunque, la distribuzione $\mathcal{N}(\mu,\sigma^2)$ è parametrizzata dalla media e dalla varianza.

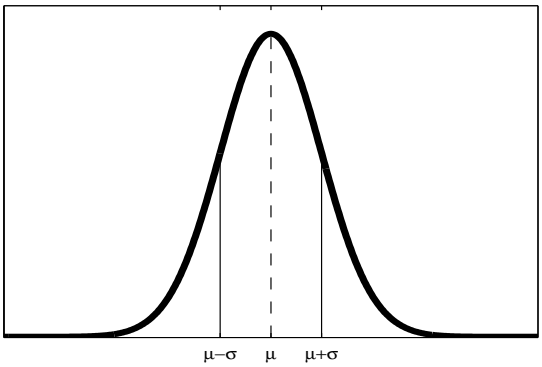


Figura 3.5: Grafico della densità gaussiana $y = g_{\mu,\sigma^2}(x)$. L'area sotto il grafico è 1, quella compresa tra $x = \mu+\sigma$ e $x = \mu-\sigma$ è circa 2/3.

Il grafico della funzione g_{μ,σ^2} è riportato in figura 3.5. La curva ha due flessi nei punti $x = \mu-\sigma$ e $x = \mu+\sigma$. Se $X \sim \mathcal{N}(\mu,\sigma^2)$, risulta che

$$P(\mu-\sigma \le X \le \mu+\sigma) \approx 0.68 \approx 2/3.$$

Questo ci permette di visualizzare immediatamente il significato della varianza (o della deviazione standard) come misura della "dispersione" di una v.a. gaussiana: la probabilità è concentrata per i 2/3 in un intevallo di ampiezza 2σ attorno al valore medio μ.

Definizione 3.15. La distribuzione $\mathcal{N}(0,1)$, normale con media 0 e varianza 1, è detta *normale standard* e, per brevità, sarà indicata solamente con $\mathcal{N}$:

$$\mathcal{N} := \mathcal{N}(0,1). \tag{3.23}$$

□

La funzione di ripartizione della distribuzione normale standard

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \mathrm{e}^{-y^2/2}\, dy \tag{3.24}$$

(per cui $\Phi(x) = P(X \le x)$, se $X \sim \mathcal{N}$, cfr. paragrafo 2.2) si trova tabulata in molti testi di probabilità e statistica (si veda ad esempio [1]). Per calcolare le probabilità di una distribuzione normale qualsiasi utilizzando la tabulazione della normale standard si usa la procedura di "standardizzazione". Sia $X \sim \mathcal{N}(\mu, \sigma^2)$ e consideriamo la sua standardizzata

$$Z = \frac{X - \mu}{\sigma}.$$

Alla fine del paragrafo 2.5 abbiamo visto che la standardizzata di una v.a. ha sempre media 0 e varianza 1. In questo caso si può dire di più: Z ha ancora distribuzione normale.

Proposizione 3.16. *Se $X \sim \mathcal{N}(\mu, \sigma^2)$ e $Z = (X - \mu)/\sigma$, allora $Z \sim \mathcal{N}$, ovvero la standardizzata di X ha distribuzione normale standard.*

Dimostrazione Usiamo le funzioni di ripartizione delle due v.a. . Si ha

$$F_Z(z) = P(Z \le z) = P\left(\frac{X - \mu}{\sigma} \le z\right) = P\left(X \le \sigma z + \mu\right).$$

Dunque, usando la proprietà (iii) della proposizione 2.7,

$$f_Z(z) = \frac{d}{dz} F_X(\sigma z + \mu) = \sigma f_X(\sigma z + \mu) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z/2}$$

e quindi $Z \sim \mathcal{N}$. □

Come si usano allora le tavole di Φ? Supponiamo di voler calcolare una probabilità del tipo $P(a \le X \le b)$, con $X \sim \mathcal{N}(\mu, \sigma)$. Sapendo che $Z = (X - \mu)/\sigma \sim \mathcal{N}$, si può scrivere

$$\begin{aligned} P(a \le X \le b) &= P\left(\frac{a - \mu}{\sigma} \le \frac{X - \mu}{\sigma} \le \frac{b - \mu}{\sigma}\right) = P\left(\frac{a - \mu}{\sigma} \le Z \le \frac{b - \mu}{\sigma}\right) \\ &= P\left(Z \le \frac{b - \mu}{\sigma}\right) - P\left(Z \le \frac{a - \mu}{\sigma}\right) = F_Z\Big(\frac{b - \mu}{\sigma}\Big) - F_Z\Big(\frac{a - \mu}{\sigma}\Big) \\ &= \Phi\Big(\frac{b - \mu}{\sigma}\Big) - \Phi\Big(\frac{a - \mu}{\sigma}\Big). \end{aligned}$$

Esercizio 3.17. *Supponiamo che la durata del viaggio in auto tra Firenze e Bologna sia una v.a. normale X con media 1 ora e deviazione standard 18 minuti. Se devo essere a Bologna alle 8:30, a che ora devo partire per essere sicuro al 90% di non arrivare in ritardo?* □

3.7 Altre distribuzioni notevoli

Concludiamo la nostra carrellata di distribuzioni notevoli presentando tre distribuzioni continue, la $\mathcal{X}^2$ ("chi-quadro"), la "t di Student" e la "F di Fisher", il cui significato, di grande importanza in statistica, sarà chiarito in seguito.

Ricordiamo innanzitutto la definizione della *funzione Gamma di Eulero*:

$$\Gamma(x) = \int_0^{+\infty} \mathrm{e}^{-s}\, s^{x-1}\, ds, \qquad x > 0, \tag{3.25}$$

e le sue proprietà fondamentali

$$\Gamma(x+1) = x\,\Gamma(x), \qquad \Gamma(1) = 1, \qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{3.26}$$

Da queste si possono ricavare tutti i valori di Γ in corrispondenza dei numeri interi, sui quali la Γ risulta coincidere coi fattoriali,

$$\Gamma(n) = (n-1)!, \qquad n = 1, 2, \ldots,$$

e in corrispondenza di tutti i numeri semi-interi,

$$\Gamma\left(\frac{3}{2}\right) = \frac{\sqrt{\pi}}{2}, \quad \Gamma\left(\frac{5}{2}\right) = \frac{3}{2}\,\frac{\sqrt{\pi}}{2}, \quad \Gamma\left(\frac{7}{2}\right) = \frac{5}{2}\,\frac{3}{2}\,\frac{\sqrt{\pi}}{2}, \quad \ldots.$$

Definizione 3.18. Si definisce $\mathcal{X}^2$ *a n gradi di libertà*, con $n = 1, 2, \ldots$, la distribuzione continua caratterizzata dalla funzione densità[8]

$$\chi^2_n(x) = \begin{cases} \dfrac{2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\, \mathrm{e}^{-\frac{x}{2}}\, x^{\frac{n}{2}-1}, & \text{se } x > 0, \\ 0, & \text{se } x \le 0. \end{cases} \tag{3.27}$$

La distribuzione $\mathcal{X}^2$ a n gradi di libertà si indica con il simbolo $\mathcal{X}^2(n)$. □

Notiamo che la distribuzione $\mathcal{X}^2(n)$ dipende da un parametro intero $n = 1, 2, \ldots$. In figura 3.6 sono mostrati i grafici delle funzioni densità χ^2_n per i valori di n da 1 a 6. Media e varianza di una variabile aleatoria $X \sim \mathcal{X}^2(n)$ sono date da

$$E[X] = n, \qquad \mathrm{Var}[X] = 2n. \tag{3.28}$$

Definizione 3.19. Si definisce t *di Student a n gradi di libertà*, con $n = 1, 2, \ldots$, la distribuzione continua caratterizzata dalla funzione densità

$$\tau_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}. \tag{3.29}$$

[8] Il particolare coefficiente numerico che appare in questa espressione è legato alla misura della ipersuperficie sferica di raggio 1 in uno spazio n-dimensionale, che risulta essere data da $2\pi^{n/2}/\,\Gamma\left(\frac{n}{2}\right)$ (per $n = 1, 2, 3, 4\ldots$ sarà dunque pari a 2, 2π, 4π, $2\pi^2\ldots$).

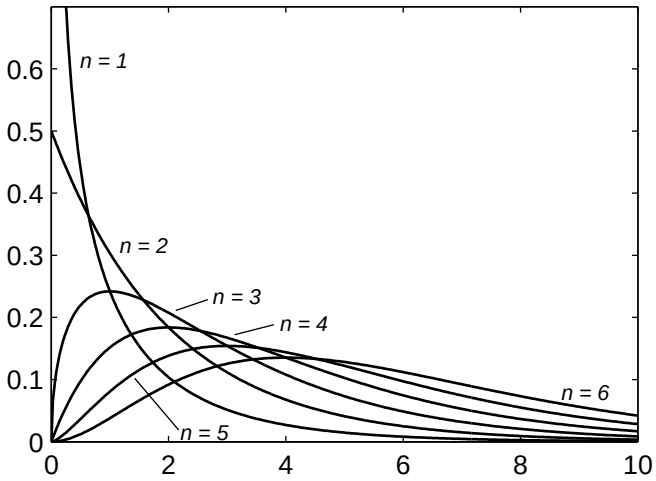


Figura 3.6: Grafici della parte $x > 0$ delle funzioni densità della distribuzione $\mathcal{X}^2(n)$, (3.27), per n da 1 a 5. Nel caso $n = 1$ la funzione ha un asintoto verticale per $x = 0$.

La distribuzione t di Student a n gradi di libertà si indica con il simbolo $t(n)$. □

Notiamo che anche la distribuzione $t(n)$ dipende da un parametro intero $n = 1, 2, \ldots$. In figura 3.7 sono mostrati i grafici delle funzioni densità τ_n per i valori $n = 1, 3$ e 10. Come vedremo nell'Esempio 4.29, al crescere di n, la distribuzione tende a una normale standard. Media e varianza di una variabile aleatoria $X \sim t(n)$ sono date da

$$E[X] = 0, \qquad \mathrm{Var}[X] = \frac{n}{n-2}, \quad \text{se } n \geq 3 \tag{3.30}$$

(la varianza non esiste finita se $n = 1, 2$).

Vediamo infine la distribuzione F di Fisher, che è una distribuzione dipendente da due gradi di libertà.

Definizione 3.20. Si definisce *F di Fisher con gradi di libertà n e m*, dove $n, m = 1, 2, \ldots$, la distribuzione continua caratterizzata dalla funzione densità

$$\varphi_{n,m}(x) = \begin{cases} \dfrac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \dfrac{n^{\frac{n}{2}}\, m^{\frac{m}{2}}\, x^{\frac{n-2}{2}}}{(nx+m)x^{\frac{n+m}{2}}} & \text{se } x > 0, \\ 0, & \text{se } x \leq 0. \end{cases} \tag{3.31}$$

La distribuzione F di Fisher con gradi di libertà n e m si indica con il simbolo $F(n, m)$ (o più spesso $F_{n,m}$). □

Media e varianza di una variabile aleatoria $X \sim F_{n,m}$ esistono se e solo se $m > 2$ e $m > 4$, rispettivamente, e sono date da

$$E[X] = \frac{m}{m-2}, \qquad \mathrm{Var}[X] = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}. \tag{3.32}$$

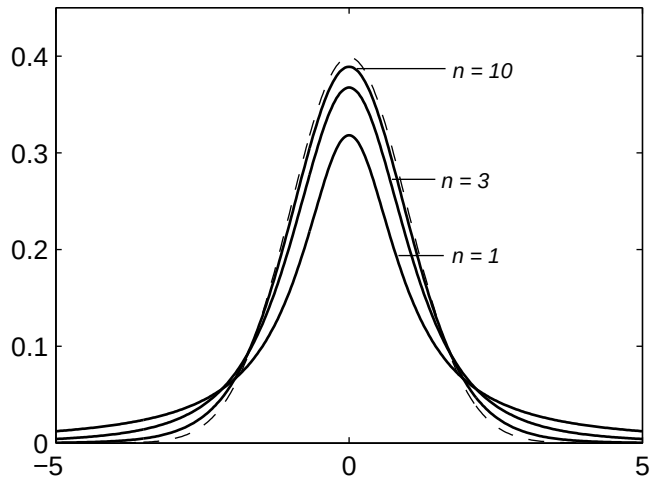


Figura 3.7: Grafici delle funzioni densità della distribuzione t-Student, (3.29), per $n = 1, 3, 10$. La curva tratteggiata è la densità normale standard, riportata per riferimento. Notare che per $n = 10$ la t di Student è già molto vicina alla normale standard, alla quale tende sempre più al crescere di n (si veda l'Esempio 4.29).

Nell'Esempio 4.19 spiegheremo il significato delle distribuzioni $\mathcal{X}^2$, t ed F, che può essere compreso solo alla luce della teoria delle v.a. multiple indipendenti. L'importanza pratica di queste tre distribuzioni apparirà chiaramente nella parte del corso dedicata alla statistica inferenziale.

Capitolo 4

Variabili aleatorie multiple

4.1 Variabili aleatorie multiple

La naturale estensione del concetto di v.a. semplice, o scalare, esaminato nei capitoli 3 e 4 è quella di *variabile aleatoria multipla*, o *vettore aleatorio*.

Definizione 4.1. Sia $P : S(\mathbb{R}^n) \to [0,1]$ una misura di probabilità sull'insieme delle n-uple di numeri reali $\mathbb{R}^n$. Allora in questo caso l'evento elementare è un vettore n-dimensionale $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ e si chiama *variabile aleatoria multipla di dimensione n*, o *vettore aleatorio n-dimensionale*. La misura di probabilità P si chiama *legge* o *distribuzione* di $\boldsymbol{X}$. □

Dunque una v.a. multipla è un punto casuale dello spazio $\mathbb{R}^n$. Osserviamo che ogni componente X_i di $\boldsymbol{X}$ è una v.a. semplice e che, ovviamente, una v.a. unidimensionale ($n = 1$) è una v.a. semplice.

In queste note ci soffermeremo spesso sulle variabili aleatorie *doppie*, ovvero con $n = 2$, per le quali useremo più spesso la notazione $\boldsymbol{X} = (X, Y)$ al posto di quella standard $\boldsymbol{X} = (X_1, X_2)$.

Esempio 4.2.
Consideriamo il lancio di due dadi come nell'esempio 1.3 e poniamo

$$X = \begin{cases} 0, & \text{se la somma dei due dadi è un numero pari,} \\ 1, & \text{se la somma dei due dadi è un numero dispari,} \end{cases}$$

$$Y = \begin{cases} 0, & \text{se la somma dei due dadi è uguale a 7,} \\ 1, & \text{se la somma dei due dadi è diversa da 7.} \end{cases}$$

Abbiamo dunque definito una v.a. doppia (X, Y) che può assumere i valori

$$(0,0), \quad (0,1), \quad (1,0), \quad (1,1).$$

È semplice verificare che

$$P((X,Y) = (0,0)) = 0, \qquad P((X,Y) = (0,1)) = 1/2,$$
$$P((X,Y) = (1,0)) = 1/6, \qquad P((X,Y) = (1,1)) = 1/3.$$

□

Esempio 4.3.: Campioni statistici
Introduciamo qui il concetto di *campione statistico*, del quale parleremo diffusamente nella seconda parte del corso.
Supponiamo, ad esempio, di voler "fare una statistica" di un certa specie di pesci che popolano un lago. Cominciamo col catturare a caso un solo pesce, del quale registriamo un certo numero m di dati, come lunghezza, peso, età, sesso, ecc. L'insieme degli eventi elementari costituiscono dunque una v.a. multipla di dimensione m:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_m)$$

(i dati sono numeri o possono essere ricondotti a numeri, come ad esempio: maschio = 1, femmina = 0). Se poi, come di fatto avviene, anziché un solo pesce se ne pescano n, e per ciascuno di questi si musurano le m caratteristiche, ci troviamo di fronte a una v.a. aleatoria di dimensione $n \times m$ che possiamo organizzare come una matrice:

$$\boldsymbol{X} = \begin{pmatrix} X_1^1 & X_2^1 & \cdots & X_m^1 \\ X_1^2 & X_2^2 & \cdots & X_m^2 \\ \vdots & \vdots & & \vdots \\ X_1^n & X_2^n & \cdots & X_m^n \end{pmatrix}.$$

Notiamo che ogni riga della matrice è una v.a. m-dimensionale

$$\boldsymbol{X}^i = (X_1^i, X_2^i, \ldots, X_m^i)$$

che contiene i dati di un singolo individuo mentre ogni colonna è una v.a. n-dimensionale

$$\boldsymbol{X}_j = \begin{pmatrix} X_j^1 \\ X_j^2 \\ \vdots \\ X_j^n \end{pmatrix}$$

che contiene i valori del j-esimo dato per tutti gli individui. La matrice $\boldsymbol{X}$ è detta *campione statistico*, le righe $\boldsymbol{X}^i$ sono dette *individui* e le colonne $\boldsymbol{X}_j$ sono dette *variabili*. In statistica si fa spesso l'ipotesi che gli individui, ovvero le v.a. $\boldsymbol{X}^1$, $\boldsymbol{X}^2$, ..., $\boldsymbol{X}^n$, siano *indipendenti* (si veda il paragrafo 4.4) e che abbiano tutti la stessa distribuzione. □

Passiamo adesso a definire l'importante concetto di distribuzione marginale. Abbiamo già osservato che ciascuna componente di una v.a. multipla è una v.a. semplice. Più in generale, data una v.a. multipla n-dimensionale $\boldsymbol{X}$, ogni sottogruppo di k delle sue componenti, con $1 \leq k \leq n$, è una v.a. multipla k-dimensionale che potremmo chiamare "sotto-variabile" della $\boldsymbol{X}$. Ad esempio, gli individui e le variabili di un campione statistico (esempio 4.3) sono sotto-variabili di dimensione m ed n, rispettivamente, della v.a. $n \times m$-dimensionale costistuita dall'intero campione.

Definizione 4.4. Data una v.a. n-dimensionale $\boldsymbol{X}$, le distribuzioni delle sue sotto-variabili k-dimensionali sono dette *distribuzioni marginali k-dimensionali* della distribuzione di $\boldsymbol{X}$. □

Molto spesso, parlando di distribuzioni marginali, si intendono quelle unidimensionali, ovvero le distribuzioni delle singole componenti: se $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ è una v.a. n-dimensionale, l'i-esima marginale è la distribuzione della v.a. scalare X_i. Notiamo che tale distribuzione (prendendo $i = 1$ per semplicità) è data da

$$P(X_1 \in A) = P\left(\boldsymbol{X} \in A \times \mathbb{R}^{n-1}\right), \quad \text{per ogni } A \in \sigma(\mathbb{R}),$$

ovvero dalla probabilità che X_1 stia nel sottoinsieme $A \subset \mathbb{R}$ e che le altre componenti abbiano un valore qualsiasi. Considerazioni del tutto analoghe si possono fare nel caso delle marginali k-dimensionali.

Osservazione 4.5. Come nel caso scalare, data una v.a. multipla $\boldsymbol{X}$ di dimensione n, se ne può considerare la *funzione di ripartizione*, che è una funzione $F_{\boldsymbol{X}} : \mathbb{R}^n \to [0, 1]$ così definita:

$$F_{\boldsymbol{X}}\left((x_1, x_2, \ldots, x_n)\right) := P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \tag{4.1}$$

e che gode di proprietà analoghe a quelle del caso scalare (si veda [3] per maggiori dettagli). □

4.2 Variabili aleatorie multiple discrete e continue.

Se si pensa all'esempio 4.3 ci si rende subito conto che l'apparire di v.a. che non sono né completamente discrete né completamente continue è ben più comune nel caso multiplo che nel caso semplice: nel campione statistico dell'esempio, che è una variabile aleatoria multipla $\boldsymbol{X}$, c'è una compresenza di componenti continue (lunghezze, pesi, ecc.) e componenti discrete (sesso, età, ecc.). Tuttavia il concentrarsi sulle due categorie principali, quella delle v.a. discrete e quella delle v.a. continue, come abbiamo già fatto nel caso scalare, è comunque di grande utilità sia teorica che pratica.

Definizione 4.6. Una variabile aleatoria n-dimensionale $\boldsymbol{X}$ si dice *discreta* se per ogni $1 \leq i \leq n$ esiste un insieme numerabile di valori $\left\{x_1^{(i)}, x_2^{(i)}, x_3^{(i)}, \ldots\right\}$ ed esistono probabilità $p_{i_1 i_2 \ldots i_n}$, con

$$p_{i_1 \ldots i_n} \geq 0, \qquad \sum_{i_1} \cdots \sum_{i_n} p_{i_1 \ldots i_n} = 1\,, \tag{4.2}$$

tali che

$$P\Big(X_1 = x_{i_1}^{(1)}, X_2 = x_{i_2}^{(2)}, \ldots, X_n = x_{i_n}^{(n)}\Big) = p_{i_1 i_2 \ldots i_n}\,. \tag{4.3}$$

□

Anche in questo caso, la distribuzione di una v.a. discreta $\boldsymbol{X}$ è descritta dall'insieme dei valori (punti di $\mathbb{R}^n$) che possono essere assunti da $\boldsymbol{X}$ e dalle loro probabilità.

Osservazione 4.7. Le sommatorie che compaiono dell'eq. (4.2) sono da intendersi come normali somme se i possibili valori assunti da ciascuna compnente sono in numero finito (in questo caso $\boldsymbol{X}$ si dirà *finita*) mentre, se almeno una componente di $\boldsymbol{X}$ può assumere infiniti valori, la corrisponendente sommatoria è da intendersi come una *serie* numerica (e in questo caso $\boldsymbol{X}$ si dirà *infinita*). □

Vediamo di illustrare la definizione 4.6 nel caso di una v.a. doppia $\boldsymbol{X} = (X, Y)$. Siano $x_1, x_2, \ldots$ i valori assunti da X e $y_1, y_2, \ldots$ i valori assunti da Y. Allora descriveremo la distribuzione mediante le probabilità

$$p_{ij} = P(\boldsymbol{X} = (x_i, y_j)) = P(X = x_i, Y = y_j)\,, \tag{4.4}$$

dove sui numeri p_{ij} ci sono le condizioni

$$p_{ij} \geq 0, \qquad \sum_i \sum_j p_{ij} = 1. \tag{4.5}$$

Notiamo che le due distribuzioni marginali sono date da

$$\begin{aligned} P(X = x_i) &= P(X = x_i, Y \in \mathbb{R}) = \sum_j p_{ij} \\ P(Y = y_i) &= P(X \in \mathbb{R}, Y = y_j) = \sum_i p_{ij} \end{aligned} \tag{4.6}$$

Nel caso a dimensione n, le distribuzioni marginali k-dimensionali sono del tipo

$$P\left(X_1 = x_{i_1}^{(1)}, \ldots, X_k = x_{i_k}^{(k)}\right) = \sum_{i_{k+1}} \cdots \sum_{i_n} p_{i_1 \ldots i_n} \tag{4.7}$$

(dove abbiamo supposto per semplicità di notazione che la sotto-variabile sia data dalle prime k componenti di $\boldsymbol{X}$).

Passiamo ora ad occuparci del caso continuo. Lo strumento matematico richiesto per trattare questo caso è l'integrale di funzioni di più variabili (o integrale multiplo): il lettore può consultare [7] come testo di riferimento.

Definizione 4.8. Una variabile aleatoria n-dimensionale $\boldsymbol{X}$ si dice *(assolutamente) continua* se esiste una funzione integrabile $f_{\boldsymbol{X}} : \mathbb{R}^n \to [0, +\infty)$ tale che

$$P(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n, \tag{4.8}$$

per ogni $A \in S(\mathbb{R}^n)$. La funzione $f_{\boldsymbol{X}}$ si chiama *densità* della v.a. $\boldsymbol{X}$. □

Dunque, nel caso n-dimensionale, la probabilità di una v.a. continua si esprime come integrale di una funzione non-negativa di n variabili. In particolare, per una v.a. continua doppia $\boldsymbol{X} = (X, Y)$, tale probabilità si esprime come integrale di una funzione di due variabili:

$$P(\boldsymbol{X} \in A) = \int_A f_{\boldsymbol{X}}(x, y)\, dx\, dy, \tag{4.9}$$

per ogni $A \in \sigma(\mathbb{R}^2)$. In questo caso la probabilità dell'insieme A può essere visualizzata come volume del sottografico della funzione $f_{\boldsymbol{X}}(x,y)$ delimitato dalla regione A, ovvero $\big\{(x,y,z) \in \mathbb{R}^3 \mid (x,y) \in A,\ 0 \le z \le f_{\boldsymbol{X}}(x,y)\big\}$.
Se $\boldsymbol{X} = (X_1, \dots, X_n)$ è una v.a. continua n dimensionale, la distribuzione della sotto-variabile k-dimensionale $(X_1, \dots, X_k)$ è data da

$$\begin{aligned} P((X_1,\dots,X_k) \in B) &= P\big((X_1,\dots,X_k,X_{k+1},\dots,X_n) \in B \times \mathbb{R}^{n-k}\big) \\ &= \int_{B\times\mathbb{R}^{n-k}} f_{\boldsymbol{X}}(x_1,x_2,\dots,x_n)\,dx_1\cdots dx_n \\ = \int_B \left\{ \int_{\mathbb{R}^{n-k}} f_{\boldsymbol{X}}(x_1,\dots,x_k,x_{k+1},\dots,x_n)\,dx_{k+1}\cdots dx_n \right\} dx_1\cdots dx_k, \end{aligned}$$

per ogni $B \in \sigma(\mathbb{R}^k)$, e pertanto la v.a. $(X_1, \dots, X_k)$ è continua, con densità

$$f_{(X_1,\dots,X_k)}(x_1,\dots,x_k) = \int_{\mathbb{R}^{n-k}} f_{\boldsymbol{X}}(x_1,\dots,x_k,x_{k+1},\dots,x_n)\,dx_{k+1}\cdots dx_n. \tag{4.10}$$

In particolare, se $\boldsymbol{X} = (X,Y)$ le due densità marginali sono date da

$$f_X(x) = \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x,y)\,dy, \qquad f_Y(y) = \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x,y)\,dx. \tag{4.11}$$

Per le v.a. multiple continue valgono proprietà analoghe a quelle enunciate per il caso scalare nella proposizione 2.7:

(i) $P(\boldsymbol{X} = (x_1,\dots,x_n)) = 0$ (la probabilità di un singolo punto è nulla);

(ii) la funzione di ripartizione $F_{\boldsymbol{X}}$ è continua;

e, se inoltre la funzione di densità $f_{\boldsymbol{X}}$ è continua in un punto $(x_1,\dots,x_n)$, si ha

(iii) $f_{\boldsymbol{X}}(x_1,\dots,x_n) = \frac{\partial}{\partial x_1}\cdots\frac{\partial}{\partial x_n} F_{\boldsymbol{X}}(x_1,\dots,x_n)$

(si veda [3] per maggiori dettagli).

4.3 Covarianza

Se abbiamo una v.a. multipla $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$, possiamo certamente prendere in considerazione valore atteso e varianza di ciascuna delle componenti. Questi si calcoleranno, a seconda dei casi discreto o continuo, usando le opportune formule, (4.6), (4.7), (4.10), (4.11), per le distribuzioni marginali. Ad esempio, per una v.a. continua doppia $\boldsymbol{X} = (X,Y)$, i valori attesi e le varianze delle due componenti saranno date da

$$\begin{aligned} E[X] = \mu_X &= \int_{-\infty}^{+\infty} x\, f_{\boldsymbol{X}}(x,y)\,dx\,dy, \\ E[Y] = \mu_Y &= \int_{-\infty}^{+\infty} y\, f_{\boldsymbol{X}}(x,y)\,dx\,dy, \\ \mathrm{Var}[X] &= \int_{-\infty}^{+\infty} (x-\mu_X)^2\, f_{\boldsymbol{X}}(x,y)\,dx\,dy, \\ \mathrm{Var}[Y] &= \int_{-\infty}^{+\infty} (y-\mu_Y)^2\, f_{\boldsymbol{X}}(x,y)\,dx\,dy, \end{aligned} \tag{4.12}$$

(purché gli integrali esistano). Tuttavia il caso multidimensionale è ben più ricco di quello scalare in quanto, oltre a medie e varianze delle singole componenti, possiamo considerare indicatori che esprimano *relazioni tra diverse componenti*. Il più importante di questi è senza dubbio la *covarianza*.

Definizione 4.9. Sia $\boldsymbol{X} = (X, Y)$ una v.a. doppia. Posto $\mu_X = E[X]$ e $\mu_Y = E[Y]$, si definisce *covarianza* di X e Y il numero

$$\mathrm{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)], \tag{4.13}$$

cioè il valore atteso della v.a. scalare $(X - \mu_X)(Y - \mu_Y)$, purché esso esista. □

Osservazione 4.10. Notiamo che se $Y = X$ la covarianza non è altro che la varianza di X:

$$\mathrm{Cov}[X, X] = E\left[(X - \mu_X)^2\right] = \mathrm{Var}[X]. \tag{4.14}$$

Dunque il concetto di covarianza contiene quello di varianza come caso particolare. □

Lasciamo per esercizio al lettore il compito di esprimere la covarianza nei casi discreto e continuo. Parlando grossolanamente, la covarianza ci dà un'idea di quanto le due variabili X e Y siano "legate" tra loro. Notiamo infatti che la covarianza tende a essere grande (in valore assoluto) quando la X e la Y variano *contemporaneamente* rispetto alle loro medie.

Vediamo adesso alcune semplici, ma importanti, proprietà di media, varianza e covarianza.

Proposizione 4.11. *Sia (X, Y, Z) una v.a. e siano a e b due numeri reali qualunque. Valgono allora la proprietà di* linearità *della media*

$$E[aX + bY] = a\,E[X] + b\,E[Y] \tag{4.15}$$

e la proprietà di bi-linearità *della covarianza*

$$\begin{aligned} \mathrm{Cov}[aX + bY, Z] &= a\,\mathrm{Cov}[X, Z] + b\,\mathrm{Cov}[Y, Z], \\ \mathrm{Cov}[Z, aX + bY] &= a\,\mathrm{Cov}[Z, X] + b\,\mathrm{Cov}[Z, Y], \end{aligned} \tag{4.16}$$

da cui segue la proprietà della varianza

$$\mathrm{Var}[aX + bY] = a^2\mathrm{Var}[X] + 2ab\,\mathrm{Cov}[X, Y] + b^2\mathrm{Var}[Y]. \tag{4.17}$$

Dimostrazione Per brevità dimostriamo la linearità della media solo nel caso discreto. Se (X, Y) assume i valori (x_i, y_j) con probabilità p_{ij}, allora la v.a. $aX + bY$ assume i valori $ax_i + by_j$ con probabilità p_{ij}. Dunque si ha

$$E[aX + bY] = \sum_i \sum_j (ax_i + by_j)p_{ij} = \sum_i \sum_j ax_i p_{ij} + \sum_i \sum_j by_j p_{ij} =$$

$$a \sum_i x_i \sum_j p_{ij} + b \sum_j y_j \sum_i p_{ij} = a \sum_i x_i\, p_i^X + b \sum_j y_j\, p_j^Y = a\,E[X] + b\,E[Y]$$

(dove con p_i^X e p_j^Y si sono indicate le due distribuzioni marginali). La prima delle (4.16) segue dalla (4.15), infatti:

$$\begin{aligned}\mathrm{Cov}\left[aX+bY,Z\right] &= E\left[(aX+bY-a\mu_X-b\mu_Y)(Z-\mu_Z)\right]\\ &= E\left[a(X-\mu_X)(Z-\mu_Z)+b(Y-\mu_Y)(Z-\mu_Z)\right]=\\ &= aE\left[(X-\mu_X)(Z-\mu_Z)\right]+bE\left[(Y-\mu_Y)(Z-\mu_Z)\right]\\ &= a\,\mathrm{Cov}\left[X,Z\right]+b\,\mathrm{Cov}\left[Y,Z\right].\end{aligned}$$

La seconda delle (4.16) si può dimostrare analogamente, oppure si può far discendere dalla simmetria della covarianza ($\mathrm{Cov}\left[X,Y\right]=\mathrm{Cov}\left[Y,X\right]$). Infine, la (4.17) si ottiene da

$$\mathrm{Var}\left[aX+bY\right]=\mathrm{Cov}\left[aX+bY,aX+bY\right]$$

usando la bi-linearità della covarianza. □

Corollario 4.12. *Sia (X,Y) una v.a. doppia tale che X e Y abbiano varianza e covarianza finite. Allora vale la* disuguaglianza di Schwartz:

$$\mathrm{Cov}\left[X,Y\right]^2\leq\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]. \tag{4.18}$$

Dimostrazione Dalla proposizione precedente segue, in particolare, che

$$\mathrm{Var}\left[tX+Y\right]=t^2\mathrm{Var}\left[X\right]+2t\,\mathrm{Cov}\left[X,Y\right]+\mathrm{Var}\left[Y\right],$$

per ogni $t\in\mathbb{R}$. Poiché $\mathrm{Var}\left[tX+Y\right]\geq 0$, il trinomio $t^2\mathrm{Var}\left[X\right]+2t\,\mathrm{Cov}\left[X,Y\right]+\mathrm{Var}\left[Y\right]$ è sempre non-negativo e perciò il discriminante è non-positivo:

$$\mathrm{Cov}\left[X,Y\right]^2-\mathrm{Var}\left[X\right]\mathrm{Var}\left[Y\right]\leq 0,$$

da cui la tesi. □

Utilizzando il corollario 4.12 si può dimostrare che l'esistenza della varianza di X e Y implica l'esistenza della covarianza.

Una quantità legata alla covarianza è il *coefficiente di correlazione*:

$$\rho[X,Y]=\frac{\mathrm{Cov}\left[X,Y\right]}{\mathrm{Std}\left[X\right]\,\mathrm{Std}\left[Y\right]} \tag{4.19}$$

Notiamo che dalla disuguaglianza di Schwartz 4.18 si ha che il coefficiente di correlazione soddifa la seguente disuguaglianza:

$$|\,\rho[X,Y]\,|\leq 1. \tag{4.20}$$

Questa ci permette di individuare due casi-limite:

- $\rho[X,Y]=\pm 1$: fra X e Y c'è una *correlazione*, o anti-correlazione, completa;
- $\rho[X,Y]=0$: fra X e Y c'è assenza di correlazione, e perciò si dicono *incorrelate* (o anche *scorrelate*).

Se ora $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ è una v.a. n-dimensionale, possiamo considerare la covarianza di due componenti qualunque:

$$\mathrm{Cov}\,[X_i, X_j] = E\,[(X_i - \mu_i)(X_j - \mu_j)] \tag{4.21}$$

(dove, chiaramente, si è posto $\mu_i = E\,[X_i]$). Con le covarianze fra tutte le possibili coppie di componenti X_i, X_j, possiamo costruire una matrice $n \times n$, detta *matrice di covarianza*:

$$\begin{pmatrix} \mathrm{Cov}\,[X_1, X_1] & \mathrm{Cov}\,[X_1, X_2] & \cdots & \mathrm{Cov}\,[X_1, X_n] \\ \mathrm{Cov}\,[X_2, X_1] & \mathrm{Cov}\,[X_2, X_2] & \cdots & \mathrm{Cov}\,[X_2, X_n] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}\,[X_n, X_1] & \mathrm{Cov}\,[X_n, X_2] & \cdots & \mathrm{Cov}\,[X_n, X_n] \end{pmatrix}.$$

Osserviamo che

(i) la matrice è *simmetrica*, poiché $\mathrm{Cov}\,[X_i, X_j] = \mathrm{Cov}\,[X_j, X_i]$;

(ii) sulla diagonale ci sono le *varianze* delle singole componenti poiché, come abbiamo già osservato in precedenza, $\mathrm{Cov}\,[X_i, X_i] = \mathrm{Var}\,[X_i]$.

Ad esempio, per $n = 2$ la matrice di convarianza è

$$\begin{pmatrix} \mathrm{Var}\,[X] & \mathrm{Cov}\,[X, Y] \\ \mathrm{Cov}\,[X, Y] & \mathrm{Var}\,[Y] \end{pmatrix}.$$

Possiamo anche considerare la matrice dei coefficienti di correlazione

$$\rho[X_i, X_j] = \frac{\mathrm{Cov}\,[X_i, X_j]}{\mathrm{Std}\,[X_i]\ \mathrm{Std}\,[X_j]}, \tag{4.22}$$

che ha le seguenti proprietà:

(i) la matrice è *simmetrica*, poiché $\rho[X_i, X_j] = \rho[X_j, X_i]$;

(ii) ogni elemento soddisfa la disuguaglianza $|\rho[X_i, X_j]| \leq 1$ (si veda la 4.20);

(iii) sulla diagonale ci sono tutti 1, infatti si ha $\rho[X_i, X_i] = 1$.

Esercizio 4.13. *Mostrare che la matrice dei coefficienti di correlazione di $\boldsymbol{X}$ coincide con la matrice di covarianza della standardizzata $\tilde{\boldsymbol{X}}$ di $\boldsymbol{X}$, ovvero la v.a.*

$$\tilde{\boldsymbol{X}} = (\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n)$$

che ha per componenti le standardizzate delle componenti di $\boldsymbol{X}$ (nel senso della definizione 2.20). □

4.4 Indipendenza di variabili aleatorie

Definizione 4.14. Sia $\boldsymbol{X} = (X, Y)$ una v.a. doppia. Le variabili aleatorie X e Y si dicono *indipendenti* se

$$P((X,Y) \in A \times B) = P(X \in A)\, P(Y \in B)\,, \tag{4.23}$$

per ogni $A \in S(\mathbb{R})$ e $B \in S(\mathbb{R})$. □

Poiché $(X, Y) \in A \times B$ significa $X \in A$ *e* $Y \in B$, notiamo che X e Y sono indipendenti se, per qualsiasi coppia di insiemi $A \in S(\mathbb{R})$ e $B \in S(\mathbb{R})$, gli eventi $X \in A$ e $Y \in B$ sono indipendenti, nel senso della definizione 1.16. In parole povere, X e Y sono indipendenti se il valore assunto dall'una non ha alcuna influenza sul valore assunto dall'altra. Dal punto di vista "geometrico", pensando la probabilità come area, l'indipendenza di X e Y si può vedere così: l'area del rettangolo $R = A \times B$ è semplicemente il prodotto "base per altezza", $P(R) = P(A)\ P(B)$ (figura (4.1)).

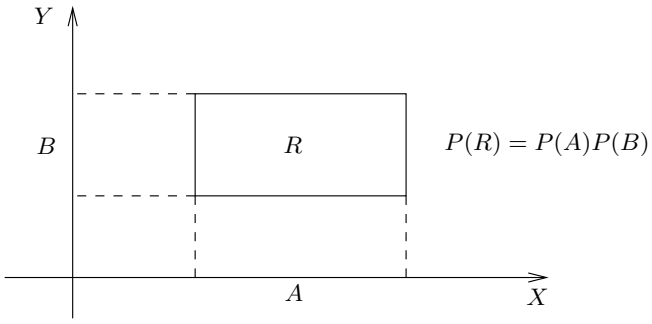


Figura 4.1: Visualizzazione geometrica dell'indipendenza.

Esempio 4.15.
Consideriamo il lancio di due dadi come negli esempi 1.3 e 4.2. È facile dimostrare che le due v.a. X e Y definite nell'esempio 4.2 *non* sono indipendenti. Infatti (prendendo ad esempio $A = \{0\}$ e $B = \{0\}$), si ha

$$P((X,Y) = (0,0)) = 0 \neq P(X = 0)\, P(Y = 0) = \frac{1}{2} \times \frac{1}{6} = \frac{1}{12}.$$

D'altra parte è chiaro, anche intuitivamente, che X e Y non possono essere indipendenti: se la somma dei due dadi è un numero pari, la somma dei due dadi non può fare 7. Se invece consideriamo le variabili aleatorie

$$X = \begin{cases} 0, & \text{se sul primo dado esce un numero pari,} \\ 1, & \text{se sul primo dado esce un numero dispari,} \end{cases}$$

$$Y = \begin{cases} 0, & \text{se sul secondo dado esce 5,} \\ 1, & \text{se sul secondo dado non esce 5,} \end{cases}$$

si verifica facilmente che queste *sono* indipendenti; ad esempio:

$$P((X,Y)=(0,0)) = \frac{1}{12} = P(X=0)\,P(Y=0) = \frac{1}{2}\times\frac{1}{6}$$

(la verifica andrebbe poi ripetuta per $(0,1)$, $(1,0)$ e $(1,1)$). D'altra parte l'indipendenza di X e Y si può intuire pensando che ciò che succede a un dado non ha influenza su ciò che accade all'altro. □

Proposizione 4.16. *Se X e Y sono indipendenti (e hanno varianza finita) si ha:*

1) $F_{(X,Y)}(x,y) = F_X(x)F_Y(y)$;

2) (caso discreto) $P((X,Y)=(x,y)) = P(X=x)\,P(Y=y)$;

2) (caso continuo) $f_{(X,Y)}(x,y) = f_X(x)\,f_Y(y)$;

3) $E[XY] = E[X]\,E[Y]$;

4) $\mathrm{Cov}[X,Y] = 0$;

5) $\rho[X,Y] = 0$;

6) $$\mathrm{Var}[aX+bY] = a^2\mathrm{Var}[X] + b^2\mathrm{Var}[Y] \tag{4.24}$$

Dimostrazione 1) Dalla definizione di funzione di ripartizione (4.1) e dalla definizione di indipendenza (4.23) si ottiene subito

$$F_{(X,Y)}(x,y) = P(X\le x,\, Y\le y) = P(X\le x)\,P(Y\le y) = F_X(x)F_Y(y).$$

2) (caso discreto) È immediata dalla definizione di indipendenza.

2) (caso continuo) Supponiamo per semplicità che la densità sia continua; allora, come osservato alla fine del paragrafo (4.2),

$$f_{(X,Y)}(x,y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{(X,Y)}(x,y).$$

Ma. poiché abbiamo già dimostrato che vale la 1), possiamo scrivere

$$\frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{(X,Y)}(x,y) = \frac{\partial}{\partial x}F_X(x)\,\frac{\partial}{\partial y}F_Y(y) = f_X(x)f_Y(y).$$

3) La si dimostra separatamente nei casi discreto e continuo. Ad esempio nel caso continuo, utilizzando la la 2), si ha

$$E[XY] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f_{(X,Y)}(x,y)\,dx\,dy = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f_X(x)f_Y(y)\,dx\,dy$$

$$= \int_{-\infty}^{+\infty} x\, f_X(x)\,dx \int_{-\infty}^{+\infty} y\, f_X(y)\,dy = E[X]\,E[Y]$$

4) Ricordando che $\mu_X = E[X]$ e $\mu_Y = E[Y]$, utilizzando la linearità della media si ha

$$\mathrm{Cov}[X,Y] = E[(X-\mu_X)(Y-\mu_Y)] = E[XY - \mu_X X - \mu_Y Y + \mu_X\mu_Y]$$

$$= E[XY] - \mu_X E[X] - \mu_Y E[Y] + \mu_X\mu_Y = E[XY] - \mu_X\mu_Y.$$

Ma dalla 3) si ha $E[XY] = \mu_X\mu_Y$ e dunque $\mathrm{Cov}[X,Y] = 0$.

5) È una conseguenza diretta della 4).

6) Segue dalla (4.17) e dalla 4). □

Osservazione 4.17. Come si è appena visto l'indipendenza di due v.a. X e Y implica la loro incorrelazione. Il viceversa, in generale, non è vero: ci sono esempi di v.a. incorrelate che non sono indipendenti. Tuttavia si può dimostrare che se (X,Y) ha distribuzione *normale* (si veda il paragrafo 4.5) allora l'incorrelazione *equivale* all'indipendenza. □

Esempio 4.18.: La v.a. binomiale come somma di v.a. Bernoulliane indipendenti
Nel paragrafo 3.1 abbiamo introdotto la ditsribuzione binomiale $\mathcal{B}(n,p)$, che ci dà la probabilità di avere k successi in n prove di tipo Bernoulliano, ciascuna con probabilità di successo p. Per ogni $i = 1,2,\ldots,n$, abbiamo definiamo la v.a. "successo all'i-esima prova":

$$X_i = \begin{cases} 1, & \text{se si ha "successo" all'}i\text{-esima prova,} \\ 0, & \text{se si ha "insuccesso" all'}i\text{-esima prova,} \end{cases}$$

che ha distribuzione Bernoulliana (3.1)

$$X_i \sim \mathcal{B}(p), \qquad i = 1,2,\ldots,n.$$

Allora l'esito delle n prove è descritto dal vettore aleatorio n-dimensionale $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ e la v.a. X "numero di successi in n prove" è data da

$$X = X_1 + X_2 + \cdots + X_n.$$

Poiché le n prove sono indipendenti, le X_i sono v.a. indipendenti. Utilizzando questo fatto si può ridimostrare per questa via che vale la (3.3). Inoltre, utilizzando la proprietà di linearità della media (4.15) e la proprietà (4.24) della varianza, dalla conoscenza di media e varianza (3.2) di una v.a. Bernoulliana, si può ricavare molto facilmente che $E[X] = np$ e $\mathrm{Var}[X] = np(1-p)$. □

Esempio 4.19.: Significato delle distribuzioni $\mathcal{X}^2$, t e F.
Abbiamo incontrato le distribuzioni $\mathcal{X}^2$, t e F nel paragrafo 3.7: ora siamo in grado precisarne il significato.[1]

Significato di $\mathcal{X}^2$. Sia $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$ una v.a. n-dimensionale tale che

[1] Senza però fornire dimostrazioni, che si possono trovare in [1].

(i) ogni componente ha distribuzione normale standard: $Z_i \sim \mathcal{N}$;

(ii) le n componenti sono *indipendenti*.

Allora la variabile aleatoria $Z_1^2 + Z_2^2 + \cdots + Z_n^2$ ha distribuzione $\mathcal{X}^2$ a n gradi di libertà:

$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \mathcal{X}^2(n).$$

Significato di $t(n)$. Sia $Z \sim \mathcal{N}$ e $Y \sim \mathcal{X}^2(n)$ e supponiamo che Z e Y siano indipendenti. Allora la v.a. $\sqrt{n}Z/\sqrt{Y}$ ha distribuzione t di Student a n gradi di libertà:

$$\frac{\sqrt{n}\,Z}{\sqrt{Y}} \sim t(n).$$

Significato di $F_{n,m}$. Se $X \sim \mathcal{X}^2(n)$ e $Y \sim \mathcal{X}^2(m)$ e supponiamo che X e Y siano indipendenti. Allora la v.a. mX/nY ha distribuzione F di Fisher con gradi di libertà n e m:

$$\frac{mX}{nY} \sim F_{n,m}\,.$$

4.5 La distribuzione normale multipla

Nel capitolo 3 abbiamo visto numerosi esempi di distribuzioni notevoli di v.a. semplici. Nel caso di v.a. multiple ci limiteremo ad illustrare un unica, ma molto imporatante, distribuzione notevole: la *distribuzione multinormale*.
Cominciamo col caso bidimensionale.

Definizione 4.20. Si chiama *binormale* (o *normale doppia* o *normale bivariata*) la distribuzione di una v.a. doppia (X, Y), continua, con densità

$$g(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$
$$\times \exp\left\{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x-\mu_1}{\sigma_1}\frac{y-\mu_2}{\sigma_2} + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right]\right\}, \quad (4.25)$$

dove μ_1, μ_2, $\sigma_1 > 0$, $\sigma_2 > 0$ e $-1 < \rho < 1$ sono parametri assegnati. □

Analogamente a quanto accade per la normale semplice (si veda il paragrafo 3.6) i parametri della distribuzione binormale hanno un significato probabilistico immediato, come asserito dalla seguente proposizione.

Proposizione 4.21. *Sia (X, Y) una v.a. con distribuzione bimormale, data dalla definizione precedente. Allora:*

(i) la $g(x, y)$ è effettivamente una densità, ovvero

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\,dx\,dy = 1;$$

(i) le due marginali sono normali semplici e precisamente:

$$X \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad Y \sim \mathcal{N}(\mu_2, \sigma_2^2);$$

(ii) ρ è il coefficiente di correlazione fra X e Y.

Dimostrazione Ponendo

$$\overline{x} = \frac{x - \mu_1}{\sigma_1}, \qquad \overline{y} = \frac{y - \mu_2}{\sigma_2} \tag{4.26}$$

possiamo scrivere

$$\begin{aligned} g(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{\overline{x}^2 - 2\rho\,\overline{x}\,\overline{y} + \overline{y}^2}{2(1-\rho^2)}\right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{(1-\rho^2)\overline{x}^2 + (\overline{y} - \rho\overline{x})^2}{2(1-\rho^2)}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{\overline{x}^2}{2}\right) \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \exp\left[-\frac{(\overline{y} - \rho\overline{x})^2}{2(1-\rho^2)}\right] \end{aligned} \tag{4.27}$$

Dunque, per dimostrare la (i) possiamo utilizzare il cambio di variabile (4.26), per cui $dx\,dy = \sigma_1\sigma_2\,d\overline{x}\,d\overline{y}$, e utilizzare la (4.26) ottenendo

$$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)\,dx\,dy = \frac{1}{2\pi\sqrt{1-\rho^2}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \exp\left\{-\frac{\overline{x}^2 - 2\rho\,\overline{x}\,\overline{y} + \overline{y}^2}{2(1-\rho^2)}\right\} d\overline{x}$$

$$= \int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(\overline{y} - \rho\overline{x})^2}{2(1-\rho^2)}\right] d\overline{y} \right\} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\overline{x}^2}{2}\right) d\overline{x}.$$

Utilizzando la proprietà (i) della proposizione (3.14) è facile dimostrare che quest'ultimo integrale fa 1.

Per dimostrare la (ii) notiamo che dall'uguaglianza (4.27) segue

$$g(x,y) = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] \frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \exp\left[-\frac{(y - \mu_2 - \rho\sigma_2\overline{x})^2}{2\sigma_2^2(1-\rho^2)}\right]$$

e che, fissata x, la funzione di y

$$\frac{1}{\sqrt{2\pi\sigma_2^2(1-\rho^2)}} \exp\left[-\frac{(y - \mu_2 - \rho\sigma_2\overline{x})^2}{2\sigma_2^2(1-\rho^2)}\right]$$

è la densità di una distribuzione normale con media $\mu_2 + \rho\sigma_2\overline{x}$ e varianza $\sigma_2^2(1-\rho^2)$ e perciò il suo integrale su y fra $-\infty$ e $+\infty$ fa 1. Si ha pertanto

$$\int_{-\infty}^{+\infty} g(x,y)\,dy = \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right],$$

ovvero $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$. La dimostrazione di $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ è analoga.

Per dimostrare la (iii), infine, ricordiamo la definizione (4.19), per cui

$$\rho(X,Y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{x-\mu_1}{\sigma_1} \frac{x-\mu_1}{\sigma_1} g(x,y)\, dx\, dy.$$

Procedendo col cambio di variabili (4.26) e utilizzando di nuovo l'uguaglianza (4.27) risulta che quest'ultimo integrale è equivalente a

$$\int_{-\infty}^{+\infty} \left\{ \int_{-\infty}^{+\infty} \frac{\overline{y}}{\sqrt{2\pi(1-\rho^2)}} \exp\left[-\frac{(\overline{y}-\rho\overline{x})^2}{2(1-\rho^2)} \right] d\overline{y} \right\} \frac{\overline{x}}{\sqrt{2\pi}} \exp\left(-\frac{\overline{x}^2}{2} \right) d\overline{x}.$$

Notiamo che l'espressione fra parentesi graffe è la media di una distribuzione normale con media $\rho\overline{x}$ e varianza $1-\rho^2$, e quindi, per la (ii) della proposizione 3.14, questa espressione è uguale a $\rho\overline{x}$. Si ha pertanto

$$\rho(X,Y) = \rho \int_{-\infty}^{+\infty} \frac{\overline{x}^2}{\sqrt{2\pi}} \exp\left(-\frac{\overline{x}^2}{2} \right) d\overline{x} = \rho,$$

dove l'ultima uguaglianza segue dalla (iii) della proposizione 3.14, in quanto l'integrale è la varianza di una distribuzione normale standard. □

Osservazione 4.22. Se $\rho = 0$ le due gaussiane marginali sono indipendenti e la $g(x,y)$ è il prodotto delle loro densità:

$$g(x,y) = \frac{1}{\sqrt{2\pi\sigma_1^2}}\, \mathrm{e}^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} \frac{1}{\sqrt{2\pi\sigma_2^2}}\, \mathrm{e}^{-\frac{(y-\mu_2)^2}{2\sigma_2^2}}. \tag{4.28}$$

Questo fatto lo si può ricavare direttamente dalla (4.25) ponendo $\rho = 0$ oppure utilizzando la (ii) e la (iii) della proposizione precedente nonché la 2 della proposizione 4.16. □

In figura 4.2 sono mostrati i grafici della funzione $g(x,y)$ per valori di μ_1, μ_2, σ_1 e σ_2 fissati ma con due diversi valori di ρ. I garfici delle due distribuzioni marginali unidimensionali sono due curve gaussiane, che abbiamo riportato per riferimento sulle pareti verticali delle figure. Notiamo che le distribuzioni marginali non cambiano nelle due figure in quanto non dipendono da ρ. Notiamo inoltre che, con l'aumentare di ρ, la densità tende a disporsi lungo la retta $y = x$ in quanto la correlazione positiva aumenta (invece, per ρ negativi la densità tenderebbe a disporsi lungo la retta $y = -x$: correlazione negativa).

Proposizione 4.23. *Sia (X,Y) una v.a. binormale con X e Y indipendenti e assumiamo $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Allora per ogni $\alpha, \beta \in \mathbb{R}$, si ha*

$$\alpha\, X + \beta\, Y \sim \mathcal{N}\left(\alpha\, \mu_1 + \beta\, \mu_2,\, \alpha^2\sigma_1^2 + \beta^2\sigma_2^2 \right).$$

Dimostrazione Diamo solo un cenno della dimostrazione, i cui dettagli sono un po' laboriosi. Posto $Z = \alpha\, X + \beta\, Y$, la funzione di ripartizione di Z è data da

$$F_Z(z) = P(Z \le z) = P(\alpha\, X + \beta\, Y \le z) = P\left(X \le \frac{z - \beta\, Y}{\alpha} \right)$$

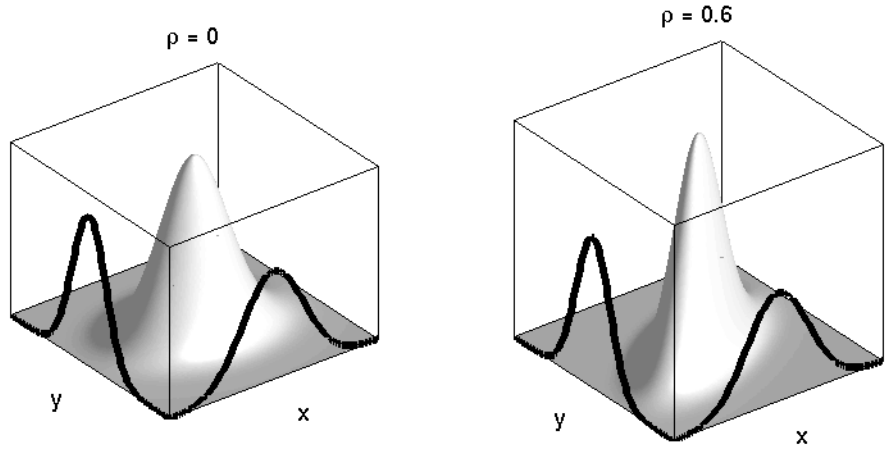


Figura 4.2: Grafici della densità normale bivariata $g(x, y)$ per due differenti valori di ρ (tutti gli altri parametri sono mantenuti costanti). Sulle pareti dei due grafici sono riportate le curve corrispondenti alle due distribuzioni marginali.

$$= \int_{-\infty}^{+\infty} \int_{-\infty}^{\frac{z-\beta\, y}{\alpha}} f_{(X,Y)}(x, y)\, dx\, dy = \int_{-\infty}^{+\infty} f_Y(y) \int_{-\infty}^{\frac{z-\beta\, y}{\alpha}} f_X(x)\, dx\, dy$$

(dove si è usata l'indipendenza di X e Y). Utilizzando la proprietà (iii) della proposizione 2.7, la densità di Z è data dalla derivata di F_Z:

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{-\infty}^{+\infty} f_Y(y) \frac{d}{dz} \int_{-\infty}^{\frac{z-\beta\, y}{\alpha}} f_X(x)\, dx\, dy$$

$$= \frac{1}{\alpha} \int_{-\infty}^{+\infty} f_Y(y)\, f_X\Big(\frac{z - \beta\, y}{\alpha}\Big)\, dy.$$

Sostituendo le due espressioni esplicite per le densità f_X e f_Y, dopo un po' di passaggi algebrici e opportuni cambi di variabile si arriva a un'espressione per $f_Z(z)$ in cui si riconosce la densità di una v.a. normale con media $\alpha\, \mu_1 + \beta\, \mu_2$ e varianza $\alpha^2\sigma_1^2 + \beta^2\sigma_2^2$. Notiamo che sarebbe sufficiente accorgersi che Z ha distribuzione normale: media e varianza sono allora date dalla proprietà di linearità della media e dalla proprietà 6), proposizione 4.16, della varianza. □

Esercizio 4.24. *Supponiamo che la durata del viaggio in auto tra Firenze e Bologna sia una v.a. normale X con media 1 ora e deviazione sandard di 18 minuti e che la durata dello stesso viaggio in treno sia una v.a. normale Y con media 1 ora e deviazione standard di 12 minuti. Se devo essere a Bologna alle 8:30 e il treno parte alle 7:25,*

1. *con che probabilità il treno arriva dopo le 8:30?*

2. *se parto in auto alle 7:10, con che probabilità arrivo prima del treno?*

[Suggerimento: si utilizzi il risultato della precedente proposizione] □

Finora abbiamo descritto la distribuzione normale in dimensione 2. Più in generale si può considerare la distribuzione *multinormale* (o *normale multipla* o *normale multivariata*) a dimensione n. Per definizione, questa è la distribuzione di una v.a. continua n-dimensionale $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ con funzione densità

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\,\Sigma^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})^T\right\} \tag{4.29}$$

dove abbiamo introdotto le seguenti notazioni:

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n), \quad \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n), \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1n} & \sigma_{n2} & \cdots & \sigma_{nn} \end{pmatrix},$$

con $\Sigma^T = \Sigma$ (simmetrica) e $\Sigma > 0$ (definita positiva[2]). Dunque la distribuzione multinormale dipende da $n + \frac{n(n+1)}{2}$ parametri (le n componenti di $\boldsymbol{\mu}$ e le $\frac{n(n+1)}{2}$ componenti indipendenti della matrice simmetrica Σ). La proposizione 4.21 ha la seguente generalizzazione.

Proposizione 4.25. *Sia $\boldsymbol{X}$ una v.a. multinormale con densità $f_{\boldsymbol{X}}$ data dalla* (4.29)*. Allora:*

(i) la $f_{\boldsymbol{X}}$ è effettivamente una densità, ovvero $\int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{x})\,d\boldsymbol{x} = 1$;

(i) le marginali unidimensionali sono normali semplici e precisamente:

$$X_i \sim \mathcal{N}(\mu_i, \sigma_i^2), \qquad i = 1, 2, \ldots, n;$$

(ii) Σ è la matrice di covarianza di $\boldsymbol{X}$.

Per la dimostrazione si veda [3].

4.6 Teoremi di convergenza

I teoremi di convergenza riguardano lo studio di *successioni* di variabili aleatorie. Una successione di variabili aleatorie è una collezione infinita e numerabile $X_1, X_2, X_3 \ldots$, di v.a. semplici X_i, che supporemo sempre *indipendenti* l'una dall'altra.[3]
In questo paragrafo presenteremo brevemente due risultati fra i più celebri: la *legge dei grandi numeri* e il *teorema centrale di convergenza*. Si tenga presente, comunque, che non enunceremo questi teoremi nella loro forma più generale,

[2] Ricordiamo che una matrice A, $n \times n$, si dice *definita positiva* se $\boldsymbol{v}A\boldsymbol{v}^T > 0$ per ogni $\boldsymbol{v} \in \mathbb{R}^n$ con $\boldsymbol{v} \neq 0$. Si veda anche il paragrafo 5.4.

[3] Questa definizione è un po' semplificata: dovremmo in realtà precisare che esiste una distribuzione "complessiva" per l'intera successione, di cui le distribuzioni di ciascuna X_i sono marginali.

essendo numerose (e difficili) le loro possibili generalizzazioni. Il lettore interessato può consultare il testo di Dall'Aglio [3] per avere una panoramica più approfondita.

Teorema 4.26. (Legge dei grandi numeri)
Sia $X_1, X_2, X_3 \dots$ una successione di v.a. indipendenti, tutte con la stessa media μ e la stessa varianza σ^2. Allora, posto

$$Y_n = \frac{X_1 + X_2 + \cdots + X_n}{n},$$

si ha che $Y_n \to \mu$ per $n \to \infty$, nel senso che

$$\lim_{n\to\infty} P(|Y_n - \mu| \geq \epsilon) = 0, \tag{4.30}$$

per ogni $\epsilon > 0$.

Dimostrazione Come si deduce facilmente dalla proprietà di linearità della media e dalla proprietà (4.17) della varianza, si ha

$$E[Y_n] = \mu, \qquad \mathrm{Var}[Y_n] = \frac{\sigma^2}{n}.$$

Perciò, usando la disuguaglianza di Cebicev (2.17), si ottiene che vale

$$P(|Y_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\,\epsilon^2}$$

per ogni $\epsilon > 0$, da cui segue subito la (4.30). □

Cosa ci dice la legge dei grandi numeri? Ci dice che "alla lunga" la media probabilistica μ è approssimata dalla "media statistica" Y_n, nel senso sopra enunciato. Ad esempio consideriamo lanci della moneta ripetuti: i risultati dei lanci sono una successione di v.a. identiche e indipendenti X_i con distribuzione

$$P(X_i = 1) = 1/2 \quad \text{(testa)}, \qquad P(X_i = 0) = 1/2 \quad \text{(croce)},$$

per cui Y_n è il rapporto fra il numero di teste e il numero di lanci. Secondo il teorema sopra dimostrato, quando il numero dei lanci diventa molto grande, la probabilità che il rapporto tra numero di teste e numero di lanci si discosti da $1/2$ per un numero ϵ, piccolo quanto si vuole, tende a 0:

$$\lim_{num.\ di\ lanci \to \infty} P\left(\left|\frac{numero\ di\ teste}{numero\ di\ lanci} - \frac{1}{2}\right| \geq \epsilon\right) = 0.$$

Attenzione: questo *non* vuol dire che ci si debba aspettare che il numero di teste tenda a essere la metà del numero dei lanci. Anzi, si può dimostrare che

$$\lim_{num.\ di\ lanci \to \infty} P\left(\left|num.\ di\ teste - \frac{num.\ di\ lanci}{2}\right| \leq \delta\right) = 0$$

per ogni $\delta > 0$. Dunque, *in senso assoluto* il numero di teste può differire anche di molto dalla metà del numero di lanci (anzi, la probabilità che differisca di

poco tende a zero, come ci dice quest'ultima formula): è *in senso relativo* (cioè *in proporzione al numero totale di lanci*) che si differisce di poco dalla metà.

Esempio 4.27.: Decadimento radioattivo II
Nell'esempio (3.10) abbiamo preso in considerazione un grande numero N di atomi, ciascuno dei quali ha una probabilità $\mathrm{e}^{-\nu t}$ di non essere ancora decaduto al tempo t, e abbiamo detto che, intuitivamente, il numero totale di atomi non decaduti al tempo t sarà dato da $N\mathrm{e}^{-\nu t}$. Questa affermazione può essere giustificata rigorosamante con la Legge dei grandi numeri. Infatti, associamo all'i-esimo atomo la v.a.

$$X_i = \begin{cases} 1, & \text{se l'atomo non è ancora decaduto al tempo } t, \\ 0, & \text{se l'atomo è già decaduto al tempo } t, \end{cases}$$

che è di tipo Bernoulliano con probabilità di successo $\mathrm{e}^{-\nu t}$. La v.a. $N_t = X_1 + X_2 + \cdots + X_N$ è il numero di atomi non decaduti al tempo t. Poichè gli atomi sono tutti indipendenti, il teorema 4.26 ci dice che

$$\frac{N_t}{N} \to \mathrm{e}^{-\nu t} \quad \text{per } N \to \infty.$$

Dunque per N molto grandi si avrà $N_t \approx N\mathrm{e}^{-\nu t}$, come previsto. □

Passiamo ora a enunciare il *teorema centrale di convergenza* che, come ci suggerisce il nome stesso, è veramente un risultato cardine del calcolo delle probabilità. Enunciamo solamente il teorema poiché la sua dimostrazione richiederebbe l'introduzione di strumenti matematici oltre lo scopo di queste note.

Teorema 4.28. (Teorema centrale di convergenza)
Sia $X_1, X_2, X_3 \ldots$ una successione di v.a. indipendenti e identicamente distribuite (in particolare, tutte con la stessa media μ e la stessa varianza σ^2). Allora la standardizzata Z_n della variabile aleatoria $Y_n = \frac{X_1+X_2+\cdots+X_n}{n}$, ovvero

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sigma\sqrt{n}},$$

tende ad avere distribuzione normale standard per $n \to \infty$, cioè

$$\lim_{n\to\infty} P(Z_n \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \mathrm{e}^{-x^2/2}\, dx. \tag{4.31}$$

Il significato di questo teorema è il seguente: l'"effetto medio" di tante v.a. indipendenti e identicamente distribuite, si manifesta come una v.a. con distribuzione normale. Questo risultato spiega il motivo dell'onnipresenza della distribuzione normale in probabilità e statistica. Ci sembrano particolarmente illuminanti in questo senso le parole di Dall'Aglio ([3], Cap. VIII.6):

È particolarmente importante l'aspetto empirico di questo risultato. Il caso viene spesso considerato come il cumularsi degli effetti di numerosi fattori, singolarmente poco rilevanti, che sarebbe troppo complesso studiare singolarmente e nel loro combinarsi; in altre parole come la somma di variabili aleatorie nessuna delle quali ha, singolarmente un peso rilevante. È facile allora ammettere che le variabili aleatorie soddisfino [le ipotesi del teorema centrale di convergenza] e concludere che l'effetto del caso si manifesta come una variabile aleatoria normale. Le numerose conferme sperimentali dell'adattamento della distribuzione normale ai più svariati fenomeni hanno confermato la fiducia in questa conclusione [...]. Questa conclusione ha senz'altro una notevole validità pratica e giustifica entro certi limiti l'utilizzazione della distribuzione normale nelle applicazioni.

Se l'ipotesi di identica distribuzione può sembrare un po' troppo riduttiva per poter suffragare una simile interpretazione (ed in effetti lo sarebbe), questa in realtà non è davvero necessaria. Una versione più generale del teorema centrale di convergenza ci dice infatti che se abbiamo una successione di variabili aleatorie $X_1, X_2, X_3 \ldots$, indipendenti ma non necessariamente identiche, tali che ciasuna ha media $\mu_i = E[X_i]$ e varianza $\sigma_i^2 = \mathrm{Var}[X_i]$ finite, allora (sotto alcune ipotesi tecniche) la v.a.

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n - (\mu_1 + \mu_2 + \cdots + \mu_n)}{\sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2}}$$

tende ad avere distribuzione normale standard per $n \to \infty$ (si veda [3]).

Esempio 4.29.: Comportamento asintotico delle distribuzioni $\mathcal{X}^2$ e t Sia $X_1, X_2, X_3 \ldots$ una successione di v.a. normali standard $X_i \sim \mathcal{N}$ indipendenti. Dall'Esempio 4.19 sappiamo che $X_i^2 \sim \mathcal{X}^2(1)$ e quindi, dall'eq. (3.28),

$$E\left[X_i^2\right] = 1, \qquad \mathrm{Var}\left[X_i^2\right] = 2, \qquad i = 1, 2, \ldots$$

Il Teorema Centrale di Convergenza implica allora che

$$\frac{X_1^2 + X_2^2 + \cdots + X_n^2 - n}{\sqrt{2n}} \longrightarrow Z,$$

per $n \to \infty$, con $Z \sim \mathcal{N}$. Dunque, per n sufficientemente grande si ha

$$X_1^2 + X_2^2 + \cdots + X_n^2 \approx \sqrt{2n}\, Z + n,$$

cioè, in termini di distribuzioni,

$$\mathcal{X}^2(n) \approx \sqrt{2n}\,\mathcal{N} + n. \tag{4.32}$$

Se ora consideriamo due variabili indipendenti $X \sim \mathcal{N}$ e $Y \sim \mathcal{X}^2(n)$, da quanto appena visto segue che, per n sufficientemente grande,

$$\frac{\sqrt{n}\, X}{\sqrt{Y}} \approx \frac{\sqrt{n}\, X}{\sqrt{\sqrt{2n}\, Z + n}} = \frac{X}{\sqrt{\frac{\sqrt{2}}{\sqrt{n}}\, Z + 1}}$$

con $Z \sim \mathcal{N}$. Ma se n è ababstanza grande, il denominatore è prossimo a 1. Dunque, ricordando il significato della t di Student, visto nell'Esempio 4.19, possiamo scrivere in termini di distribuzioni:

$$t(n) \approx \mathcal{N}, \quad \text{per } n \text{ abbastanza grande.} \tag{4.33}$$

Quanto sia "abbastanza grande", dipende un po' dal contesto; di solito in statistica si tende a utilizzare $\mathcal{N}$ al posto di $t(n)$ da $n = 30$ in poi. □

Esempio 4.30.: La "macchina di Galton"
Immaginiamo un piano inclinato su cui sono conficcati dei chiodi disposti a scacchiera regolare, come in figura 4.3. Una pallina che scende dall'alto viene deviata casualmente dai chiodi che incontra e compie una "passeggiata aleatoria" del tipo di quella rappresentata in figura. Supponiamo che le file orizzontali di chiodi siano n e che la distanza orizzontale fra chiodo e chiodo sia 2ℓ. Supponiamo che il dispositivo (detto *macchina di Galton*) sia costruito in modo tale che, ad ogni rimbalzo sui chiodi, la pallina abbia uguale probabilità di andare a destra o a sinistra. Se dunque, per $i = 1, 2, \ldots, n$, definiamo le variabili aleatorie

$$X_i = \begin{cases} 1, & \text{se l'}i\text{-esimo rimbalzo manda la pallina a destra,} \\ -1, & \text{se l'}i\text{-esimo rimbalzo manda la pallina a sinistra,} \end{cases}$$

possiamo supporre che le X_i siano indipendenti e abbiano tutte la distribuzione $P(X_i = 1) = P(X_i = -1) = \frac{1}{2}$, per cui risulta $E[X_i] = 0$ e $\mathrm{Var}[X_i] = 1$.
Notiamo che la posizione finale della pallina sull'asse orizzontale è una v.a. (che chiameremo Z_n per sottolineare la dipendenza dal numero di file di chiodi, n) data da

$$Z_n = \ell X_1 + \ell X_2 + \cdots \ell X_n = \ell\,(X_1 + X_2 + \cdots X_n)\,.$$

Adesso immaginiamo di infittire sempre più le file di chiodi, n, e, contempora-

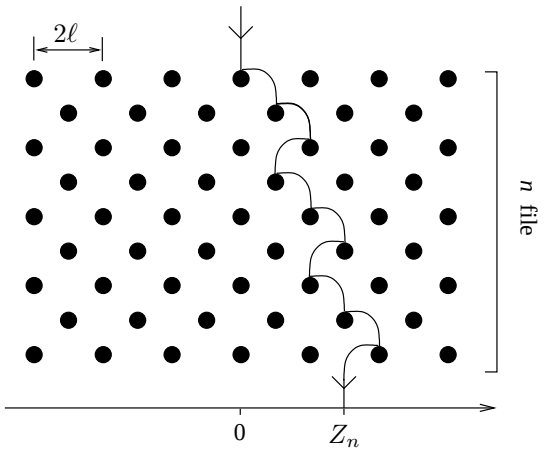


Figura 4.3: La macchina di Galton.

neamente, di diminuire la distanza ℓ in modo inversamente proporzionale a $\sqrt{n}$. Per semplicità possiamo prendere esattamente

$$\ell = 1/\sqrt{n}$$

e si ha perciò

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}.$$

Il teorema centrale di convergenza ci dice allora che

$$Z_n \to Z \sim \mathcal{N}, \qquad \text{per } n \to \infty;$$

in altre parole, al crescere di n, la posizione finale della pallina tende ad avere distribuzione normale standard.
Notiamo che se avessimo infittito i chiodi troppo poco rispetto al numero delle file prendendo, ad esempio, $\ell = 1/\sqrt[3]{n}$ o addirittura $\ell = 1$ (i chiodi rimangono sempre alla stessa distanza orizzontale), avremmo ottenuto $Z_n \to \infty$). Viceversa, se avessimo infittito troppo i chiodi, ad esempio come $\ell = 1/n$, avremmo ottenuto $Z_n \to 0$ (nel caso $\ell = 1/n$ è proprio la legge dei grandi numeri). □

Esercizio 4.31. *Un computer in ciascuna operazione di addizione commette un errore che è una v.a. continua, con distribuzione uniforme nell'intervallo* $[-0.5 \times 10^{-10}, 0.5 \times 10^{-10}]$. *Se il computer esegue un milione di somme, qual'è la probabilità che l'errore complessivo sia in valore assoluto minore di* 0.5×10^{-7}*?*

Parte II

Statistica

Sai ched'è la statistica? È 'na cosa
che serve pe' fa' un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che sposa.

Ma pe' me la statistica curiosa
è dove c'entra la percentuale,
pe' via che, lì, la media è sempre eguale
puro co' la persona bisognosa.

Me spiego, da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:

e, se nun entra ne le spese tue,
t'entra ne la statistica lo stesso
perché c'è un antro che se ne magna due.

A. Trilussa, *La statistica*

Capitolo 5

Statistica descrittiva

5.1 Concetti generali

La *statistica descrittiva* si occupa dell'analisi di dati raccolti da una "popolazione", senza porsi il problema di estrapolare o inferire alcunché al di fuori dei dati stessi. Invece la *statistica inferenziale* (di cui ci occuperemo più avanti) si avvale di metodi probabilistici per *inferire*, cioè dedurre, qualcosa che non è strettamente contenuta nei dati stessi.

Lo scopo della statistica descrittiva si può sintetizzare così: estrarre da un insieme anche molto grande di dati, delle informazioni il più possibile *sintetiche e significative* sui dati stessi. Come vedremo, a differenza di quella inferenziale, nella statistica descrittiva il calcolo delle probabilità non gioca alcun ruolo.

Le parole-chiave della statistica sono le seguenti:

Individuo Un singolo oggetto dell'indagine.

Popolazione L'insieme degli individui oggetto di indagine.

Variabile Una certa quantità misurata su ogni individuo.

Alcune precisazioni sono d'obbligo. Per "individuo" non si deve necessariamente intendere un singolo soggetto (una pianta, un fiore, un animale una persona) ma, genericamente, la singola "unità" su cui si fa la misura (si parla infatti anche di *unità di campionamento*). Ad esempio, se dividiamo un terreno in zone quadrate e siamo interessati al numero di piante di un certo tipo contenute in ognuna di esse, ogni zona sarà un individuo, il numero di piante in essa contenute sarà la variabile e l'insieme delle zone costituirà la popolazione. In questo esempio, ogni "individuo" in senso matematico può contenere più "individui" nel senso comune del termine. Anche il termine "misurare" può generare confusione: qui va inteso genericamente come "acquisire un'informazione"; in questo senso anche le interviste di un *exit-poll* sono "misurazioni".

L'insieme dei numeri[1] ottenuti misurando m variabili su n individui è detto *campione statistico* e può essere organizzato in una *matrice di dati* X, con n righe e m e colonne, in cui il numero x_j^i, che si trova all'incrocio tra la i-esima riga e la j-esima colonna, rappresenta la j-esima variabile misurata sull'i-esimo individuo:

$$X = \begin{array}{ccccccc} & & & \boldsymbol{x}_j & & & \\ & & & \downarrow & & & \\ x_1^1 & x_2^1 & \cdots & x_j^1 & \cdots & x_m^1 & \\ x_1^2 & x_2^2 & \cdots & x_j^2 & \cdots & x_m^2 & \\ \vdots & \vdots & & \vdots & & \vdots & \\ x_1^i & x_2^i & \cdots & x_j^i & \cdots & x_m^i & \leftarrow \boldsymbol{x}^i \\ \vdots & \vdots & & \vdots & & \vdots & \\ x_1^n & x_2^n & \cdots & x_j^n & \cdots & x_m^n & \end{array}$$

(si veda il paragrafo 5.4 per ulteriori informazioni sulle matrici). La j-esima *variabile* è dunque rappresentata dalla j-esima colonna della matrice X ed è un vettore n-dimensionale che indicheremo con $\boldsymbol{x}_j$ e che, per comodità, scriveremo spesso in orizzontale:

$$\boldsymbol{x}_j = (x_j^1, x_j^2, \ldots, x_j^n)$$

(si rimanda al paragrafo 5.3 per ulteriori informazioni sui vettori). L'i-esimo *individuo* è invece rappresentato dall'i-esima riga della matrice X ed è quindi un vettore m-dimensionale che indicheremo con $\boldsymbol{x}^i$:

$$\boldsymbol{x}^i = (x_1^i, x_2^i, \ldots, x_m^i).$$

La tabella 5.1 è un esempio storico di indagine statistica nelle scienze naturali. Si tratta di dati morfologici di 49 esemplari di passeri, alcuni sopravvissuti, altri deceduti, raccolti dopo il passaggio di un tifone dal biologo Hermon Bumpus nel 1898. L'indagine statistica intendeva mettere in evidenza le caratteristiche che avevano permesso ad alcuni individui di sopravvivere e ad altri no. La matrice di dati vera e propria ha $n = 49$ righe e m= 5 colonne. Torneremo su questo esempio nel paragrafo 5.5.

Concludiamo questa introduzione osservando che non abbiamo fatto (né faremo in seguito) il minimo accenno all'importantissimo argomento del *campionamento*, ovvero di come si acquisiscono i dati affinché questi siano significativi. *Un'indagine statistica non può essere considerata scientificamente valida se non si basa prima di tutto su un campionamento effettuato secondo criteri rigorosi.* D'altra parte la teoria del campionamento dipende molto dal contesto specifico dell'indagine statistica per cui in queste note, di carattere generale, preferiamo non affrontarla e rimandare il lettore a testi specifici (per la biologia e le scienze naturali. Ad esempio,si può consultare il testo di Camussi *et. al.* [2]).

[1] In realtà ci possono essere variabili, dette *nominali* o *categoriali*, che non sono numeriche (ad esempio: luogo di raccolta, titolo di studio, colore dei petali...). Evidentemente ci si può sempre ricondurre a variabili numeriche attribuendo un numero ad ogni categoria. Tuttavia in questo caso concetti come media e varianza (vedi il paragrafo 5.2) potrebbero non essere significativi.

n	lungh. totale	apert. alare	lungh. omero	lungh. becco-testa	lungh. sterno
1	156	245	31.6	18.5	20.5
2	154	240	30.4	17.9	19.6
3	153	240	31.0	18.4	20.6
4	153	236	30.9	17.7	20.2
5	155	243	31.5	18.6	20.3
6	163	247	32.0	19.0	20.9
7	157	238	30.9	18.4	20.2
8	155	239	32.8	18.6	21.2
9	164	248	32.7	19.1	21.1
10	158	238	31.0	18.8	22.0
11	158	240	31.3	18.6	22.0
12	160	244	31.1	18.6	20.5
13	161	246	32.3	19.3	21.8
14	157	245	32.0	19.1	20.0
15	157	235	31.5	18.1	19.8
16	156	237	30.9	18.0	20.3
17	158	244	31.4	18.5	21.6
18	153	238	30.5	18.2	20.9
19	155	236	30.3	18.5	20.1
20	163	246	32.5	18.6	21.9
21	159	236	31.5	18.0	21.5
22	155	240	31.4	18.0	20.7
23	156	240	31.5	18.2	20.6
24	160	242	32.6	18.8	21.7
25	152	232	30.3	17.2	19.8
26	160	250	31.7	18.8	22.5
27	155	237	31.0	18.5	20.0
28	157	245	32.2	19.5	21.4
29	165	245	33.1	19.8	22.7
30	153	231	30.1	17.3	19.8
31	162	239	30.3	18.0	23.1
32	162	243	31.6	18.8	21.3
33	159	245	31.8	18.5	21.7
34	159	247	30.9	18.1	19.0
35	155	243	30.9	18.5	21.3
36	162	252	31.9	19.1	22.2
37	152	230	30.4	17.3	18.6
38	159	242	30.8	18.2	20.5
39	155	238	31.2	17.9	19.3
40	163	249	33.4	19.5	22.8
41	163	242	31.0	18.1	20.7
42	156	237	31.7	18.2	20.3
43	159	238	31.5	18.4	20.3
44	161	245	32.1	19.1	20.8
45	155	235	30.7	17.7	19.6
46	162	247	31.9	19.1	20.4
47	153	237	30.6	18.6	20.4
48	162	245	32.5	18.5	21.1
49	164	248	32.3	18.8	20.9

Tabella 5.1: Misure di vari parametri morfologici di esemplari di passero raccolti o catturati da H. Bumpus dopo il passaggio di un tifone. Le misure sono espresse in millimetri. Gli esemplari da 1 a 21 sono sopravvissuti al tifone, gli altri sono deceduti (da B. Manly [8]).

5.2 Media, varianza e covarianza

Passiamo ora ad introdurre le più importanti fra quelle quantità "sintetiche e significative" di cui abbiamo parlato nel paragrafo precedente. Limitiamoci per il momento a *campioni monovariati*, ovvero campioni statistici con una sola variabile $\boldsymbol{x}$ (naturalmente, se abbiamo un campione *multivariato*, come quello della tabella 5.1, possiamo prendere una variabile alla volta e lavorare su ciascuna di esse come su un campione monovariato). Se la statistica è monovariata si può, ovviamente, eliminare l'indice di variabile e, più comodamente, mettere l'indice di individuo in basso, scrivendo perciò

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n).$$

Siano $V_1, V_2, \ldots V_N$, con $N \leq n$, i valori assunti dalla variabile $\boldsymbol{x}$ senza contare le ripetizioni. Ad ogni valore V_k si associa la *frequenza* f_k, cioè il numero di volte che quel valore viene assunto nella popolazione:

$$f_k = \text{numero degli } x_i \text{ uguali a } V_k. \tag{5.1a}$$

Se i valori x_i sono tutti distinti, oppure se le singole frequenze sono poco significative, si può fare una partizione delll'insieme in cui variano i valori x_i in intervalli $I_1, I_2, \ldots, I_N$ e definire la frequenza di f_k come il numero di valori x_i che cadono nell'intervallo I_k:

$$f_k = \text{ numero di } x_i \text{ che stanno nell'intervallo } I_k. \tag{5.1b}$$

Notiamo che in entrambi i casi appena descritti si ha

$$\sum_{k=1}^{N} f_k = n. \tag{5.2}$$

Quelle appena definite si dicono frequenze *assolute*. Le *frequenze relative* p_k sono invece le frequenze assolute divise per il numero di individui del campione, ovvero:

$$p_k = \frac{f_k}{n}. \tag{5.3}$$

Notiamo che, ovviamente, si ha

$$\sum_{k=1}^{N} p_k = 1. \tag{5.4}$$

Consideriamo ad esempio la seconda variabile della tabella 5.1 (apertura alare). Osserviamo che i valori variano fra un minimo di 230 e um massimo di 252. Possiamo, ad esempio, suddividere l'intervallo $[230, 252]$ nei 4 sottointervalli

$$I_1 = [230, 236),\ I_2 = [236, 242),\ I_3 = [242, 248),\ I_4 = [248, 252],$$

cui corrispondono le frequenze assolute

$$f_1 = 5,\ f_2 = 19,\ f_3 = 20,\ f_4 = 5,$$

e le frequenze relative

$$p_1 = \frac{5}{49} \approx 0.102, \quad p_2 = \frac{19}{49} \approx 0.388, \quad p_3 = \frac{20}{49} \approx 0.408, \quad p_4 = \frac{5}{49} \approx 0.102.$$

La quantità statistica per antonomasia è la *media* di una variabile $\boldsymbol{x}$. Per la verità esistono diversi tipi di medie: quella che viene detta semplicemente "media" è la *media aritmetica* $\overline{\boldsymbol{x}}$, che è così definita:

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{5.5}$$

Notiamo che la media aritmetica è quel valore che moltiplicato per n ci dà lo stesso risultato ottenuto sommando tutti i valori x_i. Usando le frequenze (ma solo nel senso della definizione (5.1a)) si può scrivere

$$\overline{\boldsymbol{x}} = \frac{1}{n} \sum_{k=1}^{N} f_k V_k = \sum_{k=1}^{N} p_k V_k. \tag{5.6}$$

Esercizio 5.1. *Verificare (utilizzando un calcolatore) che le medie delle 5 variabili della tabella* (5.1) *sono*

$$\overline{\boldsymbol{x}}_1 = 157.98 \quad \overline{\boldsymbol{x}}_2 = 241.33 \quad \overline{\boldsymbol{x}}_3 = 31.46 \quad \overline{\boldsymbol{x}}_4 = 18.47 \quad \overline{\boldsymbol{x}}_5 = 20.83.$$

□

La *media geometrica* $\overline{\boldsymbol{x}}_G$ di una variabile $\boldsymbol{x}$ è così definita:

$$\overline{\boldsymbol{x}}_G = (x_1 x_2 \cdots x_n)^{1/n}. \tag{5.7}$$

Notiamo che la media geometrica è quel valore che moltiplicato per se stesso n volte ci dà lo stesso risultato ottenuto moltiplicando tra loro tutti i valori x_i.

Esempio 5.2. Ci sono situazioni in cui la media geometrica è in indicatore migliore rispetto alla media aritmetica. Supponiamo ad esempio di osservare il tasso di inflazione per n anni. Siano F_1 il tasso d'inflazione al primo anno, F_2 il tasso d'inflazione al secondo anno, e così via. Il prezzo di un prodotto che segue l'inflazione, e che inizialmente aveva il prezzo P, dopo n anni ha subito un aumento pari a

$$\Delta P = F_1 F_2 \cdots F_n P.$$

Se per "inflazione media" intendiamo un'inflazione costante per n anni che alla fine ci dà lo stesso aumento di prezzo ΔP, allora è chiaro che questa è data dalla media geometrica $\overline{F}_G$ poiché, per definizione,

$$\overline{F}_G \, P = F_1 F_2 \cdots F_n P.$$

La media aritmetica delle F_i è un altro indicatore calcolabile, ma senza dubbio molto meno significativo. □

La *media armonica* $\overline{\boldsymbol{x}}_A$ di una variabile $\boldsymbol{x}$ è così definita:

$$\overline{\boldsymbol{x}}_A = n \left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n} \right)^{-1}, \tag{5.8}$$

ovvero

$$\frac{1}{\overline{\boldsymbol{x}}_A} = \frac{1}{n}\left(\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}\right). \tag{5.9}$$

Notiamo che la media armonica è quel valore tale che moltiplicando per n il suo inverso si ottiene la somma degli inversi degli x_i.

Esempio 5.3. Una ditta di trasporti ha n automezzi e conosce il consumo di ciascuno di essi (espresso ad esempio in kilometri per litro). Sia C_i il consumo dell'i-esimo automezzo. Ci si chiede: qual'è il "consumo medio" degli automezzi? Per ogni K kilometri percorsi da tutti gli automezzi, si consumano complessivamente

$$\frac{K}{C_1} + \frac{K}{C_2} + \cdots + \frac{K}{C_n} = K\left(\frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n}\right)$$

litri di carburante. Se tutti gli automezzi avessero lo stesso consumo C, il carburante consumato sarebbe semplicemnte nK/C. Il consumo medio è quel valore di C per cui si ha

$$\frac{nK}{C} = K\left(\frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n}\right).$$

Si vede quindi che il consumo medio è dato dalla media armonica $\overline{C}_A$. □

Esercizio 5.4. *Scrivere per la media geometrica e per quella armonica gli analoghi della formula* (5.6)*, ovvero esprimere* $\overline{\boldsymbol{x}}_G$ *e* $\overline{\boldsymbol{x}}_A$ *in termini di frequenze assolute e relative.* □

Sempre rimanendo nell'ambito della statistica monovariata, il secondo indicatore più importante dopo la media è la *varianza*[2]

$$\mathrm{Var}\,[\boldsymbol{x}] = \frac{1}{n-1}\sum_{k=1}^{n}(x_i - \overline{\boldsymbol{x}})^2 \tag{5.10}$$

La varianza indica quanto un dato tende a essere concentrato intorno alla media (varianza bassa), oppure disperso intorno ad essa (varianza alta).
La *deviazione standard* è per definizione la radice della varianza:

$$\mathrm{Std}\,[\boldsymbol{x}] = \sqrt{\mathrm{Var}\,[\boldsymbol{x}]}. \tag{5.11}$$

Il primo concetto di statistica multivariata che incontriamo è invece quello di covarianza. Se $\boldsymbol{x}$ e $\boldsymbol{y}$ sono due variabili prese da una matrice di dati statistici (può avere anche solo due colonne, ma non meno), si definisce la *covarianza di* $\boldsymbol{x}$ *e* $\boldsymbol{y}$ come

$$\mathrm{Cov}\,[\boldsymbol{x}, \boldsymbol{y}] = \frac{1}{n-1}\sum_{k=1}^{n}(x_i - \overline{\boldsymbol{x}})\,(y_j - \overline{\boldsymbol{y}}). \tag{5.12}$$

[2]Il motivo per cui nella definizione di varianza si utilizza il fattore $\frac{1}{n-1}$ (anziché, come ci si potrebbe aspettare, $\frac{1}{n}$) diverrà chiaro nella parte dedicata alla statistica inferenziale (si veda il paragrafo 6.3). In alcuni testi, tuttavia, si trova la definizione con il fattore $\frac{1}{n}$.

Notiamo che $\text{Cov}[\boldsymbol{x}, \boldsymbol{x}] = \text{Var}[\boldsymbol{x}]$. Il *coefficiente di correlazione tra* $\boldsymbol{x}$ *e* $\boldsymbol{y}$ è definito come

$$\rho[\boldsymbol{x}, \boldsymbol{y}] = \frac{\text{Cov}[\boldsymbol{x}, \boldsymbol{y}]}{\text{Std}[\boldsymbol{x}]\ \text{Std}[\boldsymbol{y}]}. \tag{5.13}$$

Se si ha una matrice di dati con m variabili, si possono calcolare le covarianze e i coefficienti di correlazione di due variabili alla volta, costruendo così la matrice di covarianza

$$\begin{pmatrix} \text{Var}[\boldsymbol{x}_1] & \text{Cov}[\boldsymbol{x}_1, \boldsymbol{x}_2] & \cdots & \text{Cov}[\boldsymbol{x}_1, \boldsymbol{x}_m] \\ \text{Cov}[\boldsymbol{x}_2, \boldsymbol{x}_1] & \text{Var}[\boldsymbol{x}_2] & \cdots & \text{Cov}[\boldsymbol{x}_2, \boldsymbol{x}_m] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[\boldsymbol{x}_m, \boldsymbol{x}_1] & \text{Cov}[\boldsymbol{x}_m, \boldsymbol{x}_2] & \cdots & \text{Var}[\boldsymbol{x}_m] \end{pmatrix}$$

e quella di correlazione

$$\begin{pmatrix} 1 & \rho[\boldsymbol{x}_1, \boldsymbol{x}_2] & \cdots & \rho[\boldsymbol{x}_1, \boldsymbol{x}_m] \\ \rho[\boldsymbol{x}_2, \boldsymbol{x}_1] & 1 & \cdots & \rho[\boldsymbol{x}_2, \boldsymbol{x}_m] \\ \vdots & \vdots & \ddots & \vdots \\ \rho[\boldsymbol{x}_m, \boldsymbol{x}_1] & \rho[\boldsymbol{x}_m, \boldsymbol{x}_2] & \cdots & 1 \end{pmatrix},$$

che sono entrambe matrici quadrate e simmetriche $m \times m$ (per la spiegazione di questa terminologia si veda il paragrafo 5.4).

5.3 Richiami di algebra lineare: vettori

Un *vettore n-dimensionale* è una n-upla di numeri reali

$$\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$$

gli n numeri reali v_i si chiamano *componenti*, o *elementi*, del vettore $\boldsymbol{v}$. L'insieme dei vettori n-dimensionali si indica con $\mathbb{R}^n$ (questa notazione è coerente col fatto che i vettori n-dimensionali sono elementi del prodotto cartesiano $\mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R}$ di $\mathbb{R}$ per se stesso n volte).
I vettori della geometria e della fisica (segmenti orientati) si identificano con i vettori algebrici 2- o 3-dimensionali: basta identificare un vettore geometrico con le sue componenti (componenti geometriche, ovvero proiezioni ortogonali sugli assi cartesiani) in un sistema di riferimento fissato. Ricordando che il modulo (lunghezza) di un vettore geometrico di componenti (v_1, v_2, v_3) è dato da $\sqrt{v_1^2 + v_2^2 + v_3^2}$ (come segue dal Teorema di Pitagora), possiamo per analogia definire il *modulo (euclideo)* di un vettore n-dimensionale qualunque $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ come

$$\|\boldsymbol{v}\| := \sqrt{v_1^2 + v_2^2 + \cdots + v_n^2}. \tag{5.14}$$

Con i vettori possiamo fare alcune operazioni.

Somma e differenza. Se $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ e $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ sono due vettori n-dimensionali, si pone

$$\boldsymbol{v} \pm \boldsymbol{w} = (v_1 \pm w_1, v_2 \pm w_2, \ldots, v_n \pm w_n) \tag{5.15}$$

(cioè la somma e la differenza si fanno "componente per componente"). Per i vettori geometrici queste operazioni corrispondono alla somma e alla differenza fatte con la regola del parallelogramma. La **distanza** (euclidea) fra due vettori $\boldsymbol{v}$ e $\boldsymbol{w}$ è definita come la norma della differenza:

$$d(\boldsymbol{v}, \boldsymbol{w}) = \|\boldsymbol{v} - \boldsymbol{w}\|. \tag{5.16}$$

Prodotto per uno scalare. Sia $\boldsymbol{v} \in \mathbb{R}^n$ un vettore e $\alpha \in \mathbb{R}$ un numero reale;[3] poniamo

$$\alpha\boldsymbol{v} = (\alpha v_1, \alpha v_2, \ldots, \alpha v_n). \tag{5.17}$$

Si può osservare che $\boldsymbol{v} - \boldsymbol{w}$ è la stessa cosa di $\boldsymbol{v} + (-1)\boldsymbol{w}$. Se $\boldsymbol{v}$ è un vettore geometrico, $\alpha\boldsymbol{v}$ è un vettore la cui direzione è quella di $\boldsymbol{v}$, il cui modulo è $|\alpha|\,\|\boldsymbol{v}\|$ e il cui verso resta quello di $\boldsymbol{v}$ se $\alpha > 0$ oppure è opposto se $\alpha < 0$.

Prodotto scalare. Se $\boldsymbol{v} = (v_1, v_2, \ldots, v_n)$ e $\boldsymbol{w} = (w_1, w_2, \ldots, w_n)$ sono due vettori n-dimensionali, il prodotto scalare di $\boldsymbol{v}$ con $\boldsymbol{w}$ si indica con $\boldsymbol{v} \cdot \boldsymbol{w}$ (e si legge "$\boldsymbol{v}$ scalare $\boldsymbol{w}$") e si definisce come il numero reale

$$\boldsymbol{v} \cdot \boldsymbol{w} = v_1 w_1 + v_2 w_2 + \cdots + v_n w_n. \tag{5.18}$$

Notiamo la differenza: il *prodotto per uno scalare* associa a un vettore $\boldsymbol{v}$ e a uno scalare α un vettore $\alpha\boldsymbol{v}$ mentre il *prodotto scalare* associa a due vettori, $\boldsymbol{v}$ e $\boldsymbol{w}$, lo scalare $\boldsymbol{v} \cdot \boldsymbol{w}$. Notiamo anche che il prodotto scalare di un vettore con se stesso è il quadrato del modulo

$$\boldsymbol{v} \cdot \boldsymbol{v} = v_1^2 + v_2^2 + \cdots + v_n^2 = \|\boldsymbol{v}\|^2.$$

Sui vettori geometrici risulta che il prodotto scalare corrisponde a

$$\boldsymbol{v} \cdot \boldsymbol{w} = \|\boldsymbol{v}\|\|\boldsymbol{w}\| \cos\theta,$$

dove θ è l'angolo formato da $\boldsymbol{v}$ e $\boldsymbol{w}$. Ma allora possiamo *definire* l'angolo formato tra due vettori di dimensione n qualunque, o meglio il suo coseno, come

$$\gamma(\boldsymbol{v}, \boldsymbol{w}) = \frac{\boldsymbol{v} \cdot \boldsymbol{w}}{\|\boldsymbol{v}\|\,\|\boldsymbol{w}\|}. \tag{5.19}$$

Due vettori di dimensione n, $\boldsymbol{v}$ e $\boldsymbol{w}$, si diranno *ortogonali* se $\boldsymbol{v} \cdot \boldsymbol{w} = 0$. Sempre in analogia col caso dei vettori geometrici, una *base ortonormale* di $\mathbb{R}^n$ viene definita come un insieme di vettori $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_n\}$ tali che

$$\begin{cases} \|\boldsymbol{e}_i\| = 1, & \text{per ogni } i, \\ \boldsymbol{e}_i \cdot \boldsymbol{e}_j = 0, & \text{per ogni } i \neq j, \end{cases}$$

ovvero di lunghezza unitaria e mutuamente ortogonali.

Consideriamo ora una matrice di dati statistici

$$X = \begin{pmatrix} x_1^1 & x_2^1 & \cdots & x_m^1 \\ x_1^2 & x_2^2 & \cdots & x_m^2 \\ \vdots & \vdots & & \vdots \\ x_1^n & x_2^n & \cdots & x_m^n \end{pmatrix}.$$

[3] In questo contesto i numeri reali si chiamano anche *scalari*, da qui il nome "prodotto per uno scalare".

Come abbiamo già detto nel paragrafo 5.1, le variabili sono gli m vettori n-dimensionali $\boldsymbol{x}_j$ costituiti dalle colonne della matrice mentre gli individui sono gli n vettori m-dimensionali $\boldsymbol{x}^i$ costituiti dalle colonne. Possiamo dare una semplice interpretazione geometrica delle quantità statistiche. Per ogni $j = 1, 2, \ldots, m$ introduciamo per comodità i vettori costanti n-dimensionali

$$\boldsymbol{\mu}_j = (\mu_j, \mu_j, \ldots, \mu_j), \quad \text{con} \quad \mu_j = \overline{\boldsymbol{x}}_j.$$

Possiamo allora scrivere:

$$\mathrm{Var}\,[\boldsymbol{x}_j] = \frac{1}{n-1}\sum_{k=1}^{n}(x_j^k - \mu_j)^2 = \frac{1}{n-1}\,\|\boldsymbol{x}_j - \boldsymbol{\mu}_j\|^2,$$

ovvero *la varianza di $\boldsymbol{x}_j$ è il quadrato del modulo di $\boldsymbol{x}_j - \boldsymbol{\mu}_j$ diviso per la dimensione n meno 1.* La deviazione standard sarà quindi

$$\mathrm{Std}\,[\boldsymbol{x}_j] = \frac{1}{\sqrt{n-1}}\,\|\boldsymbol{x}_j - \boldsymbol{\mu}_j\|.$$

Inoltre,

$$\mathrm{Cov}\,[\boldsymbol{x}_i, \boldsymbol{x}_j] = \frac{1}{n-1}\sum_{k=1}^{n}(x_i^k - \mu_i)\,(x_j^k - \mu_j) = \frac{1}{n-1}\,(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\cdot(\boldsymbol{x}_j - \boldsymbol{\mu}_j),$$

ovvero *la covarianza fra $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$ è data dal prodotto scalare di $\boldsymbol{x}_i - \boldsymbol{\mu}_i$ con $\boldsymbol{x}_j - \boldsymbol{\mu}_j$, diviso $n-1$.* Il coefficiente di correlazione fra $\boldsymbol{x}_i$ e $\boldsymbol{x}_j$ è dunque dato da

$$\rho[\boldsymbol{x}_i, \boldsymbol{x}_j] = \frac{\mathrm{Cov}\,[\boldsymbol{x}_i, \boldsymbol{x}_j]}{\mathrm{Std}\,[\boldsymbol{x}_i]\;\mathrm{Std}\,[\boldsymbol{x}_j]} = \frac{(\boldsymbol{x}_i - \boldsymbol{\mu}_i)\cdot(\boldsymbol{x}_j - \boldsymbol{\mu}_j)}{\|\boldsymbol{x}_i - \boldsymbol{\mu}_i\|\,\|\boldsymbol{x}_j - \boldsymbol{\mu}_j\|} = \gamma(\boldsymbol{x}_i - \boldsymbol{\mu}_i, \boldsymbol{x}_j - \boldsymbol{\mu}_j),$$

ovvero dal coseno dell'angolo formato da $\boldsymbol{x}_i - \boldsymbol{\mu}_i$ e $\boldsymbol{x}_j - \boldsymbol{\mu}_j$. Il coefficiente di correlazione si può dunque interpretare come il coseno di un angolo: ecco perché è sempre compreso fra -1 e 1. Osserviamo inoltre che *covarianza nulla, e quindi incorrelazione, equivalgono a ortogonalità.*

5.4 Richiami di algebra lineare: matrici

Una *matrice* $n \times m$ è una tabella di numeri reali con n righe e m colonne:

$$A = \begin{pmatrix} a_1^1 & a_2^1 & \cdots & a_m^1 \\ a_1^2 & a_2^2 & \cdots & a_m^2 \\ \vdots & \vdots & & \vdots \\ a_1^n & a_2^n & \cdots & a_m^n \end{pmatrix}.$$

I numeri a_j^i sono detti *elementi*, o *componenti*, della matrice A. L'indice in alto si chiama *indice di riga*, quello in basso *indice di colonna*.[4]
L'insieme delle matrici $n \times m$ viene indicato con $\mathbb{R}^{n,m}$, dunque scriveremo

[4] In numerosi testi si trovano entrambi gli indici in basso: a_{ij}. In questo caso il primo indice è quello di riga mentre il secondo è quello di colonna.

$A \in \mathbb{R}^{n,m}$. Le matrici $n \times n$, ovvero con un ugual numero di righe e di colonne, le diremo *quadrate* (notiamo inoltre che $\mathbb{R}^{1,1} = \mathbb{R}$).
Per risparmiare spazio è utile denotare una matrice A con $A = (a^i_j)$. Il difetto di questa notazione è che non si sa quante righe e quante colonne ci sono ma questa informazione si può fornire facilmente dicendo, ad esempio, che $A \in \mathbb{R}^{n,m}$.
I vettori si possono riguardare come casi particolari di matrici: o del tipo $\mathbb{R}^{1,n}$ (vettori-riga) o del tipo $\mathbb{R}^{n,1}$ (vettori-colonna). Chiaramente entrambi si identificano con $\mathbb{R}^n$ ma per comodità algebrica può convenire distinguere i due casi.
Se $A = (a^i_j) \in \mathbb{R}^{n,m}$, indicheremo con

$$\boldsymbol{a}^i = (a^i_1 \ a^i_1 \ \cdots \ a^i_m)$$

il suo i-esimo vettore-riga (sono n vettori m-dimensionali) e con

$$\boldsymbol{a}_j = \begin{pmatrix} a^1_j \\ a^2_j \\ \vdots \\ a^n_j \end{pmatrix}$$

il suo j-esimo vettore-colonna (sono m vettori n-dimensionali). Per quanto possibile, cercheremo di mantenere sempre le seguenti notazioni: lettere maiuscole per le matrici, lettere minuscole in grassetto per i vettori, lettere minuscole per gli scalari. Vediamo adesso la generalizzazione alle matrici delle operazioni già introdotte per i vettori.

Somma. È un'operazione che si fa fra due matrici dello stesso tipo, ottenedo ancora una matrice dello stesso tipo: se $A = (a^i_j) \in \mathbb{R}^{n,m}$ e $B = (b^i_j) \in \mathbb{R}^{n,m}$, si definisce

$$A \pm B = (a^i_j \pm b^i_j) \in \mathbb{R}^{n,m}$$

(notiamo quindi che la somma o la differenza si fanno "elemento per elemento").

Prodotto per uno scalare. Associa a un numero reale $\alpha \in \mathbb{R}$ e a una matrice $A = (a^i_j) \in \mathbb{R}^{n,m}$ la matrice

$$\alpha\, A = (\alpha\, a^i_j) \in \mathbb{R}^{n,m}$$

Notiamo, come nel caso dei vettori, che $A - B$ è la stessa cosa di $A + (-1)B$.

Prodotto. Detto anche *prodotto righe per colonne*. È un'operazione che a due matrici $A = (a^i_j) \in \mathbb{R}^{n,s}$ e $B = (b^i_j) \in \mathbb{R}^{s,m}$ associa una matrice $C = AB = (c^i_j) \in \mathbb{R}^{n,m}$ dove c^i_j è dato da

$$c^i_j = \boldsymbol{a}^i \cdot \boldsymbol{b}_j = \sum_{k=1}^{s} a^i_k b^k_j.$$

Dunque *per calcolare l'elemento c^i_j della matrice prodotto si deve fare il prodotto scalare dell'i-esima riga della matrice A con la j-esima colonna della matrice B* (perciò i vettori-riga di A devono avere la stessa dimensione dei vettori-colonna di B, ecco perché si richiede che il numero di colonne di A sia uguale al numero

di righe di B). Il risultato è una matrice $C = AB$ che ha lo stesso numero di righe di A e lo stesso numero di colonne di B.
Notiamo che *l'ordine* in cui si esegue il prodotto è fondamentale: BA può non aver senso anche se AB ce l'ha e inoltre, anche quando i due prodotti AB e BA hanno entrambi senso (il che accade se e solo se le due matrici sono quadrate), in generale si ha $AB \neq BA$. Pertanto il prodotto di matrici è *non commutativo*.
Tuttavia esso è associativo ovvero, se $A \in \mathbb{R}^{n,s}$, $B \in \mathbb{R}^{s,p}$ e $C \in \mathbb{R}^{p,m}$, si ha

$$A(BC) = (AB)C$$

(verificare per esercizio), per cui si può scrivere ABC senza ambiguità.
La *matrice identità* $n \times n$ è una matrice che ha tutti 1 sulla diagonale e tutti 0 altrove:

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Essa fa da *elemento neutro* per il prodotto di matrici quadrate $n \times n$. Se infatti $A \in \mathbb{R}^{n,n}$, si ha

$$AI = IA = A$$

(verificare per esercizio). Data una matrice quadrata $A \in \mathbb{R}^{n,n}$, se esiste un'altra matrice $B \in \mathbb{R}^{n,n}$ tale che

$$AB = BA = I,$$

diremo che B è la *matrice inversa di* A e scriveremo

$$B = A^{-1}.$$

Chiaramente, facendo il prodotto (in senso matriciale) di un vettore-riga $\boldsymbol{v} \in \mathbb{R}^{1,n}$ con un vettore-colonna $\boldsymbol{w} \in \mathbb{R}^{n,1}$ ritroviamo il prodotto scalare $\boldsymbol{v} \cdot \boldsymbol{w}$.
Un caso notevole di prodotto fra matrici è la *trasformazione lineare di un vettore*. Sia $\boldsymbol{v}$ un vettore n-dimensionale e sia A una matrice quadrata $n \times n$. Riguardando $\boldsymbol{v}$ come vettore-colonna, possiamo fare il prodotto $A\boldsymbol{v}$, ottenendo un altro vettore (colonna)

$$A\boldsymbol{v} = \begin{pmatrix} \boldsymbol{a}^1 \cdot \boldsymbol{v} \\ \boldsymbol{a}^2 \cdot \boldsymbol{v} \\ \vdots \\ \boldsymbol{a}^n \cdot \boldsymbol{v} \end{pmatrix}.$$

Trasposizione. È un'operazione che fa passare da una matrice $A \in \mathbb{R}^{n,m}$ a una matrice $A^* \in \mathbb{R}^{m,n}$, ottenuta scambiando righe con colonne e precisamente:

$$\begin{pmatrix} a_1^1 & a_2^1 & \cdots & a_m^1 \\ a_1^2 & a_2^2 & \cdots & a_m^2 \\ \vdots & \vdots & & \vdots \\ a_1^n & a_2^n & \cdots & a_m^n \end{pmatrix}^* = \begin{pmatrix} a_1^1 & a_1^2 & \cdots & a_1^n \\ a_2^1 & a_2^2 & \cdots & a_2^n \\ \vdots & \vdots & \ddots & \vdots \\ a_m^1 & a_m^2 & \cdots & a_n^m. \end{pmatrix}$$

Brevemente, se $A = (a_j^i)$, si ha $A^* = (a_i^j)$. Notiamo che la trasposizione trasforma vettori-riga in vettori-colonna e viceversa. Una matrice quadrata tale

che $A^* = A$ si dice *simmetrica*. Si dimostri per esercizio che, se $A \in \mathbb{R}^{n,s}$ e $B \in \mathbb{R}^{s,m}$, vale la proprietà

$$(AB)^* = B^* A^*.$$

Matrici ortogonali. Una matrice quadrata $A \in \mathbb{R}^{n,n}$ si dice *ortogonale* se

$$A^* A = I,$$

ovvero se la trasposta di A è anche la matrice inversa di A. Notiamo che questo corrisponde alle seguenti proprietà dei suoi vettori-riga

$$\begin{aligned} &\|\boldsymbol{a}_i\|^2 = 1, \quad \text{per ogni } i, \\ &\boldsymbol{a}_i \cdot \boldsymbol{a}_j = 0, \quad \text{per ogni } i \neq j \end{aligned}$$

(e, poiché si ha anche $AA^* = I$, valgono analoghe proprietà anche per i vettori-colonna). La proprietà fondamentale delle matrici ortogonali è che la trasformazione lineare di vettori eseguita con esse *lascia invariati i prodotti scalari fra vettori* (e quindi i moduli e gli angoli). Infatti, se A è una matrice ortogonale $n \times n$ e $\boldsymbol{v}, \boldsymbol{w}$ sono vettori n-dimensionali si ha (riguardando $\boldsymbol{v}$ e $\boldsymbol{w}$ come vettori-colonna)

$$(A\boldsymbol{v}) \cdot (A\boldsymbol{w}) = (A\boldsymbol{v})^* A\boldsymbol{w} = \boldsymbol{v}^* A^* A\boldsymbol{w} = \boldsymbol{v}^* I\boldsymbol{w} = \boldsymbol{v}^* \boldsymbol{w} = \boldsymbol{v} \cdot \boldsymbol{w}.$$

Dunque, in particolare, A trasformerà una base ortonormale in un'altra base ortonormale (tant'è vero che le matrici ortogonali sono associate a *rotazioni* dei sistemi di riferimento).

Fondamentale per la statistica multivariata è il seguente teorema, che è un notevole risultato dell'algebra lineare noto come *Teorema Spettrale*.

Teorema 5.5. (Teorema Spettrale)
Se $C \in \mathbb{R}^{n,n}$ è simmetrica, esiste $A \in \mathbb{R}^{n,n}$ ortogonale tale che la matrice $D = A^* C A \in \mathbb{R}^{n,n}$ è diagonale, ovvero

$$D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

Gli n numeri $\lambda_i \in \mathbb{R}$ si chiamano autovalori *della matrice C.*

5.5 Analisi delle componenti principali

L'analisi delle componenti principali, o PCA (acronimo di *Principal Component Analysis*) è uno dei più noti metodi della statistica multivariata. Il suo scopo è il seguente: quando si hanno molte variabili correlate, è difficile cogliere le differenze salienti fra gli individui; la PCA permette di costruire un certo numero (piccolo) di nuove variabili che mettano bene in luce differenze e similitudini fra

gli individui. Si tratta, in un certo senso di uno strumento matematico che aiuta a "sbrogliare la matassa dei dati".
Consideriamo dunque una matrice di dati con n individui (righe) e m variabili (colonne)

$$X = (x_j^i) \in \mathbb{R}^{n,m}$$

e la corrispondente matrice di covarianza

$$C_X = (c_j^i) \in \mathbb{R}^{m,m},$$

dove, posto $\mu_j = \overline{\boldsymbol{x}}_j$,

$$c_j^i := \mathrm{Cov}\,[\boldsymbol{x}_i, \boldsymbol{x}_j] = \frac{1}{n-1}\sum_{k=1}^{n}(x_i^k - \mu_i)(x_j^k - \mu_j).$$

Il primo passo è quello di *standardizzare le variabili*, ovvero passare alle nuove variabili

$$y_j^i = \frac{x_j^i - \mu_j}{\sigma_j}$$

(dove $\sigma_j = \mathrm{Std}\,[\boldsymbol{x}_j]$), che hanno media nulla e varianza unitaria. Questa operazione la si fa perché le variabili standardizzate sono meglio confrontabili. Le variabili di partenza, ad esempio, potrebbero essere fisicamente non-omogenee (ad esempio lunghezze e pesi), oppure potrebbero essere omogenee ma misurate con differenti unità. Notiamo che la matrice di covarianza C_Y di Y coincide con la matrice di correlazione di X:

$$\mathrm{Cov}\,[\boldsymbol{y}_i, \boldsymbol{y}_j] = \frac{1}{n-1}\sum_{k=1}^{n} y_i^k y_j^k = \frac{1}{n-1}\sum_{k=1}^{n}\frac{(x_i^k - \mu_i)}{\sigma_i}\frac{(x_j^k - \mu_j)}{\sigma_j} = \rho[\boldsymbol{x}_i, \boldsymbol{x}_j].$$

Il secondo passo consiste nel *diagonalizzare la matrice di covarianza* C_Y, nel senso del Teorema Spettrale. Formuliamo questo secondo passo sotto forma di teorema. A partire dalle variabili originali standardizzate $\boldsymbol{y}_j$ costruiamo m nuove variabili della forma $\boldsymbol{z}_j = \sum_{k=1}^{m} a_j^k\, \boldsymbol{y}_k$, ovvero

$$\begin{aligned} \boldsymbol{z}_1 &= a_1^1\boldsymbol{y}_1 + a_1^2\boldsymbol{y}_2 + \cdots + a_1^m\boldsymbol{y}_m \\ \boldsymbol{z}_2 &= a_2^1\boldsymbol{y}_1 + a_2^2\boldsymbol{y}_2 + \cdots + a_2^m\boldsymbol{y}_m \\ &\cdots \\ \boldsymbol{z}_m &= a_m^1\boldsymbol{y}_1 + a_m^2\boldsymbol{y}_2 + \cdots + a_m^m\boldsymbol{y}_m \end{aligned} \tag{5.20}$$

Teorema 5.6.
Possiamo scegliere i coefficienti a_j^i in modo tale che le nuove variabili $\boldsymbol{z}_j$ siano tutte incorrelate, ovvero che la loro matrice di covarianza sia diagonale, e che inoltre valga

$$\sum_{i=1}^{m}(a_j^i)^2 = 1, \qquad \textit{per ogni } j = 1, 2, \ldots, m. \tag{5.21}$$

Dimostrazione Consideriamo la matrice Z le cui colonne sono le $\boldsymbol{z}_j$ sopra definite (i coefficienti a_j^i per ora sono generici numeri reali da scegliere in seguito

in maniera opportuna). Notiamo che

$$z_j^i = \sum_{k=1}^{m} a_j^k \, y_k^i,$$

e dunque possiamo scrivere sinteticamente

$$Z = YA.$$

Notiamo inoltre che, grazie al fatto che le $\boldsymbol{y}_j$ hanno media nulla, la matrice di covarianza di Y si può scrivere

$$C_Y = \frac{1}{n-1} Y^* Y.$$

Notiamo poi che

$$\overline{\boldsymbol{z}}_j = \frac{1}{n} \sum_{i=1}^{n} z_j^i = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{m} a_j^k \, y_k^i = \sum_{k=1}^{m} a_j^k \, \frac{1}{n} \sum_{i=1}^{n} y_k^i = \sum_{k=1}^{m} a_j^k \, \overline{\boldsymbol{y}}_k = 0,$$

ovvero anche le varabili $\boldsymbol{z}_j$ hanno media nulla e perciò anche per esse vale

$$C_Z = \frac{1}{n-1} Z^* Z.$$

Ma allora, ricordando che $Z = YA$, si ha

$$C_Z = \frac{1}{n-1} Z^* Z = \frac{1}{n-1} (YA)^* YA = \frac{1}{n-1} A^* Y^* YA = A^* C_Y A.$$

Essendo C_Y simmetrica, per il Teorema Spettrale possiamo scegliere come matrice dei coefficienti A una matrice ortogonale che diagonalizza C_Y, ovvero tale che

$$A^* C_Y A = D = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{pmatrix}.$$

Pertanto, con questa scelta di A, le variabili $\boldsymbol{z}_j$ sono tutte incorrelate essendo $C_Z = D$. Poiché A è ortogonale, vale la (5.21). □

Notiamo che gli autovalori di C_Y sono le varianze delle nuove variabili:

$$\lambda_j = \mathrm{Var}\left[\boldsymbol{z}_j\right].$$

Si può inolre dimostrare che si ha sempre

$$\sum_{j=1}^{m} \lambda_j = m. \tag{5.22}$$

Abbiamo dunque costruito nuove variabili $\boldsymbol{z}_j$, combinazioni lineari delle $\boldsymbol{y}_j$, che sono incorrelate. L'interpretazione geometrica è che abbiamo costruito dei vettori $\boldsymbol{z}_j$ ortogonali. Questo era il passo matematicamente più profondo della procedura. L'ultimo passo consiste nello scegliere poche (due o tre) variabili $\boldsymbol{z}_j$ che siano significative ai fini delle variazioni, ovvero con la più alte varianze λ_j.

Ricapitolando, la PCA consiste in

1. passare dalle variabili originali $\boldsymbol{x}_j$ alle variabili standardizzate $\boldsymbol{y}_j$;
2. passare dalle variabili $\boldsymbol{y}_j$ a una loro miscela lineare $\boldsymbol{z}_j$ tale che le nuove variabili siano incorrelate (che lo si possa sempre fare ce lo assicura il Teorema Spettrale);
3. scegliere poche variabili $\boldsymbol{z}_j$ con la varianza più significativa.

Osservazione 5.7. Il fatto che A sia ortogonale, ovvero la condizione (5.21), ci assicura che la trasformazione (5.20) non ha "gonfiato" artificialmente alcune varianze facendole diventare preponderanti. Ad esempio, la trasformazione lineare $\boldsymbol{z}_1 = \alpha \boldsymbol{y}_1$ e $\boldsymbol{z}_j = \boldsymbol{y}_j$ per $j > 1$, aumenta la varianza di $\boldsymbol{z}_1$ in ragione di α^2; tuttavia questa non è una trasformazione ortogonale. Come già osservato, una trasformazione ortogonale corrisponde a nient'altro che a una rotazione del sistema di riferimento. □

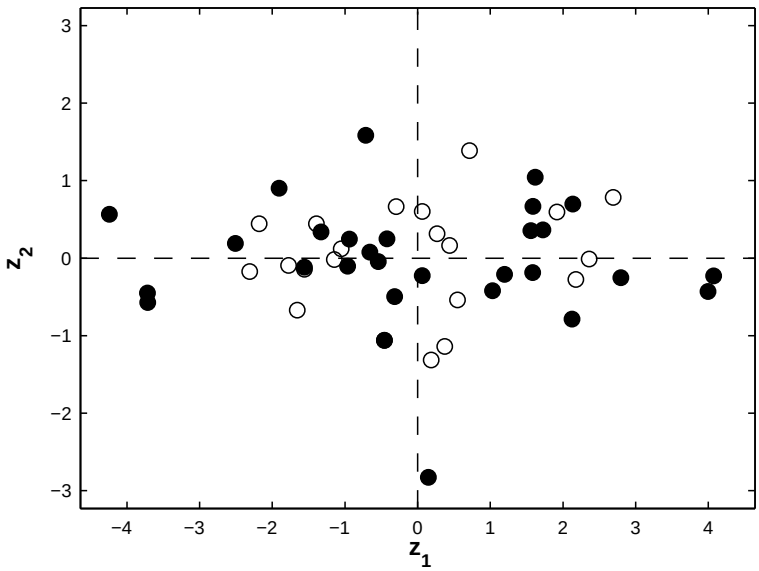


Figura 5.1: Analisi delle componenti principali relativa alla tabella 5.1. Gli esmplari indicati col pallino vuoto sono soparvvissuti, gli altri sono deceduti.

Come primo esempio di analisi delle componenti principali consideriamo la Tabella 5.1. Come già detto, si tratta delle misure di variabili morfologiche ($\boldsymbol{x}_1$ = lunghezza, $\boldsymbol{x}_2$ = apertura alare, $\boldsymbol{x}_3$ = lunghezza dell'omero, $\boldsymbol{x}_4$ = lunghezza di becco e testa, $\boldsymbol{x}_5$ = lunghezza dello sterno) di passeri raccolti o catturati dopo il passaggio di un tifone. Gli esemplari da 1 a 21 sono sopravvissuti, gli altri sono deceduti. Si vuole evidenziare, se possibile, una caratteristica saliente degli esemplari che sono sopravvissuti. Applicando la procedura descritta nel paragrafo precedente, si trovano i seguenti autovalori, qui ordinati in senso decrescente:

$$\lambda_1 = 3.6, \quad \lambda_2 = 0.53, \quad \lambda_3 = 0.39, \quad \lambda_4 = 0.30, \quad \lambda_5 = 0.16.$$

Osserviamo che la somma di questi valori è 5, in accordo con la (5.22). La prima variabile principale, quella con varianza λ_1, risulta essere la seguente:

$$\boldsymbol{z}_1 = 0.45\,\boldsymbol{y}_1 + 0.46\,\boldsymbol{y}_2 + 0.45\,\boldsymbol{y}_3 + 0.47\,\boldsymbol{y}_4 + 0.40\,\boldsymbol{y}_5.$$

Essendo responsabile da sola di circa il 72% della varianza totale, questa è quindi di gran lunga la componente principale più significativa. Si può dire che essa rappresenti la "taglia" complessiva dell'esemplare (notiamo che tutti e cinque i parametri morfologici sono pesati circa allo stesso modo nel formare $\boldsymbol{z}_1$). La seconda componente principale

$$\boldsymbol{z}_2 = -0.05\,\boldsymbol{y}_1 + 0.30\,\boldsymbol{y}_2 + 0.32\,\boldsymbol{y}_3 + 0.18\,\boldsymbol{y}_4 - 0.88\,\boldsymbol{y}_5$$

ha varianza $\lambda_2 = 0.53$, pari a circa l'11% della varianza totale, e quindi notevolmente inferiore a quella di $\boldsymbol{z}_1$. Il suo significato sembra essere quello del contrasto tra $\boldsymbol{y}_2$, $\boldsymbol{y}_3$ e $\boldsymbol{y}_4$, da una parte, e $\boldsymbol{y}_5$ dall'altra. Se riportiamo gli individui sul piano $(\boldsymbol{z}_1\boldsymbol{z}_2)$ otteniamo la distribuzione di Figura 5.1, dove i pallini vuoti corripondono agli esemplari sopravvissuti e quelli pieni agli esemplari deceduti. Si può arguire dalla distribuzione dei pallini che l'evento metereologico ha selezionato gli esemplari con valori di taglia media (né troppo grandi né troppo piccoli). Qualcosa del genere si potrebbe dire anche riguardo alla variabile $\boldsymbol{z}_2$.

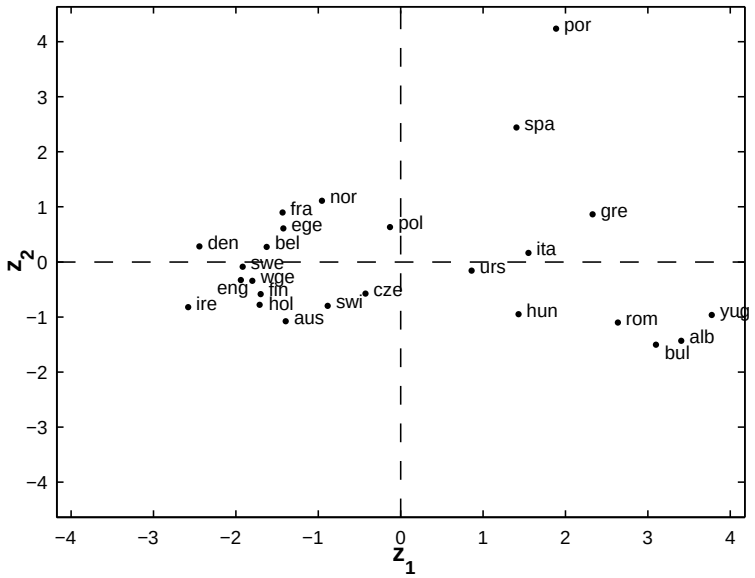


Figura 5.2: Analisi delle componenti principali relativa alla tabella 5.2.

Come secondo esempio consideriamo la Tabella 5.2, sempre tratta dal libro di Manly. Si tratta dei consumi alimentari in 25 paesi europei, secondo un'indagine statistica datata 1977. Se applichiamo l'analisi delle componenti principali otteniamo i seguenti autovalori:

$$\begin{array}{llllll} \lambda_1 = 4.1, & \lambda_2 = 1.6, & \lambda_3 = 1.1, & \lambda_4 = 0.91, & \lambda_5 = 0.43, \\ \lambda_6 = 0.35, & \lambda_7 = 0.27, & \lambda_8 = 0.13, & \lambda_9 = 0.11. & \end{array}$$

Paese	*sigla*	*carne rossa*	*carne bianca*	*uova*	*latte*	*pesce*	*cereali*	*amidi*	*legumi e oli veg.*	*frutta e verd.*
Albania	alb	10	1	1	9	0	42	1	6	2
Austria	aus	9	14	4	20	2	28	4	1	4
Belgio	bel	14	9	4	18	5	27	6	2	4
Bulgaria	bul	8	6	2	8	1	57	1	4	4
Cecoslovacchia	cze	10	11	3	13	2	34	5	1	4
Danimarca	den	11	11	4	25	10	22	5	1	2
Finlandia	fin	10	5	3	34	6	26	5	1	1
Francia	fra	18	10	3	20	6	28	5	2	7
Germania Est	ege	8	12	4	11	5	25	7	1	4
Germania Ovest	wge	11	13	4	19	3	19	5	2	4
Grecia	gre	10	3	3	18	6	42	2	8	7
Irlanda	ire	14	10	5	26	2	24	6	2	3
Italia	ita	9	5	3	14	3	37	2	4	7
Norvegia	nor	9	5	3	23	10	23	5	2	3
Paesi Bassi	hol	10	14	4	23	3	22	4	2	4
Polonia	pol	7	10	3	19	3	36	6	2	7
Portogallo	por	6	4	1	5	14	27	6	5	8
Regno Unito	eng	17	6	5	21	4	24	5	3	3
Romania	rom	6	6	2	11	1	50	3	5	3
Spagna	spa	7	3	3	9	7	29	6	6	7
Svezia	swe	10	8	4	25	8	20	4	1	2
Svizzera	swi	13	10	3	24	2	26	3	2	5
Ungheria	hun	5	12	3	10	0	40	4	5	4
URSS	urs	9	5	2	17	3	44	6	3	3
Yugoslavia	yug	4	5	1	10	1	56	3	6	3

Tabella 5.2: Dati relativi ai consumi alimentari delle nazioni europee nel 1977 (da B. Manly [8]).

Come variabili dalle varianze più significative possiamo dunque prendere $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$ (rispettivamente il 46% e il 18% della varianza totale) che sono date da

$$\begin{aligned}\boldsymbol{z}_1 = &- 0.31\,\boldsymbol{y}_1 - 0.32\,\boldsymbol{y}_2 - 0.42\,\boldsymbol{y}_3 - 0.38\,\boldsymbol{y}_4 - 0.13\,\boldsymbol{y}_5 + 0.43\,\boldsymbol{y}_6 - 0.30\,\boldsymbol{y}_7 \\ &+ 0.42\,\boldsymbol{y}_8 + 0.12\,\boldsymbol{y}_9,\\ \boldsymbol{z}_2 = &- 0.07\,\boldsymbol{y}_1 - 0.21\,\boldsymbol{y}_2 - 0.10\,\boldsymbol{y}_3 - 0.17\,\boldsymbol{y}_4 + 0.65\,\boldsymbol{y}_5 - 0.25\,\boldsymbol{y}_6 + 0.39\,\boldsymbol{y}_7 \\ &+ 0.13\,\boldsymbol{y}_8 + 0.51\,\boldsymbol{y}_9.\end{aligned}$$

Notiamo che $\boldsymbol{z}_1$ pesa negativamente in modo significativo le variabili $\boldsymbol{y}_1$ (carne rossa), $\boldsymbol{y}_2$ (carne bianca), $\boldsymbol{y}_3$ (uova), $\boldsymbol{y}_4$ (latte), $\boldsymbol{y}_7$ (amidi), mentre pesa positivamente in modo significativo le variabili $\boldsymbol{y}_6$ (cereali) e $\boldsymbol{y}_8$ (legumi, frutta secca, oli vegetali). Le variabili $\boldsymbol{y}_5$ (pesce) e $\boldsymbol{y}_9$ (frutta e verdura) paiono invece poco significative. Dunque, valori negativi di $\boldsymbol{z}_1$ indicano diete molto proteiche mentre valori positivi indicano diete basate su cereali, legumi e oli vegetali. La variabile $\boldsymbol{z}_2$, invece, "valorizza" le variabili $\boldsymbol{y}_5$, $\boldsymbol{y}_7$ e $\boldsymbol{y}_9$. Se riportiamo gli individui sul piano $(\boldsymbol{z}_1, \boldsymbol{z}_2)$ (Figura 5.2), notiamo che le variabili $\boldsymbol{z}_1$ e $\boldsymbol{z}_2$ fotografano bene la situazione geo-economica (almeno del periodo in esame). Notiamo nella parte sinistra del grafico un raggruppamento di nazioni prevalentemente nord-occidentali o mittel-europee. Nella parte destra del grafico possiamo osservare che il quadrante in alto racchiude le nazioni mediterranee (con l'eccezione della Francia) mentre il quadrante in basso raccoglie le nazioni dell'est europeo (con l'eccezione della Polonia).

5.6 Analisi dei cluster

La cosiddetta analisi dei cluster (*cluster* = gruppo, raggruppamento, agglomerato) ha lo scopo di raggruppare gli individui di un campione statistico in base a un criterio di "vicinanza". Tale vicinanza si misura con un'opportuna definizione matematica di distanza tra individui. La definizione usuale di distanza fra due individui $\boldsymbol{x}^i$ e $\boldsymbol{x}^j$ è quella *euclidea*, già introdotta nel paragrafo 5.3:

$$d(\boldsymbol{x}^i, \boldsymbol{x}^j) = \|\boldsymbol{x}^i - \boldsymbol{x}^j\| = \sqrt{\sum_{k=1}^{m} (x_k^i - x_k^j)^2}$$

ma altri tipi di distanza[5] possono essere utili in determinate circostanze, ad esempio

$$d_1(\boldsymbol{x}^i, \boldsymbol{x}^j) = \sum_{k=1}^{m} |x_k^i - x_k^j|, \qquad d_\infty(\boldsymbol{x}^i, \boldsymbol{x}^j) = \max_{1 \le k \le m} |x_k^i - x_k^j|$$

[5] In generale, $d(\boldsymbol{x}, \boldsymbol{y})$ è una distanza fra vettori se valgono le seguenti proprietà:

(i) $d(\boldsymbol{x}, \boldsymbol{y}) = d(\boldsymbol{y}, \boldsymbol{x})$;

(ii) $d(\boldsymbol{x}, \boldsymbol{y}) \ge 0$, e $d(\boldsymbol{x}, \boldsymbol{y}) = 0$ se e solo se $\boldsymbol{x} = \boldsymbol{y}$;

(iii) $d(\boldsymbol{x}, \boldsymbol{y}) \le d(\boldsymbol{x}, \boldsymbol{z}) + d(\boldsymbol{z}, \boldsymbol{y})$ (disuguaglianza triangolare).

(dette, rispettivamente, *distanza "Manhattan"* e *distanza di Cebicev*) o più in generale, per $1 \leq p < \infty$, la *p-distanza*

$$d_p(\boldsymbol{x}^i, \boldsymbol{x}^j) = \sqrt[p]{\sum_{k=1}^{m} (x_k^i - x_k^j)^p}$$

o, ancora più in generale, le *distanze pesate*

$$d_{\boldsymbol{w}}(\boldsymbol{x}^i, \boldsymbol{x}^j) = \sqrt[p]{\sum_{k=1}^{m} w_k (x_k^i - x_k^j)^p}, \qquad w_k \geq 0,$$

e altre ancora. In tutti questi casi è conveniente utilizzare variabili standardizzate (per motivi analoghi a quelli ricordati nel precedente paragrafo). Supponiamo d'ora in poi che il nostro campione sia standardizzato. Avvertiamo anche che d'ora in poi useremo le parole *cluster* e *gruppo* come sinonimi.

I metodi del'analisi di cluster sono piuttosto variegati. Una prima distinzione viene dal metodo di raggruppamento:

metodi non-gerarchici - si fissa a priori il numero di gruppi che vogliamo ottenere;

metodi gerarchici - non si fissa a priori il numero di gruppi.

I metodi gerarchici a loro volta si sudividono in:

aggregativi - si parte da un numero di gruppi pari al numero di individui (ogni individuo è un gruppo) e si riduce progressivamente il numero di gruppi, aggregandoli in base a una misura di *vicinanza fra gruppi*;

disgiuntivi - il contrario del precedente (procede per suddivisione).

In queste brevi note concentreremo la nostra attenzione sui *metodi gerarchici aggregativi*, che possiamo vedere come l'algoritmo schematizzato dal seguente diagramma a blocchi:

L'algoritmo prende avvio da una situazione in cui ogni individuo è considerato un gruppo a sé stante e procede aggregando successivamente i gruppi più vicini, finché gli individui non si trovano tutti in un solo gruppo. Notiamo quindi che a ogni passo dell'algoritmo si ottiene un numero minore di gruppi e sta a chi osserva scegliere il livello di raggruppamento ritenuto significativo. Si capisce che è fondamentale chiarire bene che cosa si intende per "vicinanza fra gruppi" (i

gruppi sono insiemi di individui, mentre finora abbiamo parlato solo di distanza fra singoli individui). Dobbiamo perciò definire una *distanza fra insiemi*: a ogni scelta di tale distanza corrisponde un possibile *criterio di raggruppamento* per il passo dell'algoritmo. Riportiamo qui di seguito una lista di possibili distanze fra gruppi (e quindi di possibili criteri di raggruppamento), supponendo di aver già scelto una distanza d fra individui (ad esempio la distanza euclidea).

Distanza del "nearest neighbour": la distanza fra due fruppi G_1 e G_2 è la minima distanza fra coppie di individui appartenenti ai due diversi gruppi:

$$D(G_1, G_2) = \min \{ d(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \in G_1,\ \boldsymbol{y} \in G_2 \}.$$

Il criterio di raggruppamento corrispondente si chiama *criterio del legame singolo*.

Distanza del "furthest neighbour": la distanza fra due fruppi G_1 e G_2 è la massima distanza fra coppie di individui appartenenti ai due diversi gruppi:

$$D(G_1, G_2) = \max \{ d(\boldsymbol{x}, \boldsymbol{y}) \mid \boldsymbol{x} \in G_1,\ \boldsymbol{y} \in G_2 \}.$$

Il criterio di raggruppamento corrispondente si chiama *criterio del legame completo*.

Distanza media intragruppo: la distanza fra due fruppi G_1 e G_2 è la media aritmetica delle distanza fra coppie di individui appartenenti ai due diversi gruppi. Il criterio di raggruppamento corrispondente si chiama *criterio del legame medio ("within groups average clustering")*.

Distanza media intergruppo: la distanza fra due fruppi G_1 e G_2 è la media aritmetica di tutte le distanze fra coppie di appartenenti all'unione dei due gruppi. Il criterio di raggruppamento corrispondente si chiama *criterio della media ("unweighted pair-groups average")*.

Distanza del baricentro: la distanza fra due fruppi G_1 e G_2 è la distanza fra i baricentri[6] dei due gruppi. Il criterio di raggruppamento corrispondente si chiama *criterio del centroide (o del baricentro)*.

Notiamo che in tutti questi casi la distanza fra due gruppi costituiti da un solo elemento ciascuno coincide con la distanza fra i due elementi.

Citiamo infine un ulteriore criterio di raggruppamento che non discende da una definizione di distanza, il *criterio di Ward*: si aggregano i gruppi in modo che l'incremento della varianza totale (cioè della somma delle varianze dei singoli gruppi) sia il minimo possibile.

[6] Il baricentro di un insieme G di N vettori è il vettore $\boldsymbol{g}$ che per coordinate la media (eventualmente pesata) delle coordinate dei punti dell'insieme:

$$\boldsymbol{g} = \frac{1}{N} \sum_{\boldsymbol{x} \in G} \boldsymbol{x}, \qquad \text{per cui} \qquad g_k = \frac{1}{N} \sum_{\boldsymbol{x} \in G} x_k.$$

Esempio 5.8. Supponiamo di avere un campione con 5 individui e un certo numero di variabili (che non ha importanza specificare). Consideriamo la matrice $\Delta = (d_{ij})$ delle distanze fra individui

$$d_{ij} = d(\boldsymbol{x}^i, \boldsymbol{x}^j),$$

che sarà una matrice 5×5 simmtrica e con tutti 0 sulla diagonale (detta anche "matrice di dissimiglianza'), e assumiamo che sia la seguente:

$$\Delta = \begin{pmatrix} 0 & & & & \\ 9 & 0 & & & \\ 3 & 7 & 0 & & \\ 6 & 5 & 9 & 0 & \\ 11 & 10 & 2 & 8 & 0 \end{pmatrix}$$

(i posti vuoti si possono completare per simmteria). Procediamo con l'analisi dei cluster gerarchica aggregativa utilizzando inizialmente il criterio del legame singolo. All'inizio i cinque individui formano cinque gruppi distinti:

$$(1) \qquad (2) \qquad (3) \qquad (4) \qquad (5).$$

Al primo passo la distanza fra gruppi coincide con la distanza fra individui che possiamo leggere direttamente sulla matrice Δ. Osservando che la distanza minore è quella fra i gruppi (3) e (5) ($d_{35} = 2$), riuniamo questi in un unico gruppo, ottenendo una prima clusterizzazione:

$$(3,5) \qquad (1) \qquad (2) \qquad (4).$$

Ora dobbiamo calcolare la distanza fra questi quattro gruppi, utilizzando la distanza "nearest neighbour":

$$\begin{aligned} &D\left((3,5),(1)\right) = \min\{d_{31}, d_{51}\} = \min\{3, 11\} = 3 \\ &D\left((3,5),(2)\right) = \min\{d_{32}, d_{52}\} = \min\{7, 10\} = 7 \\ &D\left((3,5),(4)\right) = \min\{d_{34}, d_{54}\} = \min\{9, 8\} = 8 \\ &D\left((1),(2)\right) = d_{12} = 9, \quad D\left((1),(4)\right) = d_{14} = 6, \quad D\left((2),(4)\right) = d_{24} = 5. \end{aligned}$$

I gruppi più vicini sono (3,5) e (1) (distanza = 3) e si ha così la nuova clusterizzazione:

$$(1,3,5) \qquad (2) \qquad (4).$$

Calcoliamo la distanza fra i nuovi gruppi:

$$\begin{aligned} &D\left((1,3,5),(2)\right) = \min\{d_{12}, d_{32}, d_{52}\} = \min\{9, 7, 10\} = 7 \\ &D\left((1,3,5),(4)\right) = \min\{d_{14}, d_{34}, d_{54}\} = \min\{6, 9, 8\} = 6 \\ &D\left((2),(4)\right) = d_{24} = 5, \end{aligned}$$

per cui si devono unire i cluster (2) e (4) (distanza = 5) ottenendo

$$(1,3,5) \qquad (2,4).$$

L'ultimo passo è solo formale e consiste nell'unire questi due gruppi (la cui distanza è $\min\{d_{12}, d_{32}, d_{52}, d_{14}, d_{34}, d_{54}\} = 6$) in un unico cluster:

$$(1,2,3,4,5).$$

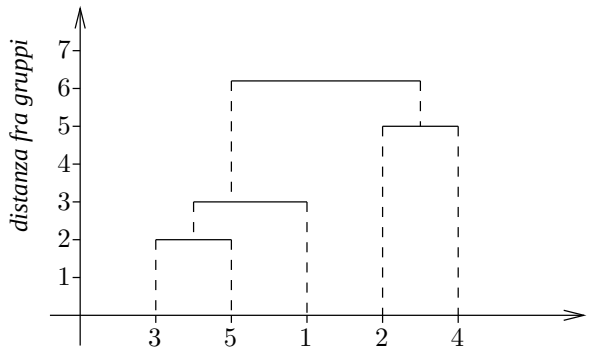


Figura 5.3: Dendrogramma riassuntivo dell'analisi dei claster esuguita sulla matrice di distanze Δ col criterio del legame semplice. Il raggruppamanto è rappresentato dalle linee orizzontali e il livello del raggruppamento, cioè, la distanza alla quale questo avviene, è data dall'altezza della linea orizzontale stessa.

Riassumiamo il rsultato di questa analisi nel cosiddetto *dendrogramma* (diagramma ad albero) di figura 5.3.
Riprendiamo lo stesso esempio e adottiamo adesso il criterio del legame completo. Poiché al primo passo le distanze fra gruppi non sono altro che le distanze fra individui, la prima clusterizzazione è identica alla precedente:

$$(3,5) \qquad (1) \qquad (2) \qquad (4) \qquad \text{(distanza 2)}.$$

Per andare avanti clacoliamo la distanza fra i quattro gruppi utilizzando la distanza "furthest neighbour":

$$D\left((3,5),(1)\right) = \max\{d_{31}, d_{51}\} = \max\{3,11\} = 11$$
$$D\left((3,5),(2)\right) = \max\{d_{32}, d_{52}\} = \max\{7,10\} = 10$$
$$D\left((3,5),(4)\right) = \max\{d_{34}, d_{54}\} = \max\{9,8\} = 9$$
$$D\left((1),(2)\right) = d_{12} = 9, \quad D\left((1),(4)\right) = d_{14} = 6, \quad D\left((2),(4)\right) = d_{24} = 5.$$

Si uniscono perciò i gruppi (2) e (4), che sono i più vicini (distanza 5):

$$(3,5) \qquad (1) \qquad (2,4).$$

Calcoliamo ora le distanze fra questi nuovi gruppi:

$$D\left((3,5),(1)\right) = \max\{d_{31}, d_{51}\} = \max\{3,11\} = 11$$
$$D\left((3,5),(2,4)\right) = \max\{d_{32}, d_{34}, d_{52}, d_{54}\} = \max\{7,9,10,8\} = 10$$
$$D\left((1),(2,4)\right) = \max\{d_{12}, d_{14}\} = \max\{9,6\} = 9$$

e uniamo quindi i gruppi (1) e (2,4), che sono i più vicini (distanza 9):

$$(3,5) \qquad (1,2,4).$$

Infine si uniscono i due gruppi rimanenti (che distano $\max\{d_{13}, d_{23}, d_{43}, d_{15}, d_{25}, d_{45}\} = 11$) in un unico gruppo:

$$(1,2,3,4,5).$$

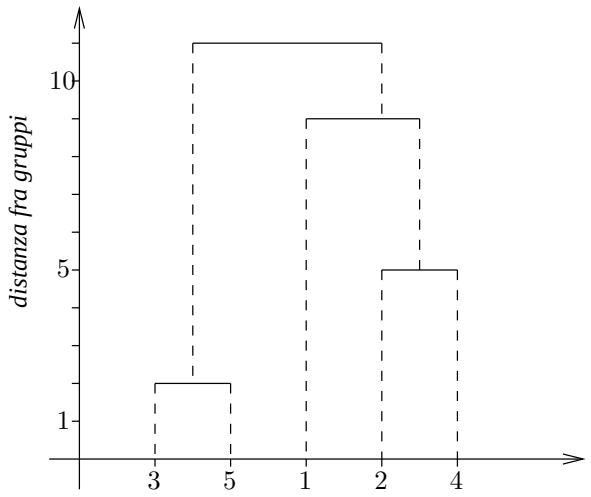


Figura 5.4: Dendrogramma riassuntivo dell'analisi dei cluster esuguita sulla matrice di distanze Δ col criterio del legame completo.

Il dendrogramma relativo a questa analisi è riportato in figura 5.4. □

Osservazione 5.9. Si potrebbe pensare di ottenere qualche vantaggio effettuando l'analisi dei cluster in combinazione con la PCA, cioè facendo l'analisi dei cluster sulla matrice $Z = XA$ anziché sulla matrice X (si veda il paragrafo precedente). Tuttavia questo risulterebbe inutile poiché, almeno usando la distanza euclidea, si ottiene

$$d(\boldsymbol{z}^i, \boldsymbol{z}^j) = d(A^*\boldsymbol{x}^i, A^*\boldsymbol{x}^j) = d(\boldsymbol{x}^i, \boldsymbol{x}^j)$$

(essendo A^* ortogonale), e dunque la matrice delle distanze calcolata su Z è identica a quella calcolata su X. □

Capitolo 6

Statistica inferenziale

6.1 Concetti generali

Come già accennato in precedenza, lo scopo della statistica inferenziale è quello di stabilire metodi rigorosi per risalire, con un "grado di certezza" calcolabile, a proprietà generali di una popolazione a partire da una raccolta di dati eseguita sulla popolazione stessa. Il rigore del metodo consiste nell'appoggiarsi alla teoria del calcolo delle probabilità, al quale la statistica inferenziale è strettamente legata. Il modello matematico della statistica inferenziale è così sintetizzabile:

- quando si effettua la misura di una variabile su n individui, il risultato della misura è un "campione estratto da una distribuzione", ovvero n variabili aleatorie $X_1, X_2, \ldots, X_n$ tutte con la stessa distribuzione $\mathcal{D}$ (e solitamente indipendenti[1]);
- la distribuzione $\mathcal{D}$ è incognita (o parzialmente incognita) e vogliamo ricavare informazioni su $\mathcal{D}$ a partire dalla misura effettuata;
- tali informazioni saranno di natura probabilistica, per cui ogni affermazione sulle proprietà di $\mathcal{D}$ sarà del tipo "la distribuzione $\mathcal{D}$ ha la data proprietà con un certo grado di certezza (cioà con una certa probabilità) ".

Il metodo statistico più comune è quello di supporre che la distribuzione $\mathcal{D}$ sia di un certo tipo (ad esempio una distribuzione normale) ma di non conoscere alcuni parametri che la caratterizzano (ad esempio la media e la varianza). A volte, invece, non si fanno ipotesi circa il tipo di distribuzione (sono i cosiddetti metodi statistici "non parametrici", solitamente più versatili ma meno precisi); nelle presenti note non tratteremo questo secondo caso.

Formalizziamo quanto detto finora con alcune definizioni precise. In queste note ci occuperemo esclusivamente del caso *monovariato*, quello cioè in cui si considera una sola variabile misurata su n individui.[2]

[1] L'indipendenza fra individui del campione è strettamente legata all'utilizzo di una corretta "tecnica di campionamento", si veda a proposito l'ultimo capoverso del paragrafo 5.1.

[2] Nel caso multivariato il campione statistico sarà una v.a. $n \times m$ (come nell'esempio 4.3), le cui n righe (individui) sono identicamente distribuite e indipendenti; si tratta della versione inferenziale del campione statistico descritto nel paragrafo 5.1.

Definizione 6.1. Un *campione statistico di dimensione (o "rango" o "numerosità") n* è una v.a. multipla

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n),$$

le cui componenti X_i sono indipendenti e identicamente distribuite. □

È ovvio che i dati concretamente raccolti in un'indagine statistica saranno semplicemente n numeri $x_1, x_2, \ldots, x_n$ ma la cosa importante da capire è che questi sono visti solo come una delle tante possibili *realizzazioni* della v.a. $\boldsymbol{X}$, realizzazione che, a priori, aveva solo una certa probabilità di verificarsi. La differenza tra statistica descrittiva e statistica inferenziale si potrebbe sintetizzare dicendo che la prima si occupa solo dei dati $x_1, x_2, \ldots, x_n$ mentre la seconda si occupa delle relazioni tra questi e la distribuzione $\mathcal{D}$.
Come già detto, solitamente si parte dall'ipotesi di conoscere la distribuzione di ogni individuo (ricordiamo che è identica per tutti) e che questa dipenda da uno o più parametri incogniti. Scriveremo perciò:

$$X_i \sim \mathcal{D}(\theta_1, \theta_2, \ldots, \theta_s), \qquad i = 1, 2, \ldots, n,$$

dove $\mathcal{D}$ è la distribuzione e $\theta_1, \theta_2, \ldots, \theta_s$ sono un certo numero di parametri incogniti. Ad esempio, potrebbe essere $X_i \sim \mathcal{N}(\mu, \sigma^2)$, con uno o entrambi i parametri μ o σ^2 incogniti. La *stima* di un parametro incognito θ è un certo numero calcolato a partire dai dati dal campione; dunque, a priori, è anch'esso una v.a. che viene detta *stimatore* del parametro θ. Un buono stimatore deve come minimo avere la proprietà che il suo valore atteso è il parametro che deve stimare.

Definizione 6.2. Uno *stimatore corretto* (o 'non distorto") del parametro θ è una v.a. Y tale che

$$E[Y] = \theta, \tag{6.1}$$

cioè che ha come valore atteso il parametro che deve stimare. □

Esempio 6.3.: Exit-poll I. Consideriamo un *exit-poll* effettuato su un campione di n elettori mediante interviste all'uscita dei seggi in una competizione elettorale fra due candidati A e B. Ogni votante può essere visto come una v.a. Bernoulliana $X_i \sim \mathcal{B}(p)$ (si veda la definizione 3.1), ovvero

$$X_i = \begin{cases} 1, & \text{se l'intervistato } i \text{ ha votato il candidato } A, \\ 0, & \text{se l'intervistato } i \text{ ha votato il candidato } B. \end{cases}$$

La v.a. X_i assume il valore 1 con probabilità p e il valore 0 con probabilità $1-p$. Assumendo l'indipendenza fra i votanti intevistati, abbiamo un campione statistico di rango n di variabili Bernoulliane

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_n).$$

Notiamo che il parametro incognito della distribuzione è la probabilità p, che è anche il valore atteso di ogni X_i (si veda la (3.2)), e che è proprio il valore che ci interessa stimare per prevedere l'esito della competizione. Lo stimatore

più naturale di p è la *media campionaria* (di cui parleremo più diffusamente nel prossimo paragrafo):

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Notiamo che questo è uno stimatore corretto di p poiché, come segue facilmente dalla proprietà di linearità della media (4.15),

$$E\left[\overline{X}\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{np}{n} = p.$$

Supponiamo di aver intervistato $n = 5000$ persone, delle quali 2347 dicharano di aver votato A e 2653 di aver votato B, la media campionaria assume il valore $2347/5000 \approx 0.469$. Si stima dunque che il candidato A otterrà il 46.9% dei voti. □

Il semplice valore numerico della stima di un parametro è di per sé un'indicazione piuttosto povera se non è accompagnata dall'espressione di un "grado di fiducia" nella stima stessa. Tale "grado di fiducia" è generalmente fornito come *probabilità* che il valore vero, incognito, del parametro si trovi in un certo intervallo di valori. Questo importante concetto è espresso dalla definizione seguente.

Definizione 6.4. Sia Y uno stimatore di un parametro θ e sia $0 \leq \alpha \leq 1$. Si chiama *intervallo di confidenza per θ di livello (di errore) α* un intervallo $I_Y \subset \mathbb{R}$ tale che

$$P(\theta \notin I_Y) \leq \alpha, \tag{6.2}$$

ovvero un intervallo di valori in cui il valore vero del parametro si trova con probabilità almeno $1 - \alpha$. □

D'ora in poi abbrevieremo "intervallo di confidenza" con IdC. Molto spesso nel linguaggio statistico si usa una terminologia del tipo "IdC al 95% per θ". Con questo si intende riferirsi non all'errore α ma al suo complementare $1 - \alpha$, espresso in percentuali. Dunque, ad esempio, un IdC al 95% è un IdC di livello $\alpha = 0.05$.
Alcuni tipi di IdC sono più usati rispetto ad altri.

Definizione 6.5. Sia $\delta > 0$ fissato. Un *IdC bilaterale per θ rispetto a Y* è un IdC del tipo:

$$I_Y = [Y - \delta, Y + \delta],$$

ovvero un intervallo che ha per centro lo stimatore Y e raggio δ. Un *IdC unilaterale destro per θ rispetto a Y* è del tipo

$$I_Y = [Y - \delta, +\infty)$$

mentre un *IdC unilaterale sinistro per θ rispetto a Y* è del tipo

$$I_Y = (-\infty, Y + \delta].$$

□

Osservazione 6.6. Riscrivendo le disuguaglianze

$$Y - \delta \leq \theta \leq Y + \delta$$

come

$$\theta - \delta \leq Y \leq \theta + \delta,$$

ci accorgiamo che l'IdC bilaterale di livello α per θ rispetto a Y può essere interpretato anche come un intervallo in cui la stima Y di θ cadrà con probabilità $1 - \alpha$ (lo chiameremo IdC di livello α *per* Y *rispetto a* θ). Analogamente, riscrivendo le disuguaglianze

$$\theta \geq Y - \delta, \qquad \theta \leq Y + \delta,$$

come

$$Y \leq \theta + \delta, \qquad Y \geq \theta - \delta,$$

vediamo che gli IdC unilaterali (destro e sinistro) di livello α per θ rispetto a Y, sono anche IdC unilaterali (sinistro e destro) per Y rispetto a θ. □

Esempio 6.7.: Exit-poll II. Riprendiamo l'esempio 6.3 e cerchiamo di fornire un intervallo di confidenza per il parametro p stimato dalla media campionaria $\overline{X}$. Poiché $\overline{X}$ è corretto, per ogni $\delta \geq 0$ possiamo scrivere

$$P\left(\left|\overline{X} - p\right| \geq \delta\right) = P\left(\left|\overline{X} - E\left[\overline{X}\right]\right| \geq \delta\right) \leq \frac{\mathrm{Var}\left[\overline{X}\right]}{\delta^2},$$

dove abbiamo usato la disuguaglianza di Cebicev (2.17). Quindi, usando la (4.24) e la (3.2), si ha

$$P\left(\left|\overline{X} - p\right| \geq \delta\right) \leq \frac{p(1-p)}{n\delta^2},$$

e dunque

$$P\left(p \notin (\overline{X} - \delta, \overline{X} + \delta)\right) \leq \frac{p(1-p)}{n\delta^2}.$$

Questa stima ancora non serve, perché dipende dallo stesso p che vogliamo stimare. Tuttavia si può utilizzare il fatto che $p(1-p) \leq 1/4$, per $p \in [0,1]$ (verificare per esercizio) e dunque possiamo scrivere

$$P\left(p \notin (\overline{X} - \delta, \overline{X} + \delta)\right) \leq \frac{1}{4n\delta^2},$$

trovando così un IdC di livello $\alpha = \frac{1}{4n\delta^2}$. Se fissiamo il livello di errore α desiderato, si dovrà prendere $\delta = \frac{1}{2\sqrt{n\alpha}}$. In definitiva, un IdC bilaterale di livello α per il parametro p è

$$I_{\overline{X}} = \left(\overline{X} - \frac{1}{2\sqrt{n\alpha}}\,,\ \overline{X} + \frac{1}{2\sqrt{n\alpha}}\right).$$

Notiamo che, fissato il livello di errore, l'intervallo si stringe sempre più al crescere di n, ovvero, come ci potevamo aspettare, maggiore è il numero di intervistati, migliore è la probabilità che il valore vero sia vicino al valore stimato

$\overline{X}$. Con i numeri dell'esempio 6.3, se chiediamo un livello di errore $\alpha = 0.1$ (ovvero un IdC al 90%), troveremo

$$\delta = \frac{1}{2\sqrt{500}} \approx 0.02$$

e quindi, ricordando che la stima di p è 0.469, si conclude che un IdC al 90% per p è dato da $(0.449, 0.489)$. Di solito qusto risultato viene presentato in questa forma:

$$p = 46.9\% \pm 2\%$$

e va letto così: *il valore stimato di p è 46.9% e comunque la probabilità che p si discosti dal valore stimato per più di due punti prcentuali è non più del* 10%. Quel $\pm 2\%$ è la famosa "forbice" che accompagna sempre i risultati degli exit-poll (o delle proiezioni): solitamente si omette di dire a quale livello di errore si riferisce (evidentemente è un livello standard, piuttosto basso). □

6.2 Criteri per la scelta degli stimatori.

Finora non abbiamo dato alcuna indicazione su come scegliere lo stimatore di un parametro, a parte la richiesta che lo stimatore sia corretto (cosa che, per di più, non è neanche sempre veramente necessaria). Talvolta la scelta è intuitiva, come nel caso della media campionaria, talvolta non lo è affatto per cui è bene lasciarsi guidare da alcuni criteri. Qui di seguito ne enunciamo alcuni fra i più noti, tenendo presente che non sono gli unici possibili.

6.2.1 Minimizzazione del rischio quadratico.

Definizione 6.8. Si definisce *rischio quadratico* dello stimatore Y di un parametro θ la funzione

$$R_Y(\theta) := E\left[(Y-\theta)^2\right]. \tag{6.3}$$

□

Osserviamo che se Y è uno stimatore corretto, $R_Y(\theta)$ non è altro che la varianza di Y (definizione 2.16). In generale, fra una certa famiglia di stimatori si cerca di scegliere quello col più basso rischio quadratico, poiché a questa maniera si riducono le fluttuazioni dello stimatore attorno al suo valore medio.
Nell'esempio 6.3 abbiamo introdotto uno stimatore corretto del parametro p per un campione Bernoulliano, la media campionaria $\overline{X}$. Ripercorrendo gli stessi passaggi svolti nell'esempio, è facile dimostrare che $\overline{X}$ è uno stimatore corretto della media μ di qualunque tipo di distribuzione (purché, ovviamente, di media finita). Notiamo però che, per ogni $k \leq n$, anche $\overline{X}^{(k)} := \frac{X_1+X_2+\cdots+X_k}{k}$ è uno stimatore corertto di μ. Possiamo però notare che, posto $\sigma^2 = \mathrm{Var}[X_1] = \mathrm{Var}[X_2] = \cdots = \mathrm{Var}[X_n]$,

$$R_{\overline{X}^{(k)}}(\mu) = \mathrm{Var}\left[\frac{X_1+X_2+\cdots+X_k}{k}\right] = \frac{k\,\sigma^2}{k^2} = \frac{\sigma^2}{k},$$

da cui segue che la scelta che minimizza il rischio quadratico è $k = n$.

6.2.2 Stimatori di massima verosimiglianza.

Sia $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ un campione statistico con distribuzione $\mathcal{D}(\theta)$, dove θ è un parametro incognito. Un possibile criterio per la scelta di uno stimatore Y di θ è il seguente: se la mia osservazione effettiva è $x_1, x_2, \ldots, x_n$, può essere ragionevole pensare che mi sia imbattuto in un'osservazione che aveva un'alta probabilità di verificarsi. Allora, in mancanza di altri criteri, la mia stima di θ sarà quella per cui la probabilità $P^\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ (dove l'apice serve a ricordare che la probabilità dipende dal parametro θ) assume il valore massimo.

Definizione 6.9. Uno stimatore Y di θ è detto *di massima verosimiglianza (maximum likelihood)* se la "funzione di verosimiglianza"

$$L(\theta) = P^\theta(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$$

raggiunge un massimo assoluto per il valore di θ stimato da Y. □

Osserviamo tre cose importanti:

1. la funzione di verosimglianza potrebbe anche avere più di un massimo assoluto per cui, in generale, lo stimatore di massima verosimiglianza non è unico;

2. Y dipende dal risultato dell'esperimento per cui è effettivamente una variabile aleatoria;

3. nel caso, che sempre consideriamo, del campione indipendente si ha la fattorizzazione

$$L(\theta) = P^\theta(X_1 = x_1)\, P^\theta(X_2 = x_2) \cdots P^\theta(X_n = x_n).$$

Se torniamo al nostro esempio elettorale (esempio 6.3), indicando con m il numero dei votanti per il candidato A, la funzione di verosimiglianza in quel caso è

$$L(p) = P^p\Big(\sum_{i=1}^{n} X_i = m\Big) = p^m(1-p)^{n-m}, \qquad 0 \le p \le 1,$$

che ha un unico massimo assoluto in $p_{\max} = \frac{m}{n}$. Dunque $Y = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}$ è l'unico stimatore di massima verosimiglianza.
Il criterio si può generalizzare in modo ovvio al caso di più parametri incogniti $\theta_1, \theta_2, \ldots, \theta_s$: la scelta degli stimatori $Y_1, Y_2, \ldots, Y_s$ sarà fatta in modo da massimizzare la funzione di verosimiglianza

$$L(\theta_1, \theta_2, \ldots, \theta_s) = P^{\theta_1, \theta_2, \ldots, \theta_s}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n).$$

In molti casi è comodo cercare i massimi non della funzione L ma del suo logaritmo:

$$\log L(\theta_1, \theta_2, \ldots, \theta_s) = \sum_{i=1}^{n} \log P^{\theta_1, \theta_2, \ldots, \theta_s}(X_i = x_i)$$

(questa uguaglianza vale nel caso di indipendenza del campione). A questo modo la ricerca dei massimi (o più in generale dei punti critici) della funzione $L(\theta)$ equivale a risolvere la cosiddetta *equazione di verosimiglianza*

$$\sum_{i=1}^{n} \frac{\partial}{\partial \theta_k} \log P^{\theta_1,\theta_2,\ldots,\theta_s}(X_i = x_i) = 0, \qquad k = 1, 2, \ldots, s.$$

6.2.3 Stimatori Bayesiani.

Supponiamo di avere fissato a priori uno stimatore Y di un parametro θ per un campione statistico $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. Abbiamo quindi una stima Y di θ con la quale, poiché la distribuzione di $\boldsymbol{X}$ dipende dal paramtero stimato, possiamo valutare le probabilità di $\boldsymbol{X}$. Supponendo per semplicità che Y e $\boldsymbol{X}$ siano v.a. discrete (ma tutto il ragionamento si può estendere al caso continuo), possiamo dire di conoscere

le probabilità a priori $P(Y = \theta_k)$,

le probabilità condizionate $P(\boldsymbol{X} = \boldsymbol{x} \mid Y = \theta_k)$.

Possiamo allora pensare di rivalutare la probabilità a priori usando la formula di Bayes (1.11)

$$P(Y = \theta_k \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{P(\boldsymbol{X} = \boldsymbol{x} \mid Y = \theta_k)\, P(Y = \theta_k)}{\sum\limits_i P(\boldsymbol{X} = \boldsymbol{x} \mid Y = \theta_i)\, P(Y = \theta_i)},$$

ottenendo così una distribuzione a posteriori per Y. La *stima Bayesiana* di θ è quel valore di θ *che rende minimo il rischio quadratico a posteriori*

$$R_{\boldsymbol{x}}(\theta) = \sum_k (\theta - \theta_k)^2 P(Y = \theta_k \mid \boldsymbol{X} = \boldsymbol{x}) \ .$$

Si può dimostrare che:

1. il minimo di $R_{\boldsymbol{x}}$ viene assunto in corrispondenza del *valore atteso a posteriori*
$$\theta_B = \sum_k \theta_k P(Y = \theta_k \mid \boldsymbol{X} = \boldsymbol{x})$$
che è dunque lo stimatore Bayesiano cercato;

2. lo stimatore Bayesiano è corretto.

Consideriamo ad esempio un'urna contenente N palline, con un certo numero incognito $0 \leq m \leq N$ di palline rosse e le rimanenti $N - m$ bianche. Se estraiamo a caso una pallilna, essa sarà rossa con probabilità m/N e dunque il risultato di un'estrazione è una v.a. Bernoulliana $X_i \sim \mathcal{B}(m/N)$, in cui compare il parametro incognito m. Estraendo $n \leq N$ palline con reimbussolamento, otteniamo un campione statistico $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ col quale vogliamo stimare il parametro incognito m. È ragionevole supporre a priori che tutti i valori di m siano equiprobabili, per cui il nostro stimatore a priori Y avrà distribuzione discreta costante

$$P(Y = m) = \frac{1}{N+1}, \qquad 0 \leq m \leq N.$$

Osserviamo ora che, supponendo di aver ottenuto r palline rosse su n estrazioni, la probabilità di m condizionata a $\boldsymbol{x}$ è data da

$$P(\boldsymbol{X} = \boldsymbol{x} \mid Y = m) = \left(\frac{m}{N}\right)^r \left(1 - \frac{m}{N}\right)^{n-r}.$$

Applichiamo la formula di Bayes (che si semplifica in quanto le probabilità a priori sono tutte uguali) per ottenere le probabilità a posteriori

$$p_{m,\boldsymbol{x}} = P(Y = m \mid \boldsymbol{X} = \boldsymbol{x}) = \frac{\left(\frac{m}{N}\right)^r \left(1 - \frac{m}{N}\right)^{n-r}}{\sum_{k=0}^{N} \left(\frac{k}{N}\right)^r \left(1 - \frac{k}{N}\right)^{n-r}} = \frac{m^r (N-m)^{n-r}}{\sum_{k=0}^{N} k^r (N-k)^{n-r}}$$

(ricordiamo che r dipende da $\boldsymbol{x}$). A questo punto lo stimatore Bayesiano è dato dal valore atteso di m rispetto alle probabilità a posteriori $p_{m,\boldsymbol{x}}$ (fornite dalla formula precedente):

$$R = \sum_{m=0}^{N} m\, p_{m,\boldsymbol{x}}\,.$$

Se scegliamo come probabilità a priori per m una distribuzione binomiale con parametro $p = r/n$,

$$P(Y = m) = \binom{N}{m} \left(\frac{r}{n}\right)^m \left(1 - \frac{r}{n}\right)^{N-m}, \qquad 0 \le m \le N,$$

(scelta che può essere ancora più ragionevole della precedente), otteniamo le probabilità a posteriori

$$\begin{aligned} P(Y = m \mid \boldsymbol{X} = \boldsymbol{x}) &= \frac{\left(\frac{m}{N}\right)^r \left(1 - \frac{m}{N}\right)^{n-r} \binom{N}{m} \left(\frac{r}{n}\right)^m \left(1 - \frac{r}{n}\right)^{N-m}}{\sum_{k=0}^{N} \left(\frac{k}{N}\right)^r \left(1 - \frac{k}{N}\right)^{n-r} \binom{N}{k} \left(\frac{r}{n}\right)^k \left(1 - \frac{r}{n}\right)^{N-k}} \\ &= \frac{m^r (N-m)^{n-r}\, r^m (n-r)^{N-m} / (N-m)!\, m!}{\sum_{k=0}^{N} k^r (N-k)^{n-r}\, r^k (n-r)^{N-k} / (N-k)!\, k!} \end{aligned}$$

e un diverso stimatore Bayesiano.

6.3 Media e varianza campionarie.

Nei paragrafi precedenti abbiamo già parlato della *media campionaria*

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}. \tag{6.4}$$

che è uno stimatore corretto della media di qualunque distribuzione (purché con media finita). È utile studiare anche la varianza della v.a. $\overline{X}$. Se $\mathrm{Var}\,[X_i] = \sigma^2$, sfruttando l'indipendenza degli individui e utilizzando la proprietà (4.24) della varianza, possiamo scrivere:

$$\mathrm{Var}\,\left[\overline{X}\right] = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\,[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

Abbiamo perciò il seguente risultato.

Proposizione 6.10. *Sia $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ un campione statistico di rango n la cui distribuzione ha media finita μ e varianza finita σ^2. Allora la media campionaria $\overline{X}$ ha media e varianza date da*

$$E\left[\overline{X}\right] = \mu, \qquad \text{Var}\left[\overline{X}\right] = \frac{\sigma^2}{n}. \tag{6.5}$$

Dopo aver visto uno stimatore corretto della media è naturale chiedersi quale sia uno stimatore corretto della varianza σ^2. Conviene distinguere due casi. Se la media μ è nota si può usare lo stimatore

$$\overline{\sigma}^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu\right)^2, \tag{6.6}$$

(verificare per esercizio che è uno stimatore corretto di σ^2). Di fatto però questo è un caso che si presenta poco frequentemente: in generale la media è anch'esso un parametro da stimare. È naturale quindi sostituire a μ nella precedente espressione la sua stima fornita dalla media campionaria (6.4). Tuttavia lo stimatore così ottenuto *risulta non corretto a meno che non si sostituisca il fattore* $\frac{1}{n}$ *con* $\frac{1}{n-1}$, ottenendo in questo modo la cosiddetta *varianza campionaria*:

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2. \tag{6.7}$$

Proposizione 6.11. *Sia $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ un campione statistico di rango n con varianza finita σ^2. Allora la varianza campionaria* (6.7) *è uno stimatore corretto di σ^2.*

Dimostrazione Utilizzando le note proprietà di media e varianza si ottiene

$$E\left[S^2\right] = \frac{1}{n-1}\sum_{i=1}^{n} E\left[X_i^2 - 2X_i\overline{X} + \overline{X}^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n} E\left[X_i^2\right] - nE\left[\overline{X}^2\right]\right).$$

Ricordando che per una v.a. Y si ha $E\left[Y^2\right] = \text{Var}\left[Y\right] + E\left[Y\right]^2$, e posto $\mu = E\left[X_i\right]$, si può perciò scrivere

$$\begin{aligned} E\left[S^2\right] &= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\sigma^2 + \mu^2\right) - n\text{Var}\left[\overline{X}\right] - nE\left[\overline{X}\right]^2\right) \\ &= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \sigma^2 - n\mu^2\right] = \sigma^2 \end{aligned}$$

□

Il fatto che la varianza campionaria sia uno stimatore corretto spiega finalmente il perché della definizione (5.10) introdotta nell'ambito della statistica descrittiva.

Per poter dire qualcosa di più sulla distribuzione della varianza campionaria dobbiamo metterci nel caso particolare di un *campione normale* e compiere un breve excursus di carattere matematico.

Definizione 6.12. Un *campione normale* (di rango n) è un campione statistico $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ con distribuzione normale, ovvero

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \qquad i = 1, 2, \ldots, n$$

(ricordiamo anche che per definizione di campione statistico le X_i devono essere tutte indipendenti tre loro). □

Nel paragrafo 3.7 abbiamo introdotto le ditribuzioni $\mathcal{X}^2$ e t *di Student* e successivamente, nell'Esempio 4.19, ne abbiamo illustrato il significato. Ricordiamo che:

1. se $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_n)$ è campione normale standard di rango n, cioè un vettore di variabili aleatorie normali standard indipendenti, allora
$$Z_1^2 + Z_2^2 + \cdots + Z_n^2 \sim \mathcal{X}^2(n);$$
2. se $Z \sim \mathcal{N}$, $Y \sim \mathcal{X}^2(n)$ e X, Y sono indipendenti, allora
$$\sqrt{n}\,\frac{Z}{\sqrt{Y}} \sim t(n).$$

Consideriamo ora la varianza campionaria (6.7) di un campione normale con media μ e varianza σ^2:

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2, \qquad X_i \sim \mathcal{N}(\mu, \sigma^2).$$

Usando la Proposizione 4.23) si ha che

$$\overline{X} \sim \mathcal{N}(\mu,\, \sigma^2/n), \qquad \Delta X_i := X_i - \overline{X} \sim \mathcal{N}(0,\, \sigma^2 + \sigma^2/n)$$

dove la v.a. ΔX_i è detta *scarto i-esimo*. Pertanto, standardizzando ΔX_i, si ha

$$\frac{\sqrt{n}}{\sigma\sqrt{n+1}}\,\Delta X_i \sim \mathcal{N}$$

e dunque saremmo (erroneamente) portati a concludere che la media campionaria, a meno di un fattore costante, ha una distribuzione di tipo $\mathcal{X}^2(n)$:

$$\frac{n(n-1)}{\sigma^2(n+1)}\,S^2 = \sum_{i=1}^{n}\left(\frac{\sqrt{n}}{\sigma\sqrt{n+1}}\,\Delta X_i\right)^2 \sim \mathcal{X}^2(n) \qquad \text{(sbagliato!)}$$

L'errore sta nel fatto di aver considerato gli scarti come variabili *indipendenti* mentre invece non lo sono; si ha infatti

$$\sum_{i=1}^{n}\Delta X_i = 0 \tag{6.8}$$

il che significa che ogni ΔX_i si può esprimere in funzione degli altri $n-1$ scarti (e dunque i ΔX_i non sono indipendenti tra loro).
La giusta risposta alla domanda su quale sia la distribuzione della media campionaria viene da un teorema "profondo", la cui dimostrazione (non semplice) si può trovare sul libro del Baldi [1].

Teorema 6.13. (Teorema di Cochran)
Sia S^2 la varianza campionaria di un campione normale di rango n con varianza σ^2. Allora si ha

$$\frac{n-1}{\sigma^2}\, S^2 \sim \mathcal{X}^2(n-1), \tag{6.9}$$

ovvero la varianza campionaria, a meno di un'opportuna costante moltiplicativa, ha distribuzione $\mathcal{X}^2$ con $n-1$ gradi di libertà.

Intuitivamente si può pensare che il numero dei gradi di libertà corrisponda al numero degli scarti indipendenti (da cui il nome "gradi di libertà"): difatti il vincolo lineare (6.8) "toglie un grado di libertà", per cui restano $n-1$ scarti indipendenti.[3]

Un importante corollario del Teorema di Cochran è il seguente.

Corollario 6.14.
Siano $\overline{X}$ e S^2 media e varianza campionarie di un campione normale di rango n con media μ e varianza σ^2. Allora si ha

$$\frac{\sqrt{n}\,(\overline{X}-\mu)}{S} \sim t(n-1). \tag{6.10}$$

Dimostrazione Osserviamo che si può scrivere

$$\frac{\sqrt{n}\,(\overline{X}-\mu)}{S} = \sqrt{n-1}\;\frac{\frac{\sqrt{n}}{\sigma}\,(\overline{X}-\mu)}{\sqrt{\frac{n-1}{\sigma^2}\,S^2}} = \sqrt{n-1}\,\frac{Z}{\sqrt{Y}},$$

dove, usando il Teorema di Cochran,

$$Z := \frac{\sqrt{n}}{\sigma}\,(\overline{X}-\mu) \sim \mathcal{N}, \qquad Y := \frac{n-1}{\sigma^2}\,S^2 \sim \mathcal{X}^2(n-1),$$

da cui, per la nota proprietà della distribuzione t, segue la tesi. □

Notiamo che la v.a. $\frac{\sqrt{n}\,(\overline{X}-\mu)}{S}$ è una sorta di standardizzazione della media campionaria fatta usando lo stimatore $S = \sqrt{S^2}$ al posto della vera deviazione standard σ.

[3] Una versione più generale del Teorema di Cochran stabilisce che la somma di n scarti quadratici soggetti a s vincoli ha distribuzione di tipo $\mathcal{X}^2(n-s)$.

6.4 Stima di media e varianza per campioni normali

Ci sono almeno tre ragioni per dare particolare rilievo al caso dei campioni normali:

1. in moltissimi casi concreti si ha a che fare con distribuzioni normali (si veda a proposito anche la discussione svolta nel paragrafo 4.6);
2. la teoria matematica delle v.a. normali è ben sviluppata e si hanno a disposizione importanti risultati quali quelli del paragrafo precedente;
3. grazie al Teorema Centrale di Convergenza (teorema 4.28) le formule che troveremo potranno essere utilizzate anche per campioni non normali, purché sufficientemente numerosi (si veda la discussione al termine del presente paragrafo).

Dividendo la discussione in quattro casi distinti, vogliamo adesso ricavare formule generali che ci danno stime e intervalli di confidenza per media e varianza di campioni normali.

Stima di μ con σ^2 nota. Come stimatore di μ si usa la media campionaria $\overline{X}$. Fissato il livello α, cerchiamo innanzitutto un IdC bilaterale per μ rispetto a $\overline{X}$ (si veda la definizione 6.5), cioè del tipo $\overline{X} - \delta \leq \mu \leq \overline{X} + \delta$, con $\delta > 0$ da individuare in funzione di α. Osservando che $\overline{X} - \delta \leq \mu \leq \overline{X} + \delta$ equivale a $\left|\overline{X} - \mu\right| \leq \delta$, si dovrà avere

$$1 - \alpha = P\left(\left|\overline{X} - \mu\right| \leq \delta\right) = P\left(\left|\frac{\sqrt{n}}{\sigma}(\overline{X} - \mu)\right| \leq \frac{\delta\sqrt{n}}{\sigma}\right) = P\left(|Z| \leq \frac{\delta\sqrt{n}}{\sigma}\right),$$

dove $Z := \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) \sim \mathcal{N}$. Osserviamo che, grazie alla simmetria della curva, chiedere $P(|Z| \leq x) = 1 - \alpha$ equivale a chiedere $P(Z < x) = 1 - \alpha/2$ (si veda la figura 6.1 (a)). Per definizione di quantile (paragrafo 2.2) si ha allora che la x che soddisfa $P(|Z| \leq x) = 1 - \alpha$ è data da $x = q^{\mathcal{N}}_{1-\alpha/2}$ (dove, ricordiamo, $q^{\mathcal{N}}_{1-\alpha/2}$ indica il quantile di ordine $1 - \alpha/2$ della distribuzione normale standard $\mathcal{N}$). Dunque, se vogliamo trovare δ tale che $P\left(\left|\overline{X} - \mu\right| \leq \delta\right) = 1 - \alpha$ basterà prendere $\frac{\delta\sqrt{n}}{\sigma} = q^{\mathcal{N}}_{1-\alpha/2}$, ovvero

$$\delta = \frac{\sigma\, q^{\mathcal{N}}_{1-\alpha/2}}{\sqrt{n}} \qquad \text{(caso bilaterale).} \tag{6.11}$$

Consideriamo ora un IdC unilaterale per μ rispetto a $\overline{X}$, che può essere destro, $\mu \geq \overline{X} - \delta$, o sinistro, $\mu \leq \overline{X} + \delta$ (si veda ancora la definizione 6.5). Nel primo caso si dovrà avere

$$1 - \alpha = P\left(\mu \geq \overline{X} - \delta\right) = P\left(\overline{X} - \mu \leq \delta\right) = P\left(Z \leq \frac{\delta\sqrt{n}}{\sigma}\right),$$

dove $Z := \frac{\sqrt{n}}{\sigma}(\overline{X} - \mu) \sim \mathcal{N}$. Pertanto (si veda la figura 6.1 (b)) si dovrà prendere $\frac{\delta\sqrt{n}}{\sigma} = q^{\mathcal{N}}_{1-\alpha}$, ovvero

$$\delta = \frac{\sigma\, q^{\mathcal{N}}_{1-\alpha}}{\sqrt{n}} \qquad \text{(caso uniterale).} \tag{6.12}$$

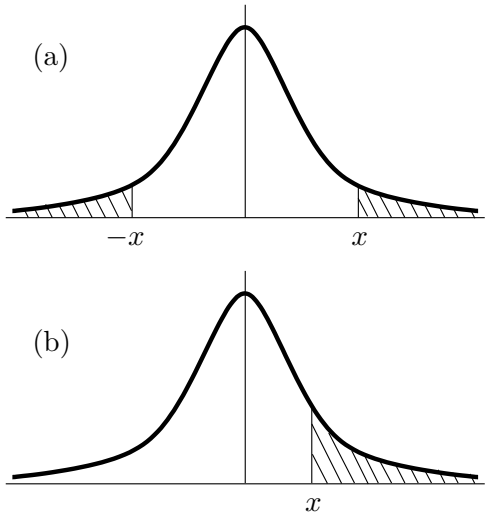


Figura 6.1: In entrambe queste figure l'area della regione tratteggiata è α e l'area della regione non tratteggiata è $1-\alpha$. Ne segue che, nel caso (a), x è il quantile di ordine $1-\alpha/2$ mentre, nel caso (b), x è il quantile di ordine $1-\alpha$.

Si verifichi per esercizio che questa stessa formula vale anche per l'IdC unilaterale sinistro.

Stima di μ con σ^2 incognita. Nel caso in cui la varianza σ^2 sia anch'essa un parametro incognito, le formule (6.11) e (6.12) sono inutilizzabili perché presuppongono la conoscenza di σ. Per trovare un IdC per μ in questo caso, l'idea più naturale è quella di ripercorrere il ragionamento svolto nel caso precedente *sostituendo a σ la sua stima data da S, radice quadrata della varianza campionaria S^2*. Dunque, nel caso bilaterale, scriviamo

$$P\left(\left|\overline{X}-\mu\right| \le \delta\right) = P\left(\left|\frac{\sqrt{n}}{S}(\overline{X}-\mu)\right| \le \frac{\delta\sqrt{n}}{S}\right).$$

Adesso la v.a. $\frac{\sqrt{n}}{S}(\overline{X}-\mu)$ non è più una normale standard ma questo non è importante perché quello che conta è che la sua distribuzione sia nota. E difatti, grazie al Teorema di Cochran e in particolare al Corollario 6.14, sappiamo che la distribuzione di $\frac{\sqrt{n}}{S}(\overline{X}-\mu)$ è $t(n-1)$, la t di Student a $n-1$ gradi di libertà. Poiché la distribuzione t ha la stessa simmetria della distribuzione $\mathcal{N}$, tutto fila come nel caso precedente, a patto di sostituire $\mathcal{N}$ con $t(n-1)$ e σ con S. Fissato il livello α si ottiene perciò

$$\delta = \frac{S\, q_{1-\alpha/2}^{t(n-1)}}{\sqrt{n}} \qquad \text{(caso bilaterale).} \tag{6.13}$$

Nel caso unilaterale si avrà

$$\delta = \frac{S\, q_{1-\alpha}^{t(n-1)}}{\sqrt{n}} \qquad \text{(caso unilaterale).} \tag{6.14}$$

La formule ottenute sono analoghe a quelle del caso di varianza nota, con la notevole differenza che si deve utilizzare il quantile della $t(n-1)$ anziché quello della $\mathcal{N}$ e che il termine S è calcolato sul campione, mentre nel caso precedente σ era un numero noto a priori.

Osservazione 6.15. Le formule (6.11), (6.12), (6.13) e (6.14) possono essere anche usate per dimensionare il campione in funzione di un'ampiezza richiesta per l'IdC. Ad esempio la (6.13) ci dà

$$n = \left(\frac{S\, q_{1-\alpha/2}^{t(n-1)}}{\delta} \right)^2 ,$$

che stabilisce quanto dev'essere numeroso il campione per avere una certa ampiezza δ dell'IdC bilaterale di livello α nel caso di varianza incognita. Nei casi concreti si può fare un pre-campionamento ridotto, per avere un'idea della stima di σ, e decidere quanti campioni raccogliere in base alla formula precedente. □

Stima di σ^2 con μ nota. Essendo la media nota, come stimatore della varianza si può usare $\overline{\sigma}^2$, dato dalla formula (6.6). Cerchiamo ora di ottenere una stima dall'alto di σ^2, ovvero un IdC per σ^2 del tipo $0 \leq \sigma^2 \leq \delta$ (un limite inferiore per σ^2, che è una quantità positiva, è sicuramente 0).[4] Possiamo allora scrivere

$$P\left(\sigma^2 \geq \delta\right) = P\left(\frac{1}{\sigma^2} \leq \frac{1}{\delta}\right) = P\left(\frac{n\,\overline{\sigma}^2}{\sigma^2} \leq \frac{n\,\overline{\sigma}^2}{\delta}\right)$$

e osservare che la v.a. $\frac{n\overline{\sigma}^2}{\sigma^2}$ ha distribuzione $\mathcal{X}^2(n)$, in quanto somma di n v.a. normali standard indipendenti (verificare per esercizio usando la definizione (6.6)). Fissato il livello α possiamo dunque scrivere che $P\left(\sigma^2 \geq \delta\right) = \alpha$ se $\frac{n\overline{\sigma}^2}{\delta} = q_\alpha^{\chi^2(n)}$, ovvero

$$\delta = \frac{n\,\overline{\sigma}^2}{q_\alpha^{\chi^2(n)}}, \tag{6.15}$$

dove $q_\alpha^{\chi^2(n)}$ è il quantile di ordine α della distribuzione $\mathcal{X}^2$ a n gradi di libertà.

Stima di σ^2 con μ incognita. Per stimare la varianza quando la media è incognita (che è il caso più frequente) si può utilizzare la varianza campionaria S^2, che ha l'espressione (6.7). Anche in questo caso ci limitiamo a cercare una stima dall'alto di σ^2, ovvero un IdC del tipo $0 \leq \sigma^2 \leq \delta$. Utilizziamo di nuovo il Teorema di Cochran, e in particolare la formula (6.9), che ci dice che la v.a. $\frac{n-1}{\sigma^2}S^2$ ha distribuzione nota e precisamente $\mathcal{X}^2(n-1)$. Scriviamo dunque

$$P\left(\sigma^2 \geq \delta\right) = P\left(\frac{1}{\sigma^2} \leq \frac{1}{\delta}\right) = P\left(\frac{n-1}{\sigma^2}\, S^2 \leq \frac{n-1}{\delta}\, S^2\right)$$

da cui, fissato il livello α, otteniamo

$$\delta = \frac{(n-1)\, S^2}{q_\alpha^{\chi^2(n-1)}}. \tag{6.16}$$

[4] Si potrebbe anche cercare una stima dal basso, migliore del semplice 0; rimandiamo a proposito al paragrafo 3.9.4 del libro di Camussi et al. [2].

ove $q_\alpha^{\chi^2(n-1)}$ è il quantile di ordine α della distribuzione $\mathcal{X}^2$ a $n-1$ gradi di libertà.

Osservazione 6.16. In statistica i quantili utilizzati per determinare gli intervalli di confidenza (o, come vedremo più avanti, nei test delle ipotesi) sono più comunemente detti "valori critici" e riportati in apposite tavole. □

Esempio 6.17. La Tabella 6.1 riporta i valori della concentrazione di cromo (espressa in microgrammi per grammo di peso della pianta) in nove campioni di piante di *Alyssum bertolonii*. Ciascun campione è stato raccolto in un sito diverso (i campioni sono numerati da *C1* a *C9*) e contiene 20 individui.[5] Di cia-

C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*
2.31	1.98	1.79	4.70	19.49	6.32	2.40	3.81	5.37
2.69	0.79	0.61	3.64	67.96	10.40	10.13	3.24	6.53
2.59	0.94	0.87	1.24	6.75	18.09	2.80	4.17	8.49
7.33	1.12	0.98	10.84	0.85	10.18	1.71	4.91	2.30
9.49	1.10	1.39	7.87	1.28	15.95	36.40	3.04	3.56
8.58	4.85	3.28	9.03	0.76	6.76	2.72	7.46	3.00
11.27	0.64	0.97	4.15	0.78	7.24	5.27	8.49	10.31
11.80	1.04	2.58	3.89	1.06	6.94	6.79	11.33	5.01
11.42	3.35	1.02	1.36	1.88	5.06	31.93	6.80	1.60
4.60	0.73	1.46	9.59	2.54	4.58	5.21	8.06	2.40
21.73	0.65	2.26	14.82	4.17	5.70	6.62	12.28	1.99
14.33	0.85	3.63	4.17	3.54	11.73	11.82	6.93	1.72
12.11	2.36	0.68	13.73	6.06	4.08	25.87	9.44	1.45
13.53	3.39	2.39	4.04	6.85	6.17	5.49	5.46	3.20
5.09	1.08	1.05	3.09	2.74	4.69	3.92	5.83	2.23
4.44	2.33	3.50	3.56	3.03	4.39	9.62	11.01	0.97
6.71	3.20	2.72	5.18	29.43	10.16	7.01	13.30	1.60
3.85	2.08	0.61	28.33	14.44	5.11	2.59	4.42	1.89
3.27	3.32	2.42	7.59	5.68	5.02	9.79	21.91	5.88
20.87	4.60	15.68	4.47	8.97	8.21	1.23	14.38	2.74

Tabella 6.1: Concentrazione di cromo (μg/g) in 9 diversi campioni di 20 piante di *Alyssum bertolonii* (dati gentilmente forniti da C. Gonnelli e F. Galardi del Dipartimento di Fisiologia Vegetale).

scuna colonna, considerata a sé come campione monovariato, possiamo stimare media e varianza. Poiché μ e σ^2 sono entrambe incognite dovremo utilizzare media e varianza campionarie come stimatori e le formule (6.13), (6.14) e (6.16) per il calcolo degli IdC. Naturalmente stiamo supponendo che i campioni siano normali, come spesso accade per misure di tipo biometrico. Osserviamo tuttavia che, essendo la numerosità dei campioni ($n = 20$) non molto elevata, la eventuale non-normalità del campione potrebbe portarci a errori grossolani. Per questo motivo, nella pratica, è indispensabile sottoporre i campioni a un *test di normalità* (cosa che anche noi faremo nel paragrafo 7.3).
La stima delle medie e delle varianze si ottengono calcolando la media campionaria (6.4) e la varianza campionaria (6.7) di ciascuna colonna. L'ampiezza

[5] Attenzione a non confondere questa tabella con una tabella multivariata, come quelle viste finora. Qui le diverse colonne non riportano diverse variabili misurate sugli stessi individui ma *la stessa* variabile misurata su *diversi gruppi* di individui.

$\delta_{\overline{X}}$ dell'IdC bilaterale per la media $\overline{X} - \delta_{\overline{X}} \leq \mu \leq \overline{X} + \delta_{\overline{X}}$ applicando la formula (6.13) ad ogni colonna mentre l'ampiezza δ_{S^2} dell'IdC per la varianza $0 \leq \sigma^2 \leq \delta_{S^2}$ si calcola applicando la formula (6.16) ad ogni colonna. Se, in particolare siamo interessati a intervalli di confidenza di livello $\alpha = 0.05$ (cioè al 95%), dovremo usare (per ogni colonna) il quantile

$$q_{1-\alpha/2}^{t(n-1)} = q_{0.975}^{t(19)} \approx 2.09$$

per il calcolo di $\delta_{\overline{X}}$ e il quantile

$$q_{\alpha}^{\chi^2(n-1)} = q_{0.05}^{\chi^2(19)} \approx 10.12$$

per il calcolo di δ_{S^2} (questi valori si trovano facilmente sulle tavole). I risultati sono riportati nella tabella (6.2) e il lettore li può verificare per esercizio. □

	C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*
$\overline{X}$	8.90	2.02	2.49	7.26	9.41	7.84	9.47	8.31	3.61
$\delta_{\overline{X}}$	2.69	0.63	1.52	2.91	7.27	1.81	4.70	2.19	1.187
S^2	33.11	1.82	10.61	38.93	242.32	15.013	101.26	21.95	6.46
δ_{S^2}	62.16	3.41	19.92	73.09	454.96	28.19	190.12	41.20	12.12

Tabella 6.2: Media e varianza campionarie, e rispettive ampiezze degli IdC al 95%, relative ai 9 campioni della tabella 6.1.

Esempio 6.18. Consideriamo adesso la prima colonna della tabella 5.1 (variabile "lunghezza totale", espressa in millimetri). Anche qui supponiamo che il campione sia normale (ma in questo caso, anche se così non fosse, la numerosità è abbastanza elevata da poter ancora utilizzare le formule dei campioni normali, come diremo brevemente alla fine di questo paragrafo). Anche in questo caso μ e σ^2 sono entrambe incognite. Media e varianza campionarie danno le stime

$$\overline{x} = 157.98, \qquad s^2 = 13.35.$$

Per calcolare un IdC bilaterale al 5% per la media usando la formula (6.13) dovremmo calcolare (o meglio, cercare sulle tavole) il quantile $q_{0.975}^{t(48)}$. D'altra parte, come abbiamo osservato in precedenza (Esempio 4.29) si può approssimare $t(48)$ con $\mathcal{N}$ e dunque utilizzeremo $q_{0.975}^{\mathcal{N}}$:

$$\delta = \frac{\sqrt{13.35}}{\sqrt{49}}\, q_{0.975}^{\mathcal{N}} = \frac{\sqrt{13.35}}{\sqrt{49}} \times 1.96 = 1.02.$$

Pertanto la lunghezza media è stimata in 157.98 ± 1.02 con una probabilità di errore del 5%.
Per calcolare un IdC al 5% per la varianza usando la formula (6.16) dobbiamo utilizzare il quantile $q_{0.05}^{\chi^2(48)}$. Se questo non fosse disponibile sulle tavole, si può usare la formula approssimata

$$q_{\alpha}^{\chi^2(n)} \approx \sqrt{2n}\, q_{\alpha}^{\mathcal{N}} + n, \tag{6.17}$$

che si ricava dalla (4.32). Si ottiene così

$$\delta = \frac{48 \times 13.35}{q_{0.05}^{\chi^2(48)}} = 20.10$$

cioè la varianza è stimata in 13.35 con limite superiore 20.10 con una probabilità di errore del 5%.

□

Concludiamo dicendo che i risulati ottenuti per campioni normali possono essere utilizzati anche per campioni statistici qualsiasi *purché sufficientemente numerosi*[6] (questo fatto può essere dimostrato utilizzando il Teorema Centrale di Convergenza).

Esempio 6.19.: Exit-poll III. Proviamo a fornire un nuovo intervallo di confidenza per i risultati dell'exit-poll che abbiamo considerato negli esempi 6.3 e 6.7. Poiché un campione di 5000 individui è senz'altro abbastanza numeroso utilizziamo la formula (6.13). Per calcolare rapidamente S usiamo il seguente trucco: sfruttando $n \approx n-1$ e $X_i^2 = X_i$ possiamo scrivere

$$S^2 \approx \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i - \overline{X}^2 = \overline{X}(1-\overline{X}).$$

Dunque, ricordando che $\overline{X}$ risultava pari a 0.469, si ha

$$S \approx \sqrt{\overline{X}(1-\overline{X})} \approx 0.499.$$

Approssimando inoltre $t(n)$ con $\mathcal{N}$ (si veda (4.33)) abbiamo $q_{0.95}^{t(4999)} \approx q_{0.95}^{\mathcal{N}} \approx 1.645$. Pertanto dalla formula (6.13) si ottiene

$$\delta = \frac{0.499 \times 1.645}{\sqrt{5000}} \approx 0.012,$$

che è un risultato assai migliore di $\delta = 0.02$ trovato in precedenza. □

[6] L'inevitabile domanda *'Quanto numerosi?'* non ha una risposta univoca e preferiamo rinviare il lettore ai vari testi applicativi per possibili risposte.

Capitolo 7

Test delle ipotesi

7.1 Concetti generali

Quella dei *test delle ipotesi* è una teoria statistica per decidere se accettare o rifiutare una certa "ipotesi". Prima di addentrarci nella generalità della teoria cominciamo con un classico esempio.

Esempio 7.1.: Test dell'efficacia di un farmaco. Un'industria farmaceutica sostiene di aver prodotto un farmaco per abbassare il colesterolo migliore di quelli in uso. Supponiamo di sapere che il farmaco attualmente in uso ha come effetto un abbassamento percentuale del tasso di colesterolo che è una v.a. normale con distribuzione $X \sim \mathcal{N}(\mu_0, \sigma^2)$ (e supponiamo, ad esempio, $\mu_0 = 25$, $\sigma = 13$).[1] Dunque la casa farmaceutica sostiene che la sua nuova cura è capace di realizzare un abbassamento $\mathcal{N}(\mu, \sigma^2)$ con $\mu > \mu_0$, cioè che l'abbassamento medio sia maggiore di quello realizzato dal farmaco in uso (supponiamo per semplicità che non ci siano differenze nella varianza delle due distribuzioni, e che questa sia nota).
Il problema è dunque quello di accettare o rifiutare la cosiddetta *ipotesi nulla* H_0 che consiste nel *negare che il nuovo farmaco sia più efficace*, in simboli:

$$\mathrm{H}_0 : \mu \leq \mu_0 \,.$$

Questo test consiste quindi nel confronto di una media "teorica" μ_0 con una media "sperimentale" o "stimata" μ. Nei test statistici l'ipotesi nulla sottoposta a test è di solito quella più "prudente": in questo esempio l'ipotesi più prudente è che il nuovo farmaco *non* sia più efficace: difatti è considerato più pericoloso attribuire a un farmaco proprietà che non ha piuttosto che non attribuirgli proprietà che ha.
Il test procede così: si considera un campione di rango n, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, ovvero n pazienti cui è stato somministrato il nuovo farmaco e su ciascuno dei quali si misura l'abbassamento del tasso di colesterolo

$$X_i \sim \mathcal{N}(\mu, \sigma^2), \qquad \text{con } \sigma \text{ nota e } \mu \text{ incognita.}$$

Sul campione si calcola la media campionaria $\overline{X}$, che stima μ, dopodiché *si respinge* H_0 *se e solo se* $\overline{X}$ *è sufficientemente più grande di* μ_0. Cosa significa

[1] Si considera X positiva quando il colesterolo diminuisce.

"sufficientemente più grande"? Significa che dobbiamo cautelarci, con un opportuno margine di sicurezza, dall'errore di interpretare come fatto significativo (per esempio la reale efficacia del farmaco) un esito del test che potrebbe essere dovuto a semplici fluttuazioni statistiche del campione. Dunque rifiuteremo H_0 se e solo se

$$\overline{X} > \mu_0 + c,$$

dove $c > 0$ è il nostro "margine di sicurezza" (detto più propriamente *valore critico*) che sarà prefissato in considerazione degli errori ai quali l'esito del test può condurre. Tali errori possono essere:

1. **Errore di prima specie** (considerato il più grave):

$$H_0 \text{ } \textit{è vera e la si respinge.}$$

La probabilità di commettere questo errore è

$$P\left(\overline{X} > \mu_0 + c \,\middle|\, H_0 \text{ } \textit{è vera}\right) = P^{\mu \leq \mu_0}\left(\overline{X} > \mu_0 + c\right),$$

dove l'apice $\mu \leq \mu_0$ serve a ricordare che, supponendo H_0 vera, le probabilità vanno calcolate usando la distribuzione $\mathcal{N}(\mu, \sigma^2)$ con $\mu \leq \mu_0$. Poiché si ha

$$P^{\mu \leq \mu_0}\left(\overline{X} > \mu_0 + c\right) \leq P^{\mu = \mu_0}\left(\overline{X} > \mu_0 + c\right),$$

notiamo che l'errore di prima specie più grande possibile corrisponde al caso $\mu = \mu_0$ Tale valore massimo è detto *livello (di significatività) del test* e verrà indicato con α:

$$\alpha := \max_{\mu \leq \mu_0} P^{\mu \leq \mu_0}\left(\overline{X} > \mu_0 + c\right) = P^{\mu = \mu_0}\left(\overline{X} > \mu_0 + c\right)$$

Il complementare $1 - \alpha$ è detto *livello di protezione* del test.

2. **Errore di seconda specie** (considerato meno grave):

$$H_0 \text{ } \textit{è falsa e la si accetta.}$$

La probabilità di commettere questo errore è

$$\beta = P\left(\overline{X} \leq \mu_0 + c \,\middle|\, H_0 \text{ } \textit{è falsa}\right) = P^{\mu > \mu_0}\left(\overline{X} \leq \mu_0 + c\right),$$

e $1 - \beta$, cioè la probabilità di accettare il nuovo farmaco se questo è davvero più efficace, è detta *potenza del test*, ed è una funzione di $\mu > \mu_0$ (figura 7.1).
Il valore critico c viene quindi scelto in base al livello di significatività e alla potenza desiderati. Come si vede dalla figura 7.1, α e β non sono indipendenti: se spostiamo c in modo da diminuire α allora cresce β, e viceversa. In altre parole, se si cerca di aumentare la protezione, il test perde potenza, e viceversa. Una volta fissato il livello α dobbiamo calcolare il valore critico c in modo che sia soddisfata la condizione

$$P^{\mu = \mu_0}\left(\overline{X} > \mu_0 + c\right) = \alpha.$$

Poiché supponiamo $\mu = \mu_0$, la precedente condizione significa che stiamo cercando un IdC unilaterale di livello α per $\overline{X}$ rispetto a μ (si vadano la definizione

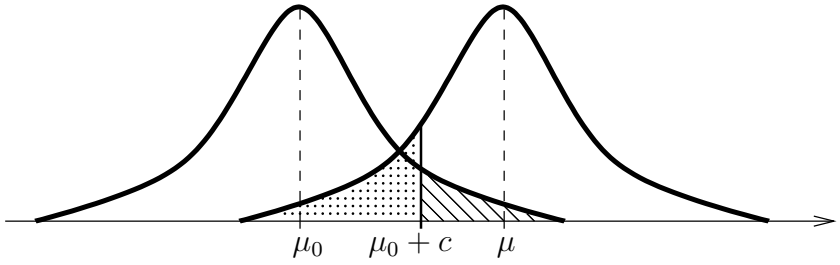


Figura 7.1: Le due curve rappresentano la distribuzione $\mathcal{N}(\mu, \sigma^2)$ nei due casi $\mu = \mu_0$ e $\mu > \mu_0$. Fissato il valore critico c, l'area tratteggiata è pari ad α, il livello del test: notiamo infatti che se $\mu < \mu_0$ l'area, che rappresenta l'errore di prima specie, diminuisce. L'area punteggiata ci dà invece la probabilità β dell'errore di seconda specie.

6.5 e l'osservazione 6.6) per un campione normale con varianza nota (si veda il paragrafo 6.4) e dunque, utilizzando la formula (6.12), si ottiene

$$c = \frac{\sigma}{\sqrt{n}} q^{\mathcal{N}}_{1-\alpha}.$$

Utilizzando i valori $\mu_0 = 25$, $\sigma = 13$, $n = 120$ e fissando il livello $\alpha = 0.01$, si ottiene

$$c = \frac{13}{\sqrt{120}} q^{\mathcal{N}}_{0.99} = 25 + \frac{13}{\sqrt{10.9}} \times 2.32 \approx 2.8.$$

In conclusione, se troveremo $\overline{X} > \mu_0 + c = 27.8$, respingeremo H_0, concludendo che il nuovo farmaco è davvero più efficace: è improbabile (l'1% di probabilità) che il valore misurato sia dovuto a fluttuazioni statistiche della distribuzione $\mathcal{N}(\mu_0, \sigma^2)$. Se invece troviamo $\overline{X} \leq 27.8$ accetteremo H_0, concludendo che il nuovo famaco non è più efficace dell'altro (c'è una probabilità non trascurabile che il valore di $\overline{X}$ sia dovuto a una fluttuazione statistica). □

In generale, la struttura di un test statistico è la seguente. Abbiamo un campione statistico di rango n: $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ con distribuzione $X_i \sim \mathcal{D}(\theta)$ dove θ rappresenta uno o più parametri incogniti della distribuzione.
Il test serve a decidere se accettare o rifiutare una *ipotesi nulla* H_0, che di solito ha la forma

$$\mathrm{H}_0 : \ \theta \in \Theta_0, \tag{7.1}$$

dove Θ_0 è un insieme di numeri reali. L'ipotesi nulla, cioè, consiste nell'affermare che il parametro incognito θ appartiene a un certo insieme Θ_0 di possibili valori. Nell'esempio precedente, θ era μ e Θ_0 era $(-\infty, \mu_0]$.
L'esito del test (accettare o rifiutare H_0) è determinato da una *regola di decisione* di questo tipo: si sceglie uno stimatore T di θ e una *regione di accettazione* $\mathcal{R} \subset \mathbb{R}$, dopodiché

$$\begin{aligned} &\textit{se } T \in \mathcal{R} \textit{ si accetta } \mathrm{H}_0 \\ &\textit{se } T \notin \mathcal{R} \textit{ si rifiuta } \mathrm{H}_0 \end{aligned} \tag{7.2}$$

(per questo motivo $\mathcal{R}$ è detta *regione di accettazione* e la sua complementare $\mathcal{R}^c$ è detta *regione di rifiuto* o *critica*). Nell'esempio precedente avevamo $T = \overline{X}$,

$\mathcal{R} = (-\infty, \mu_0 + c]$, $\mathcal{R}^c = (\mu_0 + c, +\infty)$. In casi analoghi l'estremo c della regione di accettazione si chiama *valore critico.*
Il valore critico si sceglie in base a considerazioni sulla probabilità di commettere i due tipi di errore: quello *di prima specie*

$$\mathrm{H}_0 \text{ } \textit{è vera e la si rifiuta}$$

e quello *di seconda specie*

$$\mathrm{H}_0 \text{ } \textit{è falsa e la si accetta.}$$

Le probabilità di commettere questi due errori sono date, rispettivamente, da

$$P(T \notin \mathcal{R} \mid \mathrm{H}_0 \text{ } \textit{è vera}) = P^{\theta \in \Theta_0}(T \notin \mathcal{R}), \tag{7.3a}$$

$$P(T \in \mathcal{R} \mid \mathrm{H}_0 \text{ } \textit{è falsa}) = P^{\theta \notin \Theta_0}(T \in \mathcal{R}), \tag{7.3b}$$

Il *livello (di significatività)* α del test è per definizione la massima probabilità di commettere l'errore di prima specie:

$$\textit{livello} \;=\; \alpha = \sup_{\theta \in \Theta_0} P^{\theta}(T \in \mathcal{R}). \tag{7.4}$$

La probabilità complementare $1 - \alpha$ è detto *livello di protezione* del test e si ha

$$\textit{protezione} \;=\; 1 - \alpha = \inf_{\theta \in \Theta_0} \left\{1 - P^{\theta}(T \in \mathcal{R})\right\}. \tag{7.5}$$

La *potenza* del test è per definizione il complementare della probabilità β di commettere errore di seconda specie:

$$\textit{potenza} \;=\; 1 - \beta = 1 - P^{\theta \notin \Theta_0}(T \in \mathcal{R}), \tag{7.6}$$

ed è una funzione di θ.

7.2 Alcuni test per medie e varianze

7.2.1 Confronto fra media stimata e media teorica

Il test dell'Esempio 7.1 era un confronto fra una media incognita μ, stimata con $\overline{X}$, e una media "teorica" μ_0, con varianza σ^2 nota. In generale, i test di questo genere hanno ipotesi nulle del tipo

$$\mathrm{H}_0 : \mu \leq \mu_0 \quad (\text{o } \mu \geq \mu_0), \qquad \textit{test unilaterale}$$

oppure del tipo

$$\mathrm{H}_0 : \mu = \mu_0, \qquad \textit{test bilaterale},$$

dove, in entrambi i casi, μ_0 è la media "teorica" (cioè quella che ci aspettiamo che sia) e μ è la media stimata sul campione. Le regole di decisione, corrispondenti all'*accettazione* dell'ipotesi nulla, nei due casi sono:

$$\text{test unilaterale:} \quad \overline{X} \leq \mu_0 + c, \quad (\text{o } \overline{X} \geq \mu_0 - c); \quad \text{test bilaterale:} \quad \left|\overline{X} - \mu_0\right| \leq c,$$

Come abbiamo visto nell'esempio, fissato il livello α, il valore critico c corrisponde al δ di un IdC (rispettivamente unilaterale e bilaterale) per la media di

un campione normale con varianza nota. Dalle formule (6.12) e (6.11) si ottiene perciò

$$\text{test unilaterale:} \quad c = \frac{\sigma}{\sqrt{n}}\, q^{\mathcal{N}}_{1-\alpha}, \qquad \text{test bilaterale:} \quad c = \frac{\sigma}{\sqrt{n}}\, q^{\mathcal{N}}_{1-\alpha/2}. \tag{7.7}$$

Un test di questo tipo è anche detto *Test-Z* (in quanto coinvolge i quantili di una v.a. normale standard, che spesso è indicata con Z).

Vediamo adesso come si modificano le formule precedenti nel caso, molto più comune, in cui la varianza sia incognita. Le ipotesi nulle, le regole di decisione e le regioni di accettazione rimangono le stesse. Quello che cambia sono le formule per i valori critici. Infatti, essendo la varianza incognita, non si può usare il parametro sconosciuto σ^2 ma dobbiamo ricorrere al suo stimatore S^2. Dunque, fissato il livello α, tutto si svolge come nel caso di varianza nota con la differenza che stavolta i valori critici corripondono ai δ degli IdC per la media diun campione normale con varianza incognita e sono perciò dati dalle formule (6.14) e (6.13):

$$\text{test unilaterale:} \quad c = \frac{S}{\sqrt{n}}\, q^{t(n-1)}_{1-\alpha}, \qquad \text{test bilaterale:} \quad c = \frac{S}{\sqrt{n}}\, q^{t(n-1)}_{1-\alpha/2}. \tag{7.8}$$

Test di questo tipo prendono il nome di *Test-t*, in quanto coinvolgono la distribuzione t di Student.

7.2.2 Confronto fra medie di due campioni indipendenti

Una situazione molto frequente è quella in cui si vogliono confrontare le medie di due diversi campioni. Consideriamo ad esempio la Tabella 6.1: è lecito domandarsi se la differenza fra le medie di due campioni, raccolti in siti differenti è dovuta soltanto a una fluttuazione statistica o se c'è piuttosto una differenza "sostanziale" tra i due campioni.
Consideriamo quindi, in generale, due campioni normali indipendenti

$$\begin{aligned} \boldsymbol{X} &= (X_1, X_2, \ldots, X_{n_1}), \qquad X_i \sim \mathcal{N}(\mu_1, \sigma_1^2), \\ \boldsymbol{Y} &= (Y_1, Y_2, \ldots, Y_{n_2}), \qquad Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2), \end{aligned}$$

con medie incognite. Volendo mettere in luce le differenze fra le due medie, considereremo un'ipotesi nulla di tipo bilaterale:

$$\mathrm{H}_0 : \mu_1 = \mu_2.$$

Se invece ci aspettiamo che una media sia predominante sull'altra, potremmo testare un'ipotesi nulla di tipo unilaterale $\mathrm{H}_0 : \mu_1 < \mu_2$ (lasciamo al lettore il compito di adattare a questo caso le formule che troveremo). L'ipotesi nulla bilaterale è dunque quella che *non* ci siano differenze fra i due campioni. I due parametri μ_1 e μ_2 saranno stimati dalle due medie campionarie che indicheremo, rispettivamente con $\overline{X}$ e $\overline{Y}$. La differenza fra le due stime è data dalla v.a.

$$W = \overline{X} - \overline{Y}$$

ed ha senso utilizzare una regola di decisione del tipo

$$\textit{si rifiuta } \mathrm{H}_0 \textit{ se } |W| > c,$$

ovvero, si conclude che c'e' un'effettiva differenza se la distanza fra le due stime è sufficientemente grande, maggiore di un valore critico c da fissare in base al livello α.
Consideriamo innanzitutto il caso in cui le varianze delle due popolazioni, σ_1^2 e σ_2^2, siano note. Poiché

$$\overline{X} \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right), \qquad \overline{Y} \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right),$$

allora (si veda la proposizione 4.23)

$$W \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right) = \mathcal{N}\left(\mu_1 - \mu_2, \frac{n_2\sigma_1^2 + n_1\sigma_2^2}{n_1 n_2}\right).$$

In particolare, se assumiamo vera l'ipotesi H_0, si avrà che W ha media nulla. Dunque, fissato il livello α, la condizione

$$\alpha = P^{\mu_1=\mu_2}(|W| > c)$$

ci dice che c è il δ di un IdC bilaterale di livello α per la media di un campione normale con varianza nota $\frac{n_2\sigma_1^2+n_1\sigma_2^2}{n_1 n_2}$. Mediante la solita procedura di standardizzazione si ottiene perciò

$$c = \sqrt{\frac{n_2\sigma_1^2 + n_1\sigma_2^2}{n_1 n_2}}\; q^{\mathcal{N}}_{1-\alpha/2}. \tag{7.9}$$

Se le varianze non sono note le cose diventano più difficili. Se ci aspettiamo che le due varianze (incognite) siano all'incirca uguali,[2] allora un buono stimatore per il valore comune della varianza è la cosiddetta *varianza mediata*:

$$\overline{S}^2 := \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)} = \frac{\sum_{i=1}^{n_1}(X_i - \overline{X})^2 + \sum_{i=1}^{n_2}(X_i - \overline{Y})^2}{n_1 + n_2 - 2}, \tag{7.10}$$

dove con S_1^2 e S_2^2 abbiamo indicato le varianze campionarie dei due campioni. Poiché si può dimostrare che

$$\frac{W}{\overline{S}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t(n_1+n_2-2), \tag{7.11}$$

dalla condizione $\alpha = P^{\mu_1=\mu_2}(|W| > c)$ si ottiene la formula per il valore critico:

$$c = \overline{S}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\; q^{t(n_1+n_2-2)}_{1-\alpha/2}. \tag{7.12}$$

Consideriamo ad esempio un test di confronto fra le medie dei campioni *C1* e *C4* della tabella (6.1). Le stime di medie e varianze sono riportate nella tabella 6.2: si può notare che le varianze stimate sono simili, per cui si può pensare

[2] Questa è un'ulteriore ipotesi che può essere sottoposta al test-F, studiato nel paragrafo 7.2.3

di applicare il procedimento sopra descritto (verificheremo l'ipotesi di uguale varianza nel paragrafo 7.2.3). Si ha in questo caso

$$
\begin{aligned}
&n_1 = n_2 = 20;\\
&\overline{X} = 8.90; \quad \overline{Y} = 7.26; \quad W = 1.64;\\
&S_1^2 = 33.11; \quad S_2^2 = 38.91;\\
&\overline{S}^2 = \frac{19 \times (33.11 + 38.91)}{38} = 36.01\\
&c = \sqrt{\frac{1}{10}} \times \sqrt{36.01} \times q_{0.975}^{t(38)} \approx 1.90 \times q_{0.975}^{\mathcal{N}} = 1.90 \times 1.96 = 3.72
\end{aligned}
$$

(dove si è considerato il livello $\alpha = 0.05$ e il fatto che $t(n)$ si può approssimare con $\mathcal{N}$ per $n > 30$). Poiché risulta $|W| < c$, possiamo accettare l'ipotesi nulla $\mu_1 = \mu_2$ con un grado di certezza del 95%.

7.2.3 Confronto fra varianze di due campioni indipendenti

Talvolta è di un certo interesse confrontare le varianze di due diversi campioni. Una situazione di questio tipo l'abbiamo già incontrata nel precedente paragrafo 7.2.2, in cui era necessario verificare l'ipotesi che due campioni avessero la stessa varianza per poter utilizzare lo stimatore "varianza mediata" (7.10).
Consideriamo quindi, di nuovo, due campioni normali indipendenti

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_{n_1}), \qquad \boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_{n_2}),$$

con varianze incognite σ_1^2 e σ_2^2 (e medie incognite). L'ipotesi nulla adatta a mettere in luce differenze fra le due varianze è quella di tipo bilaterale:

$$\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2.$$

Anche qui, se ci aspettiamo che una varianza sia predominante sull'altra, potremmo testare un'ipotesi nulla di tipo unilaterale $\mathrm{H}_0 : \sigma_1^2 < \sigma_2^2$ (come al solito lasciamo al lettore lo svolgimento di questo caso). Come stimatori di σ_1^2 e σ_2^2 utilizziamo le varianze campionarie

$$S_1^2 := \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_i - \overline{X}\right)^2, \qquad S_2^2 := \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(Y_i - \overline{Y}\right)^2.$$

Per capire quale regola decisionale è più adatta a questo test utilizziamo il fatto che, *nell'ipotesi nulla*, si ha $\sigma_2^2/\sigma_1^2 = 1$ e quindi possiamo scrivere

$$\frac{S_1^2}{S_2^2} = \frac{n_2 - 1}{n_1 - 1} \frac{\frac{(n_1-1)}{\sigma_1^2} S_1^2}{\frac{(n_2-1)}{\sigma_2^2} S_2^2}.$$

Osserviamo che, per il Teorema di Cochran 6.13,

$$\frac{(n_1 - 1)}{\sigma_1^2} S_1^2 \sim \mathcal{X}^2(n_1 - 1), \qquad \frac{(n_2 - 1)}{\sigma_2^2} S_2^2 \sim \mathcal{X}^2(n_2 - 1),$$

e quindi, per quanto osservato nell'Esempio 4.19, si ha

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1}\,,$$

dove F_{n_1-1,n_2-1} è la distribuzione F di Fischer con gradi di libertà n_1-1 e n_2-1 (si veda il paragrafo 3.7). Conviene perciò utilizzare il rapporto S_1^2/S_2^2 per stimare la diversità delle due varianze e adottare di conseguenza la regola decisionale:

$$\textit{supponendo } S_1^2 \geq S_2^2, \textit{ si rifiuta } \mathrm{H}_0 \textit{ se } \frac{S_1^2}{S_2^2} > c,$$

dove, al solito, c è un valore critico da calcolare in funzione del livello del test α. Ovviamente, se S_2^2 fosse la varianza maggiore, si deve utilizzare $\frac{S_2^2}{S_1^2}$ e la distribuzione F_{n_2-1,n_1-1}. Fissato dunque α, per calcolare c scriveremo

$$P^{\sigma_1^2=\sigma_2^2}\left(\frac{S_1^2}{S_2^2} > c\right)$$

dove sappiamo che, nell'ipotesi $\sigma_1^2 = \sigma_2^2$, la v.a. S_1^2/S_2^2 ha distribuzione F_{n_1-1,n_2-1}. Si ha perciò direttamente

$$c = q_{1-\alpha}^{F_{n_1-1,n_2-1}}. \tag{7.13}$$

Riprendendo l'esempio del paragrafo 7.2.2, se vogliamo testare l'ipotesi $\sigma_1^2 = \sigma_2^2$ sui campioni *C1* e *C4* della tabella 6.1 dobbiamo utilizzare (per il livello $\alpha = 0.05$)

$$\begin{aligned} &n_1 = n_2 = 20; \\ &S_1^2 = 33.11; \quad S_2^2 = 38.91; \quad S_2^2 > S_1^2; \quad S_2^2/S_1^2 = 1.17 \\ &c = q_{0.95}^{F_{19,19}} \approx 2.1\,. \end{aligned}$$

Poiché $S_2^2/S_1^2 < c$, l'ipotesi nulla $\sigma_1^2 = \sigma_2^2$ non può essere respinta. Confrontando, invece, *C1* e *C2* si ha

$$S_1^2 = 33.11; \quad S_2^2 = 1.82; \quad S_1^2 > S_2^2; \quad S_1^2/S_2^2 = 18.19$$

per cui, in questo caso, l'ipotesi nulla può essere rifiutata.
Test di questo tipo, che coinvolgono la distribuzione di Fisher, sono chiamati *Test-F*.

7.3 Test del $\mathcal{X}^2$

Il cosiddetto *test del* $\mathcal{X}^2$ è un utilissimo strumento per confrontare frequenze teoriche e frequenze misurate e verificare così l'aderenza alla realtà di un modello probabilistico con distribuzione discreta. Ad esempio, lanciando la moneta il nostro tipico modello probabilistico è basato sulla v.a. Bernoulliana

$$P(X=1) = 1/2 \ \text{(“testa”)}, \qquad P(X=0) = 1/2 \ \text{(“croce”)}.$$

Se volessimo testare statisticamente questo nostro modello, dovremmo lanciare tante volte la moneta, ottenendo una sequenza $X_1, X_2, \ldots X_n$ di v.a. Bernoulliane, e confrontare le frequenze relative sperimentali

$$F(\textit{testa}) = \frac{\textit{numero teste}}{\textit{numero lanci}} = \frac{X_1 + X_2 + \cdots + X_n}{n};$$

$$F(\textit{croce}) = \frac{\textit{numero croci}}{\textit{numero lanci}} = 1 - \frac{X_1 + X_2 + \cdots + X_n}{n};$$

con le frequenze relative teoriche che, come ci si aspetta dalla legge dei grandi numeri (Teorema 4.26), sono date da

$$F(testa) \approx 1/2; \qquad F(croce) \approx 1/2;$$

se il numero di prove è sufficientemente alto.

Consideriamo quindi in generale un campione di rango n, $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, con distribuzione discreta. Questo significa che ciascuna X_i assume m possibili valori $x_1, x_2, \ldots, x_m$ con probabilità $p_1, p_2, \ldots, p_m$:

$$P(X_i = x_k) = p_k, \qquad \text{per } \ k = 1, 2, , \ldots, m \ \text{ e } \ i = 1, 2, , \ldots, n,$$

con la condizione $\sum_{k=1}^{m} p_k = 1$. Nel contesto del test del $\mathcal{X}^2$ le probabilità p_k sono i *parametri incogniti* della distribuzione. Consideriamo le m variabili aleatorie

$$N_k = \textit{numero delle } X_i \textit{ uguali a } x_k, \qquad k = 1, 2, , \ldots, m, \tag{7.14}$$

dette *frequenze campionarie* e le m variabili aleatorie

$$F_k = \frac{N_k}{n}, \qquad k = 1, 2, , \ldots, m, \tag{7.15}$$

dette *frequenze campionarie relative* (sono esattamente le frequenze assolute e relative introdotte nel paragrafo 5.2 ma stavolta viste come variabili aleatorie). Si può verificare molto facilmente che F_k *è uno stimatore corretto di* p_k.

Osservazione 7.2. Notiamo per inciso che F_k non solo è uno stimatore corretto di p_k ma per di più tende a p_k per $n \to \infty$. Questo lo si può vedere scrivendo

$$N_k = Y_1^{(k)} + Y_2^{(k)} + \cdots + Y_n^{(k)}$$

dove, per ogni $i = 1, 2, , \ldots, n$, si definisce

$$Y_i^{(k)} := \begin{cases} 1, & \text{se } X_i = x_k; \\ 0, & \text{se } X_i \neq x_k. \end{cases}$$

Poiché $E[Y_i^{(k)}] = p_k$, dalla legge dei grandi numeri (Teorema 4.26) si ottiene che $F_k \to p_k$ quando $n \to \infty$. □

Consideriamo ora la v.a.

$$T_n := n \sum_{k=1}^{m} \frac{(F_k - p_k)^2}{p_k} = \sum_{k=1}^{m} \frac{(N_k - np_k)^2}{np_k}, \tag{7.16}$$

che dà una misura dello scarto quadratico delle frequenze teoriche da quelle campionarie. L'indice n serve a ricordare la numerosità del campione. Si ha il seguente risultato, che enunciamo senza dismostrazione.

Teorema 7.3. (Teorema di Pearson)
Se $P(X_i = x_k) = p_k$*, la successione di variabili aleatorie* $T_1, T_2, T_3, \ldots$ *definite dalla* (7.16) *tende ad avere distribuzione* $\mathcal{X}^2(m-1)$.[3] *Sinteticamente:*

$$\lim_{n\to\infty} T_n \sim \mathcal{X}^2(m-1). \tag{7.17}$$

[3] Il fatto che i gradi di libertà siano $m-1$ è dovuto al vincolo $\sum_{k=0}^{m} p_k = 1$.

Questo risultato ci permette di concepire un test dell'ipotesi nulla

$$\mathrm{H}_0 \,:\, p_1 = p_1^0,\ p_2 = p_2^0, \ldots,\ p_m = p_m^0,$$

che esprime il fatto che ci aspettiamo certe probabilità discrete (abbiamo indicato con $p_1^0, p_2^0, \ldots, p_m^0$ i valori che ci aspettiamo per $p_1, p_2, \ldots, p_m$). Poiché la v.a. T_n ci fornisce una misura della distanza fra probabilità teoriche e probabilità misurate, possiamo adottare la regola decisionale

$$\textit{si accetta } \mathrm{H}_0 \textit{ se } T_n \leq c.$$

Per stabilire il valore critico c si osserva che, se H_0 è vera, la v.a. T_n calcolata usando i nostri p_k^0 ha approssimativamente distribuzione $\mathcal{X}^2(m-1)$. Ovviamente, questo è vero *solo se* n *è abbastanza grande* (generalmente si accetta $n \geq 30$). Dunque, fissato il livello α, la condizione

$$\alpha = P^{p_1=p_1^0,\ldots,\ p_m=p_m^0}(T_n > c)$$

implica (approssimativamente)

$$c = q_{1-\alpha}^{\mathcal{X}^2(m-1)}. \tag{7.18}$$

Ricapitolando: il cosiddetto *Test del* $\mathcal{X}^2$ consiste nel calcolare l'espressione

$$T_n = \sum_{k=1}^{m} \frac{(N_k - np_k^0)^2}{np_k^0}$$

(o l'equivalente in termini delle frequenze relative F_k) e confrontarne il valore con il numero $c = q_{1-\alpha}^{\mathcal{X}^2(m-1)}$. Se T_n è minore di c si può concludere che le nostre probabilità teoriche $p_1^0, p_2^0, \ldots, p_m^0$ sono corrette.

Esempio 7.4. Non ci possiamo sottarre dal fare l'esempio più classico di utilizzo del Test del $\mathcal{X}^2$: il famoso esperimento di Mendel. Mendel osservò l'occorrenza di quattro possibili coppie, c_1, c_2, c_3, c_4, di caratteri genetici in semi di pisello. Tali coppie sono definite nella seguente tabella:

		aspetto:	
		LISCIO $_{(D)}$	GRINZOSO
colore:	GIALLO $_{(D)}$	c_1	c_3
	VERDE	c_2	c_4

dove il pedice (D) sta ad indicare il carattere dominante (l'altro è detto recessivo). In base alle osservazioni, Mendel enunciò le sue famose leggi della genetica che qui riassumiamo in una forma adatta ai nostri scopi:

(i) il rapporto dominante/recessivo è di 3 a 1;

(ii) i due caratteri genetici (aspetto e colore) sono indipendenti.

La prima legge si può tradurre, da un punto di vista probabilistico, nel modo seguente:

$$P(\textit{dominante}) = 3/4, \qquad P(\textit{recessivo}) = 1/4.$$

La seconda legge ci aiuta a calcolare, analogamente a quanto fatto nel paragrafo 1.6, la probabilità di occorrenza delle quattro coppie:

$$P(c_1) = \frac{9}{16}, \quad P(c_2) = \frac{3}{16}, \quad P(c_3) = \frac{3}{16}, \quad P(c_4) = \frac{1}{16}.$$

Vogliamo perciò testare la validità della seguente ipotesi:

$$\mathrm{H}_0 \ : \ p_1 = \frac{9}{16}, \quad p_2 = p_3 = \frac{3}{16}, \quad p_4 = \frac{1}{16}$$

basandoci sulle osservazioni dello stesso Mendel:

$$n = 556, \quad N_1 = 315, \quad N_2 = 101, \quad N_3 = 108, \quad N_4 = 32.$$

Calcoliamo perciò T_n usando le nostre probabilità teoriche:

$$T_n = \frac{(315 - 556 \times 9/16)^2}{556 \times 9/16} + \frac{(101 - 556 \times 3/16)^2}{556 \times 3/16}$$
$$+ \frac{(108 - 556 \times 3/16)^2}{556 \times 3/16} + \frac{(32 - 556 \times 1/16)^2}{556 \times 1/16} \approx 0.47\,.$$

In base alla (7.18), se scegliamo $\alpha = 0.05$, dobbiamo confrontare il valore trovato per T_n con

$$q_{1-\alpha}^{\mathcal{X}^2(m-1)} = q_{0.95}^{\mathcal{X}^2(3)} \approx 7.81,$$

da cui risulta chiaramente che H_0 può essere accettata. □

Il test del $\mathcal{X}^2$ è molto utile anche per testare l'appartenenza di un campione a una certa distribuzione teorica (anche continua). L'idea è quella di fissare un certo numero m di intervalli, $I_1, I_2, \ldots, I_m$, e di calcolare (in base alla distribuzione ipotizzata $\mathcal{D}$) le probabilità teoriche $p_1, p_2, \ldots, p_m$ che il campione ha di cadere in quegli intervalli. Così possiamo saggiare l'ipotesi

$$\mathrm{H}_0 : X_i \sim \mathcal{D}$$

con un test del $\mathcal{X}^2$ sulle probabilità discrete $P(X_i \in I_k) = p_k$. Tuttavia il problema può essere un po' più complicato se $\mathcal{D}$ dipende da parametri stimati col campione stesso (che poi è il caso più frequente). Se ad esempio vogliamo testare la normalità di un campione di cui ignoriamo media e varianza, per calcolare le probabilità teoriche $p_k = P(X_i \in I_k)$ dobbiamo utilizzare una distribuzione normale in cui μ e σ^2 sono stimati da $\overline{X}$ e S^2. In questi casi ci viene in aiuto la seguente versione raffinata del Teorema di Pearson.

Teorema 7.5. (Teorema di Pearson migliorato)
Se le probabilità teoriche $p_k = P(X_i = x_k)$ sono calcolate usando ℓ stimatori, allora

$$\lim_{n\to\infty} T_n \sim \mathcal{X}^2(m - \ell - 1). \tag{7.19}$$

Esempio 7.6.: Test di normalità. In base alle considerazioni sopra esposte, si può costruire un *test di normalità* del campione $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, ovvero un test dell'ipotesi

$$\mathrm{H}_0 : X_i \sim \mathcal{N}(\mu, \sigma^2),$$

usando il test $\mathcal{X}^2$ nel modo seguente:

1. stimo media μ e varianza σ^2 sul campione utilizzando $\overline{X}$ e S^2;
2. standardizzo il campione ponendo $Z_i = \frac{X_i - \mu}{\sigma}$;
3. suddivido la retta reale in m intervalli $I_1, I_2, \ldots, I_m$ (meglio se simmetrici rispetto all'origine);
4. utilizzando la distribuzione $\mathcal{N}$, calcolo le probabilità teoriche, $p_k^0 = P(Z_i \in I_k)$, di appartenenza agli intervalli (figura 7.2);
5. conteggio le frequenze campionarie: $N_k = \textit{numero di } Z_i \textit{ che cadono in } I_k$;
6. eseguo il test del $\mathcal{X}^2$.

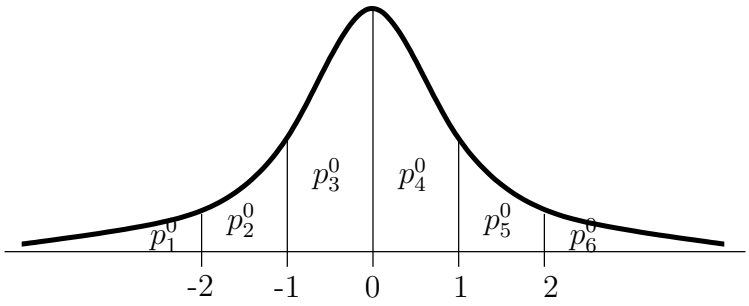


Figura 7.2: Suddivisione della retta reale in m intervalli. In questo esempio si è scelto $m = 6$, $I_1 = (-\infty, -2]$, $I_2 = [-2, -1]$, $I_3 = [-1, 0]$, $I_4 = [0, 1]$, $I_5 = [1, 2]$, $I_6 = [2, +\infty)$ e le corrispondenti probabilità discretizzate sono $p_1^0 = p_6^0 = 0.023$, $p_2^0 = p_4^0 = 0.136$, $p_3^0 = p_4^0 = 0.341$.

Sottoponiamo a questo test i campioni della concentrazione di cromo nelle piante di *Alyssum*, di cui ci siamo già occupati nell'esempio 6.17. Suddividendo la retta reale in $m = 6$ intervalli come in figura 7.2 e standardizzando i nove campioni della tabella 6.1, si trovano le frequenze campionarie e teoriche riportate nella tabella 7.1. Se calcoliamo T_{20} nei nove casi otteniamo i risultati riportati nell'ultima riga della tabella. Scegliendo $\alpha = 0.05$, si ottiene un valore critico del test $\mathcal{X}^2$ pari a

$$q_{1-\alpha}^{\mathcal{X}^2(m-\ell-1)} = q_{0.95}^{\mathcal{X}^2(3)} \approx 7.81$$

che dunque ci porta a non rifiutare l'ipotesi di normalità solo per i campioni 4 e 6. Può essere interessante visualizzare la situazione mediante istogrammi: nella figura 7.3 sono riportati in istogramma le frequenze relative dei campioni (standardizzati) *C6* e *C9* che hanno, rispettivamente, il valore di T_{20} più basso e più alto. Confrontando con la distribuzione normale standard riportata nei due grafici, si nota come, in effetti, *C6* è abbastanza compatibile con la distribuzione normale mentre il campione *C9* si discosta notevolmente da essa. □

	C1	*C2*	*C3*	*C4*	*C5*	*C6*	*C7*	*C8*	*C9*	np_k^0
N_1	0	0	0	0	0	0	0	0	0	0.46
N_2	3	0	0	2	1	2	0	0	0	2.72
N_3	8	13	12	10	13	9	14	13	16	6.82
N_4	7	4	5	5	3	6	5	4	2	6.82
N_5	0	1	2	2	2	2	0	1	1	2.72
N_6	2	2	1	1	1	1	1	2	1	0.46
T_{20}	8.6	16.2	8.4	3.4	10.1	2.3	14.6	16.2	20.7	

Tabella 7.1: Frequenze campionarie N_k e teoriche np_k^0 relative ai campione della tabella 6.1 e alla discretizzazione della figura 7.2. Nell'ultima riga sono riportati i corrispondenti valori dello stimatore T_{20}.

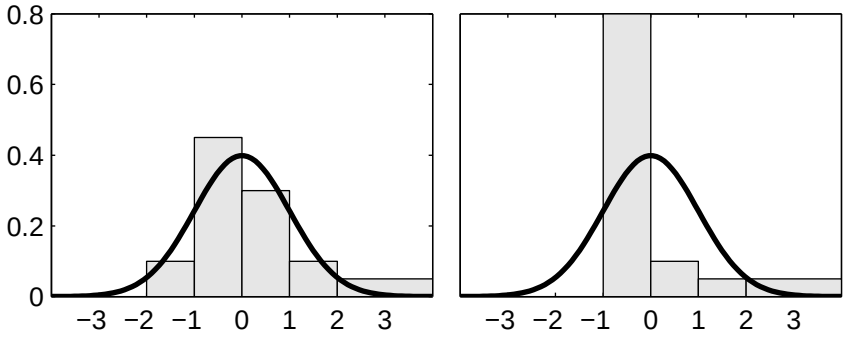


Figura 7.3: Confronto fra le frequenze relative campionarie (istogramma) e la distribuzione normale standard (linea continua). Il due grafici si riferiscono, rispettivamente, ai campioni *C6* e *C9* della tabella 7.1.

Capitolo 8

Analisi della regressione

8.1 Regressione lineare semplice

Supponiamo di effettuare la misura di una coppia di variabili (x, y) su n individui, ottenendo così n coppie di valori

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots \quad (x_n, y_n).$$

Supponiamo anche di aver buoni motivi (magari proprio perché le misurazioni stesse ce lo suggeriscono) per sospettare che la variabile y dipenda dalla variabile x secondo una legge lineare del tipo

$$y = \beta_0 + \beta_1 x. \tag{8.1}$$

Non ci possiamo ovviamente aspettare che le coppie misurate (x_i, y_i) seguano alla perfezione la legge (8.1), cioè che valga esattamente $y_i = \beta_0 + \beta_1 x_i$, ma piuttosto che questa legga sia "nascosta" da un "rumore statistico" che, nell'ipotesi più semplice, si manifesta come un termine aleatorio W_i da aggiungere alla legge (8.1). Il modello statistico di questa situazione è il seguente:

$$y_i = \beta_0 + \beta_1 x_i + W_i, \qquad i = 1, 2, \dots, n, \tag{8.2}$$

dove

- $(W_1, W_2, \dots, W_n)$ è un campione statistico di rango n con distribuzione nota;
- $(x_1, x_2, \dots, x_n)$ sono n numeri reali assegnati;
- β_0 e β_1 sono due parametri incogniti;
- $(y_1, y_2, \dots, y_n)$ sono n variabili aleatorie definite dalla (8.2).

La variabile indipendente x in questo contesto è detta *predittore* mentre β_0 e β_1 sono detti *parametri della regressione*. Possiamo osservare che :

1. le misure $x_1, x_2, \dots, x_n$ della variabile x sono trattate come deterministiche (cioè non come variabili aleatorie) in quanto ci interessa fare un'analisi statistica solamente della *dipendenza* della quantità y dalla quantità x;

2. $(y_1, y_2, \ldots, y_n)$ *non è* un campione statistico, in quanto le y_i sono indipendenti ma non sono, in generale, identicamente distribuite.

Osservazione 8.1. Il modello appena esposto si chiama *modello di regressione lineare semplice*. Il termine "lineare" non si riferisce, come si potrebbe pensare, alla dipendenza di y da x, bensì alla dipendenza lineare di y da W. □

Come abbiamo detto, le W_i sono un campione statistico con distribuzione nota; in particolare noi supporremo che tale distribuzione sia normale con media nulla:

$$W_i \sim \mathcal{N}(0, \sigma^2), \qquad i = 1, 2, \ldots, n, \tag{8.3}$$

dove la varianza σ^2 è un ulteriore parametro incognito del modello. Notiamo che le y_i hanno anch'esse ditribuzione normale e precisamente

$$y_i \sim \mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2), \qquad i = 1, 2, \ldots, n. \tag{8.4}$$

Lo scopo dell'analisi del modello di regressione è stimare i parametri incogniti β_0, β_1 e σ^2.

Scegliamo come stimatori di β_0 e β_1 le quantità b_0 e b_1 che minimizzano gli scarti quadratici fra l'andamento stimato e quello misurato. Dunque, b_0 e b_1 vanno determinati in modo da minimizzare la funzione

$$f(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2, \tag{8.5}$$

il che, da un punto di vista geometrico, equivale a scegliere la retta $y = b_0 + b_1 x$ che rende minima la somma dei quadrati delle distanze in ordinata r_i, dette *residui*, fra i punti misurati e la retta stessa (figura 8.1).[1]

I minimi della funzione $f(b_0, b_1)$ vanno cercati fra i punti (b_0, b_1) in cui il gradiente della f si annulla (si veda [7]), ovvero in cui si ha

$$\begin{cases} \dfrac{\partial f}{\partial b_0} = -2 \displaystyle\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i) = 0, \\ \dfrac{\partial f}{\partial b_1} = -2 \displaystyle\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0. \end{cases} \tag{8.6}$$

Ricaviamo b_0 dalla prima equazione del sistema: si ha $n\, b_0 = \sum_{i=1}^{n} (y_i - b_1 x_i)$, ovvero

$$b_0 = \overline{y} - b_1 \overline{x}, \tag{8.7}$$

dove[2]

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \overline{y} := \frac{1}{n} \sum_{i=1}^{n} y_i.$$

[1] Attenzione a non confondere i r_i, che sono le distanze dalla retta stimata $y = b_0 + b_1 x$, con i W_i, che sono le distanze dalla retta "vera" ma incognita $y = \beta_0 + \beta_1 x$.

[2] Notiamo che, pur riguardando le y_i come variabili aleatorie, $\overline{y}$ non è una media campionaria in senso stretto in quanto le v.a. y_i non sono identicamente distribuite.

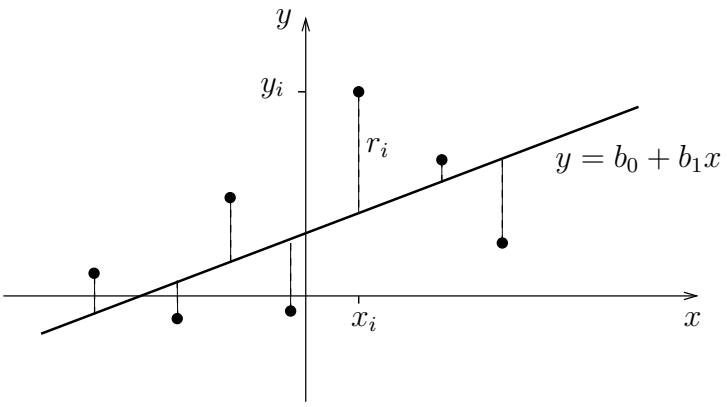


Figura 8.1: Gli stimatori b_0 e b_1 sono quelli per cui la retta di equazione $y = b_0 + b_1 x$ rende minima la somma dei quadrati delle distanze $r_i = y_i - b_0 - b_1 x_i$ tra le ordinate dei punti misurati (dischetti neri) e la retta stessa.

Sostituiamo nella seconda equazione l'espressione trovata per b_0:

$$0 = \overline{x} := \sum_{i=1}^{n} x_i(y_i - \overline{y} + b_1\overline{x} - b_1 x_i) = \sum_{i=1}^{n} x_i(y_i - \overline{y}) + b_1 \sum_{i=1}^{n} x_i(\overline{x} - x_i),$$

per cui

$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \tag{8.8}$$

dove si è posto

$$\sigma_{xy} := \sum_{i=1}^{n} x_i(y_i - \overline{y}), \qquad \sigma_x^2 := \sum_{i=1}^{n} x_i(x_i - \overline{x}), \tag{8.9}$$

Sfruttando il fatto che $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$ e $\sum_{i=1}^{n}(y_i - \overline{y}) = 0$ possiamo scrivere le seguenti espressioni alternative per σ_{xy} e σ_x^2:

$$\begin{aligned} \sigma_{xy} &= \sum_{i=1}^{n} x_i(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i - \overline{x})y_i\,, \\ \sigma_x^2 &= \sum_{i=1}^{n} x_i(x_i - \overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})^2. \end{aligned} \tag{8.10}$$

Riguardando y_i come la variabile aleatoria data dalla (8.2), le quantità b_0 e b_1 date da (8.7) e (8.8) sono due variabili aleatorie che useremo come stimatori dei parametri della regressione, β_0 e β_1.

8.2 Intervalli di confidenza per β_0, β_1 e σ^2

Nel paragrafo precedente abbiamo trovato gli stimatori

$$b_1 = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad b_0 = \overline{y} - b_1\overline{x}.$$

dei parametri della regressione β_1 e β_0. Vogliamo ora trovare valore atteso e varianza dei due stimatori. Per quanto riguarda il valore atteso di b_1, utilizzando la proprietà di linearità della media, si ha

$$E[b_1] = \frac{E\big[\sum_{i=1}^n x_i(y_i - \overline{y})\big]}{\sigma_x^2} = \frac{\sum_{i=1}^n x_i E[y_i - \overline{y}]}{\sigma_x^2}$$

$$= \frac{\beta_1 \sum_{i=1}^n x_i(x_i - \overline{x})}{\sigma_x^2} = \frac{\beta_1 \sum_{i=1}^n (x_i - \overline{x})^2}{\sigma_x^2} = \beta_1$$

dove si è utilizzato

$$E[y_i - \overline{y}] = \beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \overline{x}) = \beta_1(x_i - \overline{x}).$$

Il valore atteso di b_0 segue semplicemente:

$$E[b_0] = E[\overline{y}] - \overline{x}E[b_1] = \beta_0 + \beta_1\overline{x} - \overline{x}\beta_1 = \beta_0.$$

Abbiamo così trovato che b_1 e b_0 sono stimatori corretti. Per calcolarne le varianze conviene introdurre le quantità (non aeleatorie)

$$v_i := \frac{x_i - \overline{x}}{\sigma_x^2},$$

per le quali notiamo che vale $\sum_{i=1}^n v_i = 0$. Possiamo perciò scrivere, ricordando le note proprietà della varianza,

$$\mathrm{Var}[b_1] = \mathrm{Var}\Big[\sum_{i=1}^n v_i(y_i - \overline{y})\Big] = \mathrm{Var}\Big[\sum_{i=1}^n v_i y_i\Big] = \sum_{i=1}^n v_i^2 \,\mathrm{Var}[y_i]$$

$$= \sigma^2 \sum_{i=1}^n v_i^2 = \sigma^2 \frac{\sum_{i=1}^n (x_i - \overline{x})^2}{\sigma_x^4} = \frac{\sigma^2}{\sigma_x^2},$$

dove si è usato il fatto che, per ipotesi, le y_i hanno tutte le stessa varianza σ^2. Infine, per calcolare $\mathrm{Var}[b_0]$, calcoliamo dapprima la covarianza tra $\overline{y}$ e b_1. Sfruttando le note proprietà di bilinearità della covarianza (4.16) si ottiene

$$\mathrm{Cov}[\overline{y}, b_1] = \sum_{i=1}^n v_i \mathrm{Cov}[\overline{y}, y_i] = \sum_{i=1}^n \sum_{j=1}^n \frac{v_i}{n} \mathrm{Cov}[y_j, y_i],$$

ma poiché le y_i sono indipendenti

$$\sum_{i=1}^n \sum_{j=1}^n \frac{v_i}{n} \mathrm{Cov}[y_j, y_i] = \sum_{i=1}^n \frac{v_i}{n} \mathrm{Var}[y_i] = \frac{\sigma^2}{n} \sum_{i=1}^n v_i = 0.$$

Dunque $\overline{y}$ e b_1 sono incorrelate. Si ha perciò

$$\mathrm{Var}[b_0] = \mathrm{Var}[\overline{y}] + \overline{x}^2 \mathrm{Var}[b_1] - 2\overline{x}\mathrm{Cov}[\overline{y}, b_1] = \frac{\sigma^2}{n} + \frac{\sigma^2 \overline{x}^2}{\sigma_x^2}.$$

Osserviamo che le v.a. b_1 e b_0 sono combinazioni lineari di variabili normali indipendenti: questo lo si vede, per quanto riguarda b_1, dalla terza delle espressioni alternative per σ_{xy} nella (8.10) e, per quanto riguarda b_0, dal fatto che

$\mathrm{Cov}\,[\overline{y}, b_1] = 0$, il che per v.a. normali implica l'indipendenza (Osservazione 4.17). Pertanto, come sappiamo dalla proposizione 4.23, b_1 e b_0 sono ancora v.a. normali e precisamente, avendone già ricavato medie e varianze,

$$b_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sigma_x^2}\right), \qquad b_0 \sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2\overline{x}^2}{\sigma_x^2}\right).$$

Ricordiamo che σ^2 è un parametro incognito e che dunque va stimato. Notiamo però che la varianza campionaria

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} W_i^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

è inutilizzabile perché non conosciamo β_0 e β_1. L'idea è ovviamente quella di utilizzare le stime b_1 e b_0. Per $i = 1, 2, \ldots, n$ poniamo

$$y_i^s := b_0 + b_1 x_i \qquad \text{(valori stimati)} \tag{8.11}$$

e

$$r_i := y_i - y_i^s = y_i - b_0 - b_1 x_i \qquad \text{(residui)}. \tag{8.12}$$

Risulta allora che la v.a.

$$s^2 := \frac{1}{n-2}\sum_{i=1}^{n} r_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - y_i^s)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \tag{8.13}$$

è uno stimatore *corretto* di σ^2. Notiamo che

$$\sum_{i=1}^{n} r_i = 0, \qquad \sum_{i=1}^{n} x_i r_i = 0$$

(corrispondenti esattamente alle due condizioni $\partial f/\partial b_0 = 0$ e $\partial f/\partial b_1 = 0$ viste nel paragrafo precedente), per cui i residui r_i sono legati da *due* vincoli lineari. Ci aspettiamo perciò una distribuzione del tipo $\mathcal{X}^2(n-2)$. Più precisamente, si possono dimostrare (vedi [1]) le seguenti proprietà:

(i) s^2 è corretto;

(ii) $\frac{n-2}{\sigma^2}\, s^2 \sim \mathcal{X}^2(n-2)$;

(iii) s^2 è indipendente da b_0 e b_1;

(iv) si ha

$$\frac{b_0 - \beta_0}{s\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sigma_x^2}}} \sim t(n-2), \qquad \frac{\sigma_x\sqrt{n}(b_1 - \beta_1)}{s} \sim t(n-2);$$

(v) i residui r_i e i valori stimati y_i^s sono indipendenti.

Pertanto, se vogliamo ricavare intervalli di confidenza di livello α per b_0 e b_1 procederemo in maniera analoga a quanto fatto nel paragrafo 6.4. Ad esempio i δ degli IdC bilaterali di livello α per la media di b_0 e b_1 sono rispettivamente dati da

$$\delta_0 = s\, q_{1-\alpha/2}^{t(n-2)} \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{\sigma_x^2}} \qquad \text{e} \qquad \delta_1 = \frac{s}{\sqrt{n}\,\sigma_x}\, q_{1-\alpha/2}^{t(n-2)}. \tag{8.14}$$

A questo modo si possono implementare opportuni *test* per i valori di β_0 e β_1.

8.3 Discussione del modello ed esempi

È la legge $y_i = \beta_0 + \beta_1 x_i + W_i$ un buon modello? Un utile indicatore è la quantità

$$R^2 := \frac{\sum_{i=0}^{n}(y_i^s - \overline{y})^2}{\sum_{i=0}^{n}(y_i - \overline{y})^2}, \tag{8.15}$$

dove, ricordiamo, $y_i^s = b_0 + b_1 x_i$ sono i cosiddetti valori stimati. Notiamo che R^2 è il rapporto fra i quadrati degli scarti stimati e i quadrati degli scarti misurati. Si può dimostrare (vedi [1]) che si ha sempre $0 \leq R^2 \leq 1$. Notiamo che R^2 può essere interpretata come *la proporzione di variazione di y (rispetto alla media) spiegata dal modello* e, dunque, si può considerare il modello di regressione lineare tanto più buono quanto più R^2 si avvicina a 1.

A

x	y
5.68	11.3
7.39	15.3
6.06	13.1
6.90	15.4
5.53	12.4
7.44	14.4
6.32	13.4
8.01	15.0
7.99	15.7
7.08	14.3
6.74	13.8
8.15	15.7
7.88	16.4
7.26	13.1
7.86	15.9
7.31	15.4
6.20	13.5
6.01	13.1
6.19	13.0
6.87	14.5
7.54	15.5
6.08	12.5
7.93	15.4
6.99	13.9
6.30	11.9

B

x	y
-0.90	1.53
1.78	4.21
2.50	6.52
-0.61	1.37
1.18	-0.12
-0.54	1.50
0.08	0.26
-0.53	-0.81
-0.22	-0.74
0.39	0.59
-0.53	0.74
-0.16	-0.58
2.25	5.42
0.37	1.59
0.79	0.54
-0.67	0.79
-0.92	1.87
-0.44	1.23
1.96	4.57
2.08	3.62
-0.34	-1.02
2.47	7.32
1.49	2.07
2.05	5.35
0.68	-0.42

Tabella 8.1: Due insiemi di dati sottoposti all'analisi di regressione.

Un'altra utile indicazione può venire dal fatto che, nel modello di regressione lineare, i residui $r_i = y_i - y_i^s$ e valori stimati y_i^s risultano essere v.a. indipendenti (proprietà (v) del precedente paragrafo). Perciò, tracciando un grafico dei punti (y_i^s, r_i), si dovrebbe osservare una distribuzione completamente casuale di tali punti. Se si riscontra un qualche allineamento o una qualche dipendenza, si può cominciare a sospettare che il modello non sia adeguato. Questa procedura è detta "analisi dei residui".

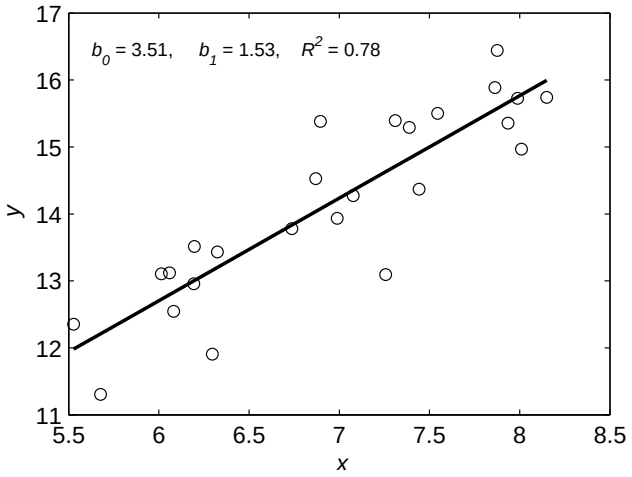


Figura 8.2: Analisi di regressione dei dati della tabella 8.1A. I cerchietti rappresentano i dati sperimentali e la retta di regressione è determinata dalle stime b_0 e b_1 riportate in alto, dove viene indicato anche il valore di R^2.

Se entrambi i due indicatori (calcolo di R^2 e analisi dei residui) depongono a sfavore del modello di regressione lineare si possono ipotizzare modelli più complicati di dipendenza di y da x, come ad esempio il modello polinomiale

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p$$

di cui daremo un breve cenno nel paragrafo seguente (ma, ovviamente, molti altri modelli sono possibili: logaritmico, esponenziale, ecc.).

Consideriamo a titolo di esempio i due insiemi di dati riportati nelle tabelle 8.1 A e B. I risultati dell'analisi dei dati delle due tabelle, effettuata col modello della regressione lineare semplice, sono riportati, rispettivamente, nelle figure 8.2 e 8.3. Nelle due figure i cerchietti indicano la posizione dei dati sperimentali (coppie (x_i, y_i)) e viene tracciata la retta di regressione $y = b_0 + b_1 x$. I valori dei parametri b_0 e b_1, calcolati usando le formule (8.7) e (8.8), sono riportati in alto a sinistra. Come si può vedere dalle figure, il modello di regressione lineare sembra più adatto ai dati A che ai dati B. Questa impressione è confermata sia dai valori di R^2, riportati nelle figure 8.2 e 8.3, sia dall'analisi dei residui che si può effettuare osservando le figure 8.4 e 8.5. In queste si vede la distribuzione dei valori (y_i^s, r_i) calcolati sui due campioni. Come sappiamo, se la regressione lineare è un buon modello, i valori stimati y_i^s e i residui $r_i = y_i - y_i^s$ hanno da essere quantità indipendenti e quindi non si dovrebbe poter notare nessun andamento regolare nella distribuzione dei punti. Se ciò è senz'altro vero per i punti relativi al campione A (figura 8.4), lo stesso non si può dire per i punti relativi al campione B (figura 8.5), dove si osserva una sospetta tendenza a disporsi a "V". Siamo perciò portati ad adottare un diverso modello di regressione per i dati della tabella B, cosa che faremo nel prossimo paragrafo dove questo esempio sarà ripreso.

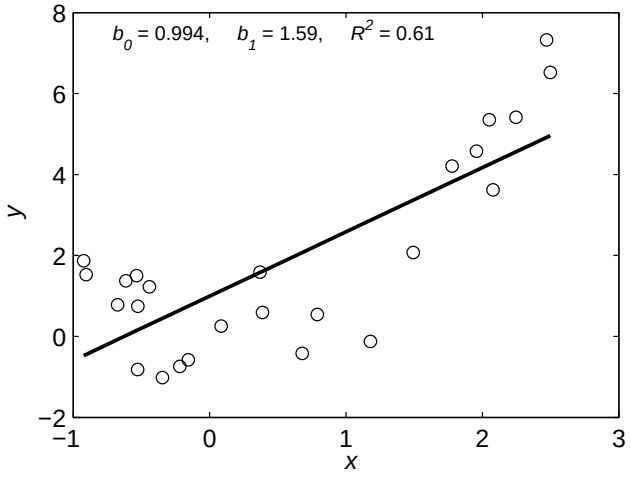


Figura 8.3: Analisi di regressione dei dati della tabella 8.1B (vedi analoga figura 8.2).

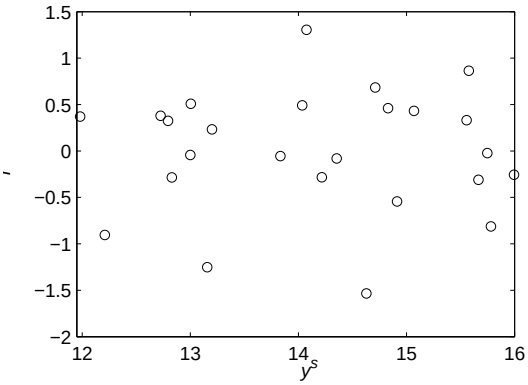


Figura 8.4: Analisi dei residui per i dati della tabella 8.1A.

8.4 Cenni sulla regressione lineare multipla

La regressione lineare multipla è un modello in cui ci sono m predittori (cioè m variabili indipendenti):

$$y = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_m x^{(m)}, \tag{8.16}$$

dove $x^{(1)}, x^{(2)}, \ldots x^{(m)}$ sono m quantità diverse, ciascuna misurata n volte:

$$x_1^{(k)}, x_2^{(k)}, \ldots x_n^{(k)}, \qquad k = 1, 2, \ldots, m.$$

Dunque, scrivendo $x_i^{(k)}$, l'indice in alto si riferisce alla variabile e quello in basso alla misurazione. Il corrispondente modello statistico è il seguente:

$$y_i = \beta_0 + \beta_1 x^{(1)} + \beta_2 x^{(2)} + \cdots + \beta_m x^{(m)} + W_i, \quad i = 1, 2, \ldots, n, \tag{8.17}$$

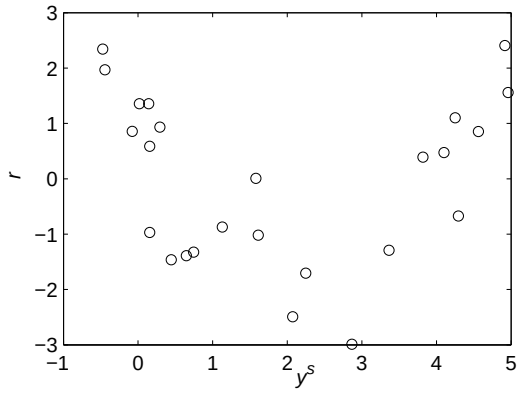


Figura 8.5: Analisi dei residui per i dati della tabella 8.1B.

dove le W_i, come nel caso della regressione semplice, sono v.a. che supporremo normali, indipendenti, identicamente distribuite, con media nulla e con varianza σ^2 incognita:

$$W_i \sim \mathcal{N}(0, \sigma^2).$$

I parametri della regressione $\beta_0, \beta_1, \cdots, \beta_m$ sono incogniti e si possono ricavare degli stimatori $b_0, b_1, \cdots, b_m$ col metodo dei minimi quadrati. Per esporre in modo semplice il risultato che si ottiene, conviene ricorrere al linguaggio dei vettori e delle matrici. Definiamo le medie degli m predittori e della variabile dipendente:

$$\overline{x}^{(k)} := \frac{1}{n}\sum_{i=1}^{n} x_i^{(k)}, \qquad k = 1, 2, \ldots, m,$$

$$\overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i.$$

Definiamo poi i vettori-colonna

$$\boldsymbol{y} = \begin{pmatrix} y_1 - \overline{y} \\ y_2 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{pmatrix} \in \mathbb{R}^{n,1}, \qquad \boldsymbol{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix} \in \mathbb{R}^{m,1}$$

e la matrice

$$X = \begin{pmatrix} x_1^{(1)} - \overline{x}^{(1)} & x_1^{(2)} - \overline{x}^{(2)} & \cdots & x_1^{(m)} - \overline{x}^{(m)} \\ x_2^{(1)} - \overline{x}^{(1)} & x_2^{(2)} - \overline{x}^{(2)} & \cdots & x_2^{(m)} - \overline{x}^{(m)} \\ \vdots & \vdots & & \vdots \\ x_n^{(1)} - \overline{x}^{(1)} & x_n^{(2)} - \overline{x}^{(2)} & \cdots & x_n^{(m)} - \overline{x}^{(m)} \end{pmatrix} \in \mathbb{R}^{n,m}.$$

Gli stimatori $b_0, b_1, \cdots, b_m$, calcolati col metodo dei minimi quadrati, sono allora dati dalle formule seguenti:

$$\boldsymbol{b} = (X^*X)^{-1}X^*\boldsymbol{y}, \tag{8.18}$$

$$b_0 = \overline{y} - b_1\overline{x}^{(1)} - b_2\overline{x}^{(2)} \cdots - b_m\overline{x}^{(m)}. \tag{8.19}$$

Notiamo che la prima formula ci fornisce $b_1, b_2, \ldots, b_m$ come trasformazione lineare del vettore $\boldsymbol{y}$ tramite la matrice $(X^*X)^{-1}X^* \in \mathbb{R}^{m,m}$. Ricordiamo (si veda il paragrafo 5.4) che $X^* \in \mathbb{R}^{m,n}$ indica la trasposta della matrice X e che $(X^*X)^{-1} \in \mathbb{R}^{m,m}$ è l'inversa della matrice $X^*X \in \mathbb{R}^{m,m}$. La seconda formula, una volta calcolato $\boldsymbol{b}$, ci permette di ricavare b_0.

Sottolineiamo ancora una volta il fatto che il termine "regressione lineare" si riferisce alla dipendenza lineare di y_i dal termine di "rumore statistico" W_i, piuttosto che alla dipendenza lineare di y_i dai predittori $x_i^{(k)}$. Il modello di regressione lineare multipla, infatti, può servire anche come modello per dipendenze *non-lineari* di y da x. Ad esempio, consideriamo la legge polinomiale

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p.$$

Per applicare l'analisi di regressione basterà trattare le varie potenze dell'*unica* variabile indipendente x come *predittori diversi* (e quindi come p variabili indipendenti):

$$x^{(1)} = x, \quad x^{(2)} = x^2, \ldots, \quad x^{(p)} = x^p,$$

per cui ci si può ricondurre al modello statistico di regressione multipla (8.17). Riprendendo ad esmpio i dati riportati nella tabella 8.1B, cerchiamo di analizzare la possibile dipendenza *quadratica* di y:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2.$$

Adottiamo perciò un modello di regressione lineare doppia

$$y_i = \beta_0 + \beta_1 x_i^{(1)} + \beta_2 x_i^{(2)} + W_i$$

in cui $x_i^{(1)} = x_i$ e $x_i^{(2)} = x_i^2$. Applicando le formule (8.18) e (8.19) otteniamo i valori degli stimatori b_0, b_1 e b_2 dei parametri β_0, β_1 e β_2 e possiamo perciò tracciare la *parabola di regressione*

$$y = b_0 + b_1 x + b_2 x^2$$

rappresentata in figura 8.6. Si osserva che la legge quadratica si adatta molto meglio ai dati sperimentali che non quella lineare tentata nel paragrafo precedente.

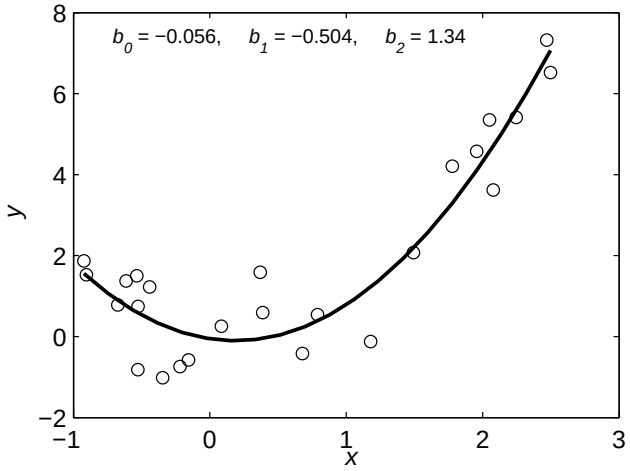


Figura 8.6: Analisi di regressione dei dati della tabella 8.1B. I cerchietti sono i punti sperimentali ed è stata tracciata la parabola di regressione $y = b_0 + b_1 x + b_2 x^2$ con i valori stimati dei parametri di regressione b_0, b_1 e b_2 riportati in alto.

Bibliografia

[1] P. Baldi, *Calcolo delle Probabilità e Statistica*, McGraw-Hill, Milano, 1992.

[2] A. Camussi, F. Möller, E. Ottaviano, M. Sari Gorla, *Metodi Statistici per la Sperimentazione Biologica* (Seconda edizione), Zanichelli, Bologna, 1995.

[3] Giorgio Dall'Aglio, *Calcolo delle Probabilità*, Zanichelli, Bologna, 2003.

[4] L. Fabbris, *Statistica multivariata : analisi esplorativa dei dati*, McGraw-Hill, Milano, 1997.

[5] J. Fowler, L. Cohen, *Statistica per ornitologi e naturalisti*, Franco Muzzio editore, 2002.

[6] E. Giusti, *Analisi Matematica 1*, Bollati-Boringhieri, Torino, 2002.

[7] E. Giusti, *Analisi Matematica 2*, Bollati-Boringhieri, Torino, 2003.

[8] B. Manly, *Multivariate Statistical Methods: a Primer*, Chapman & Hall, New York, 1994.

Tavole

Distribuzione cumulativa normale

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

La tavola riporta alcuni valori della funzione di ripartizione $F_Z(z)$ con Z normale standard. Muovendosi in verticale cambia la prima cifra decimale di z, muovendosi in orizzontale cambia la seconda cifra decimale.

Quantili della distribuzione $\mathcal{X}^2$

	0.005	0.01	0.025	0.05	0.1	0.9	0.95	0.975	0.99	0.995	0.9975	0.999
1	.00004	.00016	.00098	.00393	.0158	2.706	3.841	5.024	6.635	7.879	9.141	10.83
2	.01003	.02010	.05064	.1026	.2107	4.605	5.991	7.378	9.210	10.60	11.98	13.82
3	.0717	.1148	.2158	.3518	.5844	6.251	7.815	9.348	11.34	12.84	14.32	16.27
4	.2070	.2971	.4844	.7107	1.064	7.779	9.488	11.14	13.28	14.86	16.42	18.47
5	.4117	.5543	.8312	1.145	1.610	9.236	11.07	12.83	15.09	16.75	18.39	20.52
6	.6757	.8721	1.237	1.635	2.204	10.64	12.59	14.45	16.81	18.55	20.25	22.46
7	.9893	1.239	1.690	2.167	2.833	12.02	14.07	16.01	18.48	20.28	22.04	24.32
8	1.344	1.646	2.180	2.733	3.490	13.36	15.51	17.53	20.09	21.95	23.77	26.12
9	1.735	2.088	2.700	3.325	4.168	14.68	16.92	19.02	21.67	23.59	25.46	27.88
10	2.156	2.558	3.247	3.940	4.865	15.99	18.31	20.48	23.21	25.19	27.11	29.59
11	2.603	3.053	3.816	4.575	5.578	17.28	19.68	21.92	24.72	26.76	28.73	31.26
12	3.074	3.571	4.404	5.226	6.304	18.55	21.03	23.34	26.22	28.30	30.32	32.91
13	3.565	4.107	5.009	5.892	7.042	19.81	22.36	24.74	27.69	29.82	31.88	34.53
14	4.075	4.660	5.629	6.571	7.790	21.06	23.68	26.12	29.14	31.32	33.43	36.12
15	4.601	5.229	6.262	7.261	8.547	22.31	25.00	27.49	30.58	32.80	34.95	37.70
16	5.142	5.812	6.908	7.962	9.312	23.54	26.30	28.85	32.00	34.27	36.46	39.25
17	5.697	6.408	7.564	8.672	10.09	24.77	27.59	30.19	33.41	35.72	37.95	40.79
18	6.265	7.015	8.231	9.390	10.86	25.99	28.87	31.53	34.81	37.16	39.42	42.31
19	6.844	7.633	8.907	10.12	11.65	27.20	30.14	32.85	36.19	38.58	40.88	43.82
20	7.434	8.260	9.591	10.85	12.44	28.41	31.41	34.17	37.57	40.00	42.34	45.31
21	8.034	8.897	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40	43.78	46.80
22	8.643	9.542	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80	45.20	48.27
23	9.260	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18	46.62	49.73
24	9.886	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56	48.03	51.18
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93	49.44	52.62
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29	50.83	54.05
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64	52.22	55.48
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99	53.59	56.89
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34	54.97	58.30
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67	56.33	59.70
31	14.46	15.66	17.54	19.28	21.43	41.42	44.99	48.23	52.19	55.00	57.69	61.10
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49	56.33	59.05	62.49
33	15.82	17.07	19.05	20.87	23.11	43.75	47.40	50.73	54.78	57.65	60.39	63.87
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06	58.96	61.74	65.25
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27	63.08	66.62

La tavola riporta alcuni quantili $q_p^{\mathcal{X}^2(n)}$ della distribuzione $\mathcal{X}^2(n)$. Muovendosi in orizzontale cambia il valore di p, muovendosi in verticale cambiano i gradi di libertà n.

Quantili della distribuzione t di Student

	0.75	0.9	0.95	0.975	0.99	0.995	0.9975	0.999	0.9995	0.99975
1	1.000	3.078	6.314	12.71	31.82	63.66	127.3	318.3	636.6	1273.2
2	0.816	1.886	2.920	4.303	6.965	9.925	14.09	22.33	31.60	44.70
3	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.21	12.92	16.33
4	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610	10.31
5	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869	7.976
6	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959	6.788
7	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408	6.082
8	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041	5.617
9	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781	5.291
10	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587	5.049
11	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437	4.863
12	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318	4.716
13	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221	4.597
14	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140	4.499
15	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073	4.417
16	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015	4.346
17	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965	4.286
18	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922	4.233
19	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883	4.187
20	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850	4.146
21	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819	4.110
22	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792	4.077
23	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.768	4.047
24	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745	4.021
25	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725	3.996
26	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707	3.974
27	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690	3.954
28	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674	3.935
29	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659	3.918
30	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646	3.902
31	0.682	1.309	1.696	2.040	2.453	2.744	3.022	3.375	3.633	3.887
32	0.682	1.309	1.694	2.037	2.449	2.738	3.015	3.365	3.622	3.873
33	0.682	1.308	1.692	2.035	2.445	2.733	3.008	3.356	3.611	3.860
34	0.682	1.307	1.691	2.032	2.441	2.728	3.002	3.348	3.601	3.848
35	0.682	1.306	1.690	2.030	2.438	2.724	2.996	3.340	3.591	3.836

La tavola riporta alcuni quantili $q_p^{t(n)}$ della distribuzione t di Student con n gradi di libertà. Muovendosi in orizzontale cambia il valore di p, muovendosi in verticale cambiano i gradi di libertà n.

Valori critici della distribuzione F di Fischer per $\alpha = 0.1$

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	61.22	61.74	62.05	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.20	5.18	5.17	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.87	2.84	2.81	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.63	2.59	2.57	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.34	2.30	2.27	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.24	2.20	2.17	2.16	2.13	2.11	2.08	2.06
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	1.97	1.92	1.89	1.87	1.85	1.82	1.79	1.76
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.84	1.79	1.76	1.74	1.71	1.68	1.64	1.61
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87	1.77	1.72	1.68	1.66	1.63	1.59	1.56	1.52
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82	1.72	1.67	1.63	1.61	1.57	1.54	1.50	1.46
40	2.84	2.44	2.23	2.09	2.00	1.93	1.87	1.83	1.79	1.76	1.66	1.61	1.57	1.54	1.51	1.47	1.42	1.38
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71	1.60	1.54	1.50	1.48	1.44	1.40	1.35	1.29
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65	1.55	1.48	1.44	1.41	1.37	1.32	1.26	1.19
∞	2.71	2.30	2.08	1.94	1.85	1.77	1.72	1.67	1.63	1.60	1.49	1.42	1.38	1.34	1.30	1.24	1.17	1.00

La tavola riporta, per $\alpha = 0.1$, alcuni valori del quantile $q_{1-\alpha}^{F_{n_1,n_2}}$ della distribuzione F di Fischer con gradi di libertà n_1 e n_2. Muovendosi in orizzontale cambia il valore di n_1, muovendosi in verticale cambia il valore di n_2.

Valori critici della distribuzione F di Fischer per $\alpha = 0.05$

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	245.9	248.0	249.3	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.43	4.40	4.37
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.62	2.58	2.54
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.16	2.11	2.07
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.95	1.90	1.84
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09	2.01	1.96	1.92	1.87	1.82	1.77	1.71
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01	1.93	1.88	1.84	1.79	1.74	1.68	1.62
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92	1.84	1.78	1.74	1.69	1.64	1.58	1.51
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84	1.75	1.69	1.65	1.59	1.53	1.47	1.39
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.75	1.66	1.60	1.55	1.50	1.43	1.35	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.67	1.57	1.51	1.46	1.39	1.32	1.22	1.00

La tavola riporta, per $\alpha = 0.05$, alcuni valori del quantile $q_{1-\alpha}^{F_{n_1,n_2}}$ della distribuzione F di Fischer con gradi di libertà n_1 e n_2. Muovendosi in orizzontale cambia il valore di n_1, muovendosi in verticale cambia il valore di n_2.

Valori critici della distribuzione F di Fischer per $\alpha = 0.01$

	1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	60	120	∞
1	4052	4999	5403	5625	5764	5859	5928	5981	6022	6056	6157	6209	6240	6261	6287	6313	6339	6366
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.45	99.46	99.47	99.47	99.48	99.49	99.50
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.69	26.58	26.50	26.41	26.32	26.22	26.13
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.65	13.56	13.46
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.20	9.11	9.02
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.06	6.97	6.88
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.82	5.74	5.65
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.03	4.95	4.86
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.48	4.40	4.31
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.08	4.00	3.91
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.05	2.96	2.87
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.84	2.78	2.69	2.61	2.52	2.42
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.60	2.54	2.45	2.36	2.27	2.17
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.45	2.39	2.30	2.21	2.11	2.01
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.27	2.20	2.11	2.02	1.92	1.80
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.10	2.03	1.94	1.84	1.73	1.60
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.19	2.03	1.93	1.86	1.76	1.66	1.53	1.38
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.04	1.88	1.77	1.70	1.59	1.47	1.32	1.00

La tavola riporta, per $\alpha = 0.01$, alcuni valori del quantile $q_{1-\alpha}^{F_{n_1,n_2}}$ della distribuzione F di Fischer con gradi di libertà n_1 e n_2. Muovendosi in orizzontale cambia il valore di n_1, muovendosi in verticale cambia il valore di n_2.