Dipartimento di Informatica
Corso di Laurea in Informatica
Insegnamento di Calcolo delle Probabilità e Statistica

Nicola Cufaro Petroni

Lezioni di
Calcolo delle Probabilità
e
Statistica

anno accademico 2015/16

Università degli Studi di Bari *Aldo Moro*
Dipartimento di Matematica
via E. Orabona 4, 70125 Bari

Prefazione

La struttura di queste lezioni riflette la ben nota complementarità delle discipline che vanno sotto il nome di *Probabilità* e di *Statistica*. Per essere più chiari cominceremo con un esempio: supponiamo di voler conoscere l'orientamento politico generale dei cittadini di un determinato paese. È ben noto che in questo caso si organizzano delle *elezioni* che consistono nel raccogliere il voto di *tutti* gli elettori. Una volta esaurite le operazioni di voto si passerà allo spoglio delle schede e alla registrazione dei risultati. Tali risultati si presentano in generale come una grande quantità di dati numerici che possono essere esaminati, combinati e rappresentati in diverse maniere in modo da estrarre l'informazione rilevante ai fini elettorali. Le elezioni generali sono però tipicamente delle operazioni complesse e costose, e per questo motivo spesso si preferisce affidarsi a dei *sondaggi* per avere delle informazioni, almeno approssimative e provvisorie, sulla volontà dei cittadini. Questi sondaggi consistono nella registrazione delle opinioni di un *piccolo numero* di soggetti, a partire dal quale si ricavano delle indicazioni sulla volontà generale della popolazione intera. Ovviamente i sondaggi non possono essere sostitutivi delle elezioni, e non solo perchè bisogna dare a tutti i cittadini la possibilità di esprimere la propria opinione, ma anche per una profonda differenza fra i dati delle due operazioni. Il risultato del sondaggio, infatti, è *aleatorio*: siccome il campione di cittadini intervistato è scelto casualmente, una ripetizione del sondaggio – per quanto eseguita con i medesimi criteri – porterebbe inevitabilmente a dei risultati numerici diversi, anche se di poco. Viceversa nel caso delle elezioni l'indagine esaurisce l'intera popolazione degli elettori: una eventuale ripetizione del voto – supponendo per semplicità che non vi siano ripensamenti o errori – non modificherebbe il risultato. Noi diremo che l'esame dei risultati elettorali complessivi è compito della *Statistica descrittiva*, mentre le tecniche per ricavare informazioni su tutta la popolazione a partire dai risultati relativi a un piccolo campione sono parte della *Statistica inferenziale*. Naturalmente, come è noto, l'uso dei sondaggi comporta dei rischi dovuti alla aleatorietà dei loro esiti, per cui diventa essenziale per la Statistica inferenziale essere in grado di misurare l'affidabilità dei risultati: in questo giocheranno un ruolo essenziale i concetti e le tecniche del *Calcolo delle probabilità*.

Si noti che nel caso dell'esempio elettorale citato la possibilità di registrare il voto di tutti i cittadini esiste comunque: pertanto, in linea di principio, è sempre possibile confrontare i risultati dei sondaggi con quelli delle elezioni generali e verificarne quindi l'attendibilità. Questa possibilità, però, non sussiste sempre: in molti casi

infatti un'indagine che esaurisca l'intera popolazione semplicemente non è possibile, e ci si deve accontentate invece di esaminare le misure eseguite su un campione tentando di dedurne le caratteristiche generali del fenomeno studiato. Ad esempio in linea di principio la misura della massa di una particella elementare può essere eseguita infinite volte, e data la delicatezza della misura i risultati variano sempre in maniera aleatoria. In pratica il numero delle nostre misure sarà sempre finito, e d'altra parte, per quanto grande sia questo numero, non potremo mai dire di aver esaurito l'intera popolazione teoricamente disponibile. Allo stesso modo la determinazione della lunghezza media degli insetti di una determinata specie non potrà che essere effettuata su un campione casuale, visto che l'intera popolazione di insetti resta comunque praticamente inaccessibile. In queste occasioni, ovviamente, il raffinamento delle tecniche probabilistiche diventa essenziale.

Nasce da queste osservazioni la struttura – ormai classica – di queste lezioni divise in due parti. La prima introduce i concetti più rilevanti del Calcolo delle Probabilità (variabili aleatorie, distribuzioni, attese) e le corrispondenti procedure di calcolo; la seconda esamina invece dapprima gli strumenti principali della Statistica descrittiva (tabelle, grafici, indici di centralità e dispersione, correlazioni), e poi le tecniche più note della Statistica inferenziale (stime, intervalli di fiducia, test di ipotesi) per le quali si riveleranno essenziali le nozioni di probabilità che sono state premesse. Per ovvie ragioni di spazio e tempo gli argomenti non saranno trattati in maniera esaustiva: in particolare la parte di Calcolo delle Probabilità è principalmente rivolta a fornire quanto necessario per una corretta comprensione della parte di Statistica inferenziale. Lo scopo del corso rimane quello di mettere gli studenti in grado di usare gli strumenti più semplici e più noti della probabilità e della statistica, ma anche di stabilire alcuni pilastri concettuali che consentano loro – qualora se ne presentasse l'occasione – di estendere le loro capacità in maniera autonoma.

Bari, marzo 2016

Nicola Cufaro Petroni

Indice

Parte I

Calcolo delle Probabilità

Capitolo 1

Spazi di probabilità

1.1 Spazio dei campioni

L'origine del calcolo delle probabilità (verso la metà del XVII secolo) è strettamente legata a celebri problemi di gioco d'azzardo dai quali partiremo anche noi per introdurre i primi concetti, ma esso è oggi largamente usato in ogni ambito scientifico per produrre modelli di fenomeni naturali e risolvere i problemi associati. Inoltre la probabilità diviene uno strumento essenziale per la statistica quando si considerano campioni estratti da una popolazione mediante procedure casuali. In questo caso, infatti, i calcoli non sono più effettuati su tutta la popolazione esistente, e le stime saranno soggette a variazioni aleatorie quando il campionamento viene ripetuto. Consideriamo inizialmente degli esempi di esperimenti che diano luogo solo ad un numero finito di possibili **risultati (o eventi elementari)** casuali

Esempio 1.1. *Il caso più semplice è quello del lancio di una moneta nel quale si osserva il verificarsi di uno dei due risultati possibili: la moneta cade mostrando la faccia con la testa (T); oppure la moneta cade mostrando la faccia con la croce (C). Dire che la moneta è **equa** vuol dire che essa è non truccata, nel senso che nessuno dei due risultati è favorito rispetto all'altro ed è possibile attribuire loro le medesime possibilità di verificarsi; in tal caso diremo anche che i due eventi elementari T e C sono **equiprobabili**. Per dare una veste quantitativa a queste considerazioni si usa attribuire ad ogni evento elementare una **probabilità** intesa come frazione dell'unità, sicché nel nostro caso avremo:*

$$p = \boldsymbol{P}\{T\} = {}^1\!/_2 \qquad q = \boldsymbol{P}\{C\} = {}^1\!/_2$$

Osserviamo che $p + q = 1$, dato che con certezza (ossia con probabilità eguale ad 1) uno dei due casi, T oppure C, si verifica, e non vi sono altre possibilità.

Esempio 1.2. *Considerazioni analoghe a quelle dell'Esempio precedente applicate al caso di un dado equo conducono alla seguente attribuzione di probabilità per le sei facce che qui indicheremo con le cifre romane $I, II, ..., VI$:*

$$p_1 = \boldsymbol{P}\{I\} = {}^1\!/_6\,; \quad \ldots \quad ; p_6 = \boldsymbol{P}\{VI\} = {}^1\!/_6$$

Osserviamo che anche in questo caso si ha $p_1 + ... + p_6 = 1$.

Da quanto precede si ricava che, almeno per casi semplici, si possono attribuire delle probabilità mediante una *enumerazione*. Questa idea è alla base della cosiddetta **definizione classica** della probabilità: per attribuire una probabilità ad un evento A (in generale *non* elementare, cioè non ridotto ad un solo risultato) si enumerano i risultati *possibili* (se ritenuti, in base a qualche ipotesi, equiprobabili), e quelli *favorevoli* all'evento A (quelli, cioè, che danno luogo al verificarsi di A), e si attribuisce ad A la probabilità:

$$\boldsymbol{P}\{A\} = \frac{\text{numero dei casi favorevoli}}{\text{numero dei casi possibili}}$$

Notiamo che anche in questo caso la probabilità assegnata ad A è un numero positivo compreso fra 0 ed 1.

Esempio 1.3. *Nel lancio di un dado equo consideriamo gli eventi (non elementari)* $A =$ *"appare una faccia contrassegnata da un numero pari",* $B =$ *"appare una faccia contrassegnata da un multiplo di tre",* $C =$ *"appare una faccia diversa da* VI*". Una semplice enumerazione in base alla definizione classica ci porta a concludere che, essendo 6 i casi possibili, e rispettivamente 3, 2 e 5 i casi favorevoli ad* A, B *e* C*, si avrà:*

$$\boldsymbol{P}\{A\} = {}^{3}\!/_{6} = {}^{1}\!/_{2}\,; \qquad \boldsymbol{P}\{B\} = {}^{2}\!/_{6} = {}^{1}\!/_{3}\,; \qquad \boldsymbol{P}\{C\} = {}^{5}\!/_{6}$$

Consideriamo ora un lancio di due dadi non truccati. è facile verificare che i risultati elementari possibili sono ora 36, cioè quante sono le coppie ordinate (n, m) *dove* n *ed* m *possono assumere i 6 valori* $I, ..., VI$*. L'ipotesi che i dadi siano equi vuol dunque dire ora che i 36 eventi elementari* (I, I) ; (I, II) ; ... ; (VI, VI) *sono tutti equiprobabili e pertanto si ha*

$$\boldsymbol{P}\{I, I\} = {}^{1}\!/_{36}\,; \quad \boldsymbol{P}\{I, II\} = {}^{1}\!/_{36}\,; \quad \ldots \quad ; \boldsymbol{P}\{VI, VI\} = {}^{1}\!/_{36}$$

Sempre per enumerazione si può verificare allora ad esempio che all'evento $A =$ *"non appare la coppia* (VI, VI)*" si può attribuire una probabilità* $\boldsymbol{P}\{A\} = {}^{35}\!/_{36}$.

Dalla discussione precedente segue che la probabilità di un evento può essere pensata come un numero compreso tra 0 ed 1: il valore 1 indica la *certezza* del verificarsi, e il valore 0 la sua *impossibilità*; i valori intermedi rappresentano tutti gli altri casi. Queste probabilità possono essere calcolate nei casi semplici mediante una enumerazione di risultati equiprobabili, ma questo metodo non può in nessun modo essere considerato come generale.

Alla fine di questa sezione e nelle due sezioni seguenti riassumeremo queste considerazioni introduttive in alcune definizioni formali che preciseranno i principali concetti di base del calcolo delle probabilità

Definizione 1.4. *Chiameremo* ***spazio dei campioni*** *o* ***spazio degli eventi elementari*** *l'insieme* Ω *(finito o infinito) costituito da tutti i possibili risultati* ω *del nostro esperimento.*

Negli esempi precedenti lo spazio dei campioni $\Omega = \{\omega_1, \omega_2, \ldots, \omega_N\}$ era costituito da un *numero finito di elementi*. Ad esempio nel caso di un solo lancio di una moneta lo spazio dei campioni è composto di soli due elementi:

$$\Omega = \{T, C\}\,; \qquad N = 2\,,$$

mentre nel caso di un solo lancio di un dado si ha

$$\Omega = \{I, II, \ldots, VI\}\,; \qquad N = 6\,.$$

Se invece l'esperimento consistesse in due lanci di una moneta si avrebbe:

$$\Omega = \{TT, TC, CT, CC\}\,; \qquad N = 4\,,$$

e così via. Si noti che in generale gli elementi ω di Ω *non sono numeri*, ma oggetti astratti, anche se nel seguito avremo a che fare per lo più con insiemi numerici. I possibili spazi dei campioni Ω, inoltre, non si limitano a quelli finiti considerati finora: i casi più noti di *spazi dei campioni infiniti* che saranno adoperati nel corso di queste lezioni sono in particolare l'insieme dei numeri interi $\boldsymbol{N}$ (ad esempio nel caso in cui l'esperimento consista in conteggi che non prevedono un limite superiore), o l'insieme dei numeri reali $\boldsymbol{R}$ (nel caso di misure di generiche quantità fisiche)

1.2 Eventi

Definizione 1.5. *Chiameremo* ***evento*** *ogni sottinsieme $A \subseteq \Omega$ del quale è possibile calcolare la probabilità.*

Abbiamo già visto alcuni casi negli esempi della sezione precedente. Se poi consideriamo il caso di tre lanci di una moneta lo spazio dei campioni sarà composto di $N = 2^3 = 8$ elementi

$$\Omega = \{TTT, TTC, \ldots, CCC\}\,,$$

e il sottinsieme

$$A = \{TTT, TTC, TCT, CTT\} \subseteq \Omega$$

rappresenterà l'evento "T appare almeno due volte su tre lanci". Le osservazioni della Sezione 1.1 mostrano anche come calcolare la probabilità di tale evento sotto l'ipotesi di equiprobabilità, ma per il momento rimanderemo questo punto alla sezione successiva. In ogni caso gli eventi così definiti possono essere considerati come rappresentazioni delle *proposizioni logiche* formulabili in merito alle nostre misure, e conseguentemente le corrispondenti operazioni tra eventi (intese come operazioni tra insiemi) saranno un modello per i *connettivi logici* che uniscono delle proposizioni. Così, ad esempio, i connettivi *oppure* (OR) ed *e* (AND) sono rappresentati rispettivamente dalle operazioni di *unione* ed *intersezione*:

$$A \cup B \;=\; \{\omega \,:\, \omega \in A \text{ oppure } \omega \in B\}$$

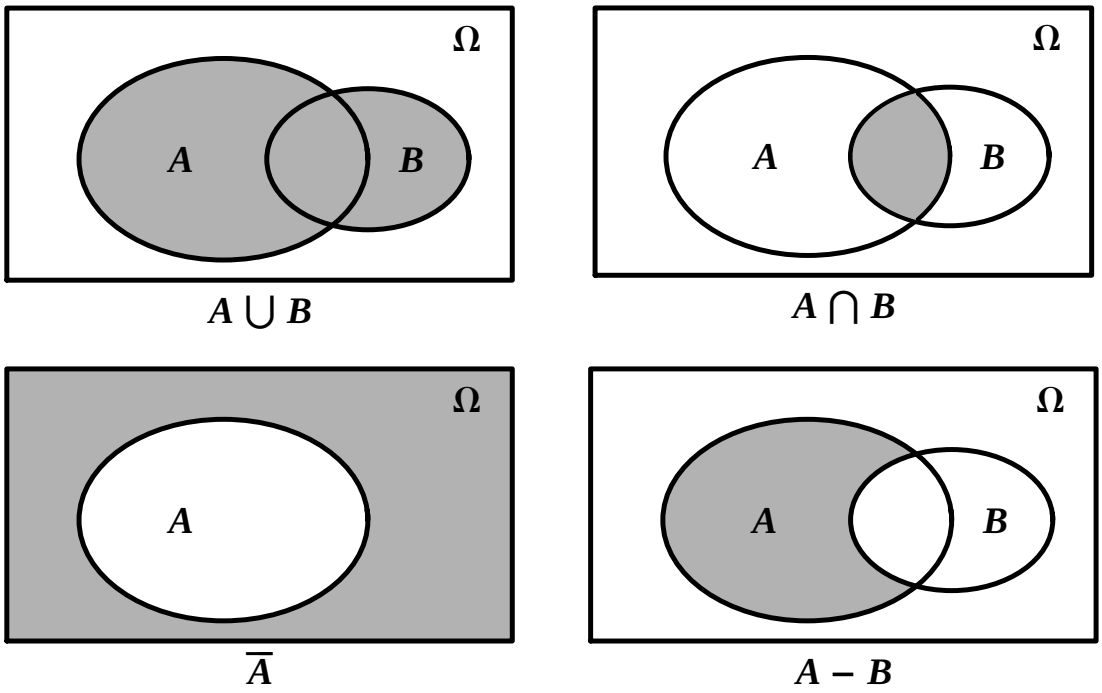


Figura 1.1: Le zone ombreggiate rappresentano i risultati delle operazioni insiemistiche indicate.

$$A \cap B \;=\; \{\omega \,:\omega \in A \,\mathrm{e}\, \omega \in B\}\,.$$

mentre il significato logico della *negazione* (complementare) e della *differenza* è facilmente deducibile tenendo presenti i diagrammi di Venn della Figura 1.1:

$$\begin{aligned} \overline{A} &= \{\omega \,:\omega \notin A\}\,; \\ A - B = A \cap \overline{B} &= \{\omega \,:\omega \in A\,,\ \mathrm{ma}\, \omega \notin B\}\,. \end{aligned}$$

Diremo anche che *un evento A si verifica* quando l'esito del nostro esperimento è un ω appartenente ad A. Si noti che in questo contesto Ω rappresenterà l'evento *certo* (nel senso che qualunque risultato cade per definizione in Ω), e $\varnothing$ rappresenterà l'evento *impossibile* (dato che nessun risultato appartiene a $\varnothing$). Diremo inoltre che i due eventi A e B sono *disgiunti* (o anche *incompatibili*) quando $A \cap B = \varnothing$ (cioè quando un risultato ω non può mai verificare contemporaneamente gli eventi A e B). Un evento può anche ridursi ad un solo elemento $A = \{\omega\}$, nel qual caso parleremo di *evento elementare*.

In generale non saremo interessati a considerare come eventi tutti i possibili sottoinsiemi di Ω: quando ad esempio $\Omega = \boldsymbol{R}$ l'insieme di tutte le parti di $\boldsymbol{R}$ sarebbe eccessivamente grande e si rivelerebbe anche poco maneggevole. Per questo motivo spesso si preferisce invece *selezionare opportune famiglie di tali sottoinsiemi da considerare come eventi*. Bisogna però, per ragioni di coerenza, garantire che tali

famiglie siano chiuse sotto le varie operazioni insiemistiche (logiche): ad esempio, se A e B sono due sottoinsiemi della nostra famiglia degli eventi, anche la loro unione o intersezione deve appartenere alla famiglia degli eventi. Diventa quindi importante aggiungere le seguenti definizioni

Definizione 1.6. *Diremo che una famiglia $\mathcal{F}$ di parti di Ω costituisce un'**algebra** quando essa è chiusa sotto tutte le operazioni insiemistiche.*

In particolare $\overline{A}$, $A \cap B$, $A \cup B$ e $A - B$ saranno sempre elementi di $\mathcal{F}$ se $A, B \in \mathcal{F}$. Si vede facilmente, ad esempio, che dato un Ω arbitrario, e $A \in \Omega$ la seguente famiglia di parti di Ω

$$\mathcal{F}_A = \{A, \overline{A}, \Omega, \varnothing\}\,.$$

è un'algebra detta algebra *generata* da A. Ci sono però anche altre importanti famiglie di sottoinsiemi

Definizione 1.7. *Diremo che una famiglia $\mathcal{D}$ di parti di Ω è una **decomposizione** di Ω se i suoi elementi D_k sono parti di Ω tutte disgiunte e tali che $\bigcup_k D_k = \Omega$*

Una decomposizione *non* è un'algebra: essa, ad esempio, non contiene le unioni dei suoi elementi. Sarà facile però convincersi del fatto che da una decomposizione si può sempre generare un'algebra aggiungendo opportunamente gli elementi mancanti. In particolare, se $A \in \Omega$, la famiglia

$$\mathcal{D}_A = \{A, \overline{A}\}$$

è una semplice decomposizione di Ω, ed è facile vedere che la $\mathcal{F}_A$ introdotta prima è un'algebra che contiene $\mathcal{D}$ ottenuta aggiungendo i due elementi Ω e $\varnothing$. Le decomposizioni giocheranno un ruolo rilevante nel capitolo sul condizionamento.

1.3 Probabilità

La probabilità $\boldsymbol{P}$ è una regola che consente di attribuire un peso probabilistico $\boldsymbol{P}\{A\}$ (un numero fra 0 e 1) ad ogni evento $A \in \mathcal{F}$. Il modo in cui tale regola viene assegnata varia secondo la natura del problema considerato. In particolare, se Ω è un insieme finito di cardinalità $\#\,\Omega = N$ (numero dei *casi possibili*) e se i suoi elementi ω_k possono essere considerati equiprobabili, si può far ricorso alla *definizione classica* (vedi Sezione 1.1): si assegna ad ogni evento elementare ω_k la probabilità $\boldsymbol{P}\{\omega_k\} = 1/N$, e ad ogni evento $A \in \mathcal{F}$ la probabilità

$$\boxed{\boldsymbol{P}\{A\} = \frac{N_A}{N}} \tag{1.1}$$

dove $N_A = \#\,A$ è la cardinalità di A, ossia il numero di elementi ω_k appartenenti ad A (numero dei *casi favorevoli*). La formula (1.1) può anche essere generalizzata al caso

in cui le ω_k non sono equiprobabili, ma hanno ognuna una probabilità $\boldsymbol{P}\{\omega_k\} = p_k$: la probabilità di un evento A sarà allora la somma delle p_k di tutti i risultati ω_k contenuti in A, cioè

$$\boxed{\boldsymbol{P}\{A\} = \sum_{\omega_k \in A} p_k} \tag{1.2}$$

Le formule (1.1) e (1.2), nonostante la loro semplicità, consentono di trattare anche problemi di una certa sofisticazione, ma non possono essere adottate in situazioni più generali. I modelli *finiti* di probabilità si rivelano infatti ben presto insufficienti perché gli spazi dei campioni sono spesso insiemi infiniti e addirittura non numerabili. In questi casi la $\boldsymbol{P}\{A\}$ non può essere costruita secondo la definizione classica, ma deve essere data per altra via, come vedremo nei capitoli seguenti. Noi qui ricorderemo solo le proprietà generali che una probabilità deve sempre avere, riservandoci di discutere più avanti il modo in cui essa viene effettivamente calcolata nei casi di nostro interesse

Definizione 1.8. *Data un'algebra $\mathcal{F}$ di eventi di Ω, chiameremo* ***probabilità*** *ogni applicazione $\boldsymbol{P} : \mathcal{F} \to [0,1]$ che sia* ***additiva****, cioè tale che, comunque scelti $n \in \boldsymbol{N}$ e $A_1, \ldots, A_n$ eventi disgiunti di $\mathcal{F}$, risulta*

$$\boldsymbol{P}\left\{\bigcup_{j=1}^{n} A_j\right\} = \sum_{j=1}^{n} \boldsymbol{P}\{A_j\}\ , \qquad \textit{con}\ \ A_i \cap A_j = \varnothing \quad \textit{se}\ i \neq j \tag{1.3}$$

Elencheremo di seguito, senza dimostrazione, le **proprietà** più note delle probabilità

1. $\boldsymbol{P}\{\varnothing\} = 0\,, \qquad \boldsymbol{P}\{\Omega\} = 1\,;$
2. $\boldsymbol{P}\{\overline{A}\,\} = 1 - \boldsymbol{P}\{A\}\ , \qquad \forall\, A \in \mathcal{F}\,;$
3. $\boldsymbol{P}\{A \cup B\} = \boldsymbol{P}\{A\} + \boldsymbol{P}\{B\} - \boldsymbol{P}\{A \cap B\}\ , \quad \forall\, A, B \in \mathcal{F}\,;$
4. $\boldsymbol{P}\{B\} \leq \boldsymbol{P}\{A\} \qquad \text{se}\ \ B \subseteq A\,, \quad \text{con}\ \ A, B \in \mathcal{F}\,;$

Definizione 1.9. *Chiameremo* ***spazio di probabilità*** *una terna $(\Omega, \mathcal{F}, \boldsymbol{P})$ in cui Ω è un insieme detto spazio dei campioni, $\mathcal{F}$ è un'algebra di eventi di Ω, e $\boldsymbol{P}$ è una probabilità su $\mathcal{F}$.*

Mostreremo infine con un esempio come le formule elementari introdotte possano già essere impiegate nella risoluzione di semplici problemi

Esempio 1.10. ***(Problema delle coincidenze)*** *Supponiamo di* ***estrarre con rimessa*** *da una scatola contenente M palline numerate una successione di n palline e di registrare i numeri estratti tenendo conto dell'ordine di estrazione. Il nostro spazio dei campioni Ω sarà allora formato dagli $N = M^n$ eventi elementari*

$\omega = (a_1, \ldots, a_n)$ costituiti dalle n-ple di numeri estratti (con possibili ***ripetizioni****). Supporremo che tali ω siano tutti equiprobabili. Consideriamo ora l'evento:*

$$\begin{aligned} A &= \{\omega : \textit{i valori delle } a_k \textit{ sono tutti diversi}\} \\ &= \textit{nelle n estrazioni non ci sono ripetizioni} \end{aligned}$$

e calcoliamone la probabilità secondo la definizione classica. Un momento di riflessione ci convincerà del fatto che

$$N_A = M\,(M-1)\,\ldots\,(M-n+1) = \frac{M!}{(M-n)!}$$

in quanto, se a_1 può essere scelto in M maniere differenti, a_2 ha solo $M-1$ possibilità di essere diverso da a_1, a_3 ne ha $M-2$ e così via fino ad a_n che potrà essere scelto in $M-n+1$ modi. Pertanto la probabilità richiesta è

$$\boldsymbol{P}\{A\} = \frac{M(M-1)\ldots(M-n+1)}{M^n} = \left(1-\frac{1}{M}\right)\left(1-\frac{2}{M}\right)\ldots\left(1-\frac{n-1}{M}\right).$$

Questo risultato permette di discutere il cosiddetto ***problema dei compleanni****: date n persone quale è la probabilità p_n che almeno due di esse celebrino il compleanno nello stesso giorno? Il modello discusso prima fornisce una risposta ponendo $M = 365$; in tal caso, essendo $\boldsymbol{P}\{A\}$ la probabilità che tutti i compleanni cadano in giorni differenti, dalla seconda delle proprietà elencate in precedenza si ha*

$$p_n = \boldsymbol{P}\{\overline{A}\} = 1 - \boldsymbol{P}\{A\} = 1 - \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\ldots\left(1-\frac{n-1}{365}\right)$$

In particolare si ottengono i seguenti sorprendenti risultati numerici:

n	*4*	*16*	*22*	*23*	*40*	*64*	...
p_n	*0.016*	*0.284*	*0.476*	*0.507*	*0.891*	*0.997*	...

è notevole infatti che già con $n = 23$ la probabilità di almeno due compleanni coincidenti supera $^1/_2$, e che con solo 64 persone tale probabilità sfiora la certezza. Si noti inoltre che con $n \geq 366$ avremo $p_n = 1$, cioè $\boldsymbol{P}\{A\} = 0$ dato che nel prodotto comparirà certamente un fattore nullo: questo corrisponde al fatto che con un numero di persone superiore alle 365 date disponibili le coincidenze diventano inevitabili. Osserviamo comunque che questi risultati appaiono meno sorprendenti se si riflette al fatto che essi sarebbero ben diversi se la domanda posta fosse stata la seguente: supponendo che io sia una delle n persone considerate nel problema precedente, quale è la probabilità q_n che almeno una celebri il suo compleanno nello stesso giorno in cui lo celebro io? Non entreremo nel dettaglio della soluzione di questo secondo problema, ma ci limiteremo a riferire che in questo secondo caso le probabilità delle coincidenze sarebbero decisamente più piccole. Per sottolineare la differenza fra i due problemi, noteremo solo che – diversamente dal primo – nel secondo caso la probabilità q_n è sempre diversa da 1 (anche per $n \geq 366$) in quanto, quale che sia il numero delle persone, può sempre capitare che nessuno celebri il suo compleanno nello stesso giorno in cui lo celebro io.

Capitolo 2

Condizionamento e indipendenza

2.1 Probabilità condizionata

Il condizionamento risponde all'esigenza di fondere una certa quantità di nuova informazione con l'informazione già contenuta in un dato spazio di probabilità $(\Omega, \mathcal{F}, \boldsymbol{P})$. L'acquisizione di nuova informazione, infatti, modifica le nostre conoscenze e quindi richiede di valutare la probabilità degli eventi in una maniera diversa da quella suggerita dalle nostre informazioni iniziali

Esempio 2.1. *Supponiamo di considerare una scatola contenente M palline delle quali m sono bianche ed $M - m$ nere ed eseguiamo due estrazioni successive. Se le palline sono estratte tutte con la medesima probabilità, e se la prima estrazione è effettuata* ***con rimessa****, è facile convincersi del fatto che – quale che sia il risultato della prima estrazione – l'evento*

$$B = \textit{la seconda pallina estratta è bianca}$$

si verifica con una probabilità $\frac{m}{M}$. Diversa sarebbe invece la nostra valutazione se la prima estrazione venisse effettuata ***senza rimessa****: se infatti nella prima estrazione venisse estratta una pallia bianca la probabilità di B sarebbe $\frac{m-1}{M-1}$; se invece nella prima estrazione si trova una pallina nera per la probabilità di B avremmo $\frac{m}{M-1}$. Pertanto, nel caso di estrazione senza rimessa, l'esito della prima estrazione influenza la probabilità dell'evento B*

Definizione 2.2. *Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \boldsymbol{P})$ e due eventi $B, D \in \mathcal{F}$ con $\boldsymbol{P}\{D\} \neq 0$, chiameremo* ***probabilità condizionata*** *di B rispetto a D (cioè probabilità che si verifichi B sapendo che si è verificato D) la quantità*

$$\boxed{\boldsymbol{P}\{B|D\} = \frac{\boldsymbol{P}\{B \cap D\}}{\boldsymbol{P}\{D\}}}$$

La quantità $\boldsymbol{P}\{B \cap D\}$ prende invece il nome di ***probabilità congiunta*** *dei due eventi B e D (cioè probabilità che si verifichino contemporaneamente B e D).*

Si potrebbe dimostrare facilmente che la nuova applicazione $\boldsymbol{P}\{\,\cdot \mid D\} : \mathcal{F} \to [0,1]$ così definita è una nuova probabilità: pertanto il condizionamento non è altro che un cambiamento di probabilità indotto dall'informazione che D si è verificato. Va inoltre ricordato che il simbolo $\boldsymbol{P}\{\,\cdot \mid \cdot\,\}$ non è simmetrico nei suoi due argomenti, nel senso che in generale avremo $\boldsymbol{P}\{B \mid D\} \neq \boldsymbol{P}\{D \mid B\}$

Teorema 2.3. (Formula della probabilità totale) *Dati un evento B e una decomposizione $\mathcal{D} = \{D_1, \ldots, D_n\}$ con $\boldsymbol{P}\{D_i\} \neq 0$, $i = 1, \ldots, n$, risulta sempre*

$$\boxed{\boldsymbol{P}\{B\} = \sum_{i=1}^{n} \boldsymbol{P}\{B \mid D_i\}\ \boldsymbol{P}\{D_i\}}$$

Dimostrazione: Siccome

$$B = B \cap \Omega = B \cap \left(\bigcup_{i=1}^{n} D_i \right) = \bigcup_{i=1}^{n} (B \cap D_i)$$

e gli eventi $B \cap D_i$ sono tutti disgiunti, dall'additività (1.3) di $\boldsymbol{P}$ risulta

$$\boldsymbol{P}\{B\} = \boldsymbol{P}\left(\bigcup_{i=1}^{n} (B \cap D_i) \right) = \sum_{i=1}^{n} \boldsymbol{P}\{B \cap D_i\}$$

D'altra parte dalla Definizione 2.2 si vede subito che

$$\boldsymbol{P}\{B \cap D_i\} = \boldsymbol{P}\{B \,|\, D_i\}\, \boldsymbol{P}\{D_i\} \qquad\qquad i = 1, \ldots, n$$

da cui segue immediatamente la formula della probabilità totale □

Osserviamo che se la decomposizione si riduce a due eventi $\mathcal{D} = \{D, \overline{D}\}$, la Formula della Probabilità Totale diviene

$$\boldsymbol{P}\{B\} = \boldsymbol{P}\{B \mid D\}\ \boldsymbol{P}\{D\} + \boldsymbol{P}\big\{B \mid \overline{D}\big\}\ \boldsymbol{P}\big\{\overline{D}\big\} \tag{2.1}$$

espressione particolarmente semplice che sarà usata spesso nel seguito

Esempio 2.4. *Riprendiamo in considerazione la scatola di palline dell'Esempio 2.1: estraiamo in successione e senza rimessa due palline e,* ***senza guardare la prima****, chiediamoci quale è la probabilità che la seconda sia bianca. Definiamo, a questo scopo, i due eventi*

$$\begin{aligned} D &= \textit{la prima pallina estratta è bianca,} \\ B &= \textit{la seconda pallina estratta è bianca;} \end{aligned}$$

e osserviamo prima di tutto che

$$\boldsymbol{P}\{D\} = \frac{m}{M}, \qquad \boldsymbol{P}\big\{\overline{D}\big\} = \frac{M-m}{M}\,.$$

Dopo aver effettuato la prima estrazione senza rimessa (ma senza guardarne il risultato), enumerando i casi possibili e i casi favorevoli otterremo che

$$\boldsymbol{P}\{B \mid D\} = \frac{m-1}{M-1} \qquad \boldsymbol{P}\big\{B \mid \overline{D}\big\} = \frac{m}{M-1}$$

Utilizzando allora la formula della probabilità totale (2.1), troviamo che

$$\boldsymbol{P}\{B\} = \frac{m-1}{M-1}\,\frac{m}{M} + \frac{m}{M-1}\,\frac{M-m}{M} = \frac{m}{M} = \boldsymbol{P}\{D\}$$

Potremo pertanto dire che la probabilità di B dipende dalle informazioni disponibili (infatti $\boldsymbol{P}\{B \mid D\}$ e $\boldsymbol{P}\big\{B \mid \overline{D}\big\}$ sono diverse da $\boldsymbol{P}\{B\}$) e dipende anche dai risultati della prima estrazione (perché $\boldsymbol{P}\{B \mid D\}$ è diversa da $\boldsymbol{P}\big\{B \mid \overline{D}\big\}$); essa tuttavia non è influenzata dal risultato della prima estrazione quando questo è sconosciuto: infatti in questo caso $\boldsymbol{P}\{B\} = \boldsymbol{P}\{D\}$, cioè la probabilità di trovare una pallina bianca alla seconda estrazione resta invariata come se non avessimo eseguito la prima estrazione

Teorema 2.5. (Teorema di Bayes) *Dati due eventi B, D con $\boldsymbol{P}\{B\} \neq 0, \boldsymbol{P}\{D\} \neq 0$, risulta*

$$\boldsymbol{P}\{D \mid B\} = \frac{\boldsymbol{P}\{B \mid D\}\ \boldsymbol{P}\{D\}}{\boldsymbol{P}\{B\}} \tag{2.2}$$

Inoltre, se $\mathcal{D} = \{D_1, \ldots, D_n\}$ è una decomposizione di Ω con $\boldsymbol{P}\{D_i\} \neq 0$, $i = 1, \ldots, n$, risulta anche

$$\boxed{\boldsymbol{P}\{D_i \mid B\} = \frac{\boldsymbol{P}\{B \mid D_i\}\ \boldsymbol{P}\{D_i\}}{\sum_{j=1}^{n} \boldsymbol{P}\{B \mid D_j\}\ \boldsymbol{P}\{D_j\}} \qquad i = 1, \ldots, n} \tag{2.3}$$

Dimostrazione: La dimostrazione della (2.2) si basa sul fatto che dalla Definizione 2.2 di probabilità condizionata si ha

$$\boldsymbol{P}\{D \mid B\}\ \boldsymbol{P}\{B\} = \boldsymbol{P}\{B \cap D\} = \boldsymbol{P}\{B \mid D\}\ \boldsymbol{P}\{D\}\ ;$$

la seconda relazione si ottiene poi dalla prima tramite il Teorema 2.3. □

Nelle applicazioni statistiche gli eventi D_i del Teorema di Bayes sono spesso chiamati *ipotesi* e le $\boldsymbol{P}\{D_i\}$ *probabilità a priori* di tali ipotesi, mentre le probabilità condizionate $\boldsymbol{P}\{D_i \mid B\}$ si chiamano *probabilità a posteriori*. Il significato di questi nomi sarà chiarito nella discussione dell'Esempio 2.8, ma a questo scopo converrà introdurre prima il concetto di *indipendenza*

2.2 Indipendenza

Due eventi sono indipendenti quando il verificarsi di uno di essi non ha alcun effetto sul valore della probabilità che viene attribuita all'altro. Sulla base del concetto di

probabilità condizionata introdotto prima potremo quindi dire che l'evento A è indipendente dall'evento B quando $\boldsymbol{P}\{A \mid B\} = \boldsymbol{P}\{A\}$ e quindi, dalla Definizione 2.2 di probabilità condizionata, se $\boldsymbol{P}\{A \cap B\} = \boldsymbol{P}\{A\}\ \boldsymbol{P}\{B\}$. È facile peraltro, data la simmetria di queste relazioni, verificare che se A è indipendente da B, anche B è indipendente da A. Il concetto di indipendenza può poi essere esteso anche al caso in cui il numero di eventi è maggiore di due, ma bisogna notare che in questo caso sarà possibile parlare e di indipendenza *due a due*, nel senso di $\boldsymbol{P}\{A \cap B\} = \boldsymbol{P}\{A\}\ \boldsymbol{P}\{B\}$, di indipendenza *tre a tre*, nel senso di $\boldsymbol{P}\{A \cap B \cap C\} = \boldsymbol{P}\{A\}\ \boldsymbol{P}\{B\}\ \boldsymbol{P}\{C\}$, e così via. Questi diversi *livelli di indipendenza*, però, non si implicano l'uno con l'altro: infatti, ad esempio, tre eventi possono essere indipendenti due a due senza esserlo tre a tre e viceversa. Per questo motivo l'indipendenza di $n \geq 3$ eventi dovrà essere sempre dichiarata esplicitamente a tutti i livelli possibili: due a due, tre a tre, e così via fino a n a n. Queste osservazioni sono raccolte nella seguente definizione

Definizione 2.6. *Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \boldsymbol{P})$ diremo che $n \geq 2$ eventi $A_1, \ldots, A_n$ sono **indipendenti** se, comunque sceltine k fra di essi (diciamo $A_{j_1}, \ldots, A_{j_k}$, con $2 \leq k \leq n$) risulta*

$$\boldsymbol{P}\{A_{j_1} \cap \ldots \cap A_{j_k}\} = \boldsymbol{P}\{A_{j_1}\} \ldots \boldsymbol{P}\{A_{j_k}\}$$

ossia se essi sono indipendenti due a due, tre a tre, ... , n a n, in tutte le combinazioni possibili. In particolare se $n = 2$, i due eventi A_1 e A_2 si dicono indipendenti quando

$$\boldsymbol{P}\{A_1 \cap A_2\} = \boldsymbol{P}\{A_1\}\ \boldsymbol{P}\{A_2\}$$

Si può mostrare – ma noi trascureremo di farlo – che se due eventi A_1, A_2 sono indipendenti, allora lo sono anche $A_1, \overline{A}_2$, e così pure $\overline{A}_1, A_2$, e $\overline{A}_1, \overline{A}_2$. Analoghe relazioni valgono anche nel caso tre o più eventi. Utilizzeremo ora questi concetti per costruire un importante modello che troverà applicazione nel seguito

Teorema 2.7. (Modello di Bernoulli) *Sia data un'urna contenente palline bianche e nere, e sia $p \in [0, 1]$ la proporzione delle palline bianche: se si eseguono n estrazioni con rimessa, e se con $0 \leq k \leq n$ si definisce l'evento*

$$B = \textit{sulle } n \textit{ estrazioni si trovano } k \textit{ palline bianche, in un ordine qualsiasi}$$

risulta

$$\boxed{\boldsymbol{P}\{B\} = \binom{n}{k} p^k (1-p)^{n-k}} \tag{2.4}$$

Dimostrazione: Per definire lo spazio dei campioni Ω adotteremo la convenzione secondo la quale il simbolo 1 indicherà il ritrovamento di una pallina bianca, e il simbolo 0 quello di una pallina nera: in questo modo Ω sarà composto da tutte le possibili n-ple di simboli $0, 1$, cioè $\omega = (a_1, \ldots, a_n)$ dove le a_j assumono i valori $0, 1$ in tutti i modi possibili. Siano poi $A_1, \ldots, A_n$ gli n eventi

$$A_j = \textit{si trova pallina bianca nella } j\textit{-ma estrazione} \qquad j = 1, \ldots, n$$

che risultano indipendenti (assieme ai loro complementari) perché le estrazioni sono effettuate *con rimessa*. Sarà facile vedere allora che nel nostro Ω ogni A_j consiste di tutte le ω con 1 al j-mo posto, e valori arbitrari di tutti gli altri simboli. Inoltre, se con $\boldsymbol{P}\{1\}$ e $\boldsymbol{P}\{0\}$ indichiamo le probabilità dei due risultati in ogni possibile estrazione, sappiamo anche che per ipotesi

$$\boldsymbol{P}\{A_j\} = \boldsymbol{P}\{1\} = p \qquad \boldsymbol{P}\{\overline{A}_j\} = \boldsymbol{P}\{0\} = 1 - p \qquad j = 1, \ldots, n$$

Per determinare la probabilità dell'evento B cominciamo allora con l'osservare che se l'evento elementare ω è una n-pla che contiene k volte 1, e $n-k$ volte 0 in un ordine ben preciso, esso coincide con l'intersezione di k eventi A_i e di $n-k$ eventi complementari $\overline{A}_j$, tutti indipendenti fra loro per cui avremo

$$\boldsymbol{P}\{\omega\} = \underbrace{\boldsymbol{P}\{1\} \cdot \ldots \cdot \boldsymbol{P}\{1\}}_{k \text{ volte}} \cdot \underbrace{\boldsymbol{P}\{0\} \cdot \ldots \cdot \boldsymbol{P}\{0\}}_{n-k \text{ volte}} = p^k (1-p)^{n-k} \,. \tag{2.5}$$

Siccome però, per un fissato k, i k simboli 1 possono essere disposti in vario modo nella n-pla ω, l'evento B consisterà di un certo numero di eventi elementari tutti con la stessa probabilità (2.5). Si può dimostrare – ma noi trascureremo di farlo – che il numero di queste diverse combinazioni è dato dal **coefficiente binomiale**

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n(n-1)\ldots(n-k+1)}{k(k-1)\ldots 2 \cdot 1} \tag{2.6}$$

e quindi dalla formula (1.2) e da (2.5) si ottiene immediatamente la (2.4) □

Si comprende facilmente che il precedente risultato resta invariato quale che sia il modello concreto (estrazioni da un'urna, lanci di monetina,...) al quale esso fa riferimento, sicché potremo in definitiva dire che (2.4) rappresenta la probabilità di ottenere k successi su n **tentativi indipendenti di verifica di un evento** che in ogni tentativo si verifica con probabilità p

Esempio 2.8. (Applicazioni del Teorema di Bayes) *Consideriamo due urne D_1 e D_2, esteriormente indistinguibili, contenenti ambedue palline bianche e nere, ma in proporzioni diverse. Più precisamente, la frazione di palline bianche in D_1 sia $^1/_2$, mentre quella in D_2 sia $^2/_3$. Supponiamo inoltre di non poter esaminare direttamente l'interno delle due urne, ma di poter eseguire un numero arbitrario di estrazioni con rimessa. Scegliamo ora una delle due urne, e chiediamoci quale delle due abbiamo preso. È evidente che $\mathcal{D} = \{D_1, D_2\}$ costituisce una decomposizione di Ω, e che, in assenza di altre informazioni, le probabilità a priori di aver preso una o l'altra urna saranno uguali per cui si avrà*

$$\boldsymbol{P}\{D_1\} = \boldsymbol{P}\{D_2\} = {}^1/_2$$

È però intuitivo pensare che per saperne di più sull'urna scelta basterà eseguire un certo numero di estrazioni con rimessa: infatti l'osservazione di un numero elevato

di palline bianche ci farebbe propendere verso l'idea di aver preso D_2, e viceversa nel caso contrario. La formula di Bayes (2.3) e i risultati del Teorema 2.7 ci permettono ora di dare una veste quantitativa a queste considerazioni rimaste fin qui puramente qualitative. Supponiamo, infatti, per fissare le idee, di eseguire $n = 10$ estrazioni con rimessa dall'urna scelta, e di trovare $k = 4$ volte una pallina bianca, e $n-k = 6$ volte una pallina nera, cioè che si verifichi l'evento

$$B \;=\; \textit{su } 10 \textit{ estrazioni si trovano } 4 \textit{ palline bianche}$$

Abbiamo già osservato che, nei due casi possibili di scelta delle nostre urne D_1 e D_2, le probabilità dell'evento B sono date dalla formula (2.4) rispettivamente con $p = \frac{1}{2}$ e con $p = \frac{2}{3}$, cioè sappiamo che

$$\begin{aligned}
\boldsymbol{P}\{B|D_1\} &= \binom{10}{4}\left(\frac{1}{2}\right)^4\left(\frac{1}{2}\right)^{10-4} = \binom{10}{4}\frac{1}{2^{10}} \\
\boldsymbol{P}\{B|D_2\} &= \binom{10}{4}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^{10-4} = \binom{10}{4}\frac{2^4}{3^{10}}
\end{aligned}$$

e pertanto dalla formula di Bayes otteniamo

$$\begin{aligned}
\boldsymbol{P}\{D_1|B\} &= \frac{\boldsymbol{P}\{B|D_1\}\;\boldsymbol{P}\{D_1\}}{\boldsymbol{P}\{B|D_1\}\;\boldsymbol{P}\{D_1\} + \boldsymbol{P}\{B|D_2\}\;\boldsymbol{P}\{D_2\}} = \frac{\frac{1}{2^{10}}}{\frac{1}{2^{10}} + \frac{2^4}{3^{10}}} \\
&= \frac{3^{10}}{3^{10} + 2^{14}} = 0.783 \\
\boldsymbol{P}\{D_2|B\} &= \frac{2^{14}}{3^{10} + 2^{14}} = 0.217
\end{aligned}$$

Come si noterà, l'aver osservato un numero relativamente scarso di palline bianche favorisce – ma ora con una stima numerica precisa delle probabilità a posteriori – l'ipotesi di aver preso l'urna D_1 che contiene la proporzione più piccola di palline bianche. Naturalmente ulteriori estrazioni produrranno delle modifiche di questa valutazione anche se, a lungo andare, ci attendiamo intuitivamente una stabilizzazione del risultato

Capitolo 3

Variabili aleatorie

3.1 Variabili e vettori aleatori

Abbiamo già osservato che in generale lo spazio dei campioni Ω non è necessariamente un insieme numerico nel senso che i suoi elementi sono oggetti astratti: ad esempio nel caso della moneta gli elementi di Ω sono T e C. D'altra parte nelle applicazioni gli aspetti più rilevanti sono legati ai valori numerici di misure empiriche. Assume quindi una certa importanza la possibilità di introdurre delle procedure che consentano di associare dei numeri ai risultati dei nostri esperimenti aleatori

Definizione 3.1. *Dato uno spazio di probabilità $(\Omega, \mathcal{F}, \boldsymbol{P})$ si dice* ***variabile aleatoria (v-a)*** *un'applicazione $X : \Omega \to \boldsymbol{R}$ tale che*

$$\{X \in \mathcal{B}\} \equiv \{\omega \,:\, X(\omega) \in \mathcal{B}\} \in \mathcal{F} \tag{3.1}$$

comunque scelto un arbitrario sottoinsieme $\mathcal{B}$ di $\boldsymbol{R}$, come ad esempio un intervallo $\mathcal{B} = [a, b]$ (vedi anche Figura 3.1). Si noti che il simbolo $\{X \in \mathcal{B}\}$ qui introdotto sarà sempre inteso come un'abbreviazione che denota un evento sottoinsieme di Ω

A prima vista la precisazione (3.1) contenuta nella precedente definizione può sembrare superflua perché $\{X \in \mathcal{B}\}$ è sempre un sottoinsieme di Ω: il punto cruciale però è che tale sottoinsieme deve anche essere uno di quelli selezionati per essere un evento di $\mathcal{F}$, e questo non è sempre garantito. Come abbiamo già osservato, infatti, in generale $\mathcal{F}$ non è la famiglia di tutte le parti di Ω, ma un'opportuna famiglia di sottoinsiemi (eventi) per i quali siamo in grado di assegnare una probabilità

Esempio 3.2. *Si consideri un dado con le facce colorate con sei colori diversi: in questo caso Ω è costituito dall'insieme dei sei colori scelti. Si supponga poi di stabilire le regole di un gioco nel quale ad ogni colore è associata una vincita in denaro: la regola che attribuisce la vincita ad ogni colore è una v-a. Un altro esempio semplice, ma molto rilevante di v-a è costituito dall'***indicatore** $I_A(\omega)$ *di un evento $A \in \mathcal{F}$, cioè la funzione con valori $0, 1$ così definita*

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

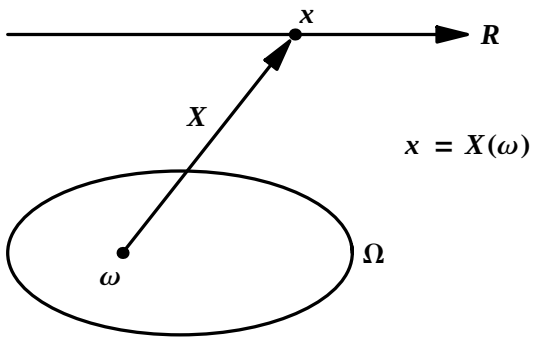


Figura 3.1: Illustrazione grafica della definizione di variabile aleatoria.

essa quindi che vale 1 per tutti i risultati ω che verificano A, e 0 in tutti gli altri casi. Si verifica facilmente che si tratta di una v-a secondo la Definizione 3.1 perché tutti gli insiemi del tipo $\{I_A \in [a,b]\}$ sono $A, \overline{A}, \varnothing$ e Ω, e quindi sono certamente elementi di $\mathcal{F}$ dato che $A \in \mathcal{F}$

La Definizione 3.1 con la precisazione (3.1) è fondamentale perché consente di associare una probabilità agli insiemi $\mathcal{B}$ di $\boldsymbol{R}$, ad esempio agli intervalli $[a,b]$: in pratica – come vedremo nella definizione seguente – la *v-a* X proietta sull'insieme $\boldsymbol{R}$ un'*immagine* P_X della probabilità $\boldsymbol{P}$ inizialmente definita su Ω

Definizione 3.3. *Chiameremo* ***legge*** *o* ***distribuzione*** *della v-a X la probabilità P_X da essa definita su $\boldsymbol{R}$ tramite la relazione*

$$P_X(\mathcal{B}) \equiv \boldsymbol{P}\{X \in \mathcal{B}\} \tag{3.2}$$

dove $\mathcal{B}$ è un arbitrario sottoinsieme di $\boldsymbol{R}$, ad esempio un intervallo $\mathcal{B} = [a,b]$. Per indicare che X segue la legge P_X useremo la notazione $X \sim P_X$

La relazione (3.2) definisce quindi una nuova P_X che permette di attribuire una probabilità agli *insiemi $\mathcal{B}$ di valori assunti da X*: per lo più essi saranno intervalli di $\boldsymbol{R}$. In questo modo l'insieme $\boldsymbol{R}$, i suoi intervalli e P_X costituiscono a tutti gli effetti un nuovo spazio di probabilità. Si noti che la Definizione 3.3 è coerente proprio perché la precedente Definizione 3.1 garantisce che $\{X \in \mathcal{B}\}$ sia un elemento di $\mathcal{F}$

Esempio 3.4. *Sia $A \in \mathcal{F}$ un evento con $\boldsymbol{P}\{A\} = p$, e sia I_A il suo indicatore. La v-a I_A assume solo i due valori 0 e 1, e pertanto eventi del tipo $\{I_A \in [2\,,4]\}$, oppure $\{I_A \in (-\infty,-3]\}$ non capitano mai, cioè*

$$\{I_A \in [2\,,4]\} = \{I_A \in (-\infty,-3]\} = \varnothing$$

per cui

$$P_{I_A}\big([2\,,4]\big) = P_{I_A}\big((-\infty\,,-3]\big) = 0$$

Viceversa si vede facilmente che ad esempio

$$\{I_A \in [{}^1/_2\,, 2]\} = A \qquad \{I_A \in [-1\,,\ {}^1/_2]\} = \overline{A} \qquad \{I_A \in [-2\,, 2]\} = \Omega$$

per cui

$$P_{I_A}([{}^1/_2\,, 2]) = p\,, \qquad P_{I_A}([-1\,,\ {}^1/_2]) = 1-p\,, \qquad P_{I_A}([-2\,, 2]) = 1$$

Ogni *v-a* attribuisce *in generale* al medesimo intervallo una probabilità diversa, per cui *v-a* diverse $X, Y, \ldots$ hanno, *in generale*, leggi diverse $P_X, P_Y, \ldots$. Non è però vietato che ci siano *v-a* differenti (nel senso della Definizione 3.1) ma dotate della medesima legge: in questo caso si parlerà di ***v-a* identicamente distribuite (*id*)**

Esempio 3.5. *Per costruire un esempio di* ***due v-a distinte, ma id*** *si consideri un dado equo e si definiscano*

$$X = \begin{cases} 1 & \textit{se esce II,IV oppure VI,} \\ 0 & \textit{altrimenti.} \end{cases} \qquad Y = \begin{cases} 1 & \textit{se esce I, II oppure III,} \\ 0 & \textit{altrimenti.} \end{cases}$$

X ed Y sono ovviamente v-a diverse: ad esempio, se esce I, X prende valore 0 mentre Y vale 1. Ciononostante esse sono id. In effetti X ed Y sono gli indicatori di due eventi, rispettivamente A = "esce II,IV oppure VI" e B = "esce esce I, II oppure III", che pur essendo diversi hanno la stessa probabilità ${}^1/_2$. Pertanto X e Y assumono ambedue valore 0 oppure 1 con la medesima probabilità ${}^1/_2$. Ragionando come nell'Esempio 3.4 si può allora mostrare che esse attribuiscono la stessa probabilità a tutti gli intervalli di $\boldsymbol{R}$, cioè $P_X = P_Y$

In molte situazioni sarà necessario associare ad ogni $\omega \in \Omega$ non un solo numero, ma un intero vettore di m numeri: questo avviene, ad esempio, quando si misurano contemporaneamente diverse quantità fisiche. In questo caso avremo una applicazione da Ω in $\boldsymbol{R}^m$, e ciò in pratica equivale a definire m *v-a* $X_1, \ldots, X_m$ che possono essere pensate come le componenti di un vettore $\boldsymbol{X}$. Riassumendo queste osservazioni sarà opportuno pertanto introdurre anche la definizione seguente

Definizione 3.6. *Chiameremo* ***variabile aleatoria*** *m-**dimensionale** (o **vettore aleatorio**) un vettore $\boldsymbol{X} = (X_1, \ldots, X_m)$ le cui componenti X_j sono m v-a nel senso della Definizione 3.1.*

Ovviamente, dalle componenti di un dato vettore aleatorio $\boldsymbol{X} = (X_1, \ldots, X_m)$, potremo sempre generare non solo gli eventi $\{X_1 \in \mathcal{B}_1\}, \ldots, \{X_m \in \mathcal{B}_m\}$ già introdotti, ma anche gli *eventi congiunti*, cioè quelli del tipo

$$\begin{aligned} \{X_1 \in \mathcal{B}_1, \ldots, X_m \in \mathcal{B}_m\} &\equiv \{\omega \in \Omega : X_1(\omega) \in \mathcal{B}_1, \ldots, X_m(\omega) \in \mathcal{B}_m\} \\ &= \{X_1 \in \mathcal{B}_1\} \cap \ldots \cap \{X_m \in \mathcal{B}_m\} \end{aligned}$$

Questo ci consente ora di assegnare una probabilità anche ai sottoinsiemi di $\boldsymbol{R}^m$ del tipo $\boldsymbol{\mathcal{B}} = \mathcal{B}_1 \times \ldots \times \mathcal{B}_m$ (in particolare, se i $\mathcal{B}_i$ sono intervalli di $\boldsymbol{R}$, il sottoinsieme $\boldsymbol{\mathcal{B}}$ sarà un rettangolo m-dimensionale)

Definizione 3.7. *Chiameremo* ***legge o distribuzione congiunta*** *del vettore* $\boldsymbol{X} = (X_1, \ldots, X_m)$ *la probabilità* $P_{\boldsymbol{X}}$ *da esso proiettata su* $\boldsymbol{R}^m$ *tramite la relazione*

$$P_{\boldsymbol{X}}(\boldsymbol{\mathcal{B}}) \equiv \boldsymbol{P}\{X_1 \in \mathcal{B}_1, \ldots, X_m \in \mathcal{B}_m\} \tag{3.3}$$

Le leggi $P_{X_i}(\mathcal{B}_i)$ *delle singole componenti* X_i *(ottenute dalla Definizione 3.3) si chiameranno invece* ***leggi o distribuzioni marginali***

Siamo ora in grado di adattare alle *v-a* il concetto di indipendenza di eventi della Definizione 2.6: come sarà precisato nella definizione che segue, infatti, due o più *v-a* sono indipendenti quando tutti gli eventi da esse generati sono indipendenti nel senso della Definizione 2.6

Definizione 3.8. *Diremo che le componenti di un vettore* $\boldsymbol{X} = (X_1, \ldots, X_m)$ *sono* ***v-a indipendenti*** *se tutti gli eventi* $\{X_1 \in \mathcal{B}_1\}, \ldots, \{X_m \in \mathcal{B}_m\}$ *da esse generati sono indipendenti nel senso della Definizione 2.6, comunque scelti i sottoinsiemi numerici* $\mathcal{B}_1, \ldots, \mathcal{B}_m$*; cioè se*

$$\boldsymbol{P}\{X_1 \in \mathcal{B}_1, \ldots, X_m \in \mathcal{B}_m\} = \boldsymbol{P}\{X_1 \in \mathcal{B}_1\} \cdot \ldots \cdot \boldsymbol{P}\{X_m \in \mathcal{B}_m\}$$

comunque scelti i sottoinsiemi numerici $\mathcal{B}_1, \ldots, \mathcal{B}_m$ *(ad esempio intervalli di* $\boldsymbol{R}$*). Se inoltre le n v-a indipendenti hanno anche tutte la stessa distribuzione diremo che esse sono* ***v-a indipendenti e identicamente distribuite (iid)***

3.2 Funzioni di distribuzione

Nelle sezioni successive esamineremo gli strumenti matematici che ci permetteranno di costruire esplicitamente le leggi delle *v-a* di uso più comune, e a questo scopo distingueremo le leggi in due grandi categorie – quelle *discrete* e quelle *continue* – che esauriscono tutti i casi di interesse pratico. In questa sezione in vece introdurremo un concetto comune ad ambedue le categorie che sarà molto utile in seguito

Definizione 3.9. *Chiameremo* ***funzione di distribuzione cumulativa (FDC)*** *di una v-a* X *la funzione definita come*

$$F_X(x) \equiv \boldsymbol{P}\{X \le x\} = \boldsymbol{P}\{X \in (-\infty, x]\} = P_X\big((-\infty, x]\big) \qquad \forall x \in \boldsymbol{R}.$$

Teorema 3.10. *La FDC* $F_X(x)$ *di una v-a* X *gode delle seguenti proprietà:*

1. $0 \le F_X(x) \le 1$*, per ogni* $x \in \boldsymbol{R}$*;*
2. $F_X(x)$ *è una funzione non decrescente di* x *che tende a 0 per* $x \to -\infty$*, e tende a 1 per* $x \to +\infty$*;*
3. $F_X(x)$ *può presentare delle discontinuità (salti): se è discontinua in* x_0*,* $F_X(x)$ *sarà continua da destra, e ammetterà limite da sinistra diverso da* $F_X(x_0)$*;*

Dimostrazione: Omessa. Osserveremo solo che da tutte queste proprietà si ricava che $F_X(x)$ è una funzione che cresce in maniera monotòna da $F(-\infty) = 0$ a $F_X(+\infty) = 1$, e che può presentare discontinuità di prima specie (salti) nel senso che, se x_0 è una di tali discontinuità, esisteranno sempre finiti ambedue i limiti da destra e da sinistra

$$F_X(x_0^+) = \lim_{x \downarrow x_0} F_X(x) \qquad F_X(x_0^-) = \lim_{x \uparrow x_0} F_X(x)$$

In tal caso, però, $F_X(x)$ sarà continua solo da destra nel senso che $F_X(x_0^+) = F_X(x_0)$, mentre invece $F_X(x_0^-) < F_X(x_0)$ □

Teorema 3.11. *La probabilità attribuita da P_X agli intervalli chiusi a destra $(a, b]$ si calcola dalla FDC $F_X(x)$ mediante la formula*

$$\boxed{\boldsymbol{P}\{a < X \leq b\} = P_X\big((a, b]\big) = F_X(b) - F_X(a)} \tag{3.4}$$

In particolare la probabilità attribuita ad un singolo punto b è

$$\boldsymbol{P}\{X = b\} = P_X\big(\{b\}\big) = F_X(b) - F_X(b^-) \tag{3.5}$$

Dimostrazione: La (3.4) è una semplice conseguenza della Definizione 3.9 e della additività (1.3) delle probabilità: infatti

$$(-\infty, b] = (-\infty, a] \cup (a, b] \qquad (-\infty, a] \cap (a, b] = \varnothing$$

e quindi

$$F_X(b) = P_X\big((-\infty, b]\big) = P_X\big((-\infty, a]\big) + P_X\big((a, b]\big) = F_X(a) + P_X\big((a, b]\big)$$

cioè la (3.4). La (3.5) deriva poi da (3.4) se si osserva che

$$P_X\big(\{b\}\big) = \lim_{a \uparrow b} P_X\big((a, b]\big)$$

In particolare da (3.5) consegue che la probabilità attribuita ad un singolo punto b è diversa da zero se e solo se $F_X(x)$ è discontinua in b; viceversa $P_X\big(\{b\}\big) = 0$ in tutti i punti b in cui $F_X(x)$ è continua □

Il concetto di funzione di distribuzione cumulativa si estende in maniera naturale anche ai vettori aleatori, ma in questo caso dovremo tenere conto sia delle *FDC* della legge congiunta del vettore aleatorio, che delle singole *FDC* delle leggi marginali delle sue componenti

Definizione 3.12. *Chiameremo **FDC congiunta** di un vettore $\boldsymbol{X} = (X_1, \ldots, X_m)$ la funzione con m variabili*

$$F_X(x_1, \ldots, x_m) = \boldsymbol{P}\{X_1 \leq x_1, \ldots, X_m \leq x_m\}$$

*e **FDC marginali** le FDC delle singole componenti X_j*

$$F_{X_j}(x_j) = \boldsymbol{P}\{X_j \leq x_j\} \qquad j = 1, \ldots, m$$

Teorema 3.13. *Le componenti di un vettore aleatorio* $\boldsymbol{X} = (X_1, \ldots, X_m)$ *sono indipendenti se e solo se*

$$F_{\boldsymbol{X}}(x_1, \ldots, x_m) = F_{X_1}(x_1) \cdot \ldots \cdot F_{X_m}(x_m)$$

cioè se la FDC congiunta si fattorizza nel prodotto delle sue FDC marginali

Dimostrazione: Omessa □

Da quanto precede si ricava che la conoscenza della *FDC* è tutto quello che serve per conoscere completamente la legge di una *v-a* o di un vettore aleatorio. Non sempre, però, la *FDC* si rivela lo strumento più comodo per eseguire i calcoli pratici. Nelle sezioni seguenti introdurremo pertanto anche gli ulteriori concetti che saranno poi utilizzati. Prima però sarà opportuno aggiungere che spesso ci capiterà anche di utilizzare *v-a* Z ottenute come *funzioni, somme o altre combinazioni di altre v-a*: ad esempio, data la *v-a* X, possiamo essere interessati ad esaminare la *v-a* $Z = X^2$, oppure $Z = \cos X$, o altre funzioni di X. Analogamente, date due (o più) *v-a* X e Y, potremmo voler eseguire dei calcoli sulla *v-a* $Z = X + Y$, oppure $Z = XY$ e così via. La *v-a* Z in tutti questi casi avrà una sua legge e una sua *FDC* $F_Z(z)$, e vi sono tecniche particolari che consentono di ricavare $F_Z(z)$ a partire dalle $F_X(x), F_Y(y), \ldots$ iniziali. Noi non entreremo in questi dettagli, ma ci limiteremo, ove necessario, a ricordare i risultati più importanti senza dimostrazioni

3.3 Leggi discrete

Definizione 3.14. *Chiameremo* ***v-a discrete*** *le v-a* X *che assumono solo un insieme finito, o infinito numerabile di valori che indicheremo con* x_k *dove* $k = 0, 1, 2, \ldots \in \boldsymbol{N}$. *Le leggi corrispondenti si chiameranno* ***leggi discrete***

In particolare nel seguito saremo interessati prevalentemente a *v-a* discrete X i cui valori x_k coincideranno proprio con i numeri interi $k = 0, 1, 2, \ldots \in \boldsymbol{N}$: in questo caso i valori $x_k = k$ di X rappresentano in genere il risultato aleatorio di un *conteggio*. Un altro importante caso particolare è l'indicatore I_A di un evento A: esso è ovviamente una *v-a* discreta dato che assume solo i due valori 0 e 1

È facile capire che la legge di una *v-a* discreta X sarà determinata non appena siano assegnate le x_k e le rispettive probabilità

$$p_k = \boldsymbol{P}\{X = x_k\}\ , \qquad k = 1, 2, \ldots$$

dove l'evento $\{X = x_k\}$ è come al solito il sottoinsieme costituito dalle $\omega \in \Omega$ per le quali $X(\omega) = x_k$. Val la pena notare che l'assegnazione dei numeri p_k può essere arbitraria purché essi soddisfino le due seguenti proprietà

$$p_k \geq 0\,, \qquad k = 1, 2, \ldots\ ; \qquad \sum_k p_k = 1 \tag{3.6}$$

con la precisazione che, nel caso in cui X assume un insieme infinito numerabile di valori, la somma nella relazione precedente è in realtà una serie della quale deve essere ovviamente assicurata la convergenza. Quando le x_k e le p_k sono note anche la legge P_X è completamente determinata: dato ad esempio intervallo $\mathcal{B} = [a, b]$, la probabilità $P_X(\mathcal{B}) = \boldsymbol{P}\{X \in \mathcal{B}\}$ si otterrà sommando le p_k relative alle x_k che cadono in $\mathcal{B}$

$$P_X(\mathcal{B}) = \sum_{k:\, x_k \in \mathcal{B}} \boldsymbol{P}\{X = x_k\}$$

Da quanto precede si ricava che la *FDC* $F_X(x)$ di una *v-a* discreta X è sempre una *funzione a scalini*: essa presenta delle discontinuità (salti) nei valori x_k assunti da X, e rimane costante fra due valori consecutivi x_k e x_{k+1}; inoltre l'altezza del salto in x_k coincide proprio con la probabilità che X assuma il valore x_k: infatti, tenendo conto della (3.5), si ha

$$\boxed{p_k = \boldsymbol{P}\{X = x_k\} = F_X(x_k) - F_X(x_k^-) = F_X(x_k) - F_X(x_{k-1})} \tag{3.7}$$

Nel seguito esamineremo alcune importanti leggi discrete di *v-a* che assumono solo valori interi $0, 1, 2, \ldots$ e mostreremo anche le loro rappresentazioni grafiche mediante **grafici a barre** consistenti in diagrammi nei quali ad ogni x_k viene semplicemente associata una *barra* verticale di altezza pari a p_k. Si noti che in realtà ogni esempio tratterà non una sola legge, ma una intera *famiglia di leggi* caratterizzate da distribuzioni che differiscono fra loro per il valore di uno o più parametri: ad esempio le leggi Binomiali $\mathfrak{B}(n; p)$ sono classificate dai due parametri $n \in \boldsymbol{N}$ e $p \in [0, 1]$; le leggi di Poisson $\mathfrak{P}(\lambda)$ sono invece classificate da un solo parametro $\lambda > 0$, e così via. La stessa osservazione si applicherà al caso delle leggi di *v-a* continue nella successiva Sezione 3.4

Definizione 3.15. *Chiameremo* ***moda (o mode)*** *di una* ***legge discreta*** *il valore (i valori) x_k in cui il grafico a barre presenta un massimo (o dei massimi relativi): si tratta quindi del valore (dei valori) più probabile (i). In caso di ambiguità (barre adiacenti della stessa altezza) si sceglie convenzionalmente un valore intermedio*

3.3.1 Legge degenere

Si dice che una *v-a* X è degenere quando essa assume con probabilità 1 sempre lo stesso valore $b \in \boldsymbol{R}$, cioè quando la X praticamente non è aleatoria dato che

$$\boldsymbol{P}\{X = b\} = 1$$

La corrispondente distribuzione P_X assegnerà quindi probabilità 1 a tutti i sottoinsiemi di $\boldsymbol{R}$ che contengono b, e 0 a tutti gli altri. Al variare del parametro $b \in \boldsymbol{R}$ queste leggi costituiscono una famiglia che indicheremo con il simbolo $\boldsymbol{\delta}_b$ e scriveremo anche $X \sim \boldsymbol{\delta}_b$. La *FDC* $F_X(x)$ in questo caso è costituita da un unico gradino alto 1 e posizionato in $x = b$, mentre il grafico a barre si riduce ad un'unica barra di altezza 1 in $x = b$

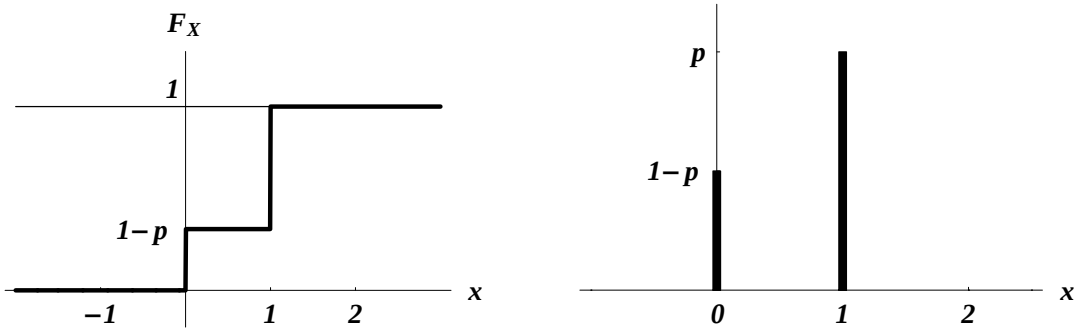


Figura 3.2: *FDC* e grafico a barre di una legge di Bernoulli. Nel caso in figura la moda è 1

3.3.2 Legge di Bernoulli

Si dice che una *v-a* X è distribuita secondo la legge di Bernoulli $\mathfrak{B}(1;p)$, e scriveremo anche $X \sim \mathfrak{B}(1;p)$, quando essa assume i seguenti valori

$$X = \begin{cases} 1 & \text{con probabilità } p, \\ 0 & \text{con probabilità } 1-p. \end{cases}$$

con $0 \le p \le 1$. In altri termini si ha

$$\boxed{p_0 = \boldsymbol{P}\{X=0\} = 1-p\,, \qquad p_1 = \boldsymbol{P}\{X=1\} = p\,.} \tag{3.8}$$

È evidente che ogni indicatore I_A di un evento A con $\boldsymbol{P}\{A\} = p$ è una *v-a* di Bernoulli $\mathfrak{B}(1;p)$: infatti

$$\boldsymbol{P}\{I_A=0\} = \boldsymbol{P}\{\overline{A}\} = 1-p\,, \qquad \boldsymbol{P}\{I_A=1\} = \boldsymbol{P}\{A\} = p\,.$$

Nella Figura 3.2 è mostrato prima di tutto il grafico della *FDC* di una legge di Bernoulli $\mathfrak{B}(1;p)$: esso presenta due discontinuità in 0 e 1; inoltre le altezze dei due salti coincidono proprio con le probabilità $1-p$ e p che X prenda rispettivamente i valori 0 e 1. Sempre nella Figura 3.2 sono poi direttamente rappresentati in un grafico a barre i valori delle probabilità p e $1-p$ che X prenda rispettivamente i valori 1 e 0. Nell'esempio considerato la moda è 1

3.3.3 Legge binomiale

Diremo che una *v-a* X segue la legge binomiale $\mathfrak{B}(n;p)$ con $n = 1, 2, \ldots$ e $p \ge 0$, e scriveremo anche $X \sim \mathfrak{B}(n;p)$, quando essa assume gli $n+1$ valori $0, 1, \ldots, n$ con le seguenti probabilità

$$\boxed{p_k = \boldsymbol{P}\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}\,, \qquad k = 0, 1, \ldots, n} \tag{3.9}$$

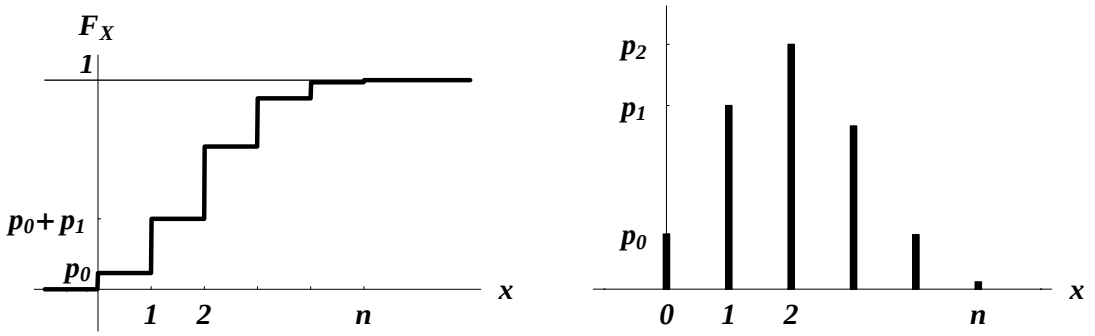


Figura 3.3: *FDC* e grafico a barre di una legge binomiale $\mathfrak{B}(n;p)$. Nel caso in figura la moda è 2

È facile verificare che le p_k in (3.9) definiscono correttamente una distribuzione perché rispettano le condizioni (3.6) quale che sia il valore di n e p: infatti dalla formula di Newton per la potenza di un binomio si ha

$$\sum_{k=0}^{n} p_k = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p+1-p)^n = 1$$

Si verifica anche facilmente che la legge di Bernoulli $\mathfrak{B}(1;p)$ della Sezione 3.3.2 non è altro (come peraltro rivela la notazione adottata) che una legge binomiale nel caso di $n = 1$

Nella Figura 3.3 è rappresentato il grafico della *FDC* della legge $\mathfrak{B}(n;p)$: esso presenta $n+1$ discontinuità nei punti $0, 1, \ldots, n$, rimane costante fra due successive discontinuità, vale 0 per $x < 0$ e 1 per $x \geq n$; inoltre l'altezza di ogni salto nel punto k coincide con la probabilità p_k. Nella medesima figura è rappresentato anche il grafico a barre dei valori p_k: per l'esempio considerato la moda è 2. L'andamento di questi grafici ovviamente cambia al variare dei valori di n e p: in particolare il grafico a barre è simmetrico quando $p = {}^1/_2$; viceversa, quando p è prossimo a 1 (rispettivamente a 0), le p_k più grandi si spostano verso i valori più elevati (rispettivamente meno elevati) di k

Sarà importante osservare a questo punto che la legge (3.9) coincide con quella già introdotta nella (2.4) del Teorema 2.7. Infatti l'evento B introdotto in quella occasione altro non è che un evento del tipo $\{X = k\}$ se X conta il numero delle palline bianche trovate in n estrazioni indipendenti e con rimessa: il Teorema (2.7) in altri termini afferma che $X \sim \mathfrak{B}(n;p)$. Per dare veste più generale a questo risultato, consideriamo un esperimento consistente in n tentativi indipendenti di verifica di un dato evento A che ogni volta si realizza con $\boldsymbol{P}\{A\} = p$: potremo allora definire da un lato n *v-a* di Bernoulli $\mathfrak{B}(1;p)$ indipendenti $X_k,\ k = 1, \ldots, n$ che assumono valore 1 se A si verifica nel tentativo k-mo e 0 in caso contrario; e dall'altro la *v-a* X che rappresenta il numero di successi sugli n tentativi. È intuitivo che fra queste *v-a* sussista la relazione $X = X_1 + \ldots + X_n$ mentre la relazione fra le loro leggi è oggetto del seguente Teorema

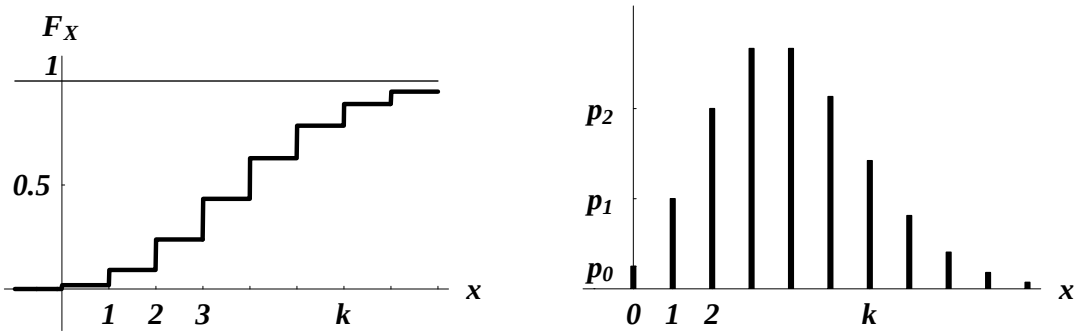


Figura 3.4: *FDC* e grafico a barre di una legge di Poisson $\mathfrak{P}(\lambda)$.

Teorema 3.16. *Se n v-a $X_1, \ldots, X_n$ sono iid con legge di Bernoulli $\mathfrak{B}(1;p)$, la loro somma $X = X_1 + \ldots + X_n$ sarà distribuita con legge binomiale $\mathfrak{B}(n;p)$. Viceversa, ogni legge binomiale $\mathfrak{B}(n;p)$ è la distribuzione di una v-a X somma di n v-a $X_1, \ldots, X_n$ iid con leggi di Bernoulli $\mathfrak{B}(1;p)$*

Dimostrazione: Omessa □

3.3.4 Legge di Poisson

Diremo che una *v-a* X segue la legge di Poisson $\mathfrak{P}(\lambda)$ con $\lambda > 0$, e scriveremo anche $X \sim \mathfrak{P}(\lambda)$, quando essa assume tutti i valori interi $k \in \boldsymbol{N}$ con le seguenti probabilità

$$\boxed{p_k = \boldsymbol{P}\{X = k\} = \frac{\lambda^k}{k!}\, e^{-\lambda}, \qquad k = 0, 1, 2, \ldots} \tag{3.10}$$

È facile verificare anche in questo caso che la somma di queste infinite p_k vale esattamente 1 quale che sia il valore di λ: infatti dal ben noto sviluppo in serie di Taylor dell'esponenziale e^λ si ha

$$\sum_{k=0}^{\infty} p_k = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda}\, e^{\lambda} = 1$$

Nella Figura 3.4 è rappresentato il grafico della *FDC* della legge $\mathfrak{P}(\lambda)$: esso presenta infinite discontinuità nei punti $0, 1, 2, \ldots$, rimane costante fra due successive discontinuità, vale 0 per $x < 0$ e tende asintoticamente verso 1 per $x \to +\infty$; inoltre l'altezza di ogni salto nei punti k coincide con la probabilità p_k di (3.10). Nella medesima figura è rappresentato anche il grafico a barre di alcuni dei valori p_k, ma per l'esempio rappresentato la moda è ambigua perché ci sono due valori (3 e 4) con la stessa probabilità: in questo caso si può prendere convenzionalmente come moda il valore 3.5. L'andamento di questi grafici ovviamente cambia al variare del valore di λ: in particolare al crescere di λ il massimo del grafico a barre si sposta verso

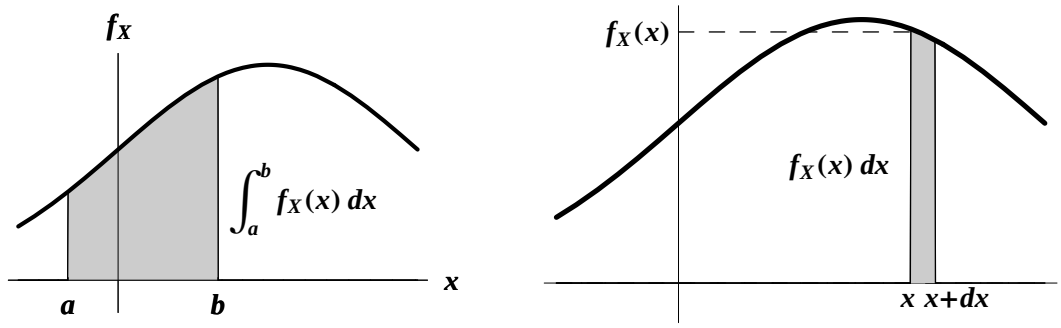


Figura 3.5: L'area fra a e b al di sotto della curva $f_X(x)$ è la probabilità che X assuma valori fra a e b (vedi equazione (3.13)); inoltre $f_X(x)\,dx$ rappresenta la probabilità infinitesima che X stia nell'intervallo $[x, x+dx]$.

valori più elevati di k. La legge di Poisson è particolarmente adatta a descrivere *v-a* che rappresentano conteggi e che possono assumere un numero illimitato di valori: numero di telefonate che arrivano ad un centralino in un dato periodo di tempo; numero di clienti che si presentano allo sportello di un ufficio durante una giornata; numero di stelle presenti in una determinata regione di cielo. Il motivo per cui questo avviene è chiarito nel successivo capitolo sui Teoremi Limite dal Teorema 5.8 e dalla discussione dell'Esempio 5.10.

3.4 Leggi continue

Definizione 3.17. *Diremo che X è una **v-a continua** se essa assume tutti i valori di un intervallo di numeri reali (non escluso l'intero insieme $\boldsymbol{R}$), e se esiste una **funzione di densità di probabilità (fdp)** $f_X(x)$, cioè una funzione definita su $\boldsymbol{R}$ con*

$$f_X(x) \geq 0 \qquad \int_{-\infty}^{+\infty} f_X(x)\,dx = 1 \tag{3.11}$$

e tale che

$$\boxed{F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt \qquad f_X(x) = F_X'(x)} \tag{3.12}$$

*dove $F_X(x)$ è la FDC di X. Le leggi delle v-a continue si chiamano **leggi continue***

Abbiamo visto nella Sezione 3.3 che la legge delle *v-a* discrete è determinata dall'associazione ai valori x_k di numeri p_k che soddisfino le proprietà (3.6). Nel caso di *v-a* continue, invece, questa procedura elementare non è più praticabile e bisogna passare invece all'uso degli strumenti del calcolo differenziale e integrale. Più precisamente la Definizione 3.17 chiarisce che la legge di una *v-a* continua X è caratterizzata dall'assegnazione di una opportuna *fdp* $f_X(x)$ mediante la quale è possibile eseguire tutti i calcoli necessari. In particolare le relazioni fra la *fdp* f_X e la *FDC*

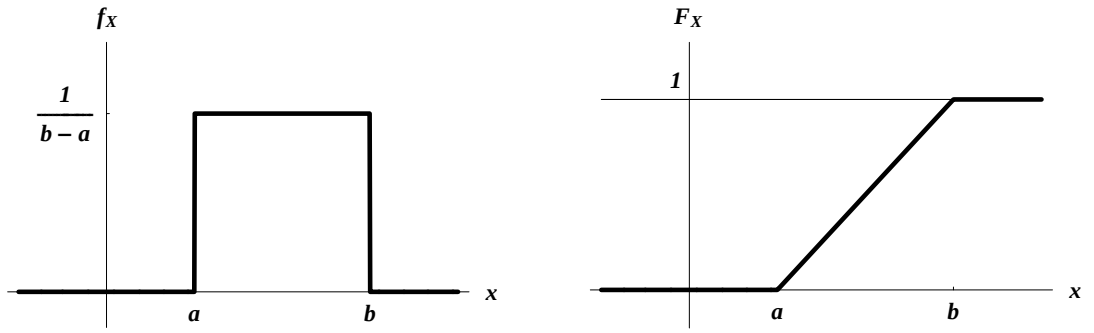


Figura 3.6: fd e *FDC* della legge Uniforme $\mathfrak{U}(a,b)$.

F_X sono riassunte in (3.12), cioè F_X è la primitiva di f_X, e quindi a sua volta f_X è la derivata di F_X. Ne deriva pertanto che, comunque scelto un intervallo $[a,b]$ (vedi Figura 3.5) tenendo conto di (3.4) e di (3.12) risulterà

$$\boxed{\boldsymbol{P}\{a \le X \le b\} = P_X\big([a,b]\big) = F_X(b) - F_X(a) = \int_a^b f_X(x)\,dx} \tag{3.13}$$

Si noti però che – a differenza da quanto avviene nel caso generale (3.4) – per le leggi continue non è più necessario precisare se l'intervallo considerato è *chiuso o aperto*. Infatti per *v-a* continue la *FDC* (3.12) è sempre una funzione continua in ogni x, e quindi da (3.5) deriva che la probabilità di assumere un singolo valore è rigorosamente zero: pertanto aggiungere o togliere gli estremi a e b all'intervallo di integrazione in (3.13) non cambia il risultato. D'altra parte è bene osservare che il valore non nullo di $f_X(x)$ in x *non rappresenta affatto la probabilità che la v-a* X *assuma il valore* x: si potrebbe infatti far vedere con dei banali esempi che una fd può assumere anche valori maggiori di 1, e quindi non può in nessun modo essere una probabilità. Piuttosto è la quantità infinitesima $f_X(x)\,dx$ che può essere interpretata come la probabilità (infinitesima) che X prenda valori nell'intervallo infinitesimo $[x, x+dx]$ (vedi Figura 3.5)

Il calcolo delle probabilità con l'integrale in (3.13) non è sempre un'operazione elementare: in mancanza di opportuni strumenti di calcolo si usano delle apposite Tavole (vedi Appendice E) nelle quali sono elencati i valori delle *FDC* F_X per le leggi più usuali. Il calcolo di $\boldsymbol{P}\{a \le X \le b\}$ potrà allora essere effettuato mediante la differenza $F_X(b) - F_X(a)$

Definizione 3.18. *Chiameremo* ***moda (o mode)*** *di una* ***legge continua*** *il valore (o i valori) di x in cui la sua fdp assume il valore massimo (o i valori massimi relativi). In caso di ambiguità (valori massimi assunti su un intervallo) si sceglie convenzionalmente un punto intermedio*

3.4.1 Legge uniforme

Diremo che una *v-a* X segue una legge $\mathfrak{U}(a,b)$ uniforme nell'intervallo $[a,b]$ (con $a,b \in \boldsymbol{R}$), e scriveremo anche $X \sim \mathfrak{U}(a,b)$, se essa è caratterizzata dalla *fdp*

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{se } a \le x \le b, \\ 0 & \text{altrimenti.} \end{cases} \tag{3.14}$$

La *FDC* si calcola poi in maniera elementare:

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

Queste due funzioni sono rappresentate nella Figura 3.6. Ovviamente le relazioni (3.11) sono sempre soddisfatte dato che l'area di un rettangolo di base $b-a$ e altezza $\frac{1}{b-a}$ è sempre 1. Si vede inoltre dall'equazione (3.13) che con c e Δ tali che $a \le c \le c+\Delta \le b$ si ha $\boldsymbol{P}\{c \le X \le c+\Delta\} = \frac{\Delta}{b-a}$ indipendentemente dal valore di c; pertanto a intervalli di larghezza Δ interni ad $[a,b]$ viene attribuita sempre la stessa probabilità $\frac{\Delta}{b-a}$ indipendentemente dalla loro collocazione: questo è in definitiva il significato della *uniformità* della distribuzione. Dalla Figura 3.6 si può anche vedere che la collocazione della moda è ambigua perché la *fdp* assume valori massimi in tutto l'intervallo $[a,b]$: in questo caso si può convenzionalmente assumere che la moda sia il punto di mezzo $\frac{a+b}{2}$

3.4.2 Legge normale o Gaussiana

Diremo che una *v-a* X segue una legge normale o Gaussiana $\mathfrak{N}(\mu,\sigma^2)$ con $\mu \in \boldsymbol{R}$ e $\sigma > 0$, e scriveremo anche $X \sim \mathfrak{N}(\mu,\sigma^2)$, se essa è caratterizzata dalla *fdp*

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2} \tag{3.15}$$

Queste *fdp* soddisfano le relazioni (3.11) per ogni valore di μ e σ, ma noi non lo verificheremo; ci limiteremo invece a dare una descrizione qualitativa del comportamento di queste funzioni che sono rappresentate nella Figura 3.7. La *fdp* di una normale $\mathfrak{N}(\mu,\sigma^2)$ è una curva a campana, simmetrica attorno ad un massimo nel punto $x = \mu$ (moda). La funzione va rapidamente verso zero allontanandosi dal centro della curva e la larghezza della campana è regolata dal valore di σ: grandi valori di σ corrispondono a curve larghe e piatte; piccoli valori di σ corrispondono a curve strette e alte. Si può mostrare che la curva presenta due flessi proprio in $x = \mu \pm \sigma$. La *FDC*

$$F_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2}\, dt \tag{3.16}$$

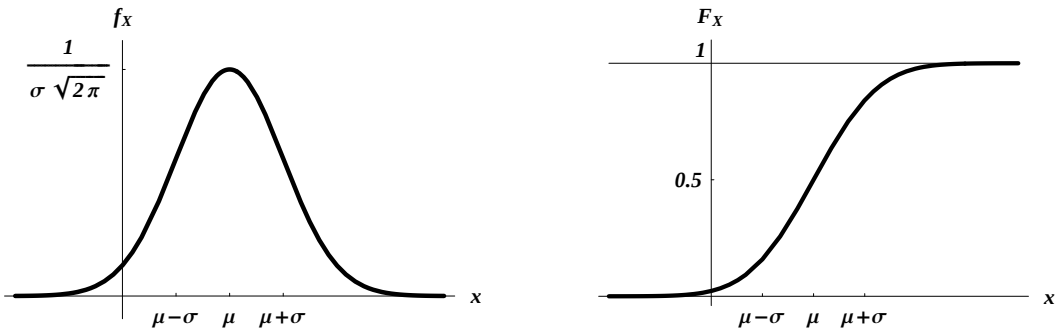


Figura 3.7: *fdp* e *FDC* della legge Normale $\mathfrak{N}(\mu, \sigma^2)$. La moda di queste leggi è μ

non ha un'espressione analitica elementare, ma il suo grafico è molto semplice, regolare e tipico delle *FDC*: ha una forma di S allungata che varia da 0 verso 1 con un punto di flesso in $x = \mu$. La *FDC* di una normale diviene sempre più ripida (e al limite approssima un gradino di altezza 1) quando $\sigma \to 0$; viceversa si allunga sempre di più con il crescere di σ

La legge $\mathfrak{N}(0, 1)$ la cui *fdp* è

$$\boxed{\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}} \tag{3.17}$$

è detta anche **legge normale standard** e riveste una importanza particolare perchè, come vedremo, il calcolo delle probabilità relative a leggi normali generiche può sempre essere facilmente ricondotto all'uso delle Tavole dell'Appendice E per la *FDC* della legge normale standard

$$\boxed{\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}}\, dt} \tag{3.18}$$

La *fdp* e la *FDC* della normale standard presentano lo stesso andamento qualitativo di Figura 3.7, ma la moda si trova in $x = 0$ e i due flessi in $x = \pm 1$. Inoltre, data la evidente simmetria di queste due funzioni attorno a $x = 0$ è anche facile verificare che

$$\boxed{\varphi(-x) = \varphi(x) \qquad \Phi(-x) = 1 - \Phi(x)} \tag{3.19}$$

Teorema 3.19.

1. *Se* $X \sim \mathfrak{N}(\mu, \sigma^2)$*, e se* a, b *sono due numeri, allora*

$$aX + b \sim \mathfrak{N}(a\mu + b,\, a^2\sigma^2)$$

2. *Se* $X \sim \mathfrak{N}(\mu, \sigma^2)$ *e* $Y \sim \mathfrak{N}(\nu, \tau^2)$ *sono v-a indipendenti, allora*

$$X + Y \sim \mathfrak{N}(\mu + \nu,\, \sigma^2 + \tau^2)$$

3. Se $X \sim \mathfrak{N}(\mu, \sigma^2)$, se a, b sono due numeri e $\Phi(x)$ è la FDC standard, allora

$$\boxed{\boldsymbol{P}\{a \leq X \leq b\} = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)} \tag{3.20}$$

Dimostrazione: Omettendo la dimostrazione dei punti 1 e 2 ci limiteremo a discutere solo la formula (3.20) del punto 3: se $X \sim \mathfrak{N}(\mu, \sigma^2)$, dal punto 1 sappiamo in particolare che

$$Y = \frac{X-\mu}{\sigma} \sim \mathfrak{N}(0,1)$$

e quindi anche $X = \sigma Y + \mu$. D'altra parte, dati i numeri a e b avremo

$$\boldsymbol{P}\{a \leq X \leq b\} = \boldsymbol{P}\{a \leq \sigma Y + \mu \leq b\} = \boldsymbol{P}\left\{\frac{a-\mu}{\sigma} \leq Y \leq \frac{b-\mu}{\sigma}\right\}$$

ed essendo $Y \sim \mathfrak{N}(0,1)$, da (3.13) e (3.18) si ricava immediatamente la (3.20) □

In base a questo Teorema, dunque, potremo sempre trasformare una nell'altra una *v-a* normale standard $Y \sim \mathfrak{N}(0,1)$, e una normale non standard $X \sim \mathfrak{N}(\mu, \sigma^2)$ mediante le relazioni

$$Y = \frac{X-\mu}{\sigma} \sim \mathfrak{N}(0,1) \qquad\qquad X = \sigma Y + \mu \sim \mathfrak{N}(\mu, \sigma^2)$$

e conseguentemente le probabilità $\boldsymbol{P}\{a \leq X \leq b\}$ potranno sempre essere calcolate mediante la (3.20): un calcolo che si riduce alla consultazione delle Tavole della *FDC* normale standard riportate nell'Appendice E. L'uso di queste Tavole sarà molto utile anche per le altre leggi che introdurremo nel seguito

Si noti che nel seguito, per indicare la probabilità che una *v-a* X cada in un certo intervallo $[a, b]$, useremo a volte il simbolo della sua legge invece che quello della *v-a*: così, ad esempio, il precedente risultato (3.20) può essere equivalentemente espresso nella forma

$$\boldsymbol{P}\big\{a \leq \mathfrak{N}(\mu, \sigma^2) \leq b\big\} = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$

facendo riferimento solo alla legge e non alla *v-a* X

3.4.3 Leggi del *chi-quadro*, di Student e di Fisher

Diremo che una *v-a* X segue una legge del ***chi-quadro*** $\boldsymbol{\chi^2}(n)$ con $n \in \boldsymbol{N}$ gradi di libertà, e scriveremo anche $X \sim \boldsymbol{\chi^2}(n)$, se essa è caratterizzata dalla *fdp*

$$f_X(x) = \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{A_n} \qquad n = 1, 2, 3, \ldots \tag{3.21}$$

dove le costanti A_n valgono

$$A_1 = \sqrt{2\pi} \qquad A_2 = 2 \qquad A_n = \begin{cases} (n-2)!!\,\sqrt{2\pi} & \text{per } n = 3, 5, 7, \ldots \\ (n-2)!!\,2 & \text{per } n = 4, 6, 8, \ldots \end{cases}$$

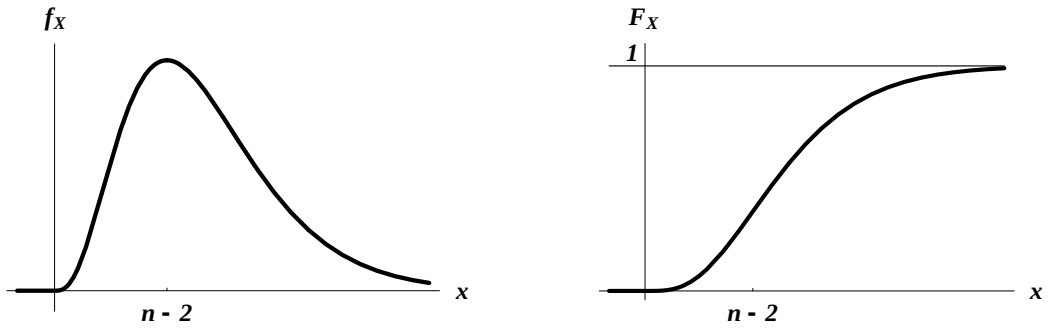


Figura 3.8: *fdp* e *FDC* della legge del *chi-quadro* $\boldsymbol{\chi^2}(n)$ con $n > 2$.

e abbiamo usato la notazione del *doppio fattoriale*

$$k!! = \begin{cases} k(k-2)\dots 5\cdot 3\cdot 1 & \text{per } k \text{ dispari} \\ k(k-2)\dots 6\cdot 4\cdot 2 & \text{per } k \text{ pari} \end{cases} \qquad \text{con} \quad 0!! = 1$$

I grafici di questa *fdp* e della sua *FDC* del tipo mostrato nella Figura 3.8 quando $n > 2$: invece per $n = 1$ la *fdp* diverge per $x \to 0^+$, mentre per $n = 2$ si ha $f_X(0) = {}^1/_2$. La *fdp* è diversa da zero solo per $x \geq 0$, mentre è rigorosamente nulla per $x < 0$; sul semiasse reale positivo il grafico è asimmetrico e presenta una una lunga coda che si annulla asintoticamente per $x \to +\infty$. La moda si trova in $x = n - 2$, e tende ad allontanarsi dall'origine per n crescenti. I valori della *FDC* di $\boldsymbol{\chi^2}(n)$ possono essere trovati sulle opportune Tavole nell'Appendice E e saranno usati per il calcolo delle probabilità (3.13)

Teorema 3.20. *Se $X_1, \dots, X_n$ sono v-a iid tutte normali standard $\mathfrak{N}(0,1)$, allora*

$$Z = X_1^2 + \dots + X_n^2 \sim \boldsymbol{\chi^2}(n)$$

cioè la v-a Z, somma dei quadrati di n normali standard indipendenti, segue la legge del chi-quadro con n gradi di libertà

Dimostrazione: Omessa □

Diremo che una *v-a* X segue una legge di **Student** $\mathfrak{T}(n)$ con $n \in \boldsymbol{N}$ gradi di libertà, e scriveremo anche $X \sim \mathfrak{T}(n)$, se essa è caratterizzata dalla *fdp*

$$f_X(x) = B_n \left(\frac{n}{n+x^2}\right)^{\frac{n+1}{2}} \qquad n = 1, 2, 3, \dots \tag{3.22}$$

dove le costanti B_n valgono

$$B_1 = \frac{1}{\pi} \qquad B_2 = \frac{1}{2\sqrt{2}} \qquad B_n \begin{cases} \frac{(n-1)!!}{(n-2)!!\,\pi\sqrt{n}} & \text{per } n = 3, 5, 7, \dots \\ \frac{(n-1)!!}{(n-2)!!\,2\sqrt{n}} & \text{per } n = 4, 6, 8, \dots \end{cases}$$

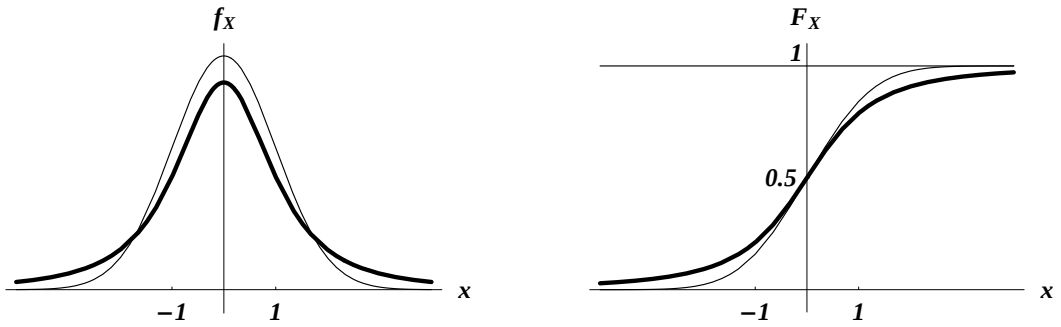


Figura 3.9: *fdp* e *FDC* della legge di Student con n gradi di libertà $\mathfrak{T}(n)$. Le curve più sottili sono la *fdp* e la *FDC* della $\mathfrak{N}(0,1)$ e sono qui riportate per confronto.

Una legge di Student $\mathfrak{T}(n)$ con $n = 1, 2, \ldots$ gradi di libertà ha una *fdp* e una *FDC* del tipo mostrato nella Figura 3.9. La *fdp* di $\mathfrak{T}(n)$ è una funzione a campana, simmetrica attorno alla moda in $x = 0$, e per alcuni versi simile alla $\mathfrak{N}(0,1)$. Come si vede dalla (3.22) e dalla Figura 3.9 la *fdp* di $\mathfrak{T}(n)$ si annulla per $x \to \pm\infty$ più lentamente della *fdp* della $\mathfrak{N}(0,1)$. Quando però il valore di n cresce la *fdp* della legge $\mathfrak{T}(n)$ diviene sempre più simile alla *fdp* normale standard, e al limite le due funzioni coincidono. I valori della *FDC* di $\mathfrak{T}(n)$ possono essere trovati sulle Tavole nell'Appendice E e saranno usati nel calcolo tramite l'equazione (3.13).

Teorema 3.21. *Se* $X \sim \mathfrak{N}(0,1)$ *e* $Z \sim \boldsymbol{\chi}^2(n)$ *sono v-a indipendenti, allora*

$$T = \frac{X}{\sqrt{Z/n}} \sim \mathfrak{T}(n)$$

cioè la v-a T *segue la legge di Student* $\mathfrak{T}(n)$ *con* n *gradi di libertà*

Dimostrazione: Omessa □

Diremo che una *v-a* X segue una legge di **Fisher** $\mathfrak{F}(n,m)$ con $n, m \in \boldsymbol{N}$ gradi di libertà, e scriveremo $X \sim \mathfrak{F}(n,m)$, se essa è caratterizzata dalla *fdp*

$$f_X(x) = C_{n,m} \frac{x^{\frac{n-2}{2}}}{(m+nx)^{\frac{n+m}{2}}} \qquad n, m = 1, 2, 3, \ldots \tag{3.23}$$

dove le costanti $C_{n,m}$ valgono

$$C_{1,1} = \frac{1}{\pi}, \quad C_{1,2} = C_{2,1} = 1, \quad C_{n,m} = \frac{(n+m-2)!!\, n^{\frac{n}{2}} m^{\frac{m}{2}}}{(n-2)!!(m-2)!!} \begin{cases} 1/\pi & n, m = 3, 5, 7 \ldots \\ 1/2 & \text{in altri casi} \end{cases}$$

e una *FDC* del tipo mostrato nella Figura 3.10. La *fdp* di $\mathfrak{F}(n,m)$ somiglia a quella di una $\boldsymbol{\chi}^2(n)$: essa è diversa da zero solo per $x \geq 0$ mentre è rigorosamente nulla per $x < 0$; sul semiasse reale positivo il grafico è asimmetrico e presenta una una

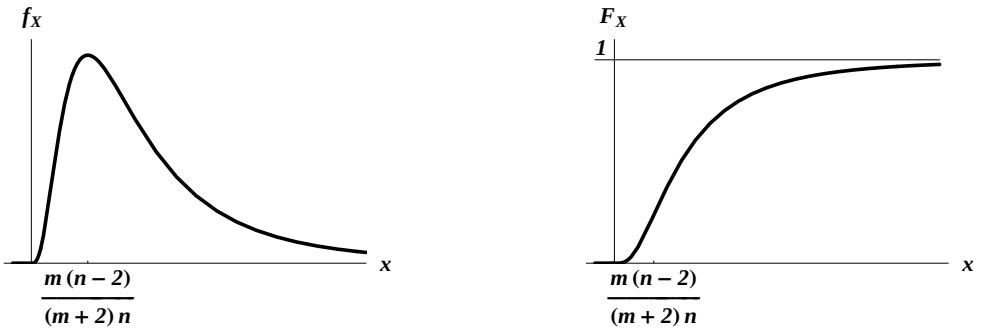


Figura 3.10: *fdp* e *FDC* della legge di Fisher con n ed m gradi di libertà $\mathfrak{F}(n,m)$.

lunga coda che si annulla asintoticamente per $x \to +\infty$. La moda si trova nel punto $\frac{m(n-2)}{(m+2)n}$. I valori della *FDC* di $\mathfrak{F}(n,m)$ possono essere trovati sulle Tavole nell'Appendice E e saranno usati nel calcolo tramite l'equazione (3.13).

Teorema 3.22. *Se* $Z \sim \boldsymbol{\chi}^{\mathbf{2}}(n)$ *e* $W \sim \boldsymbol{\chi}^{\mathbf{2}}(m)$ *sono v-a indipendenti, allora*

$$F = \frac{Z/n}{W/m} \sim \mathfrak{F}(n,m)$$

cioè la v-a F *segue la legge di Fisher* $\mathfrak{F}(n,m)$ *con* n, m *gradi di libertà*

Dimostrazione: Omessa. Si noti che il simbolo $\mathfrak{F}(n,m)$ non è simmetrico nei suoi due parametri n e m: è importante sottolineare quindi che in questo teorema il *primo* parametro n della legge di Fisher $\mathfrak{F}(n,m)$ indica sempre il numero di gradi di libertà della *v-a* $\boldsymbol{\chi}^{\mathbf{2}}(n)$ che si trova al *numeratore*, mentre il *secondo* indice m si riferisce sempre alla *v-a* $\boldsymbol{\chi}^{\mathbf{2}}(m)$ al *denominatore* □

Teorema 3.23. *Se* $X_1, \ldots, X_n$ *sono* n *v-a iid tutte normali* $\mathfrak{N}(\mu, \sigma^2)$*, e se poniamo*

$$Y_n = X_1 + \ldots + X_n \qquad \overline{X}_n = \frac{Y_n}{n} \qquad S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 \qquad S_n = \sqrt{S_n^2}$$

allora

$$\begin{aligned} Y_n^* &\equiv \frac{Y_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} \;\sim\; \mathfrak{N}(0,1) \\ Z_n &\equiv (n-1)\,\frac{S_n^2}{\sigma^2} = \sum_{k=1}^{n}\left(\frac{X_k - \overline{X}_n}{\sigma}\right)^2 \;\sim\; \boldsymbol{\chi}^{\mathbf{2}}(n-1) \\ T_n &\equiv \frac{Y_n - n\mu}{S_n\sqrt{n}} = \frac{\overline{X}_n - \mu}{S_n}\sqrt{n} \;\sim\; \mathfrak{T}(n-1) \end{aligned}$$

cioè le v-a Y_n^*, Z_n, T_n *definite come sopra seguono rispettivamente le leggi normale standard* $\mathfrak{N}(0,1)$*, del chi-quadro* $\boldsymbol{\chi}^{\mathbf{2}}(n-1)$ *e di Student* $\mathfrak{T}(n-1)$

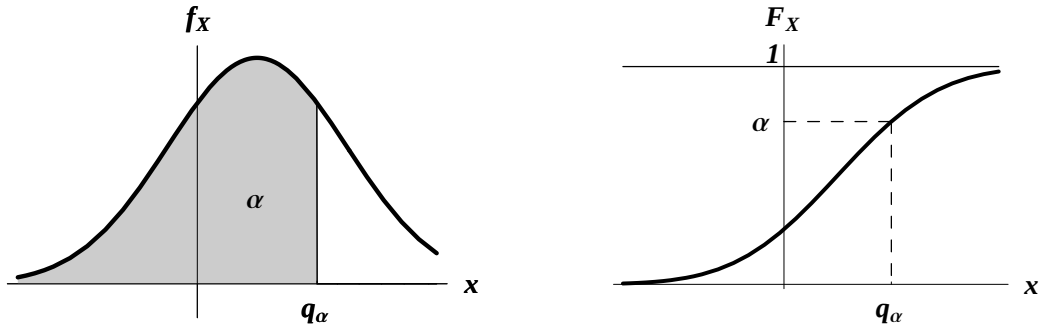


Figura 3.11: Quantile q_α di ordine α di una distribuzione con *fdp* f_X e *FDC* F_X .

Dimostrazione: Per semplicità dimostreremo solo che $Y_n^* \sim \mathfrak{N}(0,1)$: siccome le X_k sono tutte *iid* con legge $\mathfrak{N}(\mu, \sigma^2)$, da un'applicazione ripetuta del punto 2 del Teorema (3.19) si ha innanzitutto che

$$Y_n = X_1 + \ldots + X_n \sim \mathfrak{N}(n\mu, n\sigma^2)$$

e quindi utilizzando (con una opportuna identificazione dei simboli) il punto 1 del medesimo teorema

$$Y_n^* = \frac{Y_n - n\mu}{\sigma\sqrt{n}} = \frac{1}{\sigma\sqrt{n}}\, Y_n - \frac{\mu\sqrt{n}}{\sigma} \sim \mathfrak{N}(0,1)$$

Omettiamo invece la dimostrazione degli altri due risultati □

3.5 Quantili

In questa sezione, per semplificare definizioni e notazioni, supporremo sempre che X sia una *v-a* continua con *fdp* $f_X(x)$ e *FDC* $F_X(x)$: inoltre – come si vede dagli esempi della Sezione 3.4 – nei casi di nostro interesse F_X è una funzione strettamente crescente su tutto $\boldsymbol{R}$ (leggi normale e di Sudent), o almeno sul semiasse $x \geq 0$ (leggi del chi quadro e di Fisher). Preso allora un numero $0 < \alpha < 1$ uno sguardo ai grafici delle Figure1 3.7– 3.10 ci convincerà del fatto che esiste sempre una e una sola soluzione dell'equazione

$$F_X(x) = \alpha \tag{3.24}$$

In tutti questi casi avrà allora un senso preciso introdurre la seguente definizione

Definizione 3.24. *Data una v-a X continua con fdp $f_X(x)$ e FDC $F_X(x)$, chiameremo* ***quantile di ordine*** *α il numero q_α soluzione dell'equazione (3.24), cioè tale che $F_X(q_\alpha) = \alpha$. Il quantile di ordine $\alpha = {}^1/_2$ si chiama* ***mediana****; i quantili di ordine $\alpha = {}^k/_4$ con $k = 1, 2, 3$ si chiamano* ***quartili****; i quantili di ordine $\alpha = {}^k/_{10}$ con $k = 1, \ldots, 10$ si chiamano* ***decili****, e così via*

Il significato di questa definizione è illustrato nella Figura 3.11: nella immagine a destra è rappresentato il fatto che $F_X(q_\alpha) = \alpha$; in quella a sinistra, invece, è riportata la *fdp* di X, e il significato del quantile q_α è quello del punto che lascia alla sua sinistra un'area sotto la curva pari ad α. Questo è ovviamente coerente con la Definizione 3.24 dato che dalla Definizione 3.9 e dall'equazione (3.12) si ha

$$\alpha = F_X(q_\alpha) = \boldsymbol{P}\{X \le q_\alpha\} = \int_{-\infty}^{q_\alpha} f_X(x)\,dx\,.$$

In questo quadro la mediana $q_{1/2}$, che divide esattamente a metà l'area sotto la *fdp*, rappresenta un primo **indicatore di centralità** della distribuzione, mentre invece la differenza fra il primo e il terzo quartile $q_{3/4} - q_{1/4}$ costituisce un **indicatore di dispersione**

Naturalmente, siccome la *fdp* è una funzione monotòna crescente, anche i quantili q_α saranno monotòni crescenti nella variabile $\alpha \in [0,1]$. Nel seguito indicheremo con delle notazioni specifiche i quantili delle leggi più usate: così φ_α sarà il quantile di ordine α della legge normale standard $\mathfrak{N}(0,1)$; $\chi^2_\alpha(n)$ quello della legge del *chi-quadro* con n gradi di libertà $\boldsymbol{\chi^2}(n)$; $t_\alpha(n)$ quello della legge di Student con n gradi di libertà $\mathfrak{T}(n)$; e infine $f_\alpha(n,m)$ quello della legge di Fisher con n e m gradi di libertà $\mathfrak{F}(n,m)$

Teorema 3.25. *Se* $X \sim \mathfrak{N}(0,1)$*, allora i suoi quantili soddisfano le regole di simmetria*

$$\boxed{\varphi_{\frac{\alpha}{2}} = -\,\varphi_{1-\frac{\alpha}{2}}} \tag{3.25}$$

inoltre con $0 \le \alpha \le 1$ *risulterà sempre* $\varphi_{\frac{\alpha}{2}} \le 0 \le \varphi_{1-\frac{\alpha}{2}}$ *e sarà verificata la relazione*

$$\boxed{\boldsymbol{P}\big\{|X| \le \varphi_{1-\frac{\alpha}{2}}\big\} = \boldsymbol{P}\big\{\varphi_{\frac{\alpha}{2}} \le X \le \varphi_{1-\frac{\alpha}{2}}\big\} = 1-\alpha} \tag{3.26}$$

Dimostrazione: La *fdp* normale standard $\varphi(x)$ definita in (3.17) è simmetrica attorno a $x = 0$ con $\varphi(-x) = \varphi(x)$, come messo in evidenza nella Figura 3.12. Pertanto i quantili $\varphi_{\frac{\alpha}{2}}$ e $\varphi_{1-\frac{\alpha}{2}}$, che delimitano verso l'esterno due code di uguale probabilità $\frac{\alpha}{2}$ (area grigia), saranno collocati in posizioni diametralmente opposte rispetto a $x = 0$, e quindi soddisfano la (3.25). Se poi $0 \le \alpha \le 1$ si ha anche che $\varphi_{\frac{\alpha}{2}}$ sarà sempre negativo e $\varphi_{1-\frac{\alpha}{2}}$ sempre positivo. Da tutto questo deriva che potremo scrivere

$$\begin{aligned}\big\{X \le \varphi_{1-\frac{\alpha}{2}}\big\} &= \big\{X \le \varphi_{\frac{\alpha}{2}}\big\} \cup \big\{\varphi_{\frac{\alpha}{2}} \le X \le \varphi_{1-\frac{\alpha}{2}}\big\} \\ &= \big\{X \le \varphi_{\frac{\alpha}{2}}\big\} \cup \big\{-\varphi_{1-\frac{\alpha}{2}} \le X \le \varphi_{1-\frac{\alpha}{2}}\big\} = \big\{X \le \varphi_{\frac{\alpha}{2}}\big\} \cup \big\{|X| \le \varphi_{1-\frac{\alpha}{2}}\big\}\end{aligned}$$

Siccome abbiamo una unione di eventi disgiunti, segue dalla additività (1.3) della probabilità che

$$\boldsymbol{P}\big\{X \le \varphi_{1-\frac{\alpha}{2}}\big\} = \boldsymbol{P}\big\{X \le \varphi_{\frac{\alpha}{2}}\big\} + \boldsymbol{P}\big\{|X| \le \varphi_{1-\frac{\alpha}{2}}\big\}$$

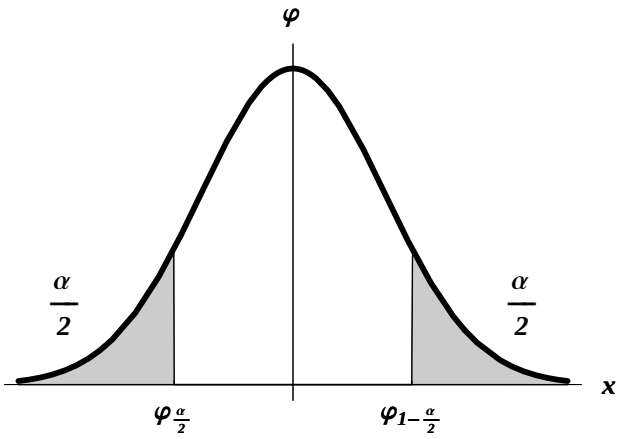


Figura 3.12: Quantili di ordine $\frac{\alpha}{2}$, e $1-\frac{\alpha}{2}$ di una legge normale standard $\mathfrak{N}(0,1)$

da cui si ottiene la (3.26) tenendo anche conto della Definizione 3.24 di quantili:

$$\boldsymbol{P}\big\{|X| \le \varphi_{1-\frac{\alpha}{2}}\big\} = \boldsymbol{P}\big\{X \le \varphi_{1-\frac{\alpha}{2}}\big\} - \boldsymbol{P}\big\{X \le \varphi_{\frac{\alpha}{2}}\big\} = \Big(1-\frac{\alpha}{2}\Big) - \frac{\alpha}{2} = 1-\alpha$$

Ovviamente nella Figura 3.12 questa probabilità corrisponde all'area sotto la *fdp* compresa fra i due quantili □

Teorema 3.26. *Se* $T \sim \mathfrak{T}(n)$*, allora i suoi quantili soddisfano le regole di simmetria*

$$\boxed{t_{\frac{\alpha}{2}}(n) = -\,t_{1-\frac{\alpha}{2}}(n)} \tag{3.27}$$

inoltre con $0 \le \alpha \le 1$ *risulterà sempre* $t_{\frac{\alpha}{2}} \le 0 \le t_{1-\frac{\alpha}{2}}$ *e sarà verificata la relazione*

$$\boxed{\boldsymbol{P}\big\{|T| \le t_{1-\frac{\alpha}{2}}(n)\big\} = \boldsymbol{P}\big\{t_{\frac{\alpha}{2}}(n) \le T \le t_{1-\frac{\alpha}{2}}(n)\big\} = 1-\alpha} \tag{3.28}$$

Infine per grandi valori di n *i quantili di* $\mathfrak{T}(n)$ *possono essere approssimati con quelli della normale standard* $\mathfrak{N}(0,1)$

$$\boxed{t_\alpha(n) \simeq \varphi_\alpha} \tag{3.29}$$

Dimostrazione: Siccome le proprietà di simmetria della *fdp* (3.22) di $\mathfrak{T}(n)$ sono del tutto analoghe a quelle della *fdp* $\varphi(x)$ della $\mathfrak{N}(0,1)$, la dimostrazione è identica a quella del Teorema 3.25 con ovvi adeguamenti della notazione. Omettiamo invece la dimostrazione di (3.29) □

Teorema 3.27. *Se* $Z \sim \boldsymbol{\chi^2}(n)$*, e se* $0 \le \alpha \le 1$*, allora sarà verificata la relazione*

$$\boxed{\boldsymbol{P}\Big\{\chi^2_{\frac{\alpha}{2}}(n) \le Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\Big\} = 1-\alpha} \tag{3.30}$$

Inoltre per grandi valori di n i quantili di $\boldsymbol{\chi^2}(n)$ *possono essere approssimati con*

$$\boxed{\chi^2_\alpha(n) \simeq \frac{\left(\varphi_\alpha + \sqrt{2n-1}\right)^2}{2}} \tag{3.31}$$

dove φ_α *sono i quantili della normale standard* $\mathfrak{N}(0,1)$

Dimostrazione: Le leggi del *chi-quadro* $\boldsymbol{\chi^2}(n)$ – vedi Figura 3.13 – non hanno le stesse proprietà di simmetria di quelle normali o di Student, e pertanto anche i loro quantili non obbediscono più a regole di simmetria e l'equazione (3.30) non può essere scritta in termini di valori assoluti. Ciononostante la sua dimostrazione segue un percorso molto simile a quello della (3.26): infatti, essendo i quantili crescenti con l'ordine, per $0 \le \alpha \le 1$ si ha innanzitutto $\chi^2_{\frac{\alpha}{2}}(n) < \chi^2_{1-\frac{\alpha}{2}}(n)$, e quindi potremo scrivere

$$\left\{Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\right\} = \left\{Z \le \chi^2_{\frac{\alpha}{2}}(n)\right\} \cup \left\{\chi^2_{\frac{\alpha}{2}}(n) \le Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\right\}$$

con i due eventi del secondo membro disgiunti. Segue allora dalla additività (1.3) della probabilità che

$$\boldsymbol{P}\left\{Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\right\} = \boldsymbol{P}\left\{Z \le \chi^2_{\frac{\alpha}{2}}(n)\right\} + \boldsymbol{P}\left\{\chi^2_{\frac{\alpha}{2}}(n) \le Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\right\}$$

da cui si ottiene la relazione (3.30) tenendo conto della Definizione 3.24 di quantili

$$\begin{aligned}\boldsymbol{P}\left\{\chi^2_{\frac{\alpha}{2}}(n) \le Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\right\} &= \boldsymbol{P}\left\{Z \le \chi^2_{1-\frac{\alpha}{2}}(n)\right\} - \boldsymbol{P}\left\{Z \le \chi^2_{\frac{\alpha}{2}}(n)\right\} \\ &= \left(1-\frac{\alpha}{2}\right) - \frac{\alpha}{2} = 1-\alpha\end{aligned}$$

Ometteremo invece la dimostrazione della relazione di approssimazione (3.31) □

Teorema 3.28. *Se* $F \sim \mathfrak{F}(n,m)$*, e se* $0 \le \alpha \le 1$*, allora sarà verificata la relazione*

$$\boxed{\boldsymbol{P}\left\{f_{\frac{\alpha}{2}}(n,m) \le F \le f_{1-\frac{\alpha}{2}}(n,m)\right\} = 1-\alpha} \tag{3.32}$$

Inoltre i quantili di $\mathfrak{F}(n,m)$ *godono della seguente proprietà di simmetria*

$$\boxed{f_\alpha(n,m) = \frac{1}{f_{1-\alpha}(m,n)}} \tag{3.33}$$

Dimostrazione: La dimostrazione di (3.32) è analoga a quella di (3.30) per le leggi del *chi-quadro*; la dimostrazione di (3.33) invece è omessa □

I valori dei quantili delle diverse leggi possono essere trovati nelle Tavole dell'Appendice E, ma come vedremo queste, per ragioni di spazio, non riportano tutti i valori necessari a discutere problemi di statistica. I valori mancanti potranno però essere ricavati utilizzando le relazioni (3.25), (3.27), (3.31) e (3.33) che sono state richiamate nei teoremi precedenti proprio a questo scopo

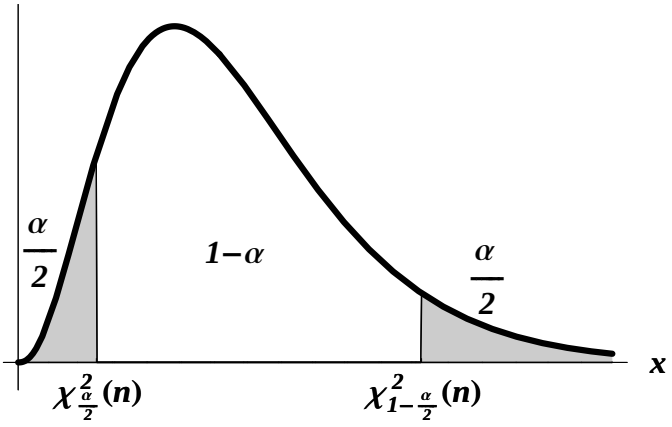


Figura 3.13: Quantili di ordine $\frac{\alpha}{2}$, e $1-\frac{\alpha}{2}$ di una legge $\boldsymbol{\chi^2}(n)$.

3.6 Leggi multivariate

Daremo infine alcune nozioni circa le leggi dei vettori aleatori $\boldsymbol{X} = (X_1, \ldots, X_m)$ introdotti con le Definizioni 3.6, 3.7 e 3.12. Se le m componenti X_i sono ***v-a* discrete**, indicando con x_i i valori discreti di ciascuna di esse, la **legge congiunta** sarà data assegnando le probabilità congiunte

$$p_{\boldsymbol{X}}(x_1, \ldots, x_m) = \boldsymbol{P}\{X_1 = x_1, \ldots, X_m = x_m\}$$

mentre le singole componenti X_i saranno dotate di **leggi marginali** assegnando le probabilità

$$p_{X_i}(x_i) = \boldsymbol{P}\{X_i = x_i\}\ , \qquad i = 1, \ldots, m$$

Le leggi congiunte e marginali di un dato vettore aleatorio non sono però assegnate separatamente in maniera arbitraria: ci sono infatti semplici regole che consentono di ricavare le leggi marginali a partire dalla sua legge congiunta (**marginalizzazione**). Ad esempio, per un vettore $\boldsymbol{X} = (X_1, X_2)$ con due componenti, si dimostra che

$$p_{X_1}(x_1) = \sum_{x_2} p_{\boldsymbol{X}}(x_1, x_2) \qquad p_{X_2}(x_2) = \sum_{x_1} p_{\boldsymbol{X}}(x_1, x_2) \tag{3.34}$$

dove le somme sono estese rispettivamente a tutti i possibili valori di X_1 e di X_2. Dal Teorema 3.13 si può infine dimostrare che le X_i sono indipendenti se e solo se

$$\boxed{p_{\boldsymbol{X}}(x_1, \ldots, x_m) = p_{X_1}(x_1) \cdot \ldots \cdot p_{X_m}(x_m)} \tag{3.35}$$

Se invece le componenti del vettore aleatorio $\boldsymbol{X} = (X_1, \ldots, X_m)$ sono ***v-a* continue** la **legge congiunta** è determinata da una *fdp* con m variabili $f_{\boldsymbol{X}}(x_1, \ldots, x_m)$, mentre le **leggi marginali** saranno date tramite le *fdp* delle singole componenti $f_{X_i}(x_i)$. Le proprietà delle *fdp* m–dimensionali sono del tutto analoghe a quelle delle *fdp* con una sola variabile introdotte nella Definizione 3.17, a parte il fatto che

la loro formulazione richiede l'uso del calcolo differenziale con m variabili. Così ad esempio per un vettore $\boldsymbol{X} = (X_1, X_2)$ con due componenti e legge continua le regole di marginalizzazione (3.34) si modificano in

$$f_{X_1}(x_1) = \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x_1, x_2)\, dx_2 \qquad\qquad f_{X_2}(x_2) = \int_{-\infty}^{+\infty} f_{\boldsymbol{X}}(x_1, x_2)\, dx_1 \tag{3.36}$$

Anche in questo caso dal Teorema 3.13 si può infine dimostrare che le componenti X_i sono indipendenti se e solo se

$$\boxed{f_{\boldsymbol{X}}(x_1, \ldots, x_m) = f_{X_1}(x_1) \cdot \ldots \cdot f_{X_m}(x_m)} \tag{3.37}$$

Esempio 3.29. (Legge multinomiale) *Per dare un esempio di legge congiunta discreta di un vettore aleatorio m–dimensionale consideriamo n ripetizioni indipendenti di un esperimento con $m \geq 2$ possibili risultati casuali $a_1, \ldots, a_m$, e supponiamo che in ogni tentativo si ottenga il risultato a_1 con probabilità q_1, ..., e il risultato a_m con probabilità q_m. Naturalmente dovranno essere verificate le relazioni*

$$0 \leq q_i \leq 1\,, \qquad i = 1, \ldots, m\,; \qquad\qquad q_1 + \ldots + q_m = 1\,.$$

Se ora X_i è la v-a che rappresenta il numero di volte in cui si è ottenuto l'i-mo valore a_i, il risultato del nostro esperimento sarà rappresentato il vettore aleatorio $\boldsymbol{X} = (X_1, \ldots, X_m)$ per il quale ovviamente deve risultare

$$X_1 + \ldots + X_m = n \tag{3.38}$$

I possibili valori di $\boldsymbol{X}$ sono quindi i vettori di numeri interi $(k_1, \ldots, k_m)$ con $k_1 + \ldots + k_m = n$ e si può dimostrare (ma noi trascureremo di farlo) che nelle nostre condizioni la legge congiunta di $\boldsymbol{X}$ è

$$\boxed{p_{\boldsymbol{X}}(k_1, \ldots, k_m) = \boldsymbol{P}\{X_1 = k_1, \ldots, X_m = k_m\} = \frac{n!}{k_1! \ldots k_m!}\, q_1^{k_1} \ldots q_m^{k_m}} \tag{3.39}$$

detta legge multinomiale. È immediato verificare peraltro che si tratta di una generalizzazione della legge binomiale dell'Esempio 3.3.3 che si ottiene come caso particolare con $m = 2$

Un ovvio esempio concreto di legge multinomiale è fornito da n lanci di un dado con $m = 6$ facce numerate $a_1, \ldots, a_6$. Se in generale il dado non è ben bilanciato le probabilità q_i di ottenere il risultato i-mo saranno diverse fra loro: detta allora X_i la v-a che rappresenta il numero di volte in cui su n lanci è uscita la i-ma faccia a_i, la legge del vettore $\boldsymbol{X} = (X_1, \ldots, X_6)$ è proprio (3.39) con $m = 6$. Se poi in particolare il dado è bilanciato, allora $q_1 = \ldots = q_6 = 1/6$ e quindi la legge multinomiale (3.39) si riduce a

$$p_{\boldsymbol{X}}(k_1, \ldots, k_6) = \boldsymbol{P}\{X_1 = k_1, \ldots, X_6 = k_6\} = \frac{n!}{k_1! \ldots k_6!}\, \frac{1}{6^n}$$

Questo ci permette allora di calcolare ad esempio la probabilità che su $n = 12$ lanci di dado ogni faccia esca esattamente due volte

$$p_{\boldsymbol{X}}(2,2,2,2,2,2) = \frac{12!}{2!\,2!\,2!\,2!\,2!\,2!}\,\frac{1}{6^{12}} \simeq 3.44 \times 10^{-3}$$

oppure la probabilità che tre delle sei facce escano 1 volta e le altre tre 3 volte è

$$p_{\boldsymbol{X}}(1,1,1,3,3,3) = \frac{12!}{1!\,1!\,1!\,3!\,3!\,3!}\,\frac{1}{6^{12}} \simeq 1.02 \times 10^{-3}$$

Per le leggi multinomiali è anche possibile calcolare le leggi marginali e verificare che le componenti X_i non sono indipendenti. Per semplicità ometteremo questa verifica e ci limiteremo solo ad osservare che è intuitivamente facile giustificare questa dipendenza se si riflette al fatto che le componenti X_i devono sempre soddisfare la relazione (3.38), per cui, fissate arbitrariamente le prime $m-1$ componenti, la m-a sarebbe immediatamente già determinata con valore $n - (X_1 + \ldots + X_{m-1})$ e quindi non è indipendente dalle altre

Esempio 3.30. (Legge Gaussiana bivariata) *Per dare un esempio di legge congiunta continua si consideri un vettore aleatorio $\boldsymbol{X} = (X_1, X_2)$ con due componenti dotato di fdp congiunta bivariata*

$$\boxed{f_{\boldsymbol{X}}(x_1,x_2) = \frac{e^{-\frac{1}{2(1-r^2)}\left[\frac{(x_1-b_1)^2}{a_1^2} - 2r\frac{(x_1-b_1)(x_2-b_2)}{a_1 a_2} + \frac{(x_2-b_2)^2}{a_2^2}\right]}}{2\pi a_1 a_2 \sqrt{1-r^2}}} \tag{3.40}$$

caratterizzata dai cinque parametri a_1, a_2, b_1, b_2, r con le limitazioni $a_1, a_2 > 0$, e $|r| < 1$: in questo caso diremo anche che $\boldsymbol{X}$ segue una legge Gaussiana (normale) bivariata. Si potrebbe mostrare da (3.36) che le leggi marginali delle due componenti sono $X_1 \sim \mathfrak{N}(b_1, a_1^2)$ e $X_2 \sim \mathfrak{N}(b_2, a_2^2)$, cioè sono anche esse normali con parametri ricavati da quelli di (3.40) e fdp

$$f_{X_1}(x_1) = \frac{1}{a_1\sqrt{2\pi}}\, e^{-(x_1-b_1)^2/2a_1^2} \qquad f_{X_2}(x_2) = \frac{1}{a_2\sqrt{2\pi}}\, e^{-(x_2-b_2)^2/2a_2^2} \tag{3.41}$$

Tenendo conto di (3.37) si vede allora che le due componenti di $\boldsymbol{X}$ saranno indipendenti se e solo se $r = 0$, caso in cui la (3.40) si fattorizza nel prodotto delle due (3.41)

Capitolo 4

Attesa, varianza e correlazione

4.1 Valore d'attesa

Definizione 4.1. *Data una* ***v-a discreta*** X*, con valori* x_k *e legge* $p_X(x_k) = \boldsymbol{P}\{X = x_k\}$*, chiameremo* ***valore d'attesa*** *di* X *(o semplicemente* ***attesa*** *o anche* ***media****) la quantità*

$$\boxed{\mu_X = \boldsymbol{E}[X] = \sum_k x_k \, p_X(x_k)} \tag{4.1}$$

se invece X *è una* ***v-a continua*** *con* fdp $f_X(x)$ *l'attesa sarà definita da*

$$\boxed{\mu_X = \boldsymbol{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\, dx} \tag{4.2}$$

Si noti che nella definizione (4.1) di attesa per una *v-a* discreta X non abbiamo indicato esplicitamente i possibili valori di k perché non è stato precisato quanti valori X assume: va però aggiunto che se X assume infiniti valori la somma è in realtà una serie, e pertanto la definizione ha senso solo quando tale serie converge verso un valore finito, oppure diverge verso $+\infty$ o $-\infty$. In questi casi diremo che la *v-a* discreta X *possiede un valore d'attesa, finito o infinito.* Nei casi in cui invece la serie (4.1) né converge, né diverge (ma nel limite continua a oscillare fra valori differenti) diremo che la *v-a* X *non possiede un valore d'attesa.* Considerazioni del tutto analoghe possono essere applicate anche al comportamento dell'integrale che compare nella definizione (4.2) del valore d'attesa di una *v-a* continua X

In pratica, sia nel caso discreto che in quello continuo, l'attesa non è altro che la somma dei valori assunti da X *pesati* con le rispettive probabilità (si ricordi a questo proposito che nel caso continuo la quantità $f_X(x)\, dx$ può essere intesa come la probabilità che X cada fra x e $x+dx$): pertanto $\boldsymbol{E}[X]$ rappresenta il *baricentro della distribuzione* di X. Nel seguito, quando non diversamente specificato, le definizioni e le proprietà enunciate con il simbolo $\boldsymbol{E}$ saranno valide sia per il caso discreto che per quello continuo

La Definizione 4.1 può essere estesa in modo naturale anche al caso di ***v-a* funzioni di una o più *v-a***. Per non appesantire le notazioni ci limiteremo al caso $Z = h(X, Y)$ in cui la *v-a* Z è funzione solo di altre due *v-a* X e Y: così ad esempio potremmo avere $Z = XY$, oppure $Z = X/Y$, o ancora $Z = \sqrt{X^2 + Y^2}$ e via dicendo. Sarà importante osservare allora che in tutte queste situazioni il valore d'attesa $\boldsymbol{E}[Z]$ – invece che con la distribuzione di Z secondo la Definizione 4.1 – può essere anche calcolato mediante la legge congiunta di X e Y nel modo seguente: se X e Y sono *v-a* discrete che assumono i valori x_k e y_ℓ con legge congiunta $p(x_k, y_\ell)$, allora avremo

$$\boldsymbol{E}[Z] = \boldsymbol{E}[h(X,Y)] = \sum_{k,\ell} h(x_k, y_\ell)\, p(x_k, y_\ell) \tag{4.3}$$

Se invece X e Y sono *v-a* continue con *fdp* congiunta $f(x, y)$, allora avremo

$$\boldsymbol{E}[Z] = \boldsymbol{E}[h(X,Y)] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} h(x,y)\, f(x,y)\, dx\, dy \tag{4.4}$$

In particolare per i prodotti di *v-a* abbiamo nei due casi

$$\boxed{\boldsymbol{E}[XY] = \sum_{k,\ell} x_k y_\ell\, p(x_k, y_\ell) \qquad \boldsymbol{E}[XY] = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} xy\, f(x,y)\, dx\, dy} \tag{4.5}$$

Queste formule si generalizzano facilmente anche al caso $Z = h(X_1, \ldots, X_n)$ di funzioni di n *v-a* usando, con qualche complicazione della notazione, le leggi congiunte del vettore aleatorio $(X_1, \ldots, X_n)$. Inoltre esse si adattano anche al caso più semplice in cui $Z = h(X)$ è funzione di una sola *v-a* X, come ad esempio $Z = X^2$, oppure $Z = 1/X$ e via dicendo. In questo caso – rispettivamente per X discreta con legge $p_X(x_k)$, o continua con *fdp* $f_X(x)$ – avremo

$$\boldsymbol{E}[Z] = \boldsymbol{E}[h(X)] = \sum_k h(x_k)\, p_X(x_k) \tag{4.6}$$

$$\boldsymbol{E}[Z] = \boldsymbol{E}[h(X)] = \int_{-\infty}^{+\infty} h(x)\, f_X(x)\, dx \tag{4.7}$$

Teorema 4.2. (Linearità delle attese) *Se $X, X_1, \ldots, X_n$ sono v-a, e se a, b sono due numeri arbitrari, allora*

$$\boldsymbol{E}[aX + b] = a\boldsymbol{E}[X] + b\,, \tag{4.8}$$

$$\boldsymbol{E}[X_1 + \ldots + X_n] = \boldsymbol{E}[X_1] + \ldots + \boldsymbol{E}[X_n]\,. \tag{4.9}$$

Dimostrazione: Limitandoci per semplicità al caso continuo (nel caso discreto la discussione è analoga sostituendo somme a integrali) da (4.7) con $h(x) = ax + b$ si ha innanzitutto la (4.8)

$$\begin{aligned}\boldsymbol{E}[aX + b] &= \int_{-\infty}^{+\infty} (ax + b)\, f_X(x)\, dx \\ &= a\int_{-\infty}^{+\infty} x\, f_X(x)\, dx + b\int_{-\infty}^{+\infty} f_X(x)\, dx = a\boldsymbol{E}[X] + b\end{aligned}$$

dove abbiamo usato (3.11), (4.2) e le usuali proprietà degli integrali. La (4.9) sarà invece discussa solo per $n = 2$ *v-a* X, Y cioè nella forma semplificata

$$\boldsymbol{E}[X + Y] = \boldsymbol{E}[X] + \boldsymbol{E}[Y]$$

In questo caso, dette $f(x, y)$ la *fdp* congiunta e $f_X(x), f_Y(y)$ rispettivamente le due marginali, avremo infatti da (4.4) con $h(x, y) = x + y$

$$\begin{aligned}
\boldsymbol{E}[X + Y] &= \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x + y)\, f(x, y)\, dx\, dy \\
&= \int_{-\infty}^{+\infty} x\, dx \int_{-\infty}^{+\infty} f(x, y)\, dy + \int_{-\infty}^{+\infty} y\, dy \int_{-\infty}^{+\infty} f(x, y)\, dx \\
&= \int_{-\infty}^{+\infty} x f_X(x)\, dx + \int_{-\infty}^{+\infty} y f_Y(y)\, dy = \boldsymbol{E}[X] + \boldsymbol{E}[Y]
\end{aligned}$$

dove abbiamo usato anche le regole di marginalizzazione (3.36) □

Teorema 4.3. *Se X e Y sono due* ***v-a indipendenti****, allora*

$$\boxed{\boldsymbol{E}[XY] = \boldsymbol{E}[X]\, \boldsymbol{E}[Y]} \tag{4.10}$$

Dimostrazione: Limitandoci infatti sempre al caso in cui X e Y siano *v-a* continue con *fdp* congiunta $f(x, y)$ e marginali $f_X(x)$ e $f_Y(y)$, sappiamo da (3.37) (cioè dal Teorema 3.13) che dalla loro indipendenza deriva

$$f(x, y) = f_X(x) f_Y(y)$$

e quindi da (4.4) con $h(x, y) = xy$ si ha

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

cioè la (4.10) □

4.2 Varianza, covarianza e correlazione

Abbiamo osservato che i valori d'attesa delle *v-a* $X, Y, \ldots$ rappresentano i baricentri delle rispettive distribuzioni: sono cioè degli **indicatori di centralità**, in maniera simile alle *mediane* e alle *mode* introdotte nei capitoli precedenti. L'importanza di tali numeri, però, non deve nascondere il fatto che essi non possono riassumere tutta l'informazione contenuta nella legge congiunta di $X, Y \ldots$ e nelle rispettive

marginali. Così ad esempio due *v-a* potrebbero avere lo stesso valore d'attesa (cioè essere distribuite attorno al medesimo numero), ma essere caratterizzate da dispersioni molto diverse attorno a tale valore: questa differenza non sarebbe catturata dalla identità dei due valori d'attesa. Allo stesso modo i valori d'attesa separati di due o più *v-a* non ci dicono nulla sulla loro dipendenza o indipendenza. Sarà quindi necessario introdurre ulteriori **indicatori di dispersione** e **indicatori di dipendenza**

Definizione 4.4. *Si dice* ***varianza*** *della v-a* X *con attesa* $\boldsymbol{E}[X] = \mu_X$ *la quantità*

$$\boxed{\sigma_X^2 = \boldsymbol{V}[X] = \boldsymbol{E}\left[\left(X - \boldsymbol{E}[X]\right)^2\right] = \boldsymbol{E}\left[(X-\mu_X)^2\right]} \tag{4.11}$$

e ***deviazione standard*** *la quantità* $\sigma_X = \sqrt{\boldsymbol{V}[X]}$. *Si chiama invece* ***covarianza*** *di due v-a* X *e* Y *con attese rispettivamente* μ_X *e* μ_Y *la quantità*

$$\boxed{\kappa_{XY} = \boldsymbol{cov}[X,Y] = \boldsymbol{E}[(X-\mu_X)(Y-\mu_Y)]} \tag{4.12}$$

e ***coefficiente di correlazione*** *la quantità*

$$\boxed{\rho_{XY} = \frac{\boldsymbol{cov}[X,Y]}{\sqrt{\boldsymbol{V}[X]}\sqrt{\boldsymbol{V}[Y]}} = \frac{\kappa_{XY}}{\sigma_X\sigma_Y}} \tag{4.13}$$

La varianza σ_X^2 di una *v-a* costituisce un **indicatore di dispersione** dei valori di X attorno al suo valore d'attesa: infatti, se si considerano gli scarti $X - \mu_X$ dei valori aleatori di X dalla sua media, si vede facilmente da (4.8) che si ha sempre

$$\boldsymbol{E}[X - \mu_X] = \boldsymbol{E}[X] - \mu_X = 0$$

per cui – essendo invariabilmente nulla – la semplice media degli scarti non può essere usata come indicatore di dispersione. In realtà questo dipende dal fatto che gli scarti attorno alla media hanno valori *positivi* e *negativi* equamente ripartiti. Per evitare questo problema, allora, si usa il valore d'attesa del *quadrato* degli scarti, cioè il cosiddetto **scarto quadratico medio**: un altro nome per la varianza definita in 4.11. Quindi una varianza grande indicherà che X tende a prendere valori anche molto lontani da μ_X, e viceversa se la varianza è piccola i valori di X saranno piuttosto concentrati attorno a μ_X

Definizione 4.5. *Diremo che due v-a* X *e* Y *sono* ***non correlate*** *quando* $\kappa_{XY} = \boldsymbol{cov}[X,Y] = 0$*, o equivalentemente quando* $\rho_{XY} = 0$

Si vede subito dalle definizioni (4.11), (4.12) e (4.13) che

$$\kappa_{XX} = \boldsymbol{V}[X] = \sigma_X^2 \qquad\qquad \rho_{XX} = 1 \tag{4.14}$$

per cui la covarianza può essere anche considerata prima di tutto come una generalizzazione del concetto di varianza al caso di due *v-a*. Ma il significato più profondo della covarianza (e del coefficiente di correlazione) risiede nel suo rapporto con il concetto di indipendenza come mostrato nel risultato seguente

Teorema 4.6. *Se due v-a X e Y sono indipendenti, allora sono anche non correlate*

Dimostrazione: Infatti dalla indipendenza e da (4.10) si vede subito che

$$\begin{aligned}\kappa_{XY} = \boldsymbol{cov}[X,Y] &= \boldsymbol{E}[(X-\mu_X)(Y-\mu_Y)] = \boldsymbol{E}[X-\mu_X]\,\boldsymbol{E}[Y-\mu_Y] \\ &= \big(\boldsymbol{E}[X]-\mu_X\big)\big(\boldsymbol{E}[Y]-\mu_Y\big) = 0\end{aligned}$$

e quindi in base alla Definizione 4.5 X e Y risultano non correlate □

Viceversa sarebbe facile far vedere con degli esempi (omessi per brevità) che esistono *v-a* non correlate che *non sono indipendenti*. In altri termini indipendenza e non correlazione non sono concetti equivalenti: l'indipendenza implica la non correlazione, ma in generale il viceversa non è vero. Ciononostante per comodità la non correlazione viene spesso utilizzata come una specie di *indipendenza debole*, nel senso che *v-a* non correlate sono considerate *quasi indipendenti*. Inoltre se $\kappa_{XY} > 0$ si parla di **correlazione positiva** e questo indica che Y tende ad assumere valori grandi (rispettivamente piccoli) quando X assume valori grandi (rispettivamente piccoli). Viceversa se $\kappa_{XY} < 0$ si parla di **correlazione negativa** e in tal caso Y tende ad assumere valori grandi (rispettivamente piccoli) quando X assume valori piccoli (rispettivamente grandi)

Il coefficiente di correlazione, infine, non è altro che una covarianza *ridotta* che gode di alcune importanti proprietà non condivise con la semplice covarianza e riassunte nel seguente teorema

Teorema 4.7. *Date due v-a X e Y risulta sempre*

$$-1 \le \rho_{XY} \le +1$$

e in particolare avremo $|\rho_{XY}| = 1$ se e solo se esistono due numeri α e β tali che $Y = \alpha X + \beta$, con $\alpha < 0$ se $\rho_{XY} = -1$, e $\alpha > 0$ se $\rho_{XY} = +1$. Inoltre, se a, b sono due numeri arbitrari e se poniamo $X' = aX + b$, $Y' = aY + b$, il coefficiente di correlazione resta invariato, cioè si avrà

$$\rho_{X'Y'} = \rho_{XY}$$

Dimostrazione: Omessa. Osserveremo solo che la covarianza κ_{XY} non gode di queste proprietà: innanzitutto κ_{XY} può assume tutti i valori reali positivi e negativi senza essere limitata fra -1 e $+1$. Questo rende difficile capire se un valore di covarianza indica una correlazione statistica grande o piccola perché manca un termine di paragone. Inoltre, diversamente da ρ_{XY}, la covarianza κ_{XY} è sensibile alle (cioé cambia valore nelle) trasformazioni lineari del tipo $X' = aX + b$ che essenzialmente rappresentano cambiamenti di unità di misura (parametro a) o di origine della scala (parametro b). Infine la terza proprietà mostra che (diversamente da κ_{XY}) ρ_{XY} è un efficace indicatore di *dipendenza lineare* fra X e Y □

Teorema 4.8. *Date due v-a X e Y si ha*

$$\kappa_{XY} = \boldsymbol{cov}[X,Y] = \boldsymbol{E}[XY] - \boldsymbol{E}[X]\,\boldsymbol{E}[Y] \tag{4.15}$$

e in particolare con $X = Y$ si ha anche

$$\sigma_X^2 = \boldsymbol{V}[X] = \boldsymbol{E}\left[X^2\right] - \boldsymbol{E}[X]^2 \tag{4.16}$$

Da (4.15) segue che X e Y sono ***non correlate*** *se e solo se $\boldsymbol{E}[XY] = \boldsymbol{E}[X]\,\boldsymbol{E}[Y]$*

Dimostrazione: Partendo dalla definizione (4.12), ponendo $\mu_X = \boldsymbol{E}[X]$ e $\mu_Y = \boldsymbol{E}[Y]$, e usando i risultati del Teorema 4.2 si ricava che

$$\begin{aligned}
\boldsymbol{cov}[X,Y] &= \boldsymbol{E}[(X-\mu_X)(Y-\mu_Y)] = \boldsymbol{E}[XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y] \\
&= \boldsymbol{E}[XY] - \mu_Y\boldsymbol{E}[X] - \mu_X\boldsymbol{E}[Y] + \mu_X\mu_Y \\
&= \boldsymbol{E}[XY] - 2\mu_X\mu_Y + \mu_X\mu_Y = \boldsymbol{E}[XY] - \mu_X\mu_Y \\
&= \boldsymbol{E}[XY] - \boldsymbol{E}[X]\,\boldsymbol{E}[Y]
\end{aligned}$$

La relazione (4.16) poi segue direttamente da (4.15) tramite (4.14) □

La (4.16) del teorema precedente si rivela utile nel calcolo delle varianze, ed è quindi opportuno ricordare in termini espliciti il suo significato: con le solite notazioni nel caso discreto avremo

$$\sigma_X^2 = \sum_k x_k^2 p(x_k) - \left(\sum_k x_k p(x_k)\right)^2 \tag{4.17}$$

mentre nel caso continuo avremo

$$\sigma_X^2 = \int_{-\infty}^{+\infty} x^2 f_X(x)\,dx - \left(\int_{-\infty}^{+\infty} x f_X(x)\,dx\right)^2 \tag{4.18}$$

Abbiamo notato che in base al Teorema 4.2 il calcolo dell'attesa $\boldsymbol{E}[X]$ di una *v-a* è un'operazione lineare; non si può dire invece la stessa cosa per la varianza nella quale evidentemente compaiono quantità elevate al quadrato. Il seguente teorema precisa alcune importanti proprietà della varianza

Teorema 4.9. *Se X è una v-a e a, b sono due numeri si ha innanzitutto*

$$\boldsymbol{V}[aX+b] = a^2\boldsymbol{V}[X] \tag{4.19}$$

Se poi $X_1, \ldots, X_n$ sono n v-a, in generale si ha

$$\begin{aligned}
\boldsymbol{V}[X_1 + \ldots + X_n] &= \sum_{i,j=1}^{n} \boldsymbol{cov}[X_i, X_j] \\
&= \boldsymbol{V}[X_1] + \ldots + \boldsymbol{V}[X_n] + \sum_{i\neq j} \boldsymbol{cov}[X_i, X_j]
\end{aligned} \tag{4.20}$$

e quindi se le X_i sono ***indipendenti*** *(o anche solo* ***non correlate****) risulta*

$$\boldsymbol{V}[X_1 + \ldots + X_n] = \boldsymbol{V}[X_1] + \ldots + \boldsymbol{V}[X_n] \tag{4.21}$$

Dimostrazione: Posto $\mu_X = \boldsymbol{E}[X]$, da (4.8) si ha $\boldsymbol{E}[aX+b] = a\mu_X + b$ e quindi

$$aX + b - \boldsymbol{E}[aX+b] = a(X - \mu_X)$$

Dalla definizione (4.11) e dal Teorema 4.2 si ha allora immediatamente la (4.19)

$$\boldsymbol{V}[aX+b] = \boldsymbol{E}\left[a^2(X-\mu_X)^2\right] = a^2\boldsymbol{E}\left[(X-\mu_X)^2\right] = a^2\boldsymbol{V}[X]$$

La (4.20) sarà invece dimostrata solo per $n = 2$ *v-a* X e Y: in questo caso essa si riduce a

$$\boldsymbol{V}[X+Y] = \boldsymbol{V}[X] + \boldsymbol{V}[Y] + 2\boldsymbol{cov}[X,Y] \tag{4.22}$$

Posto allora $\mu_X = \boldsymbol{E}[X]$ e $\mu_Y = \boldsymbol{E}[Y]$, da (4.9) si ha $\boldsymbol{E}[X+Y] = \mu_X + \mu_Y$, e quindi dalle definizioni (4.11) e (4.12) applicando il Teorema 4.2 si ha

$$\begin{aligned}
\boldsymbol{V}[X+Y] &= \boldsymbol{E}\left[\big((X+Y)-(\mu_X+\mu_Y)\big)^2\right] = \boldsymbol{E}\left[\big((X-\mu_X)+(Y-\mu_Y)\big)^2\right] \\
&= \boldsymbol{E}\left[(X-\mu_X)^2\right] + \boldsymbol{E}\left[(Y-\mu_Y)^2\right] + 2\boldsymbol{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \\
&= \boldsymbol{V}[X] + \boldsymbol{V}[Y] + 2\boldsymbol{cov}[X,Y]
\end{aligned}$$

Naturalmente la (4.22) si riduce a

$$\boldsymbol{V}[X+Y] = \boldsymbol{V}[X] + \boldsymbol{V}[Y] \tag{4.23}$$

(cioè alla (4.21) per $n = 2$) quando $\boldsymbol{cov}[X,Y] = 0$, ossia quando X e Y sono non correlate □

Definizione 4.10. *Diremo che X è una* ***v-a centrata*** *quando $\boldsymbol{E}[X] = 0$, e quindi da (4.16) risulta anche $\boldsymbol{V}[X] = \boldsymbol{E}[X^2]$; diremo invece che è una* ***v-a standardizzata*** *quando è centrata e inoltre $\boldsymbol{V}[X] = 1$*

Teorema 4.11. *Se X è una v-a con attesa $\mu_X = \boldsymbol{E}[X]$ e varianza $\sigma_X^2 = \boldsymbol{V}[X]$, e se poniamo*

$$\widetilde{X} = X - \mu_X \qquad X^* = \frac{X-\mu_X}{\sigma_X} = \frac{\widetilde{X}}{\sigma_X}$$

allora $\widetilde{X}$ risulta centrata, e X^* risulta standardizzata

Dimostrazione: Dalla (4.8) si ha innanzitutto che $\widetilde{X}$ è centrata

$$\boldsymbol{E}\big[\widetilde{X}\,\big] = \boldsymbol{E}[X-\mu_X] = \boldsymbol{E}[X] - \mu_X = 0$$

e quindi anche $\boldsymbol{E}[X^*] = 0$ sempre per effetto di (4.8) (cioè anche X^* risulterà centrata). Tenendo poi conto di (4.19) e del fatto che $\boldsymbol{E}\left[\widetilde{X}\right] = 0$ si ha

$$\boldsymbol{V}[X^*] = \frac{\boldsymbol{V}\big[\widetilde{X}\,\big]}{\sigma_X^2} = \frac{\boldsymbol{E}\big[\widetilde{X}^2\,\big]}{\sigma_X^2} = \frac{\boldsymbol{E}\left[(X-\mu_X)^2\right]}{\sigma_X^2} = \frac{\boldsymbol{V}[X]}{\sigma_X^2} = 1$$

e quindi X^* è anche standardizzata □

4.3 Momenti

Abbiamo visto nelle sezioni precedenti che il valore d'attesa di una *v-a* è un indicatore di centralità della sua distribuzione, che la sua varianza è un indicatore di dispersione e che il coefficiente di correlazione è un indicatore di dipendenza di due *v-a*. Ancora una volta però dobbiamo ricordare che l'informazione contenuta nelle leggi delle *v-a* non si esaurisce con queste informazioni, e che quindi risulta utile introdurre anche le seguenti ulteriori nozioni

Definizione 4.12. *Chiameremo rispettivamente* ***momento di ordine*** k *e* ***momento centrato di ordine*** k *di una v-a* X *le quantità*

$$\mu_k = \boldsymbol{E}\left[X^k\right] \qquad\qquad \widetilde{\mu}_k = \boldsymbol{E}\left[(X - \boldsymbol{E}[X])^k\right]$$

Chiameremo in particolare ***skewness*** *(****asimmetria****) e* ***curtòsi*** *di* X *rispettivamente le quantità*

$$\gamma_1 = \frac{\widetilde{\mu}_3}{\widetilde{\mu}_2^{\,3/2}} \qquad\qquad \gamma_2 = \frac{\widetilde{\mu}_4}{\widetilde{\mu}_2^2}$$

È facile vedere ora che $\mu_1 = \boldsymbol{E}[X] = \mu_X$ (cioè l'attesa di è il momento di ordine 1), e che $\widetilde{\mu}_2 = \boldsymbol{V}[X] = \sigma_X^2$ (cioè la varianza è il momento centrato di ordine 2), mentre *skewness* e *curtosi* sono indicatori basati su momenti di ordine 3 e 4 che forniscono ulteriori informazioni sulla *forma* della distribuzione di X. In particolare la *skewness* γ_1 è una misura della *asimmetria* (in inglese: skewness) della distribuzione di X attorno al suo valore d'attesa μ_X, e può assumere valori sia positivi che negativi: un valore $\gamma_1 < 0$ indica che la distribuzione di X ha una coda a sinistra di μ_X più lunga della coda a destra, mentre $\gamma_1 > 0$ indica un eccesso della coda verso destra. Naturalmente $\gamma_1 = 0$ indicherà che la distribuzione è perfettamente simmetrica attorno a μ_X. La *curtosi* γ_2 invece (dal greco *kyrtòs*: curvo, arcuato) misura la rapidità con cui le code della distribuzione si annullano (*appiattimento*). Essa assume solo valori positivi (è un momento di ordine *pari*, anzi si può dimostrare che $\gamma_2 \geq 1$) e i suoi valori vengono in genere confrontati con il valore 3 della *curtosi* di una legge Gaussiana: quando $\gamma_2 > 3$ la distribuzione si *appiattisce* più lentamente di una Gaussiana e si presenta più *appuntita* attorno al valore d'attesa; viceversa per $\gamma_2 < 3$ essa si *appiattisce* più velocemente della Gaussiana e si presenta più larga attorno al valore d'attesa. Attesa, varianza, *skewness* e *curtosi* tengono conto del valore dei momenti fino al quarto ordine: prendendo in considerazione gli ordini superiori al quarto si ottengono ulteriori informazioni – sempre più raffinate, ma meno rilevanti – ai fini della descrizione della distribuzione. Noi però ometteremo per brevità di farne menzione

4.4 Esempi di attese e varianze

Osserviamo innanzitutto che il valore d'attesa e la varianza (e in generale i momenti) di una *v-a* X dipendono in realtà soltanto dalla legge $\mathfrak{L}$ di X, nel senso che *v-a*

diverse hanno la stessa attesa e la stessa varianza se sono identicamente distribuite. Pertanto nel seguito, studiando alcuni semplici esempi di calcolo di queste quantità, potrà essere conveniente indicare esplicitamente solo il simbolo della distribuzione piuttosto che quello della *v-a*: così scriveremo anche $\boldsymbol{E}[\mathfrak{L}]$ o $\boldsymbol{V}[\mathfrak{L}]$ per indicare l'attesa $\boldsymbol{E}[X]$ o la varianza $\boldsymbol{V}[X]$ di una qualsiasi *v-a* $X \sim \mathfrak{L}$

4.4.1 Distribuzioni discrete

Teorema 4.13. *Per le distribuzioni binomiali* $\mathfrak{B}(n;p)$ *e di Poisson* $\mathfrak{P}(\lambda)$ *si ha*

$$\boxed{\boldsymbol{E}[\mathfrak{B}(n;p)] = np \qquad \boldsymbol{V}[\mathfrak{B}(n;p)] = np(1-p)} \tag{4.24}$$

$$\boxed{\boldsymbol{E}[\mathfrak{P}(\lambda)] = \lambda \qquad \boldsymbol{V}[\mathfrak{P}(\lambda)] = \lambda} \tag{4.25}$$

Dimostrazione: Osserviamo innanzitutto che – in base a quanto detto nelle Sezioni 3.3.2 e 3.3.3 – una *v-a* discreta con legge di Bernoulli $X \sim \mathfrak{B}(1;p)$ (cioè una binomiale con $n = 1$) assume solo i due valori 0 e 1 con probabilità rispettivamente $1-p$ e p. Pertanto dalla (4.1) si ha banalmente

$$\boldsymbol{E}[\mathfrak{B}(1;p)] = \boldsymbol{E}[X] = 0 \cdot (1-p) + 1 \cdot p = p$$

D'altra parte per X^2 dalla (4.6) con $h(x) = x^2$ risulta

$$\boldsymbol{E}\left[X^2\right] = 0^2 \cdot (1-p) + 1^2 \cdot p = p$$

e quindi da (4.16) si ha

$$\boldsymbol{V}[\mathfrak{B}(1;p)] = \boldsymbol{V}[X] = \boldsymbol{E}\left[X^2\right] - \boldsymbol{E}[X]^2 = p - p^2 = p(1-p)$$

sicché la (4.24) resta dimostrata per $n = 1$. Ricordando poi che per il Teorema 3.16 ogni *v-a* binomiale $\mathfrak{B}(n;p)$ ha la stessa legge di una somma $X_1 + \ldots + X_n$ di n *v-a indipendenti* e tutte con legge di Bernoulli $\mathfrak{B}(1;p)$, da (4.9) e (4.21) si ottiene

$$\begin{aligned} \boldsymbol{E}[\mathfrak{B}(n;p)] &= \boldsymbol{E}[X_1 + \ldots + X_n] = \boldsymbol{E}[X_1] + \ldots \boldsymbol{E}[X_n] = np \\ \boldsymbol{V}[\mathfrak{B}(n;p)] &= \boldsymbol{V}[X_1 + \ldots + X_n] = \boldsymbol{V}[X_1] + \ldots \boldsymbol{V}[X_n] = np(1-p) \end{aligned}$$

cioè (4.24) con n generico. Per una *v-a* di Poisson $X \sim \mathfrak{P}(\lambda)$ che assume tutti i valori interi $k = 0, 1, \ldots$, da (3.10), e omettendo in (4.1) il primo termine della serie perché nullo, si ha

$$\boldsymbol{E}[\mathfrak{P}(\lambda)] = \boldsymbol{E}[X] = \sum_{k=0}^{\infty} k\, p_k = \sum_{k=1}^{\infty} k\, p_k = \sum_{k=1}^{\infty} k\, e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!}$$

Ridefinendo ora l'indice di somma $\ell = k - 1$ e ricordando lo sviluppo in serie di Taylor

$$\sum_{\ell=0}^{\infty} \frac{\lambda^\ell}{\ell!} = e^\lambda$$

si ottiene infine

$$\boldsymbol{E}\left[\mathfrak{P}(\lambda)\right] = \boldsymbol{E}\left[X\right] = e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^{\ell+1}}{\ell!} = \lambda\, e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^{\ell}}{\ell!} = \lambda$$

cioè la prima delle (4.25). In maniera analoga, per la seconda si osserva innanzitutto che da (4.6) con $h(x) = x(x-1)$ risulta

$$\boldsymbol{E}\left[X(X-1)\right] = \sum_{k=2}^{\infty} k(k-1)\, e^{-\lambda}\frac{\lambda^{k}}{k!} = e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^{k}}{(k-2)!}$$

e quindi con $\ell = k-2$

$$\boldsymbol{E}\left[X(X-1)\right] = e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^{\ell+2}}{\ell!} = \lambda^2\, e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^{\ell}}{\ell!} = \lambda^2$$

Pertanto dalla linearità delle attese e dai risultati precedenti

$$\boldsymbol{E}\left[X^2\right] = \boldsymbol{E}\left[X^2\right] - \boldsymbol{E}\left[X\right] + \boldsymbol{E}\left[X\right] = \boldsymbol{E}\left[X(X-1)\right] + \boldsymbol{E}\left[X\right] = \lambda^2 + \lambda$$

e quindi

$$\boldsymbol{V}\left[\mathfrak{P}(\lambda)\right] = \boldsymbol{E}\left[X^2\right] - \boldsymbol{E}\left[X\right]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

cioè la seconda delle (4.25) □

4.4.2 Distribuzioni continue

Teorema 4.14. *Per le distribuzioni uniformi* $\mathfrak{U}(a,b)$*, Gaussiane* $\mathfrak{N}(\mu,\sigma^2)$ *e del chi-quadro* $\boldsymbol{\chi^2}(n)$ *si ha*

$$\boxed{\boldsymbol{E}\left[\mathfrak{U}(a,b)\right] = \frac{a+b}{2} \qquad \boldsymbol{V}\left[\mathfrak{U}(a,b)\right] = \frac{(b-a)^2}{12}} \tag{4.26}$$

$$\boxed{\boldsymbol{E}\left[\mathfrak{N}(\mu,\sigma^2)\right] = \mu \qquad \boldsymbol{V}\left[\mathfrak{N}(\mu,\sigma^2)\right] = \sigma^2} \tag{4.27}$$

$$\boldsymbol{E}\left[\boldsymbol{\chi^2}(n)\right] = n \qquad \boldsymbol{V}\left[\boldsymbol{\chi^2}(n)\right] = 2n \tag{4.28}$$

Dimostrazione: Omessa. Osserveremo solo che per una normale standard la (4.27) si riducono ovviamente a

$$\boldsymbol{E}\left[\mathfrak{N}(0,1)\right] = 0 \qquad \boldsymbol{V}\left[\mathfrak{N}(0,1)\right] = 1 \tag{4.29}$$

e che questo permette di dimostrare facilmente almeno la prima delle (4.28). Infatti per il Teorema 3.20 una *v-a chi-quadro* ha la stessa legge di una somma $X_1^2 + \ldots + X_n^2$ di *v-a* X_k *indipendenti* e tutte normali standard. Segue allora facilmente da (4.9) e da (4.29) che

$$\boldsymbol{E}\left[\boldsymbol{\chi^2}(n)\right] = \boldsymbol{E}\left[X_1^2\right] + \ldots + \boldsymbol{E}\left[X_n^2\right] = \boldsymbol{V}\left[X_1\right] + \ldots + \boldsymbol{V}\left[X_n\right] = n$$

cioè la prima delle (4.28). Si noti che le formule (4.27) attribuiscono ora anche un significato probabilistico preciso (attesa e varianza) ai due parametri μ e σ che inizialmente avevano solo un ruolo puramente analitico (massimo o moda, e flessi) nella descrizione del grafico della *fdp* di $\mathfrak{N}(\mu, \sigma^2)$ □

Teorema 4.15. *Per le leggi di Student $\mathfrak{T}(n)$ e di Fisher $\mathfrak{F}(n, m)$ l'esistenza delle attese e delle varianze (e più in generale dei momenti) non è sempre garantita, ma dipende dal numero di gradi di libertà n, m. Più precisamente risulta*

$$\boldsymbol{E}[\mathfrak{T}(n)] = 0, \quad \textit{se } n \geq 2 \qquad \boldsymbol{V}[\mathfrak{T}(n)] = \frac{n}{n-2}, \quad \textit{se } n \geq 3$$

$$\boldsymbol{E}[\mathfrak{F}(n,m)] = \frac{m}{m-2}, \quad \textit{se } m \geq 3 \qquad \boldsymbol{V}[\mathfrak{F}(n,m)] = \frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}, \quad \textit{se } m \geq 5$$

Dimostrazione: Omessa □

Capitolo 5

Teoremi limite

5.1 Convergenza

I teoremi limite costituiscono una famiglia di risultati della massima importanza teorica e pratica: in un certo senso infatti essi rappresentano, nel quadro dei fenomeni casuali, la manifestazione delle cosiddette *regolarità statistiche* che ci permettono di eseguire calcoli e formulare previsioni. L'enunciazione di questi teoremi richiede però che venga precisato in che senso intendiamo il *concetto di limite* in probabilità: data infatti una successione di *v-a*

$$(X_n)_{n\in\boldsymbol{N}} \equiv X_1, X_2, \ldots, X_n, \ldots$$

e la successione delle loro corrispondenti leggi (nella forma che riterremo più opportuna: *FDC* $F_n(x)$, o *fdp* $f_n(x)$ o distribuzione discreta $p_k(n)$)

$$(\mathfrak{L}_n)_{n\in\boldsymbol{N}} \equiv \mathfrak{L}_1, \mathfrak{L}_2, \ldots, \mathfrak{L}_n, \ldots$$

bisogna dire subito che ci sono molti modi non equivalenti di definire le richieste convergenze, anche se noi qui ci limteremo a dare solo i concetti strettamente necessari per esprimere compiutamente i risultati di questo capitolo

La seconda osservazione importante da fare è che in generale – quale che sia il senso preciso della convergenza adottata – il limite di una successione di *v-a* $(X_n)_{n\in\boldsymbol{N}}$ (se esiste) è ancora una *v-a* X con valori casuali, così come il limite di una successione di leggi $(\mathfrak{L}_n)_{n\in\boldsymbol{N}}$ è di nuovo una legge $\mathfrak{L}$. Non è escluso però il caso particolare della convergenza di $(X_n)_{n\in\boldsymbol{N}}$ verso un *numero non casuale*, ovvero della convergenza delle corrispondenti $(\mathfrak{L}_n)_{n\in\boldsymbol{N}}$ verso una *legge degenere* (vedi Sezione 3.3.1): in questo caso, in cui il fenomeno limite non riveste più carattere aleatorio, parleremo di *convergenza degenere*. Con queste premesse daremo ora solo le definizioni di convergenza che saranno necessarie per il seguito, limitandone peraltro gli enunciati alla forma e alle condizioni che in pratica utilizzeremo: le definizioni più generali e dettagliate, e le relazioni che intercorrono fa di esse possono comunque essere trovate su un gran numero di manuali in circolazione

Definizione 5.1. (Convergenza degenere in media quadratica - *mq*) *Data una successione di v-a* $(X_n)_{n\in\boldsymbol{N}}$ *diremo che per* $n \to \infty$ *essa converge in media quadratica (mq) verso il numero* a*, e scriveremo*

$$X_n \xrightarrow{mq} a \tag{5.1}$$

quando

$$\lim_n \boldsymbol{E}[X_n] = a \qquad\qquad \lim_n \boldsymbol{V}[X_n] = 0 \tag{5.2}$$

Questa definizione deriva il suo significato da quello di varianza (che come abbiamo visto nella Sezione 4.2 è lo scarto quadratico medio rispetto al valore d'attesa): se la successione delle *v-a* X_n deve convergere verso un numero (non aleatorio) a, è intuitivo pensare che da un lato la successione delle $\boldsymbol{E}[X_n]$ deve tendere proprio al numero a, e dall'altro che la distribuzione delle X_n deve diventare sempre più concentrata attorno al valore d'attesa limite a, cioè le loro varianze devono essere infinitesime. Sotto queste condizioni infatti è ragionevole pensare che, al linite per $n \to \infty$, le X_n diventano *v-a* che assumono invariabilmente (cioè con varianza nulla) il valore a

Definizione 5.2. (Convergenza in distribuzione - *d*) *Data una successione di v-a* $(X_n)_{n\in\boldsymbol{N}}$ *e la successione delle loro leggi* $(\mathfrak{L}_n)_{n\in\boldsymbol{N}}$ *con le corrispondenti FDC* $F_n(x)$*, diremo che per* $n \to \infty$ *essa converge in distribuzione verso la legge* $\mathfrak{L}$ *con FDC* $F(x)$*, e scriveremo*

$$\mathfrak{L}_n \xrightarrow{d} \mathfrak{L} \qquad \textit{ovvero anche} \qquad X_n \xrightarrow{d} \mathfrak{L} \tag{5.3}$$

quando per ogni x

$$\lim_n F_n(x) = F(x) \tag{5.4}$$

La (5.4) si applica anche a FDC discrete, ma quando le leggi $\mathfrak{L}_n$ *e* $\mathfrak{L}$ *sono tutte discrete con valori interi* k*, e con distribuzioni* $p_k(n)$ *e* p_k*, la convergenza in distribuzione (5.3) equivale a richiedere che per ogni* k *risulti*

$$\lim_n p_k(n) = p_k \tag{5.5}$$

5.2 Legge dei Grandi Numeri

Come vedremo meglio nella parte di Statistica Inferenziale, lo scopo principale della statistica è in generale quello di estrarre delle informazioni sulla distribuzione di una *v-a* X a partire da un certo numero di osservazioni empiriche (misure) che costituiscono quello che chiameremo un *campione statistico*. A questo proposito noi anticiperemo qui qualche idea per spiegare il ruolo nell'analisi dei dati sperimentali di alcuni importanti risultati del calcolo delle probabilità, e cominceremo con quello che è stato storicamente il primo (inizio del XVIII secolo) e va sotto il nome di Legge dei Grandi Numeri

Teorema 5.3. Legge dei Grandi Numeri (*LGN*) *Se X_k con $k \in \boldsymbol{N}$ è una successione di v-a indipendenti, tutte con la stessa attesa $\boldsymbol{E}[X_k] = \mu$ e la stessa varianza $\boldsymbol{V}[X_k] = \sigma^2$, per $n \to \infty$ avremo in media quadratica*

$$\overline{X}_n \equiv \frac{1}{n}\sum_{k=1}^{n} X_k \xrightarrow{mq} \mu \tag{5.6}$$

$$\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 \xrightarrow{mq} \sigma^2 \tag{5.7}$$

Dimostrazione: Siccome si tratta di due casi di convergenza degenere in mq nel senso della Definizione 5.1, per dimostrare (5.6) osserviamo che da (4.8) e (4.9) si ha per ogni n

$$\boldsymbol{E}\left[\overline{X}_n\right] = \frac{\boldsymbol{E}[X_1] + \ldots + \boldsymbol{E}[X_n]}{n} = \frac{n\mu}{n} = \mu$$

e quindi banalmente $\boldsymbol{E}\left[\overline{X}_n\right] \xrightarrow{n} \mu$. Ricordando poi che le X_k sono indipendenti da (4.19) e (4.21) si ha

$$\boldsymbol{V}\left[\overline{X}_n\right] = \frac{\boldsymbol{V}[X_1] + \ldots + \boldsymbol{V}[X_n]}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \xrightarrow{n} 0$$

e quindi $\overline{X}_n \xrightarrow{mq} \mu$ in base alla Definizione 5.1. Per dimostrare invece la (5.7) cominciamo con l'osservare che le *v-a* $W_k \equiv (X_k - \mu)^2$ hanno tutte la stessa attesa

$$\boldsymbol{E}[W_k] = \boldsymbol{E}\left[(X_k - \mu)^2\right] = \boldsymbol{V}[X_k] = \sigma^2$$

per cui applicando il risultato (5.6) alla successione delle W_k potremo scrivere che

$$\overline{W}_n \equiv \frac{1}{n}\sum_{k=1}^{n} W_k \xrightarrow{mq} \sigma^2$$

Inoltre dalla (5.6) si ricava anche in particolare

$$\left(\mu - \overline{X}_n\right)^2 \xrightarrow{mq} 0$$

Tenendo allora conto di questi due limiti complessivamente avremo

$$\begin{aligned}
\widehat{S}_n^2 &= \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \frac{1}{n}\sum_{k=1}^{n}[(X_k - \mu) + (\mu - \overline{X}_n)]^2 \\
&= \frac{1}{n}\sum_{k=1}^{n}(X_k - \mu)^2 + \frac{2}{n}\sum_{k=1}^{n}(X_k - \mu)(\mu - \overline{X}_n) + \frac{1}{n}\sum_{k=1}^{n}(\mu - \overline{X}_n)^2 \\
&= \overline{W}_n + 2(\mu - \overline{X}_n)\frac{1}{n}\sum_{k=1}^{n}(X_k - \mu) + \frac{n\,(\mu - \overline{X}_n)^2}{n} \\
&= \overline{W}_n + 2(\mu - \overline{X}_n)(\overline{X}_n - \mu) + (\mu - \overline{X}_n)^2 = \overline{W}_n - (\mu - \overline{X}_n)^2 \xrightarrow{mq} \sigma^2
\end{aligned}$$

come enunciato nella (5.7) □

Esempio 5.4. *Per illustrare l'importanza della LGN supponiamo di avere una popolazione di individui di due tipi che chiameremo convenzionalmente A e B: ad esempio maschi e femmine di una specie biologica, oppure divisione in due partiti di un gruppo di persone, o anche palline di due colori in un'urna ... e così via. Supporremo inoltre che sia sconosciuta la proporzione p degli individui di tipo A (ovviamente $1-p$ sarà la proporzione degli individui B, con $0<p<1$)*

*Se la popolazione è piccola ed è tutta disponibile all'osservazione, il valore di p potrà essere ottenuto semplicemente contando tutti gli individui di tipo A e B. In generale però succede che la popolazione sia molto grande (come nel caso dei cittadini di uno stato) o anche non tutta disponibile per un'osservazione (come nel caso degli individui di una specie biologica). Per attribuire un valore attendibile, ossia per **stimare** la proporzione p, si procede allora come nei sondaggi pre-elettorali: si estrae un **campione casuale** di n individui e li si esamina contando il numero N_A di quelli di tipo A. È giudicato ragionevole a questo punto supporre che il numero $\overline{p}=N_A/n$ (**frequenza relativa empirica** dei ritrovamenti di A) rappresenti una stima accettabile di p, e che tale stima sia tanto più attendibile quanto più grande è il numero n. Lasciata così, però, questa affermazione manca di una solida giustificazione. Si noti a questo proposito che p e $\overline{p}$ sono entità molto diverse: più precisamente la differenza fra di esse non sta solo nel fatto che i due valori numerici in generale non coincidono, ma anche e soprattutto nel fatto che p è un ben determinato (ancorché sconosciuto) numero, mentre $\overline{p}$ è una vera e propria v-a: infatti ogni volta che ripetiamo l'estrazione casuale degli n individui da esaminare otterremo tipicamente un diverso valore di $\overline{p}$*

Per dare al nostro problema una veste un po' più precisa dovremo allora riformularlo nel quadro della LGN*: si considerino a questo scopo n v-a indipendenti $X_1,\dots,X_n$ (il nostro **campione casuale**) in modo che ciascuna X_k rappresenti la misura sull'individuo k-mo con*

$$X_k=\begin{cases}1 & \text{se l'individuo } k\text{-mo è di tipo } A\\ 0 & \text{altrimenti}\end{cases}\qquad\qquad k=1,2,\dots,n$$

Pertanto $X_k\sim\mathfrak{B}(1;p)$ sono tutte v-a di Bernoulli con $\boldsymbol{P}\{X_k=1\}=p$, e $\boldsymbol{E}[X_k]=p$. Inoltre è chiaro che $N_A=X_1+\ldots+X_n\sim\mathfrak{B}(n;p)$ (vedi Teorema 3.16) rappresenta la frequenza dei ritrovamenti di A, sicché la stima di p intuitivamente suggerita prima si presenta ora come la v-a

$$\overline{p}=\frac{N_A}{n}=\frac{X_1+\ldots+X_n}{n}=\overline{X}_n \tag{5.8}$$

cioè la media aritmetica delle X_k. Che tale $\overline{p}$ rappresenti una buona stima di p, e che questa diventi tanto migliore quanto più grande è n, appare allora non più come un fatto intuitivo dettato da antiche consuetudini, ma chiaramente come una diretta conseguenza della LGN in quanto da (5.6) si ha $\overline{p}=\overline{X}_n\xrightarrow{mq}p$, essendo p l'attesa di tutte le v-a X_k

Dall'esempio che precede si ricava allora l'idea che in statistica – come vedremo meglio più oltre – esistono due tipi di quantità: quelle *teoriche*, e quelle *empiriche* che servono a stimare le prime come vedremo meglio nel Capitolo 8. Così in particolare il valore d'attesa (o media) $\mu = \boldsymbol{E}[X]$ di una *v-a* X è una quantità teorica (nel senso che essa dipende dal modello matematico del nostro esperimento e si calcola dalla legge di X), mentre la **media aritmetica**

$$\overline{X}_n = \frac{X_1 + \ldots + X_n}{n} = \frac{1}{n}\sum_{k=1}^{n} X_k \tag{5.9}$$

ottenuta da un campione casuale di *v-a* tutte distribuite come X, è la corrispondente quantità empirica (calcolata cioè dalle misure sperimentali effettuate sul sistema) che viene usata per stimare $\boldsymbol{E}[X]$ sulla base della *LGN*. Insisteremo inoltre sul fatto che mentre le quantità teoriche sono tipicamente dei numeri (non aleatori, come μ), le quantità empiriche usate per stimarle sono delle *v-a* (come $\overline{X}_n$) che tengono conto della inevitabile variabilità delle misure sperimentali, e sono tanto più affidabili quanto più n è grande

Naturalmente le medie, teoriche o empiriche, non sono le uniche quantità rilevanti in probabilità e statistica: in particolare è anche importante stimare le varianze. In questo caso, sempre sulla base della *LGN* che per grandi n ci consente di sostituire il calcolo empirico di medie aritmetiche al calcolo teorico dei valori d'attesa, per stimare una varianza $\sigma^2 = \boldsymbol{V}[X]$ si potrà utilizzare ad esempio la *v-a*

$$\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \overline{X_n^2} - \overline{X}_n^{\,2} \qquad \text{con} \qquad \overline{X_n^2} \equiv \frac{1}{n}\sum_{k=1}^{n} X_k^2 \tag{5.10}$$

che prende il nome di **varianza campionaria** e che – sostituendo le medie ai valori d'attesa della varianza vera e propria (4.11) – corrisponde allo scarto quadratico medio (empirico) dei valori X_k del campione dalla loro media $\overline{X}_n$. Vedremo però nel seguito che in statistica inferenziale sarà necessario introdurre una correzione della espressione (5.10): più precisamente la *v-a* usata per stimare una varianza sarà piuttosto la

$$S_n^2 \equiv \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \frac{n}{n-1}\,\widehat{S}_n^2 = \frac{n}{n-1}\,\left(\overline{X_n^2} - \overline{X}_n^{\,2}\right) \tag{5.11}$$

già introdotta nel Teorema 3.23, che prende il nome di **varianza corretta**: le ragioni di questa correzione saranno spiegate più oltre

Si noti infine che nella parte di Statistica Descrittiva introdurremo delle quantità (chiamate sempre *media* e *varianza*) molto simili a $\overline{X}_n$ e a $\widehat{S}_n^2$, ma con una importante differenza: $\overline{X}_n$ e $\widehat{S}_n^2$ sono calcolate da un campione casuale $X_1, \ldots, X_n$ estratto dalla popolazione teorica (cioè sono misure ripetute della *v-a* X) che in generale non è tutta disponibile per le osservazioni, e per questo motivo (5.9) e (5.10) sono delle vere e proprie *v-a*. In statistica descrittiva invece medie e varianze saranno calcolate a partire da campioni di misure che rappresentano tutta la popolazione disponibile, e saranno quindi semplicemente dei numeri

5.3 Teorema Limite Centrale

La *LGN* fornisce delle preziose indicazioni sulla stima dei parametri statistici, ma non permette di valutare quantitativamente l'attendibilità di queste stime. Sarebbe molto utile infatti avere degli strumenti che ci permettessero di dire, ad esempio, con quale probabilità il *valore stimato*, che è pur sempre una *v-a*, cade all'interno di un qualche intervallo preso attorno al *valore vero* del parametro in esame. In pratica abbiamo bisogno di poter valutare l'errore che si commette approssimando un parametro con una stima empirica. È peraltro evidente che per fare questo bisognerebbe avere o delle informazioni sulla distribuzione della *v-a* usata per stimare il parametro, o almeno una buona approssimazione di tale legge. Sarà utile allora ricordare un altro importante risultato noto come *Teorema Limite Centrale* che stabilisce una certa forma di *universalità delle distribuzioni Gaussiane*: se una *v-a* X è somma di un gran numero di altre piccole *v-a*, allora essa è approssimativamente distribuita secondo una legge normale. Questa proprietà è particolarmente importante in molti settori della statistica e val quindi la pena darne una formulazione più precisa

Teorema 5.5. Teorema Limite Centrale (*TLC*) *Data una successione* X_k *con* $k \in \boldsymbol{N}$ *di v-a indipendenti, tutte con la stessa attesa* μ *e la stessa varianza* $\sigma^2 > 0$*, definite le v-a* $Y_n = X_1 + \ldots + X_n = n\overline{X}_n$ *e le corrispondenti v-a standardizzate*

$$Y_n^* = \frac{Y_n - n\mu}{\sigma\sqrt{n}} = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}$$

queste convergono in distribuzione per $n \to \infty$ *verso la legge normale standard* $\mathfrak{N}(0,1)$

$$Y_n^* \xrightarrow{d} \mathfrak{N}(0,1)$$

cioè, se $F_n(y) = \boldsymbol{P}\{Y_n^* \le y\}$ *sono le FDC delle* Y_n^**, e se* $\Phi(y)$ *è la FDC (3.18) della legge normale standard* $\mathfrak{N}(0,1)$*, per ogni* $y \in \boldsymbol{R}$ *risulterà*

$$\lim_n F_n(y) = \Phi(y)$$

Dimostrazione: Omessa. Si noti che, diversamente da quanto accade nel Teorema 3.23, qui non si suppone che le X_k siano fin dall'inizio *v-a* Gaussiane: esse sono del tutto arbitrarie, tranne per il fatto di essere indipendenti e di avere varianza non nulla. Pertanto il Teorema 3.23 non si applica e le Y_n^* non sono Gaussiane per ogni dato n: è solo al limite per $n \to \infty$ che la loro distribuzione approssima quella della normale standard $\mathfrak{N}(0,1)$. Il fatto particolarmente rilevante è che questo avviene comunque sia stata scelta la legge delle X_k. Le Y_n^* però, pur non essendo Gaussiane, restano comunque tutte *standardizzate* ($\boldsymbol{E}[Y_n^*] = 0$, $\boldsymbol{V}[Y_n^*] = 1$) come le corrispondenti quantità del Teorema 3.23: infatti all'inizio della dimostrazione del Teorema 5.3 abbiamo visto che nelle nostre condizioni risulta

$$\boldsymbol{E}\left[\overline{X}_n\right] = \mu \qquad \boldsymbol{V}\left[\overline{X}_n\right] = \frac{\sigma^2}{n}$$

e quindi, tenendo conto di (4.8) e (4.19)

$$\begin{aligned} \boldsymbol{E}\left[Y_n^*\right] &= \boldsymbol{E}\left[\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}\right] = \frac{\sqrt{n}}{\sigma}\left(\boldsymbol{E}\left[\overline{X}_n\right] - \mu\right) = 0 \\ \boldsymbol{V}\left[Y_n^*\right] &= \boldsymbol{V}\left[\frac{\overline{X}_n - \mu}{\sigma}\sqrt{n}\right] = \frac{n}{\sigma^2}\boldsymbol{V}\left[\overline{X}_n\right] = 1 \end{aligned}$$

Pertanto le Y_n^* sono standardizzate, e in generale non Gaussiane, ma al limite per $n \to \infty$ esse convergono in distribuzione verso la legge normale standard $\mathfrak{N}(0,1)$ $\square$

In particolare, per n abbastanza grande, il *TLC* ci autorizza ad approssimare con la legge normale standard $\mathfrak{N}(0,1)$ la legge delle somme *standardizzate* Y_n^* di *v-a* indipendenti e arbitrarie, nel rispetto di limiti piuttosto vasti. Si noti infatti la generalità di questo risultato: nel Teorema 5.5, al di là della loro indipendenza, l'unica ipotesi richiesta sulle *v-a* X_k è che esse abbiano attesa e varianza finite, con $\sigma^2 > 0$. Per il resto la loro legge può essere del tutto arbitraria: un fatto molto utile soprattutto se tale legge non è nota. Così, se le X_k sono generiche *v-a* indipendenti, tutte con la stessa attesa μ e la stessa varianza $\sigma^2 > 0$, con le notazioni del *TLC* e per n abbastanza grande potremo sempre adottare la seguente approssimazione per la loro somma Y_n

$$\boldsymbol{P}\{Y_n \le x\} = \boldsymbol{P}\left\{\frac{Y_n - n\mu}{\sigma\sqrt{n}} \le \frac{x - n\mu}{\sigma\sqrt{n}}\right\} = \boldsymbol{P}\left\{Y_n^* \le \frac{x - n\mu}{\sigma\sqrt{n}}\right\} \simeq \Phi\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right)$$

Quando si usano relazioni di questo tipo diremo anche che stiamo adottando una **approssimazione normale**: la sua importanza sta nel fatto che, sotto le condizioni del *TLC*, essa permette di calcolare i valori della *FDC* di somme di generiche *v-a* indipendenti mediante l'uso delle tavole dei valori della *FDC* normale standard riportate nell'Appendice E.1. Si noti infine che, pur essendo $\mathfrak{N}(0,1)$ la legge di una *v-a continua*, il teorema resta vero anche se le X_k sono *v-a discrete*; in questo caso però, come vedremo in uno degli esempi seguenti, sarà bene usare dei piccoli accorgimenti per migliorare le approssimazioni

Resta da fare qualche osservazione circa il **numero minimo n** per il quale l'approssimazione normale del *TLC* si possa considerare applicabile a una somma Y_n. A questo proposito va subito detto che non ci sono risultati generali, e che bisogna adattarsi all'uso di alcune regole empiriche, peraltro piuttosto variabili secondo le fonti consultate. In genere l'approssimazione normale si considera applicabile quando n vale almeno 30-50. Anche qui però va adoperata qualche cautela nel caso in cui Y_n è una *v-a* discreta. Se ad esempio Y_n è binomiale $\mathfrak{B}(n;p)$ si può mostrare con dei calcoli espliciti che, quando p è vicino a 1 oppure a 0, l'approssimazione normale non è molto buona per i valori di n considerati accettabili con il nostro precedente criterio. In questi casi infatti bisogna tenere conto anche del valore di p, e in generale l'approssimazione normale si considera applicabile quando sono verificate *ambedue* le condizioni $np \ge 5$, e $n(1-p) \ge 5$. Così se ad esempio $p = 0.05$ si deve avere $n \ge 100$; se invece $p = 0.01$ si deve scegliere $n \ge 500$

Esempio 5.6. (Errori sperimentali) *Supponiamo di voler misurare una quantità fisica il cui valore vero μ è sconosciuto. A causa degli inevitabili errori sperimentali, però, i risultati delle osservazioni non saranno mai in generale coincidenti con μ, anzi saranno diversi ogni volta che ripeteremo la misura: abbiamo cioè a che fare con una vera e propria v-a X. Come suggerito dalla LGN, per stimare μ prenderemo allora in considerazione un campione casuale di n misure, ma per essere anche in grado di ottenere delle informazioni sull'attendibilità delle stime dovremmo anche conoscere quale è la legge di X: in questo modo sarebbe infatti possibile stabilire (come vedremo meglio nella parte di statistica inferenziale) con quale probabilità le nostre quantità aleatorie cadono in determinati intervalli*

A questo scopo, e supponendo per semplicità che non vi siano errori sistematici dovuti a difetti o distorsioni del sistema di misura, potremmo allora pensare che il valore di X sia composto di due parti: una deterministica costituita dal valore vero μ cercato, e una aleatoria W dovuta agli errori sperimentali, in modo che complessivamente risulti $X = \mu + W$. Inoltre W può essere considerata come la somma di innumerevoli piccoli disturbi dovuti alle condizioni sperimentali, e senza una direzione privilegiata: in questo caso – quale che sia la legge sconosciuta dei singoli disturbi – il TLC ci autorizza a ritenere che W segua una legge normale con media nulla e con una qualche varianza σ_W^2, cioè $W \sim \mathfrak{N}(0, \sigma_W^2)$. In base al Teorema 3.19 la v-a X risulta allora distribuita secondo la legge normale $\mathfrak{N}(\mu, \sigma_W^2)$. In questo modo abbiamo ricavato delle importanti informazioni sulla forma della legge di X la quale risulta Gaussiana, anche se restano da stimare i due parametri μ e σ_W^2. Questa è la base sulla quale si fonda la tradizionale ***legge degli errori*** *secondo la quale gli errori sperimentali si distribuiscono sempre in maniera Gaussiana*

Esempio 5.7. (Approssimazione normale di leggi discrete) *Per mostrare che nell'approssimazione normale di leggi discrete è bene adottare qualche cautela supplementare, torniamo a considerare il modello esaminato nell'Esempio 5.4 supponendo per semplicità che le proporzioni degli individui di tipo A e B siano uguali, in modo che $p = {}^1/_2$, e quindi*

$$\mu = \boldsymbol{E}[X_k] = p = \frac{1}{2} \qquad \sigma^2 = \boldsymbol{V}[X_k] = p(1-p) = \frac{1}{4}$$

Estraiamo allora un campione casuale di $n = 100$ individui, e poniamoci il problema di calcolare la probabilità che in esso ve ne siano più di 60 di tipo A. Con le notazioni dell'Esempio 5.4 sappiamo che il numero $N_A = X_1 + \ldots + X_{100}$ di individui di tipo A è una v-a binomiale $\mathfrak{B}(100; {}^1/_2)$, per cui la probabilità richiesta si scrive come

$$\boldsymbol{P}\{N_A > 60\} = \sum_{k=61}^{100} \binom{100}{k} \frac{1}{2^{100}}$$

Il valore numerico di questa probabilità non è però facilmente ricavabile, se non con l'ausilio di qualche macchina calcolatrice. L'approssimazione normale ci permette

invece di ottenerne una ragionevole stima con il semplice uso delle Tavole numeriche della FDC Normale standard Φ nell'Appendice E.1, ma con qualche opportuna precisazione

Osserviamo infatti che la v-a $N_A \sim \mathfrak{B}(100; {}^1/_2)$ non è altro che una somma analoga alla Y_n del Teorema 5.5 con $n = 100$. Applicando allora il TLC, la FDC $F_{100}(x)$ della v-a

$$N_A^* = \frac{N_A - n\mu}{\sigma\sqrt{n}} \qquad con \qquad n = 100 \qquad \mu = \sigma = \frac{1}{2}$$

potrà essere approssimata con la $\Phi(x)$ della Normale standard, e quindi

$$\begin{aligned} \boldsymbol{P}\{N_A > 60\} &= 1 - \boldsymbol{P}\{N_A \leq 60\} = 1 - \boldsymbol{P}\left\{\frac{N_A - n\mu}{\sigma\sqrt{n}} \leq \frac{60 - n\mu}{\sigma\sqrt{n}}\right\} \\ &= 1 - \boldsymbol{P}\left\{N_A^* \leq \frac{60-50}{5}\right\} = 1 - \boldsymbol{P}\{N_A^* \leq 2\} = 1 - F_{100}(2) \\ &\simeq 1 - \Phi(2) = 0.02275 \end{aligned}$$

avendo ricavato dalle Tavole dell'Appendice E.1 che $\Phi(2) = 0.97725$

Nel nostro caso, però, noi conosciamo anche il valore esatto (non approssimato) della probabilità richiesta e possiamo quindi esercitare qualche controllo sulla accuratezza della approssimazione. Calcolando allora $\boldsymbol{P}\{N_A > 60\}$ con la distribuzione binomiale otteniamo il valore 0.01760*, per cui il valore trovato con l'approssimazione normale non risulta particolarmente preciso, anche se la condizione $np = n(1-p) = 50 > 5$ è ampiamente rispettata. Questa discrepanza è dovuta al fatto che stiamo applicando il TLC a v-a discrete, cioè stiamo approssimando una funzione costante a tratti con un'altra funzione crescente e continua: infatti bisogna osservare che $N_A \sim \mathfrak{B}(100; {}^1/_2)$ assume solo i valori interi $1, 2, \ldots, 59, 60, 61, \ldots, 100$, e quindi che la sua FDC $F_{100}(x) = \boldsymbol{P}\{N_A \leq x\}$ resta costante per le x comprese fra due interi consecutivi, per cui ad esempio $\boldsymbol{P}\{N_A \leq 60\} = \boldsymbol{P}\{N_A \leq 60.5\}$. Ne segue che in particolare anche*

$$F_{100}\left(\frac{x - n\mu}{\sigma\sqrt{n}}\right) = F_{100}\left(\frac{x-50}{5}\right)$$

resta costante per $60 \leq x < 61$. Viceversa, come si vede dalle Tavole dell'Appendice E.1, nel medesimo intervallo il valore della funzione approssimante

$$\Phi\left(\frac{x-50}{5}\right)$$

cresce con continuità da $\Phi(2) = 0.97725$ a $\Phi(2.2) = 0.98610$, ed è ragionevole supporre che una funzione continua che approssima una funzione costante a tratti non ne attraversi i gradini in corrispondenza degli estremi, ma piuttosto in qualche punto compreso fra le discontinuità. È intuitivo, allora, capire che la scelta migliore

per il valore di x non è l'estremo inferiore 60 *dell'intervallo* $[60, 61)$*, usato nel calcolo precedente, ma il suo punto di mezzo* $x = 60.5$ *. Infatti, se eseguiamo di nuovo il calcolo approssimato tenendo conto di queste osservazioni, avremo*

$$\begin{aligned} \boldsymbol{P}\{N_A > 60\} &= 1 - \boldsymbol{P}\{N_A \le 60\} = 1 - \boldsymbol{P}\{N_A \le 60.5\} \\ &= 1 - \boldsymbol{P}\left\{\frac{N_A - n\mu}{\sigma\sqrt{n}} \le \frac{60.5 - n\mu}{\sigma\sqrt{n}}\right\} = 1 - F_{100}\left(\frac{60.5 - 50}{5}\right) \\ &= 1 - F_{100}(2.1) \simeq 1 - \Phi(2.1) = 0.01786 \end{aligned}$$

che costituisce un'approssimazione decisamente migliore per il valore esatto 0.01760 *. Come regola generale, quindi, se Y_n è una v-a a valori interi, si ottiene una migliore approssimazione normale se, invece di calcolare $\boldsymbol{P}\{Y_n \le k\}$ con k intero, si calcola piuttosto $\boldsymbol{P}\{Y_n \le k + \frac{1}{2}\}$*

5.4 Teorema di Poisson

Abbiamo già osservato che per leggi binomiali $\mathfrak{B}(n; p)$ con valori di p prossimi a 0 (oppure a 1) l'approssimazione normale è piuttosto problematica. D'altra parte – come vedremo in alcuni esempi – ci sono importanti casi in cui occorre considerare successioni di leggi binomiali per le quali quando $n \to \infty$ anche i valori di p variano e sono infinitesimi. Sarà importante allora enunciare il seguente risultato che fornisce uno strumento per approssimare le distribuzioni binomiali proprio nei casi in cui n è molto grande e p è molto piccola

Teorema 5.8. (Teorema di Poisson) *Dato $\lambda > 0$, e assegnate le leggi binomiali $\mathfrak{B}\left(n; {}^{\lambda}/_{n}\right)$ per n intero e $n > \lambda$, con le corrispondenti distribuzioni*

$$p_k(n) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \qquad k = 0, 1, \ldots, n$$

esse convergono in distribuzione per $n \to \infty$ verso la legge di Poisson $\mathfrak{P}(\lambda)$

$$\mathfrak{B}\left(n; {}^{\lambda}/_{n}\right) \xrightarrow{d} \mathfrak{P}(\lambda)$$

ovvero, secondo la (5.5), risulta

$$\lim_n p_k(n) = \frac{\lambda^k}{k!}\, e^{-\lambda} \qquad k = 0, 1, 2, \ldots$$

Dimostrazione: Osserviamo innanzitutto che la richiesta di avere $n > \lambda$ è motivata dal fatto che il numero ${}^{\lambda}/_{n}$ deve giocare il ruolo della probabilità p di una legge binomiale $\mathfrak{B}(n; p)$, e quindi deve sempre risultare ${}^{\lambda}/_{n} \le 1$. D'altra parte tale richiesta non è limitativa della generalità del teorema perché noi siamo interessati a studiare

un limite per $n \to \infty$, e quindi nulla ci impedisce di prendere in considerazione solo $n > \lambda$. Ora si ha

$$\begin{aligned} p_k(n) &= \binom{n}{k} \left({}^{\lambda}/_{n}\right)^k \left(1 - {}^{\lambda}/_{n}\right)^{n-k} = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!} \left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right) \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} \end{aligned}$$

e siccome inoltre, per ogni fissato k, sono noti i seguenti limiti

$$\begin{aligned} \lim_n \left[\left(1 - \frac{1}{n}\right) \dots \left(1 - \frac{k-1}{n}\right)\right] &= 1 \\ \lim_n \left(1 - \frac{\lambda}{n}\right)^{-k} &= 1 \\ \lim_n \left(1 - \frac{\lambda}{n}\right)^n &= e^{-\lambda} \end{aligned}$$

il risultato richiesto ne segue immediatamente □

Il Teorema 5.8 afferma in pratica che se n è molto grande e $p = {}^{\lambda}/_{n}$ molto piccola è possibile adottare l'**approssimazione di Poisson**: la distribuzione $p_k(n)$ di una legge binomiale $\mathfrak{B}(n; p)$ con n grande e p piccola può essere approssimata con quella di una legge di Poisson $\mathfrak{P}(\lambda) = \mathfrak{P}(np)$ con $\boldsymbol{\lambda = np}$, ovvero

$$\binom{n}{k} p^k (1-p)^{n-k} \simeq e^{-np} \frac{(np)^k}{k!} \qquad n \to \infty \qquad p \to 0$$

Illustreremo ora con alcuni esempi le conseguenze di questo risultato

Esempio 5.9. *Supponiamo di partecipare un gran numero n di volte a una lotteria per la quale la probabilità p di vincere è molto piccola, e chiediamoci quale sarà la probabilità di vincere una o più volte. Se $X_1, \dots, X_n$ sono le n v-a indipendenti di Bernoulli $\mathfrak{B}(1; p)$ tali che*

$$X_j = \begin{cases} 1 & \textit{se vinco al } j\textit{-mo tentativo} \\ 0 & \textit{altrimenti} \end{cases} \qquad j = 1, 2, \dots, n$$

il numero aleatorio di vincite $Y = X_1 + \dots + X_n$ sarà binomiale $\mathfrak{B}(n; p)$ per il Teorema 3.16. Se però prendessimo $n = 1\,000$ e $p = 0.003$, il calcolo delle binomiali $\mathfrak{B}(1\,000; 0.003)$ sarebbe molto laborioso, mentre quello delle Poisson $\mathfrak{P}(np) = \mathfrak{P}(3)$ è più semplice: ad esempio, usando Tavola E.5, la probabilità di vincere 4 volte è

$$\boldsymbol{P}\{Y = 4\} = \binom{1\,000}{4} (0.003)^4 (0.997)^{996} \simeq e^{-3} \frac{3^4}{4!} = 0.168$$

Esempio 5.10. (Istanti aleatori) *Supponiamo di voler determinare la legge del numero aleatorio X di telefonate che arrivano ad un centralino telefonico in un intervallo di tempo T in una tipica giornata lavorativa. Innanzitutto è chiaro che X è un conteggio, cioè una v-a discreta che può assumere tutti i valori interi $k = 0, 1, 2, \dots$ senza che vi sia una limitazione superiore. Per trovare la legge di X possiamo allora cominciare costruendo un modello approssimato: supponendo di sapere che mediamente arrivano $\alpha > 0$ telefonate per unità di tempo, dividiamo l'intervallo T in n sotto-intervalli di uguale lunghezza T/n, prendendo n abbastanza grande da fare in modo che il numero medio di telefonate $\alpha T/n$ in ogni sotto-intervallo sia minore di 1. Potremo supporre allora, in prima approssimazione, che in ogni intervallo di lunghezza T/n non arrivi più di una telefonata: ovviamente questa supposizione sarà tanto più realistica quanto più grande prenderemo n, e quindi anche al limite per $n \to \infty$. Se allora definiamo n v-a indipendenti $X_1, \dots, X_n$ tali che*

$$X_j = \begin{cases} 1 & \text{se nell'intervallo } j\text{-mo arriva una telefonata} \\ 0 & \text{altrimenti} \end{cases} \qquad j = 1, 2, \dots, n$$

per costruzione esse saranno tutte v-a di Bernoulli con $\boldsymbol{E}[X_j] = \alpha T/n$, e quindi anche (vedi Teorema 4.13) con $\boldsymbol{P}\{X_j = 1\} = \alpha T/n$, sicché posto $\lambda = \alpha T$ avremo $X_j \sim \mathfrak{B}\left(1; \lambda/n\right)$. In questa approssimazione quindi il numero totale di telefonate che arriva in T sarà $X_1 + \cdots + X_n$ e questa v-a, per il Teorema 3.16, sarà Binomiale $\mathfrak{B}\left(n; \lambda/n\right)$. Siccome abbiamo notato che l'approssimazione migliora al limite per $n \to \infty$, il Teorema di Poisson 5.8 ci dice che la legge di X – in quanto limite delle $\mathfrak{B}\left(n; \lambda/n\right)$ – è la legge di Poisson $\mathfrak{P}(\lambda)$. In conclusione il numero X di telefonate che arrivano al nostro centralino telefonico è una v-a di Poisson $\mathfrak{P}(\lambda)$ con $\lambda = \alpha T$: naturalmente resta da studiare in che modo si può stimare il valore del parametro λ a partire da osservazioni empiriche reali

Sarà infine utile precisare che questo risultato è molto generale e non si applica solo al numero di telefonate che arrivano a un centralino in un determinato intervallo di tempo T; infatti anche le v-a X che in T contano quante particelle sono emesse da un campione radioattivo, ovvero quanti clienti si presentano a uno sportello bancario, o ancora quanti incidenti capitano su una rete stradale e così via, seguono tutte una legge di Poisson $\mathfrak{P}(\lambda)$ con un λ opportuno. Così, se ad esempio sappiamo che un campione radioattivo emette mediamente $\alpha = 12$ particelle all'ora, e se X rappresenta il numero di particelle emesse in un intervallo di $T = 10$ minuti, avremo $X \sim \mathfrak{P}(2)$ dato che

$$\lambda = \alpha T = \frac{12}{60'} \times 10' = 2$$

e quindi

$$\boldsymbol{P}\{X = k\} = e^{-2}\,\frac{2^k}{k!} = 0.135\,\frac{2^k}{k!}$$

dove il valore (arrotondato) di e^{-2} è stato preso dalle Tavole E.5

Parte II

Statistica

Capitolo 6

Statistica descrittiva univariata

Il **calcolo delle probabilità** è una teoria che ottiene i propri risultati costruendo modelli aleatori basati su opportune ipotesi la cui validità è successivamente giudicata nel confronto con i risultati sperimentali. La **statistica**, invece, è un po' l'altra faccia della medesima medaglia: essa infatti parte dai dati empirici e cerca di estrarne l'informazione ritenuta rilevante. Bisogna però prima domandarsi se tali dati empirici sono giudicati o meno una visione esauriente del fenomeno osservato: come ricordato nella Prefazione, infatti, i risultati di una misura possono essere considerati o come la totalità degli oggetti della nostra indagine (si pensi ai risultati di una tornata elettorale), ovvero come un campione casuale estratto da una popolazione più ampia (come per i sondaggi pre-elettorali). Mentre è evidente che nel secondo caso (oggetto della **statistica inferenziale** che incontreremo più avanti) la probabilità gioca un ruolo determinante, quest'ultima è invece sostanzialmente estranea alla **statistica descrittiva** che meglio si adatta alla prima eventualità. Noi cominceremo la nostra discussione proprio con la statistica descrittiva, ma sarà comunque utile osservare preliminarmente che in essa saranno inevitabilmente introdotti concetti, notazioni e risultati che si svilupperanno in un evidente parallelismo con analoghi concetti, notazioni e risultati già definiti nella parte di calcolo delle probabilità. Sarà cura dell'autore sottolineare somiglianze e differenza, ma il lettore è fin da ora invitato ad esercitare la dovuta attenzione per evitare possibili confusioni

6.1 Dati e frequenze

Il punto di partenza della statistica è costituito dai dati empirici che distingueremo innanzitutto in due categorie: dati quantitativi e dati qualitativi. I **dati quantitativi o numerici** sono i risultati di misure che si presentano sotto forma di valori numerici: ad esempio posizioni, velocità, masse di particelle; peso, altezza, età di individui di una specie di animali; reddito dei cittadini di un paese e così via. I **dati qualitativi** viceversa si riferiscono a proprietà astratte, e non sono in genere rappresentati da numeri: ad esempio i colori delle palline estratte da un'urna; il gruppo sanguigno A, B, AB e 0 di un insieme di persone; il partito politico votato

dagli elettori di un paese e via dicendo. Si noti però che la differenza principale fra i due tipi di dati non consiste nel fatto formale di essere rappresentati o meno da numeri: in fondo potremmo convenzionalmente rappresentare anche i colori, i gruppi sanguigni e i partiti degli esempi precedenti con dei numeri. Quel che invece è profondamente diverso è il significato di questi numeri. Così ad esempio, il colore rosso delle palline di un'urna potrebbe essere per comodità rappresentato con un numero, ma questa sarebbe un'associazione convenzionale e il numero scelto (1, 2 o indifferentemente qualsiasi altro numero) non modificherebbe in nulla l'esito della nostra analisi. Invece i numeri che rappresentano i redditi dei cittadini di un paese non possono essere assegnati arbitrariamente senza perdere completamente il loro significato. Questa differenza è anche alla base del fatto che taluni indicatori statistici (medie, mediane, varianze ...) hanno un senso solo nel caso di dati quantitativi e non in quelli di dati qualitativi: ad esempio è perfettamente sensato chiedersi quale è il reddito medio dei cittadini di un paese, mentre non avrebbe alcun significato il concetto di colore medio, o di partito medio, e questo anche se i dati qualitativi fossero rappresentati da numeri

Nel linguaggio della statistica l'insieme dei soggetti presi in considerazione nella discussione di un determinato problema (animali di una specie, palline in un'urna, cittadini di un paese) costituisce una **popolazione**, mentre le caratteristiche $X, Y, \ldots$ che vengono osservate (colore, gruppo sanguigno, reddito, peso ...) prendono il nome di **caratteri** e si distinguono in caratteri **quantitativi o numerici**, e caratteri **qualitativi** in base al tipo di dati ricavati dalle osservazioni. I caratteri numerici sono poi a loro volta distinti in altre due categorie: quelli che assumano valori **discreti** (ad esempio il numero di figli delle famiglie di un dato paese), e quelli che assumono valori **continui** (il peso o l'altezza degli individui di una popolazione). I possibili valori assunti da caratteri numerici discreti (numero dei figli di una famiglia), o da caratteri qualitativi (colori delle palline in un'urna) si chiamano anche **modalità**. Così ad esempio: il peso dei cittadini di un paese è un carattere numerico continuo; il gruppo sanguigno degli individui di un gruppo è un carattere qualitativo con 4 modalità (A, B, AB e 0); il numero di figli delle famiglie di un paese è un carattere numerico discreto le cui modalità sono i numeri interi, e così via. Come vedremo meglio nel Capitolo 7, infine, potremo avere anche **dati multidimensionali**, nel senso che su ogni individuo considerato si possono misurare due o più caratteri. Ad esempio se si misurano l'età e, il peso p e il reddito r dei cittadini di un dato paese, ad ogni individuo sarà associata una terna di numeri (e, p, r): ma questo sarà oggetto della statistica descrittiva multivariata

Nell'ambito della statistica descrittiva si suppone sempre di avere a disposizione i dati relativi a tutta la popolazione di nostro interesse che, quindi, dovrà contenere un numero finito n di individui. I nostri insiemi di dati, chiamati **campioni**, saranno pertanto n-ple di simboli (eventualmente anche di numeri) del tipo $x_1, \ldots, x_n$, ed esauriranno tutta la popolazione considerata. Si noti, però, a questo proposito che in genere il procedimento avviene in senso inverso: si parte dai dati e poi si stabilisce – in base alle necessità dello sperimentatore – quale è la popolazione di

riferimento. Tipicamente nella realtà il punto di partenza è il campione $x_1, \ldots, x_n$, ma il suo significato può variare secondo il punto di vista adottato. In un certo senso è lo statistico che, in base alle proprie esigenze, stabilisce quale è la popolazione di riferimento: se egli decide che la popolazione di interesse è rappresentata *solo* dagli n dati a sua disposizione, allora egli si colloca nell'ambito della statistica descrittiva; se invece egli considera gli n dati come un campione estratto da una popolazione più vasta sulla quale vuole ricavare delle informazioni, allora si colloca nell'ambito della statistica inferenziale che studieremo in seguito

Così ad esempio potremmo supporre di avere i risultati $x_1, \ldots, x_n$ del test d'ingresso di n studenti di un corso di laurea universitario. Se il nostro scopo è quello di esaminare il livello di preparazione degli studenti che che si sono iscritti a quel corso di laurea in quell'anno accademico, è evidente che la nostra popolazione sarà ristretta proprio agli n individui che hanno sostenuto il test e non ci resterà che studiare i dati in quest'ottica. Se invece dai risultati del test volessimo ricavare delle conclusioni più generali – ad esempio relative a tutta la popolazione studentesca che accede all'università in un determinato anno accademico – è altrettanto evidente che $x_1, \ldots, x_n$ dovrà essere considerato come un campione (casuale) estratto da una popolazione più vasta. Il medesimo insieme di dati, insomma, può essere legittimamente studiato da tutti e due i punti di vista: nel primo caso, però, si tratterà di un problema di statistica descrittiva; nel secondo di un problema di statistica inferenziale

Supponiamo allora di voler studiare innanzitutto un **carattere numerico discreto, o qualitativo** X con un numero finito M di modalità esaminando una popolazione di n individui. Indicheremo nel seguito le M modalità con i simboli w_k per $k = 1, \ldots, M$: nel caso di caratteri numerici discreti tali w_k saranno numeri; per caratteri qualitativi potranno invece avere significati più astratti

Definizione 6.1. *Detto $\{j : x_j = w_k\}$ l'insieme degli individui della nostra popolazione che assumono il valore w_k, chiameremo* ***frequenza assoluta*** *della k-ma modalità di un carattere numerico discreto o qualitativo X il numero*

$$N_k = \#\{j : x_j = w_k\} \qquad k = 1, \ldots, M \tag{6.1}$$

(qui # indica la cardinalità di un dato insieme) cioè il numero delle x_j uguali a w_k; chiameremo invece ***frequenza relativa*** *il numero*

$$p_k = \frac{N_k}{n}\,, \qquad k = 1, \ldots, M \tag{6.2}$$

cioè la frazione delle x_j che assumono il valore w_k. Per i caratteri numerici si dicono poi ***frequenze assolute cumulate*** *e* ***frequenze relative cumulate*** *le quantità*

$$F_k = \sum_{i=1}^{k} N_i\,, \qquad f_k = \sum_{i=1}^{k} p_i\,, \qquad k = 1, \ldots, M \tag{6.3}$$

ovvero il numero e la frazione delle x_j con valore minore o uguale di w_k

Naturalmente banalmente sono verificate per definizione le due **relazioni di normalizzazione**

$$N_1 + \ldots + N_M = n \qquad\qquad p_1 + \ldots + p_M = 1 \tag{6.4}$$

che per le frequenze cumulate divengono semplicemente

$$F_M = n \qquad\qquad f_M = 1$$

Quando invece X è un **carattere numerico continuo** i suoi possibili valori sono infiniti, e tipicamente non numerabili per cui il concetto di *modalità* perde di senso, e per introdurre il concetto di *frequenza* bisognerà procedere in modo diverso. Siccome i valori osservati costituiscono un campione numerico $x_1, \ldots, x_n$ con n finito, essi cadranno sicuramente in qualche intervallo chiuso e limitato del tipo $[a, b]$, e noi li potremo ripartire in opportune **classi** nel modo seguente: consideriamo innanzitutto una decomposizione finita $(\mathcal{B}_k)_{k=1,\ldots,M}$ di $[a, b]$; cioè suddividiamo $[a, b]$ in M sotto-intervalli (non necessariamente tutti della stessa ampiezza, ma numerati nell'ordine) $\mathcal{B}_k = [a_k, b_k]$ con $k = 1, \ldots, M$ disgiunti e tali che la loro unione coincida con $[a, b]$. Possiamo allora adeguare al nuovo caso la Definizione 6.1

Definizione 6.2. *Chiameremo rispettivamente* ***frequenza assoluta*** *e* ***frequenza relativa*** *dei ritrovamenti di un carattere numerico continuo X nelle classi $(\mathcal{B}_k)_{k=1,\ldots,M}$ che decompongono $[a, b]$ i numeri*

$$N_k = \#\{j : x_j \in \mathcal{B}_k\} \qquad p_k = \frac{N_k}{n} \qquad k = 1, \ldots, M$$

Le corrispondenti ***frequenze cumulate assoluta*** *F_k e* ***relativa*** *f_k si definiscono poi come in Definizione 6.1 tramite le (6.3), e indicano rispettivamente il numero o la frazione di dati x_j che cadono all'interno dell'unione dei primi k sotto-intervalli, ovvero che sono minori o uguali dell'estremo destro di $\mathcal{B}_k$*

Ovviamente le relazioni di normalizzazione (6.4) sono rispettate dalle frequenze assolute e relative anche in questo caso. Si noti inoltre che una certa importanza è rivestita anche dal concetto di **valore centrale** di una classe $\mathcal{B}_k = [a_k, b_k]$, cioè la semisomma degli estremi dell'intervallo

$$\widehat{w}_k = \frac{a_k + b_k}{2} \tag{6.5}$$

Infatti, per successivi scopi di analisi statistica, tali valori centrali $\widehat{w}_k$ potranno essere utilizzati per giocare il ruolo di modalità w_k approssimate, nel senso che i numeri x_k che cadono in $\mathcal{B}_k$ saranno approssimativamente identificati proprio con $\widehat{w}_k$

In conclusione anche nel caso di caratteri X numerici continui potremo introdurre le medesime nozioni introdotte nel caso discreto/qualitativo, a patto di sostituire il conteggio del numero di volte in cui X assume la modalità k-ma con il conteggio del

3	0	3	1	1	1	2	4	1	3	2	1	0	2	1	3	3	0	2	1
3	4	3	1	3	4	1	5	0	2	0	4	1	4	2	2	2	1	2	3
2	3	2	2	3	3	2	1	2	1										

Tabella 6.1: Campione di $n = 50$ misure di un carattere con le 6 modalità $k = 0, 1, 2, 3, 4, 5$.

k	0	1	2	3	4	5
N_k	5	13	14	12	5	1
F_k	5	18	32	44	49	50
p_k	0.10	0.26	0.28	0.24	0.10	0.02
f_k	0.10	0.36	0.64	0.88	0.98	1.00

Tabella 6.2: Frequenze e frequenze cumulate, assolute e relative, per i dati riportati in Tabella 6.1.

numero di ritrovamenti di X nella classe k-ma. Bisogna però osservare subito che in questo secondo caso i valori delle frequenze dipendono dalla collocazione e dalla ampiezza delle classi $\mathcal{B}_k$, che sono scelte in maniera largamente arbitraria. Come vedremo in alcuni esempi successivi, infatti, la determinazione della collocazione e delle ampiezze delle classi può rivelarsi cruciale per mettere in evidenza (o per nascondere) alcune caratteristiche dei dati

6.2 Tabelle e grafici

L'informazione contenuta nelle frequenze assolute e relative può essere messa meglio in evidenza organizzando i dati in tabelle o anche rappresentandoli in grafici. Le tecniche di organizzazione e visualizzazione dei dati sono numerose, e la scelta di quelle più opportune dipende dal particolare problema studiato. Noi qui ne daremo solo qualche esempio senza nessuna pretesa di completezza.

Le **tabelle di frequenza** non sono altro che opportune tabelle nelle quali sono riportati in maniera organizzata i valori numerici delle varie frequenze definite in precedenza. Per la rappresentazione grafica dei *caratteri qualitativi o numerici discreti* lo strumento più usato è il **diagramma a barre** che consiste semplicemente nel riportare in corrispondenza di ogni singola modalità delle **barre** di altezza uguale ai valori delle frequenze. Su questi diagrammi possono essere rappresentate sia le frequenze assolute che quelle relative: siccome a causa di (6.2) N_k e p_k sono tutti numeri proporzionali fra loro, i diagrammi a barre dei due casi sono identici, l'unica differenza essendo la scala dei valori dell'asse verticale. Per le frequenze dei *caratteri numerici continui* invece si costruiscono degli **istogrammi**. Il principio è simile a quello dei diagrammi a barre, ma con una importante differenza: sulla classe k-ma (sottointervallo $\mathcal{B}_k = [a_k, b_k]$ dell'intervallo $[a, b]$) si costruisce un rettangolo

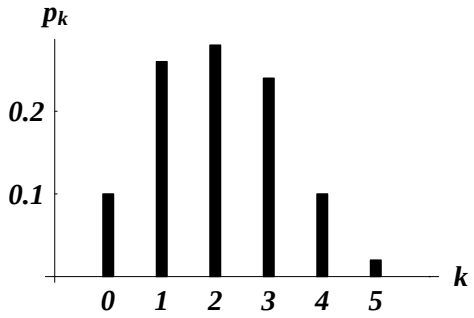


Figura 6.1: Diagramma a barre delle frequenze relative p_k dei dati della Tabella 6.1.

la cui **area** è uguale al valore della k-ma frequenza assoluta o relativa. Siccome le ampiezze $|\mathcal{B}_k| = b_k - a_k$ delle varie classi (basi dei rettangoli) possono essere diverse, in generale le altezze dei rettangoli non saranno più proporzionali alle frequenze assolute o relative, ma saranno rispettivamente

$$H_k = \frac{N_k}{b_k - a_k} \qquad h_k = \frac{p_k}{b_k - a_k} \tag{6.6}$$

A parità di frequenze, quindi, classi molto ampie tenderanno ad avere rettangoli più bassi, e viceversa. Solo nel caso in cui le ampiezze $|\mathcal{B}_k|$ fossero scelte tutte uguali le altezze dei rettangoli sarebbero nuovamente proporzionali alle frequenze (assolute o relative) delle classi. Noteremo infine che anche le *frequenze cumulate* sono ovviamente suscettibili di rappresentazioni grafiche che però noi, per brevità, trascureremo limitandoci a riportare i loro valori nelle tabelle di frequenza

Esempio 6.3. *Supponiamo di aver raccolto* $n = 50$ *misure di un carattere con* $M = 6$ *modalità che qui per comodità rappresenteremo senz'altro con i numeri* $k = 0, 1, 2, 3, 4, 5$. *Come esempio concreto possiamo pensare di aver esaminato 50 famiglie con 5 figli e di aver registrato per ciascuna di esse il numero dei figli maschi che ovviamente è un numero intero da 0 a 5; alternativamente potremmo pensare di aver lanciato 50 volte 5 monete e di aver registrato in ogni lancio il numero delle teste. I dati di partenza del nostro esempio sono mostrati nella Tabella 6.1. È facile a questo punto calcolare le frequenze assolute e relative da (6.1) e (6.2): i risultati sono riportati nella Tabella 6.2. Le frequenze possono poi essere rappresentate in un diagramma a barre come quello di Figura 6.1. Per evitare ripetizioni abbiamo scelto di riportare solo il diagramma a barre delle frequenze relative: quello delle frequenze assolute sarebbe identico, tranne che per la scala dei valori dell'asse verticale (invece di* 0.1 *e* 0.2 *troveremmo rispettivamente* 5 *e* 10*)*

Esempio 6.4. *Supponiamo di avere le* $n = 100$ *misure di un carattere continuo* X *riportate nella Tabella 6.3. Ad esempio potrebbero essere – in una opportuna unità di misura – misure di lunghezza su un gruppo di insetti, ovvero misure di massa per*

0.30	1.03	1.08	1.22	1.46	1.62	2.01	2.17	2.27	2.31
2.33	2.41	2.49	2.49	2.57	2.58	2.59	2.63	2.75	2.75
2.84	2.93	2.95	3.08	3.09	3.23	3.27	3.27	3.28	3.37
3.39	3.42	3.47	3.49	3.56	3.60	3.78	3.78	3.79	3.87
3.91	3.91	3.95	3.95	3.96	4.02	4.11	4.12	4.12	4.22
4.31	4.35	4.58	4.69	4.76	4.89	5.12	5.18	5.20	5.34
5.34	5.37	5.40	5.46	5.54	5.62	5.64	5.64	5.68	5.71
5.73	5.94	6.10	6.19	6.24	6.28	6.31	6.33	6.35	6.40
6.44	6.44	6.55	6.56	6.63	6.68	6.73	6.75	6.89	6.99
7.01	7.08	7.11	7.15	7.26	7.44	7.47	7.93	8.21	8.44

Tabella 6.3: Campione di $n = 100$ misure di un carattere continuo X. Per comodità i dati sono stati riportati in ordine crescente. La coincidenza di alcuni dei valori – particolarmente improbabile nel caso di caratteri continui – è dovuta agli arrotondamenti effettuati.

$\mathcal{B}_k$	N_k	F_k	p_k	f_k	h_k
[0.0, 2.0]	6	6	0.06	0.06	0.03
[2.0, 4.0]	39	45	0.39	0.45	0.20
[4.0, 6.0]	27	72	0.27	0.72	0.14
[6.0, 8.0]	26	98	0.26	0.98	0.13
[8.0, 10.0]	2	100	0.02	1.00	0.01

Tabella 6.4: Frequenze e altezze h_k dell'istogramma di Tabella 6.3 per 5 classi di ampiezza 2.0

le particelle elementari prodotte in un determinato esperimento e così via. Ovviamente nella realtà i valori non si ottengono nell'ordine crescente nel quale li abbiamo riportati; noi però abbiamo preferito riordinare il campione perché questo facilita il calcolo delle frequenze senza modificarne il valore. La tabella delle frequenze dipende però ora dalle classi scelte. Si vede subito che i dati cadono tutti fra 0.30 *e* 8.44*, e per rendere più simmetriche le classi possiamo, ad esempio, considerare un intervallo un po' più ampio del tipo* $[0, 10]$*. Per semplicità sceglieremo prima di tutto classi con la stessa ampiezza, e cominceremo con il dividere* $[0, 10]$ *in* 5 *sotto-intervalli di ampiezza* 2.0*. In questo caso le frequenze sono quelle della Tabella 6.4. Se invece scegliessimo come ampiezza delle classi* 0.5 *o* 0.1 *otterremmo* 20 *o* 100 *classi con frequenze piuttosto diverse che per brevità non riporteremo in tabelle, limitandoci solo alla loro successiva rappresentazione grafica. Si può passare a questo punto a costruire gli istogrammi corrispondenti a ciascuna scelta delle classi, ricordando che su ogni classe dovrà essere disegnato un rettangolo di* area *uguale alla rispettiva frequenza. Gli istogrammi ottenuti con le tre scelte delle classi (di ampiezze rispettivamente* 2.0*,* 0.5 *e* 0.1*) sono riportati nella Figura 6.2*

Si noterà che l'aspetto dei tre istogrammi è piuttosto diverso: quello con le classi più ampie (ampiezza 2.0*, in alto a sinistra) fornisce una rappresentazione piuttosto*

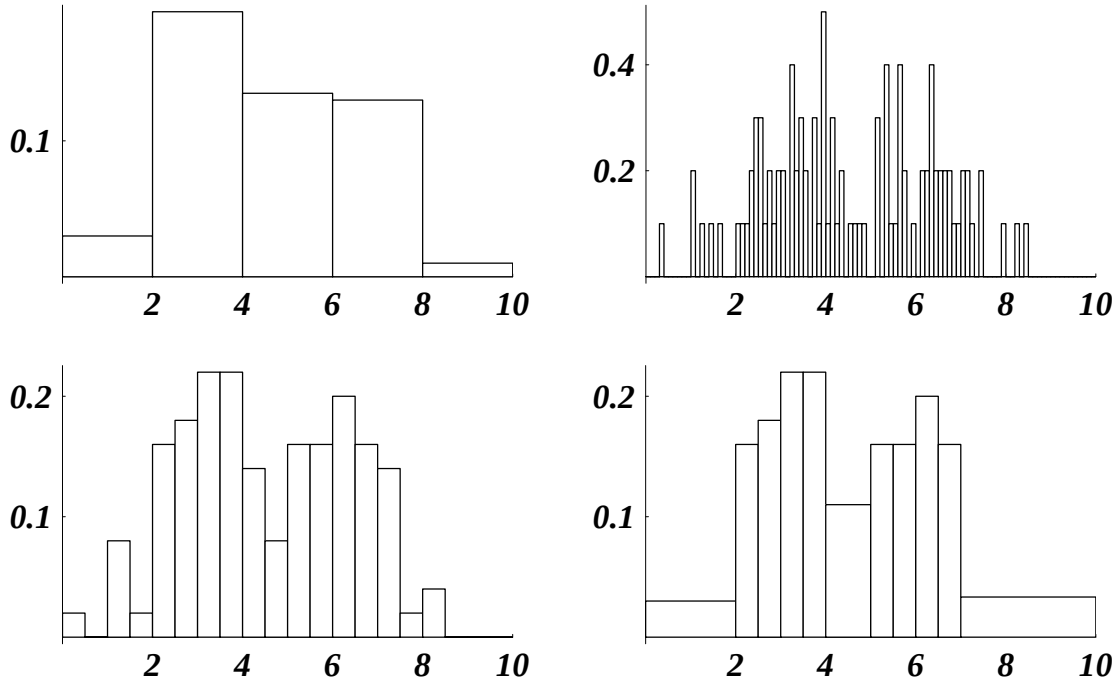


Figura 6.2: Istogrammi dei dati riportati in Tabella 6.3. I due istogrammi in alto si riferiscono a classi di ampiezze rispettivamente 2.0 e 0.1. L'istogramma in basso a sinistra è invece costruito con classi di ampiezza 0.5, mentre quello in basso a destra è costruito con classi di ampiezza variabile.

grossolana, mentre quello con le classi meno ampie (ampiezza 0.1*, in alto a destra) dà una rappresentazione piuttosto confusa. Viceversa l'istogramma con classi di ampiezza* 0.5 *(in basso a sinistra) sembra avere un aspetto più equilibrato, e mostra alcune caratteristiche dei dati che non appaiono negli altri due: in particolare esso indica che le frequenze presentano due massimi relativi in corrispondenza delle classi* [3.0, 3.5]*,* [3.5, 4.0] *e* [6.0, 6.5]*. Questa struttura dell'istogramma è interessante da un punto di vista statistico in quanto potrebbe indicare che la nostra popolazione è in realtà composta della sovrapposizione di due popolazioni con proprietà differenti: una con valori del carattere prevalentemente compresi fra* 3 *e* 4*, e l'altra con valori del carattere prevalentemente vicini a* 6*. Tale suggerimento va invece completamente perduto nelle rappresentazioni con intervalli troppo larghi o troppo stretti*

Infine sempre nella stessa Figura 6.2 è stato riportato un quarto istogramma dello stesso campione costruito con classi di ampiezze diverse fra loro. In particolare, con riferimento al terzo istogramma con i 20 *intervalli di ampiezza* 0.5*, sono stati unificati gli intervalli con le frequenze più basse che si trovano a sinistra, a destra e al centro, evitando così di riportare oscillazioni poco significative. Si vede in definitiva che la scelta delle classi modifica l'aspetto dell'istogramma, volta a volta mettendo in evidenza o nascondendo alcune caratteristiche dei dati. Non ci sono però delle*

regole per scegliere le classi nella maniera migliore, e d'altra parte non è detto che quel che viene messo in evidenza da un particolare istogramma sia poi in realtà statisticamente significativo. Il ricercatore avveduto, guidato dalla sua esperienza, farà diversi tentativi, e cercherà successivamente delle conferme per le conclusioni suggerite dalle diverse rappresentazioni dei suoi dati

6.3 Moda, media e varianza

L'analisi statistica non si esaurisce nella rappresentazione delle frequenze dei dati: un altro importante aspetto da esaminare è la ricerca di opportuni indici che permettano concentrare in pochi numeri le caratteristiche più rilevanti dei campioni. Sono di particolare importanza gli **indici di centralità** e gli **indici di dispersione**. I primi forniscono un'idea dei valori attorno ai quali sono prevalentemente concentrati i dati empirici; i secondi misurano invece la dispersione dei campioni attorno ai valori centrali. In questa e nelle successive sezioni esamineremo, senza pretesa di completezza, alcuni dei principali indici statistici, iniziando da quelli di centralità

Definizione 6.5. *Data un campione di un carattere qualitativo o numerico discreto chiameremo* ***mode*** *le modalità con frequenze più grandi di quelle immediatamente vicine. Nel caso invece di caratteri numerici continui chiameremo* ***classi modali*** *quelle che nell'istogramma corrispondono a rettangoli più alti di quelli immediatamente vicini: in questo secondo caso spesso la classe modale viene anche identificata con il suo valore centrale (6.5). Le mode e le classi modali corrispondono quindi ai* ***massimi locali*** *rispettivamente dei diagrammi a barre e degli istogrammi*

Il concetto di moda in statistica è quindi del tutto analogo a quello trovato in probabilità sulla base delle Definizioni 3.15 e 3.18 se, almeno nel caso continuo, si sostituiscono le *fdp* con gli istogrammi

Esempio 6.6. *Per il campione dell'Esempio 6.3, rappresentato nel diagramma a barre di Figura 6.1, l'unica moda è* 2*, cioè la modalità la cui frequenza assoluta* 14*, ovvero quella relativa* 0.28*, è la più grande di tutte*

Nel caso del campione dell'Esempio 6.4, invece, l'identificazione della classe modale è un po' più delicata. Intanto è chiaro dagli istogrammi di Figura 6.2 che le classi modali dipendono fortemente dalla scelta delle classi stesse. In secondo luogo questi grafici mettono bene in evidenza come possa capitare di avere campioni che presentano più di una moda o classe modale: le mode o le classi modali coincidono infatti con i massimi locali delle rappresentazioni grafiche dei campioni, piuttosto che con un unico massimo assoluto, e quindi un insieme di dati può presentare anche più di una moda o classe modale. Tornando alla Figura 6.2 vediamo allora che per l'istogramma in alto a sinistra la classe modale è l'intervallo [2.0, 4.0]*, ovvero il suo valore centrale* 3*; per ciascuno dei due istogrammi in basso, invece, ci sono due classi modali: l'intervallo* [3.0, 4.0] *(che in realtà è l'unione di due classi) ovvero il suo valore centrale* 3.5*, e l'intervallo* [6.0, 6.5] *ovvero il suo valore centrale* 6.25*.*

Infine va notato l'istogramma in alto a destra presenta un eccessivo numero di massimi locali che lo rendono confuso e poco adatto ad un'analisi statistica: in questo caso, anche se tutti quei massimi locali si qualificherebbero tecnicamente come classi modali, è piuttosto evidente che si tratterebbe di identificazioni poco significative dovute a una cattiva scelta della suddivisione in classi. Ancora una volta quindi, come già notato nell'Esempio 6.4, vediamo che classi eccessivamente larghe conducono ad un'analisi troppo grossolana che può far perdere informazioni (qui la bi-modalità del campione), mentre viceversa classi troppo ristrette producono un'analisi eccessivamente rumorosa e poco significativa. In definitiva solo una scelta ben bilanciata delle classi consente di mettere in evidenza le caratteristiche salienti del campione

Definizione 6.7. *Si chiama* ***media aritmetica****, o semplicemente* ***media*** *di un campione $x_1, \dots, x_n$ del carattere numerico X la quantità*

$$\boxed{m_X = \overline{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{j=1}^{n} x_j}$$

Se poi si assegnano anche dei ***pesi*** *$q_1 \dots, q_n$ tali che*

$$0 \le q_j \le 1 \qquad j = 1, \dots, n \qquad\qquad q_1 + \ldots + q_n = 1$$

si chiama ***media pesata*** *il numero*

$$\sum_{j=1}^{n} q_j x_j$$

I pesi di una media pesata sono una misura dell'importanza relativa dei dati nella media: la media aritmetica è un caso particolare di media pesata quando tutti i pesi sono uguali fra loro, cioè $q_j = {}^1/_n$, e i dati hanno tutti la stessa importanza

Teorema 6.8. *Dato un campione $x_1, \dots, x_n$ di un carattere numerico discreto X con modalità $w_1, \dots, w_M$, e dette p_k le frequenze relative di tali modalità, si ha*

$$\boxed{m_X = \overline{x} = \sum_{k=1}^{M} p_k w_k} \tag{6.7}$$

ovvero la media di un campione numerico discreto è anche la media pesata delle sue modalità w_k, usando come pesi le frequenze relative p_k

Dimostrazione: Basterà osservare che per (6.2) $np_k = N_k$ è il numero degli elementi del campione che assume il valore w_k, e quindi che

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{np_1 w_1 + \ldots + np_M w_M}{n} = \sum_{k=1}^{M} p_k w_k$$

come affermato nel teorema □

Il Teorema 6.8 mette in evidenza l'analogia che sussiste fra la media m_X di un carattere numerico discreto, e il valore d'attesa μ_X di una *v-a* discreta definito in (4.1): in ambedue i casi si tratta infatti di medie pesate, ma ora le *probabilità teoriche* dei possibili valori della *v-a* sono sostituiti nel loro ruolo di pesi in (6.7) dalle *frequenze relative* empiriche p_k delle modalità w_k

Teorema 6.9. *Assegnato un campione x_j con $j = 1, \ldots, n$, e dati due numeri reali a, b, i dati trasformati secondo la formula $y_j = ax_j + b$ hanno come media*

$$\overline{y} = \overline{ax + b} = a\,\overline{x} + b$$

Dimostrazione: Si ha infatti

$$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{n}\sum_{j=1}^{n}(ax_j + b) = a\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right) + \frac{1}{n}\sum_{j=1}^{n} b = a\overline{x} + b$$

che completa la dimostrazione. □

Le trasformazioni lineari di dati come $y_j = ax_j + b$ sono essenzialmente dei **riscalamenti**: il coefficiente a descrive infatti qualche cambiamento dell'**unità di misura**, mentre b rappresenta uno spostamento dello **zero** dei valori (ricentramento). Il Teorema 6.9 afferma quindi che la media dei dati riscalati coincide con il riscalamento della media dei dati originali

Esempio 6.10. *Supponiamo di sapere che un campione $x_1, \ldots, x_n$ di misure di temperatura in gradi Fahrenheit ha media $\overline{x} = 50\,°F$: come possiamo convertire questa misura in gradi centigradi? In base alle definizioni dovremmo convertire ogni misura x_j in gradi centigradi con la nota relazione*

$$y_j = \frac{100}{180}\,(x_j - 32) = \frac{5}{9}\,x_j - \frac{160}{9} \tag{6.8}$$

e poi calcolare la media $\overline{y}$. Il calcolo sarebbe ripetitivo e laborioso, e d'altra parte potrebbe essere noto solo il valore di $\overline{x}$, e non quello delle singole misure. Possiamo però applicare il Teorema 6.9 visto che la relazione (6.8) è proprio del tipo $y_j = ax_j + b$ con $a = {}^{5}/_{9}$ e $b = {}^{160}/_{9}$. Un semplice calcolo conduce allora al valore

$$\overline{y} = \frac{100}{180}\,(\overline{x} - 32) = \frac{100}{180}\,(50 - 32) = 10\,°C$$

Teorema 6.11. *Dati due campioni $x_1, \ldots, x_\ell$ e $y_1, \ldots, y_m$ con medie $\overline{x}$ e $\overline{y}$, e detto $z_1, \ldots, z_n = x_1, \ldots, x_\ell, y_1, \ldots, y_m$ il campione ottenuto unificando i primi due con $n = \ell + m$, si ha*

$$\overline{z} = \frac{\ell\,\overline{x} + m\,\overline{y}}{n}$$

Dimostrazione: La media $\overline{z}$ si esprime infatti facilmente come

$$\overline{z} = \frac{1}{n}\sum_{j=1}^{n} z_j = \frac{1}{n}\left(\sum_{j=1}^{\ell} x_j + \sum_{j=1}^{m} y_j\right) = \frac{\ell\,\overline{x} + m\,\overline{y}}{n}$$

che è il risultato richiesto. □

Si osservi come il risultato del Teorema 6.11 possa essere riformulato dicendo che la media di campioni combinati è la *media pesata* delle medie dei due campioni separati: in questo caso i pesi rispettivi sono ${}^{\ell}/_{n}$ e ${}^{m}/_{n}$, cioè le numerosità relative dei due campioni iniziali

Esempio 6.12. *Su un libretto universitario sono riportati* $n = 20$ *voti così ripartiti:* $\ell = 4$ *sono dei* 30 *con media* $\overline{x} = 30$*, mentre gli altri* $m = 16$ *sono dei* 18 *con media* $\overline{y} = 18$*. La media (calcolata dalla Definizione 6.7) è* 20.4*, e coincide con la media dei due voti* 30 *e* 18 *pesati rispettivamente con i pesi* ${}^{4}/_{20} = 0.2$ *e* ${}^{16}/_{20} = 0.8$

Definizione 6.13. *Dato un campione di un carattere numerico continuo* X *del quale sono note le frequenze relative* p_k *nelle classi* $\mathcal{B}_k = [a_k, b_k]$ *con* $k = 1, \dots, M$*, si chiama* ***media per dati raggruppati*** *la quantità*

$$\boxed{\widehat{m}_X = \sum_{k=1}^{M} p_k \widehat{w}_k} \tag{6.9}$$

dove $\widehat{w}_k$ *sono i valori centrali (6.5) degli intervalli* $\mathcal{B}_k$

La media per dati raggruppati $\widehat{m}_X$ viene spesso utilizzata come *approssimazione* del valore esatto della media m_X nel caso in cui i valori individuali x_j degli elementi del campione non sono noti (o se ne vuole evitare l'uso) mentre sono conosciuti i valori centrali $\widehat{w}_k$ e le frequenze relative p_k negli intervalli $\mathcal{B}_k$. Si noti l'analogia con il risultato (6.7) per caratteri discreti, dove però al posto dei valori centrali $\widehat{w}_k$ compaiono le *modalità* w_k che in questo modo forniscono il valore esatto m_X della media, mentre dalla (6.9) per caratteri continui si ha solo

$$\widehat{m}_X \simeq m_X$$

cioè un'approssimazione utile nel caso in cui non è nota l'intera tabella dei dati, ma solo la tabella delle frequenze relative in certe determinate classi. L'approssimazione si basa sull'identificazione di tutti i valori x_j che cadono nella classe $\mathcal{B}_k$ con il suo valore centrale $\widehat{w}_k$ pesato con la frequenza relativa nella classe k-ma

Esempio 6.14. *Se nell'Esempio 6.4 fosse conosciuta solo la Tabella 6.4 delle classi e delle frequenze relative (o un'altra analoga, basata su una diversa scelta delle classi), e non l'intera Tabella 6.3 dei dati, il calcolo esatto della media* m_X *di Definizione 6.7 non sarebbe possibile. Dalla Tabella 6.4 si ricava però facilmente la Tabella 6.5 che*

$\widehat{w}_k$	1	3	5	7	9
p_k	0.06	0.39	0.27	0.26	0.02

Tabella 6.5: Tabella dei dei valori centrali e delle frequenze relative del campione di Tabella 6.3 raggruppando i dati nelle 5 classi di ampiezza 2.0

individua anche i valori centrali, e da (6.9) si può quindi ottenere la media per dati raggruppati $\widehat{m}_X = 4.58$. *D'altra parte se utilizzassimo i dati originali della Tabella 6.3 e la Definizione 6.7 si otterrebbe il valore esatto* $m_X = 4.56$ *della media, verificando così che l'approssimazione ottenuta con* $\widehat{m}_X$ *è piuttosto buona anche se le classi scelte sono ampie. È intuitivo, comunque, riconoscere che il valore approssimato è tanto più affidabile quanto più le classi sono ristrette*

Definizione 6.15. *Si dice* ***varianza*** *di un campione* $x_1, \ldots, x_n$ *di un carattere* X *con media* $\overline{x}$ *la quantità*

$$\boxed{s_X^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = \overline{(x - \overline{x})^2}} \tag{6.10}$$

e ***scarto quadratico*** *o* ***deviazione standard*** *la sua radice quadrata* s_X. *Infine si chiama* ***coefficiente di variazione*** *il rapporto*

$$\delta = \frac{s_X}{|\overline{x}|}$$

che fornisce una misura dell'importanza relativa della deviazione standard rispetto alla media

Le quantità introdotte nella precedente definizione sono tutte misure della **dispersione** dei dati attorno alla media $\overline{x}$. In particolare grandi valori della varianza s_X^2 indicano che ci sono delle x_j anche molto lontane da $\overline{x}$, mentre piccoli valori di s_X^2 indicano che il campione è piuttosto concentrato attorno a $\overline{x}$. Il caso limite $s_X^2 = 0$, poi, implica che tutti i valori x_j coincidono con $\overline{x}$

Teorema 6.16. *Dato un campione* $x_1, \ldots, x_n$ *di un carattere numerico discreto* X *con modalità* $w_1, \ldots, w_M$, *e dette* p_k *le frequenze relative di tali modalità, si ha*

$$\boxed{s_X^2 = \sum_{k=1}^{M} p_k(w_k - \overline{x})^2} \tag{6.11}$$

ovvero la varianza di un campione numerico discreto è anche la media pesata degli scarti quadratici delle sue modalità w_k *dalla media* $\overline{x}$, *usando come pesi le frequenze relative* p_k

Dimostrazione: Come per il Teorema 6.8 basterà osservare che per (6.2) $np_k = N_k$ è il numero degli elementi del campione che assume il valore w_k, e quindi che dalla Definizione 6.15 si ha

$$s_X^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = \frac{np_1(w_1 - \overline{x})^2 + \ldots + np_M(w_M - \overline{x})^2}{n} = \sum_{k=1}^{M} p_k(w_k - \overline{x})^2$$

come affermato nel teorema □

Il Teorema 6.16 mette meglio in evidenza l'analogia che sussiste fra la varianza s_X^2 di un carattere numerico, e la varianza σ_X^2 di una *v-a* definita in (4.11): in ambedue i casi si tratta infatti di medie pesate di scarti quadratici, ma qui le *probabilità teoriche* sono sostituiti nel loro ruolo di pesi in (6.11) dalle *frequenze relative* empiriche p_k delle modalità w_k, e l'attesa μ_X dalla media $\overline{x}$

Teorema 6.17. *Dato un campione $x_1, \ldots, x_n$ di un carattere numerico X con media $\overline{x}$, si ha*

$$\boxed{s_X^2 = \overline{x^2} - \overline{x}^2 = \frac{1}{n}\sum_{j=1}^{n} x_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} x_j\right)^2} \tag{6.12}$$

dove $\overline{x^2}$ indica la media dei quadrati del campione, e $\overline{x}^2$ il quadrato della sua media. In particolare, se X fosse un carattere numerico discreto con modalità w_k, la (6.12) si scriverebbe anche come

$$s_X^2 = \sum_{k=1}^{M} p_k w_k^2 - \left(\sum_{k=1}^{M} p_k\, w_k\right)^2 \tag{6.13}$$

Dimostrazione: Infatti da (6.10) si ha

$$\begin{aligned} s_X^2 &= \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j^2 + \overline{x}^2 - 2x_j\overline{x}) \\ &= \frac{1}{n}\sum_{j=1}^{n} x_j^2 + \frac{1}{n}\sum_{j=1}^{n}\overline{x}^2 - 2\overline{x}\frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}\sum_{j=1}^{n} x_j^2 + \overline{x}^2 - 2\overline{x}^2 \\ &= \overline{x^2} - \overline{x}^2 \end{aligned}$$

che dimostra la (6.12). La dimostrazione di (6.13) per caratteri discreti ne consegue poi facilmente tenendo conto di (6.7) □

Il teorema precedente è frequentemente usato per semplificare il calcolo della varianza: una volta calcolata $\overline{x}$, infatti, è in genere più conveniente calcolare la media del campione dei quadrati e usare il Teorema 6.17, piuttosto che calcolare direttamente la varianza dalla definizione.

Teorema 6.18. *Dato un campione $x_1, \ldots, x_n$ di un carattere X con media $\overline{x}$ e varianza s_X^2, e dati due numeri a e b, la varianza del carattere Y del campione trasformato $y_j = ax_j + b$ è*

$$\boxed{s_Y^2 = a^2 s_X^2}$$

Dimostrazione: Infatti si ha dalle definizioni e dal Teorema 6.9 che

$$s_Y^2 = \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{y})^2 = \frac{1}{n}\sum_{j=1}^{n}(ax_j + b - a\overline{x} - b)^2 = \frac{a^2}{n}\sum_{j=1}^{n}(x_j - \overline{x})^2 = a^2 s_X^2$$

come volevasi dimostrare. □

Il precedente risultato è analogo al corrispondente (4.19) per le varianze di *v-a* $Y = aX + b$ che siano trasformazioni lineari di altre *v-a* X, e – in contrasto con il Teortema 6.9 per la media – mette in evidenza il carattere *non lineare* della varianza

Definizione 6.19. *Dato un campione $x_1, \ldots, x_n$ chiameremo* ***errore quadratico medio eqm rispetto al numero a*** *la quantità*

$$\mathcal{E}(a) = \frac{1}{n}\sum_{j=1}^{n}(x_j - a)^2 = \overline{(x-a)^2}$$

È evidente da questa definizione che la varianza è un caso particolare di *eqm* corrispondente alla scelta $a = \overline{x}$, cioè

$$s_X^2 = \mathcal{E}(\overline{x})$$

Ma la media $\overline{x}$ non è solo una tra le tante possibili scelte del valore di a: essa gioca infatti un ruolo particolare come mostrato dal successivo teorema

Teorema 6.20. *La media $\overline{x}$ di un campione $x_1, \ldots, x_n$ è il valore di a per il quale l'eqm $\mathcal{E}(a)$ del campione è minimo*

Dimostrazione: Per determinare il punto di minimo dell'*eqm* bisogna imporre che si annulli la derivata prima $\mathcal{E}'(a)$, cioè si deve trovare il valore di a che soddisfa l'equazione

$$\mathcal{E}'(a) = \frac{d\,\mathcal{E}(a)}{d\,a} = -\frac{2}{n}\sum_{j=1}^{n}(x_j - a) = -2(\overline{x} - a) = 0$$

ed è quindi evidente che il minimo si ottiene per $a = \overline{x}$ □

Definizione 6.21. *Diremo che $x_1, \ldots, x_n$ è un* ***campione standardizzato*** *quando*

$$m_X = \overline{x} = 0 \qquad\qquad s_X^2 = 1$$

Teorema 6.22. *Dato un campione (in generale non standardizzato) $x_1, \dots, x_n$ con media $\overline{x}$ e varianza s_X^2, il campione trasformato nel modo seguente*

$$\boxed{x_j^* = \frac{x_j - \overline{x}}{s_X}}$$

risulta sempre standardizzato

Dimostrazione: Infatti dai Teoremi 6.9 e 6.18 con $a = {}^1\!/_{s_X}$, e $b = {}^{-\overline{x}}\!/_{s_X}$ si ha

$$\overline{x^*} = a\overline{x} + b = \frac{\overline{x}}{s_X} - \frac{\overline{x}}{s_X} = 0 \qquad\qquad s_{X^*}^2 = a^2 s_X^2 = \frac{s_X^2}{s_X^2} = 1$$

per cui $x_1^*, \dots, x_n^*$ risulta standardizzato □

Definizione 6.23. *Dato un campione di un carattere numerico continuo X del quale sono note le frequenze relative p_k nelle classi $\mathcal{B}_k = [a_k, b_k]$ con $k = 1, \dots, M$, si chiama* ***varianza per dati raggruppati*** *la quantità*

$$\boxed{\widehat{s}_X^2 = \sum_{k=1}^{M} p_k \widehat{w}_k^2 - \left(\sum_{k=1}^{M} p_k \, \widehat{w}_k \right)^2} \tag{6.14}$$

dove $\widehat{w}_k$ sono i valori centrali (6.5) degli intervalli $\mathcal{B}_k$

La varianza per dati raggruppati $\widehat{s}_X^2$ viene spesso utilizzata come *approssimazione* del valore esatto della varianza s_X^2 nel caso in cui i valori individuali x_j degli elementi del campione non sono noti (o se ne vuole evitare l'uso) mentre sono conosciuti i valori centrali $\widehat{w}_k$ e le frequenze relative p_k negli intervalli $\mathcal{B}_k$. Si noti l'analogia con il risultato (6.13) per la varianza di caratteri discreti, ricordando però che in quel caso al posto dei valori centrali $\widehat{w}_k$ compaiono le *modalità* w_k che in questo modo forniscono il valore esatto s_X^2, mentre dalla (6.14) per caratteri continui si ha solo

$$\widehat{s}_X^2 \simeq s_X^2$$

cioè un'approssimazione utile nel caso in cui non è nota l'intera tabella dei dati, ma solo la tabella delle frequenze relative in certe determinate classi. Come si vedrà nell'esempio numerico seguente, comunque, l'approssimazione fornita dalla formula (6.14) è meno precisa dell'approssimazione alla media ottenuta con $\widehat{m}_X$ dalla (6.9): questo è dovuto innanzitutto alla presenza di potenze quadratiche e può essere parzialmente corretto passando alle deviazioni standard s_X e $\widehat{s}_X$

Esempio 6.24. *Riprendiamo i dati dell'Esempio 6.4 riportati in Tabella 6.3 – per i quali nell'Esempio 6.14 abbiamo già calcolato la media $m_X = 4.56$ (dopo arrotondamento alla seconda cifra decimale) – e calcoliamone ora la varianza. Un'applicazione diretta della (6.10) di Definizione 6.15 ai dati della Tabella 6.3 fornisce un valore*

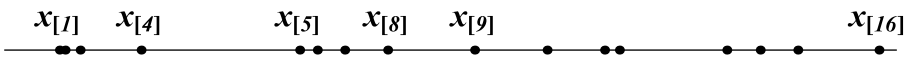


Figura 6.3: Esempio di un campione ordinato di $n = 16$ misure di un carattere numerico continuo. Il primo quartile $q_{1/4}$ si troverà fra $x_{[4]}$ e $x_{[5]}$; la mediana $q_{1/2}$ fra $x_{[8]}$ e $x_{[9]}$

(sempre arrotondato alla seconda cifra decimale) di $s_X^2 = 3.40$*, ma il calcolo è abbastanza laborioso. Per semplificarlo può essere conveniente in questi casi applicare innanzitutto il Teorema 6.17: usando la (6.12) si ottiene infatti lo stesso risultato ricavato dalla Definizione 6.15 con qualche alleviamento della procedura*

I risultati fin qui ottenuti (sebbene arrotondati alla seconda cifra decimale) sono stati tutti calcolati usando le formule esatte (6.10) e (6.12), ma prevedono l'esecuzione ripetitiva di somme con un centinaio di addendi. Usando invece i dati di Tabella 6.5 e le ***formule per dati raggruppati*** *i calcoli sono molto più veloci (le somme si riducono a soli* 5 *addendi), ma inevitabilmente approssimati. Ricordiamo infatti dall'Esempio 6.14 che con la media per dati raggruppati (6.9) avevamo già ottenuto per la media un valore approssimato di* $\widehat{m}_X = 4.58$*: calcolando ora da (6.14) anche la varianza per dati raggruppati si ottiene per la varianza il valore approssimato di* $\widehat{s}_X^2 = 3.70$*. Si noti che mentre lo scarto sull'approssimazione della media è solo di* $\widehat{m}_X - m_X = 0.02$*, per la varianza si ha* $\widehat{s}_X^2 - s_X^2 = 0.30$*; la precisione però migliora se si passa alle deviazioni standard* $\widehat{s}_X - s_X = 1.92 - 1.84 = 0.08$

6.4 Mediana, quartili e quantili

Abbiamo già osservato che è talora utile riordinare in ordine crescente i **campioni non ordinati** $x_1, \ldots, x_n$: per distinguere i due tipi di campioni indicheremo allora i **campioni ordinati** con la nuova notazione $x_{[1]}, \ldots, x_{[n]}$ in modo tale che risulti

$$x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$$

Sulla base dei campioni ordinati introdurremo ora una serie di concetti che riecheggiano quelli definiti – a partire dalle distribuzioni teoriche – nella Sezione 3.5 della parte di Probabilità. In quella discussione, però, per evitare complicazioni, ci eravamo limitati al caso di *v-a* continue: qui invece, usando campioni empirici, talune ambiguità saranno inevitabili e dovremo introdurre delle opportune procedure per dare un senso preciso alle nostre definizioni

Definizione 6.25. *Dato un campione ordinato* $x_{[1]}, \ldots, x_{[n]}$*, chiameremo* ***quantile di ordine*** α *(*$0 < \alpha < 1$*) un numero* q_α *maggiore o uguale di una frazione* α *degli*

elementi del campione, e minore o uguale della restante frazione $1-\alpha$*. Il quantile di ordine* $\alpha = {}^1/_2$ *prende il nome di* ***mediana****; i quantili di ordini* $\alpha = {}^k/_4$ *con* $k = 1, 2, 3$ *si chiamano rispettivamente primo, secondo e terzo* ***quartile*** *(naturalmente il secondo quartile coincide con la mediana). I quantili con ordini* $\alpha = {}^k/_{10}$ *con* $k = 1, \ldots, 9$ *si chiamano* ***decili****, e quelli con ordini* $\alpha = {}^k/_{100}$ *con* $k = 1, \ldots, 99$ *si chiamano* ***percentili***

Se ora prendiamo ad esempio un campione ordinato di $n = 16$ dati, come quello mostrato in Figura 6.3, è facile capire che il primo quartile $q_{{}^1/_4}$ si troverà da qualche parte fra $x_{[4]}$ e $x_{[5]}$ in quanto in questo modo lascerà alla sua sinistra ${}^1/_4$ del campione e alla sua destra il restante ${}^3/_4$. Per la stessa ragione la mediana $q_{{}^1/_2}$ cadrà fra $x_{[8]}$ e $x_{[9]}$. Questa osservazione mette però subito in evidenza le ambiguità della Definizione 6.25: innanzitutto, secondo quest'ultima, *qualunque* numero fra $x_{[4]}$ e $x_{[5]}$ si qualifica come primo quartile, e *qualunque* numero fra $x_{[8]}$ e $x_{[9]}$ è accettabile come mediana. Inoltre quella di Figura 6.3 è una situazione piuttosto particolare: in generale, infatti, il numero $\alpha\, n$ non è un numero intero, e quindi non è più chiaro cosa vuol dire essere "maggiore o uguale di una *frazione* α degli elementi del campione"

Nella pratica statistica sono note molte possibili soluzioni di queste ambiguità della Definizione 6.25, ma noi per brevità ci limteremo a descrivere solo una **procedura** semplificata che permette in ogni caso di ottenere una buona stima dei quantili richiesti: si calcola innanzitutto $\alpha(n+1)$, e poi

- se $\alpha(n+1)$ *è un numero intero*, si considera l'indice $j_\alpha = \alpha(n+1)$ e si prende come quantile di ordine α proprio $q_\alpha = x_{[j_\alpha]}$
- se $\alpha(n+1)$ *non è un numero intero*, si considera l'indice j_α tale che $j_\alpha < \alpha(n+1) < j_\alpha + 1$ e si prende come quantile di ordine α il numero che si trova a metà strada fra $x_{[j_\alpha]}$ e $x_{[j_\alpha+1]}$, cioè

$$q_\alpha = \frac{x_{[j_\alpha]} + x_{[j_\alpha+1]}}{2} \tag{6.15}$$

Esempio 6.26. *Tornando all'esempio della Figura 6.3, per il primo quartile si trova* $\alpha(n+1) = {}^{17}/_4 = 4.25$*, per cui* $q_{{}^1/_4}$ *sarà il numero esattamente a metà strada fra* $x_{[4]}$ *e* $x_{[5]}$*, cioè*

$$q_{{}^1/_4} = \frac{x_{[4]} + x_{[5]}}{2}$$

mentre per la mediana abbiamo $\alpha(n+1) = {}^{17}/_2 = 8.50$*, per cui*

$$q_{{}^1/_2} = \frac{x_{[8]} + x_{[9]}}{2}$$

In questo modo, non solo i risultati sono coerenti con le nostre precedenti osservazioni, ma i quantili assumono anche un preciso valore senza ambiguità di nessun genere. D'altra parte se il campione fosse composto di $n = 17$ *elementi, per la mediana avremmo* $\alpha(n+1) = {}^{18}/_2 = 9$*, e quindi si otterrebbe semplicemente* $q_{{}^1/_2} = x_{[9]}$*,*

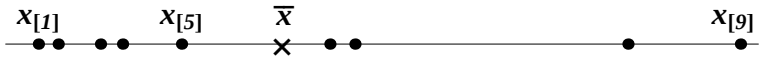


Figura 6.4: Media $\overline{x}$ e mediana $x_{[5]}$ di un campione di $n = 9$ dati: diversamente dalla media il valore della mediana non è influenzato da (limitate) variazioni dei valori di altri dati

valore che divide il campione in due parti uguali di 8 *elementi ciascuna; invece per il primo quartile avremmo* $\alpha(n+1) = {}^{18}/_{4} = 4.50$ *sicché* $q_{1/4}$ *sarebbe ancora una volta la media di* $x_{[4]}$ *e* $x_{[5]}$

Riprendendo infine la Tabella 6.3 dei dati dell'Esempio 6.4 abbiamo ora $n = 100$*: per la mediana abbiamo allora che* $\alpha(n+1) = {}^{101}/_{2} = 50.5$*, e quindi useremo (6.15) con* $j_{1/2} = 50$ *cioè*

$$q_{1/2} = \frac{x_{[50]} + x_{[51]}}{2} = \frac{4.22 + 4.31}{2} = 4.265$$

Analogamente per il primo e il terzo quartile si ha rispettivamente $\alpha(n+1) = {}^{101}/_{4} = 25.25$*, e* $\alpha(n+1) = 3 \times {}^{101}/_{4} = 75.75$*, per cui* $j_{1/4} = 25$*, e* $j_{3/4} = 75$*; pertanto i quartili sono*

$$\begin{aligned} q_{1/4} &= \frac{x_{[25]} + x_{[26]}}{2} = \frac{3.09 + 3.23}{2} = 3.16 \\ q_{3/4} &= \frac{x_{[75]} + x_{[76]}}{2} = \frac{6.24 + 6.28}{2} = 6.26 \end{aligned}$$

La **mediana** è un indice di centralità come la *media* e la *moda*: i valori di questi tre indici sono in generale differenti, e la scelta di quello più opportuno dipende dal particolare problema trattato. Anche le loro proprietà sono differenti: ad esempio per la mediana e per la moda non ci sono risultati semplici come quelli riassunti nei teoremi sulla media richiamati nella Sezione 6.3. Questo ovviamente rende il loro uso meno facile, anche se per altri versi moda e mediana presentano importanti vantaggi. In particolare – come vedremo negli esempi seguenti – la mediana, oltre ad essere a volte più significativa, è anche un indice più *robusto* della media nel senso che il suo valore risulta meno sensibile a variazioni o errori nei dati del campione

Esempio 6.27. *Si consideri il campione di* $n = 9$ *numeri rappresentato graficamente in Figura 6.4: dalla Definizione 6.25, e da una applicazione della successiva procedura di calcolo, si vede subito che in questo caso la mediana coincide con il dato* $x_{[5]}$*. Sull'asse è riportata anche la posizione della media* $\overline{x}$ *del campione, che evidentemente non coincide con la mediana. Supponiamo ora di variare (aumentare o diminuire) il valore di uno o più dati, ad esempio di* $x_{[9]}$*: è chiaro che, finché*

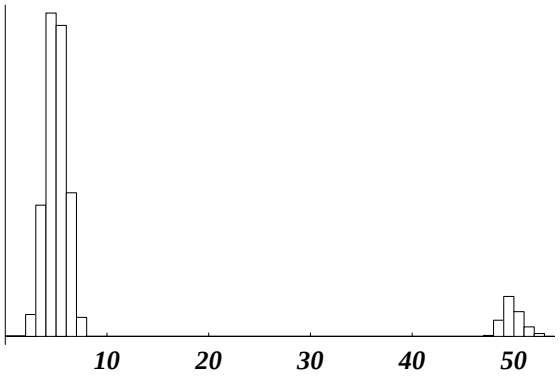


Figura 6.5: Distribuzione del salario dei dipendenti di un'azienda (Esempio 6.28)

$x_{[9]}$ rimane a destra di $x_{[5]}$, la mediana mantiene lo stesso valore $x_{[5]}$ perché questo dipende solo dal numero dagli elementi alla sua destra e alla sua sinistra. Non avviene invece la stessa cosa per la media $\overline{x}$ il cui valore dato dalla Definizione 6.7 è ovviamente sensibile a tutte le variazioni di $x_{[9]}$: in questo senso si usa dire che la mediana è un indice più ***robusto*** *della media*

Esempio 6.28. *La mediana è un indice utile soprattutto nei casi in cui la media rischia di non essere particolarmente* ***rappresentativa****. Supponiamo di considerare un'azienda con* 1 000 *impiegati e operai, e* 100 *dirigenti, e supponiamo che l'istogramma dei redditi di tutti i dipendenti sia come quello di Figura 6.5: i redditi dei* 1 000 *impiegati e operai (in qualche opportuna unità di misura che qui non è importante precisare) sono concentrati attorno a* 5*, mentre quelli dei dirigenti si distribuiscono attorno a* 50*. Per determinare un valore tipico per i salari dei dipendenti potremmo scegliere fra media e mediana, ma dai valori del campione – qui non riportati – si ottiene che la mediana è* 5.13 *mentre la media è* 9.08*: un valore quasi doppio. In questo caso la mediana è l'indicatore più significativo con un valore prossimo a quello di più del* 90% *dei dipendenti. La media invece, che risente molto dell'elevato valore dei salari del piccolo numero dei dirigenti, è meno rappresentativa*

Definizione 6.29. *Chiameremo* ***range*** *di un campione ordinato il numero $\Delta = x_{[n]} - x_{[1]}$, cioè l'ampiezza dell'intervallo $[x_{[1]}, x_{[n]}]$ che contiene tutti i dati; chiameremo* ***differenza interquartile*** *il numero $q_{3/4} - q_{1/4}$, cioè l'ampiezza dell'intervallo $[q_{1/4}, q_{3/4}]$, fra il primo e il terzo quartile, che contiene la metà centrale dei dati*

Il range e la differenza interquartile sono ovviamente degli indici di dispersione. Assieme alla mediana essi possono essere rappresentati su un grafico noto come **boxplot** a causa della sua tipica forma: dati gli estremi $x_{[1]}$, $x_{[n]}$, la mediana $q_{1/2}$ e i quartili $q_{1/4}$, $q_{3/4}$ di un campione ordinato, si disegna un rettangolo, o scatola (*box*), le cui basi inferiore e superiore coincidono con il primo e il terzo quartile, in modo che l'altezza sia pari alla differenza interquartile. All'interno della scatola

$x_{[i]}$	0.72	1.10	1.24	1.98	2.82	2.99	3.01	3.18
	3.31	8.64						
$y_{[j]}$	0.25	0.66	0.68	1.07	1.09	1.15	1.94	3.11
	4.18	4.79	6.18	7.94				
$z_{[k]}$	0.85	1.49	2.19	2.93	4.46	4.61	4.62	5.16
	5.67	6.41	6.46	7.45	7.66	8.65	9.22	

Tabella 6.6: Campioni (ordinati) utilizzati per i boxplot della Figura 6.6.

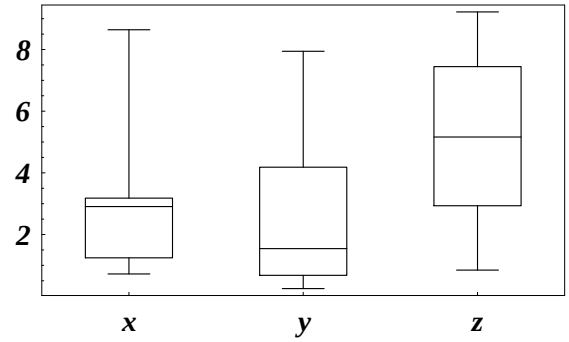


Figura 6.6: Esempi di boxplot costruiti sui tre campioni riportati nella Tabella 6.6

si traccia poi un segmento in corrispondenza della mediana. All'esterno, infine, si riportano due segmenti corrispondenti ai valori $x_{[1]}$, $x_{[n]}$ (la loro distanza è quindi il range) e due segmenti verticali che li congiungono alle basi della scatola. Nella Figura 6.6 sono disegnati i boxplot dei tre campioni di Tabella 6.6, e si può notare il contrasto fra la simmetria del campione $z_{[k]}$ e la asimmetria dei campioni $x_{[i]}$ e $y_{[j]}$: in questi due casi infatti la mediana è più vicina a uno dei due quartili, e inoltre i due dati estremi sono a distanze piuttosto diverse dai rispettivi quartili. Il grafico mostra anche in che senso il range e la distanza interquartile sono due diverse misure di dispersione: ad esempio le $z_{[k]}$, pur avendo approssimativamente lo stesso range delle $x_{[i]}$ hanno una differenza interquartile sensibilmente più elevata

6.5 Momenti, asimmetria e curtosi

Definizione 6.30. *Chiameremo rispettivamente* ***momento di ordine*** k *e* ***momento centrato di ordine*** k *di un campione* $x_1, \ldots, x_n$ *le quantità*

$$m_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k = \overline{x^k} \qquad \widetilde{m}_k = \frac{1}{n}\sum_{j=1}^{n}(x_j - \overline{x})^k = \overline{(x-\overline{x})^k}$$

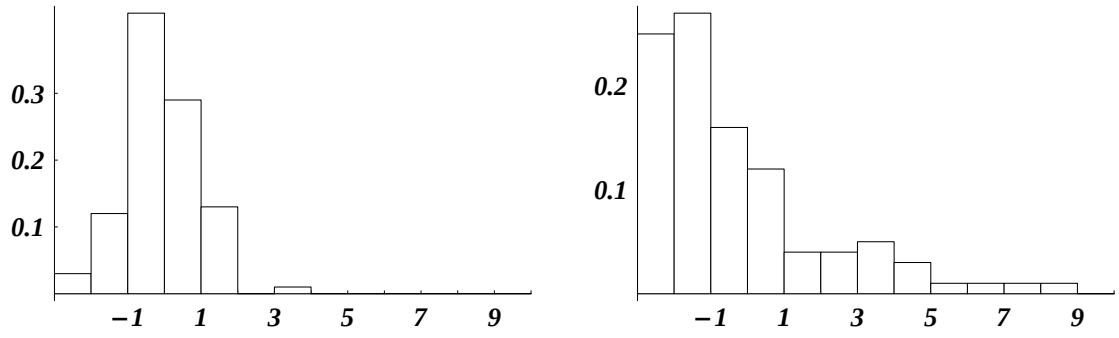


Figura 6.7: Istogrammi di dati con diversa asimmetria: $g_1 = 0.02$ per il primo, e $g_1 = 1.56$ per il secondo.

Ovviamente $m_1 = \overline{x}$, *e* $\widehat{m}_2 = s_X^2$. *Si chiamano inoltre* ***skewness*** *(****asimmetria****) e* ***curtòsi*** *del campione le quantità*

$$g_1 = \frac{\widetilde{m}_3}{\widetilde{m}_2^{\,3/2}} = \frac{\widetilde{m}_3}{s_X^3} \qquad\qquad g_2 = \frac{\widetilde{m}_4}{\widetilde{m}_2^2} = \frac{\widetilde{m}_4}{s_X^4}$$

Si noti che, scambiando i ruoli delle attese di *v-a* con quelli delle medie dei campioni, i momenti della Definizione 6.30 corrispondono esattamente agli analoghi (e omonimi) concetti della Definizione 4.12. Anche in statistica i momenti sono indici che generalizzano medie e varianze e forniscono ulteriori informazioni sulla dispersione, la simmetria e in generale la forma della distribuzione del campione. In particolare l'indice di *asimmetria* g_1 prende valori prossimi a zero se i dati si distribuiscono in maniera simmetrica attorno alla media, mentre prende valori apprezzabilmente diversi da zero se la distribuzione è asimmetrica (vedi Figura 6.7). Il valore di g_1 può essere positivo o negativo: valori positivi indicano la presenza di code verso destra; valori negativi sono invece associati a code verso sinistra. La *curtosi* g_2 invece assume solo valori positivi perché coinvolge solo medie di potenze pari dei dati: essa è legata alla velocità con cui l'istogramma tende a zero allontanandosi dal valore medio. In particolare la curtosi ha valori vicini a zero quando le code dell'istogramma sono corte, cioè quando l'istogramma si annulla rapidamente; viceversa assume valori grandi e positivi quando ci sono code lunghe, cioè quando sono presenti dati anche molto lontani dalla media (vedi Figura 6.8)

6.6 Medie generalizzate

Definizione 6.31. *Dato un campione* $x_1, \ldots, x_n$ *e una funzione* $y = h(x)$ *dotata di inversa* $x = h^{-1}(y)$, *chiameremo* ***media generalizzata*** *la quantità*

$$h^{-1}\left(\frac{h(x_1) + \ldots + h(x_n)}{n}\right) = h^{-1}\big(\,\overline{h(x)}\,\big)$$

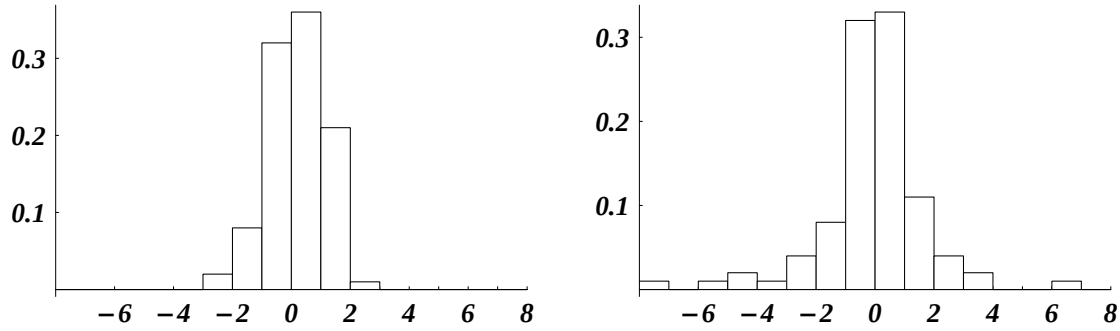


Figura 6.8: Istogrammi di dati con diversa curtosi: $g_2 = 2.59$ per il primo, e $g_2 = 7.76$ per il secondo.

In particolare scegliendo come funzione $y = h(x)$ le seguenti tre

$$\begin{cases} y = h(x) = \log x \\ x = h^{-1}(y) = e^y \end{cases} \qquad \begin{cases} y = h(x) = {}^1/_x \\ x = h^{-1}(y) = {}^1/_y \end{cases} \qquad \begin{cases} y = h(x) = x^2 \\ x = h^{-1}(y) = \sqrt{y} \end{cases}$$

si ottengono rispettivamente la ***media geometrica****, la* ***media armonica*** *e la* ***media quadratica****, cioè le quantità*

$$m_G = \sqrt[n]{x_1 \cdot \ldots \cdot x_n} \qquad m_A = \frac{1}{\frac{1}{n}\left(\frac{1}{x_1} + \ldots + \frac{1}{x_n}\right)} \qquad m_Q = \sqrt{\frac{x_1^2 + \ldots + x_n^2}{n}}$$

L'introduzione delle medie generalizzate è motivata dal fatto che in generale, per ragioni derivanti dal particolare problema discusso, può essere più significativo eseguire la media non direttamente sui dati x_j, ma sui dati trasformati con una qualche funzione $h(x_j)$: il risultato viene poi ri-trasformato all'indietro mediante la $h^{-1}(y)$. Le medie geometrica, armonica e quadratica sono i casi più noti ti tali medie generalizzate e il loro significato sarà chiarito con la discussione di alcuni esempi

Esempio 6.32. *Supponiamo che una certa quantià di denaro C sia stata investita a tassi di interesse che vengono aggiornati ogni mese, e supponiamo di indicare con $r_1, \ldots, r_n$ tali tassi di interesse nel suddetto periodo di n mesi. Questo significa che dopo il primo mese il capitale C diventa $(1 + r_1)C$ e, supponendo che questo venga interamente re-investito al tasso r_2, dopo il secondo mese troveremo $(1+r_2)(1+r_1)C$, e così via fino al tempo n. Cosa possiamo allora intendere come* ***rendimento medio*** *r del nostro investimento su n mesi? Invece di eseguire una semplice media dei coefficienti di aggiornamento $1 + r_k$ si ragiona nel modo seguente seguente: r è il tasso di interesse costante che applicato per n mesi produce lo stesso aumento di capitale prodotto dalla applicazione successiva dei tassi $r_1, \ldots, r_n$. In altre parole r*

deve soddisfare l'equazione

$$(1+r)^n C = (1+r_1)\cdot\ldots\cdot(1+r_n)C$$

e quindi in definitiva si ha

$$1+r = \sqrt[n]{(1+r_1)\cdot\ldots\cdot(1+r_n)}$$

cioè il coefficiente di aggiornamento medio $1+r$ *è la* ***media geometrica*** *dei coefficienti di aggiornamento* $1+r_k$*: da questa poi si ricava immediatamente anche il rendimento medio* r

Esempio 6.33. *Supponiamo che una ditta produttrice di automobili svolga la sua attività in* n *stabilimenti ciascuno dei quali ha un suo tempo di produzione* T_k*, nel senso che essi producono una automobile rispettivamente nei tempi* $T_1,\ldots,T_n$*: quale valore dovremmo considerare come* ***tempo medio*** T *di produzione di tutta la ditta? In questo caso adotteremo il seguente criterio:* T *è il tempo di produzione con il quale la ditta, nell'unità di tempo, produrrebbe complessivamente un numero di auto uguale a quello prodotto con i tempi* $T_1,\ldots,T_n$*. Siccome ogni stabilimento produce* $1/T_k$ *automobili nell'unità di tempo, il nostro criterio impone che*

$$\frac{n}{T} = \frac{1}{T_1} + \ldots + \frac{1}{T_n}$$

ovvero

$$T = \frac{1}{\frac{1}{n}\left(\frac{1}{T_1}+\ldots+\frac{1}{T_n}\right)}$$

Il tempo medio T *è quindi la* ***media armonica*** *dei tempi* T_k *con* $k=1,\ldots,n$

Esempio 6.34. *I batteri di una determinata specie si organizzano in colonie di forma circolare, e il numero di batteri è proporzionale alla superficie delle colonie. Si osservano* n *colonie con diametri* $d_1,\ldots,d_n$*: cosa possiamo considerare come* ***diametro medio*** d *delle colonie? Anche in questo caso ci facciamo guidare da un criterio: richiederemo che* n *colonie tutte con lo stesso diametro medio* d *abbiano la stessa superficie totale (e quindi lo stesso numero di batteri) delle* n *colonie con diametri differenti* $d_1,\ldots,d_n$*. In tal caso dovremo imporre che*

$$\frac{n\pi}{4}\,d^2 = \frac{\pi}{4}\,(d_1^2+\ldots+d_n^2)$$

e quindi avremo

$$d = \sqrt{\frac{d_1^2+\ldots+d_n^2}{n}}.$$

Il diametro medio d *quindi è la* ***media quadratica*** *dei diametri* d_j *con* $j=1,\ldots,n$

Capitolo 7

Statistica descrittiva multivariata

7.1 Statistica bivariata

Sugli individui di una popolazione possono essere eseguite osservazioni e misure di *due o più caratteri* con lo scopo di metter anche in evidenza gli eventuali *legami statistici* fra di essi. Ad esempio possiamo misurare altezza e peso dei cittadini di una determinata comunità per mettere in evidenza una relazione fra le due misure; ovvero potremmo cercare una correlazione fra il reddito pro capite e la longevità dei cittadini dei diversi paesi del mondo, e così via. In tutti questi casi gli elementi del nostro campione non saranno più dei semplici numeri, ma vettori con due o più componenti. In questa prima sezione però ci limiteremo a studiare il caso di *due soli caratteri* (X, Y), sicché i nostri campioni saranno del tipo $(x_1, y_1), \ldots, (x_n, y_n)$, cioè saranno composti di n vettori di due componenti ciascuno

Se i due caratteri sono *qualitativi o numerici discreti*, con un numero finito di modalità $A_1, \ldots, A_r$ del carattere X e $B_1, \ldots, B_s$ del carattere Y, una prima maniera di rappresentare il campione sarà quella di costruire una **tabella di contingenza** di *frequenze assolute congiunte* come quella riportata in Tabella 7.1. In essa si riportano innanzitutto le *frequenze congiunte* $N_{j,k}$, cioè il numero delle volte in cui si presenta la coppia di modalità (A_j, B_k); sui margini della tabella si riportano poi le *frequenze marginali* $N_{j,\cdot}$ e $N_{\cdot,k}$, cioè il numero di volte in cui si presentano separatamente le modalità A_j e B_k; nell'angolo destro in basso si riporta infine la numerosità totale

	B_1	$\ldots$	B_s	
A_1	$N_{1,1}$	$\ldots$	$N_{1,s}$	$N_{1,\cdot}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
A_r	$N_{r,1}$	$\ldots$	$N_{r,s}$	$N_{r,\cdot}$
	$N_{\cdot,1}$	$\ldots$	$N_{\cdot,s}$	n

Tabella 7.1: Tabella di contingenza per due caratteri X e Y rispettivamente con modalità A_j e B_k.

	GIU	ECO	LET	SCI	MED	FAR	ALTRO	
Proprietario	80	36	134	99	65	28	69	511
Contadino	6	2	15	6	4	1	5	39
Imprenditore	168	74	312	137	208	53	83	1 035
Professionista	470	191	806	400	876	164	124	3 031
Dirigente	236	99	493	264	281	56	123	1 552
Impiegato	145	52	281	133	135	30	74	850
Operaio	166	64	401	193	127	23	157	1 131
Altro	321	121	651	258	309	49	142	1 851
	1 592	639	3 093	1 490	2 005	404	777	10 000

Tabella 7.2: Tabella di contingenza per la scelta della Facoltà universitaria di $n = 10\,000$ studenti, secondo l'attività lavorativa del padre (dati relativi all'a.a. 1975/76; *INSEE*, Paris 1978)

n del campione. Si noti che, per un dato j la marginale $N_{j,\cdot}$ è la somma delle $N_{j,k}$ della sua riga, mentre per un dato k la marginale $N_{\cdot,k}$ è la somma delle $N_{j,k}$ della sua colonna; infine anche la numerosità totale n è la somma delle marginali (sia sulla riga che sulla colonna). In maniera del tutto analoga si costruisce anche la tabella di contingenza delle *frequenze relative congiunte e marginali*, cioè delle

$$p_{j,k} = \frac{N_{j,k}}{n} \qquad p_{j,\cdot} = \frac{N_{j,\cdot}}{n} \qquad p_{\cdot,k} = \frac{N_{\cdot,k}}{n}$$

In questo caso però, a causa della normalizzazione (6.4) delle frequenze relative, nell'angolo destro in basso comparirà 1 invece di n

Esempio 7.1. *Nella Tabella 7.2 sono riportati in forma di* ***tabella di contingenza*** *i dati relativi alla scelta della facoltà universitaria di $n = 10\,000$ studenti incrociandoli con quelli relativi all'attività lavorativa del padre come indicatore della loro estrazione sociale. I due caratteri sono quindi qualitativi, e le loro modalità sono le facoltà e le estrazioni sociali. I dati nella parte centrale della tabella rappresentano i numeri di studenti di una determinata estrazione sociale che hanno scelto di iscriversi a una particolare facoltà. Le frequenze marginali mettono in evidenza sia la la composizione sociale complessiva degli studenti universitari (marginali verticali), che il gradimento riscosso dalle diverse facoltà universitarie (marginali orizzontali). Infine un'analisi più attenta (anche usando strumenti che svilupperemo nel seguito) può mettere in evidenza la relazione che intercorre fra l'estrazione socio–professionale della famiglia degli studenti e la scelta della facoltà universitaria*

Una tabella di contingenza può comunque essere redatta anche per *modalità numeriche continue*, ma in questo caso – come per gli istogrammi – bisognerà raggruppare i dati in classi con una opportuna (e arbitraria) suddivisione in intervalli. Nel caso di modalità numeriche e continue, però, è possibile e molto utile rappresentare graficamente i dati in un piano Cartesiano x, y come punti con coordinate

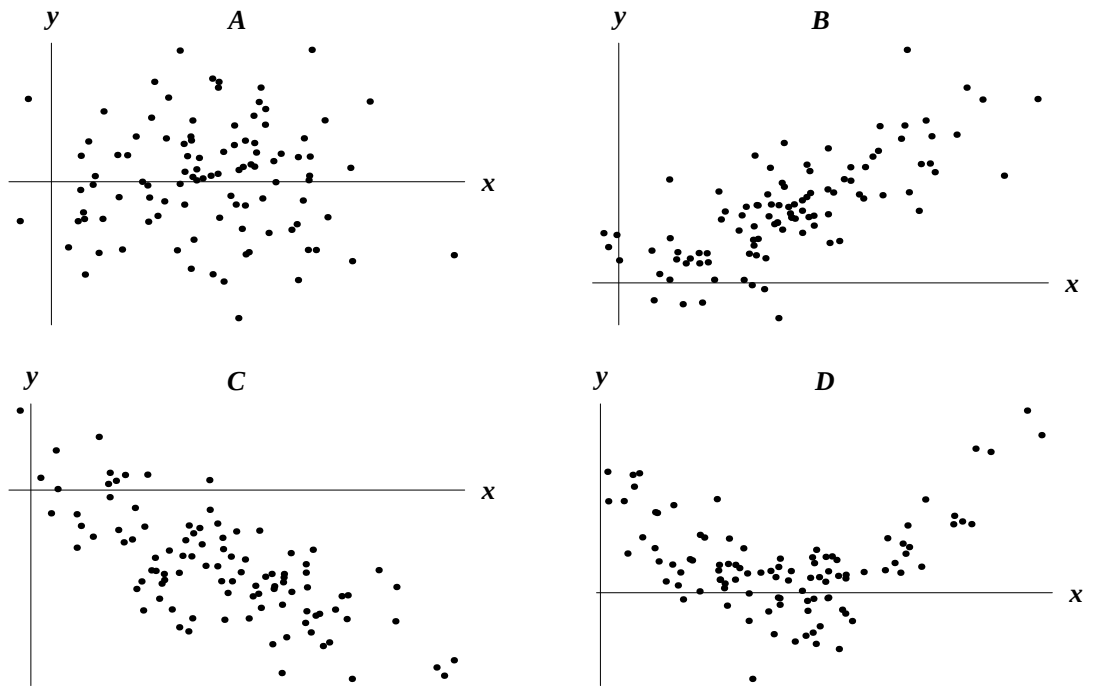


Figura 7.1: Vari esempi di conformazioni degli *scatter plot* di dati bidimensionali, numerici9 e continui

$(x_1, y_1), \ldots, (x_n, y_n)$: un grafico che porta anche il nome di ***scatter plot***. La conformazione della nuvola di punti dello *scatter plot* fornisce infatti una prima, suggestiva indicazione sulla eventuale relazione intercorrente fra i due caratteri X e Y. Nella Figura 7.1 sono riportati alcuni esempi di *scatter plot* di campioni con $n = 100$ punti, e con conformazioni differenti: innanzitutto nel caso A i punti sono disposti in modo da non suggerire nessun tipo di dipendenza tra i due caratteri X e Y, nel senso che la distribuzione delle Y è poco sensibile a variazioni del valore delle X. Invece in B si nota che i valori di Y tendono ad crescere (decrescere) quando anche i valori di X crescono (decrescono); anzi la conformazione della nuvola indica una approssimativa dipendenza funzionale lineare del tipo $Y = aX + b$ con $a > 0$. Anche nel caso C i dati mostrano una analoga dipendenza approssimativamente lineare, ma questa volta con $a < 0$: infatti ora i valori di Y tendono a crescere (decrescere) quando i valori di X decrescono (crescono). Infine il caso D suggerisce una dipendenza non lineare, approssimativamente parabolica, tra i due caratteri dato che i valori di Y crescono quando i valori di X si allontanano – nei due versi – dal centro della nuvola

7.2 Covarianza, correlazione e regressione

Definizione 7.2. *Dato un campione bivariato* $(x_1, y_1), \ldots, (x_n, y_n)$ *di due caratteri numerici* X *e* Y*, si chiama* ***covarianza*** *di* X *e* Y *la quantità*

$$\boxed{s_{XY} = \overline{(x - \overline{x})(y - \overline{y})} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}$$

dove $\overline{x}$ *e* $\overline{y}$ *sono le medie delle* x_i *e delle* y_i*. Si chiama invace* ***coefficiente di correlazione*** *la quantità*

$$\boxed{r_{XY} = \frac{s_{XY}}{s_X s_Y}}$$

dove s_X *e* s_Y *sono le deviazioni standard delle* x_i *e delle* y_i

Si tratta, come è evidente, di concetti analoghi a quelli della Definizione 4.4 per *v-a*, e come in quel caso anche qui troviamo che la varianza s_X^2 è un caso particolare di covarianze con $X = Y$, cioè

$$s_{XX} = s_X^2 \qquad\qquad r_{XX} = 1 \tag{7.1}$$

La covarianza e il coefficiente di correlazione sono indicatori numerici importanti nell'analisi della relazione che intercorre fra due caratteri X e Y. In particolare, come vedremo, essi entrano nella valutazione quantitativa della dipendenza lineare di un carattere dall'altro, cioè nella determinazione dei coefficienti a e b di una retta $Y = aX + b$ che descriva (almeno approssimativamente) l'andamento dei dati

Definizione 7.3. *Diremo che due caratteri* X *e* Y *sono* ***non correlati*** *se* $s_{XY} = 0$ *(e quindi anche* $r_{XY} = 0$*); diremo che hanno* ***correlazione positiva (negativa)*** *se* $s_{XY} > 0$ *(*$s_{XY} < 0$*)*

Teorema 7.4. *Dato un campione* $(x_1, y_1), \ldots, (x_n, y_n)$ *con medie* $\overline{x}$ *e* $\overline{y}$ *si ha*

$$\boxed{s_{XY} = \overline{xy} - \overline{x}\,\overline{y} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)\left(\frac{1}{n}\sum_{j=1}^{n} y_j\right)} \tag{7.2}$$

dove $\overline{xy}$ *indica la media dei prodotti* $x_i y_i$*, e* $\overline{x}\,\overline{y}$ *il prodotto delle due medie separate.*

Dimostrazione: Questa proprietà, che generalizza la (6.12) per la varianza, si dimostra come l'analoga proprietà (4.15) per *v-a* sostituendo le medie ai valori d'attesa: si ha infatti dalla Definizione 7.2

$$\begin{aligned} s_{XY} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i\,y_i - \overline{x}\,y_i - x_i\,\overline{y} + \overline{x}\,\overline{y}) \\ &= \overline{xy} - \frac{\overline{x}}{n}\sum_{i=1}^{n} y_i - \frac{\overline{y}}{n}\sum_{i=1}^{n} x_i + \overline{x}\,\overline{y} = \overline{xy} - 2\overline{x}\,\overline{y} + \overline{x}\,\overline{y} = \overline{xy} - \overline{x}\,\overline{y} \end{aligned}$$

come enunciato in (7.2) □

Teorema 7.5. *Dato un campione* $(x_1, y_1), \ldots, (x_n, y_n)$ *risulta sempre*

$$-1 \leq r_{XY} \leq +1$$

e in particolare avremo $|r_{XY}| = 1$ *se e solo se esistono due numeri* a *e* b *tali che* $y_i = ax_i + b$ *per* $i = 1, \ldots, n$, *con* $a < 0$ *se* $r_{XY} = -1$, *e* $a > 0$ *se* $r_{XY} = +1$. *Inoltre, se* $A_x, B_x, A_y.B_y$ *sono numeri arbitrari e se poniamo* $x'_i = A_x x_i + B_x$, $y'_i = A_y y_i + B_y$ *(cambiamento di scala e di unità di misura), il coefficiente di correlazione resta invariato, cioè si avrà*

$$r_{X'Y'} = r_{XY}$$

Infine, passando ai campioni standardizzati del Teorema 6.22

$$x_i^* = \frac{x_i - \overline{x}}{s_X} \qquad\qquad y_i^* = \frac{y_i - \overline{y}}{s_Y}$$

si trova che

$$s_{X^*Y^*} = r_{XY} \tag{7.3}$$

cioè la covarianza dei campioni standardizzati coincide con il coefficiente di correlazione dei campioni originari

Dimostrazione: Omessa. Osserveremo solo che questi risultati sono analoghi a quelli del Teorema 4.7 per il coefficiente di correlazione ρ_{XY} di due *v-a* X e Y □

Torniamo ora al problema dell'analisi degli *scatter plot* di dati bidimensionali come quelli di Figura 7.1, e domandiamoci se non sia possibile trovare una relazione analitica che descriva – almeno approssimativamente – la dipendenza delle y_i dalle x_i. L'ipotesi più semplice è che ci sia una relazione di tipo lineare $Y = aX + b$, ma un semplice sguardo ai grafici di Figura 7.1 ci convince del fatto che in generale sarà impossibile trovare due numeri a e b tali che $y_i = ax_i + b$ per tutte le $i = 1, \ldots, n$; cioè che è impossibile trovare una retta che passi per tutti i punti della nuvola, a meno che non ci si trovi proprio nel caso $|r_{XY}| = 1$ discusso nel Teorema 7.5. Potremo invece provare a determinare a e b in modo che la retta $y = ax + b$ *approssimi nel modo migliore* l'andamento dello *scatter plot*. Il senso in cui parliamo di approssimazione ottimale è precisato nelle definizioni e nei risultati seguenti

Definizione 7.6. *Dato un campione bivariato* $(x_1, y_1), \ldots, (x_n, y_n)$ *chiameremo* ***errore quadratico medio (eqm) rispetto alla retta*** $\boldsymbol{y = ax + b}$ *la quantità*

$$\mathcal{E}(a, b) = \overline{[y - (ax + b)]^2} = \frac{1}{n} \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 .$$

Siccome in generale i punti di uno *scatter plot* non saranno allineati lungo una retta, è evidente che – comunque siano scelti i numeri a e b – per ogni elemento (x_i, y_i) del campione risulterà $y_i \neq ax_i + b$. In tal caso $y_i - (ax_i + b)$ sarà lo scarto fra le due quantità e l'*eqm* $\mathcal{E}(a, b)$ sarà una misura complessiva dell'errore che si

commette approssimando le y_i mediante le $ax_i + b$. È d'altra parte ovvio che, per un fissato campione, l'*eqm* $\mathcal{E}(a, b)$ dipenderà da a e b e acquista quindi senso la seguente definizione

Definizione 7.7. *Dato un campione* $(x_1, y_1), \ldots, (x_n, y_n)$*, chiameremo* ***retta di regressione*** *la retta* $y = ax + b$ *i cui coefficienti* a *e* b *rendono minimo l'eqm* $\mathcal{E}(a, b)$

Teorema 7.8. *Dato il campione* $(x_1, y_1), \ldots, (x_n, y_n)$*, i coefficienti* a *e* b *della retta di regressione sono*

$$\boxed{a = \frac{s_{XY}}{s_X^2} = \frac{s_Y}{s_X}\, r_{XY} \qquad\qquad b = \overline{y} - a\overline{x} = \overline{y} - \frac{s_{XY}}{s_X^2}\,\overline{x} = \overline{y} - \frac{s_Y}{s_X}\, r_{XY}\,\overline{x}}$$

Dimostrazione: Per determinare le a e b che rendono minimo l'*eqm* $\mathcal{E}(a, b)$ calcoliamone le due derivate parziali

$$\frac{\partial \mathcal{E}}{\partial a} = -\frac{2}{n}\sum_{i=1}^{n} x_i[y_i - (ax_i + b)] \qquad\qquad \frac{\partial \mathcal{E}}{\partial b} = -\frac{2}{n}\sum_{i=1}^{n}[y_i - (ax_i + b)]$$

e imponiamo che si annullino ottenendo così il seguente sistema di equazioni lineari nelle incognite a e b

$$\begin{aligned} \sum_{i=1}^{n} x_i[y_i - (ax_i + b)] &= 0 \\ \sum_{i=1}^{n}[y_i - (ax_i + b)] &= 0 \end{aligned}$$

Ora la seconda equazione del sistema si scrive anche come

$$\sum_{i=1}^{n}(y_i - ax_i) - nb = 0$$

da cui si ricava subito

$$b = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i) = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{a}{n}\sum_{i=1}^{n} x_i = \overline{y} - a\overline{x} \tag{7.4}$$

Per la prima equazione osserviamo invece che, da (7.4) e dai Teoremi 6.17 e 7.4, il suo primo membro diviso per n diviene

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} x_i[y_i - (ax_i + b)] &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{a}{n}\sum_{i=1}^{n} x_i^2 - (\overline{y} - a\overline{x})\frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \overline{xy} - a\,\overline{x^2} - (\overline{y} - a\overline{x})\overline{x} = \overline{xy} - \overline{x}\,\overline{y} - a\big(\overline{x^2} - \overline{x}^2\big) \\ &= s_{XY} - a s_X^2 \end{aligned}$$

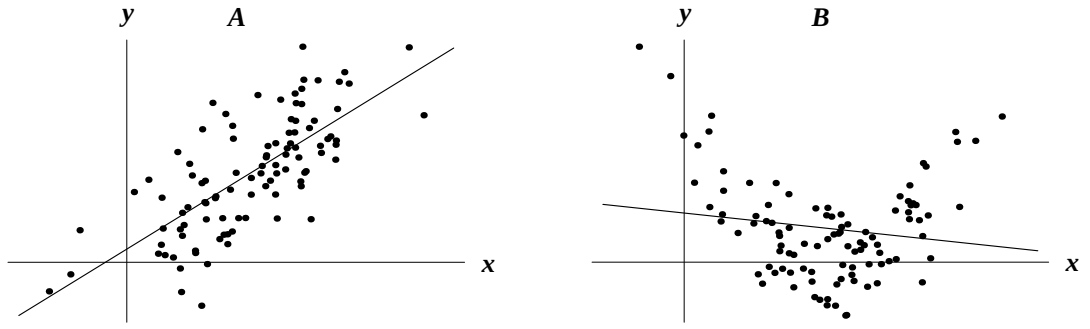


Figura 7.2: Esempi di rette di regressione per dati bivariati. Nel caso A il coefficiente di correlazione è $r_{XY} = 0.72$, mentre nel caso B è $r_{XY} = -0.14$

e quindi la prima equazione si riduce a

$$n\left(s_{XY} - as_X^2\right) = 0$$

da cui si ricava infine

$$a = \frac{s_{XY}}{s_X^2} \tag{7.5}$$

Le soluzioni (7.5) e (7.4) del nostro sistema di equazioni rendono minimo l'*eqm* e sono quindi, come stabilito dalla Definizione 7.7, i coefficienti della retta di regressione. Le espressioni in termini del coefficiente di correlazione r_{XY} si ricavano poi direttamente dalla Definizione 7.2 □

In base al Teorema 7.8 se i caratteri X e Y fossero non correlati (cioè se $s_{XY} = 0$, e $r_{XY} = 0$) allora avremmo $a = 0$, e quindi la retta di regressione sarebbe orizzontale: questo indicherebbe che i valori di Y non mostrano nessuna dipendenza dai valori di X. Inoltre il coefficiente angolare a della retta di regressione ha lo stesso segno di s_{XY}, e quindi avrà anche un andamento crescente (decrescente) se vi è correlazione positiva (negativa). Il coefficiente di correlazione r_{XY}, peraltro, si presenta come la migliore misura della *linearità* della relazione fra X e Y. Infatti, in base al Teorema 7.5, mentre la covarianza s_{XY} può assumere ogni valore positivo o negativo, r_{XY} cade sempre in $[-1, 1]$ e si può quindi stabilire facilmente se esso indica una dipendenza forte o debole. Se in particolare $r_{XY} = \pm 1$ allora i dati sono funzionalmente legati dalla relazione lineare $y_i = ax_i + b$, e la retta di regressione passa attraverso tutti i punti dello *scatter plot*

Esempio 7.9. *Due esempi di* ***rette di regressione*** *sono riportati nella Figura 7.2. Nella parte A la retta, i cui coefficienti a e b sono calcolati a partire dal Teorema 7.8, offre una descrizione approssimata ma abbastanza significativa della relazione che intercorre fra i dati del campione: c'è infatti una evidente tendenza delle y_i a crescere (linearmente) quando le x_i crescono, anche se non si può supporre una dipendenza*

strettamente funzionale fra gli elementi del campione. Nel caso in questione, peraltro, il coefficiente di correlazione $r_{XY} = 0.72$ ha un valore abbastanza elevato (vicino a 1) e positivo da suggerire una effettiva correlazione positiva fra i caratteri X e Y rappresentata da una retta con pendenza positiva. Un coefficiente di correlazione r_{XY}, e i parametri a e b di una retta di regressione posson sempre essere calcolati a partire da un dato campione bivariato $(x_1, y_1), \ldots, (x_n, y_n)$. Bisogna però evitare di pensare che cercare una qualche relazione lineare fra X e Y sia in ogni caso ragionevole. Nella parte B della Figura 7.2, infatti, si può vedere lo scatter plot di un campione in cui la relazione fra X e Y è presumibilmente non lineare (piuttosto sembra parabolica): anche in questo caso la retta di regressione può essere determinata, ma un'approssimazione lineare è ora evidentemente poco significativa

7.3 Statistica multivariata

Quando ad ogni individuo di una popolazione sono associati $p \geq 2$ caratteri numerici $X_1, \ldots, X_p$, gli n elementi del campione diventano vettori con p componenti $\boldsymbol{x}_j = (x_{j1}, \ldots, x_{jp}) \in \boldsymbol{R}^p$, $j = 1, \ldots, n$, e i dati si presentano come una matrice $p \times n$

$$\|x_{jk}\| = \begin{pmatrix} x_{11} & \ldots & x_{1p} \\ x_{21} & \ldots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \ldots & x_{np} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{pmatrix}$$

nella quale la riga j-ma è il vettore $\boldsymbol{x}_j$, mentre la colonna k-ma è l'insieme dei valori assunti dal carattere X_k. Si consiglia di consultare l'Appendice D.1 per qualche richiamo di *calcolo vettoriale*. Teoricamente, come nel caso $p = 2$ studiato nella Sezione 7.1, i vettori $\boldsymbol{x}_j$ possono essere rappresentati come uno *scatter plot* di n punti nello spazio p-dimensionale $\boldsymbol{R}^p$, ma una simile rappresentazione è irrealizzabile in pratica per $p > 3$, sicché saremo obbligati a sviluppare altri strumenti di analisi

Definizione 7.10. *Chiameremo* ***baricentro*** *dei dati il vettore $\overline{\boldsymbol{x}} = (\overline{x}_{\cdot\,1}, \ldots, \overline{x}_{\cdot\,p}) \in \boldsymbol{R}^p$ le cui componenti sono le* ***medie*** *dei valori di ciascun carattere, ossia le medie lungo le colonne di $\|x_{jk}\|$*

$$\overline{x}_{\cdot\,k} = \frac{1}{n} \sum_{j=1}^{n} x_{jk} \qquad k = 1, \ldots, p$$

Chiameremo poi ***matrice di covarianza*** *$p \times p$ la matrice $\mathbb{S} = \|s_{k\ell}\|$ i cui elementi sono le covarianze dei caratteri X_k e X_ℓ, ossia delle colonne k–ma e ℓ–ma di $\|x_{jk}\|$*

$$s_{k\ell} = \frac{1}{n} \sum_{j=1}^{n} (x_{jk} - \overline{x}_{\cdot\,k})(x_{j\ell} - \overline{x}_{\cdot\,\ell}) = \frac{1}{n} \sum_{j=1}^{n} x_{jk} x_{j\ell} - \overline{x}_{\cdot\,k} \overline{x}_{\cdot\,\ell} \qquad k, \ell = 1, \ldots, p$$

Analogamente si chiama ***matrice di correlazione*** $p \times p$ *la matrice* $\mathbb{R} = \|r_{k\ell}\|$ *i cui elementi sono i coefficienti di correlazione dei caratteri* X_k *e* X_ℓ*. Si noti in particolare che in base a (7.1) gli elementi diagonali* s_{kk} *di* $\mathbb{S}$ *sono le* ***varianze*** $s^2_{\cdot k}$ *di ciascun carattere calcolate lungo le colonne di* $\|x_{jk}\|$*, cioè*

$$s_{kk} = s^2_{\cdot k} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \overline{x}_{\cdot k})^2 = \frac{1}{n}\sum_{j=1}^{n} x^2_{jk} - \overline{x}^2_{\cdot k} \qquad k = 1, \ldots, p$$

Per gli elementi diagonali della matrice di correlazione R, invece, si ha sempre da (7.1)

$$r_{kk} = 1 \qquad k = 1, \ldots, p$$

Si chiama infine ***dispersione totale*** *dei dati la quantità*

$$\Delta = \frac{1}{n}\sum_{j=1}^{n} |\boldsymbol{x}_j - \overline{\boldsymbol{x}}|^2$$

Il baricentro e le matrici di covarianza $\mathbb{S}$ e di correlazione $\mathbb{R}$ sono strumenti fondamentali nello studio di dati p-dimensionali. Val la pena notare a questo punto che le matrici $\mathbb{S}$ e $\mathbb{R}$ sono matrici simmetriche nel senso che

$$s_{k\ell} = s_{\ell\, k} \qquad r_{k\ell} = r_{\ell\, k}$$

infatti è ovvio che la covarianza di X_k e X_ℓ coincide con la covarianza di X_ℓ e X_k. Inoltre gli elementi diagonali della matrice di correlazione sono tutti uguali a 1 dato che si tratta delle correlazioni di ciascun carattere con se stesso

7.4 Componenti principali

Torniamo ora al problema di rappresentare graficamente dei dati p-dimensionali $\|x_{jk}\|$: siccome non possiamo disegnare grafici in $\boldsymbol{R}^p$ con $p \geq 3$, dovremo ricorrere alle **proiezioni** (si veda l'Appendice D.1 per le definizioni essenziali) dei dati su rette o su piani bi-dimensionali passanti per l'origine di $\boldsymbol{R}^p$. È intuitivo però che in questo modo si perde dell'informazione: ad esempio, punti che in una proiezione cadono vicini possono anche essere proiezioni di punti che nello spazio p-dimensionale sono molto lontani. Inoltre bisogna ricordare che la scelta della retta (o del piano) di proiezione è in generale del tutto arbitraria. Dovremo quindi definire dei criteri di scelta, ed è abbastanza naturale ritenere che questi criteri debbano essere orientati a trovare in $\boldsymbol{R}^p$ le direzioni lungo le quali la proiezione risulta più fedele possibile

Iniziamo con il discutere il caso della **proiezione su una retta** individuata da un versore $\boldsymbol{v}$, ricordando che (vedi Appendice D.1) la proiezione di un vettore $\boldsymbol{x} \in \boldsymbol{R}^p$ sulla retta definita da $\boldsymbol{v}$ è il vettore nella direzione di $\boldsymbol{v}$ che ha come modulo il prodotto scalare $\boldsymbol{x} \cdot \boldsymbol{v} = |x| \cos\vartheta$. dove ϑ è l'angolo fra $\boldsymbol{x}$ e $\boldsymbol{v}$ definito in (D.2).

Pertanto la proiezione riduce il campione di n vettori $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ ad un campione *univariato* di n numeri $y_1, \ldots, y_n$ con

$$y_j = \boldsymbol{x}_j \cdot \boldsymbol{v} = \sum_{k=1}^{p} x_{jk} v_k \qquad\qquad j = 1, \ldots, n$$

che possiamo considerare come i valori di un nuovo carattere $Y = v_1 X_1 + \ldots + v_p X_p$ ottenuto come combinazione lineare dei caratteri originali $X_1, \ldots, X_p$ usando come coefficienti le componenti del versore $\boldsymbol{v} = (v_1, \ldots, v_p)$. Si noti che per noi il campione $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ è un *dato del problema*, mentre i valori di $y_1, \ldots, y_n$ dipendono dalla scelta del versore $\boldsymbol{v} = (v_1, \ldots, v_p)$ le cui componenti giocano invece il ruolo di *incognite*. Naturalmente la rappresentazione proiettata fornita dal campione numerico delle y_j conterrà meno informazione rispetto al campione iniziale composto da n vettori $\boldsymbol{x}_j$: il nostro compito sarà allora quello di determinare $\boldsymbol{v}$ in modo che la rappresentazione proiettata delle y_j sia la più fedele possibile rispetto all'informazione totale contenuta nel campione delle $\boldsymbol{x}_j$. Per determinare la migliore direzione di proiezione $\boldsymbol{v}$ adotteremo allora il seguente **Criterio di rappresentazione ottimale**:

> *La **proiezione ottimale** di un campione $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ si ottiene scegliendo come direzione di proiezione $\boldsymbol{v}$ quella che produce il campione univariato $y_1, \ldots, y_n$ con **dispersione (varianza) massima***

Infatti, siccome in una proiezione il principale rischio è quello di sovrapporre punti che nella realtà sono lontani fra loro, richiedere che la varianza dei punti proiettati sia la più grande possibile significa richiedere che questi punti siano il più possibile lontani e distinti

Per mettere in pratica questo principio ci serviremo allora di alcuni risultati richiamati anche nell'Appendice D.1: indicheremo innanzitutto con

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

gli autovalori (eventualmente anche coincidenti) della matrice di covarianza $\mathbb{S}$ dei dati, e con $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ i corrispondenti autovettori ortonormali, nel senso che supporremo verificate le seguenti equazioni agli autovalori

$$\mathbb{S}\boldsymbol{v}_k = \lambda_k \boldsymbol{v}_k \qquad\qquad k = 1, \ldots, p$$

Teorema 7.11. *Dato un campione $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ in $\boldsymbol{R}^p$, e assegnata un'arbitraria direzione di proiezione $\boldsymbol{v}$, media e varianza del campione proiettato $y_1, \ldots, y_n$ sono*

$$m_Y(\boldsymbol{v}) = \overline{\boldsymbol{x}} \cdot \boldsymbol{v} = \sum_{k=1}^{p} \overline{x}_{\cdot\, k} v_k \qquad s_Y^2(\boldsymbol{v}) = \boldsymbol{v} \cdot \mathbb{S}\boldsymbol{v} = \sum_{k,\ell=1}^{p} v_k s_{k\ell} v_\ell \tag{7.6}$$

dove $\overline{\boldsymbol{x}}$ è il baricentro e $\mathbb{S}$ è la matrice di covarianza del campione $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. In particolare, se la direzione di proiezione coincide con un autovettore $\boldsymbol{v}_k$, allora si ha

$$s_Y^2(\boldsymbol{v_k}) = \lambda_k \geq 0 \tag{7.7}$$

e inoltre per qualunque altro versore $\boldsymbol{v}$ *si avrà*

$$\lambda_1 = s_Y^2(\boldsymbol{v_1}) \geq s_Y^2(\boldsymbol{v}) \geq s_Y^2(\boldsymbol{v_p}) = \lambda_p \tag{7.8}$$

Pertanto il versore $\boldsymbol{v}$ *che rende massima la varianza* $s_Y^2(\boldsymbol{v})$ *è l'autovettore* $\boldsymbol{v}_1$ *associato all'autovalore più grande* λ_1*; limitandosi poi ai versori* $\boldsymbol{v}$ *ortogonali a* $\boldsymbol{v}_1$*, il versore per il quale* $s_Y^2(\boldsymbol{v})$ *è massima è* $\boldsymbol{v}_2$*, e così via per tutti gli altri autovettori. Infine per la dispersione totale* Δ *del campione* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *si ha*

$$\Delta = \sum_{k=1}^{p} \lambda_k \tag{7.9}$$

Dimostrazione: Tenendo conto dell'equazione (D.1) si ha innanzitutto

$$m_Y(\boldsymbol{v}) = \frac{1}{n}\sum_{j=1}^{n} y_j = \frac{1}{n}\sum_{j=1}^{n} (\boldsymbol{x}_j \cdot \boldsymbol{v}) = \left(\frac{1}{n}\sum_{j=1}^{n} \boldsymbol{x}_j\right) \cdot \boldsymbol{v} = \overline{\boldsymbol{x}} \cdot \boldsymbol{v}$$

che prova la la prima delle (7.6); trascureremo invece di dimostrare la seconda delle (7.6). Dando poi per dimostrate ambedue queste relazioni, dalle proprietà degli autovettori e del prodotto scalare (vedi Appendice D.1) si ha

$$s_Y^2(\boldsymbol{v}_k) = \boldsymbol{v}_k \cdot \mathbb{S}\boldsymbol{v}_k = \boldsymbol{v}_k \cdot (\lambda_k \boldsymbol{v}_k) = \lambda_k\, \boldsymbol{v}_k \cdot \boldsymbol{v}_k = \lambda_k |\boldsymbol{v}_k|^2 = \lambda_k$$

il che prova anche la (7.7). Inoltre per definizione $s_Y^2(\boldsymbol{v}_k)$ è non negativa, sicché anche gli autovalori della matrice di covarianza sono sempre non negativi. La (7.8), che non dimostriamo, ci dice che la varianza massima si ottiene scegliendo come direzione di proiezione l'autovettore $\boldsymbol{v}_1$ associato all'autovalore più grande λ_1. Infine la (7.9) (anch'essa non dimostrata) spiega come la dispersione totale Δ si decompone nella somma delle dispersioni associate ad ogni autovalore λ_k □

È evidente ora che il Teorema 7.11 ci permette di applicare *Criterio di rappresentazione ottimale* in maniera precisa, ma prima conviene aggiungere un'osservazione importante. Un campione multivariato $\|x_{jk}\|$ è in generale composto di **dati disomogenei**: in particolare questi potrebbero differire per i loro *ordini di grandezza*. Supponiamo ad esempio di voler compilare una statistica relativa alle condizioni meteorologiche di una località registrando pressione atmosferica (in *mmHg*), temperatura (in ^oC), velocità del vento (in *Km/h*) e copertura nuvolosa (in *ottavi* di cielo coperto). Con le unità di misura tradizionali (indicate fra parentesi) le misure di pressione saranno numeri dell'ordine di 10^3, ma la copertura nuvolosa sarà un numero intero da 1 a 8, la temperatura un numero dell'ordine delle diecine, e infine la velocità del vento potrà variare da 0 fino a numeri dell'ordine di 10^2. In queste condizioni le quantità rappresentate dai numeri più grandi assumerebbero ingiustificatamente un peso sproporzionato rispetto alle altre. Siccome però le unità di

misura sono arbitrarie, potremo opportunamente modificarle per bilanciare l'importanza relativa delle quantità osservate. A questo scopo di solito si usa *standardizzare* i dati originali $\|x_{jk}\|$, cioè li si sostituisce con la matrice delle

$$x^*_{jk} = \frac{x_{jk} - \overline{x}_{\cdot k}}{s_{\cdot k}}$$

dove abbiamo utilizzato notazioni della Definizine 7.10. I dati $\|x^*_{jk}\|$, avendo ora tutti media 0 e varianza 1, sono stati ridotti ad una scala in cui sono tutti numeri di grandezza comparabile

Applicando allora il Teorema 7.11 ai nuovi dati standardizzati dobbiamo ricordare da (7.3) che la matrice di *covarianza* delle $\|x^*_{jk}\|$ non è nient'altro che la matrice di *correlazione* delle $\|x_{jk}\|$, sicché in conclusione, per evitare problemi di disomogeneità dei dati, *è sempre consigliabile applicare il Teorema 7.11 usando la matrice di correlazione* $\mathbb{R}$ *invece che quella di covarianza* $\mathbb{S}$. In questo caso si ottengono gli stessi risultati del Teorema 7.11, con la differenza che ora bisognerà calcolare autovalori e autovettori della matrice di correlazione $\mathbb{R}$ invece che quelli della matrice di covarianza $\mathbb{S}$. Si può dimostrare, infine, che la somma degli autovalori della matrice di correlazione è sempre uguale al numero p dei caratteri $X_1, \ldots, X_p$, e quindi anche che, in base al Teorema 7.11, la dispersione totale dei dati standardizzati è sempre uguale a p

In conclusione, dovendo rappresentare un campione multivariato, adotteremo la seguente **Procedura**:

> *Dato un campione* $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ *in* $\boldsymbol{R}^p$*, per ottenere la* ***rappresentazione più fedele lungo un asse*** *bisogna*
>
> 1. *calcolare la matrice di correlazione* $\mathbb{R}$ *dei dati*
> 2. *determinare gli autovalori* $\lambda_1 \geq \ldots \geq \lambda_p$ *di* $\mathbb{R}$*, e i relativi autovettori ortonormali* $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p$
> 3. *proiettare i dati lungo l'autovettore* $\boldsymbol{v}_1$ *associato all'autovalore più grande* λ_1

Naturalmente, per conservare una maggiore quantità di informazione e ottenere una rappresentazione ancora più fedele, si può scegliere di usare delle **proiezioni su piani definiti da due direzioni** invece che lungo una sola direzione. In tal caso sempre il Teorema 7.11 ci dice che la migliore scelta per la seconda direzione è l'autovettore $\boldsymbol{v}_2$ relativo al *secondo più grande* autovalore λ_2. Pertanto la proiezione più fedele ai dati è quella eseguita nel piano definito da $\boldsymbol{v}_1$ e $\boldsymbol{v}_2$, che produce i campioni *bivariati* $(y_{j1}, y_{j\,2}) = (\boldsymbol{x}_j \cdot \boldsymbol{v}_1\,,\, \boldsymbol{x}_j \cdot \boldsymbol{v}_2)$ per $j = 1, \ldots, n$. Continuando in questo modo, ulteriore informazione si ottiene esaminando le proiezioni lungo gli altri autovettori, sempre privilegiando quelli con gli autovalori più grandi. La seconda parte del Teorema 7.11 infine ci suggerisce che ogni autovalore λ_k contribuisce alla dispersione totale Δ in proporzione al suo valore: osservazione coerente con il fatto

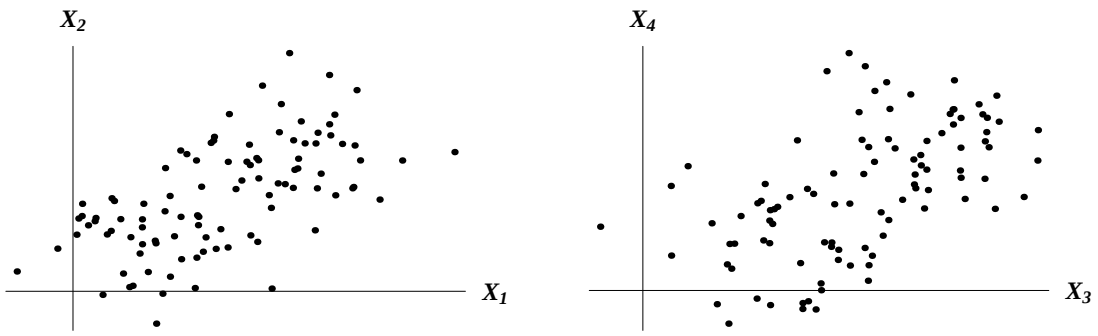


Figura 7.3: Rappresentazione di coppie di componenti dai dati della Tabella 7.3.

che le direzioni privilegiate per la proiezione sono proprio quelle degli autovettori relative agli autovalori più grandi. Tutte queste osservazioni possono infine essere riassunte nella definizione seguente

Definizione 7.12. *Chiameremo* ***direzioni o componenti principali*** *quelle degli autovettori ortonormali* $\boldsymbol{v}_k$ *della matrice di correlazione* $\mathbb{R}$*, e – ordinatamente rispetto agli autovalori* $\lambda_1 \geq \ldots \geq \lambda_p$ *– diremo* ***prima direzione principale*** *quella di* $\boldsymbol{v}_1$*,* ***seconda direzione principale*** *quella di* $\boldsymbol{v}_2$*, e così via. I piani individuati dalle coppie di autovettori* $(\boldsymbol{v}_k, \boldsymbol{v}_\ell)$ *si chiamano poi* ***piani principali****, e in particolare il piano* $(\boldsymbol{v}_1, \boldsymbol{v}_2)$ *sarà il* ***primo piano principale****. Chiameremo infine* ***fedeltà*** *della proiezione dei dati sul piano principale* $(\boldsymbol{v}_k, \boldsymbol{v}_\ell)$ *il rapporto*

$$\frac{\lambda_k + \lambda_\ell}{\lambda_1 + \ldots + \lambda_p}$$

In particolare, siccome λ_1 *e* λ_2 *sono gli autovalori più grandi, la massima fedeltà si ottiene proiettando i dati sul primo piano principale*

Esempio 7.13. *Nella Tabella 7.3 sono riportate* $n = 100$ *misure di quattro caratteri numerici continui (*$p = 4$*) ottenute con una simulazione: esse potrebbero rappresentare le misure di quattro dimensioni fisiche di 100 animali di una data specie (altezza, lunghezza, . . .), o rilevazioni di quattro parametri economici relativi a 100 paesi (popolazione, reddito pro capite, . . .), o altro ancora. Ovviamente è impossibile rappresentare graficamente i punti corrispondenti perché questi si trovano in uno spazio a 4 dimensioni; si potrebbe però pensare di rappresentarne due componenti per volta su un opportuno piano bidimensionale: un possibile scopo di questa analisi potrebbe essere, per esempio, quello di studiare se i dati mostrano la tendenza a raggrupparsi in due o più classi (*clusters*), indicando in questo modo una* classificazione *dei nostri 100 soggetti in base alle misure effettuate. Così gli animali della specie considerata potrebbero essere classificati in due o più razze sulla base delle quattro dimensioni*

X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4
3.061	2.417	3.924	3.361	2.558	3.502	2.548	3.616
3.189	3.696	1.514	4.073	2.839	1.095	2.667	3.061
3.433	3.560	2.820	5.040	3.408	3.244	2.129	3.762
3.249	2.806	2.528	1.544	2.070	2.269	4.173	2.251
3.400	3.198	3.236	4.241	3.058	2.531	3.351	3.729
2.147	1.087	1.659	3.518	3.026	3.096	2.107	3.238
1.838	1.384	1.977	2.199	3.437	3.896	2.235	3.295
2.891	3.343	4.174	4.100	2.818	2.941	3.660	3.680
3.603	3.306	2.906	3.035	3.695	3.188	3.286	3.088
3.725	1.099	3.179	2.964	3.836	3.378	2.965	3.595
2.687	2.823	2.134	2.476	0.992	3.124	1.138	4.959
2.404	3.475	2.457	3.559	3.927	3.153	1.099	1.753
3.159	2.699	2.680	2.523	4.113	1.713	2.669	2.624
2.182	2.359	3.184	3.992	2.774	2.714	3.324	2.532
4.071	3.024	2.443	3.937	2.965	2.352	2.154	1.980
3.351	4.206	2.377	2.232	1.875	4.419	3.043	3.156
0.935	3.531	3.954	1.215	2.876	2.437	2.661	3.543
3.579	3.852	2.307	3.235	3.314	3.848	2.957	2.125
2.086	3.428	3.129	4.731	2.390	3.892	2.768	3.288
0.765	3.760	3.036	2.454	2.859	2.689	2.538	2.518
3.853	1.755	2.898	2.604	3.166	3.625	2.679	2.307
4.767	3.575	1.736	2.690	2.925	3.647	3.179	3.342
3.138	2.528	2.438	4.704	1.927	4.173	3.250	2.178
1.429	2.864	3.256	2.436	3.529	4.558	2.532	3.071
3.558	3.411	3.341	1.656	2.363	3.697	2.946	2.422
5.739	4.882	4.442	5.697	3.909	5.353	5.358	4.472
4.722	3.856	5.223	5.300	6.166	6.079	4.190	5.167
5.366	5.293	6.676	3.362	4.701	5.506	4.473	4.999
4.223	5.348	5.197	6.689	3.683	5.229	3.216	5.201
4.669	5.667	7.106	5.797	4.689	4.948	5.699	5.261
5.119	6.221	3.844	5.445	4.655	4.616	4.471	5.130
4.894	5.768	5.779	5.298	4.268	5.178	6.439	4.327
4.775	5.016	3.917	5.770	4.215	7.500	4.981	4.983
5.643	3.663	5.926	5.561	4.666	4.568	5.605	3.760
4.128	3.485	4.394	4.232	4.493	5.253	3.842	6.306
5.640	4.501	5.438	4.808	4.793	5.769	5.136	5.434
3.546	6.051	5.467	6.610	5.937	4.383	5.171	6.327
6.504	5.075	6.572	5.937	4.753	6.663	3.348	5.095
4.532	4.019	5.422	3.788	4.905	5.107	4.997	5.624
4.884	5.052	5.072	4.963	5.467	4.798	4.651	4.980
4.666	5.672	5.527	5.346	4.629	4.459	5.378	4.685
4.630	3.929	4.952	4.814	3.480	4.244	4.542	4.206
5.785	5.280	5.260	3.721	3.469	7.792	5.108	3.423
4.171	5.004	5.074	4.813	5.926	5.510	4.978	5.144
5.020	4.721	6.992	4.161	4.541	3.735	4.427	4.340
3.856	5.492	5.111	4.547	3.891	4.352	3.805	4.663
5.521	4.918	4.869	3.736	5.418	4.546	4.485	5.366
5.743	4.291	3.891	5.352	5.327	4.709	4.195	5.736
4.317	4.597	5.968	4.831	6.966	5.292	4.989	5.437
4.133	5.867	5.258	5.699	4.891	4.513	5.264	5.354

Tabella 7.3: Campione di $n = 100$ misure di quattro caratteri continui.

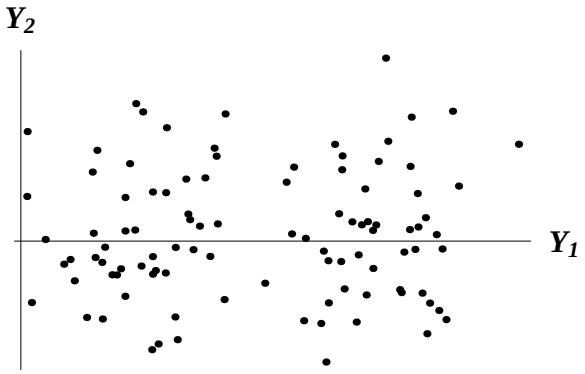


Figura 7.4: Rappresentazione dei dati della Tabella 7.3 nel primo piano principale.

fisiche considerate; oppure i 100 paesi potrebbero essere classificati in diversi livelli di sviluppo economico secondo i valori dei quattro indicatori rilevati, e così via

Per ottenere una rappresentazione bidimensionale si potrebbe allora iniziare con lo scegliere (in maniera arbitraria) due delle quattro componenti riportando le corrispondenti coordinate su un piano cartesiano. Nella Figura 7.3 sono riprodotti a titolo di esempio i punti che si otterrebbero considerando prima solo i caratteri X_1, X_2, e poi gli altri due X_3, X_4 dalla Tabella 7.3. Queste immagini – pur mettendo in evidenza una certa correlazione fra i vari caratteri – non sembrano mostrare una tendenza dei punti a raggrupparsi in classi con caratteristiche diverse. Altri grafici si potrebbero ottenere scegliendo altre coppie di coordinate, ed altri ancora considerando proiezioni su piani non coincidenti con gli originari piani coordinati, e naturalmente, modificando questi piani, si potrebbero mettere in evidenza aspetti del campione che altrimenti resterebbero nascosti. Ma è evidente che la ricerca della migliore rappresentazione non può essere eseguita per tentativi, soprattutto se – come accade – il campione fosse più complesso di quello usato in questo esempio. D'altra parte, siccome il nostro problema è quello di separare delle classi, il miglior criterio di scelta sembra proprio quello di cercare la proiezione sul piano che rende massima la dispersione dei punti proiettati, e quindi appare opportuno ricorrere all'analisi in componenti principali esposta nella presente sezione

Seguendo la procedura suggerita, si comincia quindi (usando qualche opportuno sistema di calcolo automatico) con la determinazione della matrice di correlazione $\mathbb{R}$ dei dati della Tabella 7.3:

$$\mathbb{R} = \|r_{k\ell}\| = \begin{pmatrix} 1.000 & 0.606 & 0.719 & 0.620 \\ 0.606 & 1.000 & 0.599 & 0.600 \\ 0.719 & 0.599 & 1.000 & 0.560 \\ 0.620 & 0.600 & 0.560 & 1.000 \end{pmatrix}$$

e si prosegue con il calcolo degli autovalori (ordinati)

$$\lambda_1 = 2.854\,, \qquad \lambda_2 = 0.471\,, \qquad \lambda_3 = 0.403\,, \qquad \lambda_4 = 0.273\,,$$

e dei relativi autovettori (ortonormali)

$$\boldsymbol{v}_1 = \begin{pmatrix} -0.519 \\ -0.490 \\ -0.506 \\ -0.485 \end{pmatrix} \qquad \boldsymbol{v}_2 = \begin{pmatrix} -0.369 \\ 0.372 \\ -0.580 \\ 0.624 \end{pmatrix} \qquad \boldsymbol{v}_3 = \begin{pmatrix} 0.224 \\ -0.784 \\ -0.024 \\ 0.578 \end{pmatrix} \qquad \boldsymbol{v}_4 = \begin{pmatrix} -0.738 \\ -0.079 \\ 0.638 \\ 0.204 \end{pmatrix}$$

I valori dei λ_k mostrano innanzitutto che già la prima componente principale $\boldsymbol{v}_1$ ha una fedeltà di 0.713*, che per il primo piano principale aumenta fino a* 0.831*; la restante quantità di informazione, pari a* 0.169*, si trova nelle altre due componenti $\boldsymbol{v}_3$ e $\boldsymbol{v}_4$. Nel primo piano principale le coordinate (y_{j1}, y_{j2}) con $j = 1, \ldots, n$ dei nostri n punti saranno allora ottenute mediante le combinazioni lineari seguenti*

$$\begin{aligned} y_{j1} = \boldsymbol{x}_j \cdot \boldsymbol{v}_1 &= x_{j1}v_{11} + x_{j2}v_{12} + x_{j3}v_{13} + x_{j4}v_{14} \\ &= -0.519\, x_{j1} - 0.490\, x_{j2} - 0.506\, x_{j3} - 0.485\, x_{j4} \\ y_{j2} = \boldsymbol{x}_j \cdot \boldsymbol{v}_2 &= x_{j1}v_{21} + x_{j2}v_{22} + x_{j3}v_{23} + x_{j4}v_{24} \\ &= -0.369\, x_{j1} + 0.372\, x_{j2} - 0.580\, x_{j3} + 0.624\, x_{j4} \end{aligned}$$

La rappresentazione grafica delle (y_{j1}, y_{j2}) è riportata nella Figura 7.4 e mostra – più chiaramente di quelle di Figura 7.3 – che è possibile separare i punti in due gruppi abbastanza distinti, e che in particolare è soprattutto la coordinata y_{j1} che suggerisce tale classificazione

I valori e i segni delle componenti degli autovettori $\boldsymbol{v}_k$ hanno poi anche un loro significato: esse indicano quanto e in che verso i caratteri originari X_k contribuiscono alla combinazione che definisce i nuovi caratteri Y_k. Supponiamo ad esempio che le X_k siano misure di dimensioni fisiche di animali: il fatto che le componenti di $\boldsymbol{v}_1$ abbiano valori abbastanza vicini e tutti dello stesso segno indica che Y_1 è un carattere che distingue gli n individui in base al valore di tutte le dimensioni fisiche considerate. In pratica Y_1 è una misura complessiva della grandezza *dell'animale, e distingue gli individui in animali grandi e piccoli. Negli altri autovettori, invece, le componenti hanno segni differenti: questo indica che gli altri tre caratteri mettono in contrasto i valori delle diverse dimensioni misurate e sono quindi indicatori della* forma *dell'animale. In pratica essi distingueranno ad esempio gli individui in alti e corti, bassi e lunghi e così via. L'importanza che i diversi caratteri Y_k assumono nella classificazione è infine stabilita dal valore relativo degli autovalori λ_k*

Capitolo 8

Statistica inferenziale: Stima

8.1 Stima puntuale

La *LGN* e il *TLC* sono gli strumenti principali con i quali si affrontano i problemi di stima ai quali abbiamo già accennato nel Capitolo 5. Ricorderemo innanzitutto che in generale il nostro obiettivo sarà quello di approssimare in maniera affidabile il valore di qualche indicatore numerico sconosciuto della distribuzione di una data quantità aleatoria, utilizzando a questo scopo i risultati di un numero n abbastanza grande di misure ripetute e indipendenti della quantità stessa. Si tratta quindi di un tipico argomento di *statistica inferenziale* visto che si vogliono ricavare informazioni generali su un fenomeno a partire da un numero grande, ma limitato di misure. Nel seguito la quantità che vogliamo studiare sarà descritta da una *v-a* X, e ci porremo innanzitutto il problema di stimare alcuni indicatori della sua legge come ad esempio la *probabilità* p di un evento, l'*attesa* μ, o la *varianza* σ^2 o qualunque altro momento. In altri casi supporremo che X sia dotata di una distribuzione di forma nota $\mathfrak{L}$ (binomiale, normale, Poisson e così via), ma con alcuni *parametri* sconosciuti che indicheremo genericamente con il simbolo θ. In particolare θ potrebbe coincidere l'attesa o la varianza di X, ma si danno anche situazioni più generali; ad esempio se la legge di X è di Poisson $\mathfrak{P}(\lambda)$ il parametro tipico è $\theta = \lambda$; se invece la legge è normale $\mathfrak{N}(\mu, \sigma^2)$ in genere θ è identificato con μ oppure σ^2, o anche con la coppia $\theta = (\mu, \sigma^2)$. In questi casi ci porremo l'obiettivo di stimare il valore di θ, ovvero, più in generale, di una qualche sua funzione $h(\theta)$: così nel caso di una *v-a* di Poisson potremmo essere interessati a stimare $h(\lambda) = 1/\lambda$, invece che semplicemente λ; o nel caso normale potremmo voler stimare $h(\mu, \sigma) = \mu^2$ e così via. La stima viene effettuata utilizzando un *campione* di n misure indipendenti di X: queste saranno rappresentate mediante altrettante *v-a* $X_1, \ldots, X_n$ tutte indipendenti e tutte con la medesima distribuzione di X. Che queste n misure siano ora esse stesse delle *v-a* è legato al fatto che esse possano essere *ripetute*: infatti in una replica delle n osservazioni ci aspettiamo di avere risultati differenti da quelli iniziali, e quindi siamo obbligati a trattare le misure come *v-a*. Introdurremo ora alcune definizioni un po' più formali per riordinare i concetti fin qui esposti

Definizione 8.1. *Data una v-a X con distribuzione $\mathfrak{L}$, diremo che n v-a $X_1, \ldots, X_n$ costituiscono un* ***campione (aleatorio)*** *di X se esse sono indipendenti e tutte distribuite con la stessa legge $\mathfrak{L}$; si chiama poi* ***statistica*** *ogni v-a funzione del campione*

$$U_n = u(X_1, \ldots, X_n)$$

Se la legge $\mathfrak{L}(\theta)$ contiene un parametro sconosciuto θ, e noi vogliamo stimare una sua funzione $h(\theta)$, chiameremo ***stimatore*** *ogni statistica U_n e diremo che*

- *U_n è uno* ***stimatore corretto, o non distorto di*** *$h(\theta)$ se*

  $$\boldsymbol{E}_\theta[U_n] = h(\theta)$$

  dove il simbolo $\boldsymbol{E}_\theta$ indica che l'attesa è calcolata con la legge $\mathfrak{L}(\theta)$ supponendo che θ abbia lo stesso valore che compare in $h(\theta)$ al secondo membro

- *U_n è uno* ***stimatore consistente di*** *$h(\theta)$ se per $n \to \infty$*

  $$U_n \xrightarrow{mq} h(\theta)$$

  nel senso della Definizione 5.1 di convergenza

Quando, come in questa definizione, una stima viene effettuata con il valore di un solo stimatore U_n si parla di ***stima puntuale***

In linea di principio, dunque, uno *stimatore* è una qualsiasi *v-a* funzione del campione dato, che viene usata per stimare un $h(\theta)$. Ovviamente, però, uno stimatore U_n è buono solo se i suoi valori sono vicini al valore di $h(\theta)$. Questa richiesta spiega l'introduzione dei concetti di stimatore *non distorto* e di stimatore *consistente*: essi infatti, per definizione, garantiscono che U_n prenda (sia in media, che al limite per $n \to \infty$) valori prossimi ad $h(\theta)$. In questo modo, inoltre, la scelta dei possibili stimatori accettabili si restringe molto anche se per il momento resta poco chiaro come si deve procedere per determinare lo stimatore di una $h(\theta)$. Vedremo nei risultati e negli esempi successivi che delle indicazioni più precise in merito vengono innanzitutto dalla *LGN*, cioè dal Teorema 5.3. Troveremo però successivamente anche altri metodi più generali (principio della *Massima verosimiglianza*) per determinare la forma migliore di uno stimatore

Nel seguito utilizzeremo alcune notazioni già introdotte in precedenza: dato un campione aleatorio $X_1, \ldots, X_n$ di n *v-a* indipendenti, tutte con la stessa legge $\mathfrak{L}$ con attesa μ e varianza σ^2, ricordiamo che $\overline{X}_n$ e $\widehat{S}_n^2$ rappresenteranno rispettivamente la **media aritmetica** (5.9) e la **varianza campionaria** (5.10) del campione. Utilizzeremo inoltre la **varianza corretta** (5.11) che qui richiamiamo per comodità

$$\boxed{S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X}_n)^2 = \frac{n}{n-1}\,\widehat{S}_n^2 = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^{\,2}\right)} \tag{8.1}$$

Sarà inoltre bene osservare che l'attesa μ e la varianza σ^2 sono tipiche quantità da stimare; esse sono ovviamente funzioni dei parametri θ della legge $\mathfrak{L}$, ma noi non renderemo esplicita questa dipendenza per non appesantire la notazione. Per la stessa ragione eviteremo – tranne casi particolari – di usare l'indice sottoscritto θ come nella Definizione 8.1: scriveremo cioè per lo più $\boldsymbol{E}$ invece di $\boldsymbol{E}_\theta$

Teorema 8.2. *Dato il campione aleatorio $X_1, \ldots, X_n$ di v-a tutte con la stessa attesa μ e la sessa varianza σ^2*

- *$\overline{X}_n$ e S_n^2 sono stimatori corretti e consistenti rispettivamente di μ e σ^2*
- *$\widehat{S}_n^2$ è uno stimatore consistente, ma non corretto, di σ^2*

Dimostrazione: La consistenza di $\overline{X}_n$ e $\widehat{S}_n^2$ discende direttamente dalla *LGN*, Teorema 5.3; quella di S_n^2 deriva poi da quella di $\widehat{S}_n^2$ tenendo conto della relazione (8.1). Per dimostrare poi che $\overline{X}_n$ è uno stimatore corretto di μ basterà osservare che

$$\boldsymbol{E}\left[\overline{X}_n\right] = \boldsymbol{E}\left[\frac{X_1 + \ldots + X_n}{n}\right] = \frac{\boldsymbol{E}\left[X_1\right] + \ldots + \boldsymbol{E}\left[X_n\right]}{n} = \frac{n\,\mu}{n} = \mu \tag{8.2}$$

Viceversa $\widehat{S}_n^2$ non è uno stimatore corretto di σ^2: infatti, tenendo conto della proprietàda (4.16) delle varianze, si ha innanzitutto

$$\boldsymbol{E}\left[X_k^2\right] = \boldsymbol{V}\left[X_k\right] + \boldsymbol{E}\left[X_k\right]^2 = \sigma^2 + \mu^2$$

e poi da (8.2), da (4.19) e dall'indipendenza delle X_k

$$\begin{aligned} \boldsymbol{E}\big[\overline{X}_n^{\,2}\big] &= \boldsymbol{V}\left[\overline{X}_n\right] + \boldsymbol{E}\left[\overline{X}_n\right]^2 = \boldsymbol{V}\left[\frac{1}{n}\sum_{k=1}^{n} X_k\right] + \mu^2 \\ &= \frac{1}{n^2}\sum_{k=1}^{n} \boldsymbol{V}\left[X_k\right] + \mu^2 = \frac{n\,\sigma^2}{n^2} + \mu^2 = \frac{\sigma^2}{n} + \mu^2 \end{aligned}$$

Ne segue allora da (5.10) che

$$\begin{aligned} \boldsymbol{E}\big[\widehat{S}_n^2\big] &= \boldsymbol{E}\left[\overline{X_n^2} - \overline{X}_n^{\,2}\right] = \boldsymbol{E}\left[\frac{1}{n}\sum_{k=1}^{n} X_k^2\right] - \boldsymbol{E}\big[\overline{X}_n^{\,2}\big] = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{E}\left[X_k^2\right] - \boldsymbol{E}\big[\overline{X}_n^{\,2}\big] \\ &= \frac{n\,(\sigma^2 + \mu^2)}{n} - \left(\frac{\sigma^2}{n} + \mu^2\right) = \frac{n-1}{n}\,\sigma^2 < \sigma^2 \end{aligned}$$

e quindi $\widehat{S}_n^2$ è uno stimatore distorto di σ^2. Da questa stessa relazione e da (8.1) segue invece immediatamente che S_n^2 è uno stimatore corretto di σ^2 □

8.1.1 Stima di parametri

Dal Teorema 8.2 discende che la media aritmetica (5.9) $\overline{X}_n$ è lo stimatore più naturale per l'attesa μ di un *v-a* X; inoltre sia la varianza campionaria (5.10) $\widehat{S}_n^2$ che la varianza corretta (5.11) S_n^2 sono stimatori consistenti della varianza σ^2, ma solo S_n^2 è anche uno stimatore non distorto: per questo motivo in generale nei problemi di stima si preferisce usare la varianza corretta. Discuteremo ora qualche esempio di applicazione di questi risultati riprendendo anche alcuni aspetti degli Esempi 5.4 e 5.10 del Capitolo 5

Esempio 8.3. (Stima di un valore d'attesa) *Si voglia determinare il peso medio degli individui adulti maschi di una determinata specie di animali. In questo caso descriveremo il peso degli animali con una v-a X e il nostro problema sarà quello di determinare $\mu = \boldsymbol{E}[X]$. Non abbiamo a priori informazioni sulla legge di X, anche se potremmo ragionevolmente fare delle ipotesi (almeno come approssimazione) su tale legge. La nostra richiesta, però, è solo quella di stimare μ e noi seguiremo una procedura simile a quella dell'Esempio 5.4 e ora anche suggerita dal Teorema 8.2: estrarremo un campione casuale di n individui, ne misureremo il peso e prenderemo la media aritmetica $\overline{X}_n$ di tali misure come stima di μ. Da un punto di vista formale ciò vuol dire prendere n v-a indipendenti $X_1, \ldots, X_n$ che rappresentano il peso degli animali del campione casuale – quindi tali v-a sono tutte distribuite con la stessa legge (sconosciuta) di X e hanno lo stesso valore d'attesa μ – e calcolarne la media aritmetica $\overline{X}_n$ come in (5.9). Ovviamente $\overline{X}_n$ è una v-a il cui valore cambia al variare del campione aleatorio estratto in eventuali ripetizioni delle misure: noi prenderemo il valore ricavato da uno dei campione come stima puntuale della media μ, fidandoci del fatto che, in base al Teorema 8.2, se il campione è sufficientemente grande tale stima sarà ragionevolmente vicina al valore vero μ. Si noti come la stima della proporzione p dell'Esempio 5.4 non sia altro che un caso particolare di stima puntuale di un valore d'attesa*

Esempio 8.4. (Stima di un parametro) *In altri casi, sulla base di qualche ragionamento, è possible avanzare delle ipotesi sulla forma qualitativa della distribuzione di una data v-a X, i cui parametri sono però delle quantità incognite da stimare empiricamente. Così, riprendendo la discussione dell'Esempio 5.10, potremmo dire che il numero aleatorio X di telefonate che arrivano ad un centralino telefonico in un generico intervallo di tempo T è una v-a con legge di Poisson $\mathfrak{P}(\lambda)$. Il parametro λ però è sconosciuto: esso ovviamente dipende dal particolare centralino studiato e dal particolare periodo della giornata considerato, e il nostro problema è ora quello di stimarlo. Ricorderemo a questo proposito che in base all'equazione (4.25) il parametro λ è anche il valore d'attesa $\boldsymbol{E}[X]$ della nostra v-a di Poisson. Pertanto la maniera più naturale per stimare λ consisterà nel misurare in n giorni diversi il numero di telefonate pervenute al dato centralino in un ben determinato periodo di tempo, e nel calcolare poi la media aritmetica di queste misure. Ancora una volta avremo un campione casuale composto di n v-a indipendenti $X_1, \ldots, X_n$ tutte con*

6.03	5.95	7.26	5.27	5.44	3.84	3.94	3.62	3.30	5.36
4.18	3.80	5.42	4.39	4.92	4.93	3.89	5.14	5.70	4.89

Tabella 8.1: Campione di $n = 20$ misure di una v-a X.

legge $\mathfrak{P}(\lambda)$, a partire dal quale calcoliamo la v-a $\overline{X}_n$ (5.9) fidandoci del fatto che in base al Teorema 8.2, se n è abbastanza grande, il valore numerico osservato non sarà molto diverso dal valore di λ

Esempio 8.5. (Stima di una varianza) *Supponiamo ora che sia stato assegnato il campione di $n = 20$ misure di una v-a X con i valori raccolti nella Tabella 8.1. La stima puntuale dell'attesa μ di X è fornita dalla loro media aritmetica*

$$\overline{X}_n = \frac{1}{20}\sum_{k=1}^{20} X_k = 4.86$$

mentre per stimare la varianza σ^2 è possibile calcolare sia la varianza campionaria (5.10) $\widehat{S}_n^2$ che la varianza corretta (8.1) S_n^2

$$\widehat{S}_n^2 = \overline{X_n^2} - \overline{X}_n^2 = 0.94 \qquad S_n^2 = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^2\right) = 0.99$$

Come si vede le due stime puntuali della varianza sono leggermente diverse e in generale si preferisce quella non distorta S_n^2. È evidente comunque che la differenza fra le due stime puntuali ella varianza diventa sempre più piccola al crescere di n

8.1.2 Stima di distribuzioni

Come già anticipato nell'Esempio 5.4 del Capitolo 5, anche le probabilità associate alla distribuzione di X possono essere stimate riconducendole al calcolo di particolari valori d'attesa. Nel seguito descriveremo brevemente le opportune procedure di stima relative cominciando con il caso delle distribuzioni discrete

Teorema 8.6. (Stima di una distribuzione discreta) *Sia $X_1, \ldots, X_n$ un campione di una v-a discreta X con distribuzione sconosciuta*

$$p_k = \boldsymbol{P}\{X = x_k\} \qquad k = 0, 1, 2, \ldots$$

Posto allora

$$Y(k) = \begin{cases} 1 & \text{se } X = x_k \\ 0 & \text{altrimenti} \end{cases} \qquad Y_j(k) = \begin{cases} 1 & \text{se } X_j = x_k \\ 0 & \text{altrimenti} \end{cases} \qquad j = 1, 2, \ldots, n$$

*le **frequenze relative empiriche** dei valori x_k, cioè le v-a*

$$\overline{p}_k = \frac{N_k}{n} \qquad \text{con} \quad N_k = \sum_{j=1}^{n} Y_j(k)$$

sono stimatori corretti e consistenti delle probabilità p_k

Dimostrazione: È evidente dalla definizione che $Y(k) \sim \mathfrak{B}(1; p_k)$ è una *v-a* di Bernoulli, sicché da (4.24) risulterà per ogni k

$$\boldsymbol{E}[Y(k)] = \boldsymbol{P}\{X = x_k\} = p_k$$

La stima delle p_k quindi può essere ricondotta alla stima dell'attesa di $Y(k)$. D'altronde, per ogni k fissato, le $Y_1(k), \ldots, Y_n(k)$ costituiscono un campione di $Y(k)$, per cui dal Teorema 8.2 discende che una stima corretta e consistente di $p_k = \boldsymbol{E}[Y(k)]$ è data dalla media aritmetica di tale campione, cioè con le notazioni adottate:

$$\overline{Y}(k) = \frac{1}{n}\sum_{j=1}^{n} Y_j(k) = \frac{N_k}{n} = \overline{p}_k$$

Inoltre, dalla definizione, si ha che

$$N_k = \text{ numero delle } Y_j(k) \text{ che valgono } 1 \;=\; \text{numero delle } X_j \text{ che valgono } x_k$$

cioè le N_k sono le frequenze assolute dei ritrovamenti di x_k nel campione dato, e quindi gli stimatori $\overline{p}_k$ sono le frequenze relative, come enunciato nel teorema □

Per estendere i risultati del Teorema 8.6 anche al caso di **distribuzioni continue** dovremo adottare una procedura analoga a quella introdotta nella Sezione 6.2 per passare dai *diagrammi a barre* agli *istogrammi*. Se X è una *v-a* continua con *fdp* $f(x)$, infatti, non ha più molto senso domandarsi quante volte gli elementi di un suo campione $X_1, \ldots, X_n$ coincidono con uno dei suoi possibili valori x. Divideremo invece l'intervallo dei valori di X in un numero finito di intervalli $\mathcal{B}_1, \ldots, \mathcal{B}_m$ e ci porremo il problema di utilizzare il campione dato per stimare le probabilità p_k di trovare X in $\mathcal{B}_k = [a_k, b_k]$ con $k = 1, \ldots, m$, cioè

$$p_k = \boldsymbol{P}\{X \in \mathcal{B}_k\} = \int_{\mathcal{B}_k} f(x)\,dx = \int_{a_k}^{b_k} f(x)\,dx$$

Per far questo si definiscono

$$Y(k) = \begin{cases} 1 & \text{se } X \in [a_k, b_k] \\ 0 & \text{altrimenti} \end{cases} \qquad Y_j(k) = \begin{cases} 1 & \text{se } X_j \in [a_k, b_k] \\ 0 & \text{altrimenti} \end{cases} \qquad j = 1, 2, \ldots, n$$

e, seguendo esattamente la stessa traccia del Teorema 8.6, si dimostra che le frequenze relative empiriche dei ritrovamenti del campione in $\mathcal{B}_k$, cioè le v-a

$$\overline{p}_k = \frac{N_k}{n} \qquad \text{con} \quad N_k = \sum_{j=1}^{n} Y_j(k)$$

sono stimatori corretti e consistenti delle probabilità p_k

Se vogliamo poi istituire un confronto grafico fra la *fdp* $f(x)$ di X e i dati empirici, potremo paragonare la curva $f(x)$ con l'**istogramma del campione**. Infatti

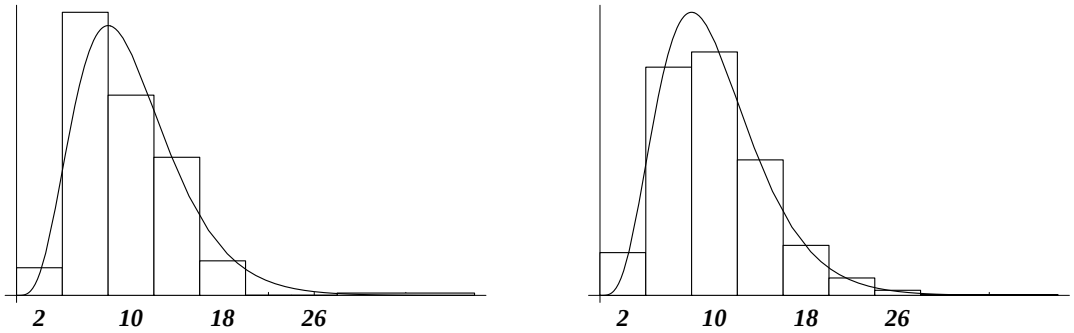


Figura 8.1: Approssimazione della *fdp* di una legge $\boldsymbol{\chi^2}(10)$ con istogrammi ottenuti da dati simulati. L'approssimazione migliora passando da $n = 100$ (primo grafico) a $n = 1\,000$ campioni (secondo grafico).

sulla base delle notazioni appena introdotte, e scegliendo in ogni $\mathcal{B}_k = [a_k, b_k]$ un opportuno punto $\overline{x}_k$, avremo

$$\overline{p}_k \simeq \int_{a_k}^{b_k} f(x)\,dx = (b_k - a_k) f(\overline{x}_k)$$

e quindi

$$f(\overline{x}_k) \simeq \frac{\overline{p}_k}{(b_k - a_k)}$$

cioè i valori di $f(x)$ in $[a_k, b_k]$ sono approssimati dalle altezze $\overline{p}_k/(b_k - a_k)$ dell'istogramma del campione

Esempio 8.7. (Stima di una distribuzione) *Un esempio di come un istogramma empirico approssimi una fdp continua assegnata è mostrato nella Figura 8.1: la curva continua rappresenta la fdp della legge* $\boldsymbol{\chi^2}(10)$*, mentre gli istogrammi sono stati ottenuti simulando dei campioni di v-a che seguono tale legge. L'altezza dei rettangoli* $\overline{p}_k/(b_k - a_k)$ *approssima i valori della* fdp*, e si noti in particolare come, per effetto della* LGN*, l'approssimazione migliori passando da un campione di* $n = 100$ *elementi (grafico a sinistra) a uno di* $n = 1\,000$ *elementi (grafico a destra)*

In molti casi però la distribuzione (discreta o continua che essa sia) della *v-a* X non è conosciuta *a priori*, mentre sono a disposizione dei campioni empirici $X_1, \ldots, X_n$ con i relativi istogrammi. Si può porre allora il problema di **individuare la distribuzione teorica** che meglio si adatta ai dati sperimentali, eventualmente procedendo con dei ragionevoli tentativi: un argomento che tratteremo anche successivamente – ma da una differente prospettiva – nella Sezione 9.5 sui Test di adattamento. Qui proporremo solo la discussione di un tipico esempio per mettere in evidenza le difficoltà che possono sorgere in questo tipo di indagini

k	N_k	p_k	$\overline{p}_k$	$p_k/\overline{p}_k$
0	3	0.00024	0.00049	0.49764
1	24	0.00293	0.00392	0.74646
2	104	0.01611	0.01701	0.94743
3	286	0.05371	0.04677	1.14840
4	670	0.12085	0.10957	1.10298
5	1 033	0.19336	0.16893	1.14462
6	1 343	0.22559	0.21962	1.02715
7	1 112	0.19336	0.18185	1.06330
8	829	0.12085	0.13557	0.89143
9	478	0.05371	0.07817	0.68712
10	181	0.01611	0.02960	0.54438
11	45	0.00293	0.00736	0.39811
12	7	0.00024	0.00114	0.21327

Tabella 8.2: Frequenze assolute N_k delle famiglie con 12 figli e k figli maschi, su $n = 6\,115$ famiglie. Le corrispondenti frequenze relative $\overline{p}_k = N_k/n$ del campione $X_1, \ldots, X_n$ sono usate per stimare la distribuzione della *v-a* X che descriva il *numero di figli maschi nelle famiglie con 12 figli*, e sono quindi confrontate con i valori teorici delle p_k di una legge binomiale $\mathfrak{B}(12; {}^1/_2)$ associata ad X in base ad alcune semplici ipotesi

Esempio 8.8. (Adattamento a una distribuzione) *Supponiamo di voler studiare la distribuzione della v-a X che rappresenta il numero di figli maschi nelle famiglie con 12 figli. Ovviamente X è una v-a discreta che assume solo i 13 valori interi $k = 0, 1, \ldots, 12$. Per determinare la legge di X possiamo iniziare costruendo un modello sulla base delle ipotesi più elementari possibili:*

1. *in ogni famiglia gli esiti di parti differenti sono indipendenti*
2. *in ogni parto nasce un maschio o una femmina con eguale probabilità ${}^1/_2$*

In base a queste ipotesi le 12 v-a $Y_1, \ldots, Y_{12}$

$$Y_j = \begin{cases} 1 & \text{se al parto } j\text{–mo nasce un figlio maschio,} \\ 0 & \text{se al parto } j\text{–mo nasce una figlia femmina,} \end{cases} \qquad j = 1, 2, \ldots, 12$$

saranno indipendenti e ciascuna con legge di Bernoulli $\mathfrak{B}(1; {}^1/_2)$: esse rappresentano gli esiti dei 12 parti indipendenti, per cui in base al Teorema 3.16 risulterà $X = Y_1 + \ldots + Y_{12} \sim \mathfrak{B}(12; {}^1/_2)$. Pertanto con le nostre ipotesi le p_k teoriche sono

$$p_k = \boldsymbol{P}\{X = k\} = \binom{12}{k} \left({}^1/_2\right)^k \left(1 - {}^1/_2\right)^{12-k} = \binom{12}{k} \frac{1}{2^{12}}$$

Confrontiamo ora questo modello con alcuni risultati sperimentali. Dall'esame dei dati anagrafici $X_1, \ldots, X_n$ di $n = 6\,115$ famiglie con 12 figli si ricavano i numeri

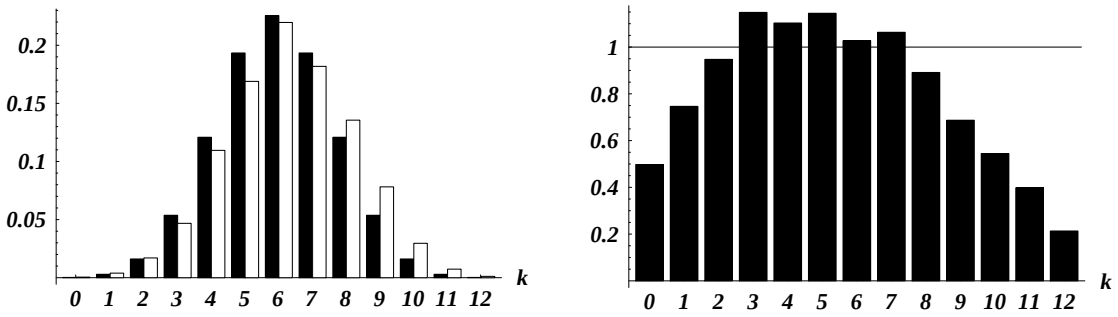


Figura 8.2: Confronto fra le frequenze teoriche p_k di $\mathfrak{B}(12; {}^1/_2)$ (barre nere), e le frequenze empiriche $\overline{p}_k$ (barre bianche). Il secondo grafico riporta i valori dei rapporti ${}^{p_k}/_{\overline{p}_k}$

N_k (frequenze assolute) di famiglie con k figli maschi; si costruiscono poi, come nel Teorema 8.6, le stime $\overline{p}_k = {}^{N_k}/_n$ (frequenze relative) e le si confronta con i valori teorici p_k. I risultati sono riportati nella Tabella 8.2 e nei grafici di Figura 8.2. Sebbene i dati teorici riproducano qualitativamente l'andamento dei dati empirici (si guardi il comportamento delle barre nere e bianche nella prima parte della Figura 8.2), ad un'analisi più accurata non sfugge che l'accordo non è particolarmente buono. Per mettere in risalto le differenze fra la distribuzione teorica e quella empirica possiamo calcolare i rapporti ${}^{p_k}/_{\overline{p}_k}$ per vedere di quanto essi si discostano dal valore ottimale 1*. I valori di questo rapporto sono riportati sempre nella Tabella 8.2 e nella seconda parte della Figura 8.2. Si può notare così ad esempio che i dati teorici sono sistematicamente un po' più grandi di quelli empirici per valori centrali di k, ma soprattutto più piccoli per i valori estremi*

Per migliorare l'accordo con i dati sperimentali potremmo allora modificare le ipotesi 1. e 2. che sono alla base del nostro modello teorico. Siccome però dall'ipotesi 1. dipende la forma binomiale della legge di X, una sua modifica produrrebbe sicuramente dei cambiamenti piuttosto profondi nella natura del nostro modello teorico. Conviene invece partire dalla discussione, più elementare, dell'ipotesi 2: se essa non fosse vera la X sarebbe sempre distribuita secondo una legge binomiale $\mathfrak{B}(12; p)$, ma con un parametro p diverso da ${}^1/_2$. Dobbiamo allora trovare un modo per stimare il valore di p in assenza dell'ipotsi 2, ma utilizzando i dati empirici a nostra disposizione. Ricordando allora che per il Teorema 8.6 un buon stimatore di una probabilità è la corrispondente frequenza relativa empirica, basterà osservare che nel nostro esempio, su $12\,n = 73\,380$ figli di ambo i sessi, i figli maschi sono

$$\sum_{i=1}^{n} X_i = \sum_{k=0}^{12} k N_k = 38\,100$$

sicché la frequenza relativa totale delle nascite maschili sarà

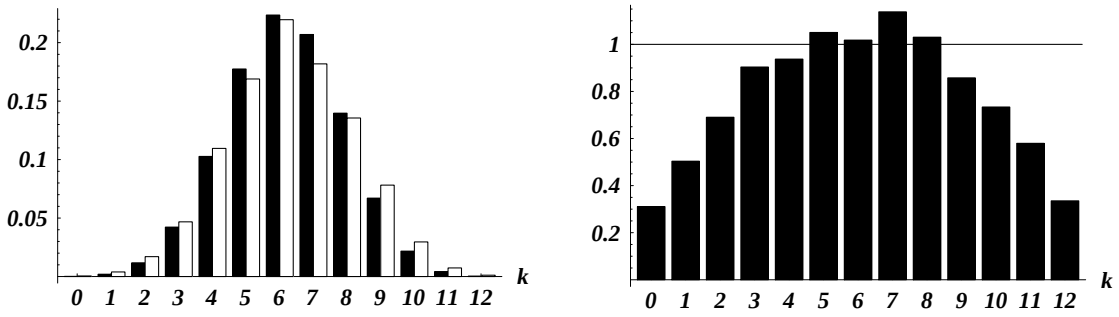


Figura 8.3: Confronto fra le frequenze teoriche p_k di $\mathfrak{B}(12;\overline{p})$ con $\overline{p} = 0.519$ stimato a partire dai dati sperimentali (barre nere), e le frequenze empiriche $\overline{p}_k$ (barre bianche). Il secondo grafico riporta i valori dei rapporti $p_k/\overline{p}_k$

$$\overline{p} = \frac{1}{12\,n}\sum_{k=0}^{12} kN_k \simeq 0.519$$

un numero che risulta leggermente diversa dal valore $1/2$ *del nostro modello iniziale. Possiamo allora provare a ripetere il confronto fra dati sperimentali e teorici nel nuovo modello binomiale* $\mathfrak{B}(12;\overline{p})$ *con* $\overline{p} = 0.519$*. Evitando per brevità di riportare i dati numerici in una nuova tabella, eseguiremo il confronto solo sui nuovi grafici che se ne ricavano e che sono riportati in Figura 8.3. Come si può notare i grafici sono ora leggermente diversi: ad esempio il grafico dei rapporti* $p_k/\overline{p}_k$ *è un po' più simmetrico, e al centro i valori sono ragionevolmente prossimi a* 1*. Non si può però dire che l'accordo con i dati empirici sia sostanzialmente migliorato per i valori estremi di* k*: le previsioni teoriche continuano infatti a sottovalutare sistematicamente i dati sperimentali sulle code della distribuzione (valori estremi di* k*)*

In conclusione questa discussione sembra suggerire che le difficoltà del nostro modello non sono nel valore del parametro p*, ma nel carattere binomiale della distribuzione di* X*. Sarebbe necessario quindi rivedere l'ipotesi 1. di indipendenza che è alla base del modello binomiale. Da un punto di vista concettuale questa revisione potrebbe portare a delle conclusioni importante: con l'abbandono dell'ipotesi 1. infatti la non indipendenza degli esiti (maschio o femmina) dei diversi parti indicherebbe che ci sono famiglie con una tendenza ad avere figli maschi, e famiglie con una tendenza ad avere figlie femmine. Discuteremo brevemente questo punto (le cui implicazioni genetiche sono fuori dell'ambito di questo corso) solo come ulteriore applicazione della LGN*

Riprendiamo le 12 *v-a* $Y_1, \ldots, Y_{12}$ *definite inizialmente e supponiamo ora che ciascuna segua una legge di Bernoulli* $\mathfrak{B}(1;\overline{p})$ *con* $\overline{p} = 0.519$ *stimato dai dati sperimentali. Se queste* Y_j *fossero indipendenti – come da noi finora ipotizzato – la* X *sarebbe binomiale* $\mathfrak{B}(12;\overline{p})$ *in base al Teorema 3.16, e in questo caso da (4.24)*

dovrebbe risultare anche

$$\boldsymbol{V}[X] = \boldsymbol{V}[Y_1] + \ldots + \boldsymbol{V}[Y_{12}] = 12\,\overline{p}\,(1-\overline{p}) = 2.996$$

Se invece le Y_j non sono indipendenti da un lato X non sarebbe più binomiale, e dall'altro non si potrebbe più applicare la (4.24). Ricordiamo infatti che in questo caso la varianza della somma $Y_1 + \ldots + Y_{12}$ non è più la semplice somma della varianze delle Y_j, ma deve essere invece usata l'equazione (4.20) che tiene conto delle rispettive covarianze. La LGN e il Teorema 8.2, però, ci permettono di stimare la varianza di X direttamente dai dati sperimentali tramite le varianze empiriche (8.1) (in questo caso, dato che $n = 6\,115$ è grande, fa poca differenza scegliere S_n^2 piuttosto che $\widehat{S}_n^2$), e soprattutto senza fare ipotesi sull'indipendenza delle Y_j. Infatti la media e la media dei quadrati del campione $X_1, \ldots, X_n$ della X riportato in Tabella 8.2 sono

$$\begin{aligned}
\overline{X}_n &= \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{k=0}^{12} kN_k \simeq 6.231 \\
\overline{X_n^2} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \frac{1}{n}\sum_{k=0}^{12} k^2 N_k \simeq 42.309
\end{aligned}$$

e quindi da (8.1) si ottiene

$$\widehat{S}_n^2 = \overline{X_n^2} - \overline{X}_n^2 \simeq 3.489$$

Il valore della varianza così stimato appare dunque abbastanza diverso dal valore 2.996 *ottenuto sommando solo le varianze delle Y_j con l'ipotesi della loro indipendenza. Tutto questo suggerisce ancora una volta una dipendenza reciproca delle Y_j. Anzi, tenendo conto di questi valori e della relazione (4.20), si ha*

$$\begin{aligned}
\sum_{i\neq j} \boldsymbol{cov}[Y_i, Y_j] &= \boldsymbol{V}[Y_1 + \ldots + Y_{12}] - \big(\boldsymbol{V}[Y_1] + \ldots + \boldsymbol{V}[Y_{12}]\big) \\
&= 3.489 - 2.996 = 0.493
\end{aligned}$$

cioè i dati sperimentali sembrano indicare che le covarianze fra gli esiti Y_j di parti diversi siano complessivamente positive, ossia che le Y_j siano positivamente correlate. Dato il significato del concetto di covarianza come discusso nella Sezione 4.1, questo suggerirebbe che vi sono famiglie con la tendenza a generare figli maschi, e famiglie con la tendenza a generare figlie femmine. L'esito dei parto potrebbe, cioè, non essere una faccenda puramente lasciata al caso: se in una famiglia si osservano nascite maschili (rispettivamente: femminili), la probabilità che anche le nascite successive siano maschili (rispettivamente:femminili) aumenta.

8.2 Stima per intervalli

La stima puntuale di $h(\theta)$ resta una risposta piuttosto grossolana al problema di determinare una ragionevole approssimazione del valore vero incognito. In particolare è evidente che il valore stimato non sarà mai uguale al valore vero $h(\theta)$, e che la teoria della stima puntuale non permette di valutare neanche probabilisticamente l'entità della differenza fra i due valori. A questa necessità risponde invece la teoria della stima per intervalli: in pratica si rinuncia a stimare $h(\theta)$ con un solo valore di uno stimatore aleatorio, e si preferisce determinare – sempre a partire dai valori del campione di misure – i **due estremi aleatori di un intervallo** prefissando in maniera opportuna il valore della probabilità dell'evento: *l'intervallo aleatorio contiene* $h(\theta)$. La differenza principale sta nel fatto che, mentre la probabilità di centrare il valore vero con una stima puntuale è sempre nulla, la probabilità che un intervallo con estremi aleatori contenga $h(\theta)$ è diversa da zero e in generale può anche essere calcolata. Nel seguito supporremo sempre di avere a disposizione un campione casuale $X_1, \ldots, X_n$ di una *v-a* X con legge dipendente da uno o più parametri θ

Definizione 8.9. *Assegnato un campione* $X_1, \ldots, X_n$*, diremo che le due v-a* $U = u(X_1, \ldots, X_n)$ *e* $V = v(X_1, \ldots, X_n)$ *sono gli estremi di un* ***intervallo di fiducia (o di confidenza)*** $[U, V]$ ***di livello*** α *(con* $0 < \alpha < 1$*) per* $h(\theta)$ *quando*

$$\boldsymbol{P}_\theta\{U \leq h(\theta) \leq V\} = 1 - \alpha$$

dove il simbolo $\boldsymbol{P}_\theta$ *indica che la probabilità è calcolata supponendo che il valore del parametro incognito sia lo stesso che compare in* $h(\theta)$

Questa definizione, però, lascia ancora dei margini di ambiguità: infatti per un dato campione e per un α fissato l'intervallo di fiducia non è unico. In particolare ci sono molti modi in cui si può ripartire la probabilità α che l'intervallo *non contenga* $h(\theta)$. In genere si risolve tale ambiguità decidendo di fissare U e V in modo che

$$\boldsymbol{P}_\theta\{h(\theta) < U < V\} = \boldsymbol{P}_\theta\{U < V < h(\theta)\} = {}^{\alpha}/_{2} \tag{8.3}$$

cioè che le probabilità di avere ambedue gli estremi o troppo grandi, o troppo piccoli per contenere $h(\theta)$ siano uguali, e valgano ${}^{\alpha}/_{2}$. Con questa precisazione l'intervallo è univocamente determinato, e – nei casi dotati di simmetria – i suoi estremi assumono la forma $W \pm \Delta$, dove il valore centrale W è un opportuno stimatore di $h(\theta)$ e l'intervallo $[W - \Delta\,,\, W + \Delta]$ ha ampiezza aleatoria e 2Δ. Naturalmente, per un dato campione, l'ampiezza dell'intervallo di fiducia dipende dalla scelta del valore di α. Tipicamente si scelgono valori piccoli di α (ad esempio 0.05 oppure 0.01), in modo che la probabilità $1 - \alpha$ che l'intervallo contenga il valore vero sia corrispondentemente grande (ad esempio 0.95 oppure 0.99). È abbastanza ovvio quindi che al diminuire di α l'intervallo di fiducia debba allargarsi, e che conseguentemente ha anche poco senso richiedere α eccessivamente piccoli, perché questi corrisponderebbero a intervalli così larghi da essere poco significativi

8.2.1 Intervallo di fiducia per l'attesa μ

Supporremo inizialmente di avere un campione $X_1, \ldots, X_n$ di una *v-a* X con legge normale $\mathfrak{N}(\mu, \sigma^2)$, e di voler stimare il valore di μ. Ricordando che la media aritmetica $\overline{X}_n$ è uno stimatore non distorto e consistente per μ, dovremo distinguere due casi sulla base delle informazioni che abbiamo su σ^2

Varianza σ^2 conosciuta

Se il valore della varianza σ^2 è noto, dal Teorema 3.23 sappiamo che

$$Y_n^* = \sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma} \sim \mathfrak{N}(0,1)$$

cioè Y_n^* è normale standard. Pertanto da (3.26) si ha

$$\boldsymbol{P}\big\{|Y_n^*| \le \varphi_{1-\frac{\alpha}{2}}\big\} = 1-\alpha$$

e quindi potremo scrivere che

$$\begin{aligned} 1-\alpha &= \boldsymbol{P}\left\{\left|\sqrt{n}\,\frac{\overline{X}_n - \mu}{\sigma}\right| \le \varphi_{1-\frac{\alpha}{2}}\right\} = \boldsymbol{P}\left\{|\overline{X}_n - \mu| \le \frac{\sigma}{\sqrt{n}}\,\varphi_{1-\frac{\alpha}{2}}\right\} \\ &= \boldsymbol{P}\left\{\overline{X}_n - \frac{\sigma}{\sqrt{n}}\,\varphi_{1-\frac{\alpha}{2}} \le \mu \le \overline{X}_n + \frac{\sigma}{\sqrt{n}}\,\varphi_{1-\frac{\alpha}{2}}\right\} \end{aligned}$$

In questo modo si riconosce subito dalla Definizione 8.9 che l'intervallo di fiducia di livello α, che soddisfa anche la condizione (8.3), è

$$\left[\overline{X}_n - \frac{\sigma}{\sqrt{n}}\,\varphi_{1-\frac{\alpha}{2}}\ ,\ \overline{X}_n + \frac{\sigma}{\sqrt{n}}\,\varphi_{1-\frac{\alpha}{2}}\right]$$

e che i suoi estremi possono essere espressi sinteticamente come

$$\boxed{\overline{X}_n \pm \frac{\sigma}{\sqrt{n}}\,\varphi_{1-\frac{\alpha}{2}}} \tag{8.4}$$

Varianza σ^2 non conosciuta

Se invece, come più spesso accade, il valore della varianza σ^2 non è noto, bisogna prima stimare σ^2 usando lo stimatore corretto S_n^2, e poi osservare che, sempre per il Teorema 3.23, si ha

$$T_n = \sqrt{n}\,\frac{\overline{X}_n - \mu}{S_n} \sim \mathfrak{T}(n-1)$$

cioè la *v-a* T_n è distribuita secondo una legge di Student $\mathfrak{T}(n-1)$. Pertanto, seguendo gli stessi passaggi del caso precedente, ma partendo da (3.28), si ha ora che l'intervallo di fiducia di livello α che soddisfa (8.3) è

$$\left[\overline{X}_n - \frac{S_n}{\sqrt{n}}\,t_{1-\frac{\alpha}{2}}(n-1)\ ,\ \overline{X}_n + \frac{S_n}{\sqrt{n}}\,t_{1-\frac{\alpha}{2}}(n-1)\right]$$

e che i suoi estremi possono essere espressi sinteticamente come

$$\boxed{\overline{X}_n \pm \frac{S_n}{\sqrt{n}}\, t_{1-\frac{\alpha}{2}}(n-1)} \tag{8.5}$$

In pratica dunque gli intervalli di fiducia per l'attesa μ hanno sempre il centro in $\overline{X}_n$, e un'ampiezza Δ_n che varia secondo i casi. Essa infatti dipende innanzitutto dal campione dato, nel senso che Δ_n cresce con σ (o con la sua stima S_n) e diminuisce con $\sqrt{n}$. Inoltre Δ_n dipende anche da un opportuno quantile della legge (normale o di Student) che meglio descrive la *v-a* standardizzata. Il fatto ovvio che tali quantili crescano quando α diminuisce, indica chiaramente che per avere una maggiore probabilità di contenere μ bisogna allargare l'intervallo di fiducia, evitando però valori eccessivi: se α è troppo piccolo, l'intervallo di fiducia diviene talmente largo da rendere banale l'evento $U \le \mu \le V$, lasciando poca informazione nel fatto che un *grande* intervallo contiene μ con grande probabilità

Strettamente parlando le formule (8.4) e (8.5) forniscono gli intervalli di fiducia per μ solo se la X studiata è *gaussiana*. Esse restano però approssimativamente valide – purché n sia abbastanza grande – anche nel caso più generale di *v-a* X *non gaussiane* per merito del Teorema 5.5 (TLC) discusso nella Sezione 5.3. In pratica esse saranno applicate in ogni caso quando $n \ge 20$

Si noti infine che le Tavole E.2 delle leggi di Student si arrestano a $n = 120$ gradi di libertà. Va ricordato però a questo proposito che la differenza fra i quantili normali e i quantili di Student diminuisce all'aumentare di n: per $n > 120$ essa diviene sostanzialmente irrilevante, e i quantili $\varphi_{1-\frac{\alpha}{2}}$ possono essere senz'altro usati al posto dei quantili $t_{1-\frac{\alpha}{2}}(n-1)$ nella (8.5) che in questo caso diviene

$$\boxed{\overline{X}_n \pm \frac{S_n}{\sqrt{n}}\, \varphi_{1-\frac{\alpha}{2}}} \qquad n > 120 \tag{8.6}$$

mescolando così in un certo senso le formule (8.4) e (8.5)

Esempio 8.10. *Supponiamo di effettuare* $n = 100$ *misure di una quantità fisica affetta da un errore sperimentale casuale* $Z \sim \mathfrak{N}(0, \sigma^2)$ *con* $\sigma^2 = 4$ *conosciuta. Se il valore vero (sconosciuto) della quantità fisica è* μ*, i risultati delle misure sono v-a del tipo* $X = \mu + Z$*, e avremo anche* $X \sim \mathfrak{N}(\mu, \sigma^2) = \mathfrak{N}(\mu, 4)$ *sicché il campione di misure sarà Gaussiano. Supponendo ora che la media delle* 100 *misure sia* $\overline{X}_n = 50$*, vogliamo determinare un intervallo di fiducia di livello* $\alpha = 0.01$ *per* μ*. Siccome in questo caso la varianza* $\sigma^2 = 4$ *è nota, possiamo semplicemente applicare (8.4), e usando le Tavole E.1 si ha*

$$\overline{X}_n \pm \frac{\sigma}{\sqrt{n}}\, \varphi_{1-\frac{\alpha}{2}} = \overline{X}_n \pm \frac{\sigma}{\sqrt{n}}\, \varphi_{0.995} = 50 \pm \frac{2}{\sqrt{100}}\, 2.58 = 50 \pm 0.52$$

cioè l'intervallo $[49.48\,,\ 50.52]$ *contiene il valore vero della nostra quantità fisica con il* 99 % *di probabilità*

Esempio 8.11. *Torniamo ora al problema – introdotto nell'Esempio 5.4 – si stimare la proporzione p di individui di tipo A in una popolazione composta di individui di tipo A e B: supponiamo di aver estratto un campione di $n = 100$ misure delle quali 57 sono di tipo A, e calcoliamo un intervallo di fiducia di livello $\alpha = 0.05$ per p. Formalmente (vedi Esempio 5.4) gli esiti possono essere rappresentati da un campione $X_1, \ldots, X_{100}$, composto di $n = 100$ v-a tutte indipendenti e Bernoulli $\mathfrak{B}(1; p)$, dal quale ricaviamo immediatamente la stima puntuale di p*

$$\overline{X}_n = \frac{1}{100} \sum_{j=1}^{100} X_j = \frac{57}{100} = 0.570$$

Il valore $n = 100$ è anche abbastanza grande per poter supporre che la media standardizzata delle X_j sia normale standard $\mathfrak{N}(0, 1)$. Siccome però non abbiamo informazioni sul valore della varianza (che per una Bernoulli implicherebbe anche la conoscenza di p) dobbiamo preventivamente stimarne il valore mediante la varianza corretta (8.1): ricordando che per una Bernoulli si ha sempre $X_j^2 = X_j$, avremo in ogni caso $\overline{X_n^2} = \overline{X}_n$, e quindi

$$S_n^2 = \frac{n}{n-1} \left(\overline{X_n^2} - \overline{X}_n^2 \right) = \frac{n}{n-1} \left(\overline{X}_n - \overline{X}_n^2 \right) = \frac{100}{99} \left(0.570 - 0.570^2 \right) \simeq 0.248$$

Usando poi l'equazione (8.5) e le Tavole E.2 delle leggi di Student si ottiene in definitiva l'intervallo

$$\begin{aligned} \overline{X}_n \pm \frac{S_n}{\sqrt{n}} t_{1-\frac{\alpha}{2}}(n-1) &= \overline{X}_n \pm \frac{S_n}{\sqrt{n}} t_{0.975}(99) \simeq 0.570 \pm \frac{0.498}{10} 1.984 \\ &\simeq 0.570 \pm 0.099 \end{aligned}$$

In realtà il quantile $t_{0.975}(99)$ necessario per il calcolo dell'intervallo non è presente nelle Tavole E.2 (che non sono ovviamente complete): per aggirare questo problema si usa il quantile più vicino ($t_{0.975}(100)$ in questo caso), osservando peraltro che la variabilità dei valori fra i quantili di Student con 90 e 100 gradi di libertà è limitata a poche unità sulla terza cifra decimale. Infine si noti che, essendo $n = 100$ abbastanza grande, il valore del quantile di Student $t_{0.975}(100) \simeq 1.984$ non differisce eccessivamente dal corrispondente quantile della normale standard $\varphi_{0.975} \simeq 1.960$ ricavabile dalle Tavole E.1. Se allora avessimo calcolato l'intervallo di fiducia con la formula mista (8.6) avremmo trovato

$$\overline{X}_n \pm \frac{S_n}{\sqrt{n}} \varphi_{1-\frac{\alpha}{2}} \simeq 0.570 \pm \frac{0.498}{10} 1.960 \simeq 0.570 \pm 0.098$$

un intervallo non molto diverso dal precedente. In conclusione possiamo dire che, per il nostro esempio, un intervallo di fiducia che contiene il valore vero di p con il $95\,\%$ di probabilità sarà approssimativamente del tipo $[0.47\,,\ 0.67]$

8.2.2 Intervallo di fiducia per la varianza σ^2

Supponiamo ora nuovamente di avere un campione $X_1, \ldots, X_n$ di una *v-a* $X \sim \mathfrak{N}(\mu, \sigma^2)$, e di voler stimare la varianza σ^2. Già sappiamo che la varianza corretta S_n^2 è uno stimatore non distorto e consistente di σ^2, e inoltre dal Teorema 3.23 abbiamo che

$$Z_n = (n-1)\,\frac{S_n^2}{\sigma^2} \sim \boldsymbol{\chi^2}(n-1)$$

segue una legge *chi-quadro*. Tenendo conto di (3.30) si ha allora

$$\begin{aligned} 1-\alpha &= \boldsymbol{P}\Big\{\chi^2_{\frac{\alpha}{2}}(n-1) \le Z_n \le \chi^2_{1-\frac{\alpha}{2}}(n-1)\Big\} \\ &= \boldsymbol{P}\left\{\chi^2_{\frac{\alpha}{2}}(n-1) \le (n-1)\,\frac{S_n^2}{\sigma^2} \le \chi^2_{1-\frac{\alpha}{2}}(n-1)\right\} \\ &= \boldsymbol{P}\left\{\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \le \sigma^2 \le \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right\} \end{aligned}$$

e quindi l'intervallo di fiducia di livello α, che soddisfi la condizione (8.3), ha la forma

$$\boxed{\left[\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\,,\ \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]} \tag{8.7}$$

Anche qui questo risultato, che è esatto quando $X \sim \mathfrak{N}(\mu, \sigma^2)$, resta approssimativamente corretto anche in casi non Gaussiani per effetto del Teorema 5.5

Esempio 8.12. *Si voglia calcolare l'intervallo di fiducia di livello* $\alpha = 0.05$ *per la varianza dell'Esempio 8.11. Da (8.7) e dai calcoli sviluppati in precedenza abbiamo che l'estremo sinistro dell'intervallo è*

$$\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} = \frac{99\,S_n^2}{\chi^2_{0.975}(99)}$$

ma siccome le Tavole E.3 si arrestano a 35 gradi di libertà dovremo usare la formula approssimata (3.31), cioè dalle Tavole E.1

$$\chi^2_{0.975}(99) \simeq \frac{\left(\varphi_{0.975} + \sqrt{199}\right)^2}{2} \simeq \frac{\left(1.960 + \sqrt{199}\right)^2}{2} \simeq 129.07$$

per cui in definitiva il valore dell'estremo sinistro è

$$\frac{99\,S_n^2}{\chi^2_{0.975}(99)} \simeq \frac{99 \times 0.248}{129.07} \simeq 0.19$$

Per l'estremo destro si ha analogamente

$$\frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} = \frac{99\,S_n^2}{\chi^2_{0.025}(99)}$$

e per il quantile richiesto (sempre da (3.31), ma ricordando anche che $\varphi_\alpha = -\varphi_{1-\alpha}$)

$$\chi^2_{0.025}(99) \simeq \frac{\left(\varphi_{0.025} + \sqrt{199}\right)^2}{2} = \frac{\left(-\varphi_{0.975} + \sqrt{199}\right)^2}{2} \simeq \frac{\left(-1.960 + \sqrt{199}\right)^2}{2} \simeq 73.77$$

Pertanto l'estremo destro è

$$\frac{99\, S_n^2}{\chi^2_{0.025}(99)} \simeq \frac{99 \times 0.248}{73.77} \simeq 0.33$$

e complessivamente il richiesto intervallo di fiducia è $[0.19\,,\ 0.33]$. *Si noti che ovviamente la stima puntuale* $S_n^2 \simeq 0.248$ *cade all'interno dell'intervallo di fiducia, ma – a causa della natura non simmetrica delle leggi* $\boldsymbol{\chi^2}$ *– non occupa più il suo centro*

8.3 Stima di Massima Verosimiglianza

Non sempre la forma dello stimatore può essere indovinata in maniera naturale, come è successo nel caso della media e della varianza per effetto della *LGN*. Sarà utile quindi avere un criterio generale per determinare opportuni stimatori di un parametro θ.

Definizione 8.13. *Dato un campione $X_1, \ldots, X_n$ di una v-a X, chiameremo* ***funzione di verosimiglianza*** *la funzione di θ*

$$L(\theta) = \boldsymbol{P}_\theta\{X_1 = x_1\} \cdot \ldots \cdot \boldsymbol{P}_\theta\{X_n = x_n\} = p_\theta(x_1) \cdot \ldots \cdot p_\theta(x_n) \tag{8.8}$$

se X è una ***v-a discreta*** *con valori x_k, ovvero la funzione di θ*

$$L(\theta) = f_\theta(x_1) \cdot \ldots \cdot f_\theta(x_n) \tag{8.9}$$

quando X è una ***v-a continua*** *con fdp $f_\theta(x)$*

Ricordando che le X_k di un campione sono sempre indipendenti e identicamente distribuite, osserviamo innanzitutto che $L(\theta)$ è rispettivamente la **distribuzione congiunta del campione** nel caso discreto come in (3.35), e la ***fdp* congiunta del campione** nel caso continuo come in (3.37). In ambedue le eventualità le probabilità e le *fdp* a secondo membro sono assegnate supponendo che θ abbia lo stesso valore che compare il $L(\theta)$

Se ora i valori $x_1, \ldots, x_n$ del campione sono considerati assegnati, $L(\theta)$ dipenderà solo da θ e sarà in generale possibile determinare il particolare valore $\widehat{\theta}$ di θ per il quale $L(\theta)$ è massima. In questo modo $\widehat{\theta}$ è il valore del parametro per il quale gli assegnati valori $x_1, \ldots, x_n$ coincidono con la **moda del campione**. In un certo senso scegliere $\widehat{\theta}$ vuol dire calibrare il parametro incognito in modo tale che quel che si misura (il campione dato) sia sempre il risultato più probabile (moda). Ovviamente $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ dipenderà dai valori di $x_1, \ldots, x_n$ anche se in genere, per semplificare la notazione, noi eviteremo di metterlo in evidenza. Potremo allora costruire la statistica $\widehat{\Theta} = \widehat{\theta}(X_1, \ldots, X_n)$ che, sulla base delle precedenti osservazioni, assumeremo proprio come stimatore di θ nella seguente definizione

Definizione 8.14. (Principio di massima verosimiglianza - *MV*) *Adotteremo come* ***stimatore di MV*** *la statistica* $\widehat{\Theta} = \widehat{\theta}(X_1, \ldots, X_n)$ *dove* $\widehat{\theta} = \widehat{\theta}(x_1, \ldots, x_n)$ *è il valore che rende massima la funzione di verosimiglianza* $L(\theta)$ *per un dato campione* $x_1, \ldots, x_n$

Il massimo della $L(\theta)$ è in genere determinato con i tradizionali metodi analitici (annullamento della derivata); siccome però $L(\theta)$ si presenta come un *prodotto* di n funzioni, e siccome non è sempre agevole derivare tali prodotti, spesso si preferisce – come faremo noi negli esempi che seguono – determinare il massimo della **verosimiglianza logaritmica** $\log L(\theta)$ che si presenta invece come una *somma* di funzioni di θ. Le due procedure sono equivalenti dato che la funzione $\log x$ è monotona crescente

Esamineremo ora alcuni risultati dell'applicazioni di questo principio osservando preliminarmente che in alcuni casi lo stimatore di MV coincide con quello che avevamo già individuato sulla base di altri principi. Questo però non accade sempre, anzi sarà opportuno osservare che, sebbene il Principio di MV selezioni buoni stimatori anche in assenza di altri suggerimenti disponibili, tuttavia **gli stimatori di *MV* non sono sempre stimatori corretti** secondo la Definizione 8.1. In molti casi, comunque, la correzione della distorsione dello stimatore di MV può essere ottenuta facilmente

Teorema 8.15. *Sulla base di un campione* $X_1, \ldots, X_n$ *di* n *misure, lo stimatore di MV del parametro* p *di una v-a binomiale* $X \sim \mathfrak{B}(m; p)$ *con* $m \geq 1$ *è*

$$\widehat{p} = \frac{1}{m\,n} \sum_{j=1}^{n} X_j = \frac{\overline{X}_n}{m} \tag{8.10}$$

Dimostrazione: Per una *v-a* binomiale $X \sim \mathfrak{B}(m; p)$ sappiamo che

$$\boldsymbol{P}\{X = k\} = \binom{m}{k} p^k (1-p)^{m-k} \qquad k = 0, 1, \ldots, m$$

per cui, detti k_j i valori delle *v-a* X_j del campione, la funzione di verosimiglianza e la verosimiglianza logaritmica saranno

$$\begin{aligned} L(p) &= \prod_{j=1}^{n} \binom{m}{k_j} p^{k_j} (1-p)^{m-k_j} \\ \log L(p) &= \sum_{j=1}^{n} \log \binom{m}{k_j} + \log p \sum_{j=1}^{n} k_j + \log(1-p) \sum_{j=1}^{n} (m - k_j) \end{aligned}$$

Siccome la derivata rispetto a p della verosimiglianza logaritmica è

$$\begin{aligned} \frac{d}{dp} \log L(p) &= \frac{1}{p} \sum_{j=1}^{n} k_j - \frac{1}{1-p} \sum_{j=1}^{n} (m - k_j) = \frac{1}{p} \sum_{j=1}^{n} k_j - \frac{1}{1-p} \left(nm - \sum_{j=1}^{n} k_j \right) \\ &= \frac{1}{p(1-p)} \sum_{j=1}^{n} k_j - \frac{nm}{1-p} \end{aligned}$$

e il massimo si ottiene imponendo che tale derivata sia nulla, lo stimatore di MV si ottiene risolvendo rispetto a p l'equazione

$$\frac{1}{p(1-p)} \sum_{j=1}^{n} k_j - \frac{nm}{1-p} = 0$$

ovvero, con qualche semplificazione,

$$\widehat{p}\,(k_1, \ldots, k_n) = \frac{1}{m\,n} \sum_{j=1}^{n} k_j$$

Il risultato (8.10) si ricava poi sostituendo ai valori k_j le rispettive *v-a* X_j del campione. Si noti che per $m = 1$ la X è una *v-a* di Bernoulli e il nostro problema si riconduce alla stima di una proporzione trattata nell'Esempio 5.4, e il risultato (5.8) coincide con la (8.10) per $m = 1$ □

Teorema 8.16. *Sulla base di un campione $X_1, \ldots, X_n$ di n misure, lo stimatore di MV del parametro λ di una v-a di Poisson $X \sim \mathfrak{P}(\lambda)$ è*

$$\widehat{\lambda} = \frac{1}{n} \sum_{j=1}^{n} X_j = \overline{X}_n \tag{8.11}$$

Dimostrazione: Per una *v-a* di Poisson $X \sim \mathfrak{P}(\lambda)$ sappiamo che

$$\boldsymbol{P}\{X = k\} = e^{-\lambda}\,\frac{\lambda^k}{k!} \qquad k = 0, 1, \ldots$$

per cui, detti k_j i valori delle *v-a* X_j del campione, la funzione di verosimiglianza e la verosimiglianza logaritmica saranno

$$\begin{aligned} L(\lambda) &= \prod_{j=1}^{n} e^{-\lambda}\,\frac{\lambda^{k_j}}{k_j!} \\ \log L(\lambda) &= -n\lambda + \log\lambda \sum_{j=1}^{n} k_j - \sum_{j=1}^{n} \log(k_j!) \end{aligned}$$

Annullando la derivata rispetto a λ della verosimiglianza logaritmica, il massimo si ottiene risolvendo l'equazione nell'incognita λ

$$\frac{d}{d\lambda} \log L(\lambda) = -n + \frac{1}{\lambda} \sum_{j=1}^{n} k_j = 0$$

ovvero

$$\widehat{\lambda}(k_1, \ldots, k_n) = \frac{1}{n} \sum_{j=1}^{n} k_j$$

e lo stimatore di MV (8.11), coincidente con la media aritmetica del campione, si ricava poi sostituendo ai valori k_j le rispettive *v-a* X_j del campione □

Teorema 8.17. *Sulla base di un campione $X_1, \ldots, X_n$ di n misure, gli stimatori di MV di μ e σ^2 di una v-a normale $X \sim \mathfrak{N}(\mu, \sigma^2)$ sono*

$$\widehat{\mu} = \frac{1}{n}\sum_{j=1}^{n} X_j = \overline{X}_n \qquad\qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2 = \widehat{S}_n^2 \tag{8.12}$$

Inoltre $\widehat{\sigma}^2$ è uno stimatore distorto, ma può essere corretto nella S_n^2 di (8.1)

Dimostrazione: La *fdp* una *v-a* normale $X \sim \mathfrak{N}(\mu, \sigma^2)$ è

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$$

per cui, detti x_j i valori delle *v-a* X_j del campione, la funzione di verosimiglianza e la verosimiglianza logaritmica saranno

$$\begin{aligned} L(\mu,\sigma) &= \prod_{j=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x_j-\mu)^2/2\sigma^2} \\ \log L(\mu,\sigma) &= -n\log(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{j=1}^{n}(x_j-\mu)^2 \end{aligned}$$

Siccome questa volta dobbiamo stimare due parametri, bisognerà risolvere le due equazioni ottenute annullando ambedue le derivate rispetto a μ e σ. Annullando la derivata della verosimiglianza logaritmica rispetto a μ si ha prima di tutto

$$\frac{d}{d\mu}\log L(\mu,\sigma) = \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j-\mu) = 0$$

da cui si ottiene per μ

$$\widehat{\mu}\,(x_1, \ldots, x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j \tag{8.13}$$

Annullando poi la derivata rispetto a σ

$$\frac{d}{d\sigma}\log L(\mu,\sigma) = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{n}(x_j-\mu)^2 = 0$$

e sostituendo la precedente soluzione $\widehat{\mu}$ al posto di μ, si ha poi per σ^2

$$\widehat{\sigma}^2(x_1, \ldots, x_n) = \frac{1}{n}\sum_{j=1}^{n}(x_j - \widehat{\mu})^2 \tag{8.14}$$

Gli stimatori di *MV* (8.11) si ricavano infine da (8.13) e (8.14) sostituendo ai valori x_j le rispettive *v-a* X_j del campione. Questi stimatori di *MV* coincidono rispettivamente con la media aritmetica $\overline{X}_n$ e la varianza campionaria $\widehat{S}_n^2$ e noi sappiamo dal Teorema 8.2 che il primo è corretto, ma il secondo è distorto, anche se la sua correzione S_n^2 è facilmente ottenibile con l'aggiunta di un fattore numerico □

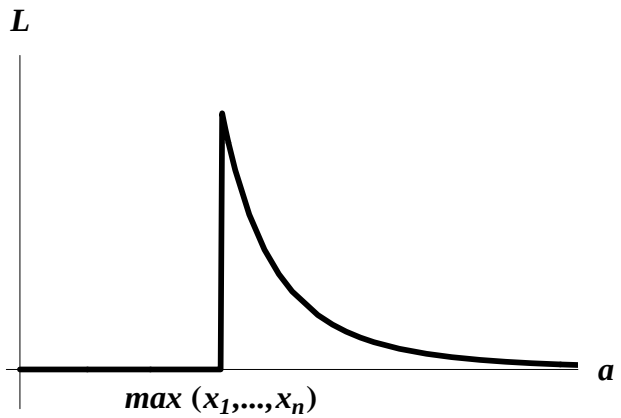


Figura 8.4: Funzione di verosimiglianza $L(a)$ del Teorema 8.18

Teorema 8.18. *Sulla base di un campione $X_1, \ldots, X_n$ di n misure, lo stimatore di MV del parametro a di una v-a uniforme $X \sim \mathfrak{U}(0, a)$ è*

$$\widehat{a} = \max\{X_1, \ldots, X_n\} \tag{8.15}$$

Inoltre $\widehat{a}$ è uno stimatore distorto, ma può essere corretto nella forma

$$\widehat{A} = \frac{n+1}{n} \max\{X_1, \ldots, X_n\} \tag{8.16}$$

Dimostrazione: Siccome la *fdp* di una *v-a* uniforme $X \sim \mathfrak{U}(0, a)$ è

$$f_X(x) = \begin{cases} {}^1\!/_a & \text{se } 0 \leq x \leq a, \\ 0 & \text{altrimenti.} \end{cases}$$

la funzione di verosimiglianza sarà diversa da zero solo se a risulta maggiore di tutte le x_j del campione, ovvero se $a > \max\{x_1, \ldots, x_n\}$, e avrà quindi la forma

$$L(a) = \begin{cases} {}^1\!/_{a^n} & \text{se } a > \max\{x_1, \ldots, x_n\} \\ 0 & \text{altrimenti} \end{cases}$$

riportata nella Figura 8.4. In questo modo si vede subito dal grafico che per un dato campione $x_1, \ldots, x_n$ la funzione di verosimiglianza $L(a)$ assume il valore massimo proprio in

$$\widehat{a}\,(x_1, \ldots, x_n) = \max(x_1, \ldots, x_n)$$

e lo stimatore di MV (8.15) si ricava sostituendo ai valori x_j le rispettive *v-a* X_j del campione. Come si potrà notare questo risultato è meno ovvio di quelli discussi in precedenza, anche se è abbastanza intuitivo che la stima dell'estremo superiore dell'intervallo $[0, a]$ sia data proprio dal più grande degli X_j. Si può inoltre dimostrare (ma noi trascureremo di farlo) che lo stimatore $\widehat{a}$ è distorto perché

$$\boldsymbol{E}\,[\widehat{a}\,] = \frac{n}{n+1}\, a < a$$

Questa distorsione può comunque essere facilmente eliminata introducendo lo stimatore corretto (8.16), ma come si capisce facilmente questa correzione diviene del tutto irrilevante per grandi valori di n □

Capitolo 9

Statistica inferenziale: Test di ipotesi

9.1 Ipotesi, decisioni ed errori

L'esito della procedura di stima di un parametro è un numero, o un intervallo di numeri; viceversa l'esito di un test statistico è una **decisione fra ipotesi alternative**. Nell'esempio che segue saranno presentati, preliminarmente e in maniera solo intuitiva, alcuni problemi pratici che possono richiedere l'uso di questo tipo di test

Esempio 9.1. *In un campione di* 11 712 *bambini nati in un paese tra il 1968 e il 1973 ci sono state* 5 934 *nascite maschili; la proporzione empirica di maschi è quindi*

$$\overline{p} = \frac{5\,934}{11\,712} = 0.507 \tag{9.1}$$

Possiamo affermare che c'è stata una prevalenza di nascite maschili? O dobbiamo solo attribuire al caso il fatto che $\overline{p}$ *non sia esattamente* $^1/_2$ *come sarebbe naturale attendersi in assenza di altre informazioni? Possiamo trovare una procedura affidabile che, sulla base dei dati disponibili, ci consenta di accettare una affermazione oppure l'altra, e di stimare la probabilità di sbagliare?*

Analogamente, riprendendo la discussione dell'Esempio 8.8, possiamo ritenere che gli esiti delle nascite successive nelle famigli di 12 *figli siano indipendenti fra loro? O dobbiamo supporre che ci siano famiglie con la tendenza a generare figli maschi, e famiglie con la tendenza a generare figlie femmine? Detto in altri termini: con quale criterio e con quale affidabilità possiamo decidere di accettare una delle due ipotesi alternative che abbiamo appena formulato?*

Una ditta farmaceutica, infine, ha prodotto un nuovo farmaco per la cura di una determinata malattia: con quale procedura e con quale affidabilità possiamo pervenire a stabilire se il farmaco è realmente efficace, o se il suo uso non produce i risultati desiderati, o se infine non è addirittura dannoso? In che maniera dovremmo rilevare i dati empirici, e come possiamo usarli per giungere alla decisione richiesta?

In tutti gli esempi precedenti lo scopo dell'analisi statistica è quello di pervenire, con un preciso livello di **affidabilità**, alla accettazione di una fra due possibili ipotesi alternative $\mathcal{H}_0$ e $\mathcal{H}_1$. Il problema della affidabilità qui è essenziale, perché bisogna subito cominciare a distinguere fra il fatto che il test mi suggerisca di accettare una determinata ipotesi, e il fatto che tale ipotesi sia quella giusta nella realtà. Infatti, anche se condotto in maniera corretta, un test non garantisce affatto che la decisione indicata sia quella esatta: trattandosi di una procedura statistica, essa potrebbe suggerire di accettare l'ipotesi sbagliata. In maniera un po' informale possiamo riassumere queste prime osservazioni nel modo seguente

> *Formulata un'ipotesi $\mathcal{H}_0$ e la sua alternativa $\mathcal{H}_1$, un* ***test di ipotesi*** *è una precisa procedura statistica basata su un campione di misure $X_1, \ldots, X_n$ il cui esito* ***indica se accettare $\mathcal{H}_0$ o la sua alternativa $\mathcal{H}_1$****. È necessario però associare anche una valutazione quantitativa della* ***probabilità dell'errore*** *che si rischia di commettere seguendo le indicazioni del test*

Inoltre bisogna ricordare che l'esito di un test dipende anche dalla sua particolare configurazione, cioè dai particolari valori che vengono inizialmente, e arbitrariamente, scelti per alcuni importanti parametri (in particolare, come vedremo fra poco, il *livello del test*) quando il test viene progettato. Naturalmente la scelta di questi valori deve essere effettuata in maniera cauta e accurate, ma è importante sottolineare che – per un dato insieme di risultati sperimentali riassunto in un campione – l'esito del test può essere capovolto modificando tale scelta. D'altra parte è anche vero che, per un dato test in una ben definita configurazione, l'esito potrebbe essere modificato semplicemente ripetendo le misure: due campioni differenti relativi allo stesso esperimento possono condurre a decisioni differenti. Tutto questo rende particolarmente delicato l'uso di questi strumenti statistici e l'interpretazione delle loro indicazioni

Nel seguito avremo a che fare innanzitutto con ipotesi che riguardano i valori di qualche parametro θ della distribuzione delle *v-a* ritenute rilevanti per il nostro esperimento. Sarà allora in generale possibile dividere l'insieme di tutti i valori di θ in due opportuni sottoinsiemi disgiunti Θ_0 e Θ_1, in modo da caratterizzare formalmente le due **ipotesi alternative** come

$$\boxed{\mathcal{H}_0 \ : \ \theta \in \Theta_0 \qquad\qquad \mathcal{H}_1 \ : \ \theta \in \Theta_1} \tag{9.2}$$

Così ad esempio, se θ può assumere arbitrari valori reali, le due ipotesi potrebbero essere del tipo

$$\mathcal{H}_0 \ : \ \theta \leq 0 \qquad\qquad \mathcal{H}_1 \ : \ \theta > 0$$

dove ovviamente abbiamo preso $\Theta_0 = (-\infty, 0]$ e $\Theta_1 = (0, +\infty)$. Nella prassi statistica $\mathcal{H}_0$ viene chiamata **ipotesi nulla**, e $\mathcal{H}_1$ viene detta **ipotesi alternativa**. In generale il ruolo di queste due ipotesi non è simmetrico: tipicamente si sceglie come

ipotesi nulla l'ipotesi più conservativa e prudente, ricordando però che lo sperimentatore è normalmente interessato a verificare se l'ipotesi $\mathcal{H}_0$ può essere rifiutata, accettando invece l'ipotesi $\mathcal{H}_1$ che dovrebbe contenere qualche elemento di novità (una *scoperta*). Così, se si sta sperimentando un nuovo farmaco, l'ipotesi $\mathcal{H}_0$ è in genere associata alla conclusione: *il farmaco è inefficace*; mentre lo sperimentatore è interessato a verificare piuttosto se può essere sostenuta l'ipotesi alternativa $\mathcal{H}_1$ che conduce alla conseguenza: *il farmaco è efficace*

Naturalmente delle due ipotesi solo una è vera nei fatti, ma noi non possiamo scoprire *con certezza* quale essa sia solo sulla base di un test statistico. Un test serve a dare un'indicazione affidabile per questa decisione, ma non dobbiamo confondere – come già sottolineato – l'esito di un test statistico a favore di un'ipotesi con un certificato di verità per quella ipotesi. L'esecuzione del test può infatti condurre a decisioni errate, e gli **errori** possibili (tutti caratterizzati da una *divergenza fra decisione e realtà*) sono di due tipi principali come dichiarato nella seguente definizione

Definizione 9.2. *Chiameremo rispettivamente*

- ***errore di prima specie*** *quello che si commette se il test mi induce a rifiutare $\mathcal{H}_0$, quando $\mathcal{H}_0$ è vera*
- ***errore di seconda specie*** *quello che si commette se il test mi induce ad accettare $\mathcal{H}_0$, quando $\mathcal{H}_0$ è falsa*

Fermi restando i contenuti, questa definizione potrebbe essere equivalentemente riformulata sostituendo a frasi come "rifiutare $\mathcal{H}_0$", oppure "$\mathcal{H}_0$ è vera", rispettivamente le espressioni equivalenti "accettare $\mathcal{H}_1$", oppure "$\mathcal{H}_1$ è falsa." Ciononostante i due tipi di errore non sono considerati in genere sullo stesso piano per due ragioni principali:

1. Gli errori di prima specie, che consistono nel fare affermazioni false e imprudenti (rifiutare $\mathcal{H}_0$, cioè accettare $\mathcal{H}_1$, quando ciò non è corretto), sono considerati **più gravi** di quelli di seconda, che consistono invece nel perdere eventualmente l'occasione di mettere in evidenza qualcosa di nuovo (accettare $\mathcal{H}_0$, cioè rifiutare $\mathcal{H}_1$ che invece è vera). Nel caso della sperimentazione di un farmaco, ad esempio, si giudica più grave mettere in circolazione un farmaco inutile (o addirittura dannoso), che perdere l'occasione di produrre un farmaco efficace

2. Gli errori di prima specie si commettono sotto l'ipotesi che $\mathcal{H}_0$ sia vera: siccome in generale $\mathcal{H}_0$ è un'ipotesi **più precisa** dell'alternativa $\mathcal{H}_1$ (che spesso è definita solo dall'essere il contrario di $\mathcal{H}_0$), supporre che $\mathcal{H}_0$ sia vera permette nella maggior parte dei casi di valutare meglio la probabilità dell'errore di prima specie. Viceversa, siccome l'errore di seconda specie si commette sotto l'ipotesi che sia vera $\mathcal{H}_1$, è in generale più difficile poter calcolare la probabilità di questo secondo tipo di errore

Sarà importante dunque essere in grado di controllare **la probabilità di commettere questi due errori**; ma prima di procedere sarà bene sottolineare che in effetti queste due probabilità sono **in competizione fra loro**. Con questo intendiamo dire che in generale non possiamo diminuirne una senza aumentare l'altra, e che quindi sarà illusorio pensare di renderle molto piccole entrambe. Questo comportamento – che sarà reso chiaro dalla discussione successiva – può essere compreso intuitivamente se si osserva che un test che renda meno probabile l'errore di prima specie, deve rendere più difficile rifiutare $\mathcal{H}_0$ e più facile accettarla; ma questo vuol dire contemporaneamente rendere più probabile l'errore di seconda specie.

Definizione 9.3. *Diremo che un evento D – definito sulla base di un dato campione statistico $X_1, \ldots, X_n$ – è un* ***evento critico*** *quando il suo verificarsi conduce al rifiuto dell'ipotesi nulla $\mathcal{H}_0$. Quando con i dati del campione l'evento critico si verifica, diremo anche che i dati sono in* ***regione critica****. Progettare il test vuol dire definire opportunamente l'evento critico D, ed eseguire il test vuol dire verificare se i dati sono o meno in regione critica. Chiameremo poi*

- ***livello*** *del test: la quantità*

$$\alpha = \sup_\theta \boldsymbol{P}_\theta\{D\} \qquad \text{con} \quad \theta \in \Theta_0 \qquad \text{cioè se è vera } \mathcal{H}_0 \tag{9.3}$$

- ***funzione potenza*** *del test: la funzione*

$$\pi(\theta) = \boldsymbol{P}_\theta\{D\} \qquad \text{con} \quad \theta \in \Theta_1 \qquad \text{cioè se è vera } \mathcal{H}_1 \tag{9.4}$$

- ***significatività*** α_s *del test: il più piccolo valore del livello α che colloca un campione dato in regione critica, cioè che conduce al rifiuto di $\mathcal{H}_0$*

In questa definizione $\boldsymbol{P}_\theta$ indica come al solito la probabilità calcolata sotto l'ipotesi che il parametro sconosciuto abbia proprio il valore θ, anche se ometteremo l'indice θ dovunque possibile. Conseguentemente $\boldsymbol{P}_\theta\{D\}$ è la probabilità di rifiutare $\mathcal{H}_0$, e se prendiamo $\theta \in \Theta_0$ come in (9.3) – cioè supponiamo che $\mathcal{H}_0$ sia vera – si vede che il **livello** α di un test è il massimo delle **probabilità di commettere errori di prima specie**. Allo stesso modo la **potenza** del test $\pi(\theta)$ è la probabilità di rifiutare $\mathcal{H}_0$ al variare di $\theta \in \Theta_1$, cioè quando $\mathcal{H}_0$ è falsa: in pratica si tratta della probabilità di mettere in evidenza la correttezza di $\mathcal{H}_1$ quando questa è vera, e quindi la **probabilità di errori di seconda specie** in realtà è $1 - \pi(\theta)$ al variare di $\theta \in \Theta_1$. Il **valore del livello** è tipicamente una scelta operata inizialmente dello sperimentatore che decide il rischio di errore di prima specie che vuole correre: i valori più usati sono $\alpha = 0.05$ e $\alpha = 0.01$, ma anche $\alpha = 0.10$. Siccome però resta aperto il problema di valutare e rendere contemporaneamente grande anche la funzione potenza $\pi(\theta)$, il livello α non può essere scelto troppo piccolo e in generale si evita di andare sotto il valore 0.01.

In genere, per un dato campione $X_1, \ldots, X_n$, un evento critico ha una forma del tipo $D = \{|U| \geq \delta\}$ con $\delta > 0$, dove $U = u(X_1, \ldots, X_n)$ è una opportuna

statistica la cui legge è nota quando l'ipotesi $\mathcal{H}_0$ è vera, cioè se $\theta \in \Theta_0$. L'estensione di D (cioè il valore di δ) è conseguenza della scelta del livello, e varia al variare di α: in particolare all'aumentare di α anche D si allarga (cioè δ diminuisce) in concomitanza con l'aumento delle probabilità $\boldsymbol{P}_\theta\{D\}$ di rifiutare $\mathcal{H}_0$. Pertanto, se da un lato con $\alpha \simeq 0$ qualunque campione si troverà fuori regione critica (cioè quasi certamente accetteremo $\mathcal{H}_0$), dall'altro aumentando progressivamente α anche l'evento critico D si allargherà fino a inglobare il campione dato portando al rifiuto di $\mathcal{H}_0$: il primo valore α_s per il quale ciò avviene è la **significatività** del test secondo la Definizione 9.3. In pratica, per un fissato campione, α_s divide i valori di α in due regioni: per $\alpha < \alpha_s$ il campione non è in regione critica; invece per $\alpha \geq \alpha_s$ il campione è in regione critica. In genere **un test è considerato buono quando α_s è piccola**: infatti se $\alpha_s < 0.01$, l'ipotesi $\mathcal{H}_0$ risulterebbe rifiutata con valori anche piccoli del livello α, e quindi con piccoli rischi di errori di prima specie. In ogni caso la significatività esprime piuttosto un giudizio sulla **adeguatezza del campione** a dare indicazioni affidabili sulle ipotesi proposte

9.1.1 Discussione dettagliata di un test

Per rendere più concreta questa prima introduzione esamineremo ora con qualche dettaglio un tipico caso tratto dall'Esempio 9.1: **decidere**, sulla base dei dati sperimentali disponibili, **se la probabilità p che in un parto si produca una nascita *maschile* possa o meno essere ritenuta uguale a $^1/_2$**. Siccome il parametro p assume i valori $0 \leq p \leq 1$, possiamo dire che le nostre due ipotesi alternative sono

$$\mathcal{H}_0 \,:\, p \,=\, {}^1/_2 \qquad\qquad \mathcal{H}_1 \,:\, p \,\neq\, {}^1/_2 \qquad\qquad 0 \leq p \leq 1$$

ovvero che i due sottoinsiemi che le definiscono sono

$$\Theta_0 = \left\{ {}^1/_2 \right\} \qquad\qquad \Theta_1 = \left[0\,,\, {}^1/_2\right) \cup \left({}^1/_2\,,1\right]$$

La procedura più intuitiva per progettare il test è ovviamente quella di confrontare il valore empirico (9.1) con il valore ipotetico $^1/_2$, esaminando il valore assunto da $|\overline{p} - {}^1/_2|$: se questa differenza sarà giudicata troppo grande rifiuteremo l'ipotesi $\mathcal{H}_0$. Per rendere precisa e quantitativa questa procedura dovremo però trovare prima di tutto un modo ragionevole per fissare una soglia $\delta > 0$ per rifiutare l'ipotesi $\mathcal{H}_0$ quando $|\overline{p} - {}^1/_2| > \delta$; e poi bisognerà essere in grado di stimare le probabilità di commettere degli errori eseguendo un test così progettato

Possiamo come al solito formalizzare meglio il problema definendo una *v-a* di Bernoulli $X \sim \mathfrak{B}(1;p)$ che prende valori 1 e 0 con probabilità p e $1-p$ secondo che in un parto si verifichi rispettivamente una nascita maschile o una femminile. Il nostro campione è composto allora di $n = 11\,712$ *v-a* di Bernoulli indipendenti $X_j \sim \mathfrak{B}(1;p)$ con $j = 1, \ldots, n$ che rappresentano gli esiti dei parti, e dai dati empirici sappiamo che fra questi 5 934 sono maschili. Se ora definiamo le *v-a*

$$Y_n = \sum_{j=1}^{n} X_j\,, \qquad\qquad \overline{X}_n = \frac{Y_n}{n} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

avremo dal Teorema 3.16 che $Y_n \sim \mathfrak{B}(n;p)$, e quindi da (4.24), (4.8) e (4.19)

$$\begin{aligned} \boldsymbol{E}[Y_n] &= np & \boldsymbol{V}[Y_n] &= np(1-p) \\ \boldsymbol{E}\left[\overline{X}_n\right] &= p & \boldsymbol{V}\left[\overline{X}_n\right] &= \frac{p(1-p)}{n} \end{aligned}$$

D'altra parte, essendo $n = 11\,712$ molto grande, il Teorema 5.5 (*TLC*) ci dice che potremo sicuramente adottare l'approssimazione normale

$$\sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \sim \mathfrak{N}(0,1)$$

Pertanto, comunque preso $\delta > 0$, e ricordando che con il simbolo Φ indichiamo la *FDC* della normale standard (3.18), potremo ricavare la seguente formula approssimata che useremo nel seguito della discussione

$$\begin{aligned} &\boldsymbol{P}\left\{\left|\overline{X}_n - {}^1\!/_2\right| \geq \delta\right\} \\ &\qquad = \; 1 - \boldsymbol{P}\left\{\left|\overline{X}_n - {}^1\!/_2\right| \leq \delta\right\} \;=\; 1 - \boldsymbol{P}\left\{{}^1\!/_2 - \delta \leq \overline{X}_n \leq {}^1\!/_2 + \delta\right\} \\ &\qquad = \; 1 - \boldsymbol{P}\left\{\sqrt{n}\,\frac{{}^1\!/_2 - \delta - p}{\sqrt{p(1-p)}} \leq \sqrt{n}\,\frac{\overline{X}_n - p}{\sqrt{p(1-p)}} \leq \sqrt{n}\,\frac{{}^1\!/_2 + \delta - p}{\sqrt{p(1-p)}}\right\} \\ &\qquad \simeq \; 1 + \Phi\left(\sqrt{n}\,\frac{1-2\delta-2p}{2\sqrt{p(1-p)}}\right) - \Phi\left(\sqrt{n}\,\frac{1+2\delta-2p}{2\sqrt{p(1-p)}}\right) \end{aligned} \tag{9.5}$$

In particolare, quando si suppone che l'ipotesi $\mathcal{H}_0$ sia vera, cioè che $p = {}^1\!/_2$, e ricordando anche le proprietà di simmetria (3.19) di Φ, la (9.5) si semplifica in

$$\begin{aligned} \boldsymbol{P}\left\{\left|\overline{X}_n - {}^1\!/_2\right| \geq \delta\right\} &\simeq 1 + \Phi(-2\delta\sqrt{n}) - \Phi(2\delta\sqrt{n}) \\ &= 2\left[1 - \Phi(2\delta\sqrt{n})\right] \qquad \text{se } \mathcal{H}_0 \text{ è vera} \end{aligned} \tag{9.6}$$

Come primo tentativo possiamo allora provare a fissare arbitrariamente una soglia $\delta = 0.01$: in questo modo l'evento critico sarà

$$D = \left\{\left|\overline{X}_n - {}^1\!/_2\right| \geq 0.01\right\}$$

e i dati del campione non saranno in regione critica perché si ha $|\overline{p} - {}^1\!/_2| = |0.507 - 0.500| = 0.007 < 0.010$. L'esito del test in questa configurazione suggerisce dunque di accettare l'ipotesi $\mathcal{H}_0$. Per calcolare poi il livello α del test così progettato, basterè osservare che, supponendo $\mathcal{H}_0$ vera, ponendo $n = 11\,712$ e $\delta = 0.01$ in (9.6) e facendo uso delle Tavole E.1, risulta

$$\alpha = \boldsymbol{P}\left\{\left|\overline{X}_n - {}^1\!/_2\right| \geq 0.01\right\} \simeq 2\left[1 - \Phi(2.164)\right] = 0.030$$

In conclusione, con il campione dato, e con $\delta = 0.01$ si ottiene un test di livello $\alpha = 0.03$ che ci induce ad accettare $\mathcal{H}_0$ con una probabilità di errore di prima specie del 3 %

Tipicamente, però, un test non si progetta fissando arbitrariamente la soglia δ, ma scegliendo invece *a priori* il **livello** α del test (cioè il rischio di errore di prima specie che si vuol correre): da questo si deduce poi il valore della soglia δ dell'evento critico D. A questo scopo si osservi che, supponendo vera $\mathcal{H}_0$, da (9.6) si ha

$$\alpha = \boldsymbol{P}\big\{\big|\overline{X}_n - {}^1\!/_2\big| \geq \delta\big\} = 2\left[1 - \Phi(2\delta\sqrt{n})\right] \qquad \text{se } \mathcal{H}_0 \text{ è vera}$$

ovvero

$$\Phi(2\delta\sqrt{n}) = 1 - {}^{\alpha}\!/_2$$

e quindi ricordando la definizione (3.24) di quantile

$$2\delta\sqrt{n} = \varphi_{1-\frac{\alpha}{2}}$$

dove come al solito φ_α indica il quantile di ordine α della normale standard. Ne consegue che per il nostro esempio, scegliendo a priori un livello α, l'evento critico del test avrà la forma

$$D = \left\{\big|\overline{X}_n - {}^1\!/_2\big| \geq \frac{\varphi_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right\}$$

Ad esempio, supponendo di voler eseguire un test di livello $\alpha = 0.05$, da $n = 11\,712$ e dalle Tavole E.1 si ottiene

$$\delta = \frac{\varphi_{1-\frac{\alpha}{2}}}{2\sqrt{n}} = \frac{\varphi_{0.975}}{2\sqrt{11\,712}} \simeq 0.009$$

per cui l'evento critico di livello $\alpha = 0.05$ è

$$D = \big\{\big|\overline{X}_n - {}^1\!/_2\big| \geq 0.009\big\} \tag{9.7}$$

Siccome per il nostro campione abbiamo $|\overline{p} - {}^1\!/_2| = |0.507 - 0.500| = 0.007 < 0.009$, anche l'esito del test di livello $\alpha = 0.05$ suggerisce di accettare l'ipotesi $\mathcal{H}_0$

Notiamo ora che scegliendo un α più grande (0.05 invece di 0.03) la soglia δ dell'evento critico D diminuisce (passa da 0.010 a 0.009), quindi la regione critica si allarga e diventa più probabile rifiutare $\mathcal{H}_0$. È evidente allora che, mantenendo sempre gli stessi dati empirici ($\overline{p} = 0.507$ e $n = 11\,712$) e aumentando progressivamente il livello α, si arriverà ad un valore α_s (detto, come sappiamo, **significatività**) tale che l'ipotesi $\mathcal{H}_0$ sia rifiutata. Dato che per noi $|\overline{p} - {}^1\!/_2| = 0.007$, è chiaro che α_s sarà il valore di α che produce per la soglia δ proprio il valore 0.007, e quindi, sempre da (9.6) e dalle Tavole E.1, la significatività del nostro test è

$$\alpha_s = \boldsymbol{P}\big\{\big|\overline{X}_n - {}^1\!/_2\big| \geq 0.007\big\} \simeq 2\left[1 - \Phi(2 \times 0.007 \times \sqrt{11\,712})\right] \simeq 0.130$$

Questa significatività non è considerata molto buona, perché l'ipotesi $\mathcal{H}_0$ potrà essere rifiutata solo da un test con livello maggiore di 0.13, ovvero con probabilità di errore del primo tipo ($\geq 13\,\%$) piuttosto alta. L'esempio però mostra anche che il valore

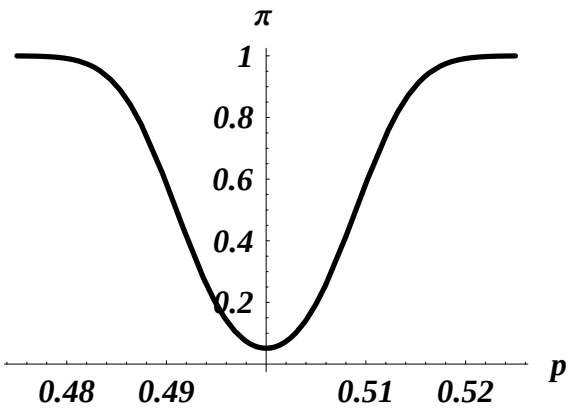


Figura 9.1: Funzione potenza $\pi(p)$ per il test di livello $\alpha = 0.05$ della Sezione 9.1.1.

della significatività dipende dal campione: se infatti la frequenza empirica $\overline{p} = 0.507$ fosse associata ad un campione più numeroso, ad esempio $n = 40\,000$, avremmo

$$\alpha_s = \boldsymbol{P}\left\{\left|\overline{X}_n - {}^1\!/_2\right| \geq 0.007\right\} \simeq 2\left[1 - \Phi(2 \times 0.007 \times \sqrt{40\,000})\right] \simeq 0.005$$

e la significatività (0.5 %) sarebbe ora molto migliorata

Infine possiamo anche calcolare la **funzione potenza** del test di livello $\alpha = 0.05$ con evento critico (9.7). Da (9.5), supponendo vera l'ipotesi $\mathcal{H}_1$ (cioè $p \neq {}^1\!/_2$) e con $\delta = 0.009$, avremo allora la funzione potenza

$$\begin{aligned}\pi(p) &= 1 + \Phi\left(\sqrt{n}\,\frac{1 - 2\delta - 2p}{2\sqrt{p\,(1-p)}}\right) - \Phi\left(\sqrt{n}\,\frac{1 + 2\delta - 2p}{2\sqrt{p\,(1-p)}}\right) \\ &= 1 + \Phi\left(\frac{108.2\,(0.491 - p)}{\sqrt{p\,(1-p)}}\right) - \Phi\left(\frac{108.2\,(0.509 - p)}{\sqrt{p\,(1-p)}}\right)\end{aligned}$$

Il grafico di questa funzione è riportato nella Figura 9.1 Da questa si vede che la potenza del test di livello 0.05 è buona quando p si discosta da ${}^1\!/_2$ di più di 0.02: in questo caso infatti il test vede che $p \neq {}^1\!/_2$ praticamente con probabilità uguale a 1. Viceversa se la differenza fra p e ${}^1\!/_2$ è minore di 0.02 la potenza diminuisce molto, pur mantenendosi sempre superiore ad un minimo che vale 0.05. In pratica quando p differisce molto poco da ${}^1\!/_2$, il test di livello $\alpha = 0.05$ vede la differenza solo con una probabilità poco più grande del 5 %. La funzione potenza, peraltro, ci permette di fare affermazioni anche nei casi intermedi: ad esempio, se p differisce da ${}^1\!/_2$ di 0.1 (cioè se nella realtà avessimo $p = 0.51$ oppure $p = 0.49$) la probabilità che il nostro test di livello $\alpha = 0.5$ lo rilevi è il 59 %, e corrispondentemente la probabilità di errori del secondo tipo (accettare $\mathcal{H}_0$ quando essa è falsa) è del 41 %

9.2 Test sulla media

Il primo tipo di test che tratteremo riguarda il valore μ – sconosciuto – dell'attesa di una *v-a* X della quale abbiamo un campione $X_1, \ldots, X_n$ di misure: tipicamente (un po' come nella Sezione 9.1.1) si chiede di conoscere se μ può essere considerato uguale ad un ben determinato valore μ_0, oppure no. In questo caso le ipotesi da valutare assumono la forma

$$\mathcal{H}_0 \,:\, \mu = \mu_0 \qquad\qquad \mathcal{H}_1 \,:\, \mu \neq \mu_0 \tag{9.8}$$

e si parla di **test bilaterale** per mettere in evidenza il fatto che nell'ipotesi alternativa μ può essere sia più grande che più piccola di μ_0. A volte invece potremmo essere interessati a mettere in evidenza solo che μ è più grande (o più piccolo) di qualche μ_0: in questo caso le ipotesi sono del tipo

$$\mathcal{H}_0 \,:\, \mu \leq \mu_0 \qquad\qquad \mathcal{H}_1 \,:\, \mu > \mu_0 \tag{9.9}$$

$$\mathcal{H}_0 \,:\, \mu \geq \mu_0 \qquad\qquad \mathcal{H}_1 \,:\, \mu < \mu_0 \tag{9.10}$$

e si parla rispettivamente di **test unilaterale destro e sinistro**. In tutte queste formulazioni si tenga presente che l'interesse dello sperimentatore risiede in generale nella possibilità di mettere in evidenza l'accettabilità di un'ipotesi alternativa $\mathcal{H}_1$ che giustifichi qualche *scoperta*: questa osservazione sarà utile per capire come progettare gli eventi critici

Esaminiamo innanzitutto il caso di un **test bilaterale** (9.8): per determinare la regione critica dobbiamo innanzitutto individuare un opportuno stimatore dell'attesa μ, e il Teorema 8.2 ci suggerisce subito la media aritmetica del campione $\overline{X}_n$ (5.9). Il test consisterà poi nell'esaminare se il valore osservato di $\overline{X}_n$ può significativamente essere considerato diverso da valore μ_0: pertanto l'evento critico prenderà la forma

$$\boxed{D = \{|\overline{X}_n - \mu_0| > \delta\}}$$

nel senso che se la differenza $|\overline{X}_n - \mu_0|$ supera la soglia $\delta > 0$, allora rifiuteremo l'ipotesi nulla $\mathcal{H}_0$ e accetteremo l'ipotesi alternativa $\mathcal{H}_1$. Per determinare poi la soglia δ a partire dal livello α del test, si impone, in base alla Definizione 9.3, che supponendo vera $\mathcal{H}_0 : \mu = \mu_0$ sia verificata la relazione

$$\boxed{\boldsymbol{P}_{\mu_0}\left\{|\overline{X}_n - \mu_0| > \delta\right\} = \alpha} \tag{9.11}$$

Questa equazione permetterà di determinare il valore della soglia δ per un test bilaterale di livello α prefissato

Nel caso di un **test unilaterale destro** (9.9), invece, la forma dell'ipotesi $\mathcal{H}_1$ suggerisce che l'evento critico sia

$$\boxed{D = \{\overline{X}_n - \mu_0 > \delta\}}$$

nel senso che ora la differenza $\overline{X}_n - \mu_0$ è presa senza valore assoluto perché siamo interessati a mettere in evidenza il fatto che μ sia *più grande* di (e non solo diverso da) μ_0: se tale differenza supera la soglia $\delta > 0$, allora rifiuteremo l'ipotesi nulla $\mathcal{H}_0$ e accetteremo l'ipotesi alternativa $\mathcal{H}_1$. Per determinare ora la soglia δ si fissa *a priori* il livello α del test e, sempre per la Definizione 9.3, supponendo vera $\mathcal{H}_0 : \mu \leq \mu_0$, si impone che sia

$$\sup_{\mu \leq \mu_0} \boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu_0 > \delta \right\} = \alpha \tag{9.12}$$

Per semplificare questa espressione si osservi che, se $\mathcal{H}_0$ è vera, avremo $\mu_0 - \mu \geq 0$ per cui

$$\begin{aligned} \boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu_0 > \delta \right\} &= \boldsymbol{P}_\mu \left\{ (\overline{X}_n - \mu) + (\mu - \mu_0) > \delta \right\} \\ &= \boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu > \delta + (\mu_0 - \mu) \right\} \leq \boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu > \delta \right\} \end{aligned}$$

Siccome però – quando il valore dell'attesa di X è proprio μ – sappiamo da (8.2) che anche l'attesa di $\overline{X}_n$ è μ, la *v-a* $\overline{X}_n - \mu$ risulta centrata (vedi Teorema 4.11) e quindi la probabilità $\boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu > \delta \right\}$ all'ultimo membro avrà sempre lo stesso valore al variare di μ. In particolare potremo porre $\mu = \mu_0$ all'ultimo membro ottenendo

$$\boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu_0 > \delta \right\} \leq \boldsymbol{P}_{\mu_0} \left\{ \overline{X}_n - \mu_0 > \delta \right\} \qquad \mu \leq \mu_0$$

che diventa banalmente un'uguaglianza esatta se $\mu = \mu_0$. In conclusione avremo

$$\sup_{\mu \leq \mu_0} \boldsymbol{P}_\mu \left\{ \overline{X}_n - \mu_0 > \delta \right\} = \boldsymbol{P}_{\mu_0} \left\{ \overline{X}_n - \mu_0 > \delta \right\}$$

e quindi la relazione (9.12) fra il livello α e la soglia δ potrà essere riscritta nella forma più agevole

$$\boxed{\boldsymbol{P}_{\mu_0} \left\{ \overline{X}_n - \mu_0 > \delta \right\} = \alpha} \tag{9.13}$$

In maniera analoga per un **test unilaterale sinistro** (9.10) l'evento critico sarà

$$\boxed{D = \{ \overline{X}_n - \mu_0 < -\delta \}}$$

e la relazione fra soglia δ e livello α sarà

$$\boxed{\boldsymbol{P}_{\mu_0} \left\{ \overline{X}_n - \mu_0 < -\delta \right\} = \alpha} \tag{9.14}$$

A questo punto, per poter procedere ulteriormente, è necessario avere anche delle informazioni sulla legge di $\overline{X}_n$, partendo da alcune considerazioni sul campione $X_1, \ldots, X_n$. Se potremo supporre che le X_k siano (almeno approssimativamente) *v-a* normali $\mathfrak{N}(\mu, \sigma^2)$, potremo usare direttamente i risultati del Teorema 3.23. Se invece non abbiamo informazioni precise sulla legge delle X_k, il Teorema 5.5 (*TLC*) ci garantisce comunque che, se n è abbastanza grande, $\overline{X}_n$ è approssimativamente Gaussiana $\mathfrak{N}(\mu, \sigma^2)$. Si noti però che in generale si ha una conoscenza limitata della legge (esatta o approssimata) $\mathfrak{N}(\mu, \sigma^2)$ di $\overline{X}_n$: non solo infatti non ci è noto il valore di μ, ma potrebbe essere sconosciuto anche il valore di σ^2. Per questo motivo dovremo ora distinguere due tipi di test

9.2.1 Test di Gauss

Supporremo ora che – o sulla base del Teorema 3.23 se si sa che il campione $X_1, \ldots, X_n$ segue la legge $\mathfrak{N}(\mu, \sigma^2)$, o almeno approssimativamente per effetto del *TLC* Teorema 5.5 – la media aritmetica del campione $\overline{X}_n$ sia normale $\mathfrak{N}(\mu, {}^{\sigma^2}/_n)$ con μ sconosciuta, ma con ***la varianza*** σ^2 ***nota***. In tal caso è facile riconoscere che la statistica

$$\boxed{U_0 = \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{\sigma} \sim \mathfrak{N}(0,1) \qquad \text{se } \mu = \mu_0} \tag{9.15}$$

è normale standard: una relazione che useremo nel seguito per applicare le formule (9.11), (9.13) e (9.14)

Test bilaterale

L'evento critico di un test bilaterale ha ora la forma

$$D = \{|\overline{X}_n - \mu_0| > \delta\} = \left\{|U_0| > \frac{\delta\sqrt{n}}{\sigma}\right\}$$

e la (9.11) si scriverà come

$$\boldsymbol{P}_{\mu_0}\left\{|\overline{X}_n - \mu_0| > \delta\right\} = \boldsymbol{P}_{\mu_0}\left\{|U_0| > \frac{\delta\sqrt{n}}{\sigma}\right\} = \boldsymbol{P}\left\{|\mathfrak{N}(0,1)| > \frac{\delta\sqrt{n}}{\sigma}\right\} = \alpha$$

Ricordando allora il risultato (3.26) si ha facilmente

$$\frac{\delta\sqrt{n}}{\sigma} = \varphi_{1-\frac{\alpha}{2}}$$

per cui in definitiva l'**evento critico** diviene semplicemente

$$\boxed{D = \left\{|U_0| > \varphi_{1-\frac{\alpha}{2}}\right\}} \tag{9.16}$$

In conclusione un test bilaterale di livello α delle ipotesi (9.8) si esegue con la seguente **procedura**:

> *Si calcola il valore empirico della statistica U_0 (9.15) e si confronta il suo valore assoluto con il quantile $\varphi_{1-\frac{\alpha}{2}}$ della normale standard ricavato dalle Tavole E.1: se risulta $|U_0| > \varphi_{1-\frac{\alpha}{2}}$, allora si rifiuta l'ipotesi $\mathcal{H}_0$ e si accetta $\mathcal{H}_1$; se invece risulta $|U_0| \leq \varphi_{1-\frac{\alpha}{2}}$ si accetta l'ipotesi $\mathcal{H}_0$*

Siccome poi la significatività α_s è il più piccolo valore del livello α per il quale i dati empirici sono in regine critica, per calcolare il suo valore bisogna solo imporre $\varphi_{1-\frac{\alpha_s}{2}} = |U_0|$ dove U_0 è il valore empirico della statistica (9.15). Dalla definizione di quantile si ha allora che $1 - \frac{\alpha_s}{2} = \Phi(|U_0|)$ dove Φ è la *FDC* normale standard, e quindi

$$\boxed{\alpha_s = 2\left[1 - \Phi(|U_0|)\right]}$$

Test unilaterale destro e sinistro

Seguendo un percorso strettamente analogo al precedente, ma partendo dalla (9.13), per il **test unilaterale destro** di livello α bisogna richiedere

$$\begin{aligned} \boldsymbol{P}_{\mu_0}\left\{\overline{X}_n - \mu_0 > \delta\right\} &= \boldsymbol{P}_{\mu_0}\left\{U_0 > \frac{\delta\sqrt{n}}{\sigma}\right\} = \boldsymbol{P}\left\{\mathfrak{N}(0,1) > \frac{\delta\sqrt{n}}{\sigma}\right\} \\ &= = 1 - \boldsymbol{P}\left\{\mathfrak{N}(0,1) \le \frac{\delta\sqrt{n}}{\sigma}\right\} = \alpha \end{aligned}$$

per cui si ha questa volta

$$\frac{\delta\sqrt{n}}{\sigma} = \varphi_{1-\alpha}$$

e l'**evento critico** diviene semplicemente

$$\boxed{D = \{U_0 > \varphi_{1-\alpha}\}} \tag{9.17}$$

In conclusione un test unilaterale destro di livello α delle ipotesi (9.9) si esegue con la seguente **procedura**:

> *Si calcola il valore empirico della statistica U_0 e lo si confronta con il quantile $\varphi_{1-\alpha}$ della normale standard ricavato dalle Tavole E.1: se risulta $U_0 > \varphi_{1-\alpha}$, allora si rifiuta l'ipotesi $\mathcal{H}_0$ e si accetta $\mathcal{H}_1$; se invece risulta $U_0 \le \varphi_{1-\alpha}$ si accetta l'ipotesi $\mathcal{H}_0$*

La significatività si ottiene poi imponendo $\varphi_{1-\alpha_s} = U_0$, cioè $1 - \alpha_s = \Phi(U_0)$ e quindi

$$\boxed{\alpha_s = 1 - \Phi(U_0)}$$

Il **test unilaterale sinistro** di livello α delle ipotesi (9.10) si costruisce infine in maniera identica con l'unica differenza che l'**evento critico** sarà ora

$$\boxed{D = \{U_0 < -\varphi_{1-\alpha}\}}$$

e la **procedura** diviene

> *Si calcola il valore empirico della statistica U_0 e lo si confronta con il quantile $\varphi_{1-\alpha}$ della normale standard ricavato dalle Tavole E.1: se risulta $U_0 < -\varphi_{1-\alpha}$, allora si rifiuta l'ipotesi $\mathcal{H}_0$ e si accetta $\mathcal{H}_1$; se invece risulta $U_0 \ge -\varphi_{1-\alpha}$ si accetta l'ipotesi $\mathcal{H}_0$*

La significatività del test unilaterale sinistro si ottiene allora da $\varphi_{1-\alpha_s} = -U_0$, cioè $1 - \alpha_s = \Phi(-U_0)$ e quindi

$$\boxed{\alpha_s = 1 - \Phi(-U_0)}$$

166.6	169.3	168.2	176.4	168.6	170.1	167.7	168.1	164.3	171.1
172.5	165.7	166.1	171.3	176.5	168.8	169.7	168.1	167.1	172.8
173.5	168.9	169.7	167.7	173.0	159.4	168.8	163.7	174.4	174.0
164.4	171.1	168.1	171.4	174.6	168.7	169.4	165.7	159.5	164.1
166.0	168.1	169.0	172.6	172.2	170.4	173.4	181.5	165.5	167.9
168.9									

Tabella 9.1: Altezze in *cm* di un campione di $n = 51$ reclute.

Esempio 9.4. *È noto che l'altezza X delle persone di un determinato paese è una v-a che segue una legge normale $\mathfrak{N}(\mu, \sigma^2)$: supponiamo ora di sapere anche, in base ai dati di un censimento del 1950, che per gli individui di sesso maschile si ha una media $\mu_0 = 168$ cm con una varianza $\sigma^2 = 19$ cm^2. Nel 1965 viene esaminato alla visita di leva un campione di $n = 51$ reclute, e se ne riportano le altezze nella Tabella 9.1. Si constata a questo punto che la media del campione è $\overline{X}_n = 169.3$ cm. Supponendo di poter considerare la varianza $\sigma^2 = 19$ cm^2 come ancora attendibile e quindi nota, si vuol sapere se al livello $\alpha = 0.05$ possiamo dire che la media delle altezze è aumentata*

Il test richiesto è dunque unilaterale destro con ipotesi del tipo (9.9): dai dati a nostra disposizione il valore empirico della statistica di riferimento (9.15) è

$$U_0 = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} = \frac{169.3 - 168}{\sqrt{19}}\sqrt{51} = 2.13$$

Siccome dalle Tavole E.1 risulta

$$\varphi_{1-\alpha} = \varphi_{0.95} = 1.65 < 2.13 = U_0$$

si vede subito che i dati sono nella regione critica per cui il test unilaterale destro di livello $\alpha = 0.05$ ci conduce a rifiutare $\mathcal{H}_0$ e ad accettare l'ipotesi $\mathcal{H}_1$ che l'altezza sia aumentata

Dal valore di $U_0 = 2.13$ possiamo anche calcolare la significatività del test che è $\alpha_s = 1 - \Phi(2.13) = 0.017$, un valore non particolarmente buono che lascia qualche dubbio sulla sicurezza del'esito del test. Infatti se avessimo svolto i calcoli con un livello $\alpha = 0.01$ avremmo trovato

$$\varphi_{1-\alpha} = \varphi_{0.99} = 2.33 > 2.13 = U_0$$

e il risultato del test sarebbe stato capovolto: l'altezza media è rimasta di 168 cm

Il problema dovrebbe essere affrontato in maniera un po' diversa se si chiedesse di verificare che l'altezza media è cambiata (non aumentata). In questo caso il test sarebbe bilaterale: il calcolo di U_0 resta invariato, e al livello $\alpha = 0.05$ avremmo

$$\varphi_{1-\frac{\alpha}{2}} = \varphi_{0.975} = 1.96 < 2.13 = U_0$$

per cui anche il test bilaterale confermerebbe il rifiuto di $\mathcal{H}_0$. La significatività è però peggiorata: $\alpha_s = 2[1 - \Phi(2.13)] = 0.033$, e quindi l'esito del test sarebbe ancora meno sicuro

9.2.2 Test di Student

Sempre supponendo che – o sulla base del Teorema 3.23 quando il campione $X_1, \ldots, X_n$ è $\mathfrak{N}(\mu, \sigma^2)$, o almeno approssimativamente dal *TLC* Teorema 5.5 – la media aritmetica del campione $\overline{X}_n$ possa essere considerata normale $\mathfrak{N}(\mu, {}^{\sigma^2}/_n)$ con μ sconosciuta, esamineremo ora il caso in cui ***la varianza*** σ^2 ***non è nota***. La differenza pratica con il caso precedente risiede nel fatto che ora il valore empirico di U_0 (9.15) non può più essere calcolato compiutamente perché manca il valore di σ. Dovremo quindi prima stimare σ^2 tramite la varianza corretta S_n^2 (8.1), e poi ricordare che – almeno approssimativamente – dal Teorema 3.23 si ha che la statistica

$$\boxed{T_0 = \sqrt{n}\,\frac{\overline{X}_n - \mu_0}{S_n} \sim \mathfrak{T}(n-1) \qquad \text{se } \mu = \mu_0} \tag{9.18}$$

segue una legge di Student $\mathfrak{T}(n-1)$ con $n-1$ gradi di libertà

Test bilaterale

Seguendo un percorso analogo a quello del Test di Gauss si perviene alla definizione dell'**evento critico** che ora prende la forma

$$\boxed{D = \left\{|T_0| > t_{1-\frac{\alpha}{2}}(n-1)\right\}} \tag{9.19}$$

dove $t_{1-\frac{\alpha}{2}}(n-1)$ è il quantile della legge di Student $\mathfrak{T}(n-1)$ che può essere trovato sulle Tavole E.2. Infatti da (3.28) si ha facilmente

$$\boldsymbol{P}_{\mu_0}\left\{|T_0| > t_{1-\frac{\alpha}{2}}(n-1)\right\} = \boldsymbol{P}\left\{|\mathfrak{T}(n-1)| > t_{1-\frac{\alpha}{2}}(n-1)\right\} = \alpha$$

Il test bilaterale di Student di livello α delle ipotesi (9.8) si esegue quindi con la seguente **procedura**:

> *Si calcola il valore empirico della statistica T_0 (9.18) e si confronta il suo valore assoluto con il quantile $t_{1-\frac{\alpha}{2}}(n-1)$ della legge di Student $\mathfrak{T}(n-1)$ ricavato dalle Tavole E.2: se risulta $|T_0| > t_{1-\frac{\alpha}{2}}(n-1)$, allora si rifiuta l'ipotesi $\mathcal{H}_0$ e si accetta $\mathcal{H}_1$; se invece risulta $|T_0| \le t_{1-\frac{\alpha}{2}}(n-1)$ si accetta l'ipotesi $\mathcal{H}_0$*

Infine la significatività del test bilaterale è

$$\alpha_s = 2\,[1 - F_{n-1}(|T_0|)]$$

dove $F_{n-1}(x)$ è la *FDC* della legge di Student $\mathfrak{T}(n-1)$: si noti però che i valori di questa, e di altre *FDC* non Gaussiane non sono dati nelle tavole in appendice, per cui un calcolo della significatività – tranne che nel caso del Test di Gauss – non sarà possibile senza l'ausilio di ulteriori strumenti di calcolo

Test unilaterale destro e sinistro

In modo simile si progettano i test unilaterali destro e sinistro di livello α per le ipotesi (9.9) e (9.10): gli **eventi critici** sono ora rispettivamente

$$\boxed{D = \begin{cases} \{T_0 > t_{1-\alpha}(n-1)\} & \text{test unilaterale destro} \\ \{T_0 < -t_{1-\alpha}(n-1)\} & \text{test unilaterale sinistro} \end{cases}} \tag{9.20}$$

dove T_0 è sempre data da (9.18), e i test si eseguono con la seguente **procedura**

> *Si calcola il valore empirico della T_0 (9.18) e lo si confronta con il quantile $t_{1-\alpha}(n-1)$ della legge di Student $\mathfrak{T}(n-1)$ ricavato dalle Tavole E.2: se risulta $T_0 > t_{1-\alpha}(n-1)$ (rispettivamente: $T_0 < -t_{1-\alpha}(n-1)$), allora si rifiuta l'ipotesi $\mathcal{H}_0$ e si accetta $\mathcal{H}_1$; se invece risulta $T_0 \leq t_{1-\alpha}(n-1)$ (rispettivamente: $T_0 \geq -t_{1-\alpha}(n-1)$) si accetta l'ipotesi $\mathcal{H}_0$*

Le significatività dei test unilaterali destro e sinistro sono infine rispettivamente

$$\alpha_s = \begin{cases} 1 - F_{n-1}(T_0) & \text{test unilaterale destro} \\ 1 - F_{n-1}(-T_0) & \text{test unilaterale sinistro} \end{cases}$$

Il caso $n > 120$

Nelle Tavole E.2 i quantili della distribuzione di Student $\mathfrak{T}(n-1)$ sono riportati solo per $n \leq 120$. Per eseguire i test nel caso $n > 120$ bisogna allora ricordare che per n molto grande la distribuzione di Student tende a coincidere con la distribuzione normale standard $\mathfrak{N}(0,1)$. Conseguentemente, per un dato valore del livello α, e per $n > 120$ avremo

$$t_\alpha(n-1) \simeq \varphi_\alpha$$

e quindi al posto dei quantili della Student $\mathfrak{T}(n-1)$ potremo usare quelli della normale standard $\mathfrak{N}(0,1)$ prendendoli dalle Tavole E.1. In questo caso il test per la media può essere effettuato calcolando il valore empirico T_0 da (9.18), ma usando i quantili della normale standard negli **eventi critici** che diventano

$$\boxed{D = \begin{cases} \{|T_0| > \varphi_{1-\frac{\alpha}{2}}\} & \text{test bilaterale} \\ \{T_0 > \varphi_{1-\alpha}\} & \text{test unilaterale destro} \\ \{T_0 < -\varphi_{1-\alpha}\} & \text{test unilaterale sinistro} \end{cases}} \tag{9.21}$$

In pratica per $n > 120$ le **procedure** per un test di Student di livello α si modificano nel modo seguente:

> *Si calcola il valore empirico di T_0 (9.18) e lo si confronta con gli opportuni quantili della normale standard delle Tavole E.1: l'ipotesi $\mathcal{H}_0$ si rifiuta rispettivamente se $|T_0| > \varphi_{1-\frac{\alpha}{2}}$ nel test bilaterale; se $T_0 > \varphi_{1-\alpha}$ nel test unilaterale destro; e se $T_0 < -\varphi_{1-\alpha}$ nel test unilaterale sinistro. Viceversa nei casi opposti*

Esempio 9.5. *Riprendendo l'Esempio 9.4 supponiamo ora di non poter considerare come attendibile il valore* 19 *per la varianza* σ^2 *della nostra v-a* X*. Dovremo allora innanzitutto stimare la varianza corretta (8.1) – che con un rapido calcolo risulta essere* $S_n^2 = 16.5$ *– e poi calcolare la statistica di Student*

$$T_0 = \frac{\overline{X}_n - \mu_0}{S_n}\sqrt{n} = \frac{169.3 - 168}{\sqrt{16.5}}\sqrt{51} \simeq 2.29$$

Siccome dalle Tavole E.2 risulta

$$t_{1-\alpha}(n-1) = t_{0.95}(50) \simeq 1.68 < 2.29 = T_0$$

anche con un test unilaterale destro di Student di livello $\alpha = 0.05$ *i dati si troveranno in regione critica, e quindi rifiuteremo l'ipotesi* $\mathcal{H}_0$ *che l'attesa sia rimasta uguale a* $\mu_0 = 168$ *cm. Da* $T_0 = 2.29$ *(ma con tavole più complete delle nostre) potremmo poi calcolare anche la significatività*

$$\alpha_s = 1 - F_{n-1}(T_0) = 1 - F_{50}(2.29) \simeq 0.013$$

un valore che insinua di nuovo qualche dubbio sulla affidabilità del test. Infatti se avessimo svolto i calcoli con un livello $\alpha = 0.01$ *avremmo trovato*

$$t_{1-\alpha}(n-1) = t_{0.99}(50) = 2.40 > 2.29 = T_0$$

per cui in questo caso l'esito del test sarebbe stato quello di accettare l'ipotesi $\mathcal{H}_0$ *secondo la quale l'altezza media è rimasta di* 168 *cm*

9.3 Test per il confronto delle medie

Un altro tipo di test riguarda il confronto fra le attese μ_X e μ_Y di due *v-a* X e Y. Dovremo qui però distinguere subito due casi, secondo che i due campioni siano *accoppiati* o *indipendenti*:

- *Chiameremo* ***campione accoppiato*** *un campione* $(X_1, Y_1), \ldots, (X_n, Y_n)$ *di coppie di misure per le quali è importante conservare la connessione fra gli elementi* X_k *e* Y_k *con lo stesso indice.* Se, ad esempio, vogliamo studiare l'effetto di un farmaco possiamo somministrandolo a n pazienti, potremmo misurare su ognuno il valore di qualche parametro rilevante prima (X) e dopo (Y) la somministrazione per mettere in evidenza eventuali differenze fra i comportamenti medi μ_X e μ_Y. In questo caso ovviamente è importante non perdere l'accoppiamento fra le misure X_k e Y_k eseguite sul paziente k-mo

- *Parleremo invece di* ***campioni indipendenti*** *quando non vi è nessuna relazione rilevante fra gli elementi* $X_1, \ldots, X_n$ *del primo campione e gli elementi* $Y_1, \ldots, Y_m$ *del secondo*, nel senso che gli elementi *corrispondenti* X_k e Y_k non

sono misurati sullo stesso individuo, e la loro collocazione al posto k-mo è convenzionale e priva di significato statistico. Peraltro due campioni indipendenti possono contenere numeri diversi n ed m di elementi, situazione evidentemente non consentita nel caso di un campione accoppiato. Un esempio di campioni indipendenti si ha quando si confronta l'effetto di un farmaco con quello di un *placebo*: in questo caso si somministrano il farmaco e il *placebo* a due gruppi distinti di pazienti (anche di numero diverso) e si misura qualche parametro rilevante sui due gruppi (X e Y) per confrontarne i valori d'attesa μ_X e μ_Y

Anche il confronto fra le medie viene formalizzato tramite l'individuazione di opportune ipotesi: avremo così un **test bilaterale** se le ipotesi sono

$$\mathcal{H}_0 : \mu_X = \mu_Y \qquad \mathcal{H}_1 : \mu_X \neq \mu_Y \tag{9.22}$$

Avremo invece un **test unilaterale** quando le ipotesi sono del tipo

$$\mathcal{H}_0 : \mu_X \leq \mu_Y \qquad \mathcal{H}_1 : \mu_X > \mu_Y \tag{9.23}$$

Per ambedue i test, comunque, il confronto deve essere fatto seguendo procedure diverse per campioni accoppiati e campioni indipendenti come vedremo nella discussione seguente. Si noti inoltre che questa volta sarebbe stato superfluo parlare di test unilaterale *destro o sinistro*: i due casi infatti si ottengono uno dall'altro scambiando i nomi X e Y delle due *v-a*

9.3.1 Campioni accoppiati

Il caso dei campioni accoppiati si riconduce facilmente alle procedure già sviluppate nella precedente Sezione 9.2: per far questo basterà costruire innanzitutto il **campione delle differenze**

$$Z_k = X_k - Y_k \qquad k = 1, \ldots, n \tag{9.24}$$

e poi si eseguono dei test sull'unico campione $Z_1, \ldots, Z_n$ confrontando la sua attesa $\mu = \mu_X - \mu_Y$ con in valore $\mu_0 = 0$. Le ipotesi per i test bilaterali (9.22) e unilaterali (9.23) diventano allora rispettivamente

$$\mathcal{H}_0 : \mu = 0 \qquad \mathcal{H}_1 : \mu \neq 0 \qquad \text{bilaterale} \tag{9.25}$$

$$\mathcal{H}_0 : \mu \leq 0 \qquad \mathcal{H}_1 : \mu > 0 \qquad \text{unilaterale} \tag{9.26}$$

e quindi si ricade sostanzialmente nei test già studiati nella Sezione 9.2 con il particolare valore $\mu_0 = 0$. Nel seguito supporremo che le *v-a* Z_k abbiano attesa μ sconosciuta, ma come nella sezione precedente dovremo distinguere i due casi in cui la loro varianza σ^2 sia nota o meno

Varianza σ^2 nota

Se le *v-a* Z_k hanno attesa μ sconosciuta, ma la loro varianza σ^2 è nota, posto

$$\overline{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k$$

potremo come al solito supporre che, nell'ipotesi $\mathcal{H}_0 : \mu = 0$, almeno approssimativamente risulti

$$\boxed{U_0 = \sqrt{n}\,\frac{\overline{Z}_n}{\sigma} \sim \mathfrak{N}(0,1) \qquad \text{se } \mu = 0} \tag{9.27}$$

cioè che U_0 sia normale standard. Le procedure per i test di livello α sono allora le stesse della Sezione 9.2.1 con **eventi critici**

$$\boxed{D = \begin{cases} \{|U_0| > \varphi_{1-\frac{\alpha}{2}}\} & \text{test bilaterale} \\ \{U_0 > \varphi_{1-\alpha}\} & \text{test unilaterale} \end{cases}} \tag{9.28}$$

e significatività

$$\boxed{\alpha_s = \begin{cases} 2\,[1 - \Phi(|U_0|)] & \text{test bilaterale} \\ 1 - \Phi(U_0) & \text{test unilaterale} \end{cases}} \tag{9.29}$$

Varianza σ^2 non nota

Se viceversa la varianza σ^2 non è nota si introduce la varianza corretta del campione delle Z_k

$$S_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Z_k - \overline{Z}_n)^2$$

e nell'ipotesi $\mathcal{H}_0 : \mu = 0$ potremo supporre che almeno approssimativamente sia

$$\boxed{T_0 = \sqrt{n}\,\frac{\overline{Z}_n}{S_n} \sim \mathfrak{T}(n-1) \qquad \text{se } \mu = 0} \tag{9.30}$$

cioè che T_0 sia una Student con $n-1$ gradi di libertà. Gli **eventi critici** dei test di livello α sono allora

$$\boxed{D = \begin{cases} \{|T_0| > t_{1-\frac{\alpha}{2}}(n-1)\} & \text{test bilaterale} \\ \{T_0 > t_{1-\alpha}(n-1)\} & \text{test unilaterale} \end{cases}} \tag{9.31}$$

(le procedure sono quelle di Sezione 9.2.2) e le significatività

$$\alpha_s = \begin{cases} 2\,[1 - F_{n-1}(|T_0|)] & \text{test bilaterale} \\ 1 - F_{n-1}(T_0) & \text{test unilaterale} \end{cases}$$

dove F_{n-1} sono le *FDC* della legge di Student $\mathfrak{T}(n-1)$

Y	X	Z	Y	X	Z	Y	X	Z
80	85	5	70	82	12	78	70	-8
80	84	4	65	73	8	75	77	2
82	87	5	83	89	6	76	76	0
75	81	6	74	85	11	78	82	4
80	79	-1	81	86	5	77	83	6
74	85	11	68	72	4	75	80	5
80	87	7	69	74	5	72	80	8
72	78	6	71	77	6	71	81	10
91	86	-5	70	75	5	75	76	1
88	80	-8	73	81	8	78	77	-1

Tabella 9.2: Pulsazioni di $n = 30$ pazienti prima (Y) e dopo (X) l'assunzione di un farmaco.

Esempio 9.6. *Si sperimenta un farmaco su un campione di $n = 30$ pazienti rilevando il numero delle pulsazioni al minuto prima (Y) e dopo (X) la somministrazione: i dati sono riportati nella Tabella 9.2. Possiamo dire in base a questi valori, e ad un livello $\alpha = 0.05$, che la frequenza delle pulsazioni è aumentata?*

Il modo di porre la domanda suggerisce subito che dovremo discutere delle ipotesi nella forma (9.23). Inoltre è evidente che il nostro campione è accoppiato dato che sarà importante conservare l'associazione fra i valori X e Y misurati su ogni singolo paziente. Cominceremo quindi con il definire il campione delle differenze (9.24) i cui valori sono riportati nella Tabella 9.2. Si noti che noi siamo interessati a scoprire se può essere ritenuta vera l'ipotesi alternativa $\mu_X > \mu_Y$ di (9.23), cioè se $\mu = \mu_X - \mu_Y > 0$: pertanto sul campione delle differenze le ipotesi prendono la forma del test unilaterale (9.26)

Siccome non ci sono state fornite informazioni sulla varianza delle Z_k, passeremo innanzitutto a calcolare media $\overline{Z}_n$, varianza corretta S_n^2 e statistica di Student T_0 (9.30) del campione delle differenze riportato in Tabella ottenendo

$$\overline{Z}_n = 4.23 \qquad S_n = 5.01 \qquad T_0 = \frac{\overline{Z}_n}{S_n}\sqrt{n} \simeq 4.63$$

A questo punto per eseguire il test di livello $\alpha = 0.05$ dobbiamo paragonare il valore di T_0 con l'opportuno quantile di Student delle Tavole E.2 ottenendo

$$t_{1-\alpha}(n-1) = t_{0.95}(29) \simeq 1.70 < 4.63 \simeq T_0$$

per cui il campione è in regione critica, e quindi il test ci suggerisce di accettare $\mathcal{H}_1$, cioè che il farmaco ha prodotto un aumento della frequenza delle pulsazioni. La significatività del test questa volta è piuttosto buona: infatti $\alpha_s = 1 - F_{29}(4.63)$ dove $F_{29}(x)$ è la FDC della legge $\mathfrak{T}(n-1) = \mathfrak{T}(29)$; questo valore non è presente sulle Tavole E.2 ma può essere calcolato in altro modo e alla fine risulta $\alpha_s = 0.00003$

9.3.2 Campioni indipendenti

Nel caso di campioni indipendenti con numerosità rispettive n ed m non ha alcun senso studiare il campione delle differenze. In questo caso supporremo che le X_j e le Y_k abbiano rispettivamente attese μ_X, μ_Y e varianze σ_X^2, σ_Y^2, definiremo separatamente le medie dei due campioni

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad\qquad \overline{Y}_m = \frac{1}{m}\sum_{k=1}^{m} Y_k$$

e supporremo come al solito che esse siano (almeno approssimativamente) Gaussiane. Tenendo poi presenti le ipotesi nelle forme (9.22) e (9.23), distingueremo due casi secondo le informazioni disponibili sulle varianze

Varianze σ_X^2 e σ_Y^2 note

Se le varianze σ_X^2 e σ_Y^2 sono note si può dimostrare che, supponendo vera $\mathcal{H}_0 = \mu_X = \mu_Y$, la statistica

$$\boxed{U_0 = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim \mathfrak{N}(0,1) \qquad \text{se } \mu_X = \mu_Y} \tag{9.32}$$

è normale standard. Pertanto le procedure per i test potranno essere definite in analogia con quanto stabilito nella Sezione 9.2: gli **eventi critici** dei test di livello α sono

$$\boxed{D = \begin{cases} \{|U_0| > \varphi_{1-\frac{\alpha}{2}}\} & \text{test bilaterasle} \\ \{U_0 > \varphi_{1-\alpha}\} & \text{test unilaterale} \end{cases}} \tag{9.33}$$

e le corrispondenti **procedure** sono:

> *Si calcola il valore empirico della U_0 (9.32) e lo si confronta con gli opportuni quantili della normale standard: l'ipotesi $\mathcal{H}_0$ viene rifiutata rispettivamente se $|U_0| > \varphi_{1-\frac{\alpha}{2}}$ nel test bilaterale, e se $U_0 > \varphi_{1-\alpha}$ nel test unilaterale*

Varianze σ_X^2 e σ_Y^2 non note

Se invece le varianze σ_X^2 e σ_Y^2 non sono note, è necessario introdurre una nuova quantità: la **varianza combinata**

$$\boxed{V^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}} \tag{9.34}$$

che è una media delle varianze corrette S_X^2 e S_Y^2 pesate con le rispettive numerosità dei campioni

X	9.2	8.3	10.3	11.0	12.0	8.6	9.3	10.3	9.7	9.0
Y	10.9	11.3	10.9	10.2	9.3	10.4	10.5	11.3	10.8	12.0
	10.6	10.4	12.2	10.9	10.4					

Tabella 9.3: Produttività mensili di una ditta prima (Y) e dopo (X) aver introdotto cambiamenti nel processo produttivo.

Si può allora dimostrare che, se le varianze σ_X^2 e σ_Y^2 sono almeno approssimativamente uguali, e se è vera $\mathcal{H}_0 : \mu_X = \mu_Y$, la statistica

$$\boxed{T_0 = \frac{\overline{X}_n - \overline{Y}_m}{V\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim \mathfrak{T}(n+m-2) \qquad \text{se } \mu_X = \mu_Y} \tag{9.35}$$

è una Student con $n+m-2$ gradi di libertà. In tal caso gli **eventi critici** dei test di livello α sono rispettivamente

$$\boxed{D = \begin{cases} \{|T_0| > t_{1-\frac{\alpha}{2}}(n+m-2)\} & \text{test bilaterale} \\ \{T_0 > t_{1-\alpha}(n+m-2)\} & \text{test unilaterale} \end{cases}} \tag{9.36}$$

con conseguente definizione delle **procedure**

> *Si calcola prima la varianza combinata (9.34), poi il valore empirico della T_0 (9.35) e lo si confronta con gli opportuni quantili della legge di Student con $n+m-2$ gradi di libertà: l'ipotesi $\mathcal{H}_0$ viene rifiutata rispettivamente se $|T_0| > t_{1-\frac{\alpha}{2}}(n+m-2)$ nel test bilaterale, e se $T_0 > t_{1-\alpha}(n+m-2)$ nel test unilaterale*

Esempio 9.7. *Una ditta introduce dei cambiamenti nel proprio processo produttivo e vuol sapere se questo ha modificato la produttività. Sono disponibili i dati delle produzioni mensili (in una opportuna unità di misura): nella Tabella 9.7 sono riportati $n = 10$ valori successivi (X), e $m = 15$ precedenti (Y) al cambiamento del processo produttivo. I campioni sono indipendenti dato che non si attribuisce particolare significato all'accoppiamento dei valori mensili (che peraltro sono in numero differente). Supponendo noto che ambedue le v-a X e Y sono normali $\mathfrak{N}(\mu_X, \sigma_X^2)$ e $\mathfrak{N}(\mu_Y, \sigma_Y^2)$, si vuol sapere se, sulla base dei dati empirici per i quali*

$$\overline{X}_n = 9.8 \qquad \overline{Y}_m = 10.8$$

e al livello $\alpha = 0.02$, possiamo affermare che la prodttività è cambiata, o se dobbiamo conservare l'ipotesi $\mathcal{H}_0 : \mu_X = \mu_Y$

Supponiamo in un primo momento di sapere che la deviazione standard della produttività non è stata modificata dai cambiamenti, e che essa vale $\sigma_X = \sigma_Y = 1.1$. In questo caso potremo eseguire un test di Gauss con la statistica (9.32) il cui valore è $U_0 = -2.23$: dalle Tavole E.1 abbiamo allora

$$\varphi_{1-\frac{\alpha}{2}} = \varphi_{0.99} = 2.33 > 2.23 = |U_0|$$

per cui i dati non sono in regione critica, e il test bilaterale di livello $\alpha = 0.02$ *suggerisce di accettare l'ipotesi* $\mathcal{H}_0 : \mu_X = \mu_Y$. *Se invece supponiamo di non conoscere nulla sulle deviazioni standard di* X *e* Y, *dovremo applicare un test di Student con la statistica (9.35). Sarà necessario allora stimare preventivamente dai dati empirici le varianze corrette e la varianza combinata (9.34)*

$$S_X^2 = 1.30 \qquad S_Y^2 = 0.51 \qquad V^2 = 0.82$$

e calcolare poi il valore della statistica $T_0 = -2.71$. *Dalle Tavole E.2 ricaviamo allora che*

$$t_{1-\frac{\alpha}{2}}(n+m-2) = t_{0.99}(23) = 2.50 < 2.71 = |T_0|$$

cioè nel test di Student di livello $\alpha = 0.02$ *i dati si trovano in regione critica, e quindi ci viene suggerito di rifiutare l'ipotesi* $\mathcal{H}_0 : \mu_X = \mu_Y$. *I risultati discordanti dei due tipi di test sono ovviamente un segno di incertezza: un esame della significatività dei due casi conferma peraltro questa intuizione. Infatti si ha* $\alpha_s = 2[1 - \Phi(|U_0|)] = 0.03$ *nel caso di Gauss e* $\alpha_s = 2[1 - F_{23}(|T_0|)] = 0.01$ *nel caso di Student: due valori piuttosto elevati che sollevano dei dubbi sulla attendibilità dei risultati*

9.4 Test di Fisher sulla varianza

Anche la stima della varianza è un aspetto importante dell'analisi statistica che abbiamo già incontrato nella Sezione 8.2.2: qui esamineremo le procedure per verificare alcune ipotesi relative alle varianze di due campioni. Supporremo a questo scopo di avere due campioni (indipendenti) $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, rispettivamente con n e m elementi, estratti da due *v-a* X e Y con varianze σ_X^2 e σ_Y^2: le ipotesi che sottoporremo a verifica – nelle solite forme bilaterale e unilaterale – saranno allora rispettivamente

$$\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 : \sigma_X^2 \neq \sigma_Y^2 \tag{9.37}$$

$$\mathcal{H}_0 : \sigma_X^2 \leq \sigma_Y^2 \qquad \mathcal{H}_1 : \sigma_X^2 > \sigma_Y^2 \tag{9.38}$$

Si noti che anche qui abbiamo indicato solo una forma di test unilaterale: sarebbe infatti superfluo distinguere i casi *destro* e *sinistro* dal momento che si può ottenere uno dall'altro *scambiando il ruolo di* X *e* Y

Naturalmente la verifica delle ipotesi (9.37) e (9.38) sarà effettuata esaminando i valori empirici delle varianze corrette S_X^2 e S_Y^2 dei due campioni. Ma questa volta – diversamente dal caso delle attese – invece di studiarne la *differenza* sarà meglio esaminare il loro *rapporto*. Infatti si può dimostrare che, supponendo $\sigma_X^2 = \sigma_Y^2$, almeno approssimativamente la statistica

$$\boxed{F_0 = \frac{S_X^2}{S_Y^2} \sim \mathfrak{F}(n-1, m-1) \qquad \text{se } \sigma_X^2 = \sigma_Y^2} \tag{9.39}$$

segue la legge di Fisher $\mathfrak{F}(n-1, m-1)$. Sarà importante sottolineare a proposito della relazione (9.39) che la **posizione dei due indici della legge di Fisher** $\mathfrak{F}(n-1, m-1)$ non è simmetrica: la legge si modifica scambiando n con m. Pertanto è fondamentale saper riconoscere quale dei due deve essere messo al primo, e quale al secondo posto. Il principio che guida questa scelta in (9.39) è il seguente

> *Al* ***primo posto*** *in* $\mathfrak{F}(n-1, m-1)$ *deve essere inserito l'indice (qui* n*) che rappresenta la numerosità del campione (qui* X_k*) la cui varianza è al* ***numeratore*** *della* F_0*. Gli indici* n *e* m *si scambiano se vengono scambiati i ruoli di* X *e* Y *in* F_0

Il test poi consiste nel confrontare il valore di F_0 con gli opportuni quantili di Fisher: le due varianze σ_X^2 e σ_Y^2 saranno giudicate uguali quando F_0 assume valori vicini a 1, mentre potremo dire che le due varianze sono diverse se F_0 è abbastanza diversa da 1. Pertanto, tenendo conto della asimmetria della *fdp* della legge di Fisher, gli eventi critici di livello α per i test bilaterale e unilaterale saranno rispettivamente

$$\begin{aligned} \{F_0 < f_{\frac{\alpha}{2}}(n-1, m-1)\} \cup \{F_0 > f_{1-\frac{\alpha}{2}}(n-1, m-1)\} \qquad & \text{test bilaterale} \\ \{F_0 > f_{1-\alpha}(n-1, m-1)\} \qquad & \text{test unilaterale} \end{aligned}$$

dove le $f_\alpha(n, m)$ indicano i quantili delle corrispondenti leggi di Fisher che possono essere ricavati dalle Tavole E.4. Per semplificare però la forma dell'evento critico *bilaterale*, osserveremo che esso consiste nell'unione di due eventi che rappresentano i casi in cui F_0 *cade fuori* dall'intervallo $[f_{\frac{\alpha}{2}}\ ,\ f_{1-\frac{\alpha}{2}}]$ dei quantili di Fisher: in questa eventualità infatti il valore di F_0 è considerato troppo diverso da 1 per poter accettare $\mathcal{H}_0$. Sarà naturale allora riformulare gli **eventi critici** nel modo seguente

$$\boxed{D = \begin{cases} \overline{\{f_{\frac{\alpha}{2}}(n-1, m-1) \le F_0 \le f_{1-\frac{\alpha}{2}}(n-1, m-1)\}} & \text{test bilaterale} \\ \{F_0 > f_{1-\alpha}(n-1, m-1)\} & \text{test unilaterale} \end{cases}} \tag{9.40}$$

ricordando che il primo di questi è ora il *negativo* (complementare) di un evento secondo quanto definito nella Sezione 1.2: esso indica infatti che F_0 *non cade* nell'intervallo $[f_{\frac{\alpha}{2}}\ ,\ f_{1-\frac{\alpha}{2}}]$

Esempio 9.8. *Riprendendo la discussione dell'Esempio 9.7 possiamo porci il problema di verificare se, al livello* $\alpha = 0.05$*, le varianze* σ_X^2 *e* σ_Y^2 *di* X *e* Y *possono essere considerate uguali. Bisogna allora eseguire un test bilaterale di Fisher: dai calcoli dell'Esempio 9.7 ricaviamo prima di tutto il valore della statistica di Fisher* $F_0 = 2.55$*. Calcoliamo poi i quantili della legge di Fisher che definiscono la regione critica: il valore* $f_{0.975}(9, 14) = 3.21$ *si trova immediatamente dalle Tavole E.4; l'altro quantile che invece non si trova sulle Tavole va calcolato da (3.33)*

$$f_{0.025}(9, 14) = \frac{1}{f_{0.975}(14, 9)} = 0.26$$

A questo punto possiamo eseguire il test, e siccome

$$f_{0.025}(9, 14) = 0.26 \le F_0 = 2.55 \le 3.21 = f_{0.975}(9, 14)$$

al livello $\alpha = 0.05$ *accetteremo l'ipotesi nulla bilaterale* $\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2$

9.5 Test del χ^2 di adattamento

Prenderemo ora in esame il problema di decidere se un dato campione di n misure possa essere considerato estratto da una *v-a* X con una distribuzione $\mathfrak{L}$. L'oggetto della nostra indagine, quindi, non è più il valore di qualche parametro θ della legge di X, ma l'eventuale **adattamento** dei dati sperimentali ad una possibile legge teorica $\mathfrak{L}$ di X

Nel seguito supporremo che la legge $\mathfrak{L}$ di X preveda che essa assuma solo un numero *finito* m di possibili valori con *probabilità teoriche* $p_1, \ldots, p_m$. Se invece X assume infiniti valori, o è continua, sceglieremo qualche opportuno metodo per raggruppare i dati in un numero finito m di classi con probabilità $p_1, \ldots, p_m$ ottenendo in questo modo una ragionevole approssimazione della legge $\mathfrak{L}$. Le ipotesi che dovremo sottoporre a verifica sono quindi

$$\begin{aligned} \mathcal{H}_0 &: \text{ il campione segue la legge } \mathfrak{L} \text{ delle } p_1, \ldots, p_m \\ \mathcal{H}_1 &: \text{ il campione segue un'altra legge} \end{aligned}$$

Siccome nella discussione che segue i valori numerici di X non giocano nessun ruolo, tutta la procedura può essere facilmente adattata anche al caso di variabili qualitative

Supponiamo allora che, in un campione di n misure, il valore j-mo di X venga sperimentalmente trovato N_j volte con $j = 1, \ldots, m$, e che quindi le $\overline{p}_j = {}^{N_j}/_n$ siano le corrispondenti frequenze relative: si noti che N_j e $\overline{p}_j$ sono *v-a* (cioè cambiano ad ogni ripetizione dell'esperimento), mentre le probabilità teoriche p_j di $\mathfrak{L}$ sono numeri dati nell'ipotesi $\mathcal{H}_0$. Ovviamente il test sarà eseguito con un confronto fra le p_j teoriche e le $\overline{p}_j$ empiriche, ed è quindi necessario individuare un'opportuna statistica che misuri una distanza complessiva fra la distribuzione empirica delle $\overline{p}_j$, e la distribuzione teorica $\mathfrak{L}$ delle p_j. Si dimostra a questo scopo che, se $\mathcal{H}_0$ è vera, la statistica di Pearson

$$\boxed{K_0 = n \sum_{j=1}^{m} \frac{(\overline{p}_j - p_j)^2}{p_j} = \sum_{j=1}^{m} \frac{(N_j - np_j)^2}{np_j} \sim \boldsymbol{\chi^2}(m-1) \qquad \text{se è vera } \mathcal{H}_0} \tag{9.41}$$

se n è abbastanza grande, segue approssimativamente una legge del *chi-quadro* $\boldsymbol{\chi^2}(m-1)$ con $m-1$ gradi di libertà. Evidentemente, per la sua definizione, K_0 tende a prendere valori piuttosto grandi se $\mathcal{H}_0$ non è verificata, perché in questo caso le differenze $\overline{p}_j - p_j$ non sono trascurabili. Pertanto l'**evento critico** di un test di livello α sarà

$$\boxed{D = \{K_0 > \chi^2_{1-\alpha}(m-1)\}} \tag{9.42}$$

dove $\chi^2_{1-\alpha}(m-1)$ sono i quantili della legge del *chi-quadro* che possono essere ricavati dalle Tavole E.3. Si ponga cura in questo test a non confondere il numero m dei possibili valori di X (numero che fissa i gradi di libertà della legge $\boldsymbol{\chi^2}(m-1)$) con il numero n delle misure contenute nel campione dei dati

La **procedura** per eseguire il test di adattamento è quindi la seguente

> *A partire dalle frequenze osservate $\overline{p}_j$ e da quelle teoriche p_j si calcola il valore di K_0 in (9.41), e poi lo si confronta con il quantile $\chi^2_{1-\alpha}(m-1)$ (dove m è il numero dei possibili valori di X): se $K_0 \leq \chi^2_{1-\alpha}(m-1)$ si accetta $\mathcal{H}_0$, cioè il campione si adatta alla legge $\mathfrak{L}$; se invece $K_0 > \chi^2_{1-\alpha}(m-1)$ l'ipotesi $\mathcal{H}_0$ viene rifiutata*

Un uso corretto di questo test richiede però alcune precisazioni

- Non sono noti risultati rigorosi circa la **grandezza di** n necessaria per rendere applicabile questo test; ci sono solo regole empiriche la più nota delle quali è la seguente

  > *Il numero n degli elementi de campione deve sempre essere abbastanza grande da avere $np_j \geq 5$ per tutti gli indici $j = 1, \ldots, m$*

  Se questo non avviene l'espediente più comune per aggirare la difficoltà è quello di **unificare le classi** con le p_j più piccole sommaando le probabilità, in modo da formarne altre con probabilità più grandi che rispettino la condizione $np_j \geq 5$

- A volte la distribuzione teorica $p_j$di $\mathfrak{L}$ non è completamente nota. Ad esempio ci si potrebbe porre il problema di un adattamento dei dati ad una legge Binomiale $\mathfrak{B}(n;p)$ per la quale p sia sconosciuto. In questo caso prima di applicare il test bisognerà **stimare i parametri** necessari a partire dai dati $X_1, \ldots, X_n$, e a seguito di ciò la **procedura** del test deve essere completata nel modo seguente

  > *Se inoltre la conoscenza completa della distribuzione teorica $\mathfrak{L}$ richiede la determinazione di $q < m-1$ parametri, bisogna prima stimare questi parametri mediante gli opportuni stimatori di MV, poi si deve calcolare K_0 di (9.41) e infine si applica il test usando come evento critico di livello α l'evento*

  $$\boxed{D = \{K_0 > \chi^2_{1-\alpha}(m-q-1)\}} \tag{9.43}$$

  In pratica, nella legge del *chi-quadro* di K_0, bisognerà **sottrarre tanti gradi di libertà quanti sono i parametri stimati**

Esempio 9.9. *Riprendiamo la discussione dell'Esempio 8.8 sul numero X di figli maschi nelle famiglie con 12 figli. Si ricorderà che le più semplici ipotesi iniziali conducono a un modello nel quale la legge $\mathfrak{L}$ della v-a X è binomiale $\mathfrak{B}(12; {}^1/_2)$: proviamo allora a sottoporre questa ipotesi $\mathcal{H}_0$ a un test del χ^2 con*

$$p_k = \binom{12}{k} \left(\frac{1}{2}\right)^k \left(1-\frac{1}{2}\right)^{12-k} = \binom{12}{k} \frac{1}{2^{12}}, \qquad k = 0, 1, \ldots, 12$$

Osserviamo però innanzitutto che per questa distribuzione $\mathfrak{B}(12;\,{}^1\!/_2)$ *le condizioni di applicabilità del test non sono rispettate dal nostro campione. Infatti per le classi estreme (quelle con probabilità più piccole) si ha* $p_0 = p_{12} = 2^{-12}$*, e quindi dato che* $n = 6\,115$ *risulta*

$$np_0 = np_{12} = 6\,115 \times 2^{-12} \simeq 1.49 < 5$$

Per adattare il test al nostro campione dovremo allora unificare le classi estreme $\{X = 0\}$ *e* $\{X = 12\}$*, rispettivamente con le classi contigue* $\{X = 1\}$ *e* $\{X = 11\}$*, e sommarne le probabilità in modo da avere*

$$\begin{aligned} n(p_0 + p_1) &= 6\,115 \times (2^{-12} + 12 \times 2^{-12}) = 6\,115 \times 13 \times 2^{-12} \simeq 19.47 > 5 \\ n(p_{11} + p_{12}) &= 6\,115 \times (12 \times 2^{-12} + 2^{-12}) = 6\,115 \times 13 \times 2^{-12} \simeq 19.47 > 5 \end{aligned}$$

In conclusione il nostro test del χ^2 *non dovrà essere eseguito per l'originaria distribuzione* $\mathfrak{B}(12;\,{}^1\!/_2)$ *a* 13 *valori con le* $p_0, \ldots, p_{12}$ *binomiali, ma su una nuova distribuzione raggruppata* $\mathfrak{L}$ *a* 11 *valori con probabilità così definite*

$$q_j = \begin{cases} p_0 + p_1 & \text{per } j = 1, \text{ cioè } k = 0, 1 \\ p_j & \text{per } j = k = 2, \ldots, 10 \\ p_{11} + p_{12} & \text{per } j = 11, \text{ cioè } k = 11, 12 \end{cases}$$

Naturalmente anche le frequenze empiriche $N_0, \ldots, N_{12}$ *riportati nella Tabella 8.2 dovranno essere raggruppate nelle* M_j *delle nuove classi secondo lo stesso schema, ottenendo*

$$M_j = \begin{cases} N_0 + N_1 & \text{per } j = 1 \\ N_j & \text{per } j = 2, \ldots, 10 \\ N_{11} + N_{12} & \text{per } j = 11 \end{cases}$$

A questo punto, con $m = 11$*, si calcola la statistica di Pearson (9.41) che vale*

$$K_0 = \sum_{j=1}^{m} \frac{(M_j - nq_j)^2}{nq_j} \simeq 242.05$$

e siccome dalle Tavole E.3 il corrispondente quantile del chi-quadro è

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.95}(10) = 18.31 \ll 242.05 = K_0$$

il test di livello $\alpha = 0.05$ *decisamente rigetta l'ipotesi che la distribuzione empirica si adatti alla distribuzione binomiale* $\mathfrak{B}(12;\,{}^1\!/_2)$

Come secondo tentativo possiamo allora provare a verificare un'ipotesi $\mathcal{H}_0$ *modificata – già avanzata nell'Esempio 8.8 – secondo la quale la distribuzione di* X *è binomiale* $\mathfrak{B}(12;\,\overline{p})$*, con* $\overline{p} \simeq 0.519$ *stima di MV di* p *ottenuta a partire dai dati del campione. Il valore di* $\overline{p}$ *è stato calcolato nell'Esempio 8.8 mediante lo stimatore di MV coincidente con la media dei valori empirici di* X*, ma in realtà, per essere*

rigorosi, qui dovremmo rideterminare lo stimatore di MV perché abbiamo modificato il modello raggruppando le classi estreme. Siccome però la correzione sarebbe piccola e non capovolgerebbe il risultato del test, noi assumeremo che la vecchia stima $\overline{p} = 0.519$ *possa essere tranquillamente usata anche ora come stima di MV del parametro* p*. A questo punto dovremo seguire la stessa procedura già seguita in precedenza, ma con due modifiche che qui riassumiamo senza fornire i dettagli numerici*

- *prima di tutto bisogna ricalcolare le* p_k *e le* q_j *tenendo conto del fatto che ora il parametro della legge binomiale vale* 0.519*, e non* $^1/_2 = 0.500$
- *in secondo luogo bisogna calcolare di nuovo il valore della statistica* K_0 *di Pearson (9.41) con i nuovi parametri, ma dovremo confrontarlo con il quantile* $\chi^2_{1-\alpha}(m-1-1) = \chi^2_{0.95}(9)$ *dato che siamo stati obbligati a stimare un parametro* $(q = 1)$ *per definire la distribuzione teorica*

Sebbene in questo secondo calcolo il valore della statistica di Pearson sia più che dimezzato, l'esito del test di livello $\alpha = 0.05$

$$\chi^2_{1-\alpha}(m-1-1) = \chi^2_{0.95}(9) = 16.92 < 105.79 \simeq K_0$$

chiaramente indica ancora una volta che la distribuzione empirica non può essere considerata ben adattata neanche alla distribuzione binomiale $\mathfrak{B}\,(12;\, 0.519)$*. Va infine notato che l'esito del test appare piuttosto affidabile nel senso che non sarebbe modificato se scegliessimo valori diversi, ma ragionevoli, per il livello* α

Esempio 9.10. (Approssimazione di Poisson) *Supponiamo di voler verificare empiricamente il Teorema 5.8 di Poisson secondo il quale, in particolari condizioni, una legge binomiale può essere ben approssimata da una legge di Poisson. Per far questo simuliamo un campione di* $n = 2\,000$ *numeri estratti da una v-a* X *Binomiale* $\mathfrak{B}\,(200;\, 0.01)$*, e siccome*

$$\lambda = n\,p = 200 \times 0.01 = 2$$

confrontiamo le frequenze empiriche trovate con le probabilità teoriche della legge di Poisson $\mathfrak{P}(2)$

$$p_k = e^{-2}\,\frac{2^k}{k!} \qquad k = 0, 1, \ldots$$

I valori N_k *delle frequenze assolute del campione sono riportati nella Tabella 9.4*

Osserviamo allora preliminarmente che i possibili valori interi di una v-a X *di Poisson sono in numero infinito, mentre il test del* χ^2 *è stato progettato solo per un numero finito* m *di valori. Inoltre per la legge* $\mathfrak{P}(2)$ *da noi considerata si ha*

$$2\,000 \times p_0 = 270.67, \quad \ldots \quad 2\,000 \times p_7 = 6.87, \quad 2\,000 \times p_8 = 1.72\,, \quad \ldots$$

per cui la condizione di applicabilità del test $(np_k \geq 5)$ *è verificata sempre tranne che per* $k \geq 7$*. Questi due problemi possono essere superati assieme raggruppando*

k	0	1	2	3	4	5	6	7	8
N_k	280	545	544	355	186	55	25	9	1

Tabella 9.4: Frequenze di un campione simulato di 2 000 numeri estratti secondo la legge binomiale $\mathfrak{B}\,(200;\,0.01)$

classi	0	1	2	3	4	5	6	≥ 7
M_j	280	545	544	355	186	55	25	10

Tabella 9.5: Rispetto alla Tabella 9.4 le frequenze dei valori 7 e 8 sono state unificate in un'unica classe ≥ 7

tutti i valori $k \geq 7$ in un'unica classe. In questo modo non solo avremo un numero $m = 8$ finito di classi di valori con le probabilità derivate da quelle di Poisson

$$q_0 = p_0 \qquad q_1 = p_1 \quad \dots \quad q_6 = p_6 \qquad q_7 = \sum_{k\geq 7} p_k = 1 - \sum_{k=0}^{6} p_k$$

ma anche la condizione di applicabilità del test sarà ora sempre rispettata perché

$$nq_7 = 2000\,(1 - p_0 - \,\dots - p_6) \simeq 9.07$$

Naturalmente anche le frequenze empiriche N_k dovranno essere raggruppate nello stesso modo, come riportato nella Tabella 9.5

A questo punto con $m = 8$, con le probabilità q_j e le frequenze M_j possiamo calcolare la statistica di Pearson (9.41) e confrontarla con il quantile del chi-quadro ottenendo

$$K_0 \simeq 4.85 < 14.07 \simeq \chi^2_{0.95}(7) = \chi^2_{1-\alpha}(m-1)$$

Pertanto, siccome i dati non si trovano in regione critica, il test di livello $\alpha = 0.05$ suggerisce di accettare l'ipotesi $\mathcal{H}_0$ secondo la quale il campione segue una legge di Poisson $\mathfrak{P}(2)$

Esempio 9.11. (Approssimazione normale) *Proviamo ora a controllare empiricamente le conclusioni del TLC, Teorema 5.5, secondo il quale somme standardizzate di un numero abbastanza grane di v-a indipendenti sono approssimativamente distribuite secondo la legge normale standard $\mathfrak{N}\,(0,1)$. In particolare, con le notazioni usate nel Teorema 5.5, considereremo la somma standardizzata Y^*_{30} di 30 v-a indipendenti e uniformi $\mathfrak{U}\,(0,1)$, ed eseguiremo un test del chi-quadro per verificare l'ipotesi $\mathcal{H}_0$ che Y^*_{30} è approssimativamente normale standard $\mathfrak{N}\,(0,1)$*

A questo scopo simuliamo un campione di $n = 1\,000$ valori di Y^*_{30}: siccome si tratta di una v-a continua, l'applicazione del test del χ^2 richiede preventivamente la definizione degli classi nelle quali vanno calcolate le frequenze empiriche. Naturalmente la scelta va operata tenendo conto delle usuali condizioni di applicabilità del

classi	≤ -2.0	[-2.0, -1.5]	[-1.5, -1.0]	[-1.0, -0.5]	[-0.5, 0.0]
N_j	27	53	100	147	191
classi	$[0.0, 0.5]$	$[0.5, 1.0]$	$[1.0, 1.5]$	$[1.5, 2.0]$	≥ 2.0
N_j	183	143	79	38	39

Tabella 9.6: Frequenze di un campione di $n = 1\,000$ valori di Y_{30}^*, somma standardizzata di 30 *v-a* uniformi $\mathfrak{U}(0,1)$ indipendenti

test ($np_k \geq 5$). Osserviamo allora che per una legge normale standard $\mathfrak{N}(0,1)$ dalle Tavole E.1 si ha ad esempio

$$\begin{aligned} n\,\boldsymbol{P}\{\mathfrak{N}(0,1) \leq -2\} &= 1\,000 \times \Phi(-2) = 1\,000 \times [1 - \Phi(2)] = 22.75 > 5 \\ n\,\boldsymbol{P}\{\mathfrak{N}(0,1) \geq -2\} &= n\,(1 - \boldsymbol{P}\{\mathfrak{N}(0,1) \leq 2\}) = 1\,000 \times [1 - \Phi(2)] = 22.75 > 5 \end{aligned}$$

e analogamente si prova che le condizioni sono rispettate da tutti gli intervalli di ampiezza 0.5 presi fra -2 e 2; così ad esempio

$$n\,\boldsymbol{P}\{1.5 \leq \mathfrak{N}(0,1) \leq 2.0\} = 1\,000 \times [\Phi(2.0) - \Phi(1.5)] = 44.06 > 5$$

Definiremo allora $m = 10$ classi utilizzando gli intervalli

$$(-\infty,\, -2.0] \quad [-2.0,\, -1.5] \quad \ldots \quad [0.0,\, 0.5] \quad \ldots \quad [1.5,\, 2.0] \quad [2.0,\, +\infty)$$

con le corrispondenti probabilità teoriche

$$p_j = \begin{cases} \boldsymbol{P}\{\mathfrak{N}(0,1) \leq -2\} = 1 - \Phi(2) & j = 1 \\ \boldsymbol{P}\left\{\frac{j-6}{2} \leq \mathfrak{N}(0,1) \leq \frac{j-5}{2}\right\} = \Phi\left(\frac{j-5}{2}\right) - \Phi\left(\frac{j-6}{2}\right) & j = 2, \ldots, 9 \\ \boldsymbol{P}\{\mathfrak{N}(0,1) > 2\} = 1 - \Phi(2) & j = 10 \end{cases}$$

i cui valori numerici possono essere ricavati usando le Tavole E.1. Le frequenze empiriche nelle 10 classi del campione simulato sono poi riportate nella Tabella 9.6

A questo punto siamo in grado di calcolare la statistica di Pearson (9.41) e di confrontarla con il quantile del chi-quadro $\boldsymbol{\chi^2}(m-1) = \boldsymbol{\chi^2}(9)$ ottenendo per un test di livello $\alpha = 0.01$

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.99}(9) = 21.67 > 18.32 \simeq K_0$$

I nostri dati non sono quindi in regione critica, e il test suggerisce di accettare l'ipotesi $\mathcal{H}_0$ secondo la quale il campione delle somme standardizzate segue una legge normale standard. Le conclusioni di questo test non sono però particolarmente solide: infatti, se avessimo scelto un livello $\alpha = 0.05$, dalla relazione

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.95}(9) = 16.92 < 18.32 \simeq K_0$$

dagli stessi dati empirici avremmo ricavato per $\mathcal{H}_0$ la conclusione opposta. Naturalmente questo non indica una debolezza del TLC Teorema 5.5, quanto piuttosto una parziale inadeguatezza dell'esperimento di simulazione. Infatti la legge della somma di appena 30 v-a indipendenti può essere approssimata da una normale standard soltanto in maniera un po' rudimentale (vedi le osservazioni in Sezione 5.3)

9.6 Test del χ^2 di indipendenza

Una versione particolare del test del χ^2 di adattamento viene usata come test di indipendenza di due campioni accoppiati $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$ estratti da due *v-a* X e Y. In questo caso le ipotesi da esaminare prendono la forma

$$\begin{aligned} \mathcal{H}_0 &: \text{ i due campioni sono indipendenti} \\ \mathcal{H}_1 &: \text{ i due campioni non sono indipendenti} \end{aligned}$$

Supponendo ora che le nostre due *v-a* possano assumere solo un numero finito di valori (anche di tipo qualitativo), rispettivamente $u_1, \ldots, u_r$ e $v_1, \ldots, v_s$, e riprendendo una notazione già introdotta nella Sezione 7.1, indicheremo con $N_{j\,k}$ la frequenza con la quale compare nel campione la coppia di valori (u_j, v_k); con $N_{j\,\cdot}$ e $N_{\cdot\,k}$ denoteremo invece le frequenze con le quali si presentano separatamente, rispettivamente i valori u_j e v_k. Le probabilità teoriche congiunte p_{jk}, e marginali p_j e q_k associate ai valori delle *v-a* X e Y non sono note, ma possono essere stimate tramite i loro stimatori *MV*, cioè tramite le frequenze relative empiriche congiunte e marginali

$$\overline{p}_{jk} = \frac{N_{jk}}{n} \qquad \overline{p}_j = \frac{N_{j\,\cdot}}{n} \qquad \overline{q}_k = \frac{N_{\cdot\,k}}{n} \qquad \begin{cases} j = 1, \ldots, r \\ k = 1, \ldots, s \end{cases}$$

Ricorderemo ora da (3.35) che l'indipendenza di due *v-a* X e Y equivale alla richiesta che fra le loro probabilità teoriche congiunte e marginali sussista la relazione

$$p_{j\,k} = p_j q_k$$

Questo suggerisce allora di progettare un test di indipendenza confrontando le frequenze congiunte empiriche $\overline{p}_{jk}$ con i prodotti $\overline{p}_j \overline{q}_k$ delle loro marginali. A questo scopo si dimostra che, quando l'ipotesi $\mathcal{H}_0$ è vera, la statistica di Pearson

$$\boxed{K_0 = \sum_{j,k} \frac{n\,(\overline{p}_{jk} - \overline{p}_j \overline{q}_k)^2}{\overline{p}_j \overline{q}_k} = \sum_{j,k} \frac{(N_{jk} - n\overline{p}_j \overline{q}_k)^2}{n\overline{p}_j \overline{q}_k} \sim \boldsymbol{\chi^2}\,[(r-1)(s-1)]} \tag{9.44}$$

segue una legge $\boldsymbol{\chi^2}\,[(r-1)(s-1)]$ con $(r-1)(s-1)$ gradi di libertà, se le frequenze relative marginali non sono troppo piccole. Pertanto in questo caso l'**evento critico** di livello α è

$$\boxed{D = \left\{K_0 > \chi^2_{1-\alpha}[(r-1)(s-1)]\right\}} \tag{9.45}$$

In pratica la statistica di Pearson K_0 è una misura globale della differenza fra le frequenze congiunte $\overline{p}_{jk}$ e i prodotti $\overline{p}_j \overline{q}_k$ delle loro marginali: se tale differenza è troppo grande il test suggerirà di non accettare l'ipotesi $\mathcal{H}_0$ di indipendenza, e viceversa in caso contrario

		Fis			
		alti	*medi*	*bassi*	marginali
	alti	56 *30.80*	71 *72.66*	12 *35.54*	139
Mat	*medi*	47 *54.95*	163 *129.64*	38 *63.41*	248
	bassi	14 *31.24*	42 *73.70*	85 *36.05*	141
	marginali	117	276	135	**528**

Tabella 9.7: Frequenze dei voti in fisica e matematica di $n = 528$ studenti

La **procedura** per il test è quindi la seguente

> *Si costruisce la tabella di contingenza delle frequenze congiunte e marginali dei due campioni e da queste si calcola la statistica K_0 (9.44): se $K_0 \leq \chi^2_{1-\alpha}[(r-1)(s-1)]$ si accetta l'ipotesi $\mathcal{H}_0$ di indipendenza, e la si rifiuta invece quando $K_0 > \chi^2_{1-\alpha}[(r-1)(s-1)]$*

Questo test di indipendenza può essere applicato anche per *v-a* continue (o comunque con un numero infinito di valori) raggruppando gli elementi dei campioni di X e Y in un piccolo numero di classi che permettano di costruire una tabella di contingenza

Esempio 9.12. *Siano date, nelle prime tre righe centrali della Tabella 9.7, le frequenze congiunte dei voti (raggruppati in tre classi: alti, medi e bassi) riportati in esami di Fisica e Matematica da $n = 528$ studenti. Si vuole esaminare l'ipotesi $\mathcal{H}_0$ secondo la quale i risultati degli esami in Fisica e Matematica sono indipendenti*

Innanzitutto bisogna completare la tabella delle frequenze congiunte $N_{j\,k}$ con le frequenze marginali $N_{j\,\cdot}$ e $N_{\cdot\,k}$ ottenute sommando lungo le righe e lungo le colonne, e con il valore del numero totale delle misure n. Nella stessa tabella, poi, sotto ogni valore delle frequenze congiunte è stato riportato il valore delle quantità $n\overline{p}_j\overline{q}_k$ che servono per calcolare la statistica K_0 (9.44). Siccome ci sono $r = s = 3$ classi per ambedue le variabili, il test del χ^2 di livello $\alpha = 0.05$ consiste semplicemente nell'osservare dalle Tavole E.3 che

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.95}(4) = 9.49 < 145.78 \simeq K_0$$

cioè che i dati del campione sono in regione critica, per cui la conclusione è che al livello $\alpha = 0.05$ si rifiuta l'ipotesi $\mathcal{H}_0$ di indipendenza fra i voti di Fisica e Matematica. Il test appare piuttosto affidabile visto che la conclusione non cambia se si passa al livello $\alpha = 0.01$: infatti

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.99}(4) = 13.28 < 145.78 \simeq K_0$$

		Età					
		18–30	31–40	41–50	51–60	61–70	marginali
Incidenti	0	748 *770.13*	821 *818.37*	786 *772.81*	720 *720.99*	672 *664.70*	3 747
	1	74 *61.87*	60 *65.74*	51 *62.08*	66 *57.92*	50 *53.40*	301
	2	31 *22.40*	25 *23.81*	22 *22.48*	16 *20.97*	15 *19.34*	109
	> 2	9 *7.60*	10 *8.08*	6 *7.63*	5 *7.12*	7 *6.56*	37
	marginali	862	916	865	807	744	**4 194**

Tabella 9.8: Numero di incidenti per classi di età di $n = 4\,194$ guidatori

Esempio 9.13. *Siano dati, nelle prime righe della parte centrale della Tabella 9.8, i risultati di un'inchiesta volta a stabilire se si può accettare l'ipotesi $\mathcal{H}_0$ che l'età di un guidatore (divisa in $r = 5$ classi) è indipendente dal numero degli incidenti automobilistici provocati (diviso in $s = 4$ classi)*

Una volta completata la tabella con le marginali e il numero totale $n = 4\,194$ di soggetti esaminati, si calcolano i valori delle quantità $n\overline{p}_j\overline{q}_k$ e li si riporta sotto le frequenze congiunte. Possiamo ora calcolare la statistica K_0 (9.44), e per un test di livello $\alpha = 0.05$ dalle Tavole E.3 si ha

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.95}(12) = 21.03 > 14.40 \simeq K_0$$

mentre per un test di livello $\alpha = 0.01$ si ha

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.99}(12) = 26.22 > 14.40 \simeq K_0$$

Pertanto in ambedue i casi il test suggerisce di accettare l'ipotesi $\mathcal{H}_0$ di indipendenza fra età e numero di incidenti. Va però detto che qualche dubbio sulla affidabilità del test deve essere avanzato a causa della prossimità dei valori di K_0 e di $\chi^2_{1-\alpha}$, soprattutto perché siamo in presenza di una tabella di contingenza con casi di frequenze (< 10) piuttosto piccole

Parte III

Appendici

Appendice A

Esercizi

A.1 Calcolo delle probabilità

Esercizio A.1.1. *Per incoraggiare la carriera tennistica di Mario suo padre gli promette un premio se egli riesce a vincere almeno due partite di seguito in una serie di tre partite giocate alternativamente con suo padre (P) e con il campione della loro associazione (C). Mario può scegliere di iniziare la serie con suo padre (PCP)o con il campione (CPC). Sapendo che il campione gioca meglio del padre, quale successione di partite gli converrà scegliere?*

Soluzione: Le due possibilità presentano vantaggi opposti: nel caso PCP Mario giocherebbe due volte con l'avversario meno abile; nel caso CPC l'avversario meno abile si trova in posizione centrale (separando così due eventuali vittorie contro il padre). Supponendo allora di indicare con p la probabilità che Mario batta suo padre e con c quella che egli batta il campione (ovviamente $p > c$), e supponendo che le vittorie (V) e le sconfitte (S) si succedano in maniera indipendente, le probabilità dei tre casi favorevoli a Mario sono le seguenti

partite	VVV	VVS	SVV
PCP	pcp	$pc(1-p)$	$(1-p)cp$
CPC	cpc	$cp(1-c)$	$(1-c)pc$

Si ricava pertanto che le probabilità di vincere il premio nei due casi sono

PCP	$pc(2-p)$
CPC	$pc(2-c)$

e pertanto, siccome $p > c$, si ha $pc(2-c) > pc(2-p)$, per cui giocare due volte col campione, ma non di seguito (CPC) risulta più favorevole che giocare una volta sola col campione, ma la seconda partita (PCP) □

Esercizio A.1.2. *Un cubo con le facce colorate viene diviso in* 1 000 *cubetti di eguali dimensioni. I cubetti così ottenuti sono poi mescolati: determinare la probabilità che un cubetto estratto a caso abbia due facce colorate*

Soluzione: Tra i 1 000 cubetti ve ne sono 8 con tre facce colorate, $8 \cdot 12 = 96$ con due facce colorate, $64 \cdot 6 = 384$ con una faccia colorata e $8^3 = 512$ non colorati. Pertanto

$$p = \frac{96}{1\,000} = 0,096$$

è la probabilità richiesta □

Esercizio A.1.3. *Un cassetto contiene calzini rossi e bianchi e si sa che, se si estraggono a caso due calzini, la probabilità che siano ambedue rossi è* $^1/_2$*. Qual è il più piccolo numero di calzini (bianchi e rossi) che rende possibile questa situazione? Qual è la risposta alla domanda precedente se si suppone che il numero dei calzini bianchi sia pari?*

Soluzione: Se r è il numero di calzini rossi e b il numero di quelli bianchi, le condizioni del problema impongono che

$$\frac{r(r-1)}{(r+b)(r+b-1)} = \frac{1}{2}$$

cioè che

$$r^2 - (2b+1)r - b(b-1) = 0$$

Ne segue che tra il numero di calzini rossi ed il numero di calzini bianchi intercorre la relazione

$$r = \frac{(2b+1) \pm \sqrt{8b^2+1}}{2}$$

Possiamo ora affrontare il problema per tentativi ricordando che, naturalmente, r e b devono essere numeri interi e che il cassetto deve contenere almeno due calzini

b	r	
0	0 e 1	i calzini sono meno di 2
1	0 e 3	*risposta alla prima domanda*
2	$\frac{5\pm\sqrt{33}}{2}$	numeri non interi
3	$\frac{7\pm\sqrt{73}}{2}$	numeri non interi
4	$\frac{9\pm\sqrt{129}}{2}$	numeri non interi
5	$\frac{11\pm\sqrt{201}}{2}$	numeri non interi
6	-2 e 15	*risposta alla seconda domanda*

Pertanto la risposta alla prima domanda è $b = 1$, $r = 3$, mentre la risposta alla seconda domanda è $b = 6$, $r = 15$ □

Esercizio A.1.4. *Una v-a X ha fdp $f_A(x)$ se si verifica l'evento A, e fdp $f_B(x)$ se si verifica l'evento B; inoltre $\boldsymbol{P}\{A\} = p$, e $\boldsymbol{P}\{B\} = q$, con $p + q = 1$.*

- *Determinare la fdp $f(x)$ della v-a X*
- *Supponendo che le leggi con densità f_A ed f_B abbiano rispettivamente attese μ_A ed μ_B, e varianze σ_A^2 e σ_B^2, calcolare l'attesa μ e la varianza σ^2 di X*

Soluzione: Definendo le *FDC* dei due casi

$$\begin{aligned} F_A(x) &= \boldsymbol{P}\{X \le x \mid A\} = \int_{-\infty}^{x} f_A(y)\,dy \qquad & f_A(x) = F_A'(x) \\ F_B(x) &= \boldsymbol{P}\{X \le x \mid B\} = \int_{-\infty}^{x} f_B(y)\,dy \qquad & f_B(x) = F_B'(x) \end{aligned}$$

avremo innanzitutto dalla Formula della probabilità totale (2.1)

$$\begin{aligned} F(x) &= \boldsymbol{P}\{X \le x\} = \boldsymbol{P}\{X \le x \mid A\}\,\boldsymbol{P}\{A\} + \boldsymbol{P}\{X \le x \mid B\}\,\boldsymbol{P}\{B\} \\ &= pF_A(x) + qF_B(x) \\ f(x) &= F'(x) = pF_A'(x) + qF_B'(x) = pf_A(x) + qf_B(x) \end{aligned}$$

che fornisce la risposta alla prima domanda. Si ricava poi facilmente che

$$\mu = \boldsymbol{E}[X] = \int_{-\infty}^{+\infty} x f(x)\,dx = p\int_{-\infty}^{+\infty} x f_A(x)\,dx + q\int_{-\infty}^{+\infty} x f_B(x)\,dx = p\mu_A + q\mu_B$$

mentre per la varianza bisogna prima osservare che

$$\begin{aligned}\boldsymbol{E}\left[X^2\right] &= \int_{-\infty}^{+\infty} x^2 f(x)\,dx = p\int_{-\infty}^{+\infty} x^2 f_A(x)\,dx + q\int_{-\infty}^{+\infty} x^2 f_B(x)\,dx \\ &= p(\mu_A^2+\sigma_A^2) + q(\mu_B^2+\sigma_B^2)\end{aligned}$$

e poi che

$$\begin{aligned}\boldsymbol{V}[X] &= \boldsymbol{E}\left[X^2\right] - \boldsymbol{E}[X]^2 = p(\mu_A^2+\sigma_A^2) + q(\mu_B^2+\sigma_B^2) - (p\mu_A+q\mu_B)^2 \\ &= p\sigma_A^2 + q\sigma_B^2 + pq(\mu_A-\mu_B)^2\end{aligned}$$

che completa anche la risposta alla seconda domanda □

Esercizio A.1.5. *Due v-a X ed Y sono legate dalla relazione $Y = 2-3X$. Sapendo che $\boldsymbol{E}[X] = -1$ e $\boldsymbol{V}[X] = 4$, calcolare l'attesa e la varianza di Y, la covarianza e il coefficiente di correlazione delle due v-a*

Soluzione: Da (4.8) e (4.19) avremo innanzitutto

$$\begin{aligned}\boldsymbol{E}[Y] &= \boldsymbol{E}[2-3X] = 2-3\boldsymbol{E}[X] = 5 \\ \boldsymbol{V}[Y] &= \boldsymbol{V}[2-3X] = 9\boldsymbol{V}[X] = 36\end{aligned}$$

Per la covarianza da (4.15) risulta

$$\begin{aligned}\boldsymbol{cov}[X,Y] &= \boldsymbol{E}[XY] - \boldsymbol{E}[X]\,\boldsymbol{E}[Y] = \boldsymbol{E}[X(2-3X)] + 5 \\ &= 2\boldsymbol{E}[X] - 3\boldsymbol{E}\left[X^2\right] + 5 = 3 - 3\boldsymbol{E}\left[X^2\right]\end{aligned}$$

e d'altra parte da (4.16) si ha

$$\boldsymbol{E}\left[X^2\right] = \boldsymbol{V}[X] + \boldsymbol{E}[X]^2 = 5$$

per cui in definitiva

$$\boldsymbol{cov}[X,Y] = 3 - 3\boldsymbol{E}\left[X^2\right] = -12$$

Per il coefficiente di correlazione (4.13) si ha infine dai risultati precedenti

$$\rho_{XY} = \frac{\boldsymbol{cov}[X,Y]}{\sqrt{\boldsymbol{V}[X]}\sqrt{\boldsymbol{V}[Y]}} = -1$$

come peraltro prevedibile sulla base del Teorema 4.7 data la dipendenza lineare sussistente fra X e Y □

Esercizio A.1.6. *Un telegrafo trasmette* punti *e* linee*, ed è noto che la frequenza dei* punti *al momento della trasmissione è* $^5/_8$ *mentre quella delle* linee *è* $^3/_8$*. Disturbi di trasmissione modificano* $^2/_5$ *dei* punti *in* linee *ed* $^1/_3$ *delle* linee *in* punti*. Supponendo di ricevere un messaggio, calcolare la probabilità che il segnale ricevuto sia proprio quello trasmesso, sia nel caso in cui si riceve un* punto *sia in quello in cui si riceve una* linea

Soluzione: Per risolvere questo problema con il Teorema di Bayes 2.5 conviene definire gli eventi

$$\begin{array}{lcllcl} R_P &=& \text{ricevo } \textit{punto} & R_L &=& \text{ricevo } \textit{linea} \\ T_P &=& \text{è stato trasmesso } \textit{punto} & T_L &=& \text{è stata trasmessa } \textit{linea} \end{array}$$

e osservare che in base ai dati del problema si ha

$$\begin{array}{lcllcllcl} \boldsymbol{P}\{T_P\} &=& {}^5/_8 & \boldsymbol{P}\{R_P|\,T_P\} &=& {}^3/_5 & \boldsymbol{P}\{R_L|\,T_P\} &=& {}^2/_5 \\ \boldsymbol{P}\{T_L\} &=& {}^3/_8 & \boldsymbol{P}\{R_P|\,T_L\} &=& {}^1/_3 & \boldsymbol{P}\{R_L|\,T_L\} &=& {}^2/_3 \end{array}$$

D'altra parte dalla formula della probabilità totale (2.1) risulta

$$\begin{aligned} \boldsymbol{P}\{R_P\} &= \boldsymbol{P}\{R_P \mid T_P\}\,\boldsymbol{P}\{T_P\} + \boldsymbol{P}\{R_P \mid T_L\}\,\boldsymbol{P}\{T_L\} = {}^1/_2 \\ \boldsymbol{P}\{R_L\} &= \boldsymbol{P}\{R_L \mid T_P\}\,\boldsymbol{P}\{T_P\} + \boldsymbol{P}\{R_L \mid T_L\}\,\boldsymbol{P}\{T_L\} = {}^1/_2 \end{aligned}$$

e quindi la formula di Bayes ci consente di ottenere

$$\begin{aligned} \boldsymbol{P}\{T_P \mid R_P\} &= \frac{\boldsymbol{P}\{R_P \mid T_P\}\,\boldsymbol{P}\{T_P\}}{\boldsymbol{P}\{R_P\}} = \frac{3}{4} \\ \boldsymbol{P}\{T_L \mid R_L\} &= \frac{\boldsymbol{P}\{R_L \mid T_L\}\,\boldsymbol{P}\{T_L\}}{\boldsymbol{P}\{R_L\}} = \frac{1}{2} \end{aligned}$$

cioè la ricezione del segno *punto* è più affidabile della ricezione del segno *linea* □

Esercizio A.1.7. *Un segnale arriva nell'intervallo di tempo* $[0,\tau]$ *con probabilità* p*. Se esso arriva in tale intervallo l'istante di arrivo è distribuito uniformemente in* $[0,\tau]$*. Preso un istante* t *con* $0 \le t \le \tau$*, e definiti gli eventi (vedi Figura A.1)*

$$\begin{array}{lcl} A &=& \textit{il segnale arriva in } [0,\tau] \\ B &=& \textit{il segnale arriva in } [0,t] \\ C &=& \textit{il segnale arriva in } [t,\tau] \end{array}$$

calcolare la probabilità che il segnale arrivi in $[t,\tau]$ *supponendo che esso non sia arrivato in* $[0,t]$

Soluzione: La traccia del problema ci dice che

$$\boldsymbol{P}\{A\} = p \qquad\qquad \boldsymbol{P}\{\overline{A}\} = 1-p$$

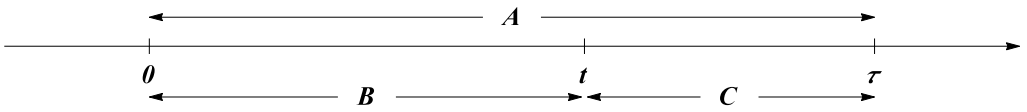


Figura A.1: Esercizio A.1.7

e ci richiede di calcolare $\boldsymbol{P}\{C \mid \overline{B}\}$: d'altra parte, siccome ovviamente $C \subseteq \overline{B}$, questo si riduce a calcolare

$$\boldsymbol{P}\{C \mid \overline{B}\} = \frac{\boldsymbol{P}\{C \cap \overline{B}\}}{\boldsymbol{P}\{\overline{B}\}} = \frac{\boldsymbol{P}\{C\}}{\boldsymbol{P}\{\overline{B}\}}$$

Inoltre, siccome $\overline{B} \cap A = C$ e $\overline{B} \cap \overline{A} = \overline{A}$, e siccome l'istante di arrivo in $[0, \tau]$ è distribuito uniformemente, abbiamo anche (vedi Figura A.1)

$$\begin{aligned} \boldsymbol{P}\{C \mid A\} &= \frac{\tau - t}{\tau} & \boldsymbol{P}\{C \mid \overline{A}\} &= 0 \\ \boldsymbol{P}\{\overline{B} \mid A\} &= \frac{\tau - t}{\tau} & \boldsymbol{P}\{\overline{B} \mid \overline{A}\} &= 1 \end{aligned}$$

Dalla formula della probabilità totale (2.1) abbiamo allora

$$\begin{aligned} \boldsymbol{P}\{C\} &= \boldsymbol{P}\{C \mid A\}\,\boldsymbol{P}\{A\} + \boldsymbol{P}\{C \mid \overline{A}\}\,\boldsymbol{P}\{\overline{A}\} = p\,\frac{\tau - t}{\tau} \\ \boldsymbol{P}\{\overline{B}\} &= \boldsymbol{P}\{\overline{B} \mid A\}\,\boldsymbol{P}\{A\} + \boldsymbol{P}\{\overline{B} \mid \overline{A}\}\,\boldsymbol{P}\{\overline{A}\} = p\,\frac{\tau - t}{\tau} + 1 - p \end{aligned}$$

e quindi in definitiva

$$\boldsymbol{P}\{C \mid \overline{B}\} = \frac{\boldsymbol{P}\{C\}}{\boldsymbol{P}\{\overline{B}\}} = \frac{p\tau - pt}{\tau - pt}$$

è la probabilità richiesta □

Esercizio A.1.8. *Quando le condizioni meteorologiche sono favorevoli (H) e il pilota può vedere la pista, un aereo atterra felicemente (A) con probabilità $\boldsymbol{P}\{A|H\} = p$. Se invece le condizioni meteorologiche non sono favorevoli ($\overline{H}$) e impediscono di vedere la pista il pilota deve eseguire un atterraggio strumentale. In tal caso l'affidabilità degli strumenti – cioè la probabilità di funzionare correttamente (T) in condizioni di tempo sfavorevole – è $\boldsymbol{P}\{T \mid \overline{H}\} = q$, e se gli strumenti funzionano correttamente la probabilità di un atterraggio felice resta invariata $\boldsymbol{P}\{A \mid T \cap \overline{H}\} = p$; se invece gli strumenti danno problemi ($\overline{T}$) la probabilità di un atterraggio felice è $\boldsymbol{P}\{A \mid \overline{T} \cap \overline{H}\} = p^\star < p$*

- *Sapendo che le condizioni meteorologiche sono favorevoli con una probabilità $\boldsymbol{P}\{H\} = s$, determinare la probabilità totale $\boldsymbol{P}\{A\}$ di un atterraggio felice*

- *Supponendo che un aereo sia atterrato felicemente, determinare la probabilità* $\boldsymbol{P}\{\overline{H} \mid A\}$ *che il pilota sia stato costretto dalle sfavorevoli condizioni meteorologiche ad un atterraggio strumentale*

Soluzione: Riassumendo le indicazioni della traccia, abbiamo

$$\boldsymbol{P}\{H\} = s \qquad \boldsymbol{P}\{\overline{H}\} = 1 - s$$
$$\boldsymbol{P}\{A \mid H\} = p \qquad \boldsymbol{P}\{T \mid \overline{H}\} = q \qquad \boldsymbol{P}\{\overline{T} \mid \overline{H}\} = 1 - q$$
$$\boldsymbol{P}\{A \mid T \cap \overline{H}\} = p \qquad \boldsymbol{P}\{A \mid \overline{T} \cap \overline{H}\} = p^\star < p$$

e da queste possiamo anche ricavare

$$\begin{aligned}
\boldsymbol{P}\{A \mid \overline{H}\} &= \frac{\boldsymbol{P}\{A \cap \overline{H}\}}{\boldsymbol{P}\{\overline{H}\}} = \frac{\boldsymbol{P}\{A \cap T \cap \overline{H}\} + \boldsymbol{P}\{A \cap \overline{T} \cap \overline{H}\}}{\boldsymbol{P}\{\overline{H}\}} \\
&= \frac{\boldsymbol{P}\{A \cap T \cap \overline{H}\}}{\boldsymbol{P}\{T \cap \overline{H}\}} \frac{\boldsymbol{P}\{T \cap \overline{H}\}}{\boldsymbol{P}\{\overline{H}\}} + \frac{\boldsymbol{P}\{A \cap \overline{T} \cap \overline{H}\}}{\boldsymbol{P}\{\overline{T} \cap \overline{H}\}} \frac{\boldsymbol{P}\{\overline{T} \cap \overline{H}\}}{\boldsymbol{P}\{\overline{H}\}} \\
&= \boldsymbol{P}\{A \mid T \cap \overline{H}\}\, \boldsymbol{P}\{T \mid \overline{H}\} + \boldsymbol{P}\{A \mid \overline{T} \cap \overline{H}\}\, \boldsymbol{P}\{\overline{T} \mid \overline{H}\} \\
&= pq + p^\star(1 - q)
\end{aligned}$$

Per la prima domanda si ha allora dalla formula della probabilità totale (2.1)

$$\boldsymbol{P}\{A\} = \boldsymbol{P}\{A \mid H\}\, \boldsymbol{P}\{H\} + \boldsymbol{P}\{A \mid \overline{H}\}\, \boldsymbol{P}\{\overline{H}\} = ps + \big(pq + p^\star(1 - q)\big)(1 - s)$$

Quanto alla seconda domanda basterà poi usare il Teorema di Bayes (2.2)

$$\boldsymbol{P}\{\overline{H} \mid A\} = \frac{\boldsymbol{P}\{A \mid \overline{H}\}\, \boldsymbol{P}\{\overline{H}\}}{\boldsymbol{P}\{A\}} = \frac{\big(pq + p^\star(1 - q)\big)(1 - s)}{ps + \big(pq + p^\star(1 - q)\big)(1 - s)}$$

per ottenere il risultato richiesto □

Esercizio A.1.9. *Due scatole esternamente identiche contengono ciascuna 24 palline di colore bianco e nero, ma con diverse composizioni:*

$$\begin{cases} \textit{scatola 1}: & \textit{8 bianche, 16 nere} \\ \textit{scatola 2}: & \textit{18 bianche, 6 nere} \end{cases}$$

Si prende una scatola a caso, nel senso che con

$$D_1 = \textit{ho preso la scatola 1} \qquad D_2 = \textit{ho preso la scatola 2}$$

le probabilità a priori sono

$$\boldsymbol{P}\{D_1\} = \boldsymbol{P}\{D_2\} = {}^1\!/_2$$

e si estraggono successivamente (rimettendole nella scatola dopo ogni estrazione) 10 palline ottenendo il seguente risultato

$$B = 4 \textit{ bianche}, 6 \textit{ nere}$$

Calcolare le probabilità a posteriori $\boldsymbol{P}\{D_1|B\}$ *e* $\boldsymbol{P}\{D_2|B\}$

Soluzione: Dalla composizione delle due scatole sappiamo che le probabilità di estrarre una pallina bianca sono rispettivamente

$$p = \begin{cases} {}^{8}/_{24} = {}^{1}/_{3} & \text{per la scatola 1 } (D_1) \\ {}^{18}/_{24} = {}^{3}/_{4} & \text{per la scatola 2 } (D_2) \end{cases}$$

per cui dalla (2.4) del Teorema 2.7 si ha

$$\begin{aligned} \boldsymbol{P}\{B|D_1\} &= \binom{10}{4} \left({}^{1}/_{3}\right)^4 \left({}^{2}/_{3}\right)^6 = \frac{10!}{4!\,6!}\,\frac{2^6}{3^{10}} \\ \boldsymbol{P}\{B|D_2\} &= \binom{10}{4} \left({}^{3}/_{4}\right)^4 \left({}^{1}/_{4}\right)^6 = \frac{10!}{4!\,6!}\,\frac{3^4}{4^{10}} \end{aligned}$$

e quindi da (2.1)

$$\boldsymbol{P}\{B\} = \boldsymbol{P}\{B|D_1\}\,\boldsymbol{P}\{D_1\} + \boldsymbol{P}\{B|D_2\}\,\boldsymbol{P}\{D_2\} = \frac{10!}{4!6!}\left(\frac{1}{2}\,\frac{2^6}{3^{10}} + \frac{1}{2}\,\frac{3^4}{4^{10}}\right)$$

Dalla (2.3) del Teorema di Bayes 2.5 avremo allora

$$\begin{aligned} \boldsymbol{P}\{D_1|B\} &= \frac{\boldsymbol{P}\{B|D_1\}\,\boldsymbol{P}\{D_1\}}{\boldsymbol{P}\{B\}} = \frac{\frac{10!}{4!\,6!}\,\frac{2^6}{3^{10}}\,\frac{1}{2}}{\frac{10!}{4!6!}\left(\frac{1}{2}\,\frac{2^6}{3^{10}} + \frac{1}{2}\,\frac{3^4}{4^{10}}\right)} = \frac{2^{26}}{2^{26}+3^{14}} \simeq 0.933 \\ \boldsymbol{P}\{D_2|B\} &= \frac{\boldsymbol{P}\{B|D_2\}\,\boldsymbol{P}\{D_2\}}{\boldsymbol{P}\{B\}} = \frac{\frac{10!}{4!\,6!}\,\frac{3^4}{4^{10}}\,\frac{1}{2}}{\frac{10!}{4!6!}\left(\frac{1}{2}\,\frac{2^6}{3^{10}} + \frac{1}{2}\,\frac{3^4}{4^{10}}\right)} = \frac{3^{14}}{2^{26}+3^{14}} \simeq 0.064 \end{aligned}$$

A posteriori, quindi, l'esito delle estrazioni favorisce decisamente l'ipotesi di aver preso la scatola 1 □

Esercizio A.1.10. *Come il precedente Esercizio A.1.9, ma supponendo che nelle* 10 *estrazioni risulti*

$$B = 7 \textit{ bianche}, 3 \textit{ nere}$$

Risposta: $\boldsymbol{P}\{D_1|B\} \simeq 0.061\,, \quad \boldsymbol{P}\{D_2|B\} \simeq 0.939$ □

Esercizio A.1.11. $n = 100$ *urne, esternamente identiche e contenenti ciascuna* 20 *palline bianche e nere, sono ripartite in quattro categorie secondo la loro composizione interna:*

30	*contengono solo palline bianche (categoria* D_1*)*
40	*contengono* 16 *palline bianche e* 4 *nere (categoria* D_2*)*
20	*contengono* 10 *palline bianche e* 10 *nere (categoria* D_3*)*
10	*contengono* 5 *palline bianche e* 15 *nere (categoria* D_4*)*

Si sceglie un'urna e si estraggono – con rimessa – 3 *palline: se si verifica l'evento*

$$B = \textit{le } 3 \textit{ palline estratte sono tutte bianche}$$

quali sono le probabiità che l'urna scelta appartenga rispettivamente alle categorie D_1, D_2, D_3 *e* D_4*?*

Risposta: $\boldsymbol{P}\{D_1|B\} \simeq 0.565$, $\boldsymbol{P}\{D_2|B\} \simeq 0.385$, $\boldsymbol{P}\{D_3|B\} \simeq 0.047$ e $\boldsymbol{P}\{D_4|B\} \simeq 0.003$ □

Esercizio A.1.12. *Come il precedente Esercizio A.1.11, ma supponendo che nelle* 3 *estrazioni risulti*

$$B = \textit{le } 3 \textit{ palline estratte sono tutte nere}$$

Risposta: $\boldsymbol{P}\{D_1|B\} = 0$, $\boldsymbol{P}\{D_2|B\} \simeq 0.046$, $\boldsymbol{P}\{D_3|B\} \simeq 0.355$ e $\boldsymbol{P}\{D_4|B\} \simeq 0.599$ □

Esercizio A.1.13. *Tre scatole esternamente identiche contengono ciascuna 4 palline di colore bianco e nero, ma con diverse composizioni:*

$$\begin{cases} \textit{scatola } 1: & 1 \textit{ bianca, } 3 \textit{ nere} \\ \textit{scatola } 2: & 2 \textit{ bianche, } 2 \textit{ nere} \\ \textit{scatola } 3: & 3 \textit{ bianche, } 1 \textit{ nera} \end{cases}$$

Si prende una scatola a caso, si estrae una pallina e si osserva che essa è bianca (evento B): definiti allora gli eventi

$D_1 = \textit{ho preso la scatola } 1 \quad D_2 = \textit{ho preso la scatola } 2 \quad D_3 = \textit{ho preso la scatola } 3$

calcolare le probabilità a posteriori $\boldsymbol{P}\{D_1 \mid B\}$, $\boldsymbol{P}\{D_2 \mid B\}$ *e* $\boldsymbol{P}\{D_3 \mid B\}$

Risposta: $\boldsymbol{P}\{D_1 \mid B\} \simeq 0.167 \quad \boldsymbol{P}\{D_2 \mid B\} \simeq 0.333 \quad \boldsymbol{P}\{D_3 \mid B\} \simeq 0.500$ □

Esercizio A.1.14. *Una partita di 100 oggetti ne contiene 5 difettosi. La partita viene sottoposta ad una verifica alla fine della quale essa può essere accettata o rifiutata. Quale delle seguenti due procedure respinge la partita con maggior probabilità?*

a) si scelgono 6 oggetti e si respinge la partita se se ne trovano 1 o più difettosi
b) si scelgono 20 oggetti e si respinge la partita se se ne trovano 2 o più difettosi

Soluzione: Definiti gli eventi

$$\begin{aligned} A &= \text{si respinge la partita con la procedura } a \\ B &= \text{si respinge la partita con la procedura } b \\ A_0 &= \text{si trovano 0 difettosi su 6 ispezionati} \\ B_0 &= \text{si trovano 0 difettosi su 20 ispezionati} \\ B_1 &= \text{si trova 1 difettoso su 20 ispezionati} \end{aligned}$$

è evidente che

$$\boldsymbol{P}\{A\} = 1 - \boldsymbol{P}\{A_0\} \qquad \boldsymbol{P}\{B\} = 1 - \boldsymbol{P}\{B_0\} - \boldsymbol{P}\{B_1\}$$

D'altra parte si ha

$$\begin{aligned}\boldsymbol{P}\{A_0\} &= \frac{95}{100}\frac{94}{99}\frac{93}{98}\frac{92}{97}\frac{91}{96}\frac{90}{95} \simeq 0.729\\ \boldsymbol{P}\{B_0\} &= \frac{95}{100}\frac{94}{99}\frac{93}{98}\cdots\frac{78}{83}\frac{77}{82}\frac{76}{81} \simeq 0.319\end{aligned}$$

mentre, tenendo conto di tutti i modi in cui si può trovare un pezzo difettoso, per B_1 risulta

$$\begin{aligned}\boldsymbol{P}\{B_1\} &= \frac{5}{100}\frac{95}{99}\frac{94}{98}\cdots\frac{78}{82}\frac{77}{81}+\frac{95}{100}\frac{5}{99}\frac{94}{98}\cdots\frac{78}{82}\frac{77}{81}+\ldots+\frac{95}{100}\frac{94}{99}\frac{93}{98}\cdots\frac{78}{82}\frac{5}{81}\\ &= 20\,\frac{95}{100}\frac{94}{99}\frac{93}{98}\cdots\frac{78}{83}\frac{77}{82}\frac{5}{81} \simeq 0.420\end{aligned}$$

per cui in definitiva $\boldsymbol{P}\{A\} \simeq 0.271$, e $\boldsymbol{P}\{B\} \simeq 0.261$ e quindi la procedura a risulta più severa della procedura b □

Esercizio A.1.15. *Un gruppo di candidati viene sottoposto ad un test con un questionario di $n = 8$ domande a risposta multipla. Ogni domanda ha 4 possibili risposte. Il test viene superato se si risponde correttamente ad almeno 6 domande. Un candidato completamente impreparato risponde in maniera del tutto casuale: in media a quante domande risponderà correttamente? E con quale probabilità supererà il test?*

Soluzione: Se il candidato risponde in maniera casuale, per ogni domanda la sua risposta sarà corretta con probabilità $p = {}^1\!/_4$; se inoltre si suppone che le risposte siano indipendenti fra loro, il numero X delle risposte esatte sarà una *v-a* binomiale $\mathfrak{B}(n;p) = \mathfrak{B}({}^1\!/_4)$ per cui

$$\begin{aligned}\boldsymbol{E}[X] &= np = \frac{8}{4} = 2\\ \boldsymbol{P}\{X \geq 6\} &= \sum_{k=6}^{8}\binom{8}{k}\left({}^1\!/_4\right)^k\left({}^3\!/_4\right)^{8-k} = \frac{8\cdot 7}{2}\frac{3^2}{4^8}+8\,\frac{3}{4^8}+\frac{1}{4^8} = \frac{277}{65\,536} \simeq 0.004\end{aligned}$$

cioè risponderà mediamente a due domande e supererà il test con una probabilità dello 0.4%, cioè circa quattro volte su mille □

Esercizio A.1.16. *Un venditore deve fare* 10 *telefonate ogni giorno per convincere dei clienti ad acquistare un prodotto; dall'esperienza precedente si sa che il cliente acquista il prodotto nel* 15% *dei casi. Supponendo che gli esiti delle telefonate siano indipendenti, calcolare:*

- *il numero medio μ di prodotti venduti al giorno e la deviazione standard σ*
- *la probabilità di vendere meno di tre prodotti al giorno*

Se invece la percentuale dei successi fosse solo del 5%*, usare l'approssimazione di Poisson per calcolare con quale probabilità* 10 *venditori venderebbero esattamente* 4 *prodotti in un giorno*

Soluzione: Siccome ogni giorno il venditore esegue $n = 10$ tentativi indipendenti con probabilità di successo $p = 0.15$, il numero di prodotti venduti sarà una *v-a* X binomiale $\mathfrak{B}(n;p) = \mathfrak{B}(10;0.15)$ e quindi

$$\boldsymbol{P}\{X = k\} = \binom{10}{k}(0.15)^k(0.85)^{10-k} \qquad k = 0, 1, \dots, 10$$

Pertanto da (4.24) avremo

$$\mu = np = 1.5 \qquad \sigma = \sqrt{np(1-p)} \simeq 1.129$$

e inoltre

$$\boldsymbol{P}\{X < 3\} = \sum_{k=0}^{2}\binom{10}{k}(0.15)^k(0.85)^{10-k} \simeq 0.820$$

Nella seconda parte abbiamo invece $n = 100$ e $p = 0.05$ per cui le probabilità per $X \sim \mathfrak{B}(100;0.05)$ saranno più complicate da calcolare: sarà quindi preferibile adottare l'approssimazione di Poisson con $\lambda = np = 100 \times 0.05 = 5$, cioè

$$\boldsymbol{P}\{X = 4\} = \binom{100}{4}(0.05)^4(0.95)^{96} \simeq e^{-5}\,\frac{5^4}{4!} = 0.006738\,\frac{5^4}{4!} \simeq 0.175$$

dove abbiamo anche fatto uso delle Tavole E.5 □

Esercizio A.1.17. *Il numero di particelle* α *emesse da un campione radioattivo in ogni intervallo di* 10 *secondi è una v-a* $X \sim \mathfrak{P}(\lambda)$ *di Poisson con* $\lambda = 2$*: calcolare* $\boldsymbol{P}\{X > 6\}$

Soluzione: Usando la legge di Poisson $\mathfrak{P}(2)$ si ha

$$\boldsymbol{P}\{X > 6\} = 1 - \boldsymbol{P}\{X \le 6\} = 1 - \sum_{k=0}^{6}\boldsymbol{P}\{X = k\} = 1 - e^{-2}\sum_{k=0}^{6}\frac{2^k}{k!} \simeq 0.0045$$

avendo preso il valore dell'esponenziale dalle Tavole E.5 □

Esercizio A.1.18. *Nella fascia oraria fra le 12:00 e le 13:00 di ogni giorno un centralino telefonico riceve un numero aleatorio* X *di chiamate con una media di* $\alpha = 10$ *telefonate all'ora: calcolare la probabilità* $\boldsymbol{P}\{X \le 4\}$ *di ricevere non più di* 4 *telefonate in quella fascia oraria*

Soluzione: In base alla discussione dell'Esempio 5.10, il numero X di telefonate nel dato intervallo di $T = 1$ ora sarà una *v-a* di Poisson $\mathfrak{P}(\lambda)$ con $\lambda = \alpha T = 10$, e pertanto

$$\boldsymbol{P}\{X \leq 4\} = \sum_{k=0}^{4} \boldsymbol{P}\{X = k\} = e^{-10} \sum_{k=0}^{4} \frac{10^k}{k!} \simeq 0.029$$

con il solito uso delle tavole □

Esercizio A.1.19. *Data la v-a* $X \sim \mathfrak{B}(n; p)$ *binomiale con* $n = 1\,000$ *e* $p = 0.003$*, calcolare con l'approssimazione di Poisson la probabilità* $\boldsymbol{P}\{2 \leq X \leq 4\}$

Soluzione: Nell'approssimazione (Teorema 5.8) useremo la legge di Poisson $\mathfrak{P}(\lambda)$ con $\lambda = np = 1\,000 \times 0.003 = 3$, e avremo

$$\begin{aligned} \boldsymbol{P}\{2 \leq X \leq 4\} &= \boldsymbol{P}\{X = 2\} + \boldsymbol{P}\{X = 3\} + \boldsymbol{P}\{X = 4\} \\ &\simeq e^{-3} \left(\frac{3^2}{2!} + \frac{3^3}{3!} + \frac{3^4}{4!} \right) \simeq 0.616 \end{aligned}$$

avendo preso il valore dell'esponenziale dalle Tavole E.5 □

Esercizio A.1.20. *Usando l'approssimazione di Poisson, calcolare la probabilità q di vincere* 1 *o* 2 *volte giocando regolarmente per venti anni a una lotteria settimanale nella quale la probabilità di vincere in ogni estrazione è* $p = 0.00025$

Risposta: $q \simeq 0.227$ □

Esercizio A.1.21. *La probabilità che si verifichi un particolare evento A è* $p = 0.001$*. Supponendo di eseguire* $2\,000$ *tentativi di verifica di A, calcolare con l'approssimazione di Poisson la probabilità q che A si verifichi almeno* 4 *volte*

Risposta: $q \simeq 0.143$ □

Esercizio A.1.22. *La probabilità che esca un particolare ambo in una estrazione di* 5 *numeri al lotto è* $p = 0.0025$*. Supponendo che ci siano* 2 *estrazioni alla settimana e usando l'approssimazione di Poisson, calcolare la probabilità q che in* 5 *anni quell'ambo esca almeno* 3 *volte*

Risposta: $q \simeq 0.141$ □

Esercizio A.1.23. *In ogni estrazione di una lotteria la probabilità di vincere il premio è* $p = 0.0004$ *. Utilizzare l'approssimazione di Poisson per calcolare la probabilità q di vincere almeno* 3 *volte (cioè:* 3 *o più volte) su* $n = 3\,650$ *estrazioni*

Risposta: $q \simeq 0.181$ □

Esercizio A.1.24. *In ogni estrazione di una lotteria la probabilità di vincere il premio è* $p = 0.002$. *Utilizzare l'approssimazione di Poisson per calcolare la probabilità* q *di vincere esattamente* 2 *volte su* $n = 2\,500$ *estrazioni*

Risposta: $q \simeq 0.084$ □

Esercizio A.1.25. *Data una v-a normale* $X \sim \mathfrak{N}(0, \sigma^2)$, *si prenda in maniera arbitraria un intervallo* $[a, b]$ *con* $0 < a < b$: *facendo uso della FDC e della fdp normali standard*

$$\Phi(x) = \int_{-\infty}^{x} \varphi(y)\, dy\,, \qquad \varphi(x) = \Phi'(x) = \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}$$

- *determinare (in funzione di* a *e* b*) il valore* $\widehat{\sigma}$ *di* σ *che rende massima la probabilità* $\boldsymbol{P}\{a \le X \le b\}$
- *calcolare esplicitamente* σ *e* $\boldsymbol{P}\{a \le X \le b\}$ *nel caso in cui* $a = 1$ *e* $b = 2$ *(usare le Tavole E.1)*

Soluzione: Se $X^* = {}^{X}\!/_{\sigma} \sim \mathfrak{N}(0, 1)$ è la *v-a* standardizzata, avremo innanzitutto

$$\begin{aligned} q(\sigma) &= \boldsymbol{P}\{a \le X \le b\} = \boldsymbol{P}\left\{ {}^{a}\!/_{\sigma} \le X^* \le {}^{b}\!/_{\sigma} \right\} = \Phi\left({}^{b}\!/_{\sigma}\right) - \Phi\left({}^{a}\!/_{\sigma}\right) \\ q'(\sigma) &= -\frac{b}{\sigma^2}\, \varphi\left({}^{b}\!/_{\sigma}\right) + \frac{a}{\sigma^2}\, \varphi\left({}^{a}\!/_{\sigma}\right) = \frac{a\, e^{-{}^{a^2}\!/_{2\sigma^2}} - b\, e^{-{}^{b^2}\!/_{2\sigma^2}}}{\sigma^2\sqrt{2\pi}} \end{aligned}$$

per cui il valore di σ che rende massima $q(\sigma)$ sarà soluzione dell'equazione $q'(\sigma) = 0$ cioè di

$$a\, e^{-{}^{a^2}\!/_{2\sigma^2}} = b\, e^{-{}^{b^2}\!/_{2\sigma^2}}$$

Passando ai logaritmi questa equazione si riscrive come

$$\log a - \frac{a^2}{2\sigma^2} = \log b - \frac{b^2}{2\sigma^2}$$

e la sua soluzione – che risponde alla prima domanda – è

$$\widehat{\sigma} = \sqrt{\frac{b^2 - a^2}{\log b^2 - \log a^2}}$$

Ponendo poi $a = 1$ e $b = 2$, dalle Tavole E.1 si ottiene

$$\begin{aligned} \widehat{\sigma} &= \sqrt{\frac{4-1}{\log 4 - \log 1}} = \sqrt{\frac{3}{\log 4}} \simeq 1.471 \\ q(\widehat{\sigma}) &= \Phi\left(2\sqrt{\frac{\log 4}{3}}\right) - \Phi\left(\sqrt{\frac{\log 4}{3}}\right) \simeq \Phi(1.36) - \Phi(0.68) \\ &= 0.91309 - 0.75175 \simeq 0.161 \end{aligned}$$

che risponde anche alla seconda domanda □

Esercizio A.1.26. *Calcolare la probabilità* $\boldsymbol{P}\{-3 \leq X \leq -1\}$ *per una v-a* $X \sim \mathfrak{N}(\mu, \sigma^2)$ *normale con media* $\mu = -2$ *e varianza* $\sigma^2 = 4$ *(usare le Tavole E.1)*

Soluzione: Detta

$$X^* = \frac{X - \mu}{\sigma} = \frac{X + 2}{2} \sim \mathfrak{N}(0, 1)$$

la *v-a* standardizzata, avremo

$$\begin{aligned} \boldsymbol{P}\{-3 \leq X \leq -1\} &= \boldsymbol{P}\left\{\frac{-3+2}{2} \leq \frac{X+2}{2} \leq \frac{-1+2}{2}\right\} = \boldsymbol{P}\left\{-{}^1\!/_2 \leq X^* \leq {}^1\!/_2\right\} \\ &= \Phi\left({}^1\!/_2\right) - \Phi\left(-{}^1\!/_2\right) \end{aligned}$$

dove $\Phi(x)$ è la *FDC* normale standard (3.18) i cui valori numerici si trovano nelle Tavole E.1. Tenendo inoltre conto delle proprietà di simmetria (3.19) di $\Phi(x)$ abbiamo che $\Phi\left(-{}^1\!/_2\right) = 1 - \Phi\left({}^1\!/_2\right)$, e quindi in definitiva con un arrotondamento finale

$$\boldsymbol{P}\{-3 \leq X \leq -1\} = 2\Phi\left({}^1\!/_2\right) - 1 = 2 \times 0.69146 - 1 \simeq 0.383$$

avendo preso dalle tavole i valori numerici necessari □

Esercizio A.1.27. *Supponiamo di estrarre* $n = 40$ *valori da una v-a* $X \sim \mathfrak{N}(\mu, \sigma^2)$ *normale con media* $\mu = 3$ *e varianza* $\sigma^2 = 6$*: quale è il numero medio di valori che cadono nell'intervallo* $[3, 4]$ *?*

Soluzione: Ogni valore estratto cade in $[3, 4]$ con probabilità $p = \boldsymbol{P}\{3 \leq X \leq 4\}$. Se inoltre

$$X^* = \frac{X - \mu}{\sigma} = \frac{X - 3}{\sqrt{6}} \sim \mathfrak{N}(0, 1)$$

è la *v-a* standardizzata, avremo

$$\begin{aligned} p &= \boldsymbol{P}\{3 \leq X \leq 4\} = \boldsymbol{P}\left\{\frac{3-3}{\sqrt{6}} \leq \frac{X-3}{\sqrt{6}} \leq \frac{4-3}{\sqrt{6}}\right\} = \boldsymbol{P}\left\{0 \leq X^* \leq {}^1\!/_{\sqrt{6}}\right\} \\ &= \Phi\left({}^1\!/_{\sqrt{6}}\right) - \Phi(0) \simeq \Phi(0.41) - \Phi(0) = 0.65910 - 0.5000 \simeq 0.159 \end{aligned}$$

Ne segue che il numero aleatorio Y di valori estratti che cadono in $[3, 4]$ sarà una *v-a* binomiale $\mathfrak{B}(n; p)$ con $n = 40$ e $p = 0.159$, per cui in definitiva

$$\boldsymbol{E}[Y] = np = 40 \times 0.159 = 6.36$$

sarà il numero medio di valori in $[3, 4]$ □

Esercizio A.1.28. *Una coda di* $n = 60$ *persone attende di ritirare del denaro ad uno sportello bancario: la quantità di denaro prelevata da ciascuno è una v-a con media* $\mu = 50$ *US$, e deviazione standard* $\sigma = 20$ *US$. I prelevamenti sono indipendenti. Usando l'approssimazione normale, determinare l'ammontare di denaro che deve essere inizialmente tenuto in cassa per soddisfare le richieste di tutti con una probabilità del* 95%

Soluzione: Se X_k con $k = 1, 2, \ldots, n$ sono le *v-a* che rappresentano i singoli prelevamenti indipendenti, con

$$\boldsymbol{E}[X_k] = \mu \qquad \boldsymbol{V}[X_k] = \sigma^2$$

il denaro ritirato dagli n clienti sarà

$$Y = \sum_{k=1}^{n} X_k \qquad \text{con} \qquad \boldsymbol{E}[Y] = n\mu \qquad \boldsymbol{V}[Y] = n\sigma^2$$

Se allora la corrispondente *v-a* standardizzata è

$$Y^* = \frac{Y - \boldsymbol{E}[Y]}{\sqrt{\boldsymbol{V}[Y]}} = \frac{Y - n\mu}{\sigma\sqrt{n}}$$

l'approssimazione normale ci dice che per ogni y avremo

$$\boldsymbol{P}\{Y \le y\} = \boldsymbol{P}\left\{Y^* \le \frac{y - n\mu}{\sigma\sqrt{n}}\right\} \simeq \Phi\left(\frac{y - n\mu}{\sigma\sqrt{n}}\right)$$

Il problema richiede ora di determinare il valore di y (denaro tenuto inizialmente in cassa) in modo che $\boldsymbol{P}\{Y \le y\} = 0.95$, e quindi di trovare il valore di y che soddisfa l'equazione

$$\Phi\left(\frac{y - n\mu}{\sigma\sqrt{n}}\right) = 0.95$$

Posto in maniera diversa (vedi Sezione 3.5) questo vuol dire che l'argomento della funzione Φ deve coincidere con il quantile $\varphi_{0.95}$ di ordine 0.95 della legge normale standard, cioè dalle Tavole E.1

$$\frac{y - n\mu}{\sigma\sqrt{n}} = \varphi_{0.95} \simeq 1.65$$

per cui in definitiva, sostituendo i valori numerici,

$$y \;=\; n\mu + \varphi_{0.95}\sigma\sqrt{n} \;=\; 60 \times 50 + 1.65 \times 20\sqrt{60} \;\simeq\; 3\,255.61 \text{ US\$}$$

è la richiesta somma che con il 95% di probabilità soddisfa le richieste di tutti i clienti □

Esercizio A.1.29. *Una fabbrica produce delle partite di* $n = 10\,000$ *pezzi e la probabilità che uno di tali pezzi sia difettoso è* $p = 0.05$. *Le cause dei difetti sono indipendenti per i diversi pezzi. I pezzi difettosi sono accumulati in un recipiente: usando l'approssimazione normale, determinare la capienza minima del recipiente che premette di contenere tutti i pezzi difettosi con una probabilità del* 99%

Soluzione: Il numero X dei pezzi difettosi è una *v-a* ottenuta sommando n *v-a* indipendenti di Bernoulli $\mathfrak{B}(1;p)$, e quindi $X \sim \mathfrak{B}(n;p)$ con

$$\boldsymbol{E}[X] = np = 500 \qquad \boldsymbol{V}[X] = np(1-p) = 475$$

Definita allora la *v-a* standardizzata

$$X^* = \frac{X - \boldsymbol{E}[X]}{\sqrt{\boldsymbol{V}[X]}} = \frac{X - np}{\sqrt{np(1-p)}}$$

l'approssimazione normale ci permette di scrivere

$$\boldsymbol{P}\{X \le x\} = \boldsymbol{P}\left\{ X^* \le \frac{x - np}{\sqrt{np(1-p)}} \right\} \simeq \Phi\left(\frac{x - np}{\sqrt{np(1-p)}} \right)$$

e quindi la richiesta capienza minima x determinata dalla condizione $\boldsymbol{P}\{X \le x\} = 0.99$ si troverà risolvendo l'equazione

$$\Phi\left(\frac{x - np}{\sqrt{np(1-p)}} \right) = 0.99$$

ovvero imponendo che l'argomento della Φ coincida con il quantile $\varphi_{0.99}$ di ordine 0.99 della normale standard. Dalle Tavole E.1 si ha allora

$$\frac{x - np}{\sqrt{np(1-p)}} = \varphi_{0.99} \simeq 2.33$$

e quindi, sostituendo i valori numerici e arrotondando i risultati,

$$x = np + \varphi_{0.99}\sqrt{np(1-p)} \simeq 551$$

è la capienza minima del recipiente che contiene tutti i pezzi difettosi con il 99% di probabilità □

Esercizio A.1.30. *Sia $Y_n = X_1 + \ldots + X_n$ la somma di $n = 192$ numeri aleatori indipendenti X_k tutti con valori compresi fra 0 e 1, e con attesa $\mu = \boldsymbol{E}[X_k] = {}^1/_2$ e varianza $\sigma^2 = \boldsymbol{V}[X_k] = {}^1/_{12}$. Facendo uso dell'approssimazione normale calcolare la probabilità $\boldsymbol{P}\{95 \le Y_n \le 100\}$*

Soluzione: Si osserva innanzitutto che

$$\boldsymbol{E}[Y_n] = n\mu = 96 \qquad \boldsymbol{V}[Y_n] = n\sigma^2 = 16$$

e quindi che la somma standardizzata è (si noti che $\sigma\sqrt{n} = \sqrt{n\sigma^2} = 4$)

$$Y_n^* = \frac{Y_n - n\mu}{\sigma\sqrt{n}} = \frac{Y_n - 96}{4}$$

Pertanto avremo

$$\boldsymbol{P}\{95 \le Y_n \le 100\} = \boldsymbol{P}\left\{\frac{95-96}{4} \le Y_n^* \le \frac{100-96}{4}\right\} = \boldsymbol{P}\left\{-{}^1\!/_4 \le Y_n^* \le 1\right\}$$

e quindi in approssimazione normale

$$\boldsymbol{P}\{95 \le Y_n \le 100\} \simeq \Phi(1) - \Phi\left(-{}^1\!/_4\right) = \Phi(1) + \Phi\left(0.25\right) - 1 \simeq 0.440$$

con arrotondamento dei valori presi dalle Tavole E.1 □

Esercizio A.1.31. *Siano date $n = 144$ v-a indipendenti, ciascuna con attesa $\mu = 0$ e deviazione standard $\sigma = 3$: calcolare in approssimazione normale la probabilità $\boldsymbol{P}\left\{-{}^1\!/_8 \le \overline{X}_n \le {}^3\!/_8\right\}$ che la loro media $\overline{X}_n$ cada fra $-{}^1\!/_8$ e ${}^3\!/_8$*

Soluzione: Siccome

$$\boldsymbol{E}\left[\overline{X}_n\right] = \mu = 0 \qquad \boldsymbol{V}\left[\overline{X}_n\right] = \frac{\sigma^2}{n} = \frac{1}{16}$$

la *v-a* standardizzata sarà

$$Y_n^* = \frac{\overline{X}_n - \mu}{\sigma}\sqrt{n} = 4\overline{X}_n$$

e quindi in approssimazione normale

$$\begin{aligned}\boldsymbol{P}\left\{-{}^1\!/_8 \le \overline{X}_n \le {}^3\!/_8\right\} &= \boldsymbol{P}\left\{-{}^1\!/_2 \le Y_n^* \le {}^3\!/_2\right\} \\ &\simeq \Phi(1.5) - \Phi(-0.5) = \Phi(1.5) + \Phi(0.5) - 1 \simeq 0.625\end{aligned}$$

con arrotondamento dei valori presi dalle Tavole E.1 □

Esercizio A.1.32. *La v-a Y_n è somma di $n = 64$ numeri aleatori indipendenti, identicamente distribuiti, ciascuno con attesa $\mu = {}^1\!/_2$ e varianza $\sigma^2 = {}^1\!/_{16}$. Facendo uso dell'approssimazione normale, calcolare la probabilità che il valore di Y_n cada fra 34 e 36*

Risposta: $\boldsymbol{P}\{34 \le Y_n \le 36\} \simeq 0.136$ □

Esercizio A.1.33. *Si sa che una v-a X ha varianza $\sigma^2 = 4$, ma non se ne conosce l'attesa μ: si eseguono $n = 64$ misure di X e se ne calcola la media $\overline{X}_n$. Facendo uso dell'approssimazione normale, calcolare la probabilità $\boldsymbol{P}\left\{|\overline{X}_n - \mu| > 0.5\right\}$ che il valore assoluto dello scarto fra media $\overline{X}_n$ e valore d'attesa μ superi 0.5*

Risposta: $\boldsymbol{P}\left\{|\overline{X}_n - \mu| > 0.5\right\} \simeq 0.0455$ □

Esercizio A.1.34. *Sia X una v-a della quale non si conosce la distribuzione, ma della quale si sa che $\mu = \boldsymbol{E}[X] = 1$ e $\sigma^2 = \boldsymbol{V}[X] = {}^1/_3$. Se $Y_n = X_1 + \ldots + X_n$ è la somma di $n = 300$ v-a indipendenti e tutte distribuite come X*

- *determinare l'attesa $\boldsymbol{E}[Y_n]$ e la varianza $\boldsymbol{V}[Y_n]$ di tale somma;*
- *calcolare la probabilità $\boldsymbol{P}\{296 \leq Y_n \leq 302\}$ facendo uso dell'approssimazione normale.*

Risposta: $\boldsymbol{E}[Y_n] = 300$ $\boldsymbol{V}[S_n] = 100$ $\boldsymbol{P}\{296 \leq Y_n \leq 302\} \simeq 0.235$ □

Esercizio A.1.35. *Un computer genera $n = 192$ numeri aleatori indipendenti $X_1, \ldots, X_n$ tutti distribuiti in maniera uniforme fra 0 e 1, e quindi con $\boldsymbol{E}[X_k] = {}^1/_2$ e $\boldsymbol{V}[X_k] = {}^1/_{12}$. Facendo uso dell'approssimazione normale, calcolare la probabilità $\boldsymbol{P}\{92 \leq Y_n \leq 100\}$ che la somma $Y_n = X_1 + \ldots + X_n$ cada fra 92 e 100*

Risposta: $\boldsymbol{P}\{92 \leq Y_n \leq 100\} \simeq 0.683$ □

Esercizio A.1.36. *Si sa che il valore di un segnale in un determinato istante è $\mu = 100$; si sa però anche che a tale valore si sovrappone un rumore casuale Z con media nulla e varianza $\sigma^2 = 9$ in modo che il risultato di una misura sia la v-a $X = \mu + Z$. Si eseguono $n = 81$ misure indipendenti di X e se ne prende la media $\overline{X}_n$: calcolare in approssimazione normale la probabilità $\boldsymbol{P}\{\overline{X}_n \geq 100.5\}$ che la media superi il valore 100.5*

Risposta: $\boldsymbol{P}\{\overline{X}_n \geq 100.5\} \simeq 0.067$ □

Esercizio A.1.37. *Una variabile aleatoria con valore d'attesa $\mu = 250$ e varianza $\sigma^2 = 1\,710$ viene misurata $n = 190$ volte: detta $\overline{X}_n$ la media aritmetica dei risultati di tali misure, usare l'approssimazione normale e le Tavole E.1 per calcolare il valore della probabilità $\boldsymbol{P}\{247.6 \leq \overline{X}_n \leq 253.9\}$*

Risposta: $\boldsymbol{P}\{247.6 \leq \overline{X}_n \leq 253.9\} \simeq 0.691$ □

Esercizio A.1.38. *Si sa che una v-a X ha varianza $\sigma^2 = 3$; si eseguono $n = 243$ misure di X e se ne calcola la media $\overline{X}_n$. Facendo uso dell'approssimazione normale, calcolare la probabilità $\boldsymbol{P}\{|\overline{X}_n - \mu| > {}^1/_6\}$ che il valore assoluto della differenza fra $\overline{X}_n$ e il suo valore d'attesa μ superi ${}^1/_6$*

Risposta: $\boldsymbol{P}\{|\overline{X}_n - \mu| > {}^1/_6\} \simeq 0.134$ □

Esercizio A.1.39. *Un computer genera $n = 3\,000$ numeri aleatori positivi X_i, $i = 1, \ldots, n$, con $\boldsymbol{E}[X_i] = 3.0$, e $\boldsymbol{V}[X_i] = 2.7$. Utilizzando l'approssimazione normale calcolare la probabilità $\boldsymbol{P}\{Y_n > 9\,108\}$ che la somma $Y_n = X_1 + \ldots + X_n$ superi 9 108*

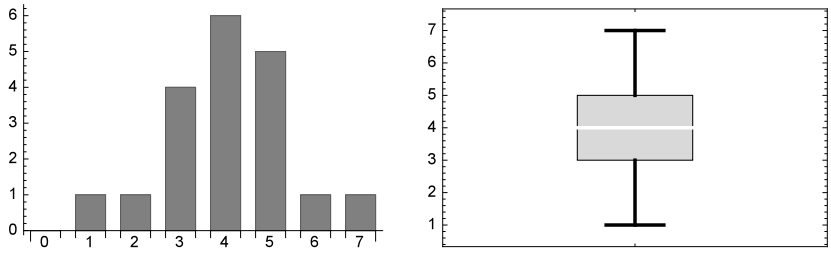


Figura A.2: Diagramma a barre e *boxplot* dell'Esercizio A.2.1

Risposta: $\boldsymbol{P}\{Y_n > 9\,108\} \simeq 0.115$ □

Esercizio A.1.40. *$n = 240$ numeri aleatori X_i (con $i = 1, \ldots, 240$) sono indipendenti e identicamente distribuiti con media $\mu = \boldsymbol{E}[X_i] = {}^1/_2$, e varianza $\sigma^2 = \boldsymbol{V}[X_i] = {}^3/_5$. Ponendo $Y_n = X_1 + \ldots + X_n$, e utilizzando l'approssimazione normale, calcolare la probabilità $\boldsymbol{P}\{114 \le Y_n \le 132\}$*

Risposta: $\boldsymbol{P}\{114 \le Y_n \le 132\} \simeq 0.533$ □

A.2 Statistica Descrittiva

Esercizio A.2.1. *Sia dato il seguente campione composto di $n = 19$ numeri interi:*

$$5,\ 4,\ 5,\ 4,\ 4,\ 1,\ 5,\ 6,\ 5,\ 4,\ 2,\ 3,\ 7,\ 5,\ 3,\ 3,\ 4,\ 3,\ 4$$

Calcolare: la media m, la varianza s^2, la deviazione standard s, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, la media armonica m_A, la media quadratica m_Q e la media geometrica m_G. Costruire la tabella delle frequenze assolute e relative, tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Soluzione: Quando, come in questo caso, il campione non è già ordinato, conviene prima di tutto riordinarlo:

$$1,\ 2,\ 3,\ 3,\ 3,\ 3,\ 4,\ 4,\ 4,\ 4,\ 4,\ 4,\ 5,\ 5,\ 5,\ 5,\ 5,\ 6,\ 7$$

Il valore della media si ottiene facilmente

$$m = \frac{77}{19} \simeq 4.05$$

Dalla Definizione 6.31 le medie armonica, quadratica e geometrica sono

$$m_A = \frac{266}{79} \simeq 3.37 \qquad m_Q = \sqrt{\frac{347}{19}} \simeq 4.27 \qquad m_G \simeq 3.77$$

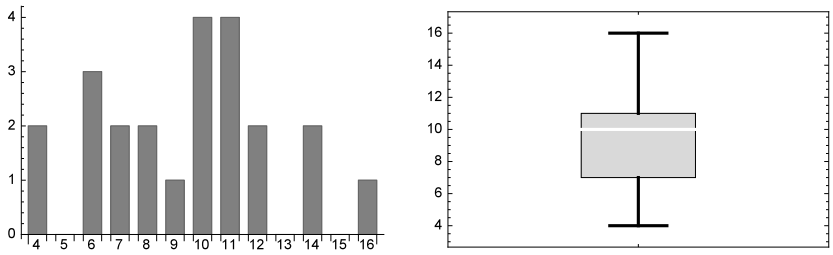


Figura A.3: Diagramma a barre e *boxplot* dell'Esercizio A.2.2

Per la varianza (e la deviasione standard) conviene prendere la media del quadrato dei dati (già usata per la media quadratica) e sottrarle il quadrato della media:

$$s^2 = \frac{347}{19} - \left(\frac{77}{19}\right)^2 = \frac{664}{361} \simeq 1.84 \qquad s = \sqrt{\frac{664}{361}} \simeq 1.36$$

Tenendo conto del campione ordinato, il range è ovviamente

$$\Delta = 7 - 1 = 6$$

mentre per i quantili si calcolano prima le loro posizioni

$$\text{mediana} \quad \frac{n+1}{2} = 10 \qquad 1^o \text{ e } 3^o \text{ quartile} \quad \frac{n+1}{4} = 5\,, \quad 3\,\frac{n+1}{4} = 15$$

e poi si cercano il 10^o il 5^o e il 15^o elemento del campione

$$q_{1/2} = 4 \qquad q_{1/4} = 3 \qquad q_{3/4} = 5$$

Le modalità w_k sono gli interi $k = 1, 2, \ldots, 7$ e la tabella delle frequenze è

k	1	2	3	4	5	6	7
N_k	1	1	4	6	5	1	1
F_k	1	2	6	12	17	18	19
p_k	0.05	0.05	0.21	0.32	0.26	0.05	0.05
f_k	0.05	0.11	0.32	0.63	0.89	0.95	1.00

Sulla base di questi risultati si disegnano facilmente il diagramma a barre (delle frequenze assolute) e il *boxplot* rappresentati in Figura A.2: la moda è 4 □

Esercizio A.2.2. *Nella fascia oraria fra le 12:00 e le 13:00 di ogni giorno un centralino telefonico riceve un numero aleatorio X di chiamate. Il valore di X è stato registrato in* 23 *giorni diversi ottenendo i seguenti risultati:*

10, 14, 11, 10, 11, 10, 8, 6, 12, 7, 16 , 10, 11, 6, 8, 4, 9, 12, 7, 6, 14, 4, 11.

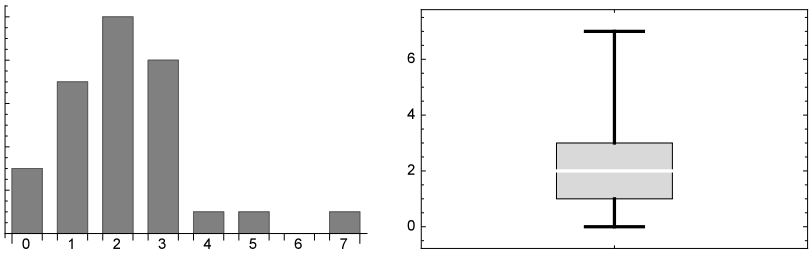


Figura A.4: Diagramma a barre e *boxplot* dell'Esercizio A.2.3

Calcolare: la media m, la varianza s^2, la deviazione standard s, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, la media armonica m_A, la media quadratica m_Q e la media geometrica m_G. Costruire la tabella delle frequenze assolute e relative, tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m = \frac{217}{23} \simeq 9.43 \qquad s^2 = \frac{5\,052}{529} \simeq 9.55 \qquad s = \sqrt{\frac{5\,052}{529}} \simeq 3.09$$
$$\Delta = 12 \qquad q_{1/2} = 10 \qquad q_{1/4} = 7 \qquad q_{3/4} = 11$$
$$m_A = \frac{55\,540}{6\,707} \simeq 8.27 \qquad m_Q = \sqrt{\frac{2\,267}{23}} \simeq 9.93 \qquad m_G \simeq 8.88$$

Il diagramma a barre e il *boxplot* sono rappresentati in Figura A.3. Strettamente parlando le mode sono molte, ma la loro significatività statistica è piuttosto modesta: quelle principali sono $6, 10$ e 11 □

Esercizio A.2.3. *Il numero di particelle α emesso da un campione radioattivo in ogni periodo di* 10 *secondi è una v-a X: supponendo che in* 31 *misurazioni (di* 10 *secondi l'una) le frequenze N_k dei valori k di X siano state:*

$k =$	0	1	2	3	4	5	6	7
$N_k =$	3	7	10	8	1	1	0	1

calcolare la media m, la varianza s^2, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 2.16 \qquad s^2 = \frac{5\,052}{529} \simeq 2.07$$
$$\Delta = 12 \qquad q_{1/2} = 2 \qquad q_{1/4} = 1 \qquad q_{3/4} = 3$$

Il diagramma a barre e il *boxplot* sono rappresentati in Figura A.4, e la moda è 2 (il massimo isolato in 7 non è significativo) □

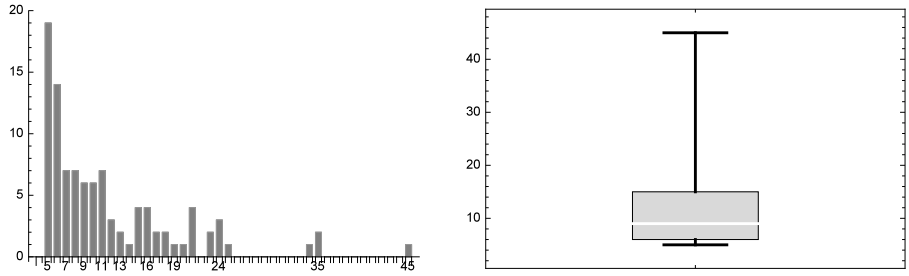


Figura A.5: Diagramma a barre e *boxplot* dell'Esercizio A.2.4

Esercizio A.2.4. *$n = 100$ giocatori di roulette partono con un capitale di 5\$ ciascuno, e alla fine del gioco hanno perduto tutto. Si registrano i valori massimi del capitale raggiunto da ogni giocatore durante il gioco ottenendo la seguente tabella*

25	9	5	5	5	9	6	5	15	45	35	6	5	6	24	21	16	5	8	7
7	5	5	35	13	9	5	18	6	10	19	16	21	8	13	5	9	10	10	6
23	8	5	10	15	7	5	5	24	9	11	34	12	11	17	11	16	5	15	5
12	6	5	5	7	6	17	20	7	8	8	6	10	11	6	7	5	12	11	18
6	21	6	5	24	7	16	21	23	15	11	8	6	8	14	11	6	9	6	10

Calcolare la media m, la varianza s^2, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m = \frac{58}{5} \simeq 11.60 \qquad s^2 = \frac{593}{10} = 59.30$$
$$\Delta = 40 \qquad q_{1/2} = 9 \qquad q_{1/4} = 6 \qquad q_{3/4} = 15$$

Il diagramma a barre e il *boxplot* sono rappresentati in Figura A.5, e la moda è 5 (gli altri massimi non sono significativi) □

Esercizio A.2.5. *Il numero k di clienti che si presentano ad uno sportello bancario fra le 12:00 e le 13:00 viene registrato in $n = 100$ giorni lavorativi ottenendo i risultati riportati nella seguente tabella di frequenze assolute N_k:*

$k =$	1	2	3	4	5	6	7	8	9	10	11
$N_k =$	2	11	15	18	20	14	10	5	2	2	1

Calcolare la media m, la varianza s^2, la deviazione standard s, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 4.84 \qquad s^2 \simeq 4.23 \qquad s \simeq 2.06$$
$$\Delta = 10 \qquad q_{1/2} = 5 \qquad q_{1/4} = 3 \qquad q_{3/4} = 6$$

Il diagramma a barre e il *boxplot* non sono riportati per brevità: la moda è 5 □

Esercizio A.2.6. *Si misura $n = 230$ volte il numero k di particelle α emesse in un periodo di 10 secondi da un campione radioattivo, e si ottengono i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$N_k =$	1	7	15	24	32	48	32	24	23	13	5	2	2	1	1

Calcolare la media m, la varianza s^2, la deviazione standard s, il coefficiente di variazione δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 5.48 \qquad s^2 \simeq 5.78 \qquad s \simeq 2.40 \qquad \delta \simeq 0.44$$
$$\Delta = 14 \qquad q_{1/2} = 5 \qquad q_{1/4} = 4 \qquad q_{3/4} = 7$$

Il diagramma a barre e il *boxplot* non sono riportati per brevità: la moda è 5 □

Esercizio A.2.7. *Si misura il numero k di clienti che telefonano ad un centralino tra le 11:00 e le 12:00 in $n = 50$ giornate lavorative tipiche, e si ottengono i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	1	2	3	4	5	6	7	8
$N_k =$	3	8	10	12	6	6	4	1

Calcolare la media m, la varianza s^2, la media geometrica m_G, la media armonica m_A, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 3.98 \qquad s^2 \simeq 3.02 \qquad m_G \simeq 3.56 \qquad m_A \simeq 3.08$$
$$q_{1/2} = 4 \qquad q_{1/4} = 3 \qquad q_{3/4} = 5$$

Il diagramma a barre e il *boxplot* non sono riportati per brevità: la moda è 4 □

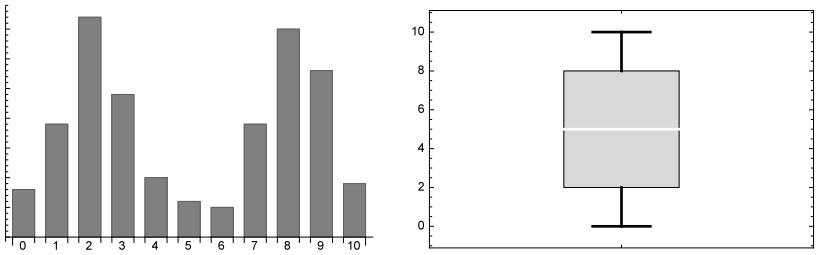


Figura A.6: Diagramma a barre e *boxplot* dell'Esercizio A.2.8

Esercizio A.2.8. *Si rileva il numero k di telefonate pervenute ad un centralino in un periodo di 2 ore in $n = 200$ giornate, e si ottengono i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	0	1	2	3	4	5	6	7	8	9	10
$N_k =$	8	19	37	24	10	6	5	19	35	28	9

Calcolare la media m, la varianza s^2, la deviazione standard s, il coefficiente di variazione δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 5.10 \qquad s^2 \simeq 10.05 \qquad s \simeq 3.17 \qquad \delta \simeq 0.62$$
$$q_{1/2} = 5 \qquad q_{1/4} = 2 \qquad q_{3/4} = 8$$

Il diagramma a barre e il *boxplot* sono riportati in Figura A.6; le mode sono 2 e 8 □

Esercizio A.2.9. *Si misura il numero k di clienti che telefonano ad un centralino tra le 11:00 e le 12:00 in $n = 47$ giornate lavorative tipiche, e si ottengono i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	1	2	3	4	5	6	7
$N_k =$	2	10	12	11	7	4	1

Calcolare la media m, la varianza s^2, la media geometrica m_G, la media armonica m_A, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 3.57 \qquad s^2 \simeq 1.99 \qquad m_G \simeq 3.27 \qquad m_A \simeq 2.94$$
$$q_{1/2} = 3 \qquad q_{1/4} = 2 \qquad q_{3/4} = 5$$

Il diagramma a barre e il *boxplot* non sono riportati per brevità: la moda è 3 □

Esercizio A.2.10. *Si misura $n = 200$ volte il numero k di particelle α emesse in un periodo di 10 secondi da un campione radioattivo, e si ottengono i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	1	2	3	4	5	6	7	8	9	10	11	12
$N_k =$	4	7	34	38	36	21	27	14	10	6	2	1

Calcolare la media m, la varianza s^2, la deviazione standard s, il coefficiente di variazione δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 5.32 \qquad s^2 \simeq 4.84 \qquad s \simeq 2.20 \qquad \delta \simeq 0.41$$
$$q_{1/2} = 5 \qquad q_{1/4} = 4 \qquad q_{3/4} = 7$$

Il diagramma a barre e il *boxplot* non sono riportati; le mode sono 3 e 6 □

Esercizio A.2.11. *Si misura il numero k di clienti che telefonano ad un centralino tra le 11:00 e le 12:00 in $n = 50$ giornate lavorative tipiche, e si ottengono i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	1	2	3	4	5	6	7	8
$N_k =$	1	2	6	14	13	6	7	1

Calcolare la media m, la varianza s^2, la media geometrica m_G, la media armonica m_A, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 4.74 \qquad s^2 \simeq 2.23 \qquad m_G \simeq 4.46 \qquad m_A \simeq 4.09$$
$$q_{1/2} = 5 \qquad q_{1/4} = 4 \qquad q_{3/4} = 6$$

Il diagramma a barre e il *boxplot* non sono riportati; le mode sono 4 e 7 □

Esercizio A.2.12. *Si misura $n = 50$ volte il numero k di particelle α emesse da un campione radioattivo in un periodo di 10 secondi ottenendo i risultati riportati nella seguente tabella delle frequenze assolute N_k:*

$k =$	1	2	3	4	5	6	7	8
$N_k =$	1	5	15	10	9	6	3	1

Calcolare la media m, la varianza s^2, la media geometrica m_G, la media armonica m_A, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

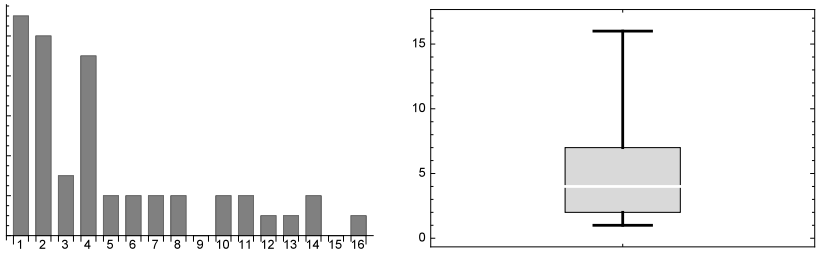


Figura A.7: Diagramma a barre e *boxplot* dell'Esercizio A.2.14

Risposta: I risultati numerici sono

$$m \simeq 4.12 \qquad s^2 \simeq 2.39 \qquad m_G \simeq 3.82 \qquad m_A \simeq 3.48$$
$$q_{1/2} = 4 \qquad q_{1/4} = 3 \qquad q_{3/4} = 5$$

Il diagramma a barre e il *boxplot* non sono riportati; la moda è 3 □

Esercizio A.2.13. *Il numero k di clienti che si presentano ad uno sportello bancario fra le 12:00 e le 13:00 viene registrato in $n = 50$ giorni lavorativi ottenendo i risultati riportati nella seguente tabella di frequenze assolute:*

$k =$	1	2	3	4	5	6	7	8	9	10	11
$N_k =$	2	3	10	8	7	4	4	5	5	1	1

Calcolare la media m, la varianza s^2, la deviazione standard s, il coefficiente di variazione δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$; tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m \simeq 5.26 \qquad s^2 \simeq 6.19 \qquad s \simeq 2.49 \qquad \delta \simeq 0.47$$
$$q_{1/2} = 5 \qquad q_{1/4} = 3 \qquad q_{3/4} = \frac{15}{2} = 7.50$$

Il diagramma a barre e il *boxplot* non sono riportati; le mode sono $3, 8$ e 9 □

Esercizio A.2.14. *Il numero di lanci di dado necessari per ottenere per la prima volta il risultato "6 " è aleatorio: si ripete l'esperimento 50 volte ottenendo i seguenti risultati*

4	4	3	14	7	4	5	11	2	8	2	4	13	1	2	6	7	2	1	2
3	4	2	2	16	1	12	3	2	5	2	4	1	11	1	1	1	10	10	4
4	6	8	2	1	1	4	14	1	1										

Calcolare: la media m, la varianza s^2, la deviazione standard s, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, la media armonica m_A, la media quadratica m_Q e la media geometrica m_G. Costruire la tabella delle frequenze assolute e relative, tracciare il diagramma a barre e il boxplot, e determinare la moda (o le mode)

Risposta: I risultati numerici sono

$$m = \frac{239}{50} \simeq 4.78 \qquad s^2 = \frac{41\,529}{2\,500} \simeq 16,61 \qquad s = \sqrt{\frac{41\,529}{2\,500}} \simeq 4.08$$
$$\Delta = 15 \qquad q_{1/2} = 4 \qquad q_{1/4} = 2 \qquad q_{3/4} = 7$$
$$m_A \simeq 2.35 \qquad m_Q \simeq 6.28 \qquad m_G \simeq 3.32$$

Il diagramma a barre e il *boxplot* sono rappresentati in Figura A.7. Le mode sono molte, ma le più significative sono $1, 2$ e 4 □

Esercizio A.2.15. *$n = 53$ misure di una quantità aleatoria forniscono i seguenti risultati riportati in ordine crescente:*

1.70	1.93	2.42	2.52	2.59	2.66	2.72	2.76	2.88	3.01
3.05	3.12	3.12	3.15	3.15	3.17	3.32	3.36	3.40	3.54
3.63	3.71	3.71	3.72	3.81	3.95	4.01	4.01	4.04	4.04
4.07	4.07	4.15	4.17	4.42	4.43	4.46	4.52	4.56	4.78
4.83	5.13	5.15	5.15	5.30	5.33	5.33	5.39	5.44	5.61
5.66	5.83	6.90							

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati nei seguenti 8 intervalli di ampiezze differenti

[0.0 , 2.0]	(2.0 , 3.0]	(3.0 , 3.5]	(3.5 , 4.0]
(4.0 , 4.5]	(4.5 , 5.0]	(5.0 , 6.0]	(6.0 , 8.0]

disegnare l'istogramma e determinare la class modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Soluzione: Per campioni di caratteri continui, come quelli di questo esercizio, conviene stilare una tabella di frequenze (per semplicità omettiamo quelle cumulate) completata con le ampiezze $b_k - a_k$ delle classi, le altezze H_k dell'istogramma delle frequenze assolute, i valori centrali $\widehat{w}_k$, e alcuni elementi necessari per il calcolo delle medie e varianze per dati raggruppati:

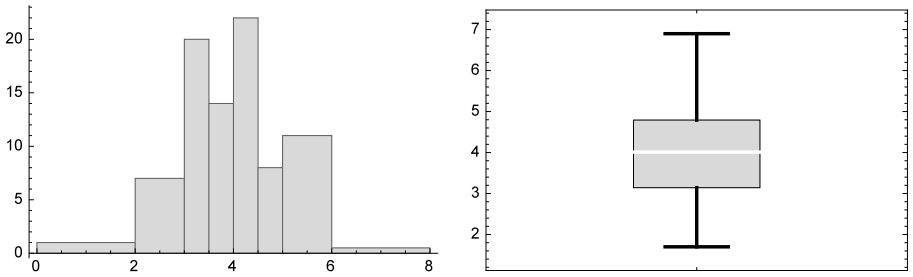


Figura A.8: Istogramma e *boxplot* dell'Esercizio A.2.15

$[a_k, b_k]$	N_k	p_k	$b_k - a_k$	H_k	$\widehat{w}_k$	$p_k\widehat{w}_k$	$p_k\widehat{w}_k^2$
[0.0, 2.0]	2	0.038	2.0	1.0	1.00	0.038	0.038
[2.0, 3.0]	7	0.132	1.0	7.0	2.50	0.330	0.825
[3.0, 3.4]	10	0.189	0.5	20.0	3.25	0.613	1.993
[3.5, 4.0]	7	0.132	0.5	14.0	3.75	0.495	1.857
[4.0, 4.5]	11	0.208	0.5	22.0	4.25	0.882	3.749
[4.5, 5.0]	4	0.075	0.5	8.0	4.75	0.358	1.703
[5.0, 6.0]	11	0.208	1.0	11.0	5.50	1.142	6.278
[6.0, 8.0]	1	0.019	2.0	0.5	7.00	0.132	0.925
						3.990	17.368

Innanzitutto si calcola facilmente la media aritmentica del campione

$$m \simeq 3.98$$

Poi si determinano il range

$$\Delta = 6.90 - 1.70 = 5.20$$

e le posizioni dei quantili

$$\text{mediana} \quad \frac{n+1}{2} = 27 \qquad 1^o \text{ e } 3^o \text{ quartile} \quad \frac{n+1}{4} = 13.5\,, \quad 3\,\frac{n+1}{4} = 40.5$$

Pertanto le mediana è il 27^o elemento del campione ordinato, mentre per i due quartili bisogna calcolare la media aritmetica del 13^o e 14^o elemento, e del 40^o e 41^o elemento: si ottiene così

$$q_{1/2} = 4.01 \qquad q_{1/4} = 3.135 \qquad q_{3/4} = 4.805$$

In fondo alle ultime due colonne sono state riportate le loro somme che corrispondono rispettivamente alla media e alla media dei quadrati dei valori centrali delle classi ($\widehat{w}_k$): da essi è possibile calcolare facilmente la media (6.9) e la varianza (6.14) per dati raggruppati

$$\widehat{m} \simeq 3.99 \qquad \widehat{s}^2 \simeq 1.44$$

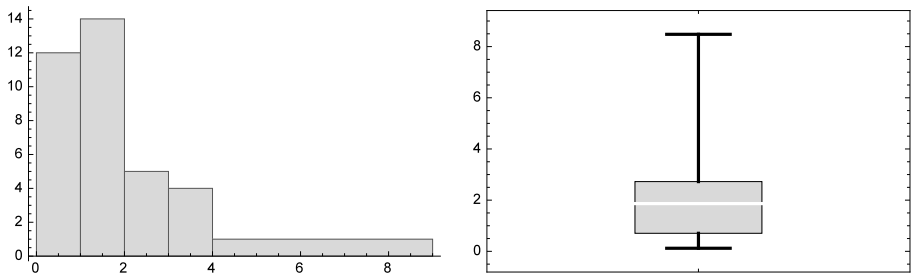


Figura A.9: Istogramma e *boxplot* dell'Esercizio A.2.16

Infine sempre dai dati della tabella delle frequenze è facile tracciare l'istogramma e il *boxplot* che sono riportati in Figura A.8; le classi modali sono (3.0 , 3.5], (4.0 , 4.5] e (5.0 , 6.0] □

Esercizio A.2.16. $n = 40$ *misure di una quantità aleatoria forniscono i seguenti risultati:*

0.12	0.13	0.18	0.18	0.21	0.25	0.30	0.35	0.46	0.54
0.87	0.92	1.10	1.19	1.43	1.45	1.47	1.67	1.79	1.84
1.89	1.90	1.91	1.91	1.97	1.98	2.09	2.26	2.35	2.70
2.75	3.39	3.58	3.62	3.89	4.20	5.50	6.43	6.96	8.48

Calcolare la media m*, il range* Δ*, la mediana* $q_{1/2}$*, i due quartili* $q_{1/4}$*,* $q_{3/4}$*, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati nei seguenti intervalli di ampiezze differenti*

$$[0,1] \quad (1,2] \quad (2,3] \quad (3,4] \quad (4,9]$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ *e la varianza* $\widehat{s}^2$ *approssimate per dati raggruppati*

Risposta: I risultati numerici sono

$m \simeq 2.16$ $\quad \Delta = 8.36$ $\quad q_{1/2} = 1.865$ $\quad q_{1/4} = 0.705$ $\quad q_{3/4} = 2.725$

$\widehat{m} \simeq 2.15$ $\quad \widehat{s}^2 \simeq 3.53$

L'istogramma e il *boxplot* sono riportati in Figura A.9 e la classe modale è (1 , 2] □

Esercizio A.2.17. $n = 40$ *misure di una quantità aleatoria forniscono i seguenti risultati:*

0.04	0.12	0.23	0.37	0.47	0.59	0.64	0.76	0.80	0.97
0.99	1.01	1.08	1.11	1.22	1.33	1.53	1.61	1.63	1.98
2.03	2.19	2.25	2.36	2.77	2.96	3.05	3.10	3.34	3.79
3.85	4.56	5.27	5.79	5.82	6.41	7.88	7.99	8.16	9.87

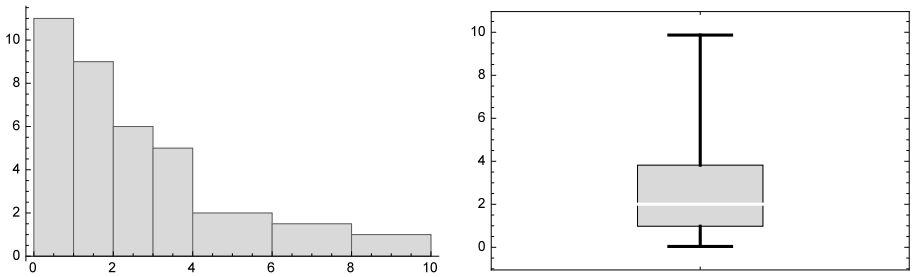


Figura A.10: Istogramma e *boxplot* dell'Esercizio A.2.17

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati nei seguenti intervalli

$$[0,1] \quad (1,2] \quad (2,3] \quad (3,4] \quad (4,6] \quad (6,8] \quad (8,10]$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$m \simeq 2.80$ $\quad \Delta = 9.83$ $\quad q_{1/2} = 2.005$ $\quad q_{1/4} = 0.98$ $\quad q_{3/4} = 3.82$
$\widehat{m} \simeq 2.76$ $\quad \widehat{s}^2 \simeq 5.64$

L'istogramma e il *boxplot* sono riportati in Figura A.10; la classe modale è $(0\,,\,1]$ □

Esercizio A.2.18. *$n = 40$ misure di velocità del vento in una stazione meteorologica forniscono i seguenti risultati:*

0.12	0.23	0.24	0.50	0.50	0.56	0.77	1.01	1.03	1.10
1.40	1.45	1.64	1.68	1.72	1.72	1.81	1.83	1.84	2.14
2.28	2.31	2.34	2.43	2.55	2.91	3.00	3.12	3.53	3.59
3.94	4.17	4.70	4.73	5.02	5.07	6.73	6.97	7.09	9.74

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati nei seguenti intervalli

$$[0\,,\,2] \qquad (2\,,\,4] \qquad (4\,,\,6] \qquad (6\,,\,8] \qquad (8\,,\,10]$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

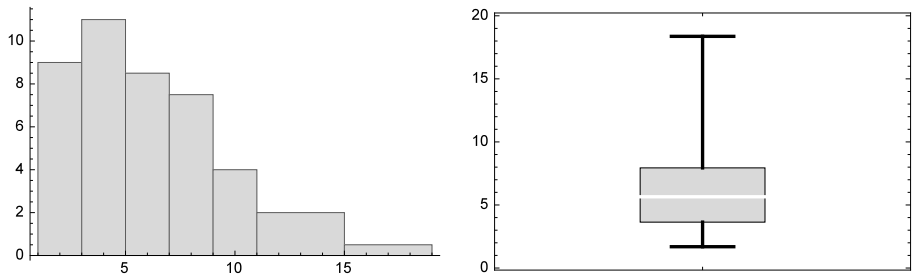


Figura A.11: Istogramma e *boxplot* dell'Esercizio A.2.19

Risposta: I risultati numerici sono

$m \simeq 2.74$ $\quad \Delta = 9.62$ $\quad q_{1/2} = 2.21$ $\quad q_{1/4} = 1.25$ $\quad q_{3/4} = 3.765$
$\widehat{m} \simeq 2.75$ $\quad \widehat{s}^2 \simeq 4.44$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $(0\,,\,2]$ □

Esercizio A.2.19. *Un'azienda vuol condurre un'indagine sulla propria clientela misurando i consumi di un determinato prodotto in* 90 *famiglie. Si ottenengono i seguenti risultati:*

1.69	1.71	1.72	1.92	1.99	2.01	2.04	2.06	2.09	2.50
2.52	2.53	2.55	2.66	2.81	2.82	2.93	2.98	3.04	3.36
3.48	3.54	3.64	3.71	3.75	3.82	3.85	3.91	4.01	4.11
4.23	4.23	4.23	4.36	4.38	4.54	4.55	4.59	4.76	4.83
5.10	5.19	5.29	5.29	5.61	5.69	5.71	5.78	5.91	5.92
5.93	6.08	6.22	6.38	6.70	6.85	6.89	7.00	7.06	7.09
7.23	7.42	7.45	7.55	7.66	7.80	7.81	7.94	7.96	7.99
8.77	8.88	9.10	9.20	9.38	9.41	9.75	10.02	10.07	10.41
11.28	11.38	11.82	11.86	12.37	12.53	13.54	14.22	15.80	18.37

Calcolare la media m*, il range* Δ*, la mediana* $q_{1/2}$*, i due quartili* $q_{1/4}$*,* $q_{3/4}$*, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti* 1, 3, 5, 7, 9, 11, 15, 19*, disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media* $\widehat{m}$ *e la varianza* $\widehat{s}^2$ *approssimate per dati raggruppati*

Risposta: I risultati numerici sono

$m \simeq 6.19$ $\quad \Delta = 16.68$ $\quad q_{1/2} = 5.65$ $\quad q_{1/4} = 3.59$ $\quad q_{3/4} = 7.95$
$\widehat{m} \simeq 6.27$ $\quad \widehat{s}^2 \simeq 13.24$

L'istogramma e il *boxplot* sono riportati in Figura A.11; la classe modale è $(3\,,\,5]$ □

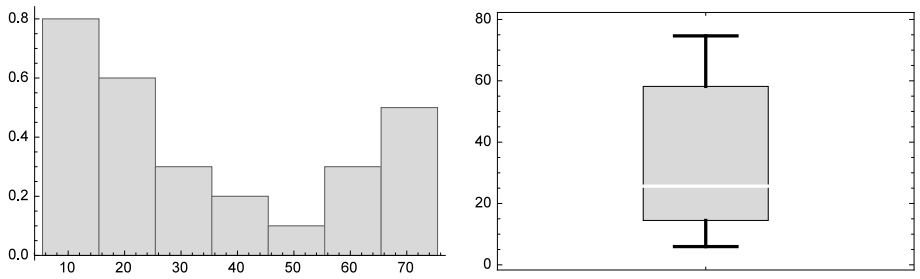


Figura A.12: Istogramma e *boxplot* dell'Esercizio A.2.21

Esercizio A.2.20. *La seguente tabella contiene i pesi in grammi di* 40 *prodotti:*

21.3	21.6	21.8	21.8	22.1	22.2	22.2	22.2	22.4	22.4
22.5	22.5	22.5	22.6	22.6	22.8	22.8	22.8	22.9	23.0
23.0	23.0	23.1	23.2	23.2	23.4	23.5	23.5	23.5	23.6
23.8	23.8	23.9	23.9	24.0	24.0	24.3	24.6	24.6	24.9

Calcolare la media m*, il range* Δ*, la mediana* $q_{1/2}$*, i due quartili* $q_{1/4}$*,* $q_{3/4}$*, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati in* 8 *classi di ampiezza* 0.5 *partendo dall'intervallo* $[20.95\,,\,21.45]$*, disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media* $\widehat{m}$ *e la varianza* $\widehat{s}^2$ *approssimate per dati raggruppati*

Risposta: I risultati numerici sono

$m \simeq 23.0$ $\quad \Delta = 3.6$ $\quad q_{1/2} = 23.0$ $\quad q_{1/4} = 22.5$ $\quad q_{3/4} = 23.7$

$\widehat{m} \simeq 23.1$ $\quad \widehat{s}^2 \simeq 0.75$

L'istogramma e il *boxplot* non sono riportati; le classi modali sono $[22.45\,,\,22.95]$ e $[23.45\,,\,23.95]$ □

Esercizio A.2.21. $n = 28$ *misure di una quantità aleatoria forniscono i seguenti risultati (non ordinati):*

14.00	5.99	26.35	35.95	15.95	24.95	19.95	32.95	59.00	9.95
69.95	61.35	14.95	12.95	16.95	10.95	57.35	29.95	5.95	41.95
66.95	19.85	11.95	15.95	50.25	74.65	68.00	69.95		

Calcolare la media m*, il range* Δ*, la mediana* $q_{1/2}$*, i due quartili* $q_{1/4}$*,* $q_{3/4}$*, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati in* 7 *classi di ampiezza* 10 *partendo dall'intervallo* $[5.455\,,\,15.455]$*, disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media* $\widehat{m}$ *e la varianza* $\widehat{s}^2$ *approssimate per dati raggruppati*

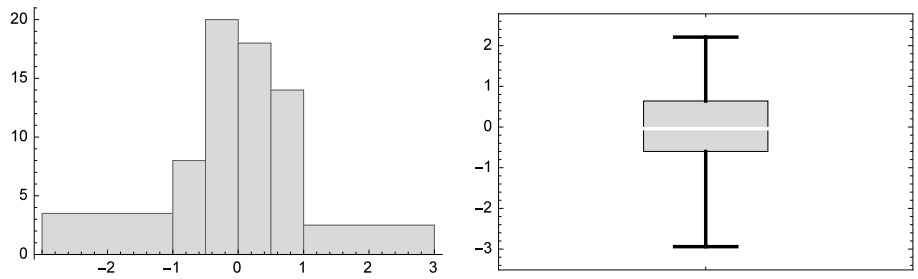


Figura A.13: Istogramma e *boxplot* dell'Esercizio A.2.23

Risposta: I risultati numerici sono

$m \simeq 33.75$ $\quad \Delta = 68.70$ $\quad q_{1/2} = 25.65$ $\quad q_{1/4} = 14.475$ $\quad q_{3/4} = 58.175$
$\widehat{m} \simeq 34.38$ $\quad \widehat{s}^2 \simeq 523.85$

L'istogramma e il *boxplot* sono riportati in Figura A.12; le classi modali sono $[5.455\,,\,15.455]$ e $[65.455\,,\,75.455]$ □

Esercizio A.2.22. *Vengono eseguite* $n = 50$ *misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.04	0.04	0.07	0.11	0.11	0.14	0.17	0.19	0.19	0.20
0.25	0.32	0.36	0.43	0.52	0.55	0.62	0.64	0.67	0.88
0.88	0.97	1.04	1.09	1.10	1.16	1.20	1.22	1.22	1.32
1.49	1.62	1.67	1.87	1.89	1.99	2.17	2.21	2.53	2.60
2.65	2.68	2.90	3.18	3.50	4.57	5.03	5.91	7.56	9.20

Calcolare la media m*, il range* Δ*, la mediana* $q_{1/2}$*, i due quartili* $q_{1/4}$*,* $q_{3/4}$*, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti* 0, 1, 2, 3, 5, 10*, disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media* $\widehat{m}$ *e la varianza* $\widehat{s}^2$ *approssimate per dati raggruppati*

Risposta: I risultati numerici sono

$m \simeq 1.70$ $\quad \Delta = 9.16$ $\quad q_{1/2} = 1.13$ $\quad q_{1/4} = 0.34$ $\quad q_{3/4} = 2.37$
$\widehat{m} \simeq 1.83$ $\quad \widehat{s}^2 \simeq 3.73$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $[0,\,1]$ □

Esercizio A.2.23. *Vengono eseguite $n = 42$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

−2.94	−1.62	−1.48	−1.41	−1.04	−1.02	−1.01	−0.86	−0.78	−0.64
−0.60	−0.46	−0.43	−0.40	−0.40	−0.30	−0.29	−0.28	−0.24	−0.22
−0.10	0.02	0.07	0.15	0.15	0.15	0.17	0.18	0.23	0.38
0.61	0.64	0.89	0.90	0.93	0.98	0.99	1.24	1.30	1.48
1.57	2.21								

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$-3.0 \qquad -1.0 \qquad -0.5 \qquad 0.0 \qquad 0.5 \qquad 1.0 \qquad 3.0$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$$m \simeq -0.03 \qquad \Delta = 5.15 \qquad q_{1/2} = -0.04 \qquad q_{1/4} = -0.62 \qquad q_{3/4} = 0.765$$
$$\widehat{m} \simeq -0.05 \qquad \widehat{s}^2 \simeq 1.32$$

L'istogramma e il *boxplot* sono riportati in Figura A.13; la classe modale è $(-0.5\,,\,0.0]$
□

Esercizio A.2.24. *Vengono eseguite $n = 50$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.02	0.04	0.10	0.16	0.17	0.23	0.28	0.34	0.35	0.36
0.43	0.56	0.56	0.64	0.64	0.80	0.83	0.85	0.87	0.91
0.96	1.05	1.07	1.19	1.21	1.23	1.38	1.68	1.74	1.77
1.81	1.86	1.87	1.95	2.18	2.32	2.52	2.59	2.59	2.68
2.93	3.31	3.80	3.97	4.12	4.46	5.05	6.44	6.69	9.82

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti 0, 1, 2, 3, 4, 5, 10*, disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati*

Risposta: I risultati numerici sono

$$m \simeq 1.91 \qquad \Delta = 9.80 \qquad q_{1/2} = 1.22 \qquad q_{1/4} = 0.56 \qquad q_{3/4} = 2.59$$
$$\widehat{m} \simeq 1.94 \qquad \widehat{s}^2 \simeq 3.85$$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $[0,\,1]$ □

Esercizio A.2.25. *Vengono eseguite $n = 38$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

1.61	1.86	2.42	2.72	2.83	2.84	2.87	2.98	3.06	3.07
3.22	3.33	3.33	3.40	3.40	3.70	3.77	3.78	3.81	3.84
3.85	3.95	3.95	4.14	4.17	4.36	4.36	4.43	4.52	4.57
4.59	4.70	4.75	4.81	5.05	5.06	5.46	5.86		

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$1.0 \qquad 2.5 \qquad 3.0 \qquad 3.5 \qquad 4.0 \qquad 4.5 \qquad 6.0$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$$m \simeq 3.80 \qquad \Delta = 4.25 \qquad q_{1/2} = 3.825 \qquad q_{1/4} = 3.065 \qquad q_{3/4} = 4.545$$
$$\widehat{m} \simeq 3.83 \qquad \widehat{s}^2 \simeq 1.11$$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $(3.5\,,\,4.0]$ □

Esercizio A.2.26. *Vengono eseguite $n = 39$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.04	0.16	0.26	0.29	0.30	0.32	0.32	0.36	0.39	0.42
0.43	0.47	0.48	0.51	0.51	0.51	0.53	0.53	0.54	0.60
0.61	0.63	0.66	0.66	0.70	0.70	0.75	0.78	0.78	0.79
0.79	0.83	0.88	0.89	0.89	0.89	0.90	0.94	0.96	

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$0.005 \qquad 0.305 \qquad 0.505 \qquad 0.605 \qquad 0.705 \qquad 0.805 \qquad 1.005$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$$m \simeq 0.59 \qquad \Delta = 0.92 \qquad q_{1/2} = 0.60 \qquad q_{1/4} = 0.42 \qquad q_{3/4} = 0.79$$
$$\widehat{m} \simeq 0.59 \qquad \widehat{s}^2 \simeq 0.06$$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $(0.505\,,\,0.605]$ □

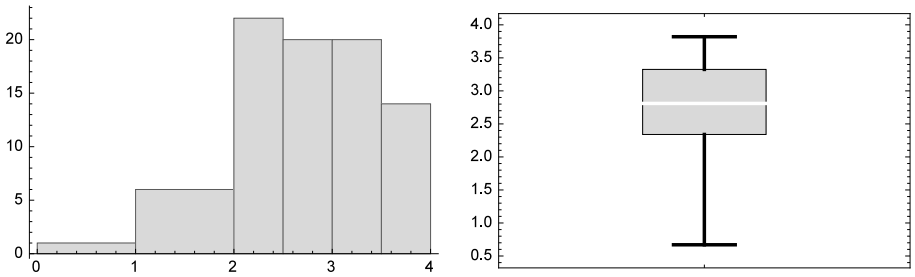


Figura A.14: Istogramma e *boxplot* dell'Esercizio A.2.27

Esercizio A.2.27. *Vengono eseguite $n = 45$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.67	1.19	1.25	1.56	1.59	1.83	1.85	2.03	2.20	2.21
2.25	2.37	2.41	2.44	2.45	2.45	2.48	2.48	2.52	2.63
2.66	2.72	2.81	2.81	2.84	2.89	2.91	2.92	3.07	3.12
3.20	3.26	3.27	3.32	3.34	3.37	3.46	3.48	3.55	3.58
3.66	3.66	3.76	3.78	3.82					

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$0.0 \qquad 1.0 \qquad 2.0 \qquad 2.5 \qquad 3.0 \qquad 3.5 \qquad 4.0$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$$m \simeq 2.71 \qquad \Delta = 3.15 \qquad q_{1/2} = 2.91 \qquad q_{1/4} = 2.31 \qquad q_{3/4} = 3.33$$
$$\widehat{m} \simeq 2.68 \qquad \widehat{s}^2 \simeq 0.59$$

Istogramma e *boxplot* sono riportati in Figura A.14; la classe modale è $(2.0\,,\,2.5]$ □

Esercizio A.2.28. *Vengono eseguite $n = 40$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

−1.64	−1.63	−1.62	−1.62	−1.43	−1.33	−1.14	−1.13	−1.08	−0.87
−0.81	−0.63	−0.57	−0.51	−0.44	−0.39	−0.27	−0.26	−0.23	−0.21
−0.20	−0.18	−0.13	−0.05	0.03	0.09	0.13	0.18	0.22	0.39
0.41	0.53	0.71	0.73	1.01	1.08	1.10	1.12	1.57	2.27

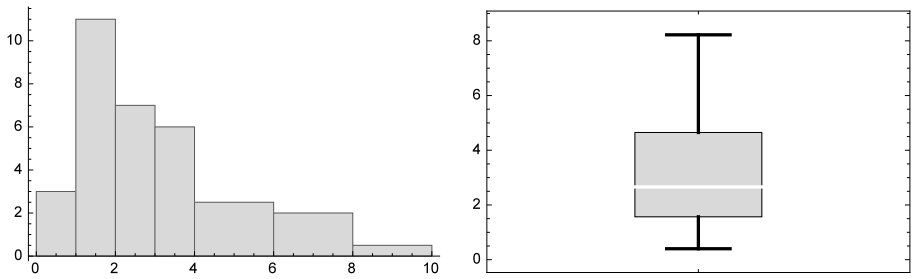


Figura A.15: Istogramma e *boxplot* dell'Esercizio A.2.29

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$-3.0 \qquad -1.0 \qquad -0.5 \qquad 0.0 \qquad 0.5 \qquad 1.0 \qquad 3.0$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$$m \simeq -0.17 \qquad \Delta = 3.91 \qquad q_{1/2} = -0.205 \qquad q_{1/4} = -0.84 \qquad q_{3/4} = 0.40$$
$$\widehat{m} \simeq 0.21 \qquad \widehat{s}^2 \simeq 1.60$$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $(-0.5\,,\,0.0]$ □

Esercizio A.2.29. *Vengono eseguite $n = 37$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.40	0.78	0.91	1.06	1.25	1.31	1.33	1.40	1.53	1.58
1.65	1.77	1.85	1.92	2.03	2.07	2.15	2.39	2.66	2.87
2.98	3.14	3.15	3.16	3.66	3.71	3.82	4.59	4.83	5.67
5.78	5.99	6.39	6.59	6.81	7.93	8.22			

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti 0, 1, 2, 3, 4, 6, 8, 10*, disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati*

Risposta: I risultati numerici sono

$m \simeq 3,23$ $\quad \Delta = 7,82$ $\quad q_{1/2} = 2.66$ $\quad q_{1/4} = 1.555$ $\quad q_{3/4} = 4.71$
$\widehat{m} \simeq 3.20$ $\quad \widehat{s}^2 \simeq 4.47$

Istogramma e *boxplot* sono riportati in Figura A.15; la classe modale è (1, 2] □

Esercizio A.2.30. *Vengono eseguite $n = 41$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.78	1.53	1.53	1.65	1.66	1.72	1.80	1.93	1.94	2.02
2.07	2.26	2.27	2.34	2.44	2.54	2.64	2.67	2.76	2.80
2.84	2.88	2.93	2.94	2.95	2.96	3.06	3.17	3.18	3.29
3.38	3.42	3.43	3.45	3.46	3.52	3.58	3.65	3.71	3.76	3.95

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$0.0 \qquad 1.0 \qquad 2.0 \qquad 2.5 \qquad 3.0 \qquad 3.5 \qquad 4.0$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

Risposta: I risultati numerici sono

$m \simeq 2.70$ $\quad \Delta = 3.17$ $\quad q_{1/2} = 2.84$ $\quad q_{1/4} = 2.045$ $\quad q_{3/4} = 3.40$
$\widehat{m} \simeq 2.63$ $\quad \widehat{s}^2 \simeq 0.65$

L'istogramma e il *boxplot* non sono riportati; la classe modale è (2.5 , 3.0] □

Esercizio A.2.31. *Vengono eseguite $n = 43$ misure di una quantità aleatoria ottenendo i seguenti risultati riportati in ordine crescente:*

0.08	0.11	0.20	0.22	0.34	0.45	0.53	0.58	0.70	0.75
0.78	0.81	0.86	1.12	1.15	1.18	1.19	1.22	1.25	1.28
1.34	1.35	1.40	1.42	1.43	1.44	1.51	1.51	1.52	1.54
1.59	1.59	1.61	1.64	1.71	1.72	1.76	1.9	2.07	2.09
2.15	2.60	2.80							

Calcolare la media m, il range Δ, la mediana $q_{1/2}$, i due quartili $q_{1/4}$, $q_{3/4}$, e disegnare il boxplot. Costruire la tabella delle frequenze (assolute e relative) dei ritrovamenti dei dati negli intervalli delimitati dai punti

$$0 \qquad {}^{1}/_{2} \qquad 1 \qquad {}^{4}/_{3} \qquad {}^{5}/_{3} \qquad 2 \qquad {}^{5}/_{2} \qquad 3$$

disegnare l'istogramma e determinare la classe modale (o le classi modali). Usando poi la tabella delle frequenze relative, calcolare la media $\widehat{m}$ e la varianza $\widehat{s}^2$ approssimate per dati raggruppati

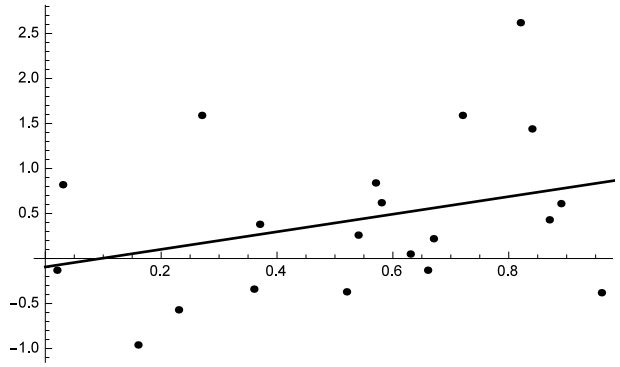


Figura A.16: *Scatterplot* e retta di regressione dell'Esercizio A.2.32

Risposta: I risultati numerici sono

$$m \simeq 1.27 \qquad \Delta = 2.72 \qquad q_{1/2} = 1.35 \qquad q_{1/4} = 0.78 \qquad q_{3/4} = 1.61$$
$$\widehat{m} \simeq 1.29 \qquad \widehat{s}^2 \simeq 0.41$$

L'istogramma e il *boxplot* non sono riportati; la classe modale è $(\,^4/_3\,,\,^5/_3]$ □

Esercizio A.2.32. *$n = 20$ misure di due caratteri X e Y forniscono i seguenti risultati*

X	Y	X	Y	X	Y	X	Y
0.02	−0.13	0.36	−0.34	0.58	0.62	0.82	2.62
0.03	0.82	0.37	0.38	0.63	0.05	0.84	1.44
0.16	−0.96	0.52	−0.37	0.66	−0.13	0.87	0.43
0.23	−0.57	0.54	0.26	0.67	0.22	0.89	0.61
0.27	1.59	0.57	0.84	0.72	1.59	0.96	−0.38

Calcolare le medie m_X, m_Y e le varianze s_X^2, s_Y^2 di X e Y, la covarianza s_{XY}, il coefficiente di correlazione r_{XY} e i coefficienti a e b della retta di regressione

Risposta: I risultati numerici sono

$$m_X \simeq 0.54 \qquad m_Y \simeq 0.43 \qquad s_X^2 \simeq 0.08 \qquad s_Y^2 \simeq 0.73$$
$$s_{XY} \simeq 0.08 \qquad r_{XY} \simeq 0.32 \qquad a \simeq 0.98 \qquad b \simeq -0.09$$

Lo *scatterplot* dei dati e la retta di regressione sono riportati a scopo illustrativo nella Figura A.16 □

Esercizio A.2.33. *$n = 28$ misure di due caratteri X e Y forniscono i seguenti risultati*

X	Y	X	Y	X	Y	X	Y
0.04	1.41	0.28	0.53	0.45	1.05	0.65	0.26
0.07	−0.31	0.29	0.39	0.45	1.23	0.65	1.08
0.18	−1.18	0.30	−0.89	0.51	0.40	0.83	1.88
0.18	−1.01	0.30	1.20	0.55	1.04	0.88	0.48
0.21	−0.15	0.39	−0.52	0.57	−0.91	0.88	1.42
0.22	−0.31	0.42	1.79	0.58	−1.33	0.89	0.36
0.22	0.32	0.45	0.74	0.62	1.73	0.97	−0.37

Calcolare le medie m_X, m_Y e le varianze s_X^2, s_Y^2 di X e Y, la covarianza s_{XY}, il coefficiente di correlazione r_{XY} e i coefficienti a e b della retta di regressione

Risposta: I risultati numerici sono

$$m_X \simeq 0.47 \qquad m_Y \simeq 0.37 \qquad s_X^2 \simeq 0.07 \qquad s_Y^2 \simeq 0.87$$
$$s_{XY} \simeq 0.06 \qquad r_{XY} \simeq 0.25 \qquad a \simeq 0.91 \qquad b \simeq -0.06$$

Lo *scatterplot* dei dati e la retta di regressione non sono riportati □

A.3 Statistica Inferenziale

Esercizio A.3.1. *Le misure di una quantità X sono soggette ad errori casuali: $n = 15$ valori delle misure sono*

2.05 1.42 6.18 6.69 4.47 5.36 3.47 6.74 5.19 0.91
3.22 9.50 5.85 3.41 5.66

Determinare l'intervallo di fiducia di livello $\alpha = 0.05$ per l'attesa

Soluzione: Non essendo nota la varianza, si calcolano innanzitutto le quantità

$$\overline{X} \simeq 4.675 \qquad S^2 = \frac{n}{n-1}\left(\overline{X^2} - \overline{X}^2\right) \simeq 5.273 \qquad S \simeq 2.296$$

poi con $\alpha = 0.05$ dalle Tavole E.2 di Student si trova

$$t_{1-\frac{\alpha}{3}}(n-1) = t_{0.975}(14) \simeq 2.145$$

e infine si calcolano gli estremi dell'intervallo di fiducia

$$\overline{X} \pm \frac{S}{\sqrt{n}}\, t_{1-\frac{\alpha}{3}}(n-1) \simeq 4.675 \pm 1.272$$

come indicato nell'equazione (8.5) □

Esercizio A.3.2. $n = 10$ *misure di una v-a* X *con attesa e varianza sconosciute danno i seguenti risultati*

$$1.01 \quad 2.25 \quad 1.60 \quad 1.75 \quad 1.49 \quad 1.45 \quad 2.51 \quad 1.87 \quad 3.95 \quad 2.10$$

Determinare l'intervallo di fiducia di livello $\alpha = 0.02$ *per l'attesa*

Soluzione: Non essendo nota la varianza, si calcolano innanzitutto le quantità

$$\overline{X} \simeq 1.998 \qquad S \simeq 0.811$$

poi con $\alpha = 0.02$ dalle Tavole E.2 di Student si trova

$$t_{1-\frac{\alpha}{3}}(n-1) = t_{0.990}(9) \simeq 2.821$$

e infine si calcolano gli estremi dell'intervallo di fiducia

$$\overline{X} \pm \frac{S}{\sqrt{n}}\, t_{1-\frac{\alpha}{2}}(n-1) \simeq 1.998 \pm 0.723$$

come indicato nell'equazione (8.5) □

Esercizio A.3.3. $n = 18$ *misure di una v-a* X *forniscono i seguenti risultati*

$$\begin{array}{cccccccccc} 4.09 & 4.56 & 5.01 & 5.49 & 4.82 & 5.56 & 3.95 & 4.04 & 2.63 \\ 3.78 & 3.58 & 4.52 & 4.86 & 3.65 & 4.44 & 4.62 & 3.97 & 3.63 \end{array}$$

Supponendo che la varianza $\sigma^2 = 1$ *sia conosciuta, determinare prima l'intervallo di fiducia di livello* $\alpha = 0.05$ *per l'attesa* μ*; eseguire poi un test bilaterale di Gauss di livello* $\alpha = 0.05$ *per decidere fra le due ipotesi*

$$\mathcal{H}_0 \,:\, \mu = 4 \qquad \mathcal{H}_1 \,:\, \mu \neq 4$$

e calcolarne la significatività α_s

Soluzione: Si calcola innanzitutto $\overline{X} \simeq 4.289$; poi con $\alpha = 0.05$ dalle Tavole E.1 della normale standard si trova che

$$\varphi_{1-\frac{\alpha}{2}} = \varphi_{0.975} \simeq 1.960$$

per cui dall'equazione (8.4) l'intervallo di fiducia è

$$\overline{X} \pm \frac{\sigma}{\sqrt{n}}\, \varphi_{1-\frac{\alpha}{2}} \simeq 4.289 \pm 0.462$$

Per il test si osserva poi innanzitutto che con $\mu_0 = 4$ dall'equazione (9.15) si ha

$$|U_0| = \left| \sqrt{n}\, \frac{\overline{X} - \mu_0}{\sigma} \right| \simeq 1.226$$

per cui essendo

$$|U_0| \simeq 1.226 < 1.960 \simeq \varphi_{1-\frac{\alpha}{2}}$$

l'evento critico D (9.16) non si verifica, cioè i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$. Sempre dalle Tavole E.1 potremo poi anche calcolare la significatività del test

$$\alpha_s = 2\,[1 - \Phi(|U_0|)] \simeq 2\,[1 - \Phi(1.226)] \simeq 0.22$$

che non è molto buona principalmente a causa del piccolo numero di dati a disposizione □

Esercizio A.3.4. *$n = 20$ misure di una quantità aleatoria X forniscono i seguenti risultati:*

2.23	4.09	3.97	5.57	3.09	3.00	2.85	2.12	3.26	2.11
3.10	1.82	2.82	1.99	3.25	1.53	2.99	1.03	3.86	2.45

Supponendo che media μ e varianza σ^2 non siano note, eseguire un test bilaterale di livello $\alpha = 0.05$ per decidere tra le due ipotesi

$$\mathcal{H}_0 \,:\, \mu = 3 \qquad\qquad \mathcal{H}_1 \,:\, \mu \neq 3$$

Soluzione: Siccome media e varianza non sono note dovremo eseguire un test di Student: calcoleremo quindi innanzitutto dai dati le stime puntuali

$$\overline{X} \simeq 2.857 \qquad\qquad S \simeq 1.031$$

poi con $\alpha = 0.05$ dalle Tavole E.2 si trova

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.975}(19) \simeq 2.093$$

che fissa i limiti dell'evento critico, e infine con $\mu_0 = 3$ si calcola la statistica di Student (9.18)

$$|T_0| = \left| \sqrt{n}\, \frac{\overline{X} - \mu_0}{S} \right| \simeq 0.623$$

A questo punto si osserva che

$$|T_0| \simeq 0.623 < 2.093 \simeq t_{1-\frac{\alpha}{2}}(n-1)$$

per cui l'evento critico D (9.19) non si verifica, cioè i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.5. *Le misure di pressione di un campione di $n = 200$ pneumatici di automobile hanno una media $\overline{X} = 33.57$ e una varianza $S^2 = 1.723$. Decidere tra le due ipotesi*

$$\mathcal{H}_0 \,:\, \mu = 34 \qquad\qquad \mathcal{H}_1 \,:\, \mu \neq 34$$

con un test bilaterale di livello $\alpha = 0.01$

Soluzione: I valori dati per la media e la varianza devono essere considerati come calcolati dal campione di 200 misure che non è riportato: pertanto la varianza deve essere considerata non come conosciuta, ma come ricavata dai dati empirici (infatti con un altro campione di 200 misure il valore di S^2 sarebbe diverso). Pertanto dovremo eseguire un test di Student, ma sulle Tavole E.2 i quantili delle leggi con $n-1=199$ gradi di libertà non sono riportati. Bisognerà quindi usare l'approssimazione (3.29) per calcolare questi quantili da quelli della normale standard delle Tavole E.1, sicché con $\alpha=0.01$ avremo

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.995}(199) \simeq \varphi_{0.995} \simeq 2.58$$

A questo punto con $\mu_0=34$ si calcola il valore della statistica di Student (9.18)

$$|T_0| = \left|\sqrt{n}\,\frac{\overline{X}-\mu_0}{S}\right| \simeq 4.63$$

e si osserva che

$$|T_0| \simeq 4.63 > 2.58 \simeq t_{1-\frac{\alpha}{2}}(n-1)$$

per cui l'evento critico D (9.19) si verifica, cioè i dati sono in regione critica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.6. *$n=20$ misure dell'energia cinetica delle particelle di un gas forniscono – nelle opportune unità di misura – i seguenti risultati:*

2.58	0.25	3.96	4.89	3.80	1.42	0.96	7.99	2.47	4.32
2.19	0.66	1.37	6.22	1.41	2.56	1.06	1.40	0.45	1.40

Supponendo che l'attesa μ e la varianza σ^2 siano sconosciute,determinare l'intervallo di fiducia di livello $\alpha=0.05$ per l'attesa μ. Eseguire poi un test bilaterale di livello $\alpha=0.05$ per decidere fra le ipotesi

$$\mathcal{H}_0 \,:\, \mu = 3 \qquad\qquad \mathcal{H}_1 \,:\, \mu \neq 3$$

Soluzione: Siccome la varianza è sconosciuta dovremo innanzitutto stimare le quantità necessarie:

$$\overline{X} \simeq 2.568 \qquad S^2 \simeq 4.210 \qquad S \simeq 2.052$$

e poi determinare con $\alpha=0.05$ l'opportuno quantile di Student dalle Tavolr E.2:

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.975}(19) \simeq 2.093$$

A questo punto l'intervallo di fiducia è

$$\overline{X} \pm \frac{S}{\sqrt{n}}\, t_{1-\frac{\alpha}{2}}(n-1) \simeq 2.568 \pm 0.960$$

mentre la statistica di Student (9.18) con $\mu_0 = 3$ vale

$$|T_0| = \left| \sqrt{n}\, \frac{\overline{X} - \mu_0}{S} \right| \simeq 0.942$$

Pertanto

$$|T_0| \simeq 0.942 < 2.093 \simeq t_{1-\frac{\alpha}{2}}(n-1)$$

per cui l'evento critico D (9.19) non si verifica, cioè i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.7. *$n = 10$ misure di una v-a X con media e varianza sconosciute danno i seguenti risultati:*

$$2.47 \quad 1.79 \quad 0.01 \quad 2.94 \quad 4.10 \quad 2.13 \quad 4.51 \quad 0.72 \quad -2.99 \quad 0.83$$

Determinare l'intervallo di fiducia di livello $\alpha = 0.05$ per la media, ed eseguire un test unilaterale destro di livello $\alpha = 0.05$ per decidere fra le ipotesi

$$\mathcal{H}_0 \,:\, \mu \leq 0 \qquad\qquad \mathcal{H}_1 \,:\, \mu > 0$$

Soluzione: Si stimano innanzitutto le quantità necessarie:

$$\overline{X} \simeq 1.651 \qquad S \simeq 2.174$$

e poi dalle Tavolr E.2 si determina l'opportuno quantile di Student per l'intervallo di fiducia di livello $\alpha = 0.05$:

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.975}(9) \simeq 2.262$$

Pertanto l'intervallo di fiducia è

$$\overline{X} \pm \frac{S}{\sqrt{n}}\, t_{1-\frac{\alpha}{2}}(n-1) \simeq 1.651 \pm 1.555$$

Per il test unilaterale destro di livello $\alpha = 0.05$, invece, il quantile di Student necessario è

$$t_{1-\alpha}(n-1) = t_{0.950}(9) \simeq 1.833$$

mentre la statistica di Student (9.18) con $\mu_0 = 0$ vale

$$T_0 = \sqrt{n}\, \frac{\overline{X} - \mu_0}{S} \simeq 2.402$$

Pertanto si avrà

$$T_0 \simeq 2.402 > 1.833 \simeq t_{1-\alpha}(n-1)$$

per cui l'evento critico destro D (9.20) si verifica, cioè i dati sono in regione critica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.8. *Sia dato il seguente campione di* $n = 10$ *misure di una v-a* X *con attesa e varianza sconosciute:*

$$2.02 \quad -0.87 \quad -1.68 \quad -1.39 \quad -0.05 \quad -2.69 \quad 3.14 \quad -2.46 \quad -0.05 \quad 1.83$$

Determinare gli intervalli di fiducia di livello $\alpha = 0.05$ *per la media e per la varianza. Con un test bilaterale di livello* $\alpha = 0.05$ *si decida poi fra le due ipotesi*

$$\mathcal{H}_0 \,:\, \mu = 0 \qquad\qquad \mathcal{H}_1 \,:\, \mu \neq 0$$

Soluzione: Una volta stimate le quantità necessarie

$$\overline{X} \simeq -0.220 \qquad S^2 \simeq 3.957 \qquad S \simeq 1.989$$

si cerchino sulle Tavole E.2 e E.3 i quantili necessati per i due intervalli di fiducia di livello $\alpha = 0,05$:

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.975}(9) \simeq 2.262$$
$$\chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.025}(9) \simeq 2.700 \qquad\qquad \chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.975}(9) \simeq 19.023$$

Gli estremi dell'intervallo (8.5) per l'attesa sono

$$\overline{X} \pm \frac{S}{\sqrt{n}}\, t_{1-\frac{\alpha}{2}}(n-1) \simeq -0.220 \pm 1.432$$

mentre quelli dell'intervallo (8.7) per la varianza sono

$$\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \simeq 1.872 \qquad\qquad \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \simeq 13.187$$

Per il test bilaterale bisogna calcolare la statistica di Student (9.18) con $\mu_0 = 0$

$$|T_0| = \left|\sqrt{n}\,\frac{\overline{X} - \mu_0}{S}\right| \simeq 0.350$$

e siccome risulta

$$|T_0| \simeq 0.350 < 2.262 \simeq t_{1-\frac{\alpha}{2}}(n-1)$$

l'evento critico D (9.19) non si verifica, cioè i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.9. *Si vuol verificare se in un determinato impianto il livello medio del rumore non superi* 90 dB*: si effettuano* $n = 9$ *misure durante una giornata e si trovano i seguenti valori in* dB*:*

$$95 \quad 98 \quad 92 \quad 84 \quad 105 \quad 92 \quad 110 \quad 86 \quad 98$$

Supponendo che il livello di rumore sia una v-a X *con media e varianza sconosciute, valutare con un test unilaterale destro di livello* $\alpha = 0.05$ *quale accettare fra le due ipotesi*

$$\mathcal{H}_0 \,:\, \mu \leq 90 \qquad \mathcal{H}_1 \,:\, \mu > 90$$

Soluzione: Dopo aver calcolato le quantità necessarie

$$\overline{X} \simeq 95.56 \qquad S \simeq 8.37$$

e dopo aver determinato il quantile di Student per un test unilaterale destro di livello $\alpha = 0.05$

$$t_{1-\alpha}(n-1) = t_{0.950}(8) \simeq 1.86$$

si calcola la statistica di Student (9.18) con $\mu_0 = 90$

$$T_0 = \sqrt{n}\,\frac{\overline{X} - \mu_0}{S} \simeq 1.99$$

Siccome

$$T_0 \simeq 1.99 > 1.86 \simeq t_{1-\alpha}(n-1)$$

l'evento critico destro D (9.20) si verifica, cioè i dati sono in regione critica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$, ma l'esiguità del margine fra i due membri della precedente disuguaglianza invitano a una certa cautela nelle conclusioni □

Esercizio A.3.10. *Una fabbrica produce automezzi che consumano in media* 10 lt *di carburante ogni* 100 Km *ad una data velocità. Per verificare gli effetti di un nuovo dispositivo si misurano i litri di carburante consumati da un campione di* $n = 10$ *automezzi ottenendo i seguenti risultati:*

9.0 12.0 11.0 7.5 10.2 9.8 13.0 12.0 12.5 10.4

Supponendo che il consumo sia una v-a con media e varianza sconosciute, decidere con un test bilaterale di livello $\alpha = 0.05$ *quale accettare fra le due ipotesi*

$$\mathcal{H}_0 \,:\, \mu = 10 \qquad \mathcal{H}_1 \,:\, \mu \neq 10$$

Soluzione: Dopo aver calcolato le quantità necessarie

$$\overline{X} \simeq 10.74 \qquad S \simeq 1.71$$

e dopo aver determinato il quantile di Student per un test bilaterale di livello $\alpha = 0.05$

$$t_{1-\frac{\alpha}{2}}(n-1) = t_{0.975}(9) \simeq 2.26$$

si calcola la statistica di Student con $\mu_0 = 10$

$$|T_0| = \left|\sqrt{n}\,\frac{\overline{X} - \mu_0}{S}\right| \simeq 1.37$$

Siccome

$$|T_0| \simeq 1.37 < 2.26 \simeq t_{1-\frac{\alpha}{2}}(n-1)$$

l'evento critico D (9.19) non si verifica, cioè i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.11. *Per verificare l'efficacia di un medicinale si misura la temperatura di $n = 30$ pazienti prima (X) e dopo (Y) l'assunzione del farmaco. I campioni accoppiati così ricavati sono i seguenti:*

$$X = \begin{array}{cccccccccc} 36.8 & 37.9 & 38.3 & 38.1 & 38.3 & 38.8 & 36.3 & 37.7 & 37.1 & 36.9 \\ 37.7 & 37.5 & 37.4 & 39.0 & 38.6 & 41.4 & 37.9 & 37.5 & 39.9 & 38.2 \\ 38.4 & 37.7 & 37.5 & 36.6 & 38.0 & 38.6 & 39.0 & 37.6 & 38.0 & 39.5 \end{array}$$

$$Y = \begin{array}{cccccccccc} 37.1 & 36.1 & 38.2 & 39.0 & 38.4 & 36.4 & 36.8 & 39.3 & 37.3 & 37.9 \\ 37.6 & 38.4 & 38.6 & 37.4 & 38.3 & 37.0 & 35.9 & 36.0 & 37.2 & 39.3 \\ 36.0 & 36.6 & 38.0 & 38.0 & 38.7 & 39.4 & 37.6 & 36.1 & 39.0 & 36.4 \end{array}$$

Decidere con un test unilaterale di livello $\alpha = 0.05$ quale accettare fra le ipotesi

$$\mathcal{H}_0 \,:\, \mu_X \le \mu_Y \qquad \mathcal{H}_1 \,:\, \mu_X > \mu_Y$$

Soluzione: Trattandosi di un test unilaterale per campioni accoppiati (vedi Sezione 9.3.1) con ipotesi alternativa $\mu_X > \mu_Y$, bisognerà innanzitutto costruire il campione delle differenze

$$Z = X - Y = \begin{array}{cccccccccc} -0.3 & 1.8 & 0.1 & -0.9 & -0.1 & 2.4 & -0.5 & -1.6 & -0.2 & -1.0 \\ 0.1 & -0.9 & -1.2 & 1.6 & 0.3 & 4.4 & 2.0 & 1.5 & 2.7 & -1.1 \\ 2.4 & 1.1 & -0.5 & -1.4 & -0.7 & -0.8 & 1.4 & 1.5 & -1.0 & 3.1 \end{array}$$

con un valore d'attesa $\mu = \mu_X - \mu_Y$, e poi eseguire il test unilaterale per le ipotesi

$$\mathcal{H}_0 \,:\, \mu \le 0 \qquad \mathcal{H}_1 \,:\, \mu > 0$$

A questo punto, siccome attese e varianze non sono note, si calcolano la media e la varianza corretta delle Z

$$\overline{Z} \simeq 0.473 \qquad S_Z^2 \simeq 2.455$$

e da queste il valore della statistica di Student (9.30)

$$T_0 = \sqrt{n}\,\frac{\overline{Z}}{S_Z} \simeq 1.654$$

Per determinare poi l'evento critico unilaterale (9.31) di livello $\alpha = 0.05$ si cerca il quantile di Student

$$t_{1-\alpha}(n-1) = t_{0.950}(29) \simeq 1.699$$

e siccome

$$T_0 \simeq 1.65 < 1.70 \simeq t_{1-\alpha}(n-1)$$

l'evento critico non si verifica, i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.12. *Siano X e Y la quantità di nicotina depositata rispettivamente da sigarette senza filtro e con filtro: da un campione di $n = 11$ misure di X si ha $\overline{X} = 1.36$ e $S_X^2 = 0.22$, mentre da un campione di $m = 9$ misure di Y si ha $\overline{Y} = 0.70$ e $S_Y^2 = 0.03$. Stabilire con un test unilaterale di livello $\alpha = 0.01$ se la media μ_X è significativamente più grande della media μ_Y decidendo quale accettare fra le ipotesi*

$$\mathcal{H}_0 \,:\, \mu_X \leq \mu_Y \qquad \mathcal{H}_1 \,:\, \mu_X > \mu_Y$$

Soluzione: Il test riguarda un confronto fra medie di campioni indipendenti con varianze sconosciute (i valori delle varianze dati nella traccia sono infatti ricavati dai campioni anche se questi non sono esplicitamente riportati): si veda a questo proposito la Sezione 9.3.2. L'evento critico unilaterale è pertanto quello in (9.36), e il quantile di Student che lo definisce per $\alpha = 0.01$ è

$$t_{1-\alpha}(n+m-2) = t_{0.99}(18) \simeq 2.552$$

Per calcolare la corrispondente statistica di Student (9.35) dovremo però preventivamente calcolare la varianza combinata (9.34) ottenendo

$$V^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2} \simeq 0.136 \qquad V \simeq 0.368$$

Per la statistica di Student avremo allora

$$T_0 = \frac{\overline{X} - \overline{Y}}{V\sqrt{\frac{1}{n} + \frac{1}{m}}} \simeq 3.99$$

e siccome

$$T_0 \simeq 3.99 > 2.552 \simeq t_{1-\alpha}(n+m-2)$$

ne concludiamo che l'evento critico unilaterale (9.36) si verifica, i dati sono in regione critica e quindi rifiuteremo $\mathcal{H}_0$ e accetteremo $\mathcal{H}_1$ □

Esercizio A.3.13. *Per controllare l'efficacia di un nuovo medicinale si paragonano i risultati X ed Y di un certo tipo di analisi clinica eseguita su due campioni indipendenti rispettivamente di $n = 10$ ed $m = 12$ pazienti: ai 10 pazienti del primo gruppo è stato somministrato il nuovo farmaco; ai 12 del secondo gruppo è stato somministrato solo un placebo. Se il farmaco è efficace la media di X deve essere più grande della media di Y. I risultati delle analisi sono i seguenti:*

$$X = \begin{array}{ccccc} 9.51 & 8.39 & 8.62 & 9.48 & 8.85 \\ 9.29 & 8.43 & 9.57 & 9.30 & 9.21 \end{array}$$

$$Y = \begin{array}{cccccc} 8.10 & 8.58 & 9.05 & 7.28 & 7.64 & 5.83 \\ 8.61 & 7.10 & 6.44 & 7.43 & 8.63 & 7.94 \end{array}$$

Decidere con un test unilaterale di livello $\alpha = 0.05$ quale accettare fra le ipotesi

$$\mathcal{H}_0 \,:\, \mu_X \leq \mu_Y \qquad \mathcal{H}_1 \,:\, \mu_X > \mu_Y$$

Soluzione: Anche in questo caso si tratta di un confronto di medie di campioni indipendenti tramite un test unilaterale con varianze sconosciute. Dopo aver calcolato dai campioni

$$\overline{X} \simeq 9.065 \qquad S_X^2 \simeq 0.206 \qquad \overline{Y} \simeq 7.719 \qquad S_Y^2 \simeq 0.927$$
$$V^2 \simeq 0.603 \qquad V \simeq 0.776$$

e dopo aver trovato il quantile di Student con $\alpha = 0.05$

$$t_{1-\alpha}(n+m-2) = t_{0.95}(20) \simeq 1.725$$

si calcola la statistica di Student

$$T_0 = \frac{\overline{X} - \overline{Y}}{V\sqrt{\frac{1}{n} + \frac{1}{m}}} \simeq 4.050$$

e siccome

$$T_0 \simeq 4.050 > 1.725 \simeq t_{1-\alpha}(n+m-2)$$

ne concludiamo che l'evento critico unilaterale (9.36) si verifica, i dati sono in regione critica e quindi rifiuteremo $\mathcal{H}_0$ e accetteremo $\mathcal{H}_1$ □

Esercizio A.3.14. *La differenza fra entrate e uscite (in milioni di Euro) di una ditta è una v-a con varianza $\sigma^2 = 1$. Vengono introdotte alcune modifiche nel sistema di vendita dei prodotti: per controllare se la situazione finanziaria è migliorata si confrontano i bilanci X di $n = 10$ mesi successivi all'introduzione di tali modifiche con quelli Y di $m = 12$ mesi precedenti, e si ottengono i seguenti risultati*

$$X = \begin{matrix} 0.61 & 0.90 & 2.76 & 1.31 & 3.33 \\ 2.08 & 1.42 & -0.67 & 2.22 & 3.28 \end{matrix}$$

$$Y = \begin{matrix} 0.80 & 2.28 & 1.11 & -1.26 & 0.70 & 1.26 \\ 0.42 & 2.24 & 1.58 & -0.21 & -0.26 & -2.02 \end{matrix}$$

Supponendo che le modifiche abbiano lasciato immutata la varianza, decidere con un test di livello $\alpha = 0.05$ quale accettare fra le ipotesi

$$\mathcal{H}_0 : \mu_X \le \mu_Y \qquad \mathcal{H}_1 : \mu_X > \mu_Y$$

Soluzione: Questa volta il confronto fra le medie di campioni indipendenti è effettuato supponendo che le varianze $\sigma_X^2 = \sigma_Y^2 = 1$ siano conosciute, per cui l'evento critico unilaterale assume la forma (9.33). Pertanto, dopo aver calcolato dai campioni

$$\overline{X} \simeq 1.724 \qquad \overline{Y} \simeq 0.553$$

si cerca sulle Tavole E.1 con $\alpha = 0.05$ il quantile della normale standard

$$\varphi_{1-\alpha} = \varphi_{0.950} \simeq 1.65$$

e si calcola il valore della statistica di Gauss (9.32)

$$U_0 = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \simeq 2.74$$

Siccome

$$U_0 \simeq 2.74 > 1.65 \simeq \varphi_{1-\alpha}$$

ne concludiamo che l'evento critico unilaterale (9.33) si verifica, i dati sono in regione critica e quindi rifiuteremo $\mathcal{H}_0$ e accetteremo $\mathcal{H}_1$

A scopo illustrativo possiamo confermare questo risultato abbandonando l'ipotesi che le varianze siano conosciute e calcolandone invece una stima dai campioni. In questo caso avremmo

$$S_X^2 \simeq 1.596 \qquad S_Y^2 \simeq 1.722 \qquad V^2 \simeq 1.665 \qquad V \simeq 1.290$$

mentre con $\alpha = 0.05$ il quantile di Student sarebbe

$$t_{1-\alpha}(n+m-2) = t_{0.950}(20) \simeq 1.725$$

Siccome ora la statistica di Student (9.35) è

$$T_0 = \frac{\overline{X} - \overline{Y}}{V\sqrt{\frac{1}{n} + \frac{1}{m}}} \simeq 2.119$$

avremo

$$T_0 \simeq 2.119 > 1.725 \simeq t_{1-\alpha}(n+m-2)$$

cioè l'evento critico unilaterale (9.36) si verifica, i dati sono in regione critica e quindi ancora una volta rifiuteremo $\mathcal{H}_0$ e accetteremo $\mathcal{H}_1$ □

Esercizio A.3.15. *Per controllare se una nuova procedura di fabbricazione ha modificato la qualità dei prodotti di una azienda si paragonano le misure X ed Y di una data caratteristica dei prodotti prima e dopo l'introduzione della nuova procedura. Si ottengono così due campioni indipendenti rispettivamente di $n = 10$ ed $m = 12$ valori:*

$$X = \begin{matrix} 9.41 & 9.71 & 10.32 & 9.05 & 8.63 \\ 9.12 & 8.65 & 8.91 & 10.36 & 8.80 \end{matrix}$$

$$Y = \begin{matrix} 8.10 & 7.58 & 8.06 & 8.43 & 8.63 & 8.69 \\ 7.61 & 8.39 & 10.57 & 9.11 & 8.60 & 8.62 \end{matrix}$$

decidere con un test bilaterale di livello $\alpha = 0.05$ quale accettare fra le ipotesi

$$\mathcal{H}_0 \,:\, \mu_X = \mu_Y \qquad \mathcal{H}_1 \,:\, \mu_X \neq \mu_Y$$

Soluzione: Si tratta di un test di confronto di medie di campioni indipendenti con varianze sconosciute per cui l'evento critico bilaterale è (9.36). Dopo aver calcolato

$$\overline{X} \simeq 9.296 \qquad \overline{Y} \simeq 8.533 \qquad S_X^2 \simeq 0.412 \qquad S_Y^2 \simeq 0.612 \qquad V \simeq 0.522$$

e dopo aver trovato con $\alpha = 0.05$ il quantile di Student

$$t_{1-\frac{\alpha}{2}}(n+m-2) = t_{0.975}(20) \simeq 2.086$$

si determina la statistica di Student

$$|T_0| = \left| \frac{\overline{X} - \overline{Y}}{V\sqrt{\frac{1}{n} + \frac{1}{m}}} \right| \simeq 3.415$$

Siccome

$$|T_0| \simeq 3.415 > 2.086 \simeq t_{1-\frac{\alpha}{2}}(n+m-2)$$

l'evento critico bilaterale (9.36) si verifica, i dati sono in regione critica e quindi rifiuteremo $\mathcal{H}_0$ e accetteremo $\mathcal{H}_1$ □

Esercizio A.3.16. *Siano dati due campioni indipendenti $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$, composti rispettivamente di $n = 21$ e $m = 31$ valori, e supponiamo che le varianze empiriche calcolate a partire dai dati siano rispettivamente*

$$S_X^2 = 2 \qquad S_Y^2 = 1$$

Determinare prima gli intervalli di fiducia di livello $\alpha = 0.05$ delle varianze σ_X^2 e σ_Y^2, e successivamente eseguire un test bilaterale di Fisher di livello $\alpha = 0.05$ per decidere fra le due ipotesi

$$\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \,:\, \sigma_X^2 \neq \sigma_Y^2$$

Soluzione: Per determinare gli intervalli di fiducia di livello $\alpha = 0.05$ delle varianze dobbiamo prima trovare sulle Tavole E.3 i quantili occorrenti

$$\chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.975}(20) \simeq 34.170 \qquad \chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.025}(20) \simeq 9.591$$
$$\chi^2_{1-\frac{\alpha}{2}}(m-1) = \chi^2_{0.975}(30) \simeq 46.979 \qquad \chi^2_{\frac{\alpha}{2}}(m-1) = \chi^2_{0.025}(30) \simeq 16.791$$

Gli estremi degli intervalli di fiducia (8.7) per le varianze σ_X^2 e σ_Y^2 sono allora rispettivamente

$$\frac{(n-1)S_X^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \simeq 1.171 \qquad \frac{(n-1)S_X^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \simeq 4.171$$
$$\frac{(m-1)S_Y^2}{\chi^2_{1-\frac{\alpha}{2}}(m-1)} \simeq 0.639 \qquad \frac{(m-1)S_Y^2}{\chi^2_{\frac{\alpha}{2}}(m-1)} \simeq 1.787$$

e quindi gli intervalli per σ_X^2 e σ_Y^2 sono rispettivamente

$$[1.171\,,\ 4.171] \qquad\qquad [0.639\,,\ 1.787]$$

Per il test bilaterale sulle varianze (vedi Sezione 9.4) useremo poi la statistica di Fisher (9.39) nella forma

$$F_0 = \frac{S_X^2}{S_Y^2}$$

Pertanto, per determinare l'evento critico bilaterale (9.40) del test di livello $\alpha = 0.05$ dovremo innanzitutto individuare sulle Tavole E.4 gli opportuni quantili

$$f_{\frac{\alpha}{2}}(n-1, m-1) = f_{0.025}(20, 30) = \frac{1}{f_{0.975}(30, 20)} \simeq \frac{1}{2.35} \simeq 0.43$$
$$f_{1-\frac{\alpha}{2}}(n-1, m-1) = f_{0.975}(20, 30) \simeq 2.20$$

Siccome per i nostri dati $F_0 = 2$, avremo che

$$f_{\frac{\alpha}{2}}(n-1, m-1) \simeq 0.43 \ < \ F_0 = 2 \ < \ 2.20 \simeq f_{1-\frac{\alpha}{2}}(n-1, m-1)$$

cioè l'evento critico bilaterale (9.40) non si verifica, i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.17. *Due serie indipendenti di $n = 9$ ed $m = 10$ misure rispettivamente di due quantità aleatorie X ed Y danno i seguenti risultati*

X	=	4.83	−2.52	−1.79	−2.85	1.45	1.09	1.87	2.03	−2.60	
Y	=	−1.34	1.32	−0.96	0.29	−1.41	0.23	−0.56	−0.32	0.66	1.27

Determinare gli intervalli di fiducia di livello $\alpha = 0.05$ per le due varianze sconosciute σ_X^2 e σ_Y^2, e successivamente eseguire un test unilaterale di livello $\alpha = 0.05$ per decidere fra le ipotesi

$$\mathcal{H}_0 \ : \ \sigma_X^2 \leq \sigma_Y^2 \qquad\qquad \mathcal{H}_1 \ : \ \sigma_X^2 > \sigma_Y^2$$

Soluzione: Bisogna prima calcolare medie e varianze dei campioni

$$\overline{X} \simeq 0.179 \qquad S_X^2 \simeq 7.346 \qquad \overline{Y} \simeq -0.082 \qquad S_Y^2 \simeq 0.998$$

e poi per determinare gli intervalli di fiducia di livello $\alpha = 0.05$ delle varianze dobbiamo trovare sulle Tavole E.3 i quantili occorrenti

$$\chi^2_{1-\frac{\alpha}{2}}(n-1) = \chi^2_{0.975}(8) \simeq 17.535 \qquad \chi^2_{\frac{\alpha}{2}}(n-1) = \chi^2_{0.025}(8) \simeq 2.180$$
$$\chi^2_{1-\frac{\alpha}{2}}(m-1) = \chi^2_{0.975}(9) \simeq 19.023 \qquad \chi^2_{\frac{\alpha}{2}}(m-1) = \chi^2_{0.025}(9) \simeq 2.700$$

Con questi valori gli estremi degli intervalli di fiducia (8.7) per le varianze σ_X^2 e σ_Y^2 sono allora rispettivamente

$$\frac{(n-1)S_X^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)} \simeq 3.352 \qquad \frac{(n-1)S_X^2}{\chi^2_{\frac{\alpha}{2}}(n-1)} \simeq 26.960$$
$$\frac{(m-1)S_Y^2}{\chi^2_{1-\frac{\alpha}{2}}(m-1)} \simeq 0.472 \qquad \frac{(m-1)S_Y^2}{\chi^2_{\frac{\alpha}{2}}(m-1)} \simeq 3.326$$

e quindi gli intervalli per σ_X^2 e σ_Y^2 sono rispettivamente

$$[3.352\,,\ 26.960] \qquad [0.472\,,\ 3.326]$$

Per il test unilaterale sulle varianze useremo la statistica di Fisher (9.39), e per determinare l'evento critico unilaterale (9.40) del test di livello $\alpha = 0.05$ cercheremo sulle Tavole E.4 il quantile

$$f_{1-\alpha}(n-1, m-1) = f_{0.950}(8, 9) \simeq 3.23$$

A questo punto, siccome

$$F_0 = \frac{S_X^2}{S_Y^2} \simeq 7.36$$

avremo

$$F_0 \simeq 7.36 \ > \ 3.23 \simeq f_{1-\alpha}(n-1, m-1)$$

cioè l'evento critico unilaterale (9.40) si verifica, i dati sono in regione critica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ e accetteremo $\mathcal{H}_1$ □

Esercizio A.3.18. *Siano dati i seguenti due campioni indipendenti:*

$$X = \begin{array}{ccccc} 3.16 & 4.37 & 6.67 & 2.49 & 2.06 \\ 4.10 & -0.88 & 3.84 & 1.45 & \end{array} \qquad Y = \begin{array}{ccccc} 3.51 & 6.94 & 5.46 & 5.27 & 7.48 \\ 5.76 & 2.21 & 7.29 & 7.80 & 6.06 \end{array}$$

Con un test bilaterale di livello $\alpha = 0.05$ *si decida quale accettare fra le ipotesi*

$$\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \,:\, \sigma_X^2 \neq \sigma_Y^2$$

Poi con un test bilaterale di livello $\alpha = 0.05$ *si decida quale accettare fra le ipotesi*

$$\mathcal{H}_0 \,:\, \mu_X = \mu_Y \qquad \mathcal{H}_1 \,:\, \mu_X \neq \mu_Y$$

Soluzione: Conviene innanzitutto calcolare dai campioni le quantità necessarie

$$\overline{X} \simeq 3.029 \qquad S_X^2 \simeq 4.485 \qquad \overline{Y} \simeq 5.778 \qquad S_Y^2 \simeq 3.215$$

poi per il primo test bilaterale sulle varianze di livello $\alpha = 0.05$ si determinano i quantili di Fisher

$$f_{\frac{\alpha}{2}}(n-1, m-1) = f_{0.025}(8, 9) = \frac{1}{f_{0.975}(9, 8)} \simeq \frac{1}{4.36} \simeq 0.23$$
$$f_{1-\frac{\alpha}{2}}(n-1, m-1) = f_{0.975}(8, 9) \simeq 4.10$$

e siccome la statistica di Fisher è

$$F_0 = \frac{S_X^2}{S_Y^2} \simeq 1.395$$

si ottiene

$$f_{\frac{\alpha}{2}}(n-1, m-1) \simeq 0.23 \ < \ F_0 = 1.395 \ < \ 4.10 \simeq f_{1-\frac{\alpha}{2}}(n-1, m-1)$$

cioè l'evento critico bilaterale (9.40) non si verifica, i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0 : \sigma_X^2 = \sigma_Y^2$. Per il test bilaterale sulle medie di livello $\alpha = 0.05$ bisogna invece prima determinare il quantile di Student

$$t_{1-\frac{\alpha}{2}}(n+m-2) = t_{0.975}(17) \simeq 2.11$$

poi calcolare la varianza combinata (9.34) ottenendo

$$V^2 \simeq 3.813 \qquad V \simeq 1.95$$

e poi la statistica di Student (9.35)

$$|T_0| \simeq 3.06$$

A questo punto si osserva che

$$|T_0| \simeq 3.06 > 2.11 \simeq t_{1-\frac{\alpha}{2}}(n+m-2)$$

per cui l'evento critico bilaterale (9.36) si verifica, i dati sono in regione critica e quindi rifiuteremo $\mathcal{H}_0 : \mu_X = \mu_Y$ e accetteremo $\mathcal{H}_1 : \mu_X \neq \mu_Y$ □

Esercizio A.3.19. *Siano X e Y due quantità aleatorie: due campioni indipendenti rispettivamente di $n = 21$ e $m = 16$ misure hanno medie $\overline{X} = 5$ e $\overline{Y} = 3$, e varianze corrette (stimate a partire dai campioni) $S_X^2 = 5$ e $S_Y^2 = 3$. Decidere prima con un test bilaterale di livello $\alpha = 0.05$ quale accettare fra le due ipotesi*

$$\mathcal{H}_0 \ : \ \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \ : \ \sigma_X^2 \neq \sigma_Y^2$$

e poi con un test unilaterale di livello $\alpha = 0.05$ quale accettare fra le due ipotesi

$$\mathcal{H}_0 \ : \ \mu_X \leq \mu_Y \qquad \mathcal{H}_1 \ : \ \mu_X > \mu_Y$$

Infine determinate gli intervalli di fiducia di livello $\alpha = 0.05$ per le attese μ_X e μ_Y

Risposta: Nel primo test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(20,15) \simeq 0.39 \ < \ F_0 = \frac{S_X^2}{S_Y^2} \simeq 1.67 \ < \ 2.76 \simeq f_{0.975}(20,15)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2$. Nel secondo test unilaterale di Student sulle attese si ha

$$V^2 \simeq 4.143$$

e quindi l'evento critico si verifica

$$T_0 = \frac{\overline{X} - \overline{Y}}{V\sqrt{{}^1\!/_n + {}^1\!/_m}} \simeq 2.96 \ > \ 1.69 \simeq t_{0.950}(35)$$

per cui si rifiuta l'ipotesi $\mathcal{H}_0 \,:\, \mu_X \le \mu_Y$. Infine

$$5 \pm 1.02 \qquad\qquad 3 \pm 0.92$$

sono gli intervalli di fiducia di μ_X e μ_Y □

Esercizio A.3.20. *Siano dati i due seguenti campioni indipendenti delle quantità aleatorie X e Y:*

$$\begin{array}{rcl} X & = & \begin{array}{cccccccccc} 1.13 & 0.11 & 0.74 & 0.97 & 0.10 & 1.14 & 0.68 & 0.21 & 0.70 & 0.41 \\ 0.82 & 0.04 & 0.85 & 0.85 & 0.84 & 0.11 & 0.55 & 0.09 & 1.01 & 0.34 \\ 1.04 & & & & & & & & & \end{array} \\ Y & = & \begin{array}{cccccccccc} 0.85 & 0.62 & 0.21 & 0.22 & 0.34 & 0.13 & 0.93 & 0.38 & 0.01 & 0.65 \\ 0.97 & 0.79 & 0.42 & 0.09 & 0.44 & 0.31 & & & & \end{array} \end{array}$$

Determinate gli intervalli di fiducia di livello $\alpha = 0.05$ per le varianze σ_X^2 e σ_Y^2, e poi decidere con un test bilaterale di livello $\alpha = 0.05$ quale accettare fra le due ipotesi

$$\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2 \qquad\qquad \mathcal{H}_1 \,:\, \sigma_X^2 \neq \sigma_Y^2$$

Risposta: Innanzitutto risulta

$$\overline{X} \simeq 0.606 \qquad \overline{Y} \simeq 0.460 \qquad S_X^2 \simeq 0.145 \qquad S_Y^2 \simeq 0.094$$

per cui gli intervalli di fiducia di σ_X^2 e σ_Y^2 sono

$$[0.08\,,\,0.30] \qquad\qquad [0.05\,,\,0.23]$$

Nel test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(20,15) \simeq 0.39 \ < \ F_0 = \frac{S_X^2}{S_Y^2} \simeq 1.53 \ < \ 2.76 \simeq f_{0.975}(20,15)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2$ □

Esercizio A.3.21. *Due campioni indipendenti delle quantità X e Y con attese μ_X, μ_Y e varianze σ_X^2, σ_Y^2 sconosciute, sono composti rispettivamente di $n = 21$ e $m = 31$ misure, e hanno medie $\overline{X} = 1.09$ e $\overline{Y} = 2.22$ e varianze corrette (calcolate dai campioni) $S_X^2 = 2.34$ e $S_Y^2 = 3.11$. Decidere prima con un test bilaterale di livello $\alpha = 0.05$ quale accettare fra le due ipotesi*

$$\mathcal{H}_0 \ : \ \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \ : \ \sigma_X^2 \neq \sigma_Y^2$$

e poi decidere con un test bilaterale di livello $\alpha = 0.01$ quale accettare fra le le due ipotesi

$$\mathcal{H}_0 \ : \ \mu_X = \mu_Y \qquad \mathcal{H}_1 \ : \ \mu_X \neq \mu_Y$$

Risposta: Nel primo test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(20, 30) \simeq 0.43 \ < \ F_0 = \frac{S_X^2}{S_Y^2} \simeq 0.75 \ < \ 2.20 \simeq f_{0.975}(20, 30)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \ : \ \sigma_X^2 = \sigma_Y^2$. Nel secondo test unilaterale di Student sulle attese si ha

$$V^2 \simeq 2.802$$

e quindi l'evento critico non si verifica

$$|T_0| = \frac{|\overline{X} - \overline{Y}|}{V\sqrt{{}^1\!/_n + {}^1\!/_m}} \simeq 2.39 \ < \ 2.68 \simeq t_{0.995}(50)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \ : \ \mu_X = \mu_Y$ □

Esercizio A.3.22. *Siano dati i due seguenti campioni indipendenti delle quantità aleatorie X e Y:*

$$\begin{array}{rcccccccccccc} X & = & 5.49 & 4.37 & 4.42 & 3.79 & 5.57 & 3.37 & 4.30 & 3.14 & 4.55 & 4.23 & \\ & & 3.44 & 2.51 & 4.46 & 2.54 & 3.59 & 4.20 & 3.14 & 3.50 & 2.26 & 4.01 & 4.06 \\ Y & = & 5.84 & 2.29 & 3.71 & 5.09 & 7.07 & 5.61 & 4.83 & 3.29 & 5.55 & 6.78 & \\ & & 5.74 & 5.58 & 5.28 & 4.89 & 4.21 & 2.54 & 3.71 & 3.81 & 4.60 & 6.81 & \\ & & 4.60 & 5.10 & 4.72 & 6.33 & 5.76 & 4.86 & 3.50 & 5.18 & 6.60 & 5.47 & 3.49 \end{array}$$

Determinate prima gli intervalli di fiducia di livello $\alpha = 0.05$ per σ_X^2 e σ_Y^2, e poi decidere con un test bilaterale di livello $\alpha = 0.05$ quale accettare fra le due ipotesi

$$\mathcal{H}_0 \ : \ \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \ : \ \sigma_X^2 \neq \sigma_Y^2$$

Risposta: Innanzitutto risulta

$$\overline{X} \simeq 3.854 \qquad \overline{Y} \simeq 4.930 \qquad S_X^2 \simeq 0.764 \qquad S_Y^2 \simeq 1.493$$

per cui gli intervalli di fiducia di σ_X^2 e σ_Y^2 sono

$$[0.45\,,\,1.59] \qquad\qquad [0.95\,,\,2.67]$$

Nel test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(20,30) \simeq 0.43 \;<\; F_0 = \frac{S_X^2}{S_Y^2} \simeq 0.51 \;<\; 2.20 \simeq f_{0.975}(20,30)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2$ □

Esercizio A.3.23. *Due serie indipendenti di $n = 10$ ed $m = 9$ misure rispettivamente di due quantità aleatorie X ed Y danno i seguenti risultati*

$$\begin{array}{rrrrrrrrrrr} X = & -1.34 & 1.32 & -0.96 & 0.29 & -1.41 & 0.23 & -0.56 & -0.32 & 0.66 & 1.27 \\ Y = & 4.83 & -2.52 & -1.79 & -2.85 & 1.45 & 1.09 & 1.87 & 2.03 & -2.60 & \end{array}$$

Eseguire un test di bilaterale di livello $\alpha = 0.05$ per decidere quale accettare fra le due ipotesi

$$\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2 \qquad\qquad \mathcal{H}_1 \,:\, \sigma_X^2 \neq \sigma_Y^2$$

Risposta: Innanzitutto risulta

$$\overline{X} \simeq -0.082 \qquad \overline{Y} \simeq 0.168 \qquad S_X^2 \simeq 0.998 \qquad S_Y^2 \simeq 7.303$$

Nel test bilaterale di Fisher sulla varianza, poi, l'evento critico si verifica

$$F_0 = \frac{S_X^2}{S_Y^2} \simeq 0.14 \;<\; 0.24 \simeq f_{0.025}(9,8) \;<\; 4.36 \simeq f_{0.975}(9,8)$$

per cui si rifiuta l'ipotesi $\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2$ □

Esercizio A.3.24. *Per giudicare il rendimento di due classi di $n = 31$ e $m = 21$ studenti si paragonano i voti (espressi in trentesimi) con i quali è stato superato un certo esame. I due campioni indipendenti sono i seguenti:*

$$\begin{array}{rcl} X & = & \begin{array}{cccccccccccccccc} 25 & 27 & 27 & 26 & 27 & 25 & 23 & 25 & 22 & 26 & 25 & 21 & 24 & 22 & 24 & 24 \\ 23 & 22 & 29 & 26 & 21 & 26 & 23 & 24 & 24 & 20 & 29 & 22 & 23 & 18 & 27 & \end{array} \\ Y & = & \begin{array}{ccccccccccc} 21 & 19 & 19 & 22 & 25 & 20 & 25 & 23 & 26 & 27 & 22 \\ 25 & 27 & 20 & 19 & 23 & 28 & 22 & 25 & 23 & 28 & \end{array} \end{array}$$

Decidere prima con un test bilaterale di livello $\alpha = 0.05$ quale accettare tra le ipotesi

$$\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2 \qquad\qquad \mathcal{H}_1 \,:\, \sigma_X^2 \neq \sigma_Y^2$$

e poi decidere con un test unilaterale di livello $\alpha = 0.01$ quale accettare tra le ipotesi

$$\mathcal{H}_0 \,:\, \mu_X \leq \mu_Y \qquad\qquad \mathcal{H}_1 \,:\, \mu_X > \mu_Y$$

Risposta: Innanzitutto risulta

$$\overline{X} \simeq 24.193 \qquad \overline{Y} \simeq 23.286 \qquad S_X^2 \simeq 6.495 \qquad S_Y^2 \simeq 8.914$$

Nel primo test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(30,20) \simeq 0.46 \ < \ F_0 = \frac{S_X^2}{S_Y^2} \simeq 0.73 \ < \ 2.35 \simeq f_{0.975}(30,20)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \, : \, \sigma_X^2 = \sigma_Y^2$. Nel secondo test unilaterale di Student sulle attese si ha

$$V^2 \simeq 7.462$$

e quindi l'evento critico non si verifica

$$T_0 = \frac{\overline{X} - \overline{Y}}{V\sqrt{{}^1\!/_n + {}^1\!/_m}} \simeq 1.18 \ < \ 2.40 \simeq t_{0.990}(50)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \, : \, \mu_X \leq \mu_Y$ □

Esercizio A.3.25. *Siano dati i seguenti due campioni aleatori di* $n = 16$ *ed* $m = 21$ *misure:*

$$\begin{array}{rcl} X & = & \begin{array}{cccccccccccc} 9.08 & 9.80 & 10.76 & 9.58 & 10.56 & 9.99 & 10.61 & 8.86 & 10.41 & 10.27 & 7.94 \\ 8.61 & 10.38 & 9.77 & 11.30 & 10.09 \end{array} \\ Y & = & \begin{array}{cccccccccc} 3.99 & 4.04 & 4.20 & 6.11 & 7.26 & 6.43 & 5.23 & 5.93 & 3.71 & 4.53 & 4.91 \\ 5.33 & 3.95 & 4.60 & 3.87 & 5.25 & 4.09 & 5.21 & 5.85 & 4.10 & 4.26 \end{array} \end{array}$$

Calcolare prima gli intervalli di fiducia di livello $\alpha = 0.01$ *per le varianze* σ_X^2 *e* σ_Y^2*, e poi decidere con un test bilaterale di livello* $\alpha = 0.05$ *quale accettare tra le ipotesi*

$$\mathcal{H}_0 \, : \, \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \, : \, \sigma_X^2 \neq \sigma_Y^2$$

Risposta: Innanzitutto risulta

$$\overline{X} \simeq 9.876 \qquad \overline{Y} \simeq 4.898 \qquad S_X^2 \simeq 0.779 \qquad S_Y^2 \simeq 0.961$$

per cui gli intervalli di fiducia di σ_X^2 e σ_Y^2 sono

$$[0.36\,,\,2.54] \qquad [0.48\,,\,2.59]$$

Nel test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(15,20) \simeq 0.36 \ < \ F_0 = \frac{S_X^2}{S_Y^2} \simeq 0.81 \ < \ 2.57 \simeq f_{0.975}(15,20)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \, : \, \sigma_X^2 = \sigma_Y^2$ □

Esercizio A.3.26. *Per confrontare le attività di due materiali radioattivi si rilevano i numeri di particelle emesse in un intervallo di* 10 *minuti. Si misurano pertanto, rispettivamente per i due materiali,* $n = 31$ *e* $m = 31$ *valori di tale conteggio ottenendo i seguenti campioni indipendenti*

$$\begin{array}{rcl} X & = & \begin{array}{cccccccccccccccc} 15 & 20 & 21 & 26 & 20 & 14 & 13 & 17 & 22 & 21 & 24 & 19 & 18 & 22 & 17 & 17 \\ 19 & 23 & 21 & 22 & 31 & 14 & 14 & 24 & 24 & 21 & 20 & 25 & 17 & 20 & 20 & \end{array} \\ Y & = & \begin{array}{cccccccccccccccc} 18 & 18 & 23 & 19 & 14 & 14 & 14 & 16 & 23 & 21 & 17 & 21 & 14 & 22 & 17 & 19 \\ 16 & 9 & 22 & 13 & 17 & 19 & 16 & 19 & 22 & 11 & 16 & 9 & 19 & 21 & 13 & \end{array} \end{array}$$

Decidere prima con un test bilaterale di livello $\alpha = 0.05$ *quale accettare tra le ipotesi*

$$\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1 \,:\, \sigma_X^2 \neq \sigma_Y^2$$

e poi decidere con un test unilaterale di livello $\alpha = 0.01$ *quale accettare tra le ipotesi*

$$\mathcal{H}_0 \,:\, \mu_X \leq \mu_Y \qquad \mathcal{H}_1 \,:\, \mu_X > \mu_Y$$

Risposta: Innanzitutto risulta

$$\overline{X} \simeq 20.032 \qquad \overline{Y} \simeq 17.161 \qquad S_X^2 \simeq 15.966 \qquad S_Y^2 \simeq 15.073$$

Nel primo test bilaterale di Fisher sulla varianza l'evento critico non si verifica

$$f_{0.025}(30,30) \simeq 0.48 \;<\; F_0 = \frac{S_X^2}{S_Y^2} \simeq 1.06 \;<\; 2.07 \simeq f_{0.975}(30,30)$$

per cui si accetta l'ipotesi $\mathcal{H}_0 \,:\, \sigma_X^2 = \sigma_Y^2$. Nel secondo test unilaterale di Student sulle attese si ha

$$V^2 \simeq 15.519$$

e quindi l'evento critico si verifica

$$T_0 = \frac{\overline{X} - \overline{Y}}{V\sqrt{{}^1\!/_n + {}^1\!/_m}} \simeq 2.87 \;>\; 2.39 \simeq t_{0.990}(60)$$

per cui si rifiuta l'ipotesi $\mathcal{H}_0 \,:\, \mu_X \leq \mu_Y$ □

Esercizio A.3.27. *Per studiare l'attività di un centralino telefonico si rileva il numero* X *di telefonate che arrivano tra le 11.00 e le 12.00 in* $n = 100$ *giorni lavorativi tipici. Le frequenze* N_j *con cui si ritrovano i diversi valori* $j = 0, 1, \ldots$ *di* X *sono le seguenti:*

$j =$	0	1	2	3	4	5	6	7	8	≥ 9
$N_j =$	1	4	15	18	22	17	10	8	3	2

Verificare con un test del chi-quadro di livello $\alpha = 0,05$ *se questi dati sono compatibili con l'ipotesi*

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge di Poisson } \mathfrak{P}(4)$$

Soluzione: Si deve innanzitutto verificare se è rispettata la condizione di applicabilità che richiede $np_j \geq 5$ per tutti gli indici j presenti nel campione: in questo esercizio le p_j sono le probabilità della legge di Poisson $\mathfrak{P}(4)$

$$p_j = e^{-4}\,\frac{4^j}{j!} \qquad\qquad j = 0, 1, \ldots$$

Se le condizioni non sono soddisfatte (come in effetti avviene in questo caso) si deve procedere a raggruppare i valori j con le probabilità più piccole ottenendo (come nell'Esempio 9.9) un nuovo elenco di probabilità q_j e di frequenze empiriche M_j sulle quali eseguire il test. Converrà pertanto stilare la seguente tabella

j	N_j	p_j	np_j	j	q_j	nq_j	M_j	$\frac{(M_j - nqj)^2}{nq_j}$
0	1	0.018	1.8					
1	4	0.073	7.3	0, 1	0.091	9.1	5	1.888
2	15	0.147	14.7	2	0.147	14.7	15	0.008
3	18	0.195	19.5	3	0.195	19.5	18	0.121
4	22	0.195	19.5	4	0.195	19.5	22	0.311
5	17	0.156	15.6	5	0.156	15.6	17	0.120
6	10	0.104	10.4	6	0.104	10.4	10	0.017
7	8	0.060	6.0	7	0.060	6.0	8	0.703
8	3	0.030	3.0	≥ 8	0.051	5.1	5	0.003
≥ 9	2	0.021	2.1					
							$K_0 =$	**3.170**

Nelle prime quattro colonne sono riportati i dati del problema con i valori dei prodotti np_j (con $n = 100$) dai quali si evince la necessità di raggruppare le classi estreme per soddisfare le condizioni di applicabilità del test. Nelle successive quattro colonne è mostrato l'effetto del raggruppamento, e nell'ultima colonna dai dati modificati sono calcolati gli addendi della statistica di Pearson (9.41) K_0 il cui valore (la somma di tutti i termini dell'ultima colonna) è infine mostrato in grassetto in fondo a destra. L'evento critico (9.42) del test di livello $\alpha = 0.05$ è poi definito da un quantile del *chi-quadro* con $m - 1 = 7$ gradi di libertà perché dopo il raggruppamento le classi sono diventate $m = 8$. Pertanto dalle Tavole E.3 avremo

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(7) \simeq 14.07$$

e siccome

$$K_0 \simeq 3.17 < 14.07 \simeq \chi^2_{1-\alpha}(m-1)$$

l'evento critico non si verifica, i dati non sono inn regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.28. *Si misura $n = 500$ volte una v-a X che assume i cinque valori $j = 0, 1, 2, 3, 4$, e si ottengono le seguenti frequenze dei risultati:*

$$\begin{array}{rccccc} j = & 0 & 1 & 2 & 3 & 4 \\ N_j = & 4 & 51 & 163 & 173 & 109 \end{array}$$

Verificare con un test del chi-quadro di livello $\alpha = 0.05$ se questi dati sono compatibili con l'ipotesi

$$\mathcal{H}_0 : \textit{il campione si adatta alla legge binomiale } \mathfrak{B}\left(4;\, {}^2\!/_3\right)$$

Soluzione: In questo caso le p_j sono le probabilità binomiali $\mathfrak{B}\left(4;\, {}^2\!/_3\right)$

$$p_j = \binom{4}{j} \left({}^2\!/_3\right)^j \left({}^1\!/_3\right)^{4-j} \qquad\qquad j = 0, 1, \ldots, 4$$

e come si può vedere dalla tabella seguente, con $n = 500$ non c'è bisogno di raggruppare le classi

j	N_j	p_j	np_j	$\frac{(N_j - npj)^2}{np_j}$
0	4	0.012	6.2	0.765
1	51	0.099	49.4	0.053
2	163	0.296	148.1	1.489
3	173	0.395	197.5	3.046
4	109	0.198	98.8	1.061
			$K_0 =$	**6.414**

Siccome il numero di gradi di libertà è $m = 5$, il quantile del *chi-quadro* che definisce l'evento critico (9.42) di livello $\alpha = 0.05$ è

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(4) \simeq 9.49$$

Avremo pertanto

$$K_0 \simeq 6.41 < 9.49 \simeq \chi^2_{1-\alpha}(m-1)$$

per cui l'evento critico non si verifica, i dati non sono inn regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.29. *Si ripete per $n = 200$ volte il lancio di 5 monetine, si conta in ogni ripetizione il numero di teste (con valori $j = 0, 1, 2, 3, 4, 5$), e si ottengono le seguenti frequenze*

$$\begin{array}{rcccccc} j = & 0 & 1 & 2 & 3 & 4 & 5 \\ N_j = & 8 & 37 & 62 & 56 & 34 & 3 \end{array}$$

Verificare con un test del chi-quadro di livello $\alpha = 0.05$ se questi dati sono compatibili con l'ipotesi

$$\mathcal{H}_0 : \textit{il campione si adatta alla legge binomiale } \mathfrak{B}\left(5;\, {}^1\!/_2\right)$$

Soluzione: Le p_j sono le probabilità binomiali $\mathfrak{B}\,(5;\,{}^1\!/_2)$

$$p_j = \binom{5}{j} \left({}^1\!/_2\right)^5 \qquad\qquad j = 0, 1, \ldots, 5$$

e come si può vedere dalla tabella seguente, con $n = 200$ non c'è bisogno di raggruppare le classi

j	N_j	p_j	np_j	$\frac{(N_j - npj)^2}{np_j}$
0	8	0.031	6.3	0.490
1	37	0.156	31.3	1.058
2	62	0.313	62.5	0.004
3	56	0.313	62.5	0.676
4	34	0.156	31.3	0.242
5	3	0.031	6.3	1.690
			$K_0 =$	**4.160**

Siccome il numero di gradi di libertà è $m = 6$, il quantile del *chi-quadro* che definisce l'evento critico (9.42) di livello $\alpha = 0.05$ è

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(5) \simeq 11.07$$

Avremo pertanto

$$K_0 \simeq 4.16 < 11.07 \simeq \chi^2_{1-\alpha}(m-1)$$

per cui l'evento critico non si verifica, i dati non sono inn regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.30. *Con $n = 200$ misure di una v-a discreta X che prende valori $j = 0, \ldots, 4$ si ottengono le seguenti frequenze*

$j =$	0	1	2	3	4
$N_j =$	30	72	70	24	4

Verificare con un test del chi-quadro di livello $\alpha = 0.05$ se questi dati sono compatibili con l'ipotesi

$$\mathcal{H}_0 : \textit{ il campione si adatta alla legge binomiale } \mathfrak{B}\,(4;\,0.4)$$

Soluzione: Le p_j sono le probabilità binomiali $\mathfrak{B}\,(4;\,0.4)$

$$p_j = \binom{4}{j} (0.4)^j\,(0.6)^{4-j} \qquad\qquad j = 0, 1, \ldots, 4$$

e come si può vedere dalla tabella seguente, con $n = 200$ non c'è bisogno di raggruppare le classi

j	N_j	p_j	np_j	$\frac{(N_j-npj)^2}{np_j}$
0	30	0.130	25.9	0.642
1	72	0.346	69.1	0.120
2	70	0.346	69.1	0.011
3	24	0.154	30.7	1.470
4	4	0.026	5.1	0.245
			$K_0 =$	**2.488**

Siccome il numero di gradi di libertà è $m = 5$, il quantile del *chi-quadro* che definisce l'evento critico (9.42) di livello $\alpha = 0.05$ è

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(4) \simeq 9.49$$

Avremo pertanto

$$K_0 \simeq 2.49 < 9.49 \simeq \chi^2_{1-\alpha}(m-1)$$

per cui l'evento critico non si verifica, i dati non sono inn regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.31. *Si lanciano* 5 *dadi* $n = 2\,000$ *volte, e si conta il numero* X *dei* sei *(*X *prende i valori* $j = 0, 1, 2, 3, 4, 5$*) ottenendo le seguenti frequenze*

$$\begin{array}{rcccccc} j = & 0 & 1 & 2 & 3 & 4 & 5 \\ N_j = & 822 & 804 & 300 & 67 & 7 & 0 \end{array}$$

Verificare con un test del chi-quadro di livello $\alpha = 0.05$ *se questi dati sono compatibili con l'ipotesi*

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge binomiale } \mathfrak{B}\left(5;\, {}^1\!/_6\right)$$

Soluzione: Le p_j sono le probabilità binomiali $\mathfrak{B}\,(5;\, {}^1\!/_6)$

$$p_j = \binom{5}{j} \left({}^1\!/_6\right)^j \left({}^5\!/_6\right)^{5-j} \qquad\qquad j = 0, 1, \ldots, 5$$

e come si può vedere dalla tabella seguente, con $n = 2\,000$ c'è bisogno di raggruppare le classi estreme 4 e 5

j	N_j	p_j	np_j	j	q_j	nq_j	M_j	$\frac{(M_j-nqj)^2}{nq_j}$
0	822	0.4019	803.8	0	0.4019	803.8	822	0.4141
1	804	0.4019	803.8	1	0.4019	803.8	804	0.0001
2	300	0.1608	321.5	2	0.1608	321.5	300	1.4381
3	67	0.0322	64.3	3	0.0322	64.3	67	0.1133
4	7	0.0032	6.4	4, 5	0.0033	6.7	7	0.0505
5	0	0.0001	0.3					
							$K_0 =$	**2.0161**

Siccome il numero di gradi di libertà ore è $m = 5$, il quantile del *chi-quadro* che definisce l'evento critico (9.42) di livello $\alpha = 0.05$ è

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(4) \simeq 9.49$$

Avremo pertanto

$$K_0 \simeq 2.02 < 9.49 \simeq \chi^2_{1-\alpha}(m-1)$$

per cui l'evento critico non si verifica, i dati non sono inn regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.32. *Un apparecchio contiene* 4 *componenti elettronici identici. Dopo aver funzionato per un certo tempo ogni componente ha una probabilità p (non nota) di essere ancora funzionante. Un campione di* $n = 50\,000$ *apparecchi viene esaminato, e per ciascuno di essi si conta il numero* ($j = 0, 1, 2, 3, 4$) *di componenti ancora funzionanti ottenendo i seguenti risultati*

$j =$	0	1	2	3	4
$N_j =$	6	201	2 400	14 644	32 749

Calcolare la stima $\overline{p}$ *del parametro p, e poi verificare con un test del chi-quadro di livello* $\alpha = 0.05$ *se questi dati sono compatibili con l'ipotesi*

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge binomiale } \mathfrak{B}\,(4;\,\overline{p})$$

Soluzione: In questo caso bisogna preliminarmente stimare il parametro p con il suo stimatore di MV che in base alla (8.10) del Teorema 8.15, per $\mathfrak{B}\,(4;p)$ con $n = 50\,000$, nel nostro caso vale

$$\overline{p} = \frac{1}{4\,n}\sum_{j=0}^{4} j N_j \simeq 0.90$$

Le p_j sono quindi le probabilità binomiali $\mathfrak{B}\,(4;\,0.90)$

$$p_j = \binom{4}{j}(0.90)^j\,(0.10)^{4-j} \qquad\qquad j = 0, 1, \ldots, 4$$

e sulla base i queste la seguente tabella indica che non è necessario eseguire nessun raggruppamento di classi, e permette quindi di calcolare il valore della statistica di Pearson K_0

j	N_j	p_j	np_j	$\frac{(N_j - npj)^2}{np_j}$
0	6	0.0001	5.1	0.170
1	201	0.0036	181.9	2.016
2	2 400	0.0489	2 445.4	0.841
3	14 644	0.2923	14 614.5	0.060
4	32 749	0.6551	32 753.3	0.001
			$K_0 =$	**3.088**

Ricordando ora che abbiamo stimato $q = 1$ parametri della distribuzione, l'evento critico di livello $\alpha = 0.05$ assumerà la forma (9.43) e quindi con $m = 5$ possibili valori il quantile del *chi-quadro* necessario è

$$\chi^2_{1-\alpha}(m-q-1) = \chi^2_{0.950}(3) \simeq 7.81$$

Avremo pertanto

$$K_0 \simeq 3.09 < 7.81 \simeq \chi^2_{1-\alpha}(m-q-1)$$

per cui l'evento critico non si verifica, i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.33. *$n = 40$ misure di una quantità aleatoria X danno i seguenti risultati:*

0.44	0.37	0.87	1.73	−0.41	2.84	1.40	−0.17	0.29	1.59
0.39	2.39	1.68	−0.05	1.01	1.17	0.62	2.83	0.73	0.91
0.31	−0.92	2.28	0.74	1.02	0.70	2.06	2.56	0.94	2.56
−0.34	1.40	1.42	−0.09	2.17	1.83	1.80	−0.14	1.40	0.91

Usando le frequenze dei ritrovamenti dei valori del campione nei seguenti $m = 4$ intervalli

$$(-\infty, 0] \qquad (0, 1] \qquad (1, 2] \qquad (2, +\infty)$$

stabilire con un test del chi-quadro di livello $\alpha = 0.05$ quali delle seguenti ipotesi sono accettabili

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge normale } \mathfrak{N}(1,1)$$
$$\mathcal{H}'_0 \,:\, \textit{il campione si adatta alla legge uniforme } \mathfrak{U}(-1,3)$$

Soluzione: La tabella delle frequenze dei ritrovamenti nei quattro intervalli è

intervalli	$(-\infty, 0]$	$(0, 1]$	$(1, 2]$	$(2, +\infty)$
N_j	7	13	12	8

mentre le probabilità assegnate agli stessi quattro intervalli dalla legge $\mathfrak{N}(1,1)$ sono ricavati dalle Tavole E.1 ricordando che $\Phi(-1) = 1 - \Phi(1)$

$$p_1 = \boldsymbol{P}\{-\infty \le \mathfrak{N}(1,1) \le 0\} = \boldsymbol{P}\{-\infty \le \mathfrak{N}(0,1) \le -1\} = \Phi(-1) \simeq 0.159$$
$$p_2 = \boldsymbol{P}\{0 \le \mathfrak{N}(1,1) \le 1\} = \boldsymbol{P}\{-1 \le \mathfrak{N}(0,1) \le 0\} = \Phi(0) - \Phi(-1) \simeq 0.341$$
$$p_3 = \boldsymbol{P}\{1 \le \mathfrak{N}(1,1) \le 2\} = \boldsymbol{P}\{0 \le \mathfrak{N}(0,1) \le 1\} = \Phi(1) - \Phi(0) \simeq 0.341$$
$$p_4 = \boldsymbol{P}\{2 \le \mathfrak{N}(1,1) \le +\infty\} = \boldsymbol{P}\{1 \le \mathfrak{N}(0,1) \le +\infty\} = 1 - \Phi(1) \simeq 0.159$$

e le corrispondenti quattro probabilità assegnate dalla legge uniforme $\mathfrak{U}(-1,3)$ sono banalmente tutte uguali

$$p'_1 = p'_2 = p'_3 = p'_4 = {}^1/_4$$

Le nostre tabelle saranno allora

intervalli	j	N_j	p_j	np_j	$\frac{(N_j-npj)^2}{np_j}$	p'_j	np'_j	$\frac{(N_j-np'j)^2}{np'_j}$
$(-\infty, 0]$	1	7	0.159	6.3	0.067	0.250	10	0.900
$(0, 1]$	2	13	0.341	13.7	0.031	0.250	10	0.900
$(1, 2]$	3	12	0.341	13.7	0.200	0.250	10	0.400
$(2, +\infty)$	4	8	0.159	6.3	0.431	0.250	10	0.400
				$K_0 =$	**0.730**		$K'_0 =$	**2.600**

e l'evento critico di livello $\alpha = 0.05$ è lo stesso per ambedue i test con $m = 4$ classi e

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(3) \simeq 7.81$$

Avremo pertanto

$$K_0 \simeq 0.73 < 7.81 \simeq \chi^2_{1-\alpha}(m-1)$$
$$K'_0 \simeq 2.60 < 7.81 \simeq \chi^2_{1-\alpha}(m-1)$$

per cui l'evento critico non si verifica in nessun caso, e quindi accetteremo sia l'ipotesi $\mathcal{H}_0$ che l'ipotesi $\mathcal{H}'_0$: in pratica i dati sono compatibili con ambedue le distribuzioni nel senso che la loro informazione non permette di distinguere i due casi □

Esercizio A.3.34. *$n = 20$ misure di una quantità aleatoria X danno i seguenti risultati:*

−3.601	1.064	3.370	1.535	1.017
1.933	3.100	−0.569	1.141	1.815
2.267	0.195	−0.506	−0.167	−2.936
−0.211	−0.659	−0.375	0.024	2.765

Usando le frequenze dei ritrovamenti dei valori del campione nei seguenti $m = 3$ intervalli

$$(-\infty, 0] \qquad (0, 2] \qquad (2, +\infty)$$

stabilire con un test del chi-quadro di livello $\alpha = 0.05$ se è accettabile la seguente ipotesi

$$\mathcal{H}_0 \ : \ \textit{il campione si adatta alla legge normale } \mathfrak{N}(1, 4)$$

Soluzione: La tabella delle frequenze dei ritrovamenti nei quattro intervalli è

intervalli	$(-\infty, 0]$	$(0, 2]$	$(2, +\infty)$
N_j	8	8	4

mentre le probabilità assegnate agli stessi tre intervalli dalla legge $\mathfrak{N}(1, 4)$ sono ricavati dalle Tavole E.1 e dalla solita procedura di standardizzazione del Teorema 4.11

$$p_1 = \boldsymbol{P}\{-\infty \le \mathfrak{N}(1,4) \le 0\} = \boldsymbol{P}\{-\infty \le \mathfrak{N}(0,1) \le -{}^1\!/_2\} = \Phi(-0.5) \simeq 0.309$$
$$p_2 = \boldsymbol{P}\{0 \le \mathfrak{N}(1,4) \le 2\} = \boldsymbol{P}\{-{}^1\!/_2 \le \mathfrak{N}(0,1) \le {}^1\!/_2\} = \Phi(0.5) - \Phi(-0.5) \simeq 0.383$$
$$p_3 = \boldsymbol{P}\{2 \le \mathfrak{N}(1,4) \le +\infty\} = \boldsymbol{P}\{{}^1\!/_2 \le \mathfrak{N}(0,1) \le +\infty\} = 1 - \Phi(0.5) \simeq 0.309$$

Avremo pertanto

intervalli	j	N_j	p_j	np_j	$\frac{(N_j-npj)^2}{np_j}$
$(-\infty, 0]$	1	8	0.309	6.2	0.542
$(0, 2]$	2	8	0.383	7.7	0.015
$(2, +\infty)$	3	4	0.309	6.2	0.764
				$K_0 =$	**1.321**

mentre l'evento critico di livello $\alpha = 0.05$ con $m = 3$ classi sarà definito dal quantile

$$\chi^2_{1-\alpha}(m-1) = \chi^2_{0.950}(2) \simeq 5.99$$

Avremo così

$$K_0 \simeq 1.32 < 5.99 \simeq \chi^2_{1-\alpha}(m-1)$$

per cui l'evento critico non si verifica, i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.35. *Le misure della larghezza del cranio (in mm) effettuate su un campione di $n = 84$ scheletri etruschi hanno una media $\overline{x} = 143.8$ e una varianza $s^2 = 36.0$. La tabella delle frequenze assolute negli intervalli indicati è*

larghezza (mm)	frequenza
$(-\infty, 135]$	5
$[135, 140]$	10
$[140, 145]$	33
$[145, 150]$	24
$[150, +\infty)$	12

Decidere con un test del chi-quadro di livello $\alpha = 0.05$ se è accettabile la seguente ipotesi

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge normale } \mathfrak{N}(\overline{x}, s^2) = \mathfrak{N}(143.8\,,\, 36.0)$$

Soluzione: Le probabilità assegnate ai $m = 5$ intervalli dalla legge $\mathfrak{N}(\overline{x}, s^2) = \mathfrak{N}(143.8\,,\, 36.0)$ sono ricavati dalle Tavole E.1 e dalla procedura di standardizzazione del Teorema 4.11

$$\begin{aligned}
p_1 &= \boldsymbol{P}\big\{-\infty \le \mathfrak{N}(\overline{x}, s^2) \le 135\big\} = \boldsymbol{P}\{-\infty \le \mathfrak{N}(0,1) \le -1.47\} \simeq 0.071\\
p_2 &= \boldsymbol{P}\big\{135 \le \mathfrak{N}(\overline{x}, s^2) \le 140\big\} = \boldsymbol{P}\{-1.47 \le \mathfrak{N}(0,1) \le -0.63\} \simeq 0.192\\
p_3 &= \boldsymbol{P}\big\{140 \le \mathfrak{N}(\overline{x}, s^2) \le 145\big\} = \boldsymbol{P}\{-0.63 \le \mathfrak{N}(0,1) \le 0.20\} \simeq 0.316\\
p_4 &= \boldsymbol{P}\big\{145 \le \mathfrak{N}(\overline{x}, s^2) \le 150\big\} = \boldsymbol{P}\{0.20 \le \mathfrak{N}(0,1) \le 1.03\} \simeq 0.270\\
p_5 &= \boldsymbol{P}\big\{150 \le \mathfrak{N}(\overline{x}, s^2) \le +\infty\big\} = \boldsymbol{P}\{1.03 \le \mathfrak{N}(0,1) \le +\infty\} \simeq 0.151
\end{aligned}$$

Avremo pertanto

intervalli	j	N_j	p_j	np_j	$\frac{(N_j-npj)^2}{np_j}$
$(-\infty, 135]$	1	5	0.071	6.0	0.162
$[135, 140]$	2	10	0.192	16.1	2.330
$[140, 145)$	3	33	0.316	26.5	1.570
$[145, 150]$	4	24	0.270	22.7	0.077
$[150, +\infty)$	5	12	0.151	12.7	0.034
				$K_0 =$	**4.173**

mentre l'evento critico di livello $\alpha = 0.05$ con $m = 5$ classi e $q = 2$ parametri stimati sarà definito dal quantile

$$\chi^2_{1-\alpha}(m - q - 1) = \chi^2_{0.950}(2) \simeq 5.99$$

Avremo così

$$K_0 \simeq 4.17 < 5.99 \simeq \chi^2_{1-\alpha}(m - q - 1)$$

per cui l'evento critico non si verifica, i dati non sono in regione critica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.36. *Con $n = 300$ misure di una v-a X che prende solo valori interi si ottiene la seguente tabella di frequenze*

$j =$	0	1	2	3	4	5	6
$N_j =$	40	81	83	52	27	13	4

Decidere con un test del chi-quadro di livello $\alpha = 0.01$ se è accettabile la seguente ipotesi

$$\mathcal{H}_0 : \textit{ il campione si adatta alla legge binomiale } \mathfrak{B}\left(6; {}^1/_3\right)$$

Risposta: Siccome $n\,\boldsymbol{P}\{X = 6\} \simeq 0.42$ e $n\,\boldsymbol{P}\{X = 5\} \simeq 4.94$, è necessario prima raggruppare i valori $j = 5$ e $j = 6$ riducendo le classi da 7 a 6. Con i nuovi dati si ottiene

$$K_0 \simeq 38.15 > 15.09 \simeq \chi^2_{0.990}(5)$$

per cui l'evento critico si verifica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.37. *Le misure in cm dell'altezza X di $n = 100$ persone si distribuiscono con le seguenti frequenze assolute*

altezza	N_j
≤ 168	*10*
$[168\,,\,169]$	*17*
$[169\,,\,170]$	*24*
$[170\,,\,171]$	*27*
$[171\,,\,172]$	*17*
≥ 172	*5*

Supponendo note la media $\mu = 170$ *e la varianza* $\sigma^2 = 2$*, decidere con un test del chi-quadro di livello* $\alpha = 0.05$ *se è accettabile la seguente ipotesi*

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge normale } \mathfrak{N}(\mu, \sigma^2) = \mathfrak{N}(170,\, 2)$$

Risposta: Non c'è bisogno di raggruppamenti di classi: si trova

$$K_0 \simeq 1.92 < 11.07 \simeq \chi^2_{0.950}(5)$$

per cui l'evento critico non si verifica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.38. *Quattro dadi vengono lanciati assieme per* $n = 10\,000$ *volte, e in ogni lancio si osserva quante volte esce "*6*". I valori* $j = 0, 1, 2, 3, 4$ *del* numero dei "6" in ogni lancio *si presentano con le seguenti frequenze empiriche assolute*

$j =$	0	1	2	3	4
$N_j =$	4 775	3 919	1 143	156	7

Decidere con un test del chi-quadro di livello $\alpha = 0.05$ *se è accettabile la seguente ipotesi*

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge binomiale } \mathfrak{B}\left(4;\, {}^1\!/_6\right)$$

Risposta: Non c'è bisogno di raggruppamenti di classi: si trova

$$K_0 \simeq 1.70 < 9.49 \simeq \chi^2_{0.950}(4)$$

per cui l'evento critico non si verifica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.39. *Tre dadi vengono lanciati assieme per* $n = 8\,000$ *volte, e in ogni lancio si osserva quante volte esce "*6*". I quattro valori* $j = 0, 1, 2, 3$ *del* numero dei "6" in ogni lancio *si presentano con le seguenti frequenze empiriche assolute*

$j =$	0	1	2	3
$N_j =$	4 481	2 868	603	48

Decidere con un test del chi-quadro di livello $\alpha = 0.05$ *se è accettabile la seguente ipotesi*

$$\mathcal{H}_0 \,:\, \textit{il campione si adatta alla legge binomiale } \mathfrak{B}\left(3;\, {}^1\!/_6\right)$$

Risposta: Non c'è bisogno di raggruppamenti di classi: si trova

$$K_0 \simeq 14.00 > 7.81 \simeq \chi^2_{0.950}(3)$$

per cui l'evento critico si verifica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.40. *Si misurano la lunghezza X e il peso Y di $n = 600$ pezzi prodotti da una fabbrica per controllare se sono:* troppo lunghi, giusti, *o* troppo corti *in X;* troppo pesanti, giusti, troppo leggeri *in Y. I risultati della verifica sono riassunti nella seguente tabella di contingenza*

	corti	giusti	lunghi
leggeri	6	48	8
giusti	52	402	36
pesanti	6	38	4

Stabilire con un test del χ^2 di livello $\alpha = 0.05$ se è accettabile l'ipotesi

$$\mathcal{H}_0 \ : \ \textit{le quantià misurate } X \textit{ e } Y \textit{ sono indipendenti}$$

Soluzione: Conviene riprodurre innanzitutto la tabella di contingenza completandola con le marginali e il numero totale degli oggetti misurati che viene riportato nell'angolo in basso a destra. Successivamente sotto ognuna delle frequenze della tabella (numeri in grassetto) conviene riportare il relativo valore delle quantità

$$n\overline{p}_j\overline{q}_k = \frac{N_{j\,\cdot}\, N_{\cdot\, k}}{n}$$

necessarie per calcolare poi gli addendi della statistica di Pearson (9.44)

$$\frac{(N_{jk} - n\overline{p}_j\overline{q}_k)^2}{n\overline{p}_j\overline{q}_k}$$

i cui valori sono registrati nella seconda riga sotto ciascuna frequenza

	corti	*giusti*	*lunghi*	marg
leggeri	***6***	***48***	***8***	***62***
	6.61	50.43	4.96	
	0.057	0.117	1.863	
giusti	***52***	***402***	***36***	***490***
	52.27	398.53	39.20	
	0.001	0.030	0.261	
pesanti	***6***	***38***	***4***	***48***
	5.12	39.04	3.84	
	0.151	0.028	0.007	
marg	***64***	***488***	***48***	***600***

A questo punto la statistica di Pearson si ottiene sommando tutti i termini riportati nella seconda riga sotto le frequenze

$$K_0 = \sum_{j,k} \frac{(N_{jk} - n\overline{p}_j\overline{q}_k)^2}{n\overline{p}_j\overline{q}_k} \simeq 2.515$$

L'evento critico (9.45) di livello $\alpha = 0.05$ con $r = s = 3$ classi di possibili valori di X e Y è d'altra parte definito dal quantile del *chi-quadro*

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.950}(4) \simeq 9.488$$

e siccome risulta

$$K_0 \simeq 2.515 < 9.488 \simeq \chi^2_{1-\alpha}[(r-1)(s-1)]$$

l'evento critico non si verifica e quindi accetteremo l'ipotesi $\mathcal{H}_0$ di indipendenza □

Esercizio A.3.41. *Si sceglie un campione di $n = 250$ individui che usano quotidianamente un automezzo privato per raggiungere il posto di lavoro: ogni soggetto è classificato in base alla potenza X della propria vettura e alla distanza Y in Km che percorre ogni giorno ottenendo i dati della seguente tabella di contingenza*

	0/10 *Km*	10/20 *Km*	> 20 *Km*
molto potente	6	27	19
potente	8	36	17
normale	21	45	33
piccola	14	18	6

Stabilire con un test del χ^2 di livello $\alpha = 0.05$ se è accettabile l'ipotesi

$$\mathcal{H}_0 : \textit{le quantià misurate } X \textit{ e } Y \textit{ sono indipendenti}$$

Soluzione: La tabella di contingenza completata come nel caso precedente è

	0/10 *Km*	10/20 *Km*	> 20 *Km*	marg
molto potente	***6***	***27***	***19***	***52***
	10.19	26.21	15.60	
	1.724	0.024	0.741	
potente	***8***	***36***	***17***	***61***
	11.96	30.74	18.30	
	1.309	0.899	0.092	
normale	***21***	***45***	***33***	***99***
	19.40	49.90	29.70	
	0.131	0.480	0.367	
piccola	***14***	***18***	***6***	***38***
	7.45	19.15	11.40	
	5.764	0.069	2.558	
marg	***49***	***126***	***75***	***250***

e quindi la statistica di Pearson (9.44) vale

$$K_0 \simeq 14.158$$

Siccome il quantile del *chi-quadro* che definisce l'evento critico di livello $\alpha = 0.05$ con $r = 4, s = 3$ è

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.950}(6) \simeq 12.592$$

avremo

$$K_0 \simeq 14.16 > 12.59 \simeq \chi^2_{1-\alpha}[(r-1)(s-1)]$$

sicché l'evento critico si verifica, i dati sono in regione critica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ di indipendenza □

Esercizio A.3.42. *Viene effettuata un'indagine per sapere quale mezzo di comunicazione è considerato più affidabile: ad ogni individuo viene richiesta età, sesso, titolo di studio e mezzo di comunicazione ritenuto più affidabile. I risultati sono riassunti nelle seguenti tre tabelle di contingenza*

	giornale	televisione	radio
<35 anni	30	68	10
35/54 anni	61	78	21
>54 anni	98	43	21

	giornale	televisione	radio
maschio	92	108	20
femmina	97	81	32

	giornale	televisione	radio
media inferiore	45	22	6
media superiore	95	115	33
università	49	52	13

Stabilire con un test del χ^2 di livello $\alpha = 0.05$ se sono separatamente accettabili le tre ipotesi

$\mathcal{H}_0$: *il giudizio è indipendente rispettivamente da età, sesso o titolo di studio*

Soluzione: Le tre tabelle completate sono le seguenti

	giornale	*televisione*	*radio*	marg
<35 *anni*	***30***	***68***	***10***	***108***
	47.47	47.47	13.06	
	6.429	8.879	0.717	
35/54 *anni*	***61***	***78***	***21***	***160***
	70.33	70.33	19.35	
	1.237	0.837	0.141	
>54 *anni*	***98***	***43***	***21***	***162***
	71.20	71.20	19.59	
	10.083	11.172	0.101	
marg	***189***	***189***	***52***	***430***

	giornale	*televisione*	*radio*	marg
maschio	***92***	***108***	***20***	***220***
	96.70	96.70	26.60	
	0.228	1.321	1.640	
femmina	***97***	***81***	***32***	***210***
	92.30	92.30	25.40	
	0.239	1.384	1.718	
marg	***189***	***189***	***52***	***430***

	giornale	*televisione*	*radio*	marg
media inferiore	***45***	***22***	***6***	***73***
	32.09	32.09	8.83	
	5.198	3.170	0.906	
media superiore	***95***	***115***	***33***	***243***
	106.81	106.81	29.39	
	1.305	0.628	0.444	
università	***49***	***52***	***13***	***114***
	50.11	50.11	13.79	
	0.024	0.072	0.045	
marg	***189***	***189***	***52***	***430***

I valori della statistica di Pearson e dei quantili del *chi-quadro* per eventi critici di livello $\alpha = 0.05$ sono per le tre tabelle

$$K_0 \simeq \begin{cases} 39.60 > 9.49 \simeq \chi^2_{0.95}(4) & (r=3, s=3) \quad \text{età} \\ 6.53 > 5.99 \simeq \chi^2_{0.95}(2) & (r=3, s=2) \quad \text{sesso} \\ 11.79 > 9.49 \simeq \chi^2_{0.95}(4) & (r=3, s=3) \quad \text{titolo} \end{cases}$$

per cui gli eventi critici si verificano in tutti e tre i casi e quindi rifiuteremo tutte e tre le ipotesi $\mathcal{H}_0$ di indipendenza dei giudizi da età, sesso e titolo di studio □

Esercizio A.3.43. *Un'azienda vuole verificare l'affidabilità di tre diverse configurazioni (A, B e C) di una macchina industriale esaminando i guasti a cui essa è soggetta: sapendo che ci sono quattro possibili tipi di guasti (1, 2, 3 e 4) e che i dati sono quelli della seguente tabella*

	1	*2*	*3*	*4*
A	20	44	17	9
B	4	17	7	12
C	10	31	14	5

Stabilire con un test del χ^2 *di livello* $\alpha = 0.05$ *se è accettabile l'ipotesi*

$\mathcal{H}_0$: *il tipo di guasti è indipendente dalla configurazione della macchina*

Soluzione: La tabella di contingenza completata è

	1	2	3	4	marg
A	***20***	***44***	***17***	***9***	***90***
	16.11	43.58	18.00	12.32	
	0.942	0.004	0.056	0.893	
B	***4***	***17***	***7***	***12***	***40***
	7.16	19.37	0.00	5.47	
	1.393	0.290	0.125	7.781	
C	***10***	***31***	***14***	***5***	***60***
	10.74	29.05	12.00	8.21	
	0.051	0.131	0.333	1.255	
marg	***34***	***92***	***38***	***26***	***190***

e quindi la statistica di Pearson (9.44) vale

$$K_0 \simeq 12.739$$

Siccome il quantile del *chi-quadro* che definisce l'evento critico di livello $\alpha = 0.05$ con $r = 3, s = 4$ è

$$\chi^2_{1-\alpha}[(r-1)(s-1)] = \chi^2_{0.950}(6) \simeq 12.592$$

avremo

$$K_0 \simeq 12.74 > 12.59 \simeq \chi^2_{1-\alpha}[(r-1)(s-1)]$$

sicché l'evento critico si verifica, i dati sono in regione critica e quindi rifiuteremo l'ipotesi $\mathcal{H}_0$ di indipendenza. Data la ristrettezza del margine, però, l'esito del test appare poco significativo □

Esercizio A.3.44. *Il direttore di una scuola vuol sapere se l'opinione delle famiglie su un certo cambiamento di orario scolastico dipende dal fatto che la loro abitazione è situata in una zona urbana o in una zona rurale. Si chiede pertanto il parere di* $n = 500$ *famiglie ottenendo i risultati riportati nella seguente tabella*

	favorevoli	*contrari*	*indifferenti*
urbana	123	36	41
rurale	145	85	70

Stabilire con un test del χ^2 *di livello* $\alpha = 0.01$ *se è accettabile l'ipotesi*

$\mathcal{H}_0$: *l'opinione delle famiglie è indipendente dalla collocazione della loro abitazione*

Risposta: Si verifica l'evento critico: $K_0 \simeq 9.610 > 9.210 \simeq \chi^2_{0.990}(2)$, per cui si rifiuta l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.45. *In un paese si vuol sapere se le risposte (favorevole o contrario) dei cittadini ad una determinata questione sono o meno influenzate dall'età . Si chiede pertanto il parere di* $n = 600$ *persone ottenendo i risultati riportati nella seguente tabella*

	giovani	*maturi*	*anziani*
favorevoli	100	150	30
contrari	126	120	74

Stabilire con un test del χ^2 *di livello* $\alpha = 0.05$ *se è accettabile l'ipotesi*

$$\mathcal{H}_0 : \textit{l'opinione dei cittadini è indipendente dall'età}$$

Risposta: Si verifica l'evento critico: $K_0 \simeq 22.37 > 5.99 \simeq \chi^2_{0.950}(2)$, per cui si rifiuta l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.46. *Una ditta che vende automobili vuol sapere se l'età degli acquirenti (giovani, maturi e anziani) influenza la scelta del colore (bianco, rosso o nero) delle vetture vendute. Si esaminano i dati relativi a* $n = 500$ *contratti di vendita ottenendo i risultati riportati nella seguente tabella*

	giovani	*maturi*	*anziani*
bianco	56	60	40
rosso	47	87	38
nero	23	64	85

Stabilire con un test del χ^2 *di livello* $\alpha = 0.05$ *se è accettabile l'ipotesi*

$$\mathcal{H}_0 : \textit{il colore scelto è indipendente dall'età degli acquirenti}$$

Risposta: Si verifica l'evento critico: $K_0 \simeq 44.40 > 9.49 \simeq \chi^2_{0.950}(4)$, per cui si rifiuta l'ipotesi $\mathcal{H}_0$ □

Esercizio A.3.47. *Un professore vuol sapere se i voti (in trentesimi) registrati per il suo esame dipendono o meno dall'anno di immatricolazione degli studenti. Egli esamina i voti* $n = 378$ *studenti immatricolati in diversi anni accademici e ottiene i risultati riportati nella seguente tabella*

	18-20	*21-23*	*24-26*	*27-30*
2000/01	17	30	50	15
2001/02	15	26	34	10
2002/03	11	25	38	19
2003/04	9	40	29	10

Stabilire con un test del χ^2 *di livello* $\alpha = 0.01$ *se è accettabile l'ipotesi*

$$\mathcal{H}_0 : \textit{il voto è indipendente dall'anno di immatricolazione}$$

Risposta: Non si verifica l'evento critico: $K_0 \simeq 14.03 < 21.67 \simeq \chi^2_{0.990}(9)$, per cui si accetta l'ipotesi $\mathcal{H}_0$ □

Appendice B

Schemi

B.1 Formulario di Statistica Inferenziale

Stime e test per l'attesa di una *v-a* X

Sia $X_1, \ldots, X_n$ un campione di una *v-a* X con attesa μ e varianza σ^2. Notazioni:

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad \overline{X_n^2} = \frac{1}{n}\sum_{j=1}^{n} X_j^2 \qquad S_n^2 = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^2\right)$$

φ_α e $t_\alpha(n)$ saranno i quantili di ordine α rispettivamente delle leggi $\mathfrak{N}(0,1)$ e $\mathfrak{T}(n)$; infine poniamo:

$$U_0 = \frac{\overline{X}_n - \mu_0}{\sigma}\sqrt{n} \qquad T_0 = \frac{\overline{X}_n - \mu_0}{S_n}\sqrt{n}$$

Intervalli di fiducia di livello α per μ:

1. varianza σ^2 nota: $\overline{X}_n \pm \varphi_{1-\frac{\alpha}{2}}\sigma/\sqrt{n}$
2. varianza σ^2 non nota, e $n \leq 120$: $\overline{X}_n \pm t_{1-\frac{\alpha}{2}}(n-1)S_n/\sqrt{n}$
3. varianza σ^2 non nota, e $n > 120$: $\overline{X}_n \pm \varphi_{1-\frac{\alpha}{2}}S_n/\sqrt{n}$

Test bilaterale per le seguenti ipotesi sui valori dell'attesa μ:

$$\mathcal{H}_0: \ \mu = \mu_0 \qquad \mathcal{H}_1: \ \mu \neq \mu_0$$

Eventi critici di livello α:

1. σ^2 nota: $\left\{|U_0| > \varphi_{1-\frac{\alpha}{2}}\right\}$
2. σ^2 non nota, e $n \leq 120$: $\left\{|T_0| > t_{1-\frac{\alpha}{2}}(n-1)\right\}$
3. σ^2 non nota, e $n > 120$: $\left\{|T_0| > \varphi_{1-\frac{\alpha}{2}}\right\}$

Test unilaterali per le seguenti ipotesi sui valori dell'attesa μ:

$$\begin{array}{ll} \mathcal{H}_0: \ \mu \leq \mu_0 & \mathcal{H}_1: \ \mu > \mu_0 \\ \mathcal{H}_0: \ \mu \geq \mu_0 & \mathcal{H}_1: \ \mu < \mu_0 \end{array}$$

Eventi critici di livello α: rispettivamente

1. σ^2 nota: $\{U_0 > \varphi_{1-\alpha}\}$ e $\{U_0 < -\varphi_{1-\alpha}\}$
2. σ^2 non nota, e $n \leq 120$: $\{T_0 > t_{1-\alpha}(n-1)\}$ e $\{T_0 < -t_{1-\alpha}(n-1)\}$
3. σ^2 non nota, e $n > 120$: $\{T_0 > \varphi_{1-\alpha}\}$ e $\{T_0 < -\varphi_{1-\alpha}\}$

Confronto fra le attese di due *v-a* X e Y

Ipotesi per i **test bilaterale e unilaterale** sull'eguaglianza delle attese:

$$\mathcal{H}_0: \ \mu_X = \mu_Y \qquad \mathcal{H}_1: \ \mu_X \neq \mu_Y$$
$$\mathcal{H}_0: \ \mu_X \leq \mu_Y \qquad \mathcal{H}_1: \ \mu_X > \mu_Y$$

Campioni accoppiati $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_n$; σ^2 varianza di $Z = X - Y$

$$Z_k = X_k - Y_k \qquad \overline{Z}_n = \frac{1}{n}\sum_{k=1}^{n} Z_k \qquad \overline{Z_n^2} = \frac{1}{n}\sum_{k=1}^{n} Z_k^2 \qquad S_n^2 = \frac{n}{n-1}\left(\overline{Z_n^2} - \overline{Z}_n^2\right)$$
$$U_0 = \frac{\overline{Z}_n}{\sigma}\sqrt{n} \qquad T_0 = \frac{\overline{Z}_n}{S_n}\sqrt{n}$$

Se σ^2 è nota, eventi critici bilaterale e unilaterale di livello α rispettivamente

$$\left\{|U_0| > \varphi_{1-\frac{\alpha}{2}}\right\} \qquad \left\{U_0 > \varphi_{1-\alpha}\right\}$$

Se σ^2 non è nota, eventi critici bilaterale e unilaterale di livello α rispettivamente

$$\left\{|T_0| > t_{1-\frac{\alpha}{2}}(n-1)\right\} \qquad \left\{T_0 > t_{1-\alpha}(n-1)\right\}$$

Campioni indipendenti $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$ di X e Y con attese μ_X e μ_Y e varianze σ_X^2 e σ_Y^2. Notazioni:

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad \overline{X_n^2} = \frac{1}{n}\sum_{j=1}^{n} X_k^2 \qquad \overline{Y}_m = \frac{1}{n}\sum_{k=1}^{m} Y_k \qquad \overline{Y_m^2} = \frac{1}{m}\sum_{k=1}^{m} Y_k^2$$
$$S_X^2 = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^2\right) \qquad S_Y^2 = \frac{m}{m-1}\left(\overline{Y_m^2} - \overline{Y}_m^2\right)$$
$$V^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$
$$U_0 = \frac{\overline{X}_n - \overline{Y}_m}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \qquad T_0 = \frac{\overline{X}_n - \overline{Y}_m}{V\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

Se σ_X^2 e σ_Y^2 sono note: eventi critici bilaterale e unilaterale di livello α

$$\left\{|U_0| > \varphi_{1-\frac{\alpha}{2}}\right\} \qquad \left\{U_0 > \varphi_{1-\alpha}\right\}$$

Se σ_X^2 e σ_Y^2 non sono note: eventi critici bilaterale e unilaterale di livello α

$$\left\{|T_0| > t_{1-\frac{\alpha}{2}}(n+m-2)\right\} \qquad \left\{T_0 > t_{1-\alpha}(n+m-2)\right\}$$

Stime e test per le varianze

$X_1, \ldots, X_n$ campione di X con attesa μ e varianza σ^2 sconosciute. Notazioni:

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad \overline{X_n^2} = \frac{1}{n}\sum_{j=1}^{n} X_j^2 \qquad S_n^2 = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^2\right)$$

Inoltre $\chi^2_\alpha(n)$ siano i quantili di ordine α della legge $\boldsymbol{\chi^2}(n)$
Intervallo di fiducia di livello α per σ^2:

$$\left[\frac{(n-1)S_n^2}{\chi^2_{1-\frac{\alpha}{2}}(n-1)}\ ,\ \frac{(n-1)S_n^2}{\chi^2_{\frac{\alpha}{2}}(n-1)}\right]$$

Per $n > 35$ i quantili $\chi^2_\alpha(n)$ non tabulati si possono calcolare dai quantili normali:

$$\chi^2_\alpha(n) \simeq \frac{1}{2}\left(\varphi_\alpha + \sqrt{2n-1}\right)^2$$

Confronto fra le varianze di due campioni indipendenti $X_1, \ldots, X_n$ e $Y_1, \ldots, Y_m$

$$\overline{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j \qquad \overline{X_n^2} = \frac{1}{n}\sum_{j=1}^{n} X_k^2 \qquad \overline{Y}_m = \frac{1}{n}\sum_{k=1}^{m} Y_k \qquad \overline{Y_m^2} = \frac{1}{m}\sum_{k=1}^{m} Y_k^2$$

$$S_X^2 = \frac{n}{n-1}\left(\overline{X_n^2} - \overline{X}_n^2\right) \qquad S_Y^2 = \frac{m}{m-1}\left(\overline{Y_m^2} - \overline{Y}_m^2\right)$$

$$F_0 = \frac{S_X^2}{S_Y^2}$$

$f_\alpha(n,m)$ quantili di ordine α della legge di Fisher $\mathfrak{F}(n,m)$; ricordare anche che

$$f_\alpha(n,m) = \frac{1}{f_{1-\alpha}(m,n)}$$

Test bilaterale di Fisher: ipotesi sull'eguaglianza delle varianze

$$\mathcal{H}_0:\ \sigma_X^2 = \sigma_Y^2 \qquad \mathcal{H}_1:\ \sigma_X^2 \neq \sigma_Y^2$$

evento critico di livello α:

$$\begin{aligned}\left\{F_0 < f_{\frac{\alpha}{2}}(n-1,m-1)\right\} \cup \left\{F_0 > f_{1-\frac{\alpha}{2}}(n-1,m-1)\right\} \\ = \overline{\left\{f_{\frac{\alpha}{2}}(n-1,m-1) \le F_0 \le f_{1-\frac{\alpha}{2}}(n-1,m-1)\right\}}\end{aligned}$$

Test unilaterale di Fisher: ipotesi sull'eguaglianza delle varianze:

$$\mathcal{H}_0:\ \sigma_X^2 \le \sigma_Y^2 \qquad \mathcal{H}_1:\ \sigma_X^2 > \sigma_Y^2$$

evento critico di livello α:

$$\left\{F_0 > f_{1-\alpha}(n-1,m-1)\right\}$$

Test di adattamento e di indipendenza

Test del χ^2 per l'adattamento di campioni a distribuzioni teoriche
v-a con m valori, campione di numerosità n, frequenze assolute $N_1, \ldots, N_m$, distribuzione teorica $p_1, \ldots, p_m$. Notazione:

$$\overline{p}_j = \frac{N_j}{n} \qquad K_0 = n \sum_{j=1}^{m} \frac{(\overline{p}_j - p_j)^2}{p_j} = \sum_{j=1}^{m} \frac{(N_j - np_j)^2}{np_j}$$

Inoltre $\chi^2_\alpha(m)$ siano i quantili di ordine α della legge $\boldsymbol{\chi^2}(m)$

> Requisito: n deve essere abbastanza grande da avere $np_j \geq 5$ per $j = 1, \ldots, m$; altrimenti bisogna unificare le classi più povere

Evento critico di livello α:

$$\{K_0 > \chi^2_{1-\alpha}(m-1)\}$$

Se la conoscenza della distribuzione teorica richiede la stima (di MV) di q) parametri, la regione critica diviene

$$\{K_0 > \chi^2_{1-\alpha}(m-q-1)\}$$

Test del χ^2 per l'indipendenza di campioni
X prende valori $u_1, \ldots, u_r$, Y prende valori $v_1, \ldots, v_s$
Valori del campione accoppiato (u_j, v_k) con $j = 1, \ldots, r$ e $k = 1, \ldots, s$

N_{jk} frequenza congiunta assoluta dei valori (u_j, v_k)

$N_{j\cdot}$ frequenza marginale assoluta dei valori u_j

$N_{\cdot k}$ frequenza marginale assoluta dei valori v_k

$$\overline{p}_j = \frac{N_{j\cdot}}{n} \qquad \overline{q}_k = \frac{N_{\cdot k}}{n} \qquad K_0 = \sum_{j,k} \frac{\left(N_{jk} - n\overline{p}_j\overline{q}_k\right)^2}{n\overline{p}_j\overline{q}_k}$$

Ricordare anche che

$$n\overline{p}_j\overline{q}_k = \frac{N_{j\cdot}\, N_{\cdot k}}{n}$$

Evento critico di livello α:

$$\{K_0 > \chi^2_{1-\alpha}\left[(r-1)(s-1)\right]\}$$

Appendice C

Domande

In questa appendice sono elencate alcune tipiche domande teoriche con l'indicazione delle sezioni del testo alle quali esse fanno riferimento: lo scopo è quello di far esercitare lo studente nella comprensione precisa dei quesiti, e nella formulazione compiuta delle relative risposte

C.1 Calcolo delle probabilità

Domanda C.1.1. *Definire la convergenza in distribuzione di una successione di v-a* $(X_n)_{n\in\boldsymbol{N}}$*; enunciare il Teorema Limite Centrale mostrando in particolare che le v-a* Y_n^* *sono standardizzate; discutere l'idea di approssimazione normale e i requisiti su* n *che ne garantiscono l'applicabilità*

Risposta: Definizione 5.2; Teorema 5.5 e discussione successiva fino all'Esempio 5.6 escluso □

Domanda C.1.2. *Definire la varianza di una v-a* X*, la covarianza e il coefficiente di correlazione di due v-a* X, Y*; dimostrare le proprietà delle varianze di* $aX + b$ *(dove* a, b *sono numeri) e di* $X + Y$

Risposta: Definizione 4.4; Teorema 4.9 □

Domanda C.1.3. *Definire la convergenza in distribuzione di una successione di v-a* $(X_n)_{n\in\boldsymbol{N}}$*; enunciare e dimostrare il Teorema di Poisson sulla approssimazione per* $n \to \infty$ *delle distribuzioni binomiali* $\mathfrak{B}\left(n; {}^{\lambda}/_{n}\right)$ *con la legge di Poisson* $\mathfrak{P}(\lambda)$

Risposta: Definizione 5.2; Teorema 5.8 □

Domanda C.1.4. *Definire il coefficiente di correlazione* ρ_{XY} *di due v-a* X, Y *e discutere le sue proprietà; definire la non correlazione di due v-a* X, Y*, e discutere il rapporto fra indipendenza e non correlazione*

Risposta: Definizione 4.4; Teorema 4.7; Definizioni 4.5; Teorema 4.6 □

Domanda C.1.5. *Definire l'indipendenza delle componenti di un vettore aleatorio* $\boldsymbol{X} = (X_1, \ldots, X_m)$*, e spiegare le conseguenze dell'indipendenza sulla FDC congiunta, sulle probabilità congiunte (caso discreto) e sulle fdp congiunte*

Risposta: Definizione 3.8; Teorema 3.13; Sezione 3.6 esclusi gli esempi □

Domanda C.1.6. *Definire la varianza* σ_X *di una v-a* X*, e la covarianza* κ_{XY} *di due v-a* X, Y*; enunciare e dimostrare una procedura di calcolo di* σ_X *e* κ_{XY} *diversa dalla definizione*

Risposta: Definizione 4.4; Teorema 4.8 □

Domanda C.1.7. *Definire la probabilità condizionata* $\boldsymbol{P}\{A \mid B\}$*; enunciare e dimostrare la Formula della probabilità totale e il Teorema di Bayes*

Risposta: Definizione 2.2, Teoremi 2.3 e 2.5 □

Domanda C.1.8. *Enunciare (senza dimostrazione) il Teorema di Poisson, illustrare con degli esempi il suo ruolo nell'approssimazione di leggi binomiali, e nella determinazione della distribuzione di istanti aleatori*

Risposta: Teorema 5.8; e discussione seguente; Esempio 5.9; Esempio 5.10 □

Domanda C.1.9. *Enunciare e dimostrare le formule per le attese e le varianze della legge binomiale* $\mathfrak{B}(n;p)$ *e della legge di Poisson* $\mathfrak{P}(\lambda)$

Risposta: Teorema 4.13 □

Domanda C.1.10. *Definire uno spazio di probabilità* $(\Omega, \mathcal{F}, \boldsymbol{P})$ *spiegando il significato dei simboli usati; discutere il Problema delle Coincidenze*

Risposta: Definizioni 1.9, 1.4, 1.5, 1.6, 1.8; Esempio 1.10 □

Domanda C.1.11. *Dare la definizione di v-a centrate e standardizzate, e mostrare come si centra e si standardizza una generica v-a* X*. Definire i momenti e i momenti centrati di una v-a e discutere il significato di asimmetria e curtosi*

Risposta: Definizione 4.10; Teorema 4.11; Definizione 4.12 e sua discussione □

Domanda C.1.12. *Enunciare il Teorema Limite Centrale, discuterne il ruolo nella formulazione della legge degli errori, e spiegare infine quali accorgimenti numerici è opportuno adottare nel caso di approssimazioni normali di leggi discrete*

Risposta: Teorema 5.5; Esempio 5.6; Esempio 5.7 □

Domanda C.1.13. *Definire i quantili di una v-a, e dimostrare che per una v-a normale standard* $X \sim \mathfrak{N}(0,1)$ *si ha* $\boldsymbol{P}\big\{|X| \le \varphi_{1-\frac{\alpha}{2}}\big\} = \boldsymbol{P}\big\{\varphi_{\frac{\alpha}{2}} \le X \le \varphi_{1-\frac{\alpha}{2}}\big\} = 1-\alpha$

Risposta: Definizione 3.24; Teorema 3.25 □

Domanda C.1.14. *Scrivere esplicitamente la fdp e la FDC di una legge normale* $\mathfrak{N}(\mu,\sigma^2)$*, e di una legge normale standard* $\mathfrak{N}(0,1)$*; enunciare le loro proprietà di simmetria; dire quali sono le leggi di* $aX+b$ *e di* $X+Y$ *quando* X *e* Y *sono normali e indipendenti; mostrare come si calcola* $\boldsymbol{P}\{a \le X \le b\}$ *mediante la FDC normale standard* $\Phi(x)$ *quando* $X \sim \mathfrak{N}(\mu,\sigma^2)$

Risposta: Sezione 3.4.2; Teorema 3.19 □

Domanda C.1.15. *Enunciare (senza dimostrazione) la Legge dei grandi Numeri, illustrare con un esempio il suo ruolo in un problema di stima di parametri, e discutere il corrispondente uso della media aritmetica $\overline{X}_n$ e della varianza campionaria $\widehat{S}_n^2$*

Risposta: Teorema 5.3; Esempio 5.4 e discussione seguente fino alla Sezione 5.3 esclusa □

Domanda C.1.16. *Definire la FDC $F_X(x)$ di una v-a X e discuterne le proprietà; definire una v-a continua e la sua fdp $f_X(x)$, e spiegare in che relazione sono la FDC e la fdp*

Risposta: Definizione 3.9; Teoremi 3.10 e 3.11; Definizione 3.17 e discussione fino alla Definizione 3.18 esclusa □

Domanda C.1.17. *Definire il concetto di v-a X e discuterne un esempio; definire la distribuzione di una v-a X e discuterne un esempio; mostrare con un esempio che v-a diverse possono avere la stessa distribuzione*

Risposta: Definizione 3.1; Esempio 3.2; Definizione 3.3; Esempi 3.4 e 3.5 □

Domanda C.1.18. *Definire il concetto di valore d'attesa di una v-a X, estenderlo anche al caso di funzioni di una ($h(X)$) o più ($h(X,Y)$) v-a, e dimostrare le proprietà di linearità delle attese*

Risposta: Definizione 4.1 e discussione seguente; Teorema 4.2 □

Domanda C.1.19. *Definire l'indipendenza di due o più eventi; costruire il Modello di Bernoulli per descrivere n estrazioni indipendenti (con rimessa) di palline bianche e nere da un'urna; utilizzare infine tale modello per costruire un esempio di applicazione del Teorema di Bayes*

Risposta: Definizione 2.6; Teorema 2.7; Esempio 2.8 □

Domanda C.1.20. *Definire la convergenza degenere in media quadratica di una successione di v-a $(X_n)_{n\in\boldsymbol{N}}$; enunciare e dimostrare la Legge dei Grandi Numeri per la media aritmetica $\overline{X}_n$ e la varianza campionaria $\widehat{S}_n^2$*

Risposta: Definizione 5.1; Teorema 5.3 □

C.2 Statistica

Domanda C.2.1. *Dato un campione $(x_1, y_1), \ldots, (x_n, y_n)$ di due caratteri X e Y, definire l'errore quadratico medio $\mathcal{E}(a, b)$ rispetto a una retta. Definire poi la retta di regressione, ed enunciare e dimostrare le formule che permettono di calcolarne i coefficienti a e b a partire dai dati del campione*

Risposta: Definizioni 7.6 e 7.7; Teorema 7.8 □

Domanda C.2.2. *Assegnata una matrice $p \times n$ di dati $\|x_{jk}\|$ con $j = 1, \ldots, n$ e $k = 1, \ldots, p$, definire il baricentro $\overline{\mathbf{x}}$, la matrice di covarianza $\mathbb{S}$, la matrice di correlazione $\mathbb{R}$ e la dispersione totale Δ. Discutere il problema della rappresentazione dei dati e il Criterio di rappresentazione ottimale*

Risposta: Definizioni 7.10 e discussione dall'inizio della Sezione 7.4 fino al Teorema 7.11 escluso □

Domanda C.2.3. *Dato un campione $x_1, \ldots, x_n$ di un carattere X, definire moda, media $\overline{x}$, varianza s_X^2, scarto quadratico s_X e coefficiente di variazione δ; dimostrare poi che*

$$s_X^2 = \overline{x^2} - \overline{x}^2 \qquad \text{e che} \qquad s_Y^2 = a^2 s_X^2$$

dove abbiamo definito il campione trasformato $y_i = ax_i + b$

Risposta: Definizioni 6.5, 6.7 e 6.15; Teoremi 6.17 e 6.18 □

Domanda C.2.4. *Dato un campione $x_1, \ldots, x_n$ di un carattere numerico X, definire i momenti m_k di ordine k e i momenti centrati $\widetilde{m}_k$ di ordine k; definire l'asimmetria g_1 e la curtosi g_2; spiegare infine con qualche esempio il significato statistico dell'asimmetria e della curtosi*

Risposta: Sezione 6.5 □

Domanda C.2.5. *Dato un campione $(x_1, y_1), \ldots, (x_n, y_n)$ di due caratteri numerici X e Y, definire covarianza s_{XY} e coefficiente di correlazione r_{XY}, e spiegare cosa si intende per non correlazione di X e Y; dimostrare poi che*

$$s_{XY} = \overline{xy} - \overline{x}\,\overline{y}$$

elencare infine le proprietà principali del coefficiente di correlazione r_{XY}

Risposta: Definizioni 7.2 e 7.3 ; Teoremi 7.4 e 7.5 □

Domanda C.2.6. *Dato un campione $x_1, \ldots, x_n$, definire i quantili di ordine α, la mediana, i quartili, i decili e i percentili; piegare con degli esempi cosa vuol dire che la mediana è un indice di centralità più robusto della media, e a volte anche più rappresentativo*

Risposta: Definizione 6.25; discussione dei due Esempi 6.27 e 6.28 □

Domanda C.2.7. *Dato un campione dei caratteri $X_1, \ldots, X_p$, e dato un versore $\boldsymbol{v} = (v_1, \ldots, v_p)$, si consideri $Y = v_1X_1 + \ldots + v_pX_p$: mostrare che il versore $\boldsymbol{v}$ che rende massima la varianza $s^2_Y(\boldsymbol{v})$ è l'autovettore $\boldsymbol{v}_1$ associato all'autovalore più grande λ_1 della matrice di covarianza $\mathbb{S}$ dei dati. Esporre poi la Procedura per ottenere la rappresentazione più fedele, e definire le direzioni e i piani principali, e la fedeltà della rappresentazione*

Risposta: Discussione della Sezione 7.4 da Teorema 7.11 a Definizione 7.12 □

Domanda C.2.8. *Dato un campione $x_1, \ldots, x_n$ definire l'errore quadratico medio $\mathcal{E}(a)$ rispetto al numero a, e dimostrare che la media $\overline{x}$ è il valore di a che rende minimo $\mathcal{E}(a)$; spiegare cosa si intende per campione standardizzato, e mostrare come si standardizza un arbitrario campione $x_1, \ldots, x_n$*

Risposta: Definizioni 6.19 e 6.21; Teoremi 6.20 e 6.22 □

Domanda C.2.9. *Dato un campione $x_1, \ldots, x_n$ di un carattere numerico, definire la media pesata, e dimostrare che la media aritmetica $\overline{x}$ di un carattere numerico discreto è la media delle sue modalità w_k pesate con le frequenze relative. Definire le medie geometrica, armonica e quadratica, e la media generalizzata tramite una generica funzione $h(x)$; discutere almeno un esempio di applicazione*

Risposta: Definizione 6.7; Teorema 6.8; Sezione 6.6 □

Domanda C.2.10. *A partire da un campione $X_1, \ldots, X_n$ di una legge di Poisson $\mathfrak{P}(\lambda)$ con parametro λ sconosciuto, determinare – con dimostrazione – la stima di massima verosimiglianza per il parametro λ*

Risposta: Teorema 8.16 □

Domanda C.2.11. *Definire l'evento critico D, il livello α, la funzione potenza $\pi(\theta)$ e la significatività α_s di un test e spiegare il loro significato. Scrivere e giustificare la forma della regione critica per un test bilaterale di Gauss per la media (precisando il significato dei simboli usati), e descrivere la procedura per l'esecuzione del test*

Risposta: Definizione 9.3 e discussione seguente fino alla Sezione 9.1.1 esclusa; Sezione 9.2.1 fino ai test unilaterali esclusi □

Domanda C.2.12. *Sia X una v-a Binomiale $\mathfrak{B}(m; p)$ con m assegnato e p sconosciuto, e sia $X_1, \ldots, X_n$ un campione di n misure di X: determinare – con dimostrazione – la stima di massima verosimiglianza per il parametro p*

Risposta: Teorema 8.15 □

Domanda C.2.13. *Definire i concetti di campione aleatorio e di statistica. Dato poi un campione aleatorio $X_1, \dots, X_n$ con legge dipendente da un parametro θ, definire i concetti di stimatore, di stimatore corretto e di stimatore consistente di una funzione $h(\theta)$ del parametro; dimostrare infine che la media $\overline{X}_n$ e la varianza corretta S_n^2 del campione sono stimatori corretti e consistenti dell'attesa μ e della varianza σ^2, ma che la varianza campionaria $\widehat{S}_n^2$ è uno stimatore consistente ma non corretto di σ^2*

Risposta: Definizione 8.1 e Teorema 8.2 □

Domanda C.2.14. *Sia X una v-a normale $\mathfrak{N}(\mu, \sigma^2)$ con μ e σ^2 sconosciuti, e sia $X_1, \dots, X_n$ un campione di n misure di X: determinare – con dimostrazione – le stime di massima verosimiglianza per i parametri μ e σ^2. Si tratta di stimatori corretti?*

Risposta: Teorema 8.17 □

Domanda C.2.15. *Dato un campione $X_1, \dots, X_n$ di v.a. con legge dipendente da un parametro θ, discutere la necessità di adottare delle stime per intervalli, e definire l'intervallo di fiducia di livello α per la stima di di una funzione $h(\theta)$ del parametro; chiarire anche quale convenzione si adotta per determinare in maniera unica l'intervallo di fiducia. Spiegare infine come si costruiscono gli intervalli si fiducia dell'attesa μ*

Risposta: Definizione 8.9 con discussione e Sezione 8.2.1 □

Domanda C.2.16. *Spiegare in che modo la Legge dei Grandi Numeri permette di eseguire la stima di una distribuzione (discreta o continua) di una v-a X della quale sia dato un campione aleatorio $X_1, \dots, X_n$*

Risposta: Sezione 8.1.2 con Teorema 8.6 esclusi gli esempi □

Domanda C.2.17. *Spiegare cosa è un test di ipotesi e come si formulano le ipotesi; definire i possibili tipi di errori e discuterne l'importanza, chiarendo perché le rispettive probabilità di errore sono in competizione fra loro*

Risposta: Sezione 9.1 fino a Definizione 9.3 esclusa □

Domanda C.2.18. *Sia X una v-a uniforme $\mathfrak{U}(0, a)$ con parametro a sconosciuto, e sia $X_1, \dots, X_n$ un campione di n misure di X: determinare – con dimostrazione – lo stimatore di massima verosimiglianza del parametro a. Si tratta di uno stimatore corretto?*

Risposta: Teorema 8.18 □

Appendice D

Richiami

D.1 Calcolo vettoriale

In uno spazio reale p-dimensionale $\boldsymbol{R}^p$ i punti $\boldsymbol{x} = (x_1, \ldots, x_p)$ possono essere considerati come **vettori** applicati nell'origine dello spazio e con componenti x_k. Per questo motivo parleremo indifferentemente di *vettori* o di *punti* in $\boldsymbol{R}^p$. Chiameremo **vettore nullo** il vettore $\mathbf{0}$ le cui componenti sono tutte uguali a 0: ovviamente la sua rappresentazione coincide con l'origine di $\boldsymbol{R}^p$. I concetti geometrici che si introducono non sono altro che la naturale generalizzazione dei concetti usati nel ben noto caso $p = 3$ dello spazio tridimensionale naturale.

Chiameremo **modulo** del vettore $\boldsymbol{x}$ il numero non negativo

$$|\boldsymbol{x}| = \sqrt{\sum_{k=1}^{p} x_k^2} = \sqrt{x_1^2 + \ldots + x_p^2}$$

Si introducono poi tre operazioni principali:

- la **somma di due vettori**, $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$: è il vettore le cui componenti $z_k = x_k + y_k$ sono le somme delle corrispondenti componenti dei due addendi
- il **prodotto di un vettore per un numero**, $\boldsymbol{z} = a\,\boldsymbol{x}$: è il vettore le cui componenti $z_k = ax_k$ sono il prodotto delle corrispondenti componenti di $\boldsymbol{x}$ per il numero a
- il **prodotto scalare di due vettori**, $\boldsymbol{x} \cdot \boldsymbol{y}$: è il numero reale dato da

$$\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{k=1}^{p} x_k y_k = x_1 y_1 + \ldots + x_p y_p$$

È facile vedere che il prodotto scalare è simmetrico, cioè $\boldsymbol{x} \cdot \boldsymbol{y} = \boldsymbol{y} \cdot \boldsymbol{x}$, e inoltre che, dati due numeri reali a e b e tre vettori $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \boldsymbol{R}^p$, si ha

$$\boldsymbol{z} \cdot (a\boldsymbol{x} + b\boldsymbol{y}) = a\,\boldsymbol{z} \cdot \boldsymbol{x} + b\,\boldsymbol{z} \cdot \boldsymbol{y} \tag{D.1}$$

Naturalmente si ha anche dalle definizioni che $|\boldsymbol{x}| = \sqrt{\boldsymbol{x} \cdot \boldsymbol{x}}$, e che $|a\boldsymbol{x}| = |a||\boldsymbol{x}|$. Un ruolo importante è giocato dai vettori $\boldsymbol{v}$ di modulo $|\boldsymbol{v}| = 1$ detti anche **versori**. In particolare i punti della forma $a\,\boldsymbol{v}$ (cioè i vettori di direzione $\boldsymbol{v}$ e modulo $|a|$) descrivono, al variare di a, una intera retta passante per l'origine con la direzione di $\boldsymbol{v}$: in pratica i punti di ogni retta passante per l'origine hanno tutti questa forma

Ricordando le tipiche proprietà dei prodotti scalari negli spazi a 2 e 3 dimensioni, si può anche definire un **angolo** ϑ fra i vettori $\boldsymbol{x}$ e $\boldsymbol{y}$ dalla relazione $\boldsymbol{x} \cdot \boldsymbol{y} = |\boldsymbol{x}||\boldsymbol{y}| \cos\vartheta$: in pratica si pone

$$\cos\vartheta = \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|} \tag{D.2}$$

Questo permette di introdurre dei concetti geometrici, e in particolare si dice che due vettori $\boldsymbol{x}$ e $\boldsymbol{y}$ sono **ortogonali** se $\boldsymbol{x} \cdot \boldsymbol{y} = 0$

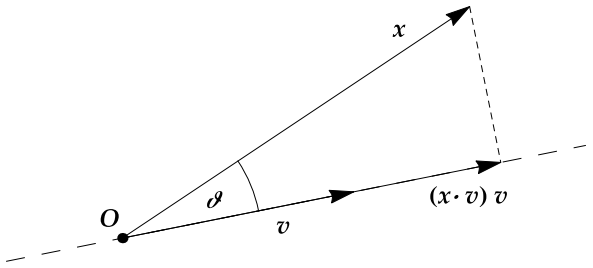


Figura D.1: Proiezione di $\boldsymbol{x}$ lungo la direzione di $\boldsymbol{v}$: il modulo del vettore proiettato è il prodotto scalare $\boldsymbol{x} \cdot \boldsymbol{v} = |\boldsymbol{x}| \cos\vartheta$

La **proiezione** di un vettore $\boldsymbol{x}$ sulla retta individuata dal versore $\boldsymbol{v}$ è un altro vettore del tipo $a\,\boldsymbol{v}$ (cioè ha la stessa direzione di $\boldsymbol{v}$, ma modulo a) con modulo $a = \boldsymbol{x} \cdot \boldsymbol{v} = |\boldsymbol{x}| \cos\vartheta$: il significato geometrico di questa operazione è illustrato nella Figura D.1

Nello spazio $\boldsymbol{R}^p$ le **matrici** $p \times p$ indicate con la notazione $\mathbb{A} = \|a_{k\ell}\|$ permettono di definire delle **trasformazioni** dei vettori $\boldsymbol{x}$ mediante il *prodotto righe per colonne* nel senso che con la notazione $\mathbb{A}\boldsymbol{x}$ indicheremo il vettore con componenti

$$(\mathbb{A}\boldsymbol{x})_k = \sum_{\ell=1}^{p} a_{k\ell} x_\ell$$

Chiameremo **matrice trasposta** la matrice $\mathbb{A}^{\mathrm{T}} = \|a_{\ell k}\|$ che si ottiene da $\mathbb{A} = \|a_{k\ell}\|$ scambiando le righe con le colonne, e diremo che $\mathbb{A}$ è una **matrice simmetrica** quando coincide con la sua trasposta, cioè se $a_{k\ell} = a_{\ell k}$

Data una matrice $\mathbb{A}$ si consideri ora l'**equazione ad autovalori**

$$\mathbb{A}\boldsymbol{x} = \lambda \boldsymbol{x} \tag{D.3}$$

Chiameremo **autovalore** di $\mathbb{A}$ ogni numero λ tale che esista un vettore $\boldsymbol{x} \neq \boldsymbol{0}$ soluzione di (D.3); in tal caso $\boldsymbol{x}$ si chiama **autovettore** di $\mathbb{A}$ associato all'autovalore λ. Se $\boldsymbol{x}$ è autovettore di $\mathbb{A}$ associato a λ, si prova facilmente che anche tutti i vettori $a\boldsymbol{x}$, con a numero reale arbitrario, sono autovettori associati allo stesso autovalore λ. In generale una matrice $\mathbb{A}$ in $\boldsymbol{R}^p$ ha p autovalori, anche non tutti distinti, e si dimostra che quando $\mathbb{A}$ è simmetrica tali autovalori sono reali. Converrà in questo caso riordinare gli autovalori in ordine decrescente

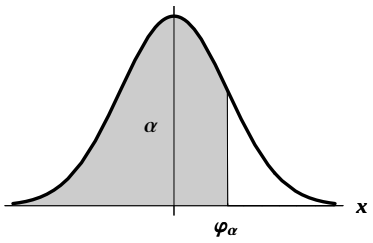
$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$$

indicando con $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_p$ i corrispondenti autovettori che potranno sempre essere scelti in modo da essere **ortonormali**, cioè ortogonali fra loro e tutti di modulo 1 secondo la relazione

$$\boldsymbol{v}_k \cdot \boldsymbol{v}_\ell = \begin{cases} 0 & \text{se } k \neq \ell \\ |\boldsymbol{v}_k|^2 = 1 & \text{se } k = \ell \end{cases}$$

Appendice E

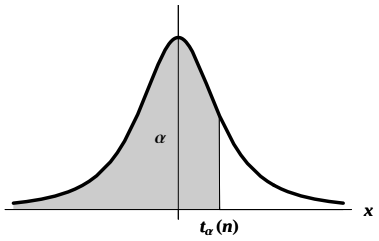
Tavole



E.1 Legge Normale standard $\mathfrak{N}(0,1)$

x	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861

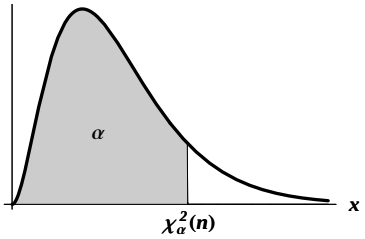
$$\boxed{-\varphi_\alpha = \varphi_{1-\alpha}}$$



E.2 Legge di Student $\mathfrak{T}(n)$

n	**0.950**	**0.975**	**0.990**	**0.995**	**n**	**0.950**	**0.975**	**0.990**	**0.995**
1	6.31375	12.70620	31.82050	63.65670	**31**	1.69552	2.03951	2.45282	2.74404
2	2.91999	4.30265	6.96456	9.92484	**32**	1.69389	2.03693	2.44868	2.73848
3	2.35336	3.18245	4.54070	5.84091	**33**	1.69236	2.03452	2.44479	2.73328
4	2.13185	2.77645	3.74695	4.60409	**34**	1.69092	2.03224	2.44115	2.72839
5	2.01505	2.57058	3.36493	4.03214	**35**	1.68957	2.03011	2.43772	2.72381
6	1.94318	2.44691	3.14267	3.70743	**36**	1.68830	2.02809	2.43449	2.71948
7	1.89458	2.36462	2.99795	3.49948	**37**	1.68709	2.02619	2.43145	2.71541
8	1.85955	2.30600	2.89646	3.35539	**38**	1.68595	2.02439	2.42857	2.71156
9	1.83311	2.26216	2.82144	3.24984	**39**	1.68488	2.02269	2.42584	2.70791
10	1.81246	2.22814	2.76377	3.16927	**40**	1.68385	2.02108	2.42326	2.70446
11	1.79588	2.20099	2.71808	3.10581	**41**	1.68288	2.01954	2.42080	2.70118
12	1.78229	2.17881	2.68100	3.05454	**42**	1.68195	2.01808	2.41847	2.69807
13	1.77093	2.16037	2.65031	3.01228	**43**	1.68107	2.01669	2.41625	2.69510
14	1.76131	2.14479	2.62449	2.97684	**44**	1.68023	2.01537	2.41413	2.69228
15	1.75305	2.13145	2.60248	2.94671	**45**	1.67943	2.01410	2.41212	2.68959
16	1.74588	2.11991	2.58349	2.92078	**46**	1.67866	2.01290	2.41019	2.68701
17	1.73961	2.10982	2.56693	2.89823	**47**	1.67793	2.01174	2.40835	2.68456
18	1.73406	2.10092	2.55238	2.87844	**48**	1.67722	2.01063	2.40658	2.68220
19	1.72913	2.09302	2.53948	2.86093	**49**	1.67655	2.00958	2.40489	2.67995
20	1.72472	2.08596	2.52798	2.84534	**50**	1.67591	2.00856	2.40327	2.67779
21	1.72074	2.07961	2.51765	2.83136	**55**	1.67303	2.00404	2.39608	2.66822
22	1.71714	2.07387	2.50832	2.81876	**60**	1.67065	2.00030	2.39012	2.66028
23	1.71387	2.06866	2.49987	2.80734	**65**	1.66864	1.99714	2.38510	2.65360
24	1.71088	2.06390	2.49216	2.79694	**70**	1.66691	1.99444	2.38081	2.64790
25	1.70814	2.05954	2.48511	2.78744	**75**	1.66543	1.99210	2.37710	2.64298
26	1.70562	2.05553	2.47863	2.77871	**80**	1.66412	1.99006	2.37387	2.63869
27	1.70329	2.05183	2.47266	2.77068	**90**	1.66196	1.98667	2.36850	2.63157
28	1.70113	2.04841	2.46714	2.76326	**100**	1.66023	1.98397	2.36422	2.62589
29	1.69913	2.04523	2.46202	2.75639	**110**	1.65882	1.98177	2.36073	2.62126
30	1.69726	2.04227	2.45726	2.75000	**120**	1.65765	1.97993	2.35782	2.61742

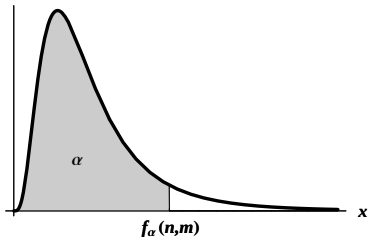
$$\boxed{-t_\alpha(n) = t_{1-\alpha}(n)}$$



E.3 Legge del *chi-quadro* $\boldsymbol{\chi^2}(n)$

n	0.005	0.010	0.025	0.050	0.950	0.975	0.990	0.995
1	0.00004	0.00016	0.00098	0.00393	3.84146	5.02389	6.63490	7.87944
2	0.01003	0.02010	0.05064	0.10259	5.99146	7.37776	9.21034	10.59663
3	0.07172	0.11483	0.21580	0.35185	7.81473	9.34840	11.34487	12.83816
4	0.20699	0.29711	0.48442	0.71072	9.48773	11.14329	13.27670	14.86026
5	0.41174	0.55430	0.83121	1.14548	11.07050	12.83250	15.08627	16.74960
6	0.67573	0.87209	1.23734	1.63538	12.59159	14.44938	16.81189	18.54758
7	0.98926	1.23904	1.68987	2.16735	14.06714	16.01276	18.47531	20.27774
8	1.34441	1.64650	2.17973	2.73264	15.50731	17.53455	20.09024	21.95495
9	1.73493	2.08790	2.70039	3.32511	16.91898	19.02277	21.66599	23.58935
10	2.15586	2.55821	3.24697	3.94030	18.30704	20.48318	23.20925	25.18818
11	2.60322	3.05348	3.81575	4.57481	19.67514	21.92005	24.72497	26.75685
12	3.07382	3.57057	4.40379	5.22603	21.02607	23.33666	26.21697	28.29952
13	3.56503	4.10692	5.00875	5.89186	22.36203	24.73560	27.68825	29.81947
14	4.07467	4.66043	5.62873	6.57063	23.68479	26.11895	29.14124	31.31935
15	4.60092	5.22935	6.26214	7.26094	24.99579	27.48839	30.57791	32.80132
16	5.14221	5.81221	6.90766	7.96165	26.29623	28.84535	31.99993	34.26719
17	5.69722	6.40776	7.56419	8.67176	27.58711	30.19101	33.40866	35.71847
18	6.26480	7.01491	8.23075	9.39046	28.86930	31.52638	34.80531	37.15645
19	6.84397	7.63273	8.90652	10.11701	30.14353	32.85233	36.19087	38.58226
20	7.43384	8.26040	9.59078	10.85081	31.41043	34.16961	37.56623	39.99685
21	8.03365	8.89720	10.28290	11.59131	32.67057	35.47888	38.93217	41.40106
22	8.64272	9.54249	10.98232	12.33801	33.92444	36.78071	40.28936	42.79565
23	9.26042	10.19572	11.68855	13.09051	35.17246	38.07563	41.63840	44.18128
24	9.88623	10.85636	12.40115	13.84843	36.41503	39.36408	42.97982	45.55851
25	10.51965	11.52398	13.11972	14.61141	37.65248	40.64647	44.31410	46.92789
26	11.16024	12.19815	13.84390	15.37916	38.88514	41.92317	45.64168	48.28988
27	11.80759	12.87850	14.57338	16.15140	40.11327	43.19451	46.96294	49.64492
28	12.46134	13.56471	15.30786	16.92788	41.33714	44.46079	48.27824	50.99338
29	13.12115	14.25645	16.04707	17.70837	42.55697	45.72229	49.58788	52.33562
30	13.78672	14.95346	16.79077	18.49266	43.77297	46.97924	50.89218	53.67196
31	14.45777	15.65546	17.53874	19.28057	44.98534	48.23189	52.19139	55.00270
32	15.13403	16.36222	18.29076	20.07191	46.19426	49.48044	53.48577	56.32811
33	15.81527	17.07351	19.04666	20.86653	47.39988	50.72508	54.77554	57.64845
34	16.50127	17.78915	19.80625	21.66428	48.60237	51.96600	56.06091	58.96393
35	17.19182	18.50893	20.56938	22.46502	49.80185	53.20335	57.34207	60.27477

$$\chi^2_\alpha(n) \simeq \frac{1}{2}(\varphi_\alpha + \sqrt{2n-1})^2 \qquad n > 35$$



E.4 Legge di Fisher $\mathfrak{F}(n,m)$

$\alpha = 0.950$

	1^o	1	2	3	4	5	6	7	8	9	10	15	20	30	60	∞
2^o																
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	8.62	8.57	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	5.75	5.69	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	4.50	4.43	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	3.81	3.74	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	3.38	3.30	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	3.08	3.01	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	2.86	2.79	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	2.70	2.62	2.54
11		4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.57	2.49	2.40
12		4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.47	2.38	2.30
13		4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.38	2.30	2.21
14		4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.31	2.22	2.13
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	2.25	2.16	2.07
16		4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.19	2.11	2.01
17		4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.15	2.06	1.96
18		4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.11	2.02	1.92
19		4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.07	1.98	1.88
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	2.04	1.95	1.84
21		4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18	2.10	2.01	1.92	1.81
22		4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	1.98	1.89	1.78
23		4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13	2.05	1.96	1.86	1.76
24		4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11	2.03	1.94	1.84	1.73
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	1.92	1.82	1.71
26		4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22	2.07	1.99	1.90	1.80	1.69
27		4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20	2.06	1.97	1.88	1.79	1.67
28		4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19	2.04	1.96	1.87	1.77	1.65
29		4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18	2.03	1.94	1.85	1.75	1.64
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	1.84	1.74	1.62
31		4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15	2.00	1.92	1.83	1.73	1.61
32		4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14	1.99	1.91	1.82	1.71	1.59
33		4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13	1.98	1.90	1.81	1.70	1.58
34		4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12	1.97	1.89	1.80	1.69	1.57
35		4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11	1.96	1.88	1.79	1.68	1.56
36		4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15	2.11	1.95	1.87	1.78	1.67	1.55
37		4.11	3.25	2.86	2.63	2.47	2.36	2.27	2.20	2.14	2.10	1.95	1.86	1.77	1.66	1.54
38		4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14	2.09	1.94	1.85	1.76	1.65	1.53
39		4.09	3.24	2.85	2.61	2.46	2.34	2.26	2.19	2.13	2.08	1.93	1.85	1.75	1.65	1.52
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	1.74	1.64	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	1.65	1.53	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	1.55	1.43	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	1.46	1.32	1.00

$$\boxed{f_\alpha(n,m) = \frac{1}{f_{1-\alpha}(m,n)}}$$

$\alpha = 0.975$

	1^o	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**15**	**20**	**30**	**60**	∞
2^o																
4		12.22	10.65	9.98	9.60	9.36	9.20	9.07	8.98	8.90	8.84	8.66	8.56	8.46	8.36	8.26
5		10.01	8.43	7.76	7.39	7.15	6.98	6.85	6.76	6.68	6.62	6.43	6.33	6.23	6.12	6.02
6		8.81	7.26	6.60	6.23	5.99	5.82	5.70	5.60	5.52	5.46	5.27	5.17	5.07	4.96	4.85
7		8.07	6.54	5.89	5.52	5.29	5.12	4.99	4.90	4.82	4.76	4.57	4.47	4.36	4.25	4.14
8		7.57	6.06	5.42	5.05	4.82	4.65	4.53	4.43	4.36	4.30	4.10	4.00	3.89	3.78	3.67
9		7.21	5.71	5.08	4.72	4.48	4.32	4.20	4.10	4.03	3.96	3.77	3.67	3.56	3.45	3.33
10		6.94	5.46	4.83	4.47	4.24	4.07	3.95	3.85	3.78	3.72	3.52	3.42	3.31	3.20	3.08
11		6.72	5.26	4.63	4.28	4.04	3.88	3.76	3.66	3.59	3.53	3.33	3.23	3.12	3.00	2.88
12		6.55	5.10	4.47	4.12	3.89	3.73	3.61	3.51	3.44	3.37	3.18	3.07	2.96	2.85	2.72
13		6.41	4.97	4.35	4.00	3.77	3.60	3.48	3.39	3.31	3.25	3.05	2.95	2.84	2.72	2.60
14		6.30	4.86	4.24	3.89	3.66	3.50	3.38	3.29	3.21	3.15	2.95	2.84	2.73	2.61	2.49
15		6.20	4.77	4.15	3.80	3.58	3.41	3.29	3.20	3.12	3.06	2.86	2.76	2.64	2.52	2.40
16		6.12	4.69	4.08	3.73	3.50	3.34	3.22	3.12	3.05	2.99	2.79	2.68	2.57	2.45	2.32
17		6.04	4.62	4.01	3.66	3.44	3.28	3.16	3.06	2.98	2.92	2.72	2.62	2.50	2.38	2.25
18		5.98	4.56	3.95	3.61	3.38	3.22	3.10	3.01	2.93	2.87	2.67	2.56	2.44	2.32	2.19
19		5.92	4.51	3.90	3.56	3.33	3.17	3.05	2.96	2.88	2.82	2.62	2.51	2.39	2.27	2.13
20		5.87	4.46	3.86	3.51	3.29	3.13	3.01	2.91	2.84	2.77	2.57	2.46	2.35	2.22	2.09
21		5.83	4.42	3.82	3.48	3.25	3.09	2.97	2.87	2.80	2.73	2.53	2.42	2.31	2.18	2.04
22		5.79	4.38	3.78	3.44	3.22	3.05	2.93	2.84	2.76	2.70	2.50	2.39	2.27	2.14	2.00
23		5.75	4.35	3.75	3.41	3.18	3.02	2.90	2.81	2.73	2.67	2.47	2.36	2.24	2.11	1.97
24		5.72	4.32	3.72	3.38	3.15	2.99	2.87	2.78	2.70	2.64	2.44	2.33	2.21	2.08	1.94
25		5.69	4.29	3.69	3.35	3.13	2.97	2.85	2.75	2.68	2.61	2.41	2.30	2.18	2.05	1.91
26		5.66	4.27	3.67	3.33	3.10	2.94	2.82	2.73	2.65	2.59	2.39	2.28	2.16	2.03	1.88
27		5.63	4.24	3.65	3.31	3.08	2.92	2.80	2.71	2.63	2.57	2.36	2.25	2.13	2.00	1.85
28		5.61	4.22	3.63	3.29	3.06	2.90	2.78	2.69	2.61	2.55	2.34	2.23	2.11	1.98	1.83
29		5.59	4.20	3.61	3.27	3.04	2.88	2.76	2.67	2.59	2.53	2.32	2.21	2.09	1.96	1.81
30		5.57	4.18	3.59	3.25	3.03	2.87	2.75	2.65	2.57	2.51	2.31	2.20	2.07	1.94	1.79
31		5.55	4.16	3.57	3.23	3.01	2.85	2.73	2.64	2.56	2.50	2.29	2.18	2.06	1.92	1.77
32		5.53	4.15	3.56	3.22	3.00	2.84	2.71	2.62	2.54	2.48	2.28	2.16	2.04	1.91	1.75
33		5.51	4.13	3.54	3.20	2.98	2.82	2.70	2.61	2.53	2.47	2.26	2.15	2.03	1.89	1.73
34		5.50	4.12	3.53	3.19	2.97	2.81	2.69	2.59	2.52	2.45	2.25	2.13	2.01	1.88	1.72
35		5.48	4.11	3.52	3.18	2.96	2.80	2.68	2.58	2.50	2.44	2.23	2.12	2.00	1.86	1.70
36		5.47	4.09	3.50	3.17	2.94	2.78	2.66	2.57	2.49	2.43	2.22	2.11	1.99	1.85	1.69
37		5.46	4.08	3.49	3.16	2.93	2.77	2.65	2.56	2.48	2.42	2.21	2.10	1.97	1.84	1.67
38		5.45	4.07	3.48	3.15	2.92	2.76	2.64	2.55	2.47	2.41	2.20	2.09	1.96	1.82	1.66
39		5.43	4.06	3.47	3.14	2.91	2.75	2.63	2.54	2.46	2.40	2.19	2.08	1.95	1.81	1.65
40		5.42	4.05	3.46	3.13	2.90	2.74	2.62	2.53	2.45	2.39	2.18	2.07	1.94	1.80	1.64
60		5.29	3.93	3.34	3.01	2.79	2.63	2.51	2.41	2.33	2.27	2.06	1.94	1.82	1.67	1.48
120		5.15	3.80	3.23	2.89	2.67	2.52	2.39	2.30	2.22	2.16	1.94	1.82	1.69	1.53	1.31
∞		5.02	3.69	3.12	2.79	2.57	2.41	2.29	2.19	2.11	2.05	1.83	1.71	1.57	1.39	1.00

E.5 Valori di $e^{-\lambda}$

λ	**0.00**	**0.01**	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
0.0	1.0000	0.9900	0.9802	0.9704	0.9608	0.9512	0.9418	0.9324	0.9231	0.9139
0.1	0.9048	0.8958	0.8869	0.8781	0.8694	0.8607	0.8521	0.8437	0.8353	0.8270
0.2	0.8187	0.8106	0.8025	0.7945	0.7866	0.7788	0.7711	0.7634	0.7558	0.7483
0.3	0.7408	0.7334	0.7261	0.7189	0.7118	0.7047	0.6977	0.6907	0.6839	0.6771
0.4	0.6703	0.6637	0.6570	0.6505	0.6440	0.6376	0.6313	0.6250	0.6188	0.6126
0.5	0.6065	0.6005	0.5945	0.5886	0.5827	0.5769	0.5712	0.5655	0.5599	0.5543
0.6	0.5488	0.5434	0.5379	0.5326	0.5273	0.5220	0.5169	0.5117	0.5066	0.5016
0.7	0.4966	0.4916	0.4868	0.4819	0.4771	0.4724	0.4677	0.4630	0.4584	0.4538
0.8	0.4493	0.4449	0.4404	0.4360	0.4317	0.4274	0.4232	0.4190	0.4148	0.4107
0.9	0.4066	0.4025	0.3985	0.3946	0.3906	0.3867	0.3829	0.3791	0.3753	0.3716
1.0	0.367879									
2.0	0.135335									
3.0	0.049787									
4.0	0.018316									
5.0	0.006738									
6.0	0.002479									
7.0	0.000912									
8.0	0.000335									
9.0	0.000123									
10.0	0.000045									

$$\lambda = \lfloor\lambda\rfloor + r \qquad \lfloor\lambda\rfloor = 0, 1, 2, \ldots \qquad 0 < r < 1$$

$$e^{-\lambda} = e^{-\lfloor\lambda\rfloor} e^{-r}$$

Indice analitico