Università degli Studi di Torino

A. Durio, E. D. Isaia

Elementi di Statistica Descrittiva per l'Analisi dei Dati

a.a 2009/10

Dipartimento di Statistica e Matematica Applicata "Diego de Castro"

È senza dubbio inevitabile che questi appunti presentino errori materiali; inoltre, seppur gli autori si siano sforzati di conciliare il rigore con la chiarezza espositiva, alcune parti del testo possono risultare poco comprensibili.
Saremo grati a tutti coloro, e specialmente agli Studenti, che vorranno segnalarci qualunque problema, dai più banali errori tipografici alle oscurità nell'esposizione.

Avvertenza

DIPARTIMENTO DI STATISTICA E MATEMATICA APPLICATA "DIEGO DE CASTRO"
Corso Unione Sovietica, 218/bis
10134, Torino (Italy)

Indice

CAPITOLO 1

INTRODUZIONE ALLA STATISTICA DESCRITTIVA

In questo capitolo, a parte un breve cenno storico, verrà data una definizione di statistica descrittiva e del suo oggetto di indagine, cioè i fenomeni collettivi. Verrà, infine, proposta una breve bibliografia.

1.1. GENESI DELLA STATISTICA

Da sempre l'essere umano si è caratterizzato per la sua attitudine a raccogliere informazioni su se stesso e sull'ambiente nel quale egli è vissuto, al duplice fine di prendere coscienza di sé e di aumentare le possibilità di sopravvivenza della propria specie. Le prime iniziative di raccolta finalizzata di dati, si ebbero non appena la vita sociale dei primi uomini si organizzò in forma più evoluta, dando luogo a popolazioni, dotate di regole di comportamento e di autorità incaricate di farle rispettare.
Ad esempio uno dei primi libri della Bibbia, I Numeri, ha tale titolo perché comincia con l'enumerazione del popolo. Nelle prime pagine viene descritto un rudimentale censimento delle tribù di Israele: numero dei maschi secondo le stirpi e le case.
Altri censimenti di popolazione, a fini militari e fiscali, furono attuati dai Romani (Censimento di Servio Tullio intorno alla metà dell'VIII a.C., Censimento di Erode in occasione della nascita di Cristo, ...), tanto per non citare più remote esperienze dei popoli egiziani e cinesi.
Con il passare del tempo l'osservazione finalistica e volontaria dei fenomeni, la loro rappresentazione simbolica e la registrazione delle informazioni acquisite si estese allo studio dei fenomeni empirici di pertinenza delle scienze naturali e sociali, ad esempio: alle piene di fiumi, alle maree, ai moti degli astri, alle precipitazioni piovose, alla produzione agricola, alla durata di vita delle persone, alla composizione, per sesso e per età, della popolazione, alla variabilità territoriale o temporale dei prezzi, alla produzione industriale, alla distribuzione del reddito delle famiglie,...

La Statistica attiene appunto alla raccolta ed alla analisi dei dati per studiare i *fenomeni collettivi*, ovvero quei fenomeni che possono essere percepiti, nella loro interezza, solo mediante numerose osservazioni.
Le applicazioni della Statistica rivolte a quello speciale fenomeno collettivo costituito da popolazioni umane stanziate in un certo territorio, furono talmente diffuse e generalizzate, che ancora oggi l'insieme delle unità portatrici delle informazioni elementari, costituenti il generico fenomeno collettivo, le cosiddette "unità statistiche", sono indicate con i sinonimi di *collettivo statistico* e *popolazione*.
Ad ogni buon conto la paternità del termine Statistica può essere fatta risalire all'italiano Ghislini, che nel 1589 definisce la Statistica come la *descrizione delle qualità che caratterizzano e degli elementi che compongono uno Stato*, al tedesco Conring che nel 1660 tenne presso l'Università di Helmstad un corso chiamato Staatskunde avente per oggetto la descrizione sistematica delle cose più notevoli per la vita degli stati, mentre il merito di aver dato origine alla più numerosa collezione di dati statistici della storia passata può essere attribuito al Concilio di Trento, che impose ai parroci la regolare trascrizione dei matrimoni, dei battesimi nonché dei morti.

1.2. I FENOMENI STATISTICI

La Statistica entra in gioco o almeno dovrebbe essere tenuta presente, non appena ci si accinge ad attuare una raccolta finalizzata di informazioni su di un prefissato fenomeno collettivo; in genere essa diventa tanto più indispensabile quanto più sono ingenti le risorse necessarie alla raccolta in oggetto.
A titolo esemplificativo proviamo ad elencare alcuni problemi reali, esposti talvolta in forma interrogativa, che, coinvolgendo lo studio di fenomeni di massa, richiedono l'uso della moderna metodologia statistica:

- *quali sono stati gli effetti, a livello provinciale, dei recenti provvedimenti del governo a favore della occupazione giovanile*?
- *è possibile evidenziare l'incidenza del fumo sulla salute degli italiani? In particolare, esiste una dipendenza tra fumo e cancro polmonare?*
- *la direzione commerciale di una azienda deve scegliere tra due diverse campagne pubblicitarie che hanno il comune obiettivo di un aumento del ricordo della marca presso i consumatori.*
- *il responsabile di produzione deve assicurarsi che le dimensioni geometriche dei pezzi prodotti in grande serie soddisfino le specifiche di disegno.*

– *un revisore aziendale deve emettere un giudizio professionale circa la correttezza o meno della voce contabile "scorte di magazzino composto da circa 15000 voci.*

L'azienda stessa da luogo a fenomeni collettivi, essa è infatti costituita da uomini, strutture organizzative e tecniche e crea beni e servizi, per cui molti aspetti aziendali possono essere percepiti solo mediante numerose osservazioni. Inoltre ogni ramo d'industria può essere costituito da una miriade di imprese.
La Statistica assume quindi un ruolo importante all'interno dell'impresa; essa infatti è *strumento di sintesi* in quanto fornisce fotografie dei vari aspetti aziendali, è *strumento di analisi*, in quanto consente il confronto tra diverse realtà aziendali o tra diverse funzioni aziendali evidenziando le eventuali differenze.

1.3. ANALISI STATISTICA DEI DATI

Se pensiamo che lo scopo della Statistica sia lo studio quantitativo e sistematico delle informazioni concernenti un fenomeno collettivo di interesse, allora possiamo pensare alla Statistica come ad una disciplina che ha a che fare con la rilevazione e la rappresentazione sintetica di insiemi di dati, donde il termine di *analisi statistica dei dati* o *statistica descrittiva.*
In tale ottica, dunque, l'osservazione si concentra sul caso individuale con cui un certo fenomeno collettivo si manifesta e pertanto l'insieme dei casi individuali viene a costituire la popolazione statistica di riferimento; i residenti in un certo paese nel caso di censimento demografico, i prezzi dei singoli beni offerti sul mercato nelle indagini sul costo della vita, gli individui di cui si desidera rilevare l'orientamento circa una data questione nel caso di indagini di opinione, ...
L'analisi statistica dei dati fornisce una sintesi quantitativa dei fenomeni oggetto di studio rendendone possibile, ricorrendo per lo più a strumenti matematici che spesso si risolvono in semplici indici di sintesi, una visione semplificata.
I metodi della statistica descrittiva, applicati ai diversi caratteri mediante i quali il fenomeno in esame si manifesta, consentono di evidenziare eventuali tendenze di fondo o regolarità nelle loro manifestazioni e mettere in luce i loro eventuali legami di dipendenza.
Scopo dei capitoli che seguono è pertanto quello di offrire al Lettore una, seppur breve e limitata, panoramica dei principali strumenti che consentono un primo approccio all'analisi statistica dei dati prendendo le mosse dallo studio di un solo carattere, sia esso qualitativo o quantitativo, per giungere infine all'analisi congiunta di due caratteri rilevati su una medesima popolazione, introducendo così il Lettore a quella che viene abitualmente detta *analisi statistica dei dati multivariati*, per la quale rimandiamo ai testi specifici citati al paragrafo 1.5.

1.4. PRINCIPALI FONTI STATISTICHE

E' quasi ovvio segnalare che in campo aziendale la raccolta dei dati è in genere finalizzata alla risoluzione di qualche specifico problema, ad esempio di produzione, di vendita, di marketing, di miglioramento della qualità, di monitoraggio degli infortuni, o più in generale all'aumento delle conoscenze su prodotti, clienti, concorrenza, ...
A tal fine si cerca di far ricorso, quando ciò è possibile, a informazioni già raccolte da enti istituzionali, associazioni di categoria, aziende specializzate e così via.
In Italia primo fra tutti per la quantità di statistiche fornite è l'ISTAT, acronimo di Istituto Centrale di Statistica, che, costituito nel 1926 ed alle dirette dipendenze del Consiglio dei Ministri, provvede a raccogliere direttamente o a coordinare la raccolta di dati da parte di altri enti, quali i Comuni, le Regioni, le CCIAA, le ASL, ..., su settori quali il *territorio, climatologia e ambiente naturale*, la *popolazione*, *sanità e sicurezza sociale*, l'*istruzione*, le *statistiche sociali e culturali varie*, la *giustizia*, il *lavoro*, i *conti economici nazionali*, l'*agricoltura, foresta e pesca*, l'*industria*, le *costruzioni ed opere pubbliche*, il *commercio interno, turismo e commercio con l'estero*, i *trasporti e comunicazioni*, il *credito*.
Per renderci conto della vastità e complessità di tale raccolta di documentazione statistica, è sufficiente osservare che sotto la voce *istruzione* l'elaborazione ISTAT concerne aspetti relativi a le scuole materne, le scuole elementari, le scuole secondarie superiori, l'età della popolazione scolastica, l'istruzione artistica e musicale, l'istruzione universitaria, i laureati, le statistiche delle scuole para-universitarie, i corsi di formazione professionale, le scuole speciali per minorati fisici, psichici e sensoriali.
Non va comunque dimenticato che, oltre alle precedenti rilevazioni correnti, cioè effettuate con scadenza annuale, l'ISTAT si occupa anche dei *censimenti*; tra questi quelli della *popolazione*, con cadenza decennale dal 1861, degli *esercizi industriali e commerciali*, degli *esercizi e della produzione industriale e del commercio*, dell'*agricoltura*.
Dopo l'ISTAT, in un elenco ideale di fornitori di informazioni statistiche, seguono: enti previdenziali ed assicurativi (INPS, INAM, INPDAP, ISVAP, ...), i Ministeri con competenza su Industria, Lavoro, Sanità, Motorizzazione Civile, ..., Associazioni territoriali dell'Industria, del Commercio, dell'Artigianato, Istituti di ricerche economiche e sociali, Istituti ed Osservatori per lo studio della congiuntura, della occupazione, ...
Sono anche disponibili numerose banche date (CERVED, SARIN, Centrale dei Bilanci, Pagine Gialle, Pagine Utili, NIELSEN, ...) in grado di fornire, a titolo oneroso e per scopi commerciali, fiscali, creditizi, di marketing, ... numerose informazioni su aziende, famiglie, consumatori, ecc. ... italiani.
In altri casi le aziende provvedono, direttamente o tramite società specializzate, quali ad esempio: DOXA, Demoscopea, CIRM, SWG, CEMIT, ASSIRM, ..., a procurarsi le informazioni di cui hanno bisogno, ad esempio il livello di soddisfazione della clientela, le

quote di mercato detenute della concorrenza, le linee di tendenza, le abitudini e le aspirazioni dei potenziali clienti,... A tale fine alle tradizionali indagini mediante interviste dirette, telefoniche e postali si sono recentemente aggiunte indagini condotte con nuovi strumenti di raccolta dati quali CATI, acronimo di Computer Aided Telephone Interview, fax, e-mail e siti Internet.

1.5. NOTA BIBLIOGRAFICA

Nel seguito forniamo al Lettore una breve bibliografia di testi di analisi statistica dei dati, così come intesa nei paragrafi precedenti. Spesso, inoltre, i testi, nonostante il titolo, offrono solo cenni di Statistica descrittiva per dedicarsi in modo più approfondito a quella che comunemente và sotto il nome di Statistica inferenziale. Questo è certamente il caso dei testi di Scuola anglosassone, che con il termine Statistics intendono Statistica inferenziale; la Scuola francese distingue, peraltro, tra Statistique, la Statistica inferenziale, e Analyse des Donnés, la Statistica descrittiva.
Per tali motivi, ci limitiamo a citare alcune pubblicazioni, in lingua italiana, che a nostro giudizio possono essere di aiuto al Lettore per approfondimenti e analisi successive.

- ⋆ sull'analisi statistica dei dati si possono consultare i testi di Frosini (1990), Frosini (1995), Jalla (1989), Jalla (1991), Landenna (1997), Leti (1983), Naddeo (1972), Parpinel and Provasi (2004), Piccolo (2000);
- ⋆ per quanto attiene alla analisi statistica dei dati multivariati il Lettore può trarre spunto dai testi di Fabbris (1997), Mignani and Montanari (2001) nonché di Vitali (1997).

CAPITOLO 2

MUTABILI E VARIABILI STATISTICHE

In questo capitolo verranno date le prime definizioni della statistica descrittiva. Si tratta di concettio semplici e tuttavia non trascurabili. Una attenta lettura dei paragrafi seguenti porterà il Lettore a conoscere i concetti di collettivo e unità statistiche, di carattere e insieme delle modalità, di scale di misura per la classificazione dei caratteri fino a giungere alle definizioni di variabile e mutabile statistica nonché ai rispettivi insiemi dei dati individuali e alla matrice dei dati.

2.1. PRIME DEFINIZIONI

Ogni attività di osservazione e di analisi di un fenomeno che faccia ricorso alla statistica comporta come primo passo la corretta delimitazione nello spazio e nel tempo delle unità portatrici di informazioni circa il fenomeno stesso, un'esaustiva individuazione delle caratteristiche che concorrono a spiegare il fenomeno nonché le espressioni con cui queste si manifestano.
Così ad esempio volendo indagare circa la qualità dell'insegnamento di un corso universitario appena terminato, potremmo pensare di intervistare tutti gli Studenti che nel corrente anno accademico hanno frequentato il corso onde ottenere informazioni su alcune caratteristiche che reputiamo possano concorrere al giudizio complessivo circa il fenomeno di interesse. Tra queste ad esempio la chiarezza espositiva del docente, la sua puntualità alle lezioni, il numero di ore dedicate alle esercitazioni, ecc.
Per formalizzare e al contempo avvicinarci alla corretta terminologia statistica è oportuno introdurre alcune definizioni che sono alla base dei concetti che verranno via via esposti nel seguito.

Definizione 2.1 (Collettivo statistico)
definiamo collettivo statistico o popolazione, l'insieme Ω delle entità mediante le quali è possibile ottenere informazioni sul fenomeno stesso.

□

Per analogia, chiamiamo *unità statistiche* gli elementi ω_α dell'insieme Ω. Indicando con n la numerosità di Ω, possiamo dunque scrivere

$$\Omega = \{\omega_\alpha\}_{\alpha=1,\dots,n} = \{\omega_1, \omega_2, \dots, \omega_n\}$$

Ciascuna unità statistica ω_α è portatrice dell'informazione statistica elementare, che verrà rilevata mediante un'attività di "misurazione".
Va ben tenuto a mente che il collettivo statistico rappresenta il presupposto di qualsiasi futura analisi. Gli strumenti statistici ci consetiranno di descrivere come il fenomeno si manifesta sulle unità della popolazione, pertanto molta attenzione va posta in questa fase iniziale di scelta del collettivo.
Così ad esempio con un'intervista telefonica agli abitanti di un comune per indagare circa la loro soddisfazione in seguito alla costruzione di una pista ciclabile non potremo affermare che i risultati dell'indagine riguardino la soddisfazione dei residenti nel comune, ma piuttosto di tutti gli abbonati telefonici del comune!

Definizione 2.2 (Carattere)
definiamo carattere la grandezza che, rilevata su ciascuna unità statistica, sarà di aiuto nella comprensione del fenomeno collettivo in esame.

□

Un carattere può essere una grandezza di differente natura, ad esempio *fisica* (statura, peso, colore dei capelli di una persona, ...), *economica* (reddito annuo imponibile, prezzo di una merce, ...), *demografica* (numero di matrimoni, numero di nati vivi, ...), *psicologica* (stato d'animo di un elettore, propensione all'acquisto di un consumatore, ...)
Precisato il fenomeno su cui si intende indagare, definito il collettivo di riferimento e scelti i caratteri da rilevare, occorre individuare le diverse possibili manifestazioni o *modalità* dei singoli caratteri in esame. A tal proposito valga la seguente

Definizione 2.3 (Insieme delle modalità di un carattere)
definiamo insieme delle modalità di un carattere l'insieme M di tutte le possibili espressioni con cui il carattere può manifestarsi.

□

▷ ESEMPIO 2.1

Volendo chiarire i concetti introdotti, immaginiamo di volere indagare circa *la capacità ricettiva degli alberghi di una comunità alpina.* Se tale è il fenomeno collettivo

su cui desideriamo indagare, il collettivo statistico (Ω) viene a coincidere con l'insieme degli esercizi alberghieri situati sul territorio della comunità. Le unità statistiche sono pertanto rappresentate dai singoli esercizi alberghieri.
Ai fini dell'indagine, possiamo pensare di rilevare, su ciasuna unità del collettivo, più caratteri, ad esempio

- ⋆ *il numero di posti letto*, le cui modalità vengono a coincidere con un sottoinsieme degli interi positivi, ad esempio $M = \{10, 11, \ldots, 120\}$;
- ⋆ *la superficie dei locali di ristorazione*, le cui modalità vengono a coincidere con un sottoinsieme dei numeri reali, ad esempio $M = [100; 1000]\ m^2$;
- ⋆ *la categoria alberghiera*, le cui modalità vengono a coincidere con un insieme di attributi, ad esempio $M = \{1°, 2°, 3°, \ldots\}$.

◁

2.2. LE SCALE DI MISURA

Come si è visto una qualsiasi attività statistica ha il suo logico presupposto in una rilevazione di informazioni sulle unità del collettivo associate al fenomeno in studio e questo naturalmente comporta un'attività di osservazione e/o di misurazione.
Per ciascun carattere in esame, è necessario stabilire la scala di misura che si intende adottare per la rilevazione.
A tale proposito scegliamo di adottare la distinzione introdotta dallo psicologo Stevens nel 1946 tra le quattro differenti scale di misura "nominali", "ordinali", "per intervalli" e "per rapporti".

2.2.1 SCALA NOMINALE

Un carattere è misurato su *scala nominale* se le sue modalità si identificano in "attributi" tra i quali è impossibile individuare una relazione d'ordine naturale, cioè prese due qualsiasi modalità non è in alcun modo possibile affermare che l'una precede l'altra.
Sono esempi di caratteri rilevabili mediante scala nominale:

- ⋆ il *sesso*, con insieme delle modalità $\{$Maschio, Femmina$\}$;
- ⋆ lo *stato civile*, con insieme delle modalità $\{$Celibe/Nubile, Coniugato/a, ...$\}$.

2.2.2 SCALA ORDINALE

Un carattere è misurato su *scala ordinale* o *per ranghi* se le sue modalità si identificano in "attributi" che presentano una relazione d'ordine naturale. Per cui prese due qualsiasi modalità è sempre possibile affermare che l'una precede l'altra.
Sono esempi di caratteri rilevabili su scala ordinale:

- ⋆ il *titolo di studio*, con insieme delle modalità $\{$Elementare, Media Inferiore, Media Superiore, Laurea, $\dots\}$;
- ⋆ l'*ordine di arrivo ad un GP di F1*, con insieme delle modalità $\{1°, 2°, 3°, \dots, 6°\}$.

Va osservato che, ricorrendo ad una scala ordinale, nulla viene precisato circa la *distanza* esistente fra i diversi attributi. Ad esempio, con riferimento al carattere *ordine di arrivo ad un GP di F1*, non abbiamo alcuna informazione circa il "distacco", ad esempio, tra il $1°$ ed il $2°$ pilota classificato, tra il $2°$ ed il $3°$ e così via.

2.2.3 SCALA PER INTERVALLI

Un carattere è misurato su *scala per intervalli* se le sue modalità si identificano con numeri nell'ambito di un sistema di riferimento dotato di origine arbitraria. Caratteristiche, dunque, di un carattere misurato su scala per intervalli sono:

- ⋆ prese due qualsiasi sue modalità è sempre possibile affermare che l'una precede l'altra;
- ⋆ la distanza intercorrente tra due sue modalità non dipende dall'origine adottata per il sistema di riferimento;
- ⋆ alla modalità nulla non necessariamente corrisponde assenza del carattere.

Esempio di scala per intervalli è la *quota altimetrica.* Si immagini che un pilota di un aereo di linea, all'inizio della manovra di atterraggio all'aereoporto di Torino, comunichi ai passeggeri che si trovano in quell'istante a 750 metri sul livello del mare e che nello stesso istante su un secondo aereo, diretto a Milano e sulla stessa verticale del primo, il pilota stia comunicando ai passeggeri di trovarsi a 1250 metri sul livello del mare. Ovviamente i due aerei distano tra loro di 500 metri. L'origine di riferimento che interessa ai piloti non è certamente quella del livello del mare, bensì quella del suolo sottostante, che sarà 250 metri (altezza sul livello del mare dell'aereoporto di Torino). Rispetto a tale origine i due aerei si trovano ad una altezza rispettivamente di 500 e 1000 metri, e tra loro la distanza continua ad essere di 500 metri. Qualunque sia l'origine del sistema di

riferimento adottata per la quota altimetrica la distanza tra i due aereomobili è sempre la stessa. Ugualmente non può dirsi per il rapporto tra le due altezze; se l'origine è quella del suolo potremmo affermare che il secondo aereo si trova ad un'altezza doppia rispetto a quella del primo ($1000/500 = 2$), mentre scegliendo quale origine il livello del mare il precedente rapporto sarebbe solo 1.67 ($= 1250/750$).
Inoltre, una volta atterrato il primo aereo si trova ad un'altezza di zero metri rispetto al sistema di riferimento del pilota, ciò non implica assenza di carattere, infatti il pilota, mutando sistema di riferimento può comunicare ai passeggeri di trovarsi a 250 metri sul livello del mare.
Sono altri esempi classici di scala per intervalli la *temperatura* oppure l'*anno di nascita.*

2.2.4 Scala per rapporti

Un carattere è misurato su *scala per rapporti* se le sue modalità si identificano con numeri di un sistema di riferimento dotato di origine assoluta, valore quest'ultimo cui è associata l'assenza di carattere. Tale scala possiede dunque tutte le caratteristiche di una scala per intervalli ed in più il rapporto fra due modalità qualsiasi è indice di proporzionalità tra le stesse. La maggior parte dei caratteri di tipo quantitativo vengono misurati su scala per rapporti; ne sono esempi:

- ⋆ il *peso alla nascita dei neonati*, con insieme delle modalità l'intervallo $[1.5; 5]$ kg;
- ⋆ il *numero di figli delle famiglie italiane*, con insieme delle modalità $\{0, 1, \ldots, 20\}$.

A commento consideriamo il carattere *reddito mensile netto* in euro: un soggetto con reddito mensile di 1000 euro ovviamente percepisce un reddito doppio di un soggetto che ne guadagna 500. Essendo il carattere misurato su scala per rapporti, il rapporto tra i loro due redditi ha senso è può ben essere interpretato. Valuteremmo certamente nullo il reddito di un terzo soggetto che percepisse zero euro al mese!

2.3. Classificazione dei caratteri

I caratteri statistici possono essere suddivisi nelle due categorie: *caratteri qualitativi* e *caratteri quantitativi.* In particolare

Definizione 2.4 (Carattere qualitativo)
un carattere viene detto qualitativo o mutabile se le sue modalità sono espresse in termini di attributi.

□

Definizione 2.5 (Carattere quantitativo)
un carattere viene detto quantitativo o variabile se le sue modalità sono espresse in termini numerici.

□

Sono pertanto caratteri qualitativi quelli misurati su una scala nominale ovvero ordinale, mentre sono caratteri quantitativi quelli misurati su scala per intervalli ovvero per rapporti.

Volendo raffinare le definizioni precedenti si potrà distinguere ancora in:

- ⋆ nel caso di *caratteri qualitativi* tra:
  - — *caratteri qualitativi sconnessi*: quando le modalità non presentano alcuna relazione d'ordine naturale, per cui la scala impiegata è nominale.
  - — *caratteri qualitativi ordinali*: quando le modalità presentano una relazione d'ordine naturale, per cui la scala impiegata è quella ordinale.
- ⋆ nel caso di *caratteri quantitativi* tra:
  - — *caratteri quantitativi discreti*: quando fissata una modalità esiste un intervallo all'interno del quale nessun altro valore costituisce una modalità. L'insieme M delle modalità è pertanto un insieme finito o al più infinito numerabile. Generalmente caratteri quantitativi discreti derivano da operazioni di conteggio.
  - — *caratteri quantitativi continui*: qualora comunque scelte due possibili modalità tra loro esistono infiniti valori che sono altrettante modalità. L'insieme M delle modalità è pertanto un insieme con la potenza del continuo. Generalmente caratteri quantitativi continui derivano da operazioni di misurazione.

OSSERVAZIONE: per i caratteri continui è bene insistere sul fatto che le possibili modalità possano coincidere con un qualsiasi punto all'interno di un segmento reale nonostante in realtà i processi di misurazione diano come risultato solo l'appartenenza o meno del vero valore ad un certo intervallo reale.
Così ad esempio rilevare che la lunghezza di un componente meccanico è di 7.0 cm avendo usato uno strumento di misura con la precisione del centimetro ci informa cha la lunghezza effettiva appartiene all'intervallo $[6.5,\ 7.5[$ cm. Avessimo utilizzato uno strumento di misura con la precisione del millimetro la precedente misurazione ci direbbe cha la lunghezza effettiva appartiene all'intervallo $[6.95,\ 7.05[$ cm.
In altri termini nella realtà, e ciò indipendentemente dalla precisione dello strumento di misura a cui si fa ricorso, si introduce di fatto una *discretizzazione dei caratteri continui.*

⋆

▷ ESEMPIO 2.2

Le prestazioni di un Personal Computer da tavolo (il fenomeno oggetto di studio) possono essere descritte mediante alcuni caratteri quali ad esempio il *tipo di processore*, il *tipo di unità ottica*, la *capacità del disco fisso*, nonché la *velocità di clock* Tali caratteri verranno rilevati su un collettivo statistico costituito da un certo numero di modelli presenti in un dato istante sul mercato.
Manifestamente i primi due caratteri sono di tipo qualitativo e per essi vengono adottate scale di misura rispettivamente nominale e ordinale; infatti i loro insiemi delle modalità sono:

$$M = \{\text{Pentium IV, Celeron, Sempron, Athlon, Opteron}\} = \{\text{P, C, S, A, O}\}$$
$$M = \{\text{CD-R, CD-RW}\} = \{\text{R, RW}\}$$

Quanto ai restanti due caratteri, di tipo quantitativo discreto il primo, continuo il secondo, verranno adottate scale di misura per rapporti. I loro insiemi delle modalità corrispondono ad esempio ai seguenti insiemi numerici:

$$M = \{20, 30, 40, 50, 60, 70, 80\}\text{Gb}\, M = \qquad [0.80; 2.00]\text{Ghz} \tag{2.1}$$

◁

▷ ESEMPIO 2.3

Si immagini di disporre di una partita di rondelle in ferro in uscita da un processo produttivo e che interessi indagare circa il loro diametro interno, che supponiamo costituisca un parametro di idoneità all'uso o di qualità delle stesse. Ai fini della rilevazione del carattere è del tutto naturale ricorrere ad una scala per rapporti e misurare, pertanto, il diametro ad esempio in millimetri.
Peraltro, disponendo delle singole misurazioni, è possibile interpretare il carattere mediante una scala nominale, cioè qualitativa, scegliendo, ad esempio, di classificare una generica rondella come:

$$\begin{cases} \text{rettificabile} & \text{se il diametro è minore di } 10 \text{ mm} \\ \text{idonea} & \text{se il diametro è compreso tra } 10 \text{ e } 12 \text{ mm} \\ \text{non idonea} & \text{se il diametro è maggiore di } 12 \text{ mm} \end{cases}$$

Si osservi che si potrebbe giungere a una classificazione su scala nominale, come quella testé proposta, disponendo di un semplice strumento del tipo passa-non passa. È evidente che è la scelta della scala di misura a consentire la classificazione del carattere in esame.

◁

2.4. MUTABILI E VARIABILI STATISTICHE

Definito il collettivo statistico Ω di riferimento, nonché le scale di misura che si intendono adottare per i diversi caratteri di interesse, per ciascuno di questi si definisce l'insieme M delle modalità. Evidentemente, nel caso di caratteri variabili l'insieme M sarà costituito da elementi di $\mathbb{R}$ (ed in simboli potremmo scrivere $M \subseteq \mathbb{R}$), mentre nel caso di caratteri mutabili esso risulterà sempre un insieme finito i cui elementi risultano essere degli attributi.
Ciò premesso, si procede alla rilevazione dei caratteri in esame su ciascuna unità statistica; *tale operazione individua una corrispondenza (ovvero una applicazione) tra collettivo statistico Ω e insieme M delle modalità del carattere in esame* (vedasi figura 2.1).

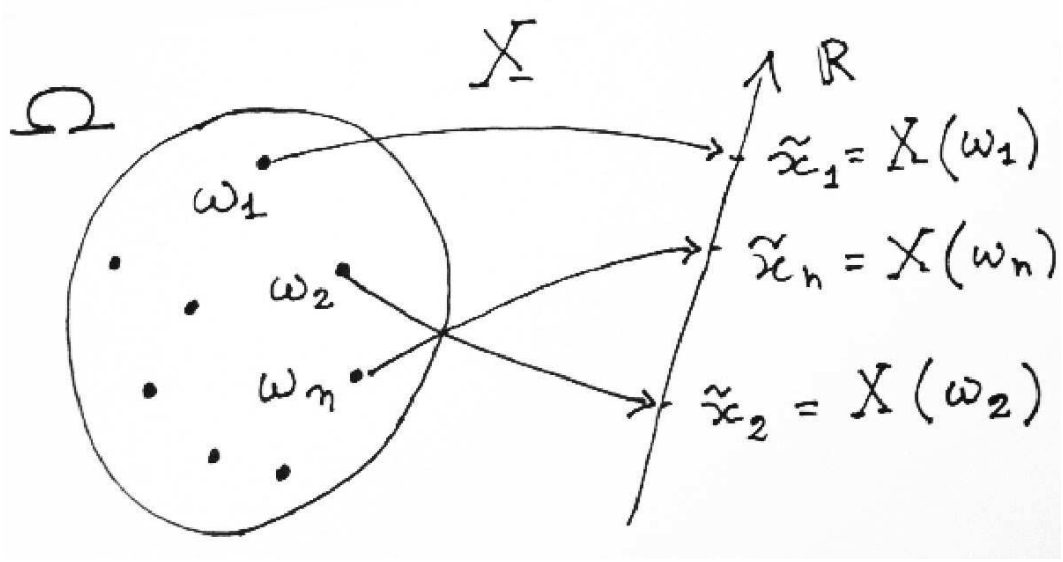


Figura 2.1 La variabile statistica $X(\cdot)$

Pertanto, valga la seguente

Definizione 2.6 (Mutabile e variabile statistica)
l'applicazione che associa a ciascun elemento del collettivo statistico Ω uno ed uno solo elemento dell'insieme M delle modalità del carattere in esame viene detta mutabile statistica (m.s.) se gli elementi di M sono rappresentati da attributi, variabile statistica (v.s.) se l'insieme M è costituito da elementi di $\mathbb{R}$.

□

OSSERVAZIONE: la classificazione dei caratteri viene ora estesa alle m.s. e alle v.s. per cui, in base alla natura dell'insieme M, parleremo di *mutabile sconnessa* o *ordinale*, oppure di *variabile discreta* o *continua.*

$\star$

Abitualmente sia le mutabili che le variabili statistiche vengono indicate mediante le lettere maiuscole dell'alfabeto anglosassone. Nel seguito riserveremo alle m.s. le prime lettere dell'alfabeto (A, B, C,...) e alle v.s. le ultime (..., X, Y, Z). In alcune situazioni le v.s. possono essere etichettate ricorrendo ad un'unica lettera maiuscola indicizzata, ad esempio $X_1, X_2, \ldots$, oppure $Y_1, Y_2, \ldots$

Nel seguito indicheremo con $\tilde{a}_\alpha$ il valore associato dalla mutabile statistica A alla α-esima unità statistica, in simboli $A(\omega_\alpha) = \tilde{a}_\alpha$. In modo del tutto analogo $\tilde{x}_\alpha$ indicherà il valore associato dalla variabile statistica X alla α-esima unità statistica, cioè $X(\omega_\alpha) = \tilde{x}_\alpha$.
Se consideriamo ora le n unità del collettivo statistico

Definizione 2.7 (Insieme dei dati individuali)
definiamo insieme dei dati individuali di una m.s. A o di una v.s. X l'insieme cositituito dalle n loro determinazioni. Cioè in simboli

$$\{\tilde{a}_\alpha\}_{\alpha=1,\ldots,n} = \{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n\} \tag{2.2}$$

per la m.s. A, e

$$\{\tilde{x}_\alpha\}_{\alpha=1,\ldots,n} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\} \tag{2.3}$$

per la v.s. X

$\square$

È bene tenere a mente che gli insiemi dei dati individuali, corrispondenti alle (2.2) e (2.3), contengono tutte le informazioni circa i caratteri rilevati. Le successive elaborazioni statistiche atte a descrivere il fenomeno di interesse avranno come oggetto primario i suddetti insiemi. Ciò nel senso che, una volta rilevati i dati, perde ogni interesse conservarne i legami con le singole unità statistiche che li hanno espressi.
Osserviamo, inoltre, che gli insiemi (2.2) e (2.3)

- $\star$ *possono non esaurire i rispettivi insiemi M delle modalità dei caratteri*; cioè a rilevazione avvenuta potrà accadere che tra i valori osservati non compaiano tutte le modalità presenti in M. Sicuramente ciò avverrà per i caratteri continui, ma non è escluso accada anche per gli altri, siano essi quantitativi o qualitativi;

⋆ *non necessariamente presentano elementi tutti distinti tra loro*; in altri termini, a rilevazione avvenuta, si potranno osservare $k \leq n$ distinte determinazioni della m.s. o della v.s. oggetto di studio. Su tale concetto ritorneremo allorché affronteremo il problema delle distribuzioni di frequenze.

▷ ESEMPIO 2.4

Con riferimento all'esempio (2.2), una volta condotta la rilevazione su dieci Personal Computer da tavolo, siamo in grado di individuare:

⋆ l'insieme dei dati individuali della m.s. sconnessa $A = \{$*tipo di processore*$\}$:

$$\{\tilde{a}_\alpha\}_{\alpha=1,\ldots,10} = \{\text{P, P, S, C, A, C, A, A, S, P}\}$$

⋆ l'insieme dei dati individuali della m.s. ordinale $B = \{$*tipo di unità ottica*$\}$:

$$\left\{\tilde{b}_\alpha\right\}_{\alpha=1,\ldots,10} = \{\text{R, RW, RW, R, RW, RW, R, R, RW, RW}\}$$

⋆ l'insieme dei dati individuali della v.s. discreta $X = \{$*capacità disco fisso*$\}$:

$$\{\tilde{x}_\alpha\}_{\alpha=1,\ldots,10} = \{40, 60, 20, 60, 20, 80, 60, 60, 20, 60\}$$

⋆ l'insieme dei dati individuali della v.s. continua $Y = \{$*velocità di clock*$\}$

$$\{\tilde{y}_\alpha\}_{\alpha=1,\ldots,10} = \{1.20, 0.85, 1.25, 1.65, 1.50, 1.75, 1.45, 1.55, 1.70, 1.85\}$$

A commento osserviamo che alcuni di tali insiemi non contengono tutti gli elementi dei rispettivi insiemi M delle modalità del carattere (cfr. esempio 2.2) ed inoltre che il solo insieme $\{\tilde{y}_\alpha\}_{\alpha=1,\ldots,10}$ della v.s. di tipo continuo non presenta valori ripetuti.

◁

OSSERVAZIONE: a volte, per svariati motivi, le determinazioni $\{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_n\}$ di una m.s. A vengono "ricodificate" in valori numerici, generalmente coincidenti con un sottoinsieme dei numeri naturali.
Si tenga ben presente che la "ricodifica" delle modalità di una generica mutabile statistica in valori numerici *non* deve portare a pensare che la mutabile statistica sia stata, in qualche modo artificioso, trasformata in una variabile statistica.

⋆

▷ ESEMPIO 2.5

A tal proposito considerando l'insieme dei dati individuali della m.s. A di cui all'esempio (2.4) e volendole riclassificare in accordo alla seguente strategia

$$\begin{cases} \text{se } A(\omega_\alpha) = \text{P} & \rightarrow \quad \tilde{a}_\alpha = 0 \\ \text{se } A(\omega_\alpha) = \text{S} & \rightarrow \quad \tilde{a}_\alpha = 1 \\ \text{se } A(\omega_\alpha) = \text{A} & \rightarrow \quad \tilde{a}_\alpha = 2 \\ \text{se } A(\omega_\alpha) = \text{C} & \rightarrow \quad \tilde{a}_\alpha = 3 \end{cases}$$

per la m.s. A otterremmo il seguente insieme dei dati individuali:

$$\{\tilde{a}_\alpha\}_{\alpha=1,\ldots,10} = \{0, 0, 1, 3, 2, 3, 2, 2, 1, 0\}$$

Sebbene compaiono dei numeri al posto delle determinazioni della mutabile A non si è in alcun modo giustificati dal considerare la stessa quale variabile statistica e non si dovrà mai scordare di indicare congiuntamente all'insieme dei suoi dati individuali la codifica adottata per gli attributi.

◁

2.5. LA MATRICE DEI DATI

Una rilevazione statistica effettuata su un prefissato collettivo rispetto ai diversi caratteri che si desidera analizzare, darà luogo, come si è detto ad altrettante mutabili e/o variabili statistiche, le cui determinazioni vengono abitualmente organizzate ricorrendo alla cosiddetta *matrice dei dati*; trattasi di una matrice le cui colonne contengono gli insiemi dei dati individuali di ciascuna m.s. e/o v.s.. Ogni riga della matrice contiene tutte le informazioni relative alla stessa unità statistica. Una generica matrice dei dati può dunque assumere il seguente aspetto:

unità statistiche	$m.s.\,A$	$m.s.\,B$	$\ldots$	$v.s.\,X$	$v.s.\,Y$	$\ldots$
ω_1	$\tilde{a}_1$	$\tilde{b}_1$	$\ldots$	$\tilde{x}_1$	$\tilde{y}_1$	$\ldots$
ω_2	$\tilde{a}_2$	$\tilde{b}_2$	$\ldots$	$\tilde{x}_2$	$\tilde{y}_2$	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_α	$\tilde{a}_\alpha$	$\tilde{b}_\alpha$	$\ldots$	$\tilde{x}_\alpha$	$\tilde{y}_\alpha$	$\ldots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
ω_n	$\tilde{a}_n$	$\tilde{b}_n$	$\ldots$	$\tilde{x}_n$	$\tilde{y}_n$	$\ldots$

È opportuno ricordare che, generalmente, tra le unità statistiche ω_α non esiste un ordine progressivo; l'indice attribuito a ciascuna di esse nella matrice dei dati si riferisce semplicemente alla riga che esse occupano.

▷ ESEMPIO 2.6

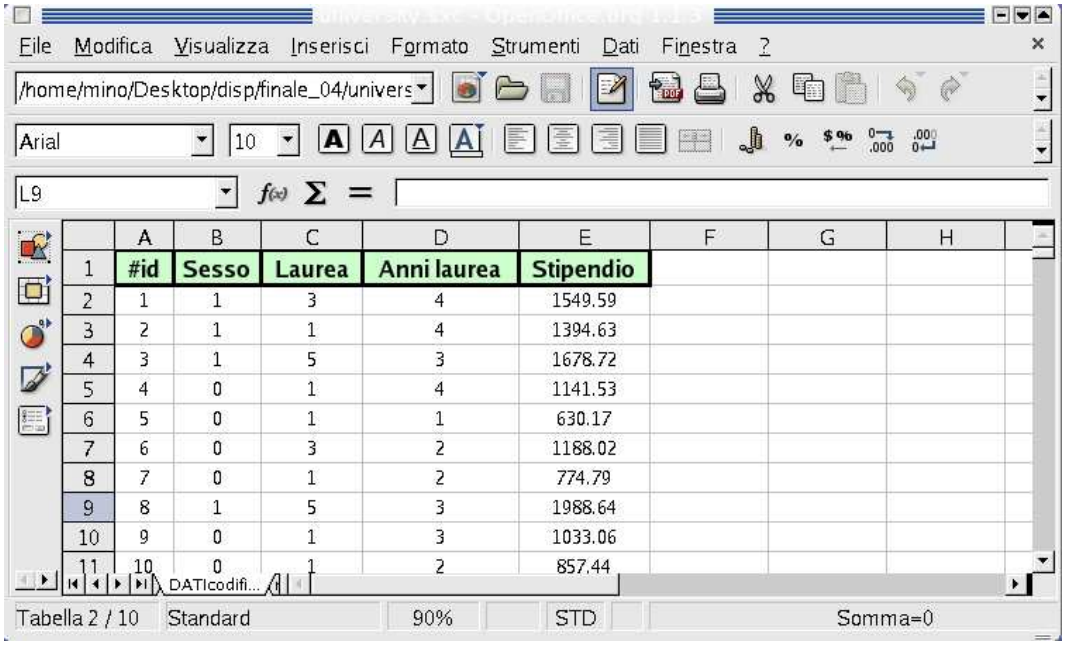
Riprendendo la situazione di cui all'esempio (2.4), il collettivo statistico Ω risulta costituito da $n = 10$ unità sulle quali sono stati rilevati quattro caratteri (il tipo di processore, il tipo di unità ottica, la capacità disco fisso nonché la velocità di clock) ed i risultati ottenuti sono stati organizzati nella seguente matrice dei dati individuali:

unità stat.	*tipo di processore*	*tipo di unità ottica*	*capacità disco fisso*	*velocità di clock*
ω_1	P	R	40	1.20
ω_2	P	RW	60	0.85
ω_3	S	RW	20	1.25
ω_4	C	R	60	1.65
ω_5	A	RW	20	1.50
ω_6	C	RW	80	1.75
ω_7	A	R	60	1.45
ω_8	A	R	60	1.55
ω_9	S	RW	20	1.70
ω_{10}	P	RW	60	1.85

◁

2.6. IL FOGLIO ELETTRONICO

In questo paragrafo e nel corso di tutti i capitoli successivi illustreremo come gli argomenti via via introdotti possano essere trattati ricorrendo all'uso del foglio elettronico *OpenOffice 1.1.2*, versione italiana.
A tal proposito ricordiamo che la quasi totalità dei comandi e delle funzioni che illustreremo sono compatibili con quelle proprie di altri fogli elettronici in commercio, quali ad esempio Microsoft Excel, Claris Works, ecc.
L'organizzazione a matrice dei dati individuali risultanti da una qualsiasi indagine statistica rappresenta la forma comune di registrazione dei dati su un foglio elettronico.
Il file `university.sxc` contiene i dati relativi ad un'indagine condotta circa il reddito di 1100 laureati che hanno trovato impiego nella provincia di Ancona.
La Figura (2.2) riporta la videata OpenOffice delle prime righe del foglio di lavoro. Si noti che sono stati rilevati i seguenti caratteri:

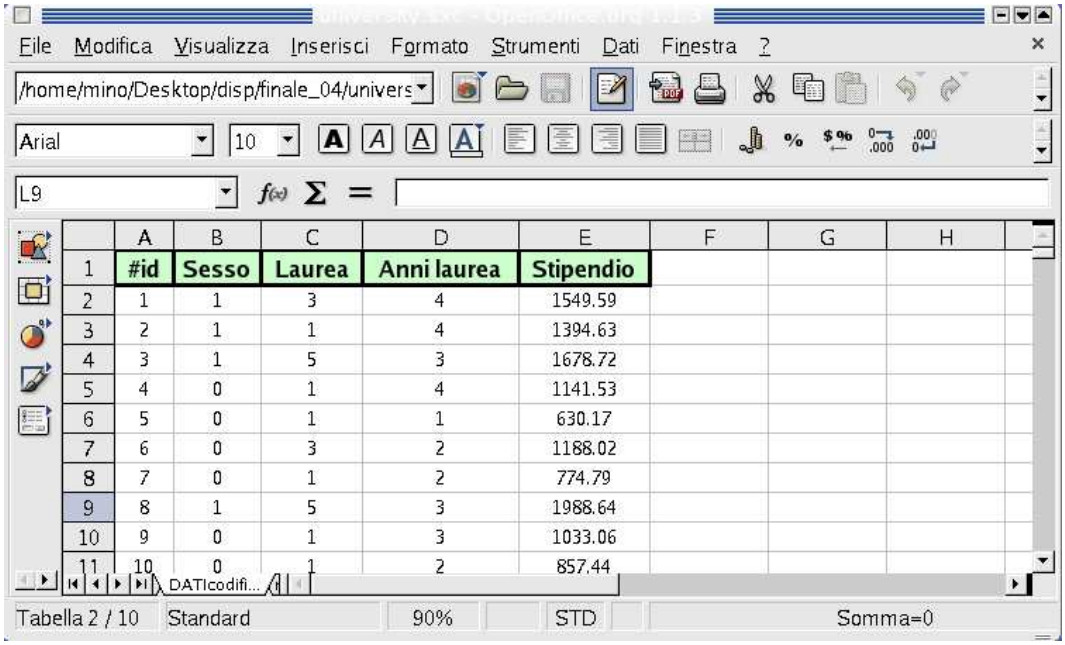


	A	B	C	D	E
1	#id	Sesso	Laurea	Anni laurea	Stipendio
2	1	1	3	4	1549.59
3	2	1	1	4	1394.63
4	3	1	5	3	1678.72
5	4	0	1	4	1141.53
6	5	0	1	1	630.17
7	6	0	3	2	1188.02
8	7	0	1	2	774.79
9	8	1	5	3	1988.64
10	9	0	1	3	1033.06
11	10	0	1	2	857.44

Figura 2.2 Videata OpenOffice, file `university.sxc`

- ⋆ il *sesso*, codificato in 0 se Femmina, 1 se Maschio;
- ⋆ il *tipo di laurea*, codificata in 1 se Agraria, 2 se Architettura, 3 se Economia, 4 se Ingegneria Civile, 5 se Ingegneria Meccanica, 6 se Lettere Classiche, 7 se Magistero, 8 se Scienze Ambientali;
- ⋆ il *numero di anni intercorsi dalla laurea*;
- ⋆ lo *stipendio attuale in euro*;

che hanno dato luogo a due mutabili (con modalità codificate) e a due variabili statistiche che possiamo osservare nelle prime cinque colonne del foglio proposto in figura (2.2).
Per inciso notiamo che la colonna A contiene i numeri progressivi delle unità statistiche e pertanto non potendo essere considerata nè mutabile nè variabile statistica mai sarà oggetto di eleborazione.
Osserviamo, infine, che il foglio è organizzato in modo da contenere su ciascuna riga tutte le informazioni rilevate sulla singola unità statistica e che tale struttura è tipica di ogni data base che verrà utilizzato a fini statistici.



	A	B	C	D	E
1	#id	Sesso	Laurea	Anni laurea	Stipendio
2	1	1	3	4	1549.59
3	2	1	1	4	1394.63
4	3	1	5	3	1678.72
5	4	0	1	4	1141.53
6	5	0	1	1	630.17
7	6	0	3	2	1188.02
8	7	0	1	2	774.79
9	8	1	5	3	1988.64
10	9	0	1	3	1033.06
11	10	0	1	2	857.44

Figura 2.3 File `university.sxc` modificato

Come primo approccio al foglio elettronico, invitiamo il Lettore a modificare il file in modo che le mutabili *sesso* e *tipo di laurea* risultino espresse in termini degli attributi originari. In figura (2.3) è proposta la soluzione ottenuta per la ricodifica mediante un uso appropriato della funzione `SE()`.

2.7. ESERCIZI

▷ ESERCIZIO 2.1

Con riferimento ai seguenti caratteri statistici:

a) il reddito annuo netto dei ricercatori di ruolo presso le Università Italiane;

b) l'ammontare delle vendite mensili di sigarette nazionali nell'area torinese;

c) il numero dei nati vivi nel mese di maggio 2000 negli ospedali piemontesi;

d) il numero di dipendenti delle industrie tessili operanti in Italia al 10.09.2000.

indicare per ciascuno di essi il corrispondente Collettivo Statistico.

◁

▷ ESERCIZIO 2.2

Per ciascuno dei seguenti caratteri statistici, indicare *l'insieme delle modalità*, la *scala di misura che si ritiene più idonea* e quindi classificare il carattere.

a) il reddito annuo lordo, in euro, delle famiglie italiane di operai ed impiegati;

b) l'ammontare delle vendite settimanali dei grandi magazzini operanti nell'area torinese;

c) il numero giornaliero di pezzi difettosi in uscita da un processo produttivo;

d) la superficie forestale delle Regioni italiane;

e) le principali cause di morte degli italiani nell'ultimo decennio;

◁

▷ ESERCIZIO 2.3

Con riferimento all'esercizio 2.2, considerata l'applicazione dal relativo collettivo statistico all'insieme delle modalità di ciascun carattere, si indichi se questa individua una *Mutabile* o una *Variabile* Statistica.

◁

▷ ESERCIZIO 2.4

Indicare di quali informazioni si dovrebbe disporre nel caso si desideri indagare circa:

a) la capacità ricettiva della Valle d'Aosta;

b) la giacenza annua dei depositi presso un Istituto di Credito;

c) la produttività del settore energetico nazionale;

d) la liquidità di un'Azienda ad un prefissato istante temporale;

e) le interruzioni volontarie della gravidanza in Italia in un particolare anno.

◁

CAPITOLO 3

PRIME ELABORAZIONI DI SINTESI

Questo capitolo, ricco di argomenti, descrive i primi strumenti che consentono di sintetizzare l'informazione contenuta nell'insieme dei dati individuali di una mutabile o di una variabile statistica. Definita la distribuzione di frequenza di mutabili e di variabili statistiche, di tipo sia discreto che continuo, si vedrà come essa può essere opportunamente presentata nelle forme tabellare e grafica.

3.1. DISTRIBUZIONE DI FREQUENZE

Come si è detto, gli insiemi dei dati individuali della mutabile statistica A e della variabile statstica X contengono tutte le informazioni circa i caratteri rilevati. Un primo modo per visualizzare sinteticamente l'insieme dei dati individuali è ricorrere alla cosiddetta *distribuzione di frequenze.*
Nel seguito illustriamo i quattro passi necessari all'individuazione della distribuzione di frequenze di una mutabile e di una variabile statistica, introducendo nuovi concetti quali: l'insieme delle modalità distinte, le frequenze assolute e quelle relative.

3.1.1 PRIMO PASSO: L'INSIEME DELLE MODALITÀ DISTINTE

Consideriamo, per la mutabile statistica A, l'insieme:

$$\{a_i\}_{i=1,\dots,k} = \{a_1, a_2, \dots, a_k\} \tag{3.1}$$

costituito dai $k \leq n$ *elementi distinti* presenti in $\{\tilde{a}_\alpha\}_{\alpha=1,\dots,n}$ e, in modo del tutto analogo, per la variabile statistica X consideriamo l'insieme:

$$\{x_i\}_{i=1,\dots,k} = \{x_1, x_2, \dots, x_k\} \tag{3.2}$$

costituito dai $k \leq n$ *valori distinti e posti in ordine crescente* presenti in $\{\tilde{x}_\alpha\}_{\alpha=1,\dots,n}$.
Tali insiemi sono rispettivamente detti *insiemi delle modalità distinte* della m.s. A e della v.s. X e per essi vale la pena osservare che:

- ⋆ nel caso A fosse una m.s. ordinale, gli elementi contenuti nella (3.1) verrebbero ad essere posti secondo il loro ordine naturale;
- ⋆ qualora i dati individuali di una m.s. o di una v.s. fossero tutti distinti tra loro sussisterebbe, evidentemente, l'uguaglianza tra il numero k delle modalità distinte e la numerosità n dell'insieme dei dati individuali ovvero quella del collettivo statistico.

▷ ESEMPIO 3.1

Compiamo il primo passo nel caso della m.s. $A = \{\text{inquadramento professionale}\}$ rilevata su di un collettivo statistico Ω formato da 10 dipendenti della Regione Piemonte. Supponiamo che il suo insieme dei dati individuali sia:

$$\{\tilde{a}_\alpha\}_{\alpha=1,\dots,10} = \{\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_{10}\} = \{\text{I, F, I, D, I, F, F, D, I, I}\}$$

dove si è indicato con I l'attributo impiegato, con F l'attributo funzionario e con D quello di dirigente. Per individuare l'insieme $\{a_i\}_{i=1,\dots,k}$ delle modalità distinte della m.s. A dobbiamo, scorrendo l'insieme $\{\tilde{a}_i\}_{i=1,\dots,10}$, individuare gli elementi distinti e porli secondo il loro ordine naturale poichè la m.s. è in questo caso ordinale. Così facendo avremo pertanto:

$$\{a_i\}_{i=1,2,3} = \{a_1, a_2, a_3\} = \{\text{I}, \text{F}, \text{D}\}$$

◁

Poiché nei passi successivi, per semplicità di esposizione, faremo riferimento unicamente ad una v.s. X prima di procedere proponiamo un esempio di individuazione dell'insieme dei dati individuali per il caso di una variabile statistica discreta.

▷ ESEMPIO 3.2

Consideriamo la v.s. $X = \{\textit{anzianità di inquadramento in ruolo}\}$ dei 10 dipendenti della Regione Piemonte dell'esempio precedente, e supponiamo che il suo insieme dei dati individuali sia:

$$\{\tilde{x}_\alpha\}_{\alpha=1,\dots,10} = \{\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_{10}\} = \{4,\ 2,\ 6,\ 2,\ 4,\ 6,\ 5,\ 4,\ 4,\ 2\}$$

L'insieme delle modalità distinte della v.s. X, che si ottiene inidviduando tra i valori dell'insieme dei dati individuali quelli che non si ripetono, sarà pertanto:

$$\{x_i\}_{i=1,\dots,4} = \{x_1, x_2, x_3, x_4\} = \{2, 4, 5, 6\}\,.$$

◁

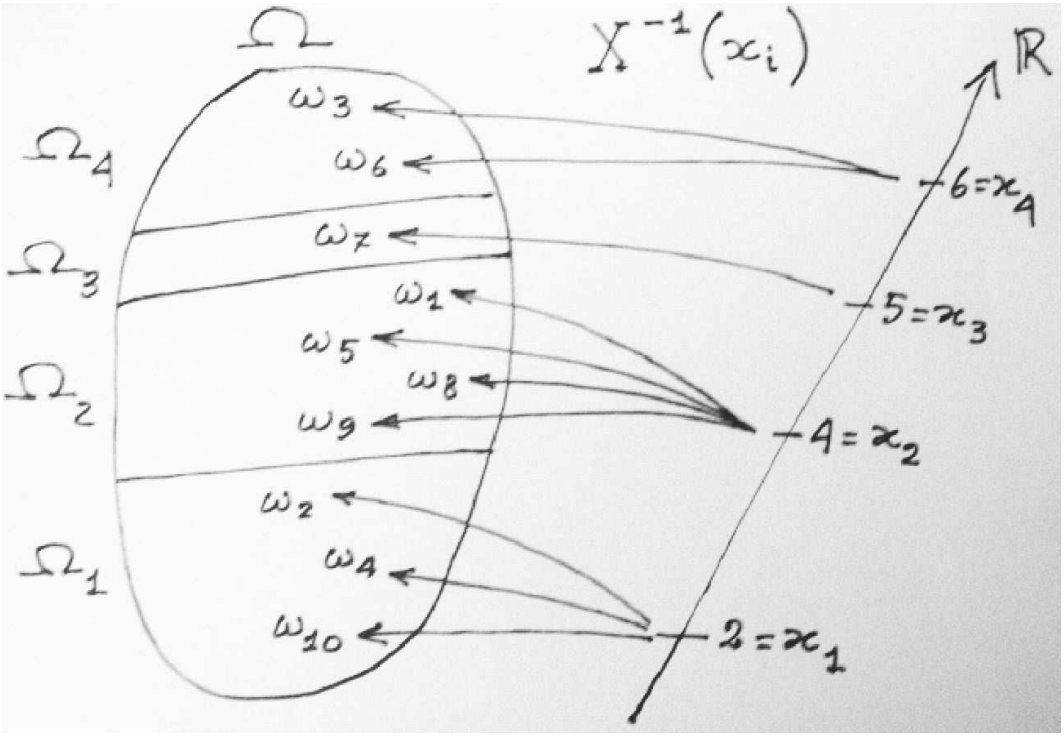


Figura 3.1 Insiemi Ω_i della v.s. *anzianità di inquadramento in ruolo.*

3.1.2 SECONDO PASSO: I SOTTOINSIEMI Ω_i

L'insieme delle modalità distinte fornisce informazioni su quali e quanti siano i valori distinti che la v.s. (o la m.s.) assume nell'insieme M delle modalità del carattere. Per descrivere completamente il contenuto dell'insieme dei dati individuali è necessario associare a ciascuna modalità distinta il numero di unità statistiche che presentano tale modalità. Si tratterà di individuare nel collettivo statistico Ω dei sottoinsiemi i cui elementi avranno la caratteristica di essere associati dalla applicazione X allo stesso valore x_i dell'insieme delle modalità distinte.
A tal fine consideriamo gli insiemi Ω_i (cfr. figura 3.1) costituiti dalle controimmagini di x_i nella applicazione $X(\cdot)$ e cioè $\forall i = 1, \ldots, k$:

$$\Omega_i = \{\omega_\alpha : X(\omega_\alpha) = x_i\}$$

È facile dedurre che essi formano una *partizione* di Ω, infatti:

- $\star$ essi sono insiemi non vuoti, risultando, $\forall\, i$, $\Omega_i \neq \emptyset$;

- ⋆ risultano essere insiemi mutualmente disgiunti, cioè per qualsiasi coppia $i,\ j$, con $i \neq j$, risulterà sempre $\Omega_i \bigcap \Omega_j = \emptyset$;
- ⋆ la loro unione porge Ω, cioè $\bigcup_{i=1}^{k} \Omega_i = \Omega$.

▷ ESEMPIO 3.3

Riprendiamo la v.s. dell'esempio (3.2) e individuiamo i sottoisiemi Ω_i del collettivo formato dai dieci dipendenti. Ricordiamo che l'insieme dei dati individuali è

$$\{\tilde{x}_\alpha\}_{\alpha=1,\ldots,10} = \{\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{10}\} = \{4,\ 2,\ 6,\ 2,\ 4,\ 6,\ 5,\ 4,\ 4,\ 2\}$$

mentre l'insieme delle modalità distinte è $\{x_1, x_2, x_3, x_4\} = \{2, 4, 5, 6\}$.
Avendo a disposizione $k = 4$ modalità distinte applichiamo la definizione per individuare i 4 sottoinsiemi Ω_i di Ω che risultano essere:

$$\begin{aligned}
\Omega_1 &= \{\omega_\alpha : X(\omega_\alpha) = x_1\} = \{\omega_\alpha : X(\omega_\alpha) = 2\} = \{\omega_2, \omega_4, \omega_{10}\} \\
\Omega_2 &= \{\omega_\alpha : X(\omega_\alpha) = x_2\} = \{\omega_\alpha : X(\omega_\alpha) = 4\} = \{\omega_1, \omega_5, \omega_8, \omega_9\} \\
\Omega_3 &= \{\omega_\alpha : X(\omega_\alpha) = x_3\} = \{\omega_\alpha : X(\omega_\alpha) = 5\} = \{\omega_7\} \\
\Omega_4 &= \{\omega_\alpha : X(\omega_\alpha) = x_4\} = \{\omega_\alpha : X(\omega_\alpha) = 6\} = \{\omega_3, \omega_6\}
\end{aligned}$$

Come possiamo notare osservando la figura (3.1) i quattro sottoinsiemi ora definiti sono disgiunti e formano una partizione del collettivo statistico Ω si ha infatti che:

$$\begin{aligned}
\Omega = \bigcup_{i=1}^{4} \Omega_i &= \underbrace{\{\omega_2,\ \omega_4,\ \omega_{10}\}}_{\Omega_1} \bigcup \underbrace{\{\omega_1,\ \omega_5,\ \omega_8,\ \omega_9\}}_{\Omega_2} \bigcup \underbrace{\{\omega_7\}}_{\Omega_3} \bigcup \underbrace{\{\omega_3,\ \omega_6\}}_{\Omega_4} = \\
&= \underbrace{\{\omega_1,\ \omega_2,\ \omega_3, \omega_4,\ \omega_5,\ \omega_6, \omega_7,\ \omega_8,\ \omega_9,\ \omega_{10}\}}_{\Omega}
\end{aligned}$$

◁

3.1.3 TERZO PASSO: FREQUENZA ASSOLUTA E FREQUENZA RELATIVA

Per come sono stati definiti i sottoinsiemi Ω_i possiamo affermare, ad esempio, che gli elementi di Ω_1 sono tutte e sole le unità statistiche del collettivo che possegono la modalità x_1. Più in generale qualunque sia $i \in \{1, \ldots, k\}$ gli elementi di Ω_i sono tutte e sole le unità statistiche del collettivo che possegono la modalità x_i.
I sottoinsiemi Ω_i sono strettamente legati alle modalità distinte della variabile statistica in esame e la loro numerosità riveste un significato così importante da avere un nome proprio dato nella seguente:

Definizione 3.1 (Frequenza assoluta)
$\forall\, i = 1, \ldots, k \leq n$ definiamo frequenza assoluta associata alla modalità x_i della v.s. X il numero n_i di elementi posseduti dall'insieme $\Omega_i \subset \Omega$, cioè:

$$n_i = Nu\{\Omega_i\}$$

□

Poiché risulterà utile conoscere la dimensione di ciascun sottoinsieme Ω_i rispetto alla numerosità dell'intero collettivo statistico diamo anche la seguente:

Definizione 3.2 (Frequenza relativa)
definiamo frequenza relativa f_i associata alla modalità x_i il rapporto:

$$f_i = \frac{Nu\{\Omega_i\}}{Nu\{\Omega\}} = \frac{n_i}{n}$$

□

A commento ci limitiamo ad osservare che:

- ⋆ per quanto già detto circa la natura dei sottoinsiemi Ω_i si ha che $\sum_{i=1}^{k} n_i = n$ e che $\sum_{i=1}^{k} f_i = 1$;
- ⋆ qualora si avesse $k = n$, si avrebbe $n_i = 1$ nel qual caso $f_i = n^{-1}$ e ciò accadrebbe $\forall\, i = 1, \ldots, k$.

3.1.4 Quarto passo: distirbuzione di frequenze

Alla luce di quanto precede, una generica v.s. X potrà così essere univocamente individuata dall'insieme delle modalità distinte $\{x_i\}_{i=1,\ldots,k}$ e dalle corrispondenti frequenze, assolute e/o relative, che rivelano come le n unità del collettivo si distribuiscono fra i diversi valori assunti dalla variabile. L'ultimo passo consiste dunque nel sintetizzare l'informazione contenuta nell'insieme dei dati individuali di una v.s. per mezzo della distrubuzione di frequenze che è definita come segue:

Definizione 3.3 (Distribuzione di frequenze)
con riferimento ad una generica v.s. X, definiamo distribuzione di frequenze assolute l'insieme di coppie $\{(x_i; n_i)\}_{i=1,\ldots,k}$ e, in modo del tutto equivalente, distribuzione di frequenze relative l'insieme di coppie $\{(x_i; f_i)\}_{i=1,\ldots,k}$.

□

In modo esteso scriveremo che la v.s. X ha distribuzione di frequenze assolute:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\dots,k} = \begin{Bmatrix} x_1 & x_2 & \dots & x_k \\ n_1 & n_2 & \dots & n_k \end{Bmatrix}$$

ovvero distribuzione di frequenze relative:

$$X \equiv \begin{Bmatrix} x_i \\ f_i \end{Bmatrix}_{i=1,\dots,k} = \begin{Bmatrix} x_1 & x_2 & \dots & x_k \\ f_1 & f_2 & \dots & f_k \end{Bmatrix}$$

▷ ESEMPIO 3.4

Riprendendo l'esempio della v.s. $X = \{$*anzianità di inquadramento in ruolo*$\}$ dei dipendenti della Regione Piemonte, determiniamo le frequenze assolute associate alle modalità x_i contando gli elementi degli insiemi Ω_i, e otteniamo:

$$n_1 = Nu\{\Omega_1\} = 3 \qquad n_2 = Nu\{\Omega_2\} = 4$$
$$n_3 = Nu\{\Omega_3\} = 1 \qquad n_4 = Nu\{\Omega_4\} = 2$$

La distribuzione di frequenze assolute della v.s. X sarà pertanto:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\dots,4} = \begin{Bmatrix} 2 & 4 & 5 & 6 \\ 3 & 4 & 1 & 2 \end{Bmatrix}$$

Leggendo la distribuzione di frequenze assolute ricaviamo l'informazione che dei nostri 10 dipendenti 3 hanno anzianità in ruolo di 2 anni, 4 dipendenti hanno anzianità in ruolo di 4 anni e così via. Lo stesso si sarebbe potuto ricavare, ma con minor immediatezza, dall'insieme dei dati individuali .
La distribuzione di frequenze relative risulta essere:

$$X \equiv \begin{Bmatrix} x_i \\ f_i \end{Bmatrix}_{i=1,\dots,4} = \begin{Bmatrix} 2 & 4 & 5 & 6 \\ 0.3 & 0.4 & 0.1 & 0.2 \end{Bmatrix}$$

dalla quale immediatamente affermiamo, ad esempio, che il 30% dei dipendenti ha anzianità in ruolo di 2 anni

◁

OSSERVAZIONE: si noti che per una v.s. X vale la seguente uguaglianza:

$$\sum_{\alpha=1}^{n} \tilde{x}_\alpha = \sum_{i=1}^{k} x_i \, n_i$$

In altri termini l'*intensità totale* con cui si è manifestato il carattere in esame, che corrisponde alla somma dei dati individuali $\tilde{x}_\alpha$, può essere espressa come somma dei prodotti tra ciascuna modalità x_i e la corrispondente frequenza n_i.

⋆

▷ ESEMPIO 3.5

Concludiamo questo paragrafo con un esempio che individua la distribuzione di frequenze assolute di una mutabile statistica.
Riprendiamo la m.s. $A = \{$*inquadramento professionale*$\}$ dell'esempio (3.1) e ricordiamo che per essa $\{\tilde{a}_1, \tilde{a}_2, \ldots, \tilde{a}_{10}\} = \{\text{I, F, I, D, I, F, F, D, I, I}\}$ e che l'insieme delle modalità distinte è $\{a_1, a_2, a_3\} = \{\text{I, F, D}\}$.
Avendo $k = 3$ modalità distinte individuiamo inizialmente i 3 sottoinsiemi Ω_i che risultano essere:

$$\Omega_1 = \{\omega_\alpha : A(\omega_\alpha) = a_1\} = \{\omega_\alpha : A(\omega_\alpha) = \text{I}\} = \{\omega_1, \omega_3, \omega_5, \omega_9, \omega_{10}\}$$
$$\Omega_2 = \{\omega_\alpha : A(\omega_\alpha) = a_2\} = \{\omega_\alpha : A(\omega_\alpha) = \text{F}\} = \{\omega_2, \omega_6, \omega_7\}$$
$$\Omega_3 = \{\omega_\alpha : A(\omega_\alpha) = a_3\} = \{\omega_\alpha : A(\omega_\alpha) = \text{D}\} = \{\omega_4, \omega_8\}$$

Banalmente contiamo gli elementi degli insiemi Ω_i e ricaviamo le frequenze assolute, cioè:

$$n_1 = Nu\{\Omega_1\} = 5, \quad n_2 = Nu\{\Omega_2\} = 3 \quad \text{e} \quad n_3 = Nu\{\Omega_3\} = 2$$

Siamo ora in grado di associare a ciascuna modalità distinta della mutabile statistica la corrispondente frequenza assoluta e indicare la sua distribuzione di frequenze che risulta essere:

$$A \equiv \begin{Bmatrix} a_i \\ n_i \end{Bmatrix}_{i=1,2,3} = \begin{Bmatrix} \text{I} & \text{F} & \text{D} \\ 5 & 3 & 2 \end{Bmatrix}$$

◁

3.2. TABELLE DI FREQUENZE

In questo breve paragrafo illustriamo come solitamente vengono presentate le distribuzioni di frequenze. È comune scrivere la distribuzione di frequenze di una variabile statistica, o di una mutabile statistica, in forma tabellare tramite la *tabella di frequenze assolute* formata da due colonne e k righe, che in generale assume la seguente forma

Modalità x_i	freq. ass. n_i
x_1	n_1
$\vdots$	$\vdots$
x_k	n_k

o equivalentemente mediante la *tabella di frequenze relative* qui sotto riportata:

Modalità x_i	freq. rel. f_i
x_1	f_1
$\vdots$	$\vdots$
x_k	f_k

Non è raro che le due tabelle vengano unificate così da avere contemporaneamente le frequenze assolute e relative eventualmente espresse in percentuale:

Modalità x_i	freq. ass. n_i	freq. rel. f_i
x_1	n_1	f_1
$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k

Sovente, tra i primi risultati delle indagini statistiche rese pubbliche compaiono le tabelle delle frequenze delle v.s. o delle m.s. analizzate. Un errore comune è quello di considerare tale tabelle come punto iniziale dell'indagine eseguita piuttosto che come primo importante risultato di sintesi dei dati. Gli studenti dimenticano facilmente che le tabelle di frequenze sono un modo di rappresentare la distribuzione di frequenze e non ricordano il significato degli elementi che le compongono. Ricordare che in questo testo le tabelle di frequenze sono state introdotte nel secondo paragrafo del terzo capitolo dovrebbe aiutare a non considerarle come base di partenza di una indagine statistica.
Ricordiamo inoltre che, mentre una distribuzione di frequenze può sempre essere posta in forma tabellare, non tutte le tabelle pubblicate riflettono in effetti distribuzioni di frequenze. Si pensi ad esempio ad una tabella ISTAT che riporti i dati relativi alla produzione annua di soia nelle regioni italiane; lasciando al Lettore la verifica, osserviamo che una siffatta tabella riporta non già una distribuzione di frequenze bensì una distribuzione di quantità. Così anche, disponendo della tabella che riporta il fatturato annuo di una società nell'ultimo triennio va tenuto a mente che si sta analizzando una "serie storica" piuttosto che la distribuzione di ferquenze di una variabile statistica.

▷ ESEMPIO 3.6

Con riferimento alla m.s. $A = \{$*inquadramento professionale*$\}$ introdotta all'esempio (3.1), le corrispondenti distribuzioni di frequenze assolute e relative potranno essere rappresentate mediante la semplice tabella:

a_i	n_i	f_i
Impiegato	5	0.5
Funzionario	3	0.3
Dirigente	2	0.2

Leggendo congiuntamente le prime due colonne della tabella abbiamo la distribuzione di frequenze assolute e non ci scordiamo di interpretare, ad esempio, la coppia $(\mathrm{F}, 3)$ dicendo che 3 sono i dipendenti che hanno qualifica da Funzionario. Se consideriamo invece la prima e l'ultima colonna otteniamo la distribuzione di frequenze relative e la coppia $(\mathrm{F}, 0.3)$ ci consente di affermare che il 30% delle unità statistiche considerate ha qualifica di Funzionario.

◁

Come vedremo nel seguito, alla tabella che esprime le distribuzioni di frequenze assolute e relative vengono talvolta aggiunte altre colonne contenenti particolari valori associati a ciascuna modalità in modo da avere in forma compatta e facilmente leggibile ulteriori informazioni sulla variabile statistica.

3.3. IL PROBLEMA DEL RAGGRUPPAMENTO IN CLASSI

Nei paragrafi precedenti abbiamo volutamente omesso esempi riguardanti le variabili statistiche continue. Trattando infatti con v.s. continue spesso accade che l'insieme dei dati individuali sia costituito da elementi tutti diversi tra loro; di conseguenza le distribuzioni di frequenza non fornirebbero alcuna sintesi e non consentirebbero di cogliere come le unità del collettivo si distribuiscano fra i diversi valori assunti dalla variabile in esame.
In tali situazioni abitualmente si ricorre a *raccogliere i dati individuali in classi di misure* e si presenta la distribuzione di frequenze dei dati raccolti in classi, così come illustrato brevemente dall'esempio seguente.

▷ ESEMPIO 3.7

Si supponga di disporre della v.s. $X = \{\textit{quantità di pane}\}$ (in kg) venduta in una data giornata in nove piccoli supermercati della stessa catena. Dall'insieme dei dati inidviduali

$$\{\tilde{x}_\alpha\}_{\alpha=1,\dots,9} = \{98.4,\ 51.3,\ 87.2,\ 42.9,\ 62.7,\ 48.5,\ 74.8,\ 69.2,\ 71.6\}$$

osservando valori tutti distinti sarà $k = n = 9$ e, pertanto, otteniamo la seguente distribuzione di frequenze assolute che, seppur conservando la totale informazione rilevata, non favorisce la sintesi dei dati:

$$\begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\dots,9} = \begin{Bmatrix} 42.9 & 48.5 & 51.3 & 62.7 & 69.2 & 71.6 & 74.8 & 87.2 & 98.4 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{Bmatrix}$$

Volendo una forma più compatta e accettando di perdere una parte di informazione contenuta nell'insieme dei dati individuali è possibile definire ad esempio le tre

classi di peso $]40;60]$, $]60;80]$ e $]80;100]$ e associare a ciascuna il numero di unità statistiche per cui la variabile X assume determinazione all'interno della classe stessa. Così facendo avremo la seguente distribuzione di frequenze dei dati raccolti in classi:

$$\left\{\begin{matrix}]40;60] &]60;80] &]80;100] \\ 3 & 4 & 2\end{matrix}\right\}$$

$\triangleleft$

Da un punto di vista assai generale, volendo raccogliere i dati individuali in k classi chiuse a destra, scelto l'intervallo $L =]a;\, b] \in \mathbb{R}$, i cui estremi soddisfano le relazioni $a < \min_\alpha\{\tilde{x}_\alpha\}$ e $b > \max_\alpha\{\tilde{x}_\alpha\}$, l'insieme delle classi sarà dato da una sua qualsiasi partizione $\{L_i\}_{i=1,\dots k}$ con:

$$L_i =]l_i; l_{i+1}] \qquad l_1 = a \qquad l_{k+1} = b$$

Il numero di unità statistiche "appartenenti" a ciascuna classe L_i e cioè:

$$n_i = Nu\{\omega_\alpha : X(\omega_\alpha) \in L_i\} = Nu\{\omega_\alpha : l_i < X(\omega_\alpha) \leq l_{i+1}\}$$

verrà ad indicare la frequenza associata alla classe stessa e la v.s. in esame avrà pertanto distribuzione di frequenze assolute:

$$\left\{\begin{matrix}L_i \\ n_i\end{matrix}\right\}_{i=1,\dots,k} = \left\{\begin{matrix}l_i \dashv l_{i+1} \\ n_i\end{matrix}\right\}_{i=1,\dots,k}$$

Si noti che nella distribuzione di frequenze qui sopra definita si è utilizzato il simbolo $\dashv$ per indicare una generica classe chiusa a destra, da ora e nel seguito varrà pertanto l'uguaglianza $]l_i; l_{i+1}] = l_i \dashv l_{i+1}$.
Ovviamente è possibile anche il raccoglimento dei dati in classi chiuse a sinistra scegliendo una partizione $\{L_i\}_{i=1,\dots k}$ dell'intervallo $L = [a; b[$ per cui $L_i = [l_i; l_{i+1}[$ e attribuendo a ciascuna classe la frequenza assoluta n_i data da:

$$n_i = Nu\{\omega_\alpha : X(\omega_\alpha) \in L_i\} = Nu\{\omega_\alpha : l_i \leq X(\omega_\alpha) < l_{i+1}\}$$

Per indicare una classe chiusa a sinistra utilizzeremo il simbolo $\vdash$, nel seguito varrà pertanto l'uguaglianza: $[l_i; l_{i+1}[= l_i \vdash l_{i+1}$
Osserviamo infine che in termini tecnici:

- $\star$ i valori l_i e l_{i+1} rappresentano i *limiti di classe* rispettivamente inferiore e superiore,
- $\star$ la differenza $w_i = l_{i+1} - l_i$ viene detta *modulo* o più semplicemente *ampiezza* della classe i-esima.

▷ ESEMPIO 3.8

Si immagini che l'analisi chimica effettuata su 40 bottiglie di acqua minerale in relazione al contenuto di Calcio abbia fornito i seguenti risultati, espressi in mg/l:

64.5	56.7	58.9	65.5	70.5	85.4	70.0	72.5	57.7	63.5
43.5	45.7	74.5	46.5	52.4	68.3	70.5	77.3	60.3	82.5
67.4	67.8	47.5	55.7	62.8	56.2	57.5	65.2	49.9	75.6
70.5	72.5	87.6	53.5	73.5	65.7	83.2	80.2	87.5	57.7

Dal momento che $\min_\alpha\{\tilde{x}_\alpha\} = 43.5$ e $\max_\alpha\{\tilde{x}_\alpha\} = 87.6$, poniamo in modo arbitrario $a = 40$ e $b = 90$ mg/l. Posto di voler raccogliere i dati individuali in cinque classi di ugual ampiezza ($w_i = w = 10$) effettuiamo la seguente partizione dell'intervallo $]40; 90]$:

$$L_1 =]40; 50] \quad L_2 =]50; 60] \quad L_3 =]60; 70] \quad L_4 =]70; 80] \quad L_5 =]80; 90]$$

Operando il tal modo, si ottiene:

$$\{\tilde{x}_\alpha : \tilde{x}_\alpha \in L_1\} = \{43.5,\ 45.7,\ 46.5,\ 47.5,\ 49.9\}$$
$$\{\tilde{x}_\alpha : \tilde{x}_\alpha \in L_2\} = \{56.7,\ 58.9,\ 57.7,\ 52.4,\ 55.7,\ 56.2,\ 57.5,\ 53.5,\ 57.7\}$$
$$\{\tilde{x}_\alpha : \tilde{x}_\alpha \in L_3\} = \{64.5,\ 65.5,\ 63.5,\ 68.3,\ 60.3,\ 67.4,\ 67.8,\ 62.8,\ 65.2,\ 65.7\}$$
$$\{\tilde{x}_\alpha : \tilde{x}_\alpha \in L_4\} = \{70.5,\ 70.0,\ 72.5,\ 74.5,\ 70.5,\ 77.3,\ 75.6,\ 70.5,\ 72.5\}$$
$$\{\tilde{x}_\alpha : \tilde{x}_\alpha \in L_5\} = \{85.4,\ 82.5,\ 87.6,\ 80.2,\ 87.5,\ 83.2\}$$

da cui la distribuzione di frequenze assolute:

$$\begin{Bmatrix} L_i \\ n_i \end{Bmatrix}_{i=1,\ldots,5} = \begin{Bmatrix} 40 \dashv 50 & 50 \dashv 60 & 60 \dashv 70 & 70 \dashv 80 & 80 \dashv 90 \\ 5 & 9 & 10 & 9 & 6 \end{Bmatrix}$$

Osserviamo che è sconsigliabile indicare nella distribuzione di frequenze classi aperte, cioè non sarebbe opportuno porre in luogo della prima classe la dicitura *fino a 50 mg/l* e per l'ultima *oltre 80 mg/l*;

◁

La tecnica di raccoglimento dei dati individuali in classi può tornare utile per la presentazione di distribuzioni di frequenza di v.s. di tipo discreto quando le modalità distinte siano in numero elevato.
Questo è il caso ad esempio delle tabelle pubblicate dall'ISTAT relativamente alla v.s. età degli Italiani, o al reddito che, derivando da un procedimento di conteggio, è da considerarsi carattere discreto ma che viene assimilato a continuo poiché la distribuzione di frequenze della v.s. rilevata è sempre presentata facendo ricorso ai dati raccolti in classi.

▷ ESEMPIO 3.9

Diamo ora un esempio di raccoglimento in classi per una v.s. di tipo discreto.
Da un'indagine effettuata su 200 automobilisti italiani al fine di indagare circa il numero di infrazioni al Codice della Strada commesse nel corso dell'anno corrente, risultano i seguenti valori individuali:

0	1	0	0	4	4	0	7	0	1	0	0	0	1	1	0	2	0	0	0
1	0	3	1	0	4	0	0	1	0	2	0	3	0	0	0	0	0	0	0
0	0	0	3	0	1	0	0	2	0	0	0	1	0	0	4	0	0	1	0
0	0	6	0	0	3	0	0	5	0	0	0	0	0	0	0	0	0	0	0
0	0	0	2	0	0	0	3	0	0	1	0	1	2	0	3	0	0	0	1
1	0	0	0	0	3	0	0	0	3	0	0	2	0	0	0	0	0	0	0
0	1	0	0	1	0	1	3	1	0	1	1	1	0	1	1	1	1	1	1
1	0	0	1	2	1	0	0	0	1	2	1	1	1	0	2	1	1	1	0
1	1	0	2	1	0	2	1	0	0	0	2	0	0	2	2	1	2	2	2
2	1	3	0	0	0	0	0	4	0	3	0	0	5	6	0	3	0	7	3

Manifestamente la v.s. discreta $X = \{n^\circ$ *di infrazioni al C.d.S.*$\}$ presenta $k = 8$ modalità distinte; al fine di presentarne la distribuzione di frequenze assolute, scegliamo di raccogliere i dati in 5 classi chiuse a sinistra. In definitiva si ha la distribuzione di frequenze assolute:

$$\left\{\begin{array}{ccccc} 0 \vdash 1 & 1 \vdash 2 & 2 \vdash 3 & 3 \vdash 4 & 4 \vdash 8 \\ 115 & 43 & 18 & 13 & 11 \end{array}\right\}$$

Come il lettore facilmente intuisce 115 automobilisti non hanno commesso alcuna infrazione mentre 43 automobilisti hanno commesso . . . e così via.

◁

Ciò premesso, sia dal punto di vista concettuale che sotto il profilo operativo sorgono spontanee alcune domande concernenti essenzialmente il modulo di classe, il numero delle classi nonché la modalità attribuibile a ciascuna classe. Diciamo subito che non esiste una soluzione unica a tali problemi, ma piuttosto vi sono regole a cui è bene attenersi. Sebbene su tale argomento ritorneremo nel corso della trattazione, per il momento osserviamo:

- ⋆ la dimensione del modulo delle classi dipende ovviamente dalla natura del particolare carattere che descrive il fenomeno oggetto di studio. Generalmente, perlomeno nelle prime fasi dell'elaborazione statistica, è bene ricorrere a classi di modulo costante;

- ⋆ anche se in linea di massima il numero k delle classi può essere scelto in modo arbitrario, esso è strettamente legato alla distribuzione della variabile in esame. Secondo la regola di Sturges, adottata e implementata dalla maggior parte dei software statistici, il numero di classi viene determinato considerando l'intero più prossimo a $1 + log_2(n)$;
- ⋆ quanto al valore rappresentativo o modalità di ciascuna classe abitualmente, supponendo che le n_i unità statistiche della classe siano uniformemente distribuite, si sceglie il *valore centrale della classe* stessa, cioè $x_i = (l_i + l_{i+1})/2$.

A riguardo di quest'ultimo punto, deve sottolinearsi che a volte si dispone unicamente di distribuzioni di frequenze con dati già raccolti in classi (si pensi ad esempio alle pubblicazioni ISTAT). In tali situazioni, ai fini di successive elaborazioni, è giocoforza ricorrere ai valori centrali di classe ma non si deve scordare che ciò facendo si ottengono valori tanto più approssimati quanto più ci si allontana dall'ipotesi di equidistribuzione dei dati all'interno di ciascuna classe. L'esempio successivo illustra come nel passaggio dai dati individuali alle classi a fronte di guadagno in sinteticità si ha tuttavia perdita di informazione.

▷ ESEMPIO 3.10

Si immagini che la v.s. $X = \{$*importo delle fatture emesse il 24.09.05 da una piccola Società commerciale*$\}$ abbia assunto i seguenti valori individuali espressi in euro:

91.50	91.50	91.50	91.50	93.50
93.50	97.45	97.45	97.45	97.45
97.45	100.00	100.00	100.00	100.00
100.00	100.00	104.50	104.50	104.50

per cui l'ammontare delle fatture emesse risulta $\sum_{\alpha=1}^{n} \tilde{x}_\alpha = 1953.75$ euro.
Evidentemente, saremmo giunti allo stesso risultato se avessimo fatto riferimento alla distribuzione di frequenze assolute della v.s. X; infatti, essendo:

$$X \equiv \left\{ \begin{matrix} x_i \\ n_i \end{matrix} \right\}_{i=1,\ldots,5} = \left\{ \begin{matrix} 91.50 & 93.50 & 97.45 & 100.00 & 104.50 \\ 4 & 2 & 5 & 6 & 3 \end{matrix} \right\}$$

si otterrebbe $\sum_{i=1}^{5} \tilde{x}_i \, n_i = 1953.75$ euro.

Si supponga ora di disporre per la v.s. in esame della seguente distribuzione di frequenze assolute con dati raccolti in classi:

$$X \equiv \left\{ \begin{matrix} L_i \\ n_i \end{matrix} \right\}_{i=1,\ldots,3} = \left\{ \begin{matrix} 90 \dashv 95 & 95 \dashv 100 & 100 \dashv 105 \\ 6 & 11 & 3 \end{matrix} \right\}$$

Scegliendo per ciascuna classe il centro, cioè $x_1 = 92.5$, $x_2 = 97.5$ e $x_3 = 102.5$, l'ammontare totale delle fatture emesse risulterebbe pari a $\sum_{i=1}^{3} x_i\, n_i = 1935$ euro; risultato "piuttosto distante" da quello esatto precedentemente calcolato.

$\triangleleft$

3.4. RAPPRESENTAZIONI GRAFICHE

Visualizzare la distribuzione di frequenza attraverso un grafico consente di coglierne alcuni aspetti caratteristici in modo immediato, ad esempio individuare la modalità che viene assunta meno frequentemente e quella che si presenta più frequentemente, e così via.

Ciò che segue è una breve descrizione dei principali grafici utilizzabili per rappresentare distribuzioni di frequenze di mutabili e variabili statistiche. Tale rassegna non ha la pretesa di essere esaustiva di tutti i possibili grafici, ma è data al fine di evidenziare al Lettore il diverso impiego dei grafici a seconda della distribuzione di frequenza che si vuole rappresentare.

3.4.1 RAPPRESENTAZIONI GRAFICHE PER MUTABILI STATISTICHE

Le più comuni e ricche di varianti rappresentazioni grafiche di mutabili statistiche sono i *diagrammi a torta* ed i *diagrammi a barre* che possiamo definire come segue.

Definizione 3.4 (Diagramma a torta)
è un grafico formato da un cerchio suddiviso in k spicchi le cui aree sono proporzionali alle frequenze associate a ciascuna modalità della mutabile statistica.

$\square$

Per disegnare il diagramma a torta della distribuzione di frequenze di una m.s., occorre stabilire l'angolo α_i di ogni "spicchio" (cioè settore circolare). Infatti dalla proporzione $\alpha_i : 360^\circ = n_i : n$ si ricava $\alpha_i = 360^\circ\ n_i/n = 360^\circ\ f_i$.

Definizione 3.5 (Diagramma a barre)
è un grafico formato da k rettangoli non contigui posti sull'asse orizzontale con basi uguali e altezze proporzionali alle frequenze assolute (o relative) associate a ciascuna modalità distinta della mutabile statistica.

$\square$

▷ ESEMPIO 3.11

Si immagini che la rilevazione del carattere $\{$*titolo di studio*$\}$ sui 35 dipendenti di una piccola azienda abbia dato luogo alla m.s. A con la seguente distribuzione di frequenze che abbiamo posto per comodità in forma tabellare:

Titolo di studio	n_i	f_i
Elementare	10	0.29
Media inf.	5	0.14
Media sup.	15	0.43
Laurea	5	0.14

Per una corretta rappresentazione grafica della distribuzione della m.s. A possiamo ricorrere ad un diagramma a torta oppure a barre, sebbene quest'ultimo sia generalmente preferibile perché di più immediata lettura, vedasi figura (3.2).

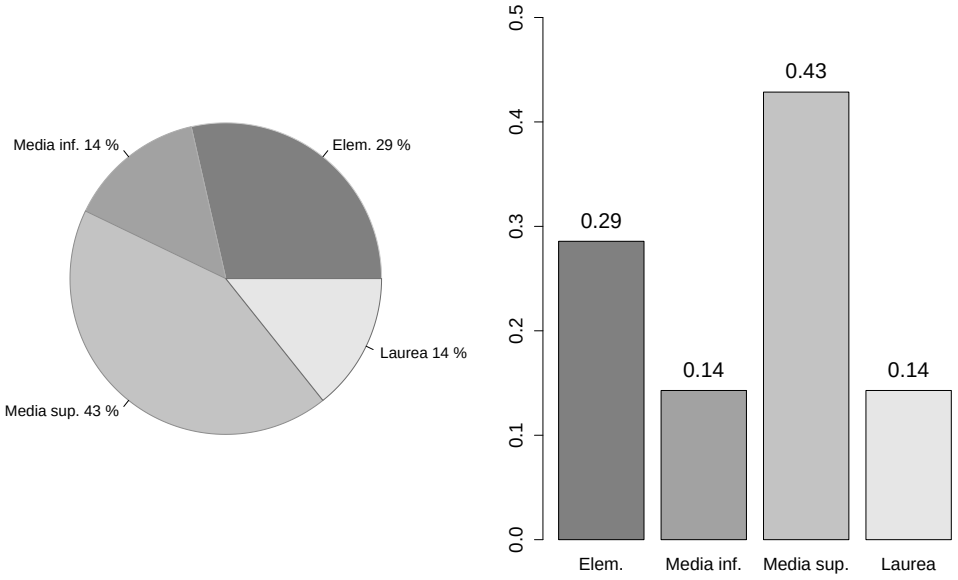


Figura 3.2 Diagramma a torta e diagramma a barre, esempio 3.11.

Osserviamo che avremmo potuto rappresentare la distribuzione di frequenze della m.s. in termini di frequenze assolute, tuttavia ciò è sconsigliabile, in modo particolare qualora si vogliano confrontare mutabili rilevate su collettivi differenti.

◁

▷ ESEMPIO 3.12

Da un'indagine circa il {*settore di occupazione*} condotta su 2694 occupati, di cui 534 Femmine e 2160 Maschi, risultano le seguenti distribuzioni di frequenze:

Settore di occupazione	Femmine n_i	Maschi n_i
Agricultura	34	120
Industria	180	1420
Servizi	320	620

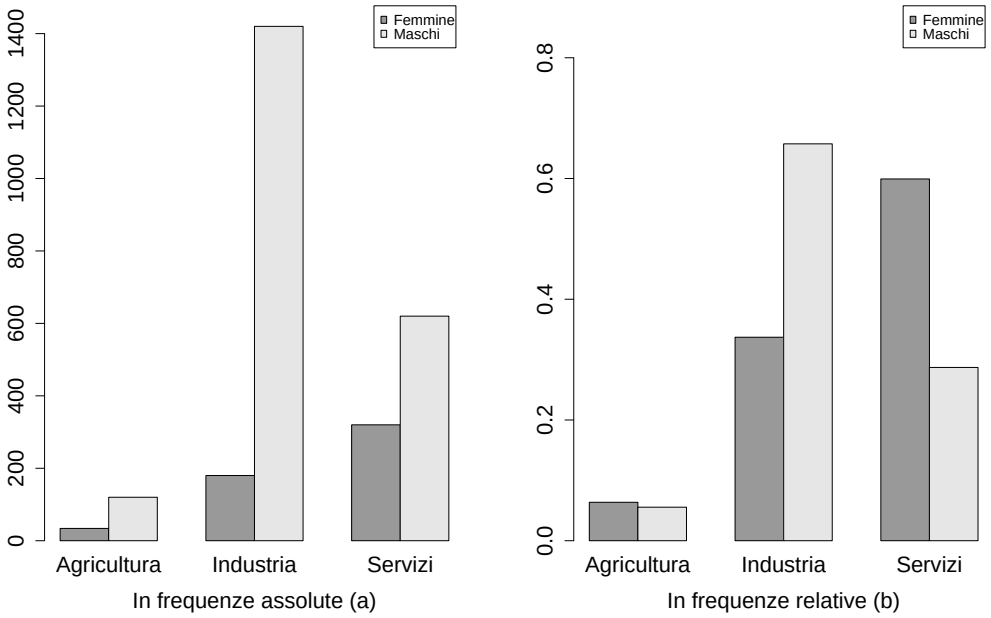


Figura 3.3 Diagrammi a barre, esempio 3.12.

Desiderando rappresentare graficamente le distribuzioni della m.s. sia per i Maschi che per le Femmine, ricorriamo al digramma a barre di figura (3.3, pannello a).
Un confronto tra le due distribuzioni in esame può essere fatto unicamente scegliendo di lavorare in termini di frequenze relative. In tal caso:

Settore di occupazione	Femmine f_i	Maschi f_i
Agricultura	0.06	0.06
Industria	0.34	0.66
Servizi	0.60	0.29

Già dalla tabella risulta evidente, ad esempio che il 60% delle femmine è occupato nel settore dei servizi, mentre il 66% dei maschi in quello dell'industria, tuttavia il grafico riportato in figura (3.3, pannello b) ci fornisce una corretta visione d'insieme delle due distribuzioni.

◁

3.4.2 RAPPRESENTAZIONI GRAFICHE PER V.S. DISCRETE

Per le v.s. di tipo discreto la più comune rappresentazione grafica, anch'essa con diverse varianti, è offerta dal *grafico ad ordinate.*

Definizione 3.6 (Grafico a bastoncini)
è un grafico formato da k segmenti, paralleli all'asse delle ordinate, posizionati in ascissa in corrispondenza delle modalità x_i e di altezza pari alle frequenze assolute (o relative) associate.

□

▷ ESEMPIO 3.13

Si immagini che la v.s. $X = \{$*numero di dipendenti extracomunitari*$\}$, rilevata su un collettivo di 73 medio-piccole aziende agricole operanti nella provincia di Imperia, abbia la seguente distribuzione di frequenze:

# di dipendenti extracomunitari	n_i	f_i
0	28	0.38
1	15	0.21
2	12	0.16
3	7	0.10
4	5	0.07
5	3	0.04
6	2	0.03
7	1	0.01

◁

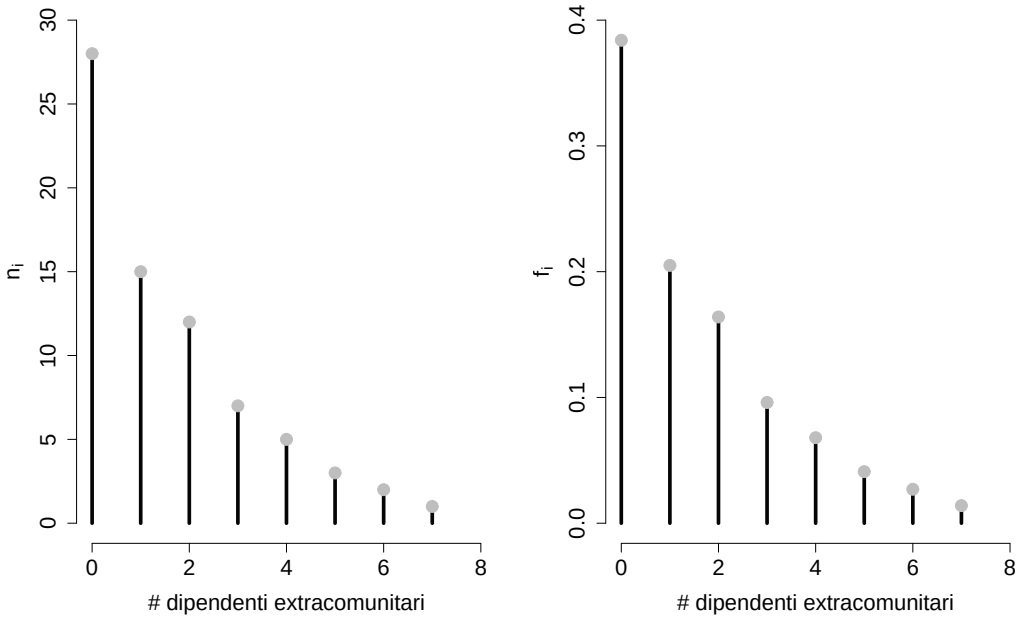


Figura 3.4 Diagramma a bastoncini, esempio 3.13.

La figura (3.4) riporta, sottoforma di grafico a bastoncini, la distribuzione della v.s. X sia in termini di frequenze assolute che relative.
Come già osservato, la rappresentazione grafica della distribuzione in termini di frequenze relative è di immediata interpretazione ogniqualvolta si debbano confrontare tra loro medesime variabili statistiche rilevate su collettivi differenti.

3.4.3 Rappresentazioni grafiche per v.s. con dati raccolti in classi

Lavorando con una variabile statistica di tipo continuo, e avendo raccolto i dati individuali in classi, per rappresentare graficamente la sua distribuzione è necessario disporre di uno strumento che contemporaneamente metta in evidenza l'ampiezza delle classi e le frequenze corrispondenti; tale strumento è l'istogramma.

Definizione 3.7 (Istogramma)
è un grafico formato da k rettangoli contigui ciascuno con base coincidente con una classe, e con area proporzionale alla frequenza assoluta della classe medesima.

□

In pratica, per costruire l'istogramma è necessario, per ogni classe i, individuare l'ampiezza di classe w_i e determinare l'altezza h_i del rettangolo imponendo le seguenti condizioni sull'area dello stesso:

$$\begin{cases} A_i = w_i \cdot h_i & \text{dalla definizione di area di un rettangolo} \\ A_i = c \cdot n_i & \text{dalla definizione di istogramma (con c costante di proporzionalità)} \end{cases}$$

dalle quali si ricava immediatamente

$$h_i = \frac{c\, n_i}{w_i} \tag{3.3}$$

Evidentemente la scelta del valore per la costante c induce diversi tipi di istogramma caratterizzati ciascuno da una diversa area totale.
Casi più comuni sono essenzialmente due:

- ⋆ desiderando un'area totale pari a n, occorrerà porre $c = 1$ per cui i singoli rettangoli dell'istogramma avranno altezze:

$$h_i = \frac{n_i}{w_i}$$

- ⋆ desiderando un'area totale unitaria, occorrerà porre $c = n^{-1}$ per cui i singoli rettangoli dell'istogramma avranno altezze:

$$h_i = \frac{n_i}{n\, w_i} = \frac{f_i}{w_i}$$

Osserviamo che alcune analisi statistiche pubblicate riportano tra i risultati istogrammi con rettangoli di altezza pari alle frequenze assolute o relative. Tali grafici possono essere denominati istogrammi *se e solo se le classi presentano ampiezza costante.*
Quando l'ampiezza di classe è costante, cioè per ogni i si ha $w_i = w$, dall'equazione (3.3) imponendo $c = w$ otteniamo $h_i = n_i$ oppure imponendo $c = w/n$ otteniamo $h_i = f_i$.
Va tuttavia osservato che gli istogrammi così costruiti hanno area totale rispettivamente pari a $w\, n$ e w.

▷ ESEMPIO 3.14

Con questo esempio ci proponiamo di costruire l'istogramma per una variabile statistica di tipo continuo con dati raccolti in classi di modulo non costante e di sottolineare l'importanza di una corretta determinazione delle altezze dei rettangoli che lo compongono.

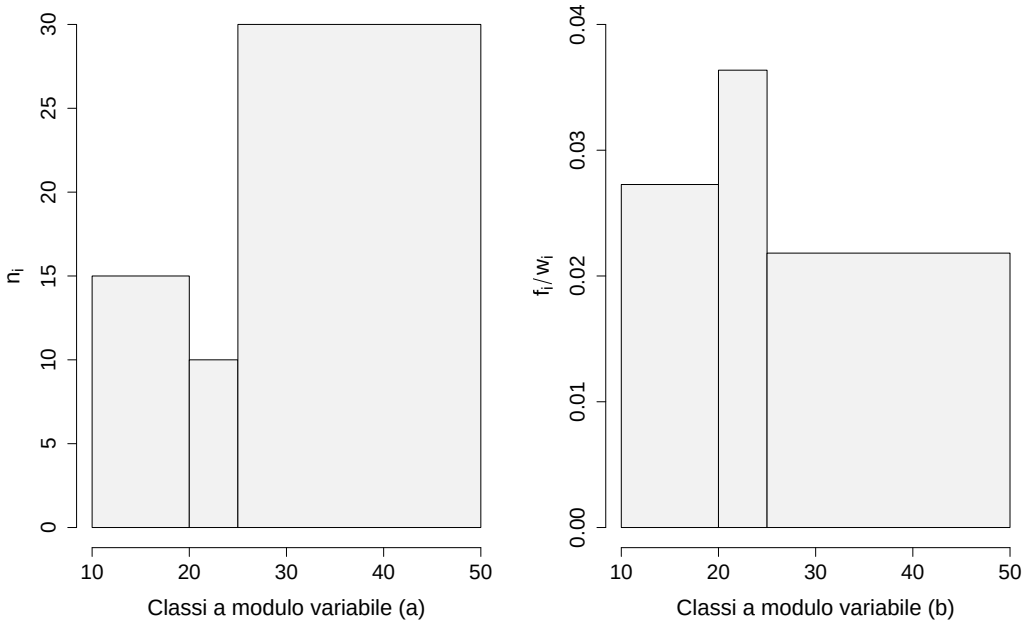


Figura 3.5 Istogrammi con classi a modulo variabile, esempio 3.14.

A tal fine si immagini che una variabile statistica abbia la seguente distribuzione di frequenze:

Classi	n_i	w_i
$10 \dashv 20$	15	10
$20 \dashv 25$	10	5
$25 \dashv 50$	30	25

Il grafico (a) di figura (3.5), i cui rettangoli hanno altezza pari alla frequenza assoluta di ciascuna classe, oltre a non potersi definire istogramma, fornisce una rappresentazione distorta della distribuzione dei dati.

Per una corretta costruzione dell'istogramma di area totale unitaria occorre porre che i rettangoli che lo compongono abbiano altezza pari a f_i/w_i. Con semplici calcoli otteniamo nella tabella seguente le altezze dell'istogramma di figura (3.5, a):

Classi	n_i	w_i	f_i	$h_i = f_i/w_i$
$10 \dashv 20$	15	10	0.273	0.027
$20 \dashv 25$	10	5	0.182	0.036
$25 \dashv 50$	30	25	0.545	0.022

Confrontando i due istogrammi della figura (3.5) è immediato notare come quello a destra consenta di apprezzare la "densità" delle unità statistiche di ciascuna classe.

$\triangleleft$

$\triangleright$ ESEMPIO 3.15

La rilevazione della statura di $n = 100$ individui ha dato luogo alla v.s. $X = \{\textit{statura}\}$ di cui la tabella che segue riporta la distribuzione di frequenza avendo scelto di raccogliere i dati individuali in quattro classi a modulo costante e pari a 10 cm:

Classi di statura (in cm)	Frequenze assolute	Frequenze relative
$160 \dashv 170$	10	0.10
$170 \dashv 180$	49	0.49
$180 \dashv 190$	36	0.36
$190 \dashv 200$	5	0.05

La costruzione del corrispondente istogramma, così come evidenziato in figura (3.6), può avvenire indifferentemente nei seguenti due modi:

- $\star$ posto $c = w$ l'istogramma sarà espresso in termini di frequenze assolute. I rettagoli che lo compongono avranno base uguale all'ampiezza di classe ($w_i = w = 10$) e altezza uguale alla corrispondente frequenza assoluta n_i. Ciò facendo l'area totale dell'istogramma sarà pari a $w \cdot n = 1000$;
- $\star$ posto $c = 1/n$ l'istogramma sarà espresso in termini di "altezze". Infatti i rettagoli che lo compongono avranno base uguale all'ampiezza di classe ($w_i = w = 10$) ed altezza uguale alla corrispondente frequenza relativa divisa per l'ampiezza di classe (f_i/w). In tal modo l'area totale dell'istogramma è unitaria.

$\triangleleft$

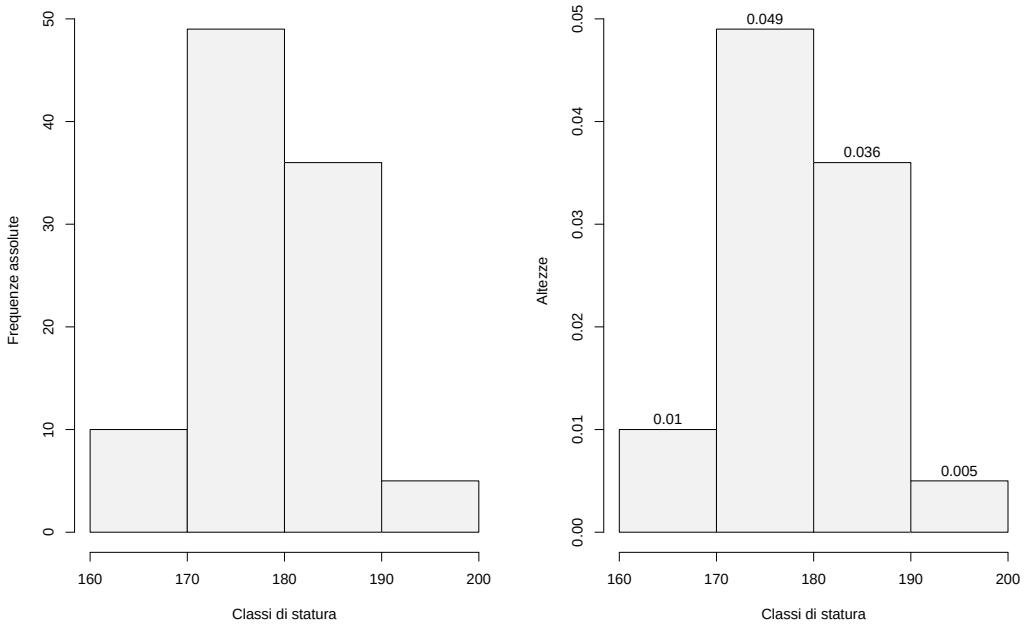


Figura 3.6 Istogrammi, esempio 3.15.

▷ ESEMPIO 3.16

Data una v.s. di tipo continuo, ci prefiggiamo ora il problema di determinare un congruo numero di classi, posto che esse abbiano modulo costante.
Evidentemente potremmo andare ad occhio, cioè per tentativi successivi, imponendo dapprima un esiguo numero di classi e aggiungendovene via via una sino a pervenire ad un compromesso tra "poche" e "troppe" classi, aggettivi che riflettono rispettivamente le antitetiche situazioni di poche e troppe informazioni.
La rappresentazione mediante istogramma delle distribuzioni derivanti da ciascuna prova può essere un valido aiuto "per l'occhio".
Gli istogrammi proposti in figura (3.7) derivano da quattro particolari scelte sul numero di classi di una ipotetica v.s. di tipo continuo. Possiamo da questi dedurre che il numero congruo di classi potrebbe essere ricercato tra quello del secondo e del terzo istogramma.
Se ricorressimo alla già citata regola di Sturges, secondo il quale il numero delle classi corrisponde all'intero più prossimo a $1+log_2(n)$, otterremmo $1+log_2(200) = 8.644$ e dunque il congruo numero di classi risulterebbe 9. Va da sè che tale algoritmo

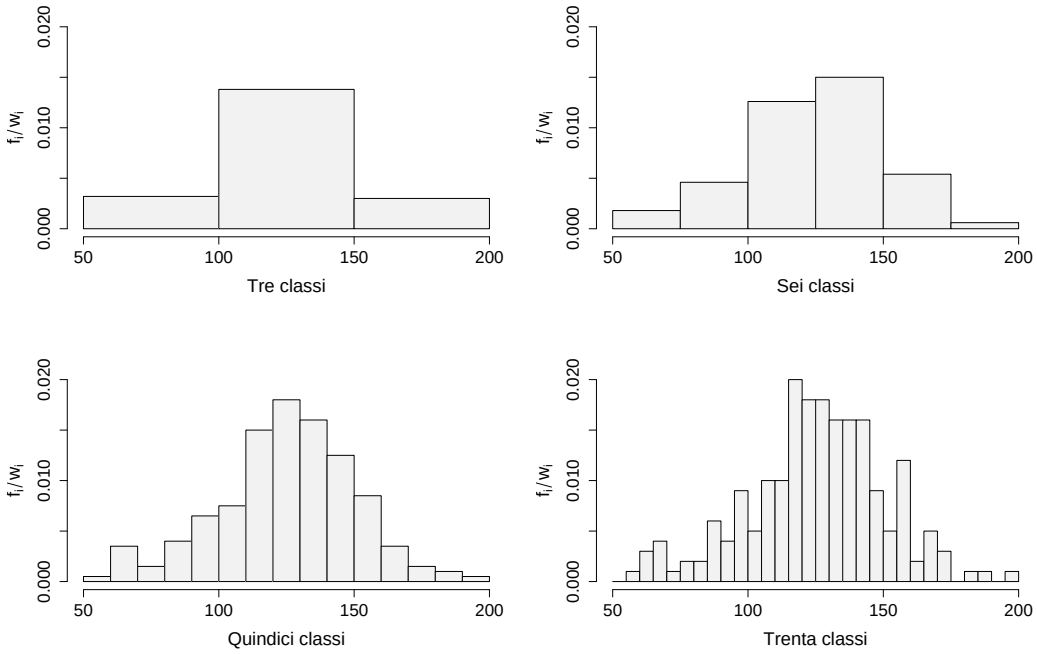


Figura 3.7 Istogrammi con diverso numero di classi, esempio 3.16.

non tiene conto che del numero di unità statistiche sulle quali si è rilevato il carattere in esame e non già dei valori assunti dalla v.s. che ne è scaturita, pertanto non sempre restituisce risultati ottimali.

◁

3.5. CON IL FOGLIO ELETTRONICO

Vediamo come poter ottenere la distribuzione di frequenze e costruire alcuni grafici utilizzando il foglio elettronico.
Supponiamo di lavorare con la matrice dei dati presentata nel capitolo precedente relativa ai 1100 laureati e di voler ottenere la distribuzione di frequenze della m.s. $A = \{\text{sesso}\}$ così come in figura (3.8)
Per ottenere 631, cioè la frequenza assoluta associata alla modalità “Maschio”, nella cella `H6` si è inserita la funzione

```
=CONTA.SE(B2:B1101;G6)
```

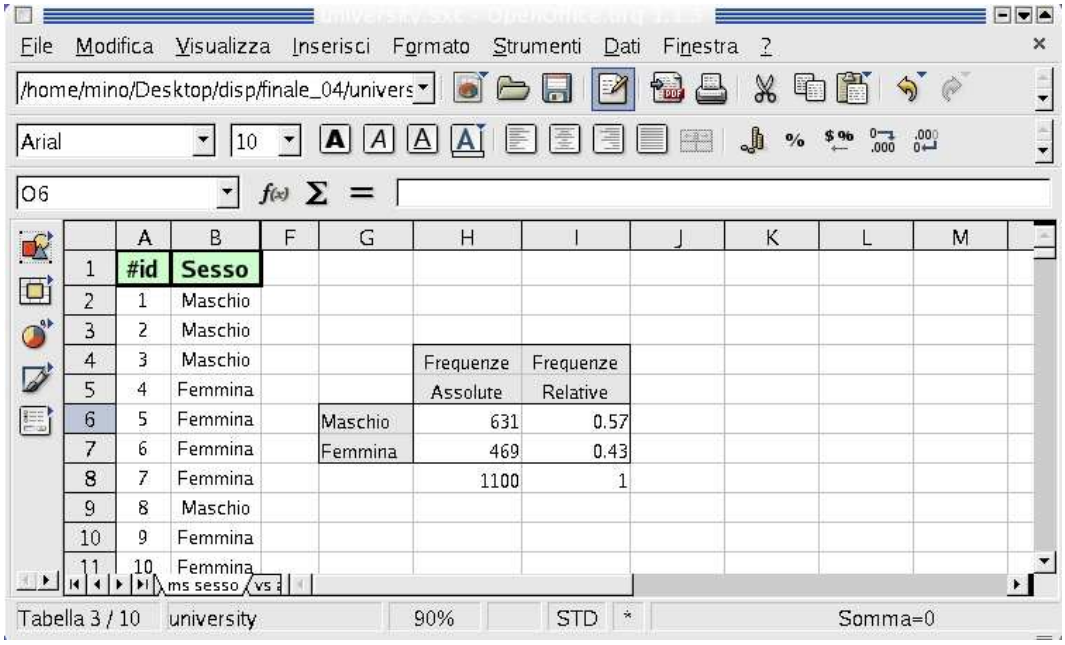


Figura 3.8 Distribuzione di frequenze della mutabile *sesso.*

la funzione restituisce il numero dei valori uguali al contenuto della cella `G6` tra i dati dell'intervallo di celle da `B2` a `B1101`.
Ovviamente, 469 della cella `H7` è il risultato restituito dalla funzione

```
=CONTA.SE(B2:B1101;G7)
```

e fornisce la frequenza assoluta associata alla modalità "Femmina".
Nella cella `H8` abbiamo inserito la semplice formula `=H6+H7` che ci dà la numerosità del collettivo statistico e che utilizziamo per il calcolo delle frequenze relative. Nella cella `I6` compare il risultato di `=H6/$H$8` mentre 0.43 deriva da `=H7/$H$8`.

Diversa è la funzione impiegata per determinare la distribuzione di frequenze della variabile statistica "Anni dalla laurea" (cfr. figura 3.9), in questo caso abbiamo utilizzato la funzione `FREQUENZA`. Tale funzione si dice, nel linguaggio dei fogli elettronici, funzione di matrice perché opera su più celle nel senso che deve essere inserita in più di una cella in quanto restituisce le frequenze assolute associate a tutte le modalità distinte della v.s. nello stesso momento. Nel nostro esempio abbiamo selezionato tutte le celle dell'intervallo `H4:H7` e abbiamo inserito la funzione:

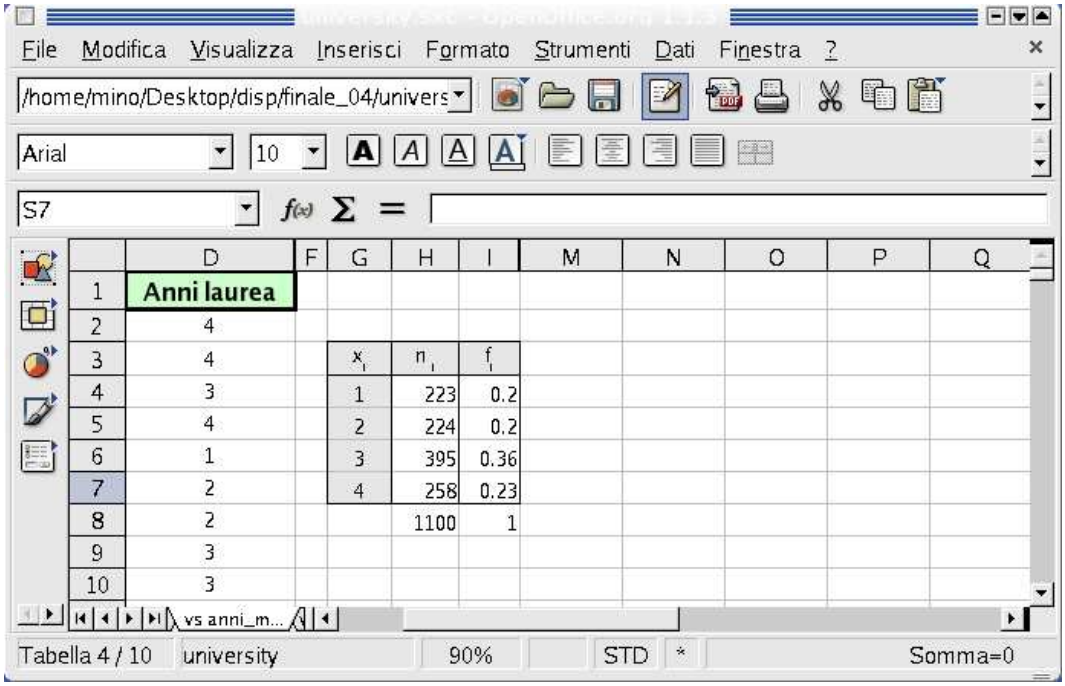
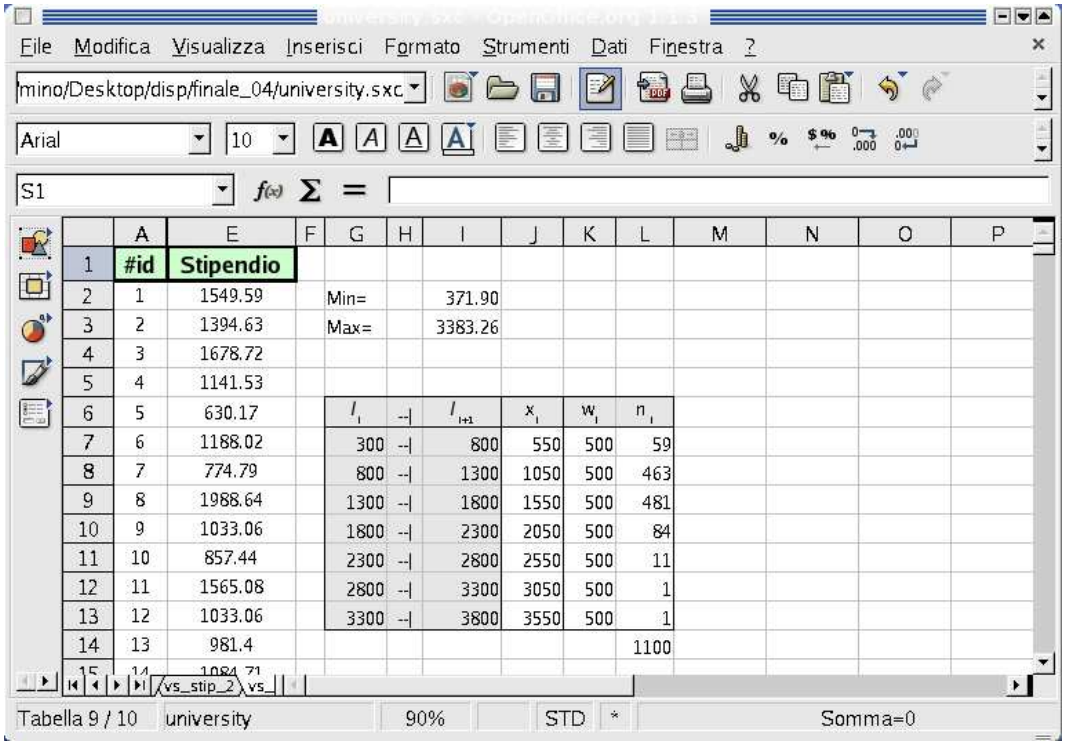


Figura 3.9 Distribuzione di frequenze della variabile *anni dalla laurea.*

```
{ =FREQUENZA(D2:D1101;G4:G7) }
```

Si noti che la funzione inserita è racchiusa tra parentesi graffe; queste sono necessarie perché vengano correttamente riempite tutte le celle dell'intervallo `H4:H7` nelle quali è stata inserita la funzione. Per ottenere tali parentesi graffe è sufficiente, una volta digitato `=FREQUENZA(D2:D1101;G4:G7)`, chiudere il comando non con il solo tasto invio ma con la sequenza di tasti, premuti contemporaneamente, control, maiuscolo e invio (Ctrl+Shift+invio).
In tale modo la funzione restituisce nella cella `H4` il numero dei valori che risultano minori o uguali a quello che si trova nella cella `G4` tra quelli dell'intervallo di celle `D2:D1101`. Nella cella `H5` restituisce il numero di valori maggiori di quello della cella `G4` e minori o uguali di quello in cella `G5` e così via. Data la natura discreta della nostra v.s., il 223 che si trova nella cella `H4` corrisponde al numero di unità statistiche che assumono modalità esattamente uguale a 1; il 224 che si trova nella cella `H5` corrisponde al numero di unità statistiche che assumono modalità esattamente uguale a 2 e così via.
Per quanto riguarda la variabile stipendio è necessario raccogliere i dati in classi (cfr. figura 3.10). Per determinare il numero e l'ampiezza delle classi iniziamo a valutare il



	A	E	G	H	I	J	K	L
1	#id	Stipendio						
2	1	1549.59	Min=		371.90			
3	2	1394.63	Max=		3383.26			
4	3	1678.72						
5	4	1141.53						
6	5	630.17	l_i	--	l_{i+1}	x_i	w_i	n_i
7	6	1188.02	300	--	800	550	500	59
8	7	774.79	800	--	1300	1050	500	463
9	8	1988.64	1300	--	1800	1550	500	481
10	9	1033.06	1800	--	2300	2050	500	84
11	10	857.44	2300	--	2800	2550	500	11
12	11	1565.08	2800	--	3300	3050	500	1
13	12	1033.06	3300	--	3800	3550	500	1
14	13	981.4						1100

Figura 3.10 Distribuzione di frequenze della v.s. *stipendio* -sette classi-.

campo di variabilità degli stipendi e nelle celle `I2` e `I3` inseriamo rispettivamente le funzioni `=min(E2:E1101)` e `=max(E2:E1101)` che restituiscono il minimo (371.90) ed il massimo (3383.26) della colonna degli stipendi.

Decidiamo in un primo momento di dividere i dati in 7 classi di ampiezza 500 a partire da 300. Le classi sono indicate nelle colonne `G`, `H` e `I` dalla riga 7 alla 13, la scelta di utilizzare tre colonne per ciascuna classe è dettata dall'esigenza di considerare i limiti inferiore e superiore quali numeri che possono essere utilizzati dalle funzioni del foglio elettronico. Infatti, ad esempio, nella colonna `J7` compare il centro della prima classe che è il risultato della funzione `=(I7+G7)/2` e così via per le righe successive, mentre nella cella `K7` vi è l'ampiezza di classe calcolata mediante `=I7-G7`. Per ottenere le frequenze assolute associate a ciascuna classe facciamo nuovamente ricorso alla funzione `FREQUENZA`, selezionando l'intervallo di celle `L7:L13` vi inseriamo la funzione:

```
{ =FREQUENZA(E2:E1101;I7:I13) }
```

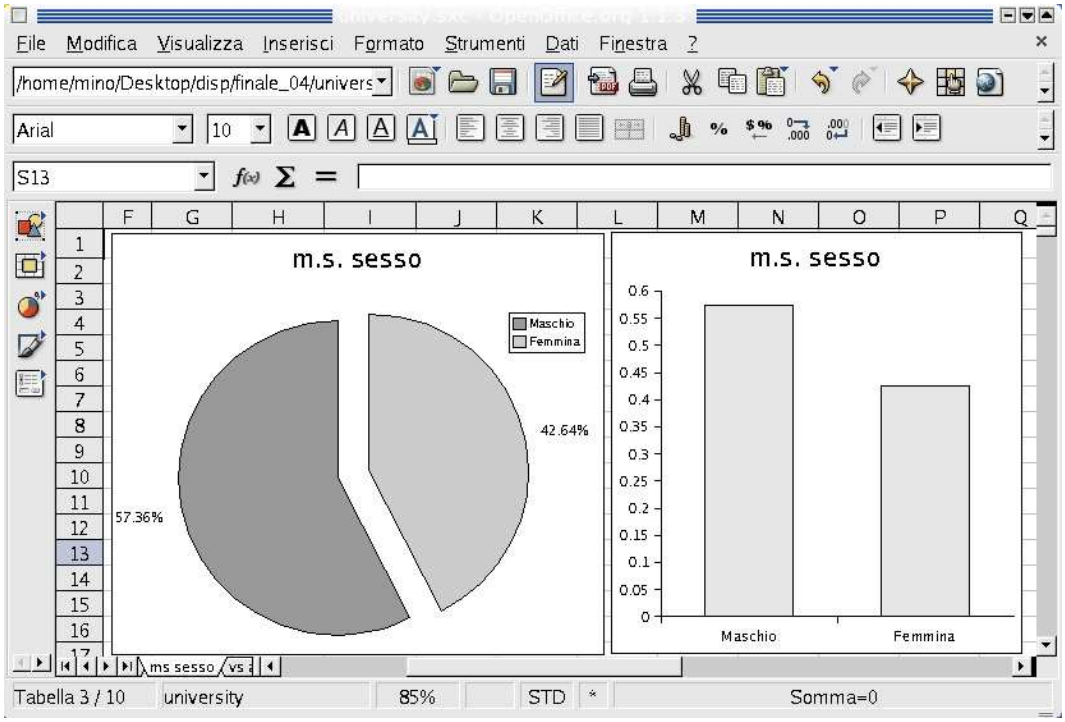


Figura 3.11 Diagrammi a torta e a barre per la m.s. *sesso*

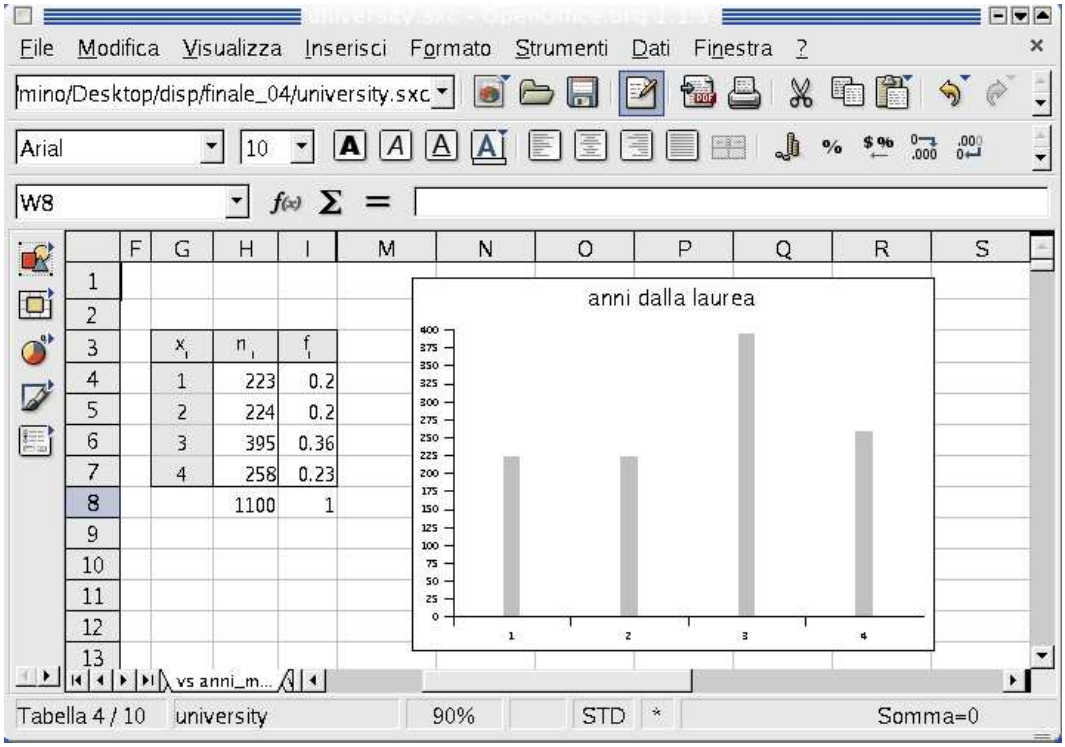
Si invita il Lettore a ripercorrere i passi sopra descritti in modo da riprodurre la tabella di frequenze di figura (3.14) proposta più avanti allorchè tratteremo il problema dell'istogramma con dati raccolti in classi a modulo non costante.

Ora che si sono ottenute le distribuzioni di frequenza delle m.s. e delle v.s. presenti nel file, vediamo brevemente quali sono i grafici opportuni per ciascuna di esse.

Per la m.s. $A = \{\text{sesso}\}$ otteniamo il diagramma a torta oppure il diagramma a barre di figura (3.11).

Entrambi i grafici si sono ottenuti, dopo aver selezionato le celle `G6:G7` e `I6:I7` contenenti rispettivamente le modalità e le frequenze relative (cfr. figura 3.8), usando il menù `inserisci --> diagramma`. Scelto il tipo di grafico tra quelli proposti si è indicato nella prima videata di autocomposizione del grafico di voler usare la `prima colonna come dicitura` e in quelle successive si sono compilati i campi riguardanti titoli, legenda, assi ecc. così come appaiono in figura.

Per ciò che riguarda la v.s. *anni dalla laurea*, secondo quanto indicato nel paragrafo pre-



x_i	n_i	f_i
1	223	0.2
2	224	0.2
3	395	0.36
4	258	0.23
	1100	1

Figura 3.12 Rappresentazione grafica della v.s. *anni dalla laurea*

cedente si dovrebbe usare un diagramma a bastoncini. Sfortunatamente i fogli elettronici non contemplano tra i grafici disponibili quello adatto per una v.s. di tipo discreto. Si può sopperire a tale mancanza usando, così come in figura (3.12), un diagramma a barre avendo l'accortezza di distanziare le barre e renderle il più piccolo possibile.

Più complessa è la costruzione dell'istogramma per la v.s. *stipendio*. Anche in questo caso i fogli elettronici non forniscono nella loro rassegna grafica ciò che è stato definito istogramma e sarà necessario pertanto adattare il solito diagramma a barre alle esigenze della definizione statistica di istogramma.
Se la distribuzione di frequenze è stata raccolta in classi di modulo costante, si ottiene facilmente un istogramma da un diagramma a barre semplicemente cliccando sulle barre e scegliendo tra le opzioni `distanza` (tra le barre) `uguale a zero`.
L'istogramma presentato in figura (3.13) è quello della v.s. *stipendio* con i dati raccolti in sette classi di ampiezza costante pari a 500; si noti che si è scelto di usare come altezza dei rettangoli la frequenza assoluta di ciascuna classe che in questo caso è consentito dalla

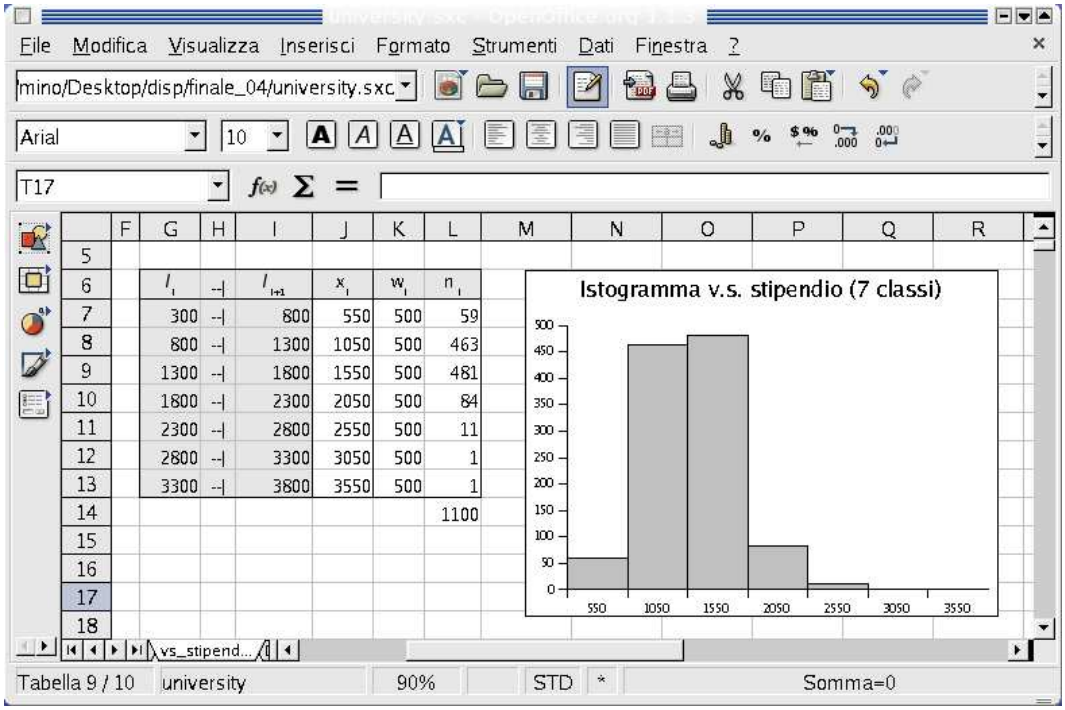


Figura 3.13 Distribuzione di frequenze della v.s. *stipendio* -sette classi-.

definizione di istogramma.
Per la costruzione del grafico si sono selezionate le celle `J7:J13` dei centri di classe (da utilizzare quali etichette dell'asse delle categorie) contemporaneamente alle celle `L7:L13` delle frequenze assolute e si sono opportunamente compilate i campi richiesti dalle videate di autocomposizione del diagramma a barre.

Per concludere il paragrafo presentiamo in figura (3.14) un esempio di istogramma nel caso di classi di modulo non costante. In questo caso è d'obbligo costruire rettangoli che abbiano base pari all'ampiezza di classe e altezza la frequenza (relativa o assoluta) diviso l'ampiezza di classe.
Poiché i rettangoli del diagramma a barre fornito dal foglio elettronico hanno la base uguale, siamo ricorsi ad a un piccolo stratagemma al fine di riprodurre un grafico che mostri rettangoli di ampiezza diversa. L'istogramma di figura (3.14) è stato ottenuto richiedendo ad Open Office un diagramma a barre sulla base dei dati che si trovano nell'intervallo di celle `I14:J20`.
Osserviamo che per le prime tre classi i rettangoli devono avere base uguale (per tutte il

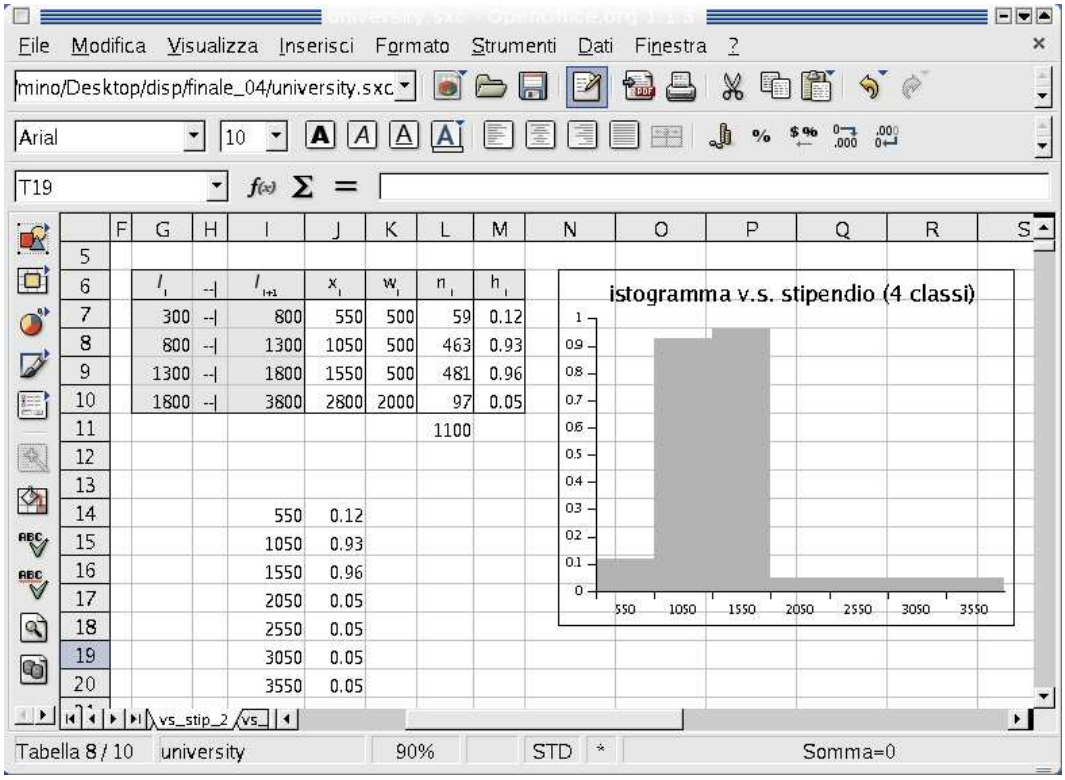


Figura 3.14 Distribuzione di frequenze della v.s. *stipendio* -cinque classi-.

modulo è 500) mentre per l'ultima il rettangolo dovrà avere base quattro volte più grande delle precedenti (il modulo di classe è infatti 2000).
Lo stratagemma adottato consiste nell'aver riportato l'altezza dell'ultima classe quattro volte nelle celle `J17:J20` in modo che i sette rettangoli restituiti dal diagramma a barre del foglio elettronico appaiano, eliminando il bordo delle barre, in effetti soltanto quattro poiché gli ultimi quattro hanno la stessa altezza.

3.6. ESERCIZI

▷ ESERCIZIO 3.1

Si immagini di avere rilevato lo *stato civile* dei $n = 100$ maschi residenti in un piccolo comune alpino e che le singole osservazioni abbiano dato luogo alla m.s. A con dati individuali:

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

dove, per praticità, è si posto:

$1 \rightarrow$ *celibe* $\quad 2 \rightarrow$ *coniugato* $\quad 3 \rightarrow$ *convivente* $\quad 4 \rightarrow$ *divorziato* $\quad 5 \rightarrow$ *vedovo*

Si proceda all'individuazione delle corrispondenti *distribuzioni di frequenze assolute e relative* e se ne dia una opportuna *rappresentazione grafica.*

◁

▷ ESERCIZIO 3.2

Si immagini che le misurazioni, in mm, della lunghezza dei chiodi in acciaio, contenuti in una confezione di $n = 20$ unità, abbiano dato luogo alla v.s. X con valori individuali:

12.4	11.3	11.6	9.5	13.6	10.9	11.5	15.5	11.6	10.3
10.1	10.3	9.8	12.6	11.5	11.1	10.7	11.7	11.1	12.2

si individui la distribuzione di frequenze assolute e relative della m.s. A definita come:

$$A \equiv \begin{cases} \text{difettoso} & \text{se } X(\omega_\alpha) < 10 \text{ mm} \\ \text{conforme} & \text{se } 10 \leq X(\omega_\alpha) \leq 12 \text{ mm} \\ \text{rettificabile} & \text{se } X(\omega_\alpha) > 12 \text{ mm} \end{cases}$$

◁

▷ ESERCIZIO 3.3

La misurazione del peso corporeo di un gruppo di coscritti al servizio di leva ha dato luogo alla v.s. X con valori individuali, espressi in Kg:

64.5	59.7	58.9	61.5	70.5	88.4	70.1	72.5	58.7	62.5
42.5	42.7	70.0	41.5	52.4	68.3	70.2	77.3	60.0	82.5
69.4	68.8	42.5	52.7	61.8	51.2	59.5	67.4	49.9	75.6
70.5	71.5	87.6	80.5	79.5	69.7	51.3	78.2	79.9	88.6

Con riferimento alla v.s. X, si completi la seguente tabella di frequenza, ove i pesi sono stati raccolti in classi di modulo costante:

Classi di peso	n_i	f_i	*Centri di classe* x_i
40 - 50	.....	.....	.....
50 - 60	.....	.....	.....
60 - 70	.....	.....	.....
70 - 80	.....	.....	.....
80 - 90	.....	.....	.....
n	.....		

◁

▷ ESERCIZIO 3.4

Con riferimento alla distribuzione di frequenze di cui all'esercizio 3.3, indicare se l'ipotesi di equidistribuzione dei dati all'interno di ciascuna classe di peso proposta può ritenersi "ragionevole". In caso contrario, proporre una diversa suddivisione in classi.

◁

▷ ESERCIZIO 3.5

La misurazione della lunghezza (in metri) delle matasse di fili elettrici giacenti in un magazzino ricambi, ha fornito i seguenti risultati:

165	167	172	177	167	165	162	178	175	170
165	171	162	174	180	185	191	181	165	170
190	195	184	180	174	175	167	165	161	162
163	168	165	173	177	167	185	169	182	183
185	175	180	172	173	164	182	171	180	162

Si completi, in modo opportuno la seguente tabella e si rappresentino i dati mediante istogramma.

Classi di Lunghezza	centro di classe	freq. assolute	freq. relative	ampiezza classe	altezze istogramma
$160 \dashv 165$	...	...	...	...	...
$165 \dashv 170$	...	...	...	...	...
$170 \dashv 180$	...	...	...	...	...
$180 \dashv 195$	...	...	...	...	...

◁

▷ ESERCIZIO 3.6

Su un collettivo statistico di $n = 30$ unità, i dipendenti dell'azienda WHY, si sono rilevati i seguenti caratteri: sesso, titolo di studio, anzianità in anni di servizio e reddito netto mensile. I dati individuali sono stati organizzati nella seguente matrice dei dati:

sesso	*titolo di studio*	*anzianità di servizio (in anni)*	*reddito netto mensile (milioni di lire)*
M	5	10	1478.71
F	1	10	933.92
M	3	15	1333.43
M	1	10	985.80
F	1	5	933.92
M	2	15	1193.34
M	4	7	1141.46
M	1	10	1089.57
F	1	10	933.92
M	5	10	1338.62
M	1	10	933.92
F	1	5	933.92
M	2	10	1089.57
M	1	7	1037.69
M	2	13	1011.75
M	2	12	985.80
F	2	4	778.27
M	1	10	1011.75
M	1	10	1400.88
F	2	10	1089.57
M	5	15	1489.08
M	1	10	933.92
F	5	12	1297.11
M	2	10	1089.57
M	1	7	1037.69
M	2	13	985.80
M	2	12	1022.12
F	5	4	1089.57
M	1	10	985.80
F	4	8	1089.57

dove per il carattere *titolo di studio* si è scelto di codificare le sue diverse modalità *Licenza elementare*, *Licenza Media inf.*, *Licenza Media sup.*, *Diploma breve* e *Laurea*

con gli interi positivi da 1 a 5.
Si individuino le distribuzioni di frequenze assolute e relative delle m.s. A={*sesso*}, B={*titolo di studio*}, X={*anzianità di servizio*} e Y={*reddito netto mensile*}

◁

▷ ESERCIZIO 3.7

Completare la seguente tabella che riporta la distribuzione del titolo di studio posseduto dai dipendenti di una certa azienda:

Modalità	*Freq. Assolute*	*Freq. Relative*
Licenza elementare	25	...
Licenza Media inf.	42	...
Licenza Media sup.	123	...
Diploma breve	5	...
Laurea	15	...

Rappresentare, inoltre, tale distribuzione mediante:

(a) un diagramma a "torta" (b) un diagramma a canne d'organo.

◁

▷ ESERCIZIO 3.8

La tabella che segue riporta la distribuzione di 100 lotti di untà omogenee prodotti in un certo giorno a seconda del numero di unità difettose ivi rinvenute:

Nůnità difettose	*Freq. Assolute*	*Freq. Relative*
0	42	...
1	21	...
2	14	...
3	9	...
4	7	...
5	4	...
6	2	...
7	1	...

Rappresentare tale distribuzione di frequenze mediante il diagramma che si ritiene più idoneo.

◁

▷ ESERCIZIO 3.9

Gli incidenti stradali avvenuti nelle province del piemonte nel corso dell'anno 1982 sono risultati (Fonte: Statistica degli incidenti stradali – Istat 1982 –):

Torino	5.129	Vercelli	1.134	Novara	2.723
Cuneo	2.196	Asti	1.259	Alessandria	2.740

In relazione ai dati della tabella, prima di rappresentarli graficamente con il diagramma che si ritiene più idoneo, descrivere il collettivo statistico, le unità statistiche ed il tipo di carattere che ha consentito la rilevazione

◁

▷ ESERCIZIO 3.10

Si immagini che la v.s. X, che rappresenta l'ammontare delle puntate giornaliere espresse in euro di 40 frequentatori abituali di un Casinò, abbia assunto valori individuali:

94	86	86	86	86	86	105	105	105	119
75	75	76	76	75	76	78	78	78	78
80	80	75	86	80	80	80	86	80	80
94	72	80	94	72	72	72	80	75	80

Si calcoli l'ammontare totale delle puntate basandosi:

a — sui dati individuali;

b — sulla distribuzione di frequenze assolute;

c — sulla distribuzione di frequenze assolute raccogliendo i dati in classi di modulo costante e pari a 10 euro e limite inferiore della prima classe pari a 70.0 euro.

◁

CAPITOLO 4

LA FUNZIONE DI RIPARTIZIONE E I QUANTILI

Per analizzare sempre più approfonditamente la distribuzione di frequenze di una variabile statistica, diamo in questo capitolo la definizione di funzione di ripartizione e di quantile. La costruzione del grafico della funzione di ripartizione, la valorizzazione delle sue peculiari proprietà nonché l'interpretazione del suo significato verranno proposti mediante alcuni esempi di specie. Il metodo di calcolo dei quantili sarà illustrato per il caso di dati individuali e di distribuzione di frequenze, anche qualora i dati siano raccolti in classi.

4.1. LA FUNZIONE DI RIPARTIZIONE DEFINIZIONE E PROPRIETÀ

La funzione di ripartizione cumulativa delle frequenze, in simboli $F_X(x)$, detta in breve funzione di ripartizione, è uno degli strumenti cardine della statistica descrittiva; essa costituisce un'alternativa alla distribuzione di frequenze relative di una variabile statistica e consente di individuare alcune sue misure di sintesi.
La funzione di ripartizione è definita su tutto l'asse reale e mette in relazione i valori che la v.s. potenzialmente può assumere con la proporzione di unità statistiche che effetivamente hanno assunto valori minori o uguali ad essi.
Premesso che non sarà possibile pervenire ad una forma analitica della funzione di ripartizione, prima di dare una definizione formale vediamo come esprimere in simboli la frequenza relativa con cui si osservano valori inferiori o uguali ad un qualunque numero reale.
Data una v.s. X con distribuzione di frequenze relative $\{x_i, f_i\}_{i=1,\dots,k}$, fissato $x \in \mathbb{R}$, indichiamo con $\sum_{x_i \leq x} f_i$ la somma delle frequenze relative associate a ciascuna modalità x_i di valore non superiore a x.
Tale quantità verrà manifestamente a dipendere dal valore x prescelto e, naturalmente, dalle modalità assunte dalla v.s. X.
Estendendo il concetto a tutti i valori dell'asse reale è possibile dare la seguente

Definizione 4.1 (Funzione di ripartizione)
data una v.s. X con distribuzione di frequenze relative $\{x_i, f_i\}_{i=1,\dots,k}$, per ogni $x \in \mathbb{R}$ definiamo funzione di ripartizione cumulativa delle frequenze l'insieme delle infinite coppie $(x; F_X(x))$ con

$$F_X(x) = \sum_{x_i \leq x} f_i \tag{4.1}$$

□

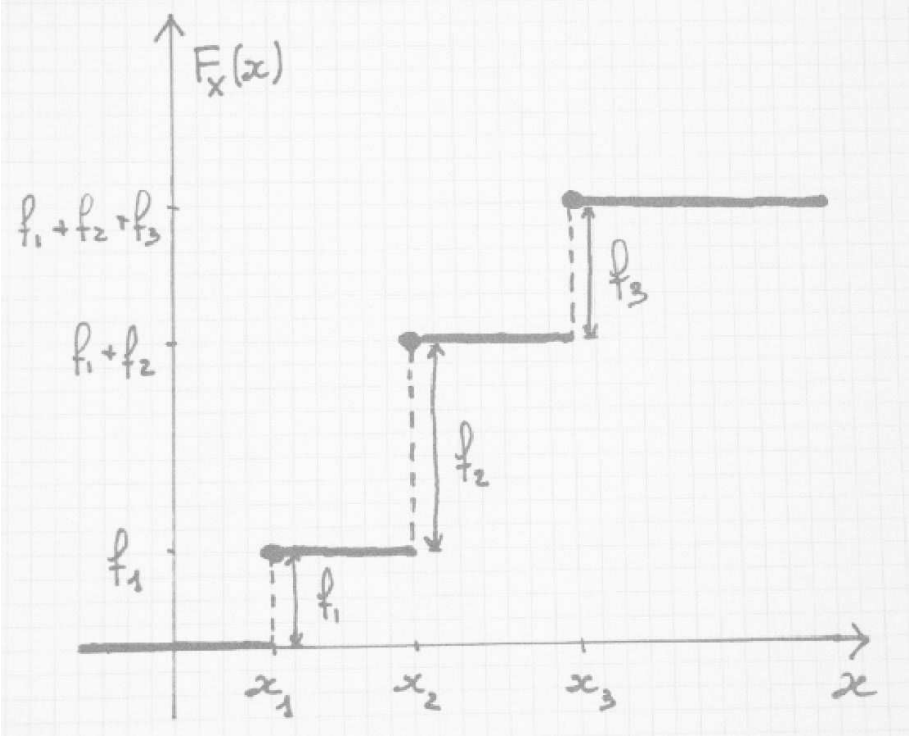


Figura 4.1 Funzione di ripartizione.

Proviamo ora a costruire la funzione di ripartizione, basandoci sulla definizione appena data, per una generica v.s. X che assume $k = 3$ modalità distinte cosicché la sua distribuzione di frequenze relative risulta

$$X \equiv \begin{Bmatrix} x_i \\ f_i \end{Bmatrix}_{i=1,\dots,k} = \begin{Bmatrix} x_1 & x_2 & x_3 \\ f_1 & f_2 & f_3 \end{Bmatrix}$$

Per qualsiasi x appartenente all'asse reale siamo in grado di individuare il valore di $F_X(x)$ applicando la definizione; infatti:

- $\star$ $\forall x < x_1 \Rightarrow \qquad F_X(x) = \sum_{x_i \leq x} f_i = 0$

  nessuna unità di Ω è stata associata da X a valori minori o uguali di x;

- $\star$ $\forall x_1 \leq x < x_2 \Rightarrow \qquad F_X(x) = \sum_{x_i \leq x} f_i = f_1$

  ovvero una proporzione pari a f_1 unità di Ω è stata associata da X a valori minori o uguali di x;

- $\star$ $\forall x_2 \leq x < x_3 \Rightarrow \qquad F_X(x) = \sum_{x_i \leq x} f_i = f_1 + f_2$

  ovvero una proporzione pari a $f_1 + f_2$ unità di Ω è stata associata da X a valori più piccoli di x;

- $\star$ $\forall x \geq x_3 \Rightarrow \qquad F_X(x) = \sum_{x_i \leq x} f_i = f_1 + f_2 + f_3 = 1$

  tutte le unità di Ω sono state associate da X a valori più piccoli di x.

La rappresentazione grafica della funzione $F_X(x)$ di tale v.s. è data in figura (4.1).

Direttamente dalla definizione (4.1) ricaviamo che, qualunque sia la v.s. X, la funzione di ripartizione $F_X(x)$ soddisfa le seguenti proprietà:

- $\star$ ha dominio l'intero asse reale, è definita cioè per ogni valore $x \in \mathbb{R}$;
- $\star$ ha codominio l'intervallo $[0; 1]$, assume, cioè, solo valori compresi tra zero ed uno;
- $\star$ è una funzione monotona non decrescente, cioè $\forall \; x, x' \in \mathbb{R}$ se $x < x'$ allora $F_X(x) \leq F_X(x')$;
- $\star$ è continua a tratti e possiede k punti di discontinuità (in corrispondenza delle modalità x_i) nei quali risulta continua a destra e compie un salto di ampiezza f_i;
- $\star$ è nulla per ogni x minore della più piccola modalità osservata e vale uno per valori di x maggiori (o uguali) alla più grande modalità osservata. In simboli:

$$F_X(x) = \begin{cases} 0 & \text{per } x < \min_i(x_i) \\ \sum_{i=1}^{k} f_i = 1 & \text{per } x \geq \max_i(x_i) \end{cases}$$

4.2. ESEMPI DI FUNZIONI DI RIPARTIZIONE

In questo paragrafo verranno presentati due semplici esempi di costruzione della funzione di ripartizione e della sua rappresentazione grafica.

▷ ESEMPIO 4.1

Consideriamo la v.s. $X = \{\textit{numero di telefonate}\}$ ricevute da 44 operatori di un call center in una data giornata di gennaio. Usiamo una tabella per indicare la sua distribuzione di frequenze relative ed a essa aggiungiamo una terza colonna nella quale riportiamo la somma cumulata delle frequenze relative f_i, ottenendo:

x_i	f_i	$F(x_i)$
90	0.4	0.4
95	0.2	0.6
100	0.3	0.9
105	0.1	1

Abbiamo in tal modo valorizzato la $F_X(x)$ nei suoi punti di salto e possiamo procedere alla costruzione del suo grafico ponendo sulle ascisse le modalità x_i in corrispondenza alle quali segnamo immediatamente in ordinata i valori della $F_X(x_i)$, così come evidenziato in figura (4.2, a). Completiamo il grafico tracciando un tratto orizzontale ad ordinata $F_X(x_i)$ per tutti gli intervalli $[x_i; x_{i+1}[$. Ponendo uguale ad uno la $F_X(x)$ per tutti i valori maggiori di 105 ed a zero per tutti i valori minori di 90 si realizza, infine, il tipico grafico a scala proposto in figura (4.2, b).
Se già dalla seconda riga della tabella avremmo potuto ricavare $F_X(95) = 0.6$ e affermare che il 60% degli operatori ha ricevuto al più 95 telefonate, tuttavia il grafico della funzione di ripartizione ci consente di rispondere immediatatmente a domande del tipo: "Quanti operatori hanno ricevuto al più 98 telefonate?"
Osserviamo ancora che, per x uguale, ad esempio, a 87 risulta $F_X(87) = 0$, infatti nessun operatore ha ricevuto meno di ottantasette telefonate, mentre per $x = 107$ risulta $F(107) = 1$, infatti tutti gli operatori hanno ricevuto meno di centosette telefonate.

◁

▷ ESEMPIO 4.2

I dati che seguono si riferiscono al consumo energetico bimensile (espresso in kWh) risultante dalle bollette di 40 famiglie lombarde:

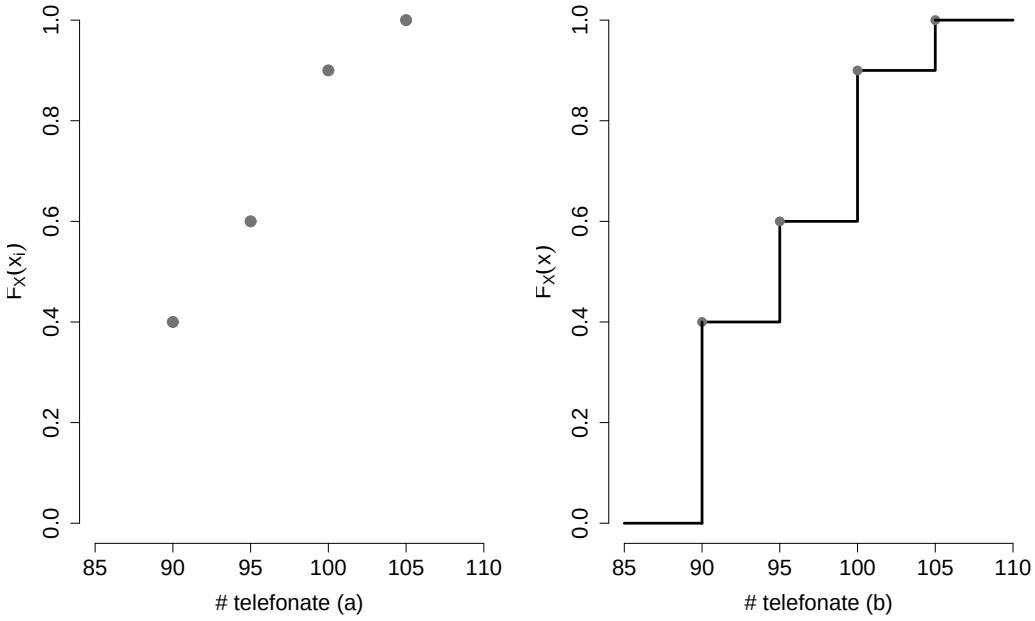


Figura 4.2 Funzione di ripartizione n° di telefonate, esempio 4.1.

409.5	432.5	494.0	454.7	446.8	476.9	467.5	493.0	409.0	429.8
212.2	279.6	325.9	382.9	367.0	336.2	317.5	368.9	353.5	325.9
439.4	583.4	528.4	534.9	526.9	593.7	530.7	559.4	529.0	585.1
367.2	399.4	328.9	413.5	498.1	424.1	432.4	404.0	447.6	483.2

La v.s. $X = \{\textit{consumo energetico}\}$ assume in questo caso determinazioni tutte distinte (tutte le famiglie del collettivo in esame hanno consumo differente), la sua distribuzione di frequenze relative è perciò costituita da $k = n = 40$ modalità distinte ciascuna con frequenza relativa $f_i = 1/40 = 0.025$.

Per rappresentare la funzione di ripartizione è necessario porre inizialmente sull'asse delle ascisse i dati individuali posti in ordine crescente e sull'asse delle ordinate i corrispondenti valori delle frequenze cumulate (cfr. figura 4.3, a). Congiungendo i punti con segmenti orrizontali otteniamo la funzione a scala con $k = 40$ salti di altezza costante pari a 0.025 così come in figura (4.3, b).

$\triangleleft$

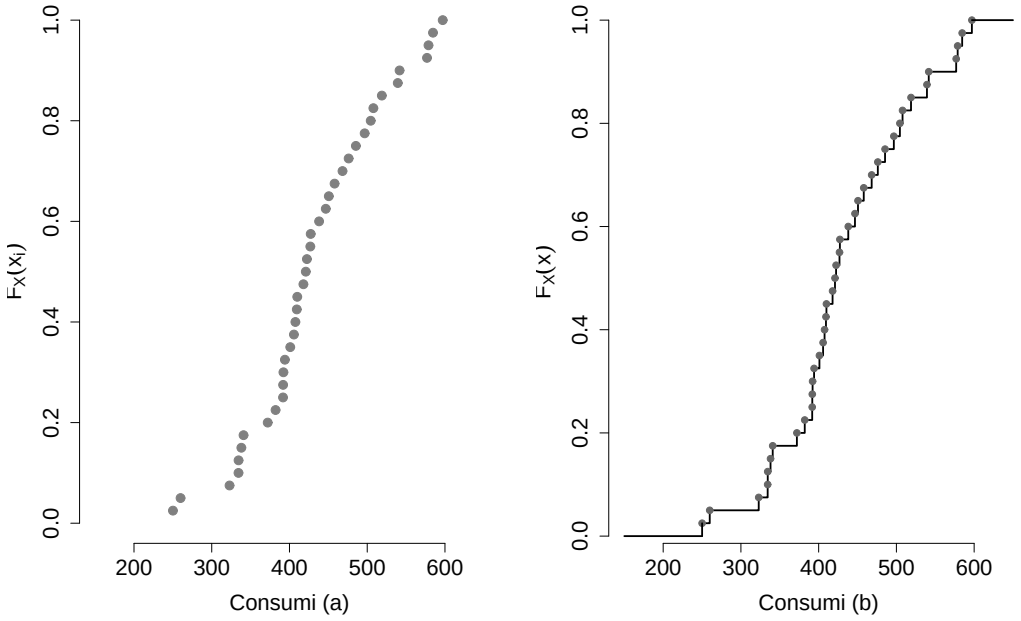


Figura 4.3 Funzione di ripartizione consumo energetico, esempio 4.2.

Osserviamo che nei capitoli successivi, qualora non sussistano ambiguità, per semplicità di notazione indicheremo il valore della funzione di ripartizione in corrispondenza di una qualsiasi modalità con F_i cioè varrà l'uguaglianza tra simboli $F(x_i) = F_i$.

4.2.1 Il caso di dati raccolti in classi

Se di una v.s. si dispone unicamente di dati raccolti in classi non è possibile costruire la funzione di ripartizione a partire dalla definizione (4.1). Avendo raccolto i dati individuali in classi si è persa l'informazione originaria del valore assunto da ciascuna unità statistica. In questa nuova situazione sappiamo che le n_i unità attribuite alla i-esima classe assumono valori appartenenti alla classe, ma non conosciamo il modo in cui esse sono "posizionate" all'interno della stessa. Ne consegue che siamo in grado di valorizzare la funzione di ripartizione unicamente nei limiti di classe, mentre nulla può dirsi del valore della stessa nei punti interni alle classi.
Per poter procedere occorre formulare qualche ulteriore ipotesi circa la distribuzione delle unità all'interno delle classi. La più naturale, già accennata nel capitolo precedente e che

adotteremo d'ora in avanti, è che le unità siano *distribuite in modo uniforme sull'intera classe*; sotto tale ipotesi la $F_X(x)$ cresce in modo lineare all'interno di ogni classe, ed il suo grafico assume l'aspetto di una curva continua spezzata.

▷ ESEMPIO 4.3

Consideriamo la v.s. $X = \{$*consumo energetico*$\}$, di cui all'esempio (4.2), e raccogliamo i dati individuali in 4 classi, in modo da ottenere la distribuzione di frequenze:

Classi di consumo $l_i \dashv l_{i+1}$	fr. ass. n_i	fr. rel. f_i	fr. cum. $F(l_{i+1})$
$200 \dashv 300$	2	0.050	0.050
$300 \dashv 400$	11	0.275	0.325
$400 \dashv 500$	18	0.450	0.775
$500 \dashv 600$	9	0.225	1.000

La colonna delle frequenze cumulate della tabella precedente fornisce il valore della funzione di ripartizione in corrispondenza al limite superiore di ogni classe (l_{i+1}).
Per rappresentare graficamente la $F_X(x)$ ricordiamo che essa fornisce per ogni $x \in \mathbb{R}$ la proporzione di unità del collettivo che assumono modalità non superiore a x, pertanto per ogni valore x minore di 200 (il limite inferiore della prima classe) avremo $F_X(x) = 0$. Ipotizzando che le 2 unità attribuite alla prima classe siano distribuite uniformemente fra i valori 200 e 300, la funzione $F_X(x)$ cresce, nella classe $]200; 300]$ in modo lineare dal valore zero al valore 0.050. Procedendo in modo analogo per le altre classi si ottiene il grafico di figura (4.4, a).
La funzione di ripartizione è in questo caso continua e risulta *un'approssimazione della vera* funzione di ripartizione. Per visualizzare come l'ipotesi di distribuzione uniforme dei dati all'interno di ciascuna classe abbia influenza sulla funzione di ripartizione si osservi la figura (4.4,b) nella quale sono messe a confronto le funzioni di ripartizione calcolate sui dati individuali e sugli stessi raccolti in classi e ciò per la classe $]300; 400]$.

◁

4.3. DEFINIZIONE DI QUANTILE

Le indagini statistiche hanno il compito di sintetizzare le molteplici informazioni contenute nella distribuzione di frequenze di una variabile statistica individuando i valori di alcuni parametri caratteristici che possano essere d'aiuto nella comprensione del carattere oggetto di studio. In questo paragrafo verranno illustrate le procedure necessarie

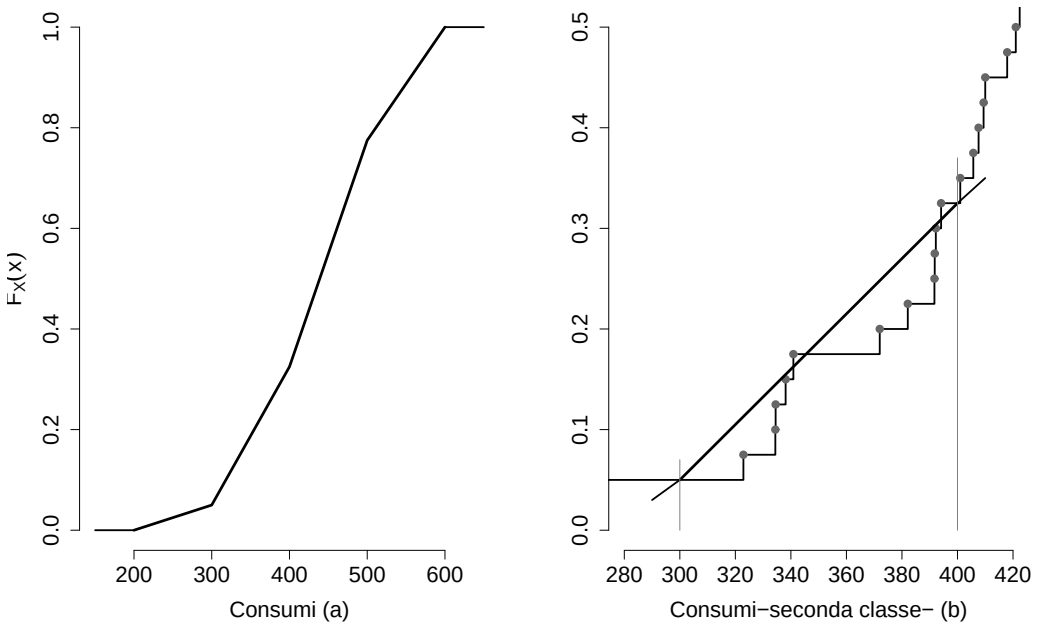


Figura 4.4 Funzione di ripartizione del consumo energetico, esempio 4.3.

per sintetizzare la distribuzione di frequenze di un v.s. attraverso la determinazione dei *quantili.*

Nel corso della vita di tutti i giorni il lettore potrebbe essersi inconsciamente già imbattuto nell'impiego dei quantili.

In occasione ad esempio di un'analisi del sangue il referto ospedaliero comunemente mostra, congiuntamente ai parametri riscontrati per il paziente in osservazione, dei valori "soglia" che sono appunto ciò che in statistica viene definito quantile. Se nel referto osserviamo, ad esempio, in corrispondenza al contenuto di glucosio la dicitura "valore di riferimento < 110 mg/dl" sappiamo che i medici, osservando numerosi pazienti sani, hanno stabilito il valore soglia (110 mg/dl) individuando il limite al di sotto del quale si colloca il contenuto di glucosio nel sangue di una "alta percentuale" dei soggetti osservati. Tale limite dipende ovviamente dalla scelta fatta circa la percentuale di soggetti da considerarsi a "norma", se ad esempio essa è pari a 95% diremo che 110 è il quantile di ordine 0.95 della v.s. contenuto di glucosio e potremmo affermare che il 95% dei soggetti sani ha nel sangue non più di 110 mg/dl di glucosio.

Più in generale, data una v.s. X il quantile di ordine α corrisponde al valore non superato

dal $\alpha\%$ delle unità statistiche.
Ricordando che la funzione di ripartizione mette in relazione i valori assunti da una v.s. con le frequenze cumulate è ovvio che per definire ed individuare i quantili sarà possibile fare ricorso ad essa. Infatti diamo la seguente

Definizione 4.2 (Quantili)
data una v.s. X, qualunque sia α scelto nell'intervallo $]0;1[$, si dicono quantili di ordine α le radici x_α dell'equazione

$$F_X(x) = \alpha \tag{4.2}$$

□

Prima di vedere in dettaglio come determinare le radici dell'equazione (4.2) osserviamo che per particolari valori dell'ordine α il quantile corrispondente viene in letteratura citato con un proprio nome.
Ad esempio i quantili di ordine $\alpha = 0.25$, $\alpha = 0.5$ e $\alpha = 0.75$ vengono detti rispettivamente *Primo, Secondo* e *Terzo Quartile*, nome dovuto alla caratteristica di $x_{0.25}$, $x_{0.5}$ e $x_{0.75}$ di essere quei valori che dividono in quattro parti la distribuzione della v.s.
Secondo lo stesso criterio, si dicono *Decili* i quantili di ordine α multiplo di 0.10 e *Percentili* i quantili di ordine α multiplo di 0.01.
I quantili rientrano tra le misure di posizione che saranno argomento del prossimo capitolo ove ne verranno evidenziate alcune caratteristiche peculiari. Come sarà evidenziato nelle applicazioni, una misura di posizione a cui sovente si farà ricorso è la *"Mediana"*, questa corrisponde al secondo quartile cioè a quel valore che divide in due parti uguali la distribuzione di frequenze (mediana = secondo quartile = $x_{0.5}$).

4.3.1 CALCOLO DEI QUANTILI

Ricordando che le radici di una generica equazione $g(x) = c$ si possono individuare dal grafico della funzione $g(x)$ come le ascisse dei punti di intersezione fra $g(x)$ e la retta $y = c$, nel caso dell'equazione (4.2) dovremo riferirci al grafico della funzione di ripartizione.
Qualunque sia la v.s., fissato $\alpha \in]0;1[$, possono verificarsi due situazioni: esistono *infinite soluzioni* oppure *non esiste alcuna soluzione.*

Si ha il primo caso quando, tracciando una linea parallela all'asse delle ascisse ad ordinata α questa interseca il grafico della funzione di ripartizione lungo tutta la "pedata di un suo gradino". Le radici dell'equazione (4.2) sono gli infiniti valori compresi fra le due modalità consecutive che delimitano il gradino (cfr. figura 4.5).

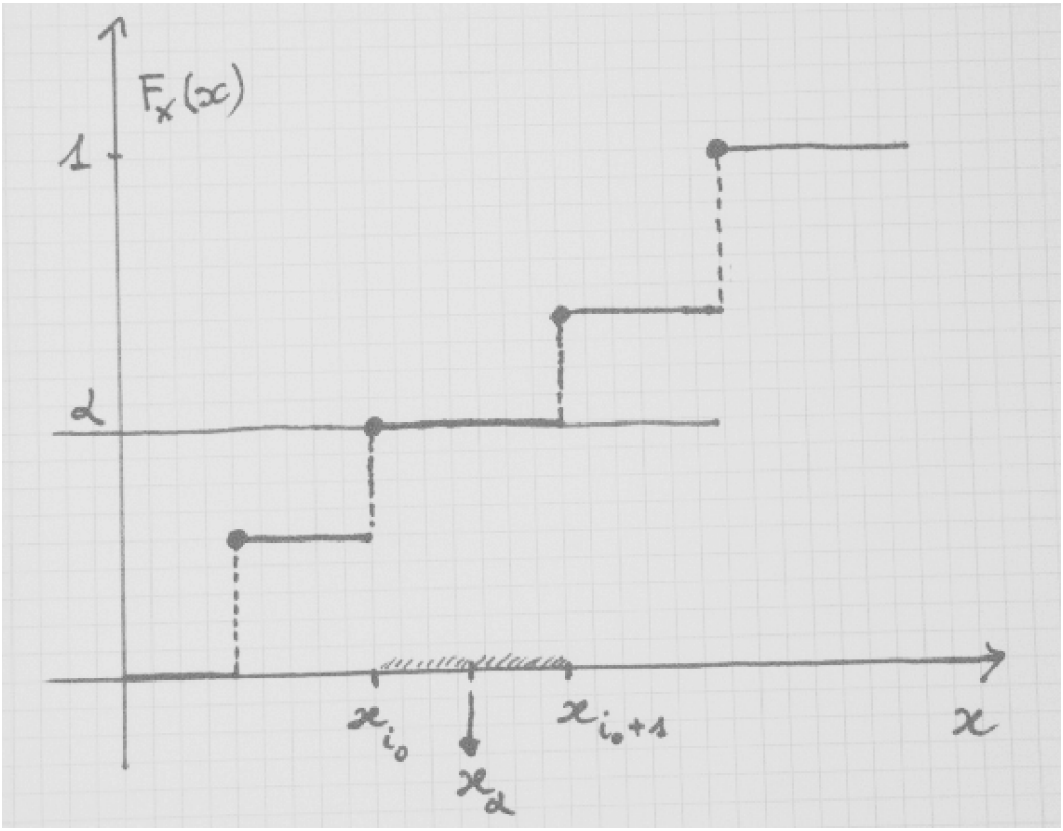


Figura 4.5 Infinite radici dell'equazione (4.2).

In altri termini, ciò significa che esiste un indice $i_0 \in (1, \ldots, k)$ per il quale risulta:

$$F_X(x_{i_0}) = \sum_{i=1}^{i_0} f_i = \alpha \tag{4.3}$$

le radici dell'equazione (4.2) sono gli infiniti valori $x \in [x_{i_0} \ ; \ x_{i_0+1}[$.
Per convenzione si assume come quantile x_α il valore centrale dell'intervallo cioè:

$$x_\alpha = \frac{x_{i_0} + x_{i_0+1}}{2}$$

▷ ESEMPIO 4.4

Si consideri un collettivo statistico formato dai 100 studenti di un corso di statistica e la v.s. $X = \{$*voto della prova scritta di statistica*$\}$ con distribuzione di frequenze:

x_i	n_i	f_i	$F_X(x_i)$
15	40	0.40	0.40
16	10	0.10	0.50
17	25	0.25	0.75
29	5	0.05	0.80
30	20	0.20	1

Si pensi di voler individuare il quantile di ordine $\alpha = 0.75$ (cioè il terzo quartile): è immediato notare che l'indice $i_0 = 3$ soddisfa la (4.3), infatti $F(x_3) = 0.75$, per cui il quantile corrispondente sarà:

$$x_{0.75} = \frac{x_3 + x_4}{2} = \frac{17 + 29}{2} = 23.0$$

Allo stesso risultato si perviene osservando il grafico della funzione di ripartizione presentato in figura (4.6, pannello a).
Come esercizio lasciamo al Lettore determinare il valore della mediana e interpretare il risultato della prova scritta.

◁

Il secondo caso si ha quando, tracciando una linea parallela all'asse delle ascisse ad ordinata α, questa incontra il grafico di $F_X(x)$ nell'"alzata" di un suo gradino (cfr. figura 4.7), per cui non esiste alcun valore che soddisfa l'equazione (4.2).
In altri termini, esiste un indice i_o per il quale si hanno simultaneamente:

$$\begin{aligned} F_X(x_{i_0-1}) &= \sum_{i=1}^{i_0-1} f_i < \alpha \\ F_X(x_{i_0}) &= \sum_{i=1}^{i_0} f_i > \alpha \end{aligned} \tag{4.4}$$

Per convenzione si assume come quantile x_α la modalità che ha determinato il punto di discontinuità, cioè:

$$x_\alpha = x_{i_0}$$

Una tale scelta è giustificabile, ricordando la definizione di funzione di ripartizione nonché quella di quantile, in base al significato delle equazioni poste in (4.4).

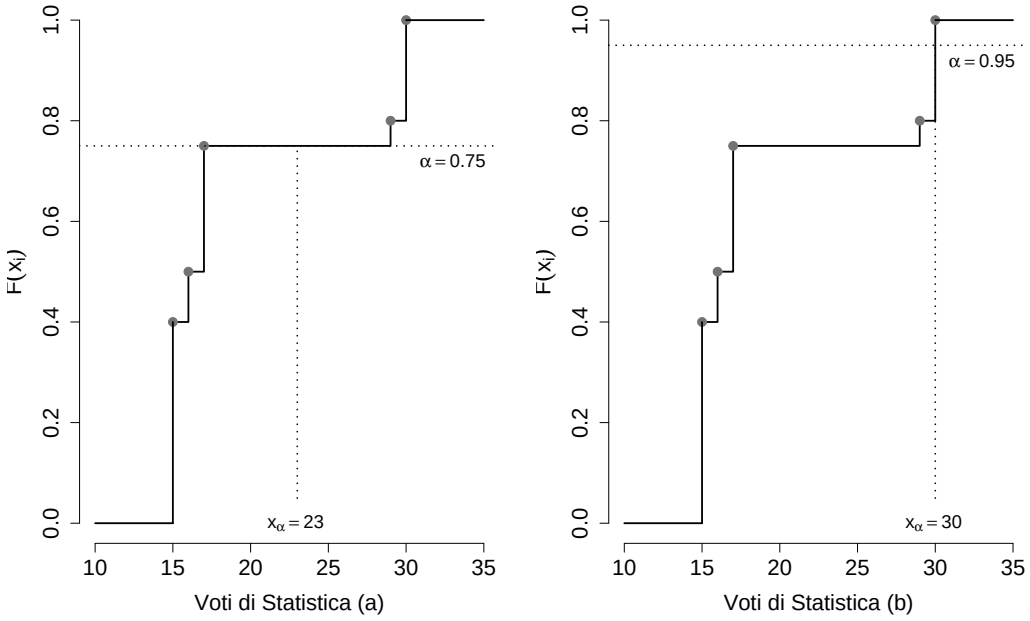


Figura 4.6 Quantili $x_{0.75}$ e $x_{0.95}$, esempi 4.4 e 4.5.

▷ ESEMPIO 4.5

Tornando alla distribuzione di frequenze della v.s. dell'esempio (4.4), scorrendo la colonna delle frequenze cumulate con l'intento di individuare il quantile di ordine 0.95 notiamo che:

$$F_X(x_4) = F_X(29) = 0.80 < 0.95$$
$$F_X(x_5) = F_X(30) = 1.00 > 0.95$$

Per quanto detto sarà $i_0 = 5$ e $x_{0.95} = x_4 = 30$. Anche in questo caso può essere di aiuto il grafico proposto in figura (4.6, pannello b).

◁

4.3.2 I QUANTILI NEL CASO DI DATI RACCOLTI IN CLASSI

Se la v.s. X è di tipo continuo con i dati raccolti in classi, sul grafico della $F_X(x)$, come illustrato in figura (4.8), individuiamo *la classe* in cui cade il quantile di ordine α

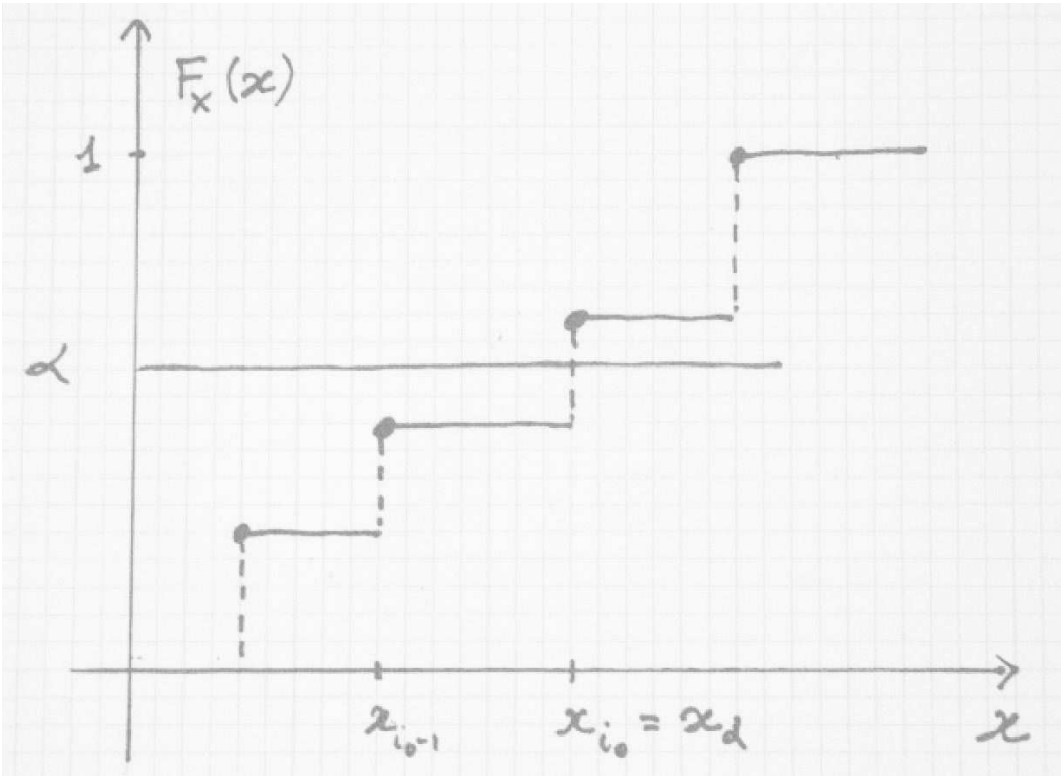


Figura 4.7 Nessuna radice dell'equazione (4.2).

tracciando una parallela all'asse delle ascisse ad ordinata α.
Per determinare il valore puntuale del quantile si procede imponendo la condizione di similitudine tra i triangoli $\widehat{ABC}$ ed $\widehat{ADE}$ evidenziati in figura (4.8) per la generica i-esima classe:

$$\overline{AC} : \overline{AE} = \overline{BC} : \overline{DE} \qquad \rightarrow \qquad \overline{AC} = \frac{\overline{AE} \cdot \overline{BC}}{\overline{DE}} \tag{4.5}$$

Ricordando la simbologia già introdotta nei paragrafi precedenti, abbiamo:

- $\star$ $\overline{AE} = l_{i+1} - l_i = w_i$, cioè l'ampiezza di classe;
- $\star$ $\overline{BC} = \alpha - F(l_i)$, cioè la differenza tra l'ordine del quantile e il valore della funzione di ripartizione in corrispondenza al limite superiore della classe immediatamente precedente;

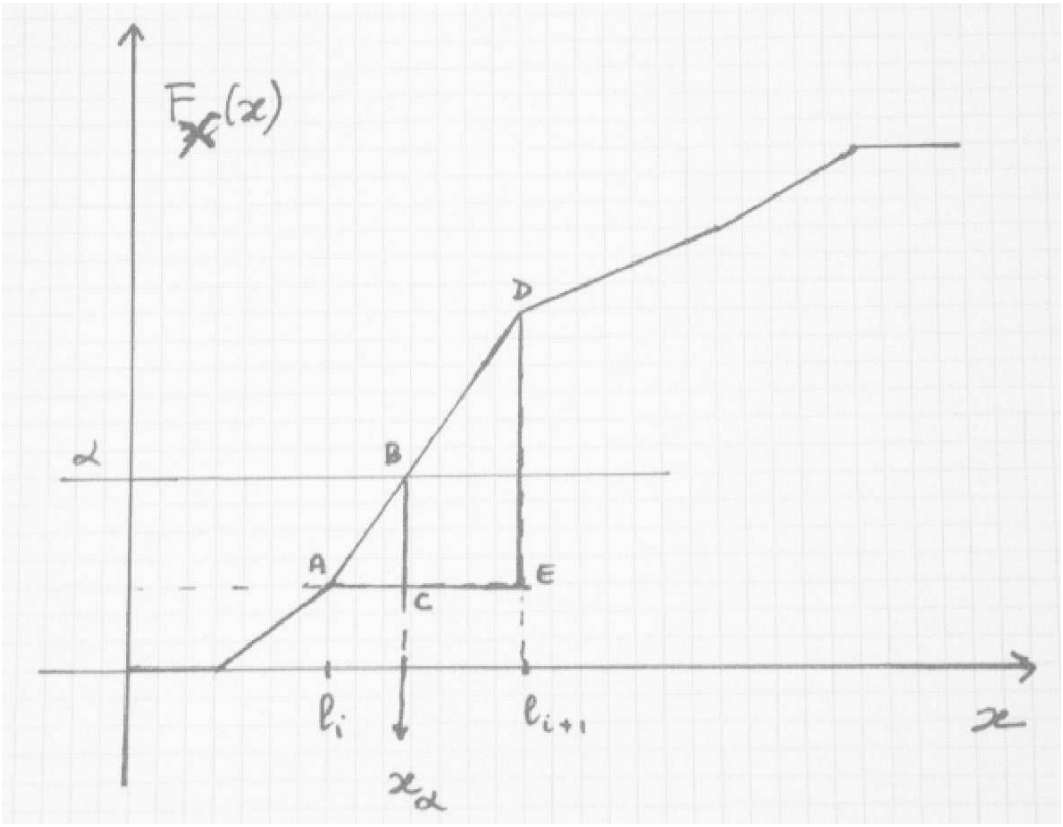


Figura 4.8 Quantile per dati raccolti in classi.

⋆ $\overline{DE} = F(l_{i+1}) - F(l_i) = f_i$, cioè la frequenza relativa della classe;

e quindi:

$$\overline{AC} = \frac{w_i \cdot (\alpha - F(l_i))}{f_i}.$$

Stando così le cose, il quantile ricercato sarà:

$$x_\alpha = l_i + \overline{AC} = l_i + \frac{w_i \cdot (\alpha - F(l_i))}{f_i}. \tag{4.6}$$

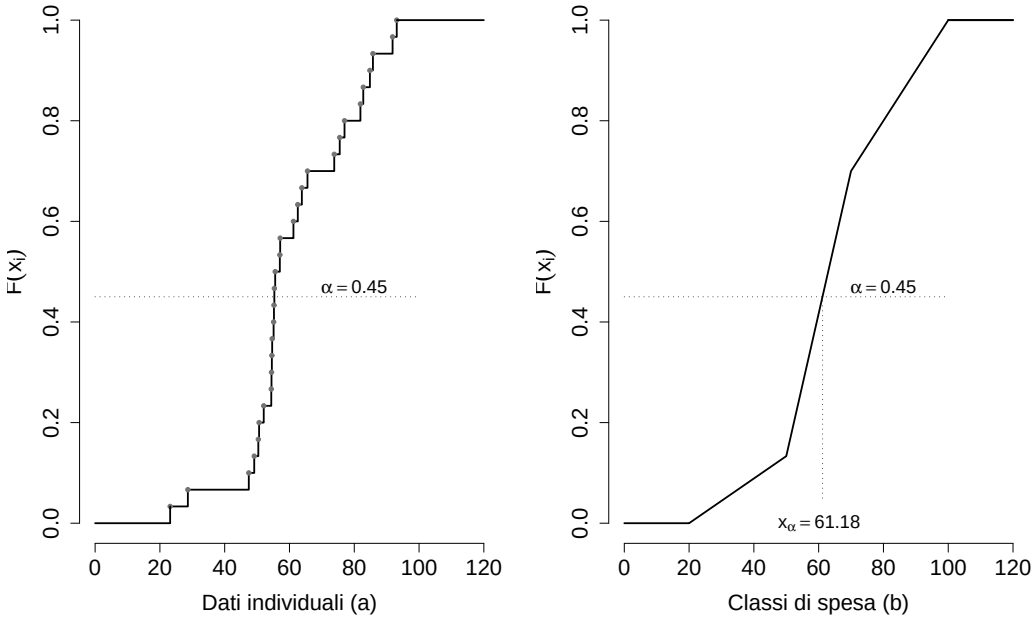


Figura 4.9 Quantile $x_{0.45}$, esempio 4.6.

▷ ESEMPIO 4.6

I valori che seguono sono le determinazioni, poste in ordine crescente, della v.s. $X = \{$*importo dello scontrino fiscale*$\}$ rilevata su un collettivo di 30 clienti all'uscita di un supermercato di periferia:

30.71	38.09	44.93	47.64	51.70	51.73	52.80	52.99	53.71	55.46
56.68	61.04	61.05	61.08	62.08	63.90	64.60	67.22	67.60	68.05
68.19	73.62	78.90	80.09	88.18	88.72	91.40	92.32	94.47	98.20

Volendo individuare il quantile di ordine $\alpha = 0.45$ notiamo che $F_X(x_{13}) = 0.43$ mentre $F_X(x_{14}) = 0.47$. Come peraltro si evince dalla funzione di ripartizione riportata in figura (4.9, pannello a), ci troviamo nella situazione descritta dall'equazione (4.4), pertanto il quantile cercato sarà $x_{0.45} = x_{14} = 61.08$.

Se consideriamo ora la seguente distribuzione di frequenze con dati raccolti in classi:

Classi di spesa	n_i	f_i	$F(l_{i+1})$
$20 \dashv 50$	4	0.133	0.133
$50 \dashv 70$	17	0.567	0.700
$70 \dashv 100$	9	0.300	1.000

Osserviamo, in primo luogo, che il quantile di ordine $\alpha = 0.45$ cade nella classe $]50; 70]$ per cui, applicando la relazione (4.6) si ottiene:

$$x_{0.45} = 50 + \frac{20 \cdot (0.45 - 0.133)}{0.567} = 61.18$$

Ciò conduce ad affermare che "il 45% dei clienti ha speso al più 61.18 euro", ma anche che "il 55% dei clienti ha fatto una spesa di importo maggiore di 61.18 euro". Come possiamo notare il quantile così ottenuto è diverso da quello calcolato a partire dai dati individuali. Tale discrepanza è dovuta unicamente alla perdita di informazioni cui ci si espone ogniqualvolta si operi su dati raccolti in classi.

$\triangleleft$

4.4. IL FOGLIO ELETTRONICO

Sebbene nei più comuni fogli elettronici non esistano grafici predefiniti per rappresentare la funzione di ripartizione di una v.s. in questo paragrafo vedremo come sia possibile, con alcuni piccoli accorgimenti, ottenere i grafici descritti in questo capitolo.
Iniziamo considerando la v.s. *anni dalla laurea* del solito file `university.sxc` aggiungendo alla tabella della distribuzione di frequenza già presentata nel precedente capitolo la colonna delle frequenze cumulate così come appare nell'intervallo di celle `J4:J7` della videata OpenOffice di figura (4.10). Se nella cella `J4` si è semplicemente introdotta la funzione `=I4` per ottenere la frequenza cumulata associata alla prima modaltà $x_1 = 1$ nella cella `J5` vi è la funzione `=J4+I5` che fornisce la frequenza cumulata associata alla seconda modalità e così via.
Per ottenere il grafico a scala della funzione di ripartizione abbiamo scelto tra quelli disponibili un grafico a dispersione, detto da OpenOffice `Diagramma XY`, che rappresenta sul piano cartesiano i punti individuati dalle coppie di coordinate selezionate. Nel nostro caso abbiamo inserito un `Diagramma XY` per i 10 punti di coordinate poste nell'intervallo di celle `H9:I18`, in un momento successivo cliccando sul grafico abbiamo chiesto di unire i punti con una linea. Il grafico riportato nella figura è proprio quello della funzione di ripartizione poichè nelle celle `H9:I18` abbiamo inserito i punti di coordinate $(0; 0)$ e $(5; 1)$ e per ciascuna modalità i punti di coordinate $(x_i; F(x_{i-1}))$ nonché $(x_i; F(x_i))$; questo accorgimento fa sì che la linea congiungente i punti risulti a gradini.

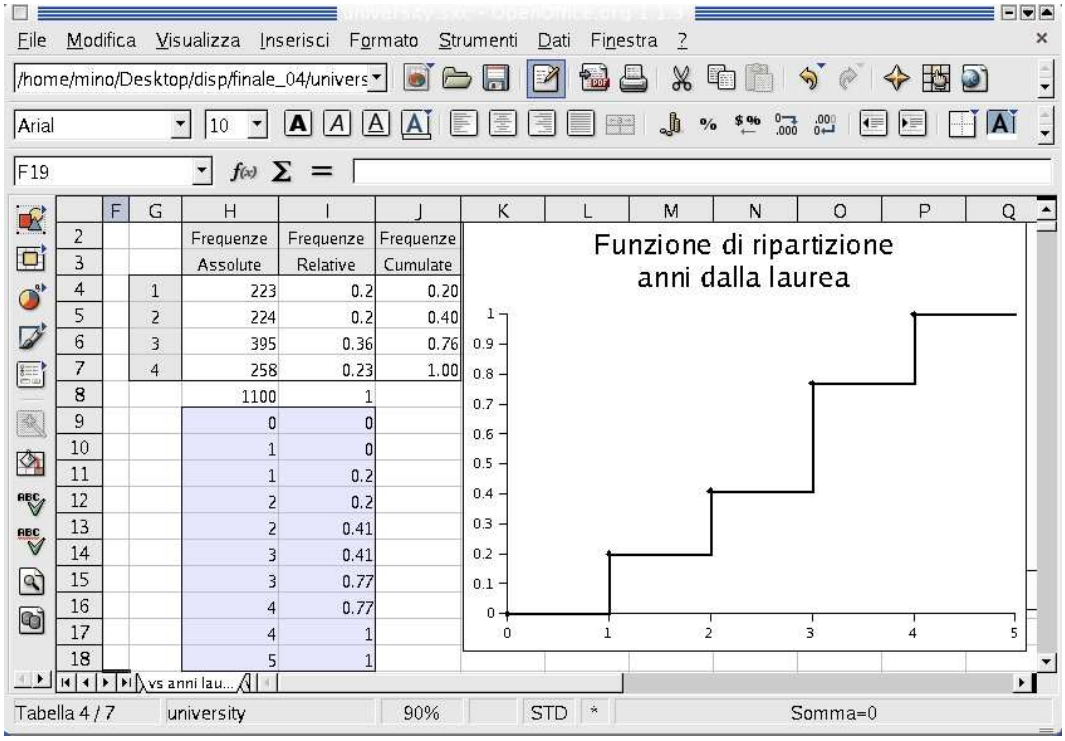


Figura 4.10 Funzione di ripartizione della v.s. *anni dalla laurea.*

Per la v.s. *stipendio* presentiamo in figura (4.11) la funzione di ripartizione della distribuzione con i dati raccolti in 4 classi. Anche in questo caso, dopo aver calcolato le frequenze cumulate, abbiamo richiesto un diagramma a dispersione per i punti di coordinate inserite nelle celle `I12:J18`. Si noti che le ascisse dei punti indicati nelle celle `I14:J17` corrispondono al limite superiore di ciascuna classe e le ordinate alla frequenza cumulata corrispondente.

Concludiamo questo paragrafo osservando che tra le funzioni di OpenOffice ne esiste una che restitusce i percentili di qualsiasi ordine. Si invita il Lettore ad inserire in una cella libera qualunque del foglio di lavoro presentato in figura (4.10) la funzione `=PERCENTILE(D2:D101;0.2)` e confrontare il risultato da essa restituito con il valore del quantile di ordine $\alpha = 0.2$ della variabile *anni dalla laurea* calcolato in base alla definizione data. Come si potrà notare i due risultati non coincidono; ciò è dovuto al fatto che la funzione `=PERCENTILE` restituisce il *quantile campionario* che è altro rispetto a quanto ci siamo prefissi in questa trattazione di analisi dei dati.

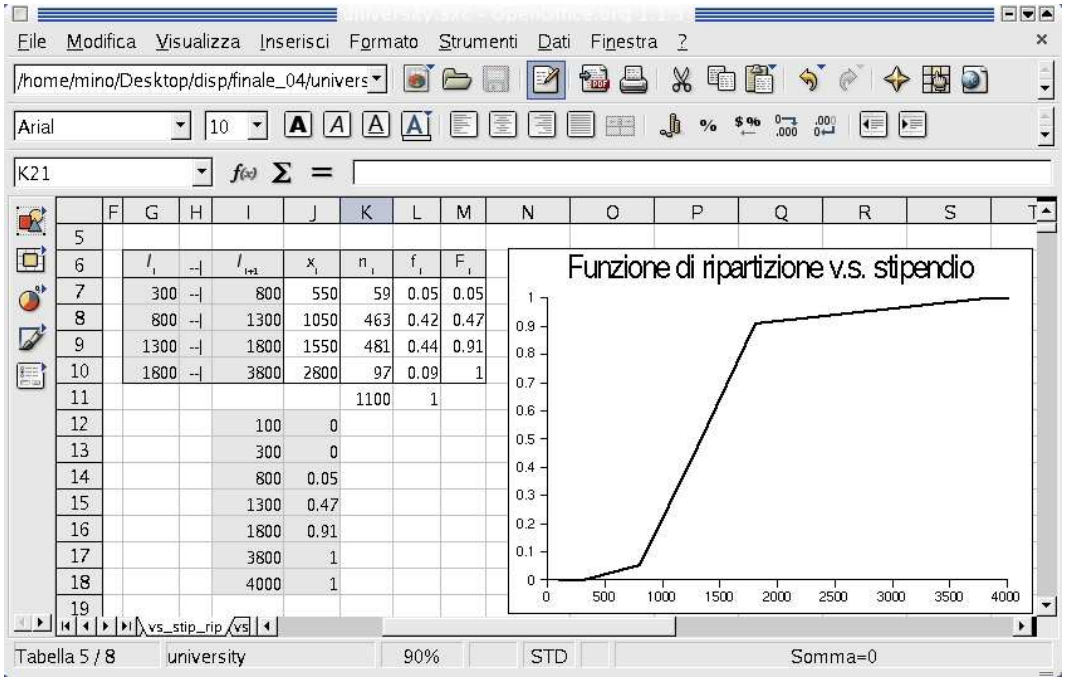


Figura 4.11 Funzione di ripartizione della v.s. *stipendio.*

4.5. ESERCIZI

▷ ESERCIZIO 4.1

Si immagini che la variabile statistica $X = \{$*numero di unità difettose per lotto*$\}$ rinvenute in 100 lotti, di 2000 unità omogenee ciascuno, in uscita da un processo produttivo, abbia la seguente distribuzione di frequenze assolute

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\ldots,6} = \begin{Bmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 44 & 22 & 11 & 4 & 3 & 2 & 14 \end{Bmatrix}$$

a) Costruire il grafico della funzione di ripartizione della v.s. X.

b) Valutare il valore di $F_X(x)$ in corrispondenza a $x = 2$, $x = 2.5$ e $x = 10$.

c) Calcolare, infine, i quantili $x_{0.50}$ e $x_{0.66}$.

◁

▷ ESERCIZIO 4.2

Un revisore contabile esamina le pratiche di pagamento sinistri effettuati da una compagnia di assicurazione nel settore R.C. auto, e sottopone al responsabile dell'agenzia a cui fanno capo i clienti rimborsati, la seguente distribuzione di frequenze per la v.s. $X = \{\textit{importo liquidato}\}$, in euro:

Classi	di	Importo	n_i	f_i	$F(l_{i+1})$
0	⊣	4	1850	...	...
4	⊣	6	2500	...	...
6	⊣	10	1200	...	...
10	⊣	40	950	...	...

Completata la tabella costruire il grafico della funzione di ripartizione. In quale classe cade il valore non superato dal 65% delle pratiche esaminate? È corretto affermare che il 15% dei rimborsi effettuati supera 1000 euro? Calcolare, infine la mediana e i quantili di ordine rispettivamente $\alpha = 0.25$, $\alpha = 0.75$

◁

▷ ESERCIZIO 4.3

Un sondaggio condotto sugli acquisti nel reparto gastronomia di un grande ipermercato ha fornito, relativamente alla quantità (in grammi) di prosciutto crudo acquistata dai 50 clienti di un sabato mattina, i seguenti dati:

200	198	270	178	180	181	111	160	167	235
203	170	207	230	119	196	220	205	132	163
143	192	250	174	252	155	240	258	197	120
210	150	184	273	130	260	254	204	277	236
179	161	221	217	140	172	201	190	185	168

Compilare la tabella di distribuzione di frequenze e rappresentare graficamente la funzione di ripartizione. Raccogliere successivamente i dati in classi e costruire il grafico della funzione di ripartizione della rispettiva distribuzione.

◁

▷ ESERCIZIO 4.4

Per i due casi proposti nell'esercizio 4.3, calcolare i quantili di ordine $\alpha = 0.20$ e $\alpha = 0.75$, confrontando i risultati ottenuti per le due situazioni.

◁

CAPITOLO 5

MISURE DI POSIZIONE

Procedendo con l'introduzione di misure di sintesi di una variabile statistica, dedichiamo questo capitolo alle misure di posizione. Definite secondo Cauchy e secondo Chisini le medie algebriche, ampio spazio sarà dedicato alla media aritmetica ed alle sue principali proprietà. Altre misure di posizione quali il massimo e il minimo, i quantili e la mediana, nonché la moda verranno definite e confrontate rispetto alla loro robustezza per particolari distribuzioni simmetriche, asimmetriche e di misture.

5.1. LE MEDIE ALGEBRICHE

I quantili di una variabile statistica, introdotti nel capitolo precedente, sono valori di sintesi che forniscono una misura della posizione delle unità statistiche rispetto al carattere quantitativo esaminato; essi rientrano in una più ampia classe di parametri caratteristici di una variabile statistica, la classe appunto delle medie.
Nel seguito assumiamo quale definizione generale di media la seguente:

Definizione 5.1 (Media secondo Cauchy)
si dice media della successione di dati individuali $\{\tilde{x}_\alpha\}_{\alpha=1,\dots,n}$ *della v.s.* X *un qualunque numero reale* x^* *interno ai dati, tale cioè che:*

$$\min_\alpha(\tilde{x}_\alpha) \leq x^* \leq \max_\alpha(\tilde{x}_\alpha)$$

□

Chiaramente la definizione del Cauchy non dice in modo esplicito quale media scegliere nè come questa debba essere calcolata, ma si limita ad affermare sotto quali condizioni un numero reale può essere considerato quale media di una successione di dati. In tale ottica, dunque, i quantili di qualsiasi ordine per definizione possono essere considerati medie.

Alla classe delle medie appartengono le cosidette medie algebriche, definibili attraverso funzioni algebriche dei valori dei dati individuali o equivalentemente della distribuzione di frequenze. Esse, al pari dei quantili, rappresentano valori di sintesi che forniscono informazioni circa l'ordine di grandezza con cui il carattere quantitativo esaminato è presente sul collettivo.
La definizione di Chisini individua un'ampia classe di medie algebriche e al contempo ci è di aiuto dal punto di vista operativo. Pertanto:

Definizione 5.2 (Media secondo Chisini)
data la successione di dati individuali $\{\tilde{x}_\alpha\}_{\alpha=1,\dots,n}$ *si dice media della v.s.* X *rispetto ad una funzione* $\varphi(\tilde{x}_1,\dots,\tilde{x}_n)$ *invertibile quella costante* M *che soddisfa la condizione di invarianza*

$$\varphi(\tilde{x}_1,\dots,\tilde{x}_n) = \varphi(M,\dots,M) \tag{5.1}$$

□

Per impiegare la definizione del Chisini al fine di individuare una media, si deve scegliere in primo luogo la funzione $\varphi(\tilde{x}_1,\dots,\tilde{x}_n)$ e successivamente individuare la costante M che soddisfa la condizione di invarianza, posta in (5.1), rispetto alla funzione scelta.

In tale ottica, dunque molteplici sono le medie che possono essere individuate. Tra quelle di maggior utilizzo ne citiamo alcune, e precisamente

- ⋆ **media aritmetica**: essa è quella costante che soddisfa la condizione di invarianza (5.1) per la funzione *somma dei dati individuali*, infatti:

$$\varphi(\tilde{x}_1,\dots,\tilde{x}_n) = \sum_{\alpha=1}^{n} \tilde{x}_\alpha \qquad \rightarrow \qquad \sum_{\alpha=1}^{n} \tilde{x}_\alpha = \sum_{\alpha=1}^{n} M$$

  per cui $M = \dfrac{1}{n}\sum_{\alpha=1}^{n} \tilde{x}_\alpha$.

- ⋆ **media armonica**: essa è quella costante che soddisfa la condizione di invarianza (5.1) per la funzione *somma del reciproco dei dati individuali*, posto che questi siano tutti positivi; infatti:

$$\varphi(\tilde{x}_1,\dots,\tilde{x}_n) = \sum_{\alpha=1}^{n} \tilde{x}_\alpha^{-1} \qquad \rightarrow \qquad \sum_{\alpha=1}^{n} \tilde{x}_\alpha^{-1} = \sum_{\alpha=1}^{n} M^{-1}$$

  per cui $M = \dfrac{n}{\sum_{\alpha=1}^{n} \tilde{x}_\alpha^{-1}}$.

⋆ **media quadratica**: essa è quella costante che soddisfa la condizione di invarianza (5.1) per la funzione *somma del quadrato dei dati individuali*, posto che questi non siano negativi; infatti

$$\varphi(\tilde{x}_1, \ldots, \tilde{x}_n) = \sum_{\alpha=1}^{n} \tilde{x}_\alpha^2 \qquad \rightarrow \qquad \sum_{\alpha=1}^{n} \tilde{x}_\alpha^2 = \sum_{\alpha=1}^{n} M^2$$

per cui $M = \sqrt{\frac{1}{n} \sum_{\alpha=1}^{n} \tilde{x}_\alpha^2}$.

⋆ **media geometrica**: essa è quella costante che soddisfa la condizione di invarianza (5.1) per la funzione *prodotto dei dati individuali*, posto che questi siano tutti positivi; infatti:

$$\varphi(\tilde{x}_1, \ldots, \tilde{x}_n) = \prod_{\alpha=1}^{n} \tilde{x}_\alpha \qquad \rightarrow \qquad \prod_{\alpha=1}^{n} \tilde{x}_\alpha = \prod_{\alpha=1}^{n} M$$

per cui $M = \left(\prod_{\alpha=1}^{n} \tilde{x}_\alpha\right)^{1/n}$.

Qualora la v.s. X assuma valori negativi o nulli, tutti o in parte, la media geometrica può non avere valore reale; essa è pertanto definita solo per v.s. a valori positivi.

Quelle sopra definite sono solo quattro delle possibili medie algebriche individuabili a partire dalla definizione proposta dal Chisini.
Nel caso in cui le determinazioni della v.s. X fossero tutte positive è possibile definire una famiglia di medie alla quale appartengono tutte quelle sopra citate. Infatti qualora ponessimo $\varphi(\tilde{x}_1, \ldots, \tilde{x}_n) = \sum_{\alpha=1}^{n} \tilde{x}_\alpha^r$, con $r \in \mathbb{R}$, e applicassimo la condizione di invarianza di Chisini, perverremmo alla seguente

Definizione 5.3 (Media di r-esima potenza)
data una v.s. X con modalità osservate $\tilde{x}_\alpha > 0$ per ogni $\alpha = 1, \ldots, n$ si dice media di r-esima potenza la radice r-esima della media aritmetica delle potenze r-esime dei valori $\tilde{x}_\alpha$, cioè

$$M_r = \left(n^{-1} \sum_{\alpha=1}^{n} \tilde{x}_\alpha^r \right)^{1/r} \tag{5.2}$$

□

Le medie precedentemente introdotte possono essere ricavate dalla (5.2) ponendo, di volta in volta, $r = 1$, $r = -1$, $r = 2$ e $r \to 0$.
Non desiderando, volutamente, trattare in modo esaustivo il concetto di medie di $r-$esima potenza, si consiglia al Lettore interessato di consultare i testi di Jalla (1991) e Landenna (1997).

Prima di concludere il paragrafo, osserviamo che qualora disponessimo della distribuzione di frequenze $\{x_i; n_i\}_{i=1,\dots,k}$ della v.s. X, la condizione di invarianza posta dalla (5.1) diverrebbe:

$$\varphi(x_1, \dots, x_k; n_1, \dots, n_k) = \varphi(M, \dots, M; n_1, \dots, n_k) \tag{5.3}$$

Se tale è il caso, allora:

- ⋆ essendo $\sum_{\alpha=1}^{n} \tilde{x}_\alpha = \sum_{i=1}^{k} x_i n_i$, la media aritmetica sarà $M = \dfrac{1}{n} \sum_{i=1}^{k} x_i n_i$
- ⋆ essendo $\sum_{\alpha=1}^{n} \tilde{x}_\alpha^{-1} = \sum_{i=1}^{k} x_i^{-1} n_i$, la media armonica sarà $M = \dfrac{n}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$
- ⋆ essendo $\sum_{\alpha=1}^{n} \tilde{x}_\alpha^2 = \sum_{i=1}^{k} x_i^2 n_i$, la media quadratica sarà $M = \sqrt{\dfrac{1}{n} \sum_{i=1}^{k} x_i^2 n_i}$
- ⋆ essendo $\prod_{\alpha=1}^{n} \tilde{x}_\alpha = \prod_{i=1}^{k} x_i n_i$, la media geometrica sarà $M = \left(\prod_{i=1}^{k} {x_i}^{n_i} \right)^{1/n}$

Nel seguito dedicheremo la nostra attenzione alla media aritmetica che, per svariati motivi che risulteranno chiari nel corso dell'esposizione, è la più largamente impiegata in statistica sì da rivestire il ruolo di "prima donna". A commento di quanto detto in questo paragrafo, valga l'esempio che segue.

▷ ESEMPIO 5.1

Data la v.s. X con distribuzione di frequenze assolute:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\dots,4} = \begin{Bmatrix} 0.2 & 0.5 & 0.9 & 1.3 \\ 6 & 3 & 2 & 4 \end{Bmatrix}$$

ci proponiamo di calcolare le medie sopra definite.

⋆ media aritmetica:

$$\frac{0.2 \cdot 6 + 0.5 \cdot 3 + 0.9 \cdot 2 + 1.3 \cdot 4}{15} = 0.647$$

⋆ media armonica:

$$\frac{15}{0.2^{-1} \cdot 6 + 0.5^{-1} \cdot 3 + 0.9^{-1} \cdot 2 + 1.3^{-1} \cdot 4} = 0.363$$

⋆ media quadratica:

$$\sqrt{\frac{0.2^2 \cdot 6 + 0.5^2 \cdot 3 + 0.9^2 \cdot 2 + 1.3^2 \cdot 4}{15}} = 0.790$$

⋆ media geometrica:

$$\sqrt[15]{0.2^6 \cdot 0.5^3 \cdot 0.9^2 \cdot 1.3^4} = 0.484$$

◁

OSSERVAZIONE: a proposito della media geometrica (M_g) è bene tenere presente che, impiegando strumenti di calcolo quali calcolatrici scientifiche o personal computer, è possibile che l'operazione di produttoria porti ad errori di overflow, soprattutto quando i dati hanno ordine di grandezza consistente. In tali casi è utile calcolare il logaritmo della media geometrica e successivamente ritornare a questa con l'operazione di antilogaritmo. Dal momento che il logaritmo della media geometrica corrisponde alla media aritmetica del logaritmo dei dati, dalla definizione applicando le proprietà del logaritmo otteniamo:

$$\ln(M_g) = \ln(\prod_{i=1}^{k} {x_i}^{n_i})^{\frac{1}{n}} = \frac{1}{n} \sum_{i=1}^{k} \ln(x_i^{n_i}) = \frac{1}{n} \sum_{i=1}^{k} n_i \ \ln(x_i)$$

Così nel caso proposto all'esempio (5.1), la procedura di calcolo per la media geometrica risulterebbe:

$$\ln(M_g) = \frac{6\ln(0.2) + 3\ln(0.5) + 2\ln(0.9) + 4\ln(1.3)}{15} = -0.726$$

da cui il risultato $M_g = e^{-0.726} = 0.484$.

⋆

5.1.1 La media aritmetica

Nel linguaggio comune, la *media* di una serie di dati corrisponde a quella che nel paragrafo precedente è stata definita come media aritmetica. Tale parametro di sintesi è senza dubbio noto anche in ambiti non strettamente statistici; qualunque studente ha, ad esempio, calcolato almeno una volta la media dei voti degli esami sostenuti; la spesa media mensile viene assunta come sintesi di condizione finanziaria da molte famiglie, ed ancora, il consumo medio di carburante è una caratteristica che viene valutata al momento dell'acquisto di una autovettura.
Ricordando la definizione del Chisini proposta in (5.2), la media aritmetica di una v.s. è quella costante che soddisfa la condizione di invarianza rispetto alla funzione *somma dei dati individuali* e tale peculiarità della media aritmetica è intrinsecamente nota nella coscienza comune.
Lo studente a cui risulta una media dei voti, diciamo, pari a 28, interpreta infatti tale valore come quello che potrebbe attribuire a ciascun esame sostenuto se gli fosse possibile accumulare, cioè sommare, i voti conseguiti e distribuirli in modo invariante fra gli esami sostenuti.

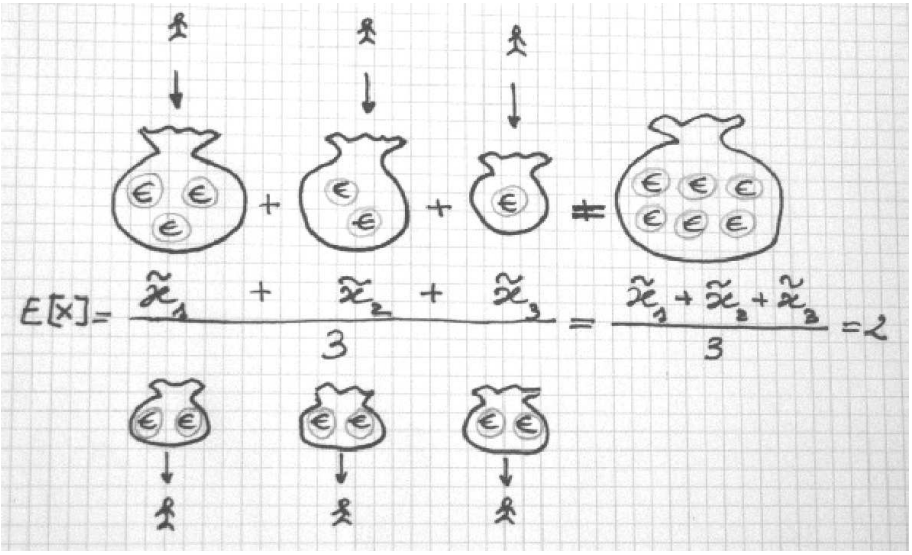


Figura 5.1 Media aritmetica e redistribuzione del reddito.

Risulta evidente che la media aritmetica è maggiormente significativa se il carattere in esame è *trasferibile*, ossia se esso è tale da poter essere ceduto da una unità all'altra del collettivo; così ad esempio, reddito pro-capite di una comunità di individui è un carattere,

almeno idealmente, trasferibile, ciascun individuo può cedere, totalmente o in parte, ad un'altro il proprio reddito (si veda figura 5.1), viceversa non sono trasferibili caratteri quali ad esempio la statura o l'età.

Tornando ora in un contesto propriamente statistico formalizziamo quanto detto sulla media aritmetica dando la seguente

Definizione 5.4 (Media aritmetica)
definiamo media aritmetica della v.s. X il valore numerico risultante dall'operazione

$$E[X] = \frac{1}{n}\sum_{\alpha=1}^{n} \tilde{x}_\alpha \tag{5.4}$$

che nel caso la v.s. X abbia distribuzione di frequenze $\{x_i; n_i\}_{i=1,\dots,k}$ verrà calcolato come

$$E[X] = \frac{1}{n}\sum_{i=1}^{k} x_i\, n_i = \sum_{i=1}^{k} x_i\, f_i \tag{5.5}$$

□

Il simbolo $E[\,\cdot\,]$ rappresenta, in questo contesto, un *operatore* che applicato ad una qualsiasi variabile statistica X restituisce uno ed uno solo numero reale, abitualmente indicato con μ_X, o più semplicemente con μ qualora non sussistano dubbi di ambiguità con altre variabili statistiche.
In altri termini, data la variabile statistica X, potremmo dire che l'operatore E ad essa applicato la trasforma in un numero reale μ_X che corrisponde appunto alla sua media aritmetica o più semplicemente valor medio.

▷ ESEMPIO 5.2

Si immagini che la v.s. $X = \{$*voto di una prova scritta di statistica*$\}$, rilevata su un colettivo di 60 studenti, possegga distribuzione di frequenze assolute:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\dots,7} = \begin{Bmatrix} 15 & 16 & 17 & 19 & 21 & 23 & 25 \\ 5 & 15 & 10 & 8 & 12 & 7 & 3 \end{Bmatrix}$$

Desiderando calcolare il valor medio di X, applicando la (5.5) si ottiene:

$$E[X] = \frac{1}{60}\sum_{i=1}^{7} x_i\, n_i = \\ = \frac{15\cdot 5 + 16\cdot 15 + 17\cdot 10 + 19\cdot 8 + 21\cdot 12 + 23\cdot 7 + 25\cdot 3}{60} = \\ = 18.75 = \mu_X$$

Osserviamo che $\mu_X = 18.75$ è un valore medio secondo la definizione di Cauchy, infatti $x_1 = 15 < 18.75 < 25 = x_7$.

$\triangleleft$

Nel caso ci si trovi ad operare su variabili statistiche con dati raccolti in classi, la corrispondente distribuzione di frequenze:

$$X \equiv \begin{Bmatrix} l_i \dashv l_{i+1} \\ n_i \end{Bmatrix}_{i=1,\ldots,k}$$

viene ricondotta alla forma:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\ldots,k} = \begin{Bmatrix} x_1 & x_2 & \ldots & x_k \\ n_1 & n_2 & \ldots & n_k \end{Bmatrix}$$

scegliendo un valore rappresentativo x_i per ogni classe, ad esempio ponendo:

$$x_i = \frac{l_i + l_{i+1}}{2}$$

Abbiamo già osservato che il raccoglimento dei dati in classi comporta una perdita di informazioni, la qualcosa risulta ancora più evidente se si confronta la media aritmetica calcolata a partire dai dati individuali con quella relativa alla distribuzione di frequenza della v.s. con dati raccolti in classi.

$\triangleright$ ESEMPIO 5.3

La misurazione del peso corporeo di un gruppo di 40 pazienti, di entrambi i sessi di un reparto ospedaliero ha fornito i seguenti valori espressi in kg:

80.3	77.0	75.5	75.6	80.0	74.0	74.2	78.5	80.5	75.0
72.5	70.5	68.0	70.0	72.3	67.0	67.3	71.5	73.0	67.9
95.5	85.7	84.5	85.2	91.3	81.2	82.0	90.5	96.0	83.5
65.6	65.0	62.6	63.0	65.5	61.0	62.0	65.3	65.8	62.5

Il valor medio della v.s. $X = \{\textit{peso dei pazienti}\}$ in accordo alla (5.4) è pertanto:

$$E[X] = \frac{80.3 + 77.0 + \ldots + 65.8 + 62.5}{40} = 74.61 = \mu_X$$

Se per la v.s. X disponessimo unicamente della seguente distribuzione di frequenza, con dati raccolti in classi di peso:

Classi di peso	x_i	n_i
$60 \dashv 70$	65	15
$70 \dashv 80$	75	13
$80 \dashv 90$	85	8
$90 \dashv 100$	95	4

per la v.s. X risulterebbe, sfruttando la (5.5):

$$E[X] = \frac{65 \cdot 15 + 75 \cdot 13 + 85 \cdot 8 + 95 \cdot 4}{40} = 75.25 = \mu_X$$

Pur essendo dati relativi allo stesso collettivo statistico (cfr. figura 5.2, a), le due medie risultano differenti. Mentre la prima porta, correttamente, ad affermare che, potendo distribuire il peso totale equamente, ciascun paziente peserebbe 74.61 kg, la seconda attribuirebbe, erroneamente, ad ogni paziente il peso di 75.25 kg.
Resta inteso che, qualora non si posseggano i valori individuali, la media della distribuzione dei dati in classi è comunque un utile parametro di sintesi.
Per completezza, in figura (5.2, pannello b), è riportato l'istogramma della distribuzione di frequenza proposta con traccia dei dati individuali. Appare del tutto evidente come il raccoglimento in classi proposto non sia soddisfacente, perlomeno per valori di peso maggiori di 85 kg. Di qui la perdita di informazione e la discrepanza tra le medie calcolate nelle due situazioni.

$\triangleleft$

A commento di quanto sopra, possiamo ancora osservare che la media aritmetica:

- $\star$ rappresenta il *baricentro* di una distribuzione di frequenze. Potremmo dire che essa costituisce l'ago della bilancia che sostiene l'area rappresentata dall'istogramma (cfr. ad esempio figura 5.2, a);

- $\star$ risente, come del resto tutte le medie algebriche essendo queste calcolate sulla base di *tutte* le osservazioni, della presenza nei dati di valori "anomali", abitualmente detti in letteratura "outliers". Trattasi di singoli valori troppo grandi o troppo piccoli rispetto all'insieme dei dati che si presentano per cause non strettamente collegate al fenomeno sotto osservazione, ad esempio errori di rilevazione o di trascrizione dei dati, presenza di unità statistiche con caratteristiche non omogenee rispetto alla totalità del collettivo ed anche influenza di cause rare e sporadiche.

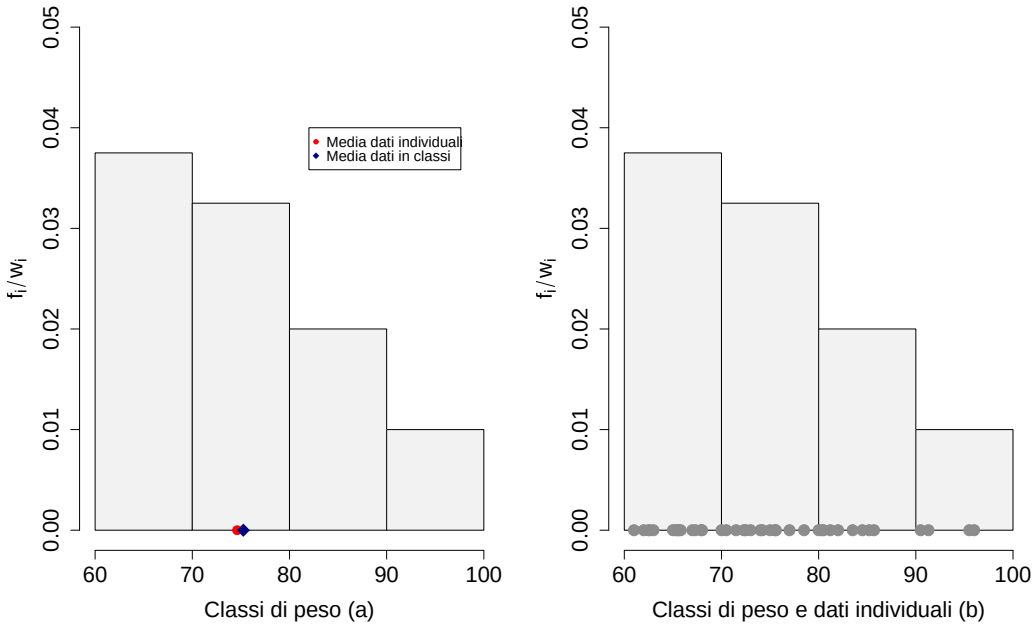


Figura 5.2 Istogrammi della distribuzione del peso, esempio 5.3.

5.1.2 Principali proprietà della media aritmetica

Proprio per l'importanza che riveste nell'ambito della statistica il concetto di media aritmetica, ne evidenziamo alcune proprietà, ipotizzando che la v.s. X abbia distribuzione di frequenze assolute $\{x_i; n_i\}_{i=1,\dots,k}$ o in modo del tutto equivalente distribuzione di frequenze relative $\{x_i; f_i\}_{i=1,\dots,k}$.

Proprietà 5.1 La somma degli scarti tra ciascuna modalità x_i e la media aritmetica μ_X è nulla, cioè:

$$\sum_{i=1}^{k}(x_i - \mu_X)n_i = 0$$

◁

Dimostrazione: applicando le proprietà dell'operatore sommatoria:

$$\sum_{i=1}^{k}(x_i - \mu_X)n_i = \sum_{i=1}^{k} x_i n_i - \mu_X \sum_{i=1}^{k} n_i = n\,\mu_X - n\,\mu_X = 0$$

□

Proprietà 5.2 La somma dei quadrati degli scarti tra ciascuna modalità x_i e una costante arbitraria $a \in \mathbb{R}$, cioè:

$$\varphi(a) = \sum_{i=1}^{k}(x_i - a)^2 n_i$$

è minima per $a = \mu_X$.

◁

Dimostrazione: osserviamo innanzitutto che la funzione $\varphi(a)$ è, per definizione, una parabola con la concavità rivolta verso l'alto; pertanto essa possiede un unico punto di minimo assoluto determinabile, come di consueto, annullando la sua derivata prima.

$$\frac{d\,\varphi(a)}{d\,a} = \frac{d}{d\,a}\sum_{i=1}^{k}(x_i - a)^2 n_i = \sum_{i=1}^{k}\frac{d}{d\,a}(x_i - a)^2 n_i =$$

$$= \sum_{i=1}^{k} 2\left[\frac{d}{d\,a}(x_i - a)\right](x_i - a)n_i = -2\sum_{i=1}^{k}(x_i - a)n_i$$

Il punto di minimo di si otterrà annullando tale derivata, cioè:

$$\sum_{i=1}^{k} x_i n_i - a\sum_{i=1}^{k} n_i = 0 \qquad \Rightarrow \qquad n\,a = \sum_{i=1}^{k} x_i n_i$$

e pertanto $a = \mu_X$.

□

OSSERVAZIONE: si vedrà in seguito che $\varphi(a = \mu_X)$ fornisce una misura della dispersione dei dati, ovvero della variabilità della v.s. X.

⋆

Proprietà 5.3 La media aritmetica soddisfa la condizione del Cauchy, ovvero gode della proprietà di internalità, cioè:

$$x_1 \leq \mu_X \leq x_k$$

◁

Dimostrazione: ricordando che per definizione $x_1 \leq x_i \leq x_k$, e ciò $\forall\, i$, moltiplicando tutti i termini per la costante positiva f_i la precedente relazione d'ordine rimane immutata, cioè:

$$x_1 f_i \leq x_i f_i \leq x_k f_i$$

Sommando ora rispetto all'indice i ciascun termine della diseguaglianza, l'ordine si mantiene e si ha:

$$\sum_{i=1}^{k} x_1 f_i \leq \sum_{i=1}^{k} x_i f_i \leq \sum_{i=1}^{k} x_k f_i \quad \Rightarrow \quad x_1 \sum_{i=1}^{k} f_i \leq \mu_X \leq x_k \sum_{i=1}^{k} f_i$$

e dunque $x_1 \leq \mu_X \leq x_k$.

□

Proprietà 5.4 La media della trasformata lineare di una v.s. corrisponde alla trasformata lineare della sua media. In altri termini, date la v.s. X e la trasformata $Y = a + b\,X$, con $a, b \in \mathbb{R}$, si ha $\mu_Y = a + b\,\mu_X$.

◁

Dimostrazione: è sufficiente applicare la definizione di media aritmetica alla nuova v.s. Y e tenere a mente che $\sum_{i=1}^{k} f_i = 1$:

$$\mu_Y = \sum_{i=1}^{k} (a + b\,x_i) f_i = a \sum_{i=1}^{k} f_i + b \sum_{i=1}^{k} x_i f_i = a + b\,\mu_X$$

□

OSSERVAZIONE: date la v.s. X e la sua trasformata $Y = a + b\,X$, con $a, b \in \mathbb{R}$, quest'ultima è a sua volta una v.s. Infatti è sufficiente osservare che $\forall\, \alpha = 1, \ldots, n$:

$$\tilde{y}_\alpha = a + b\,\tilde{x}_\alpha = a + b\,X(\omega_\alpha) = Y(\omega_\alpha)$$

Dunque Y è un'applicazione che associa a ciascun elemento di Ω uno ed un solo numero reale.

⋆

OSSERVAZIONE: considerando l'operatore $E[\,\cdot\,]$, che applicato ad una v.s. fornisce la corrispondente media aritmetica, questa ultima proprietà consente di affermare che esso è un *operatore lineare*, nel senso che se applicato ad una trasformata lineare di una v.s. fornisce la trasformata lineare della sua media aritmetica.

Nel seguito ci avvarremo sovente della seguente

Proprietà 5.5 L'operatore $E[\,\cdot\,]$ è un'operatore lineare, cioè data la v.s. $Y = a + b\,X$:

$$E[Y] = E[a + b\,X] = E[a] + E[b\,X] = a + b\,E[X] \tag{5.6}$$

◁

Dimostrazione: è sufficiente ricordare che, con $a, b \in \mathbb{R}$, $E[a] = \sum_{i=1}^{k} a\,f_i = a$ e $E[b\,X] = \sum_{i=1}^{k} b\,x_i\,f_i = b\,E[X]$.

□
⋆

▷ ESEMPIO 5.4

Con questo esempio ci proponiamo di verificare numericamente le proprietà della media aritmetica, già dimostrate analiticamente.
A tal scopo sia X una v.s. con distribuzione di frequenze assolute:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\ldots,5} = \begin{Bmatrix} 15 & 16 & 17 & 29 & 30 \\ 40 & 10 & 25 & 5 & 20 \end{Bmatrix}$$

e valor medio $\mu_X = 19.3$.
Verifichiamo la proprietà (5.1), ossia quella che asserisce la nullità della somma degli scarti:

$$\sum_{i=1}^{k}(x_i - \mu_X)\,n_i = (15 - 19.3)\,40 + \ldots + (30 - 19.3)\,20 =$$
$$= (15 \cdot 40 + 16 \cdot 10 + \ldots + 30 \cdot 20) - 100 \cdot 19.3 = 0$$

Verifichiamo la proprietà (5.2), cioè quella relativa al minimo della somma dei quadrati degli scarti:

$$\varphi(a) = \sum_{i=1}^{k}(x_i - a)^2\,n_i = a^2 \sum_{i=1}^{k} n_i - 2\,a \sum_{i=1}^{k} x_i\,n_i + \sum_{i=1}^{k} x_i^2\,n_i =$$
$$= n\,a^2 - 2\,n\,\mu_X\,a + \sum_{i=1}^{k} x_i^2\,n_i = 100 \cdot a^2 - 3860 \cdot a + 40990$$

Evidentemente la derivata prima di $\varphi(a)$ risulta $\varphi'(a) = 200\,a - 3860$ ed essa si annulla per $a = 3860/200 = 19.3$.

Immediato è verificare la proprietà (5.3) di internalità, infatti $15 < 19.3 < 30$.
Al fine di verificare la proprietà (5.4), supponiamo di voler esprimere i voti in centesimi e di voler aggiungere al voto di ciascuno studente un "bonus" di dieci centesimi, in modo da ottenere una nuova v.s. Y legata in modo lineare alla v.s. X. Dalla relazione di equivalenza $X : 30 = Y : 100$ ricaviamo la corrispondenza fra i voti X espressi in trentesimi ed i voti Y espressi in centesimi: $Y = 3.3\bar{3}\, X$. I voti in centesimi corretti dal bonus saranno pertanto $Y = 3.3\bar{3}\, X + 10$. La v.s. Y così costruita possiede la distribuzione di frequenze assolute:

$$Y \equiv \begin{Bmatrix} y_i \\ n_i \end{Bmatrix} = \begin{Bmatrix} 60 & 63.3\bar{3} & 66.6\bar{6} & 106.6\bar{6} & 110 \\ 40 & 10 & 25 & 5 & 20 \end{Bmatrix}$$

da essa ricaviamo la media aritmetica $\mu_Y = \frac{60\cdot 40 + \ldots + 30\cdot 20}{100} = 74.3\bar{3}$.
Risultato che avremmo potuto ricavare direttamente utilizzando la proprietà (5.4):

$$E[Y] = 33.33\, E[X] + 10 = 3.3\bar{3} \cdot 19.3 + 10 = 74.3\bar{3}$$

◁

▷ ESEMPIO 5.5

Si immagini che la v.s. X abbia la seguente distribuzione di frequenze con dati raccolti in classi:

Classi v.s. X	n_i
$10 \dashv 30$	5
$30 \dashv 50$	10
$50 \dashv 70$	20
$70 \dashv 90$	5

e che, introdotta la trasformata $Y = X - E[X]$ se ne desideri calcolare valor medio e mediana.
Innanzitutto per la v.s. X, con semplici calcoli, si ha $E[X] = 52.5$ e $x_{0.5} = 55.0$. Quanto al valor medio di Y, dalla proprietà (5.5) dell'operatore $E[\,\cdot\,]$, si ha:

$$E[Y] = E[X - E[X]] = E[X] - E[E[X]] = E[X] - E[X] = 0$$

risultato a cui saremmo potuti giungere direttamente sfruttando la proprietà (5.1) degli scarti dalla media aritmetica, semplicemente osservando che la v.s. Y trasforma la v.s. X in *scarti dalla sua media.* Come si evince dagli istogrammi di figura (5.3), la distribuzione di Y ha la stessa *forma* di quella di X traslata sull'asse delle ascisse sì da avere baricentro pari a zero.
A ben vedere, la distribuzione di frequenza di Y è la seguente:

Classi v.s. Y	n_i
$-42.5 \dashv -22.5$	5
$-22.5 \dashv -2.5$	10
$-2.5 \dashv 17.5$	20
$17.5 \dashv 37.5$	5

Quanto alla mediana, non abbiamo dimostrato un'analoga proprietà, ma avendo osservato che la *forma* della distribuzione di Y è uguale a quella di X, potremmo arguire che la mediana di Y sia $y_{0.5} = x_{0.5} - 52.5 = 2.5$. Lasciamo al Lettore la verifica numerica di quanto asserito.

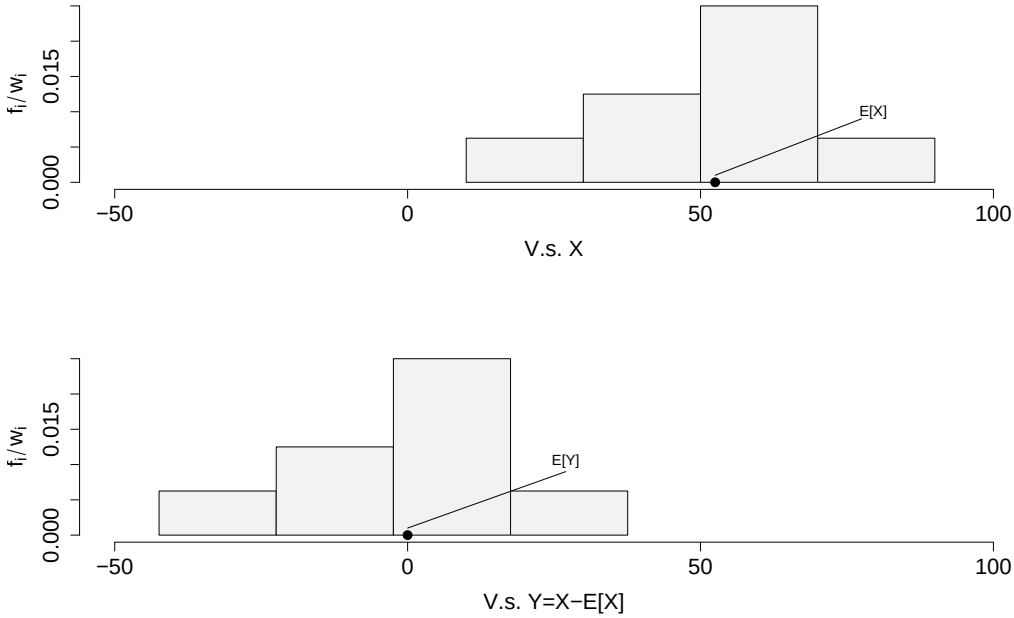


Figura 5.3 Istogrammi delle v.s. X e $Y = X - E[X]$, esempio 5.5.

◁

5.2. ALTRE MISURE DI POSIZIONE

A fianco delle medie algebriche, introdotte nei paragrafi precedenti, si possono considerare altre misure di posizione che, se impiegate congiuntamente a queste, permettono di

evidenziare altre caratteristiche dalla variabile statistica sotto osservazione.
A ben vedere, tali misure:

- ⋆ soddisfano tutte la definizione del Cauchy e pertanto possono essere considerate medie, sebbene non nel senso algebrico;
- ⋆ sono calcolate, a differenza delle medie algebriche, tenendo conto solo di parte delle realizzazioni della variabile statistica in esame.

5.2.1 IL MINIMO E IL MASSIMO

Data una v.s. X con distribuzione di frequenze assolute $\{x_i; n_i\}_{i=1,\dots,k}$ o equivalentemente distribuzione di frequenze relative $\{x_i; f_i\}_{i=1,\dots,k}$, possiamo considerare, quali misure di posizione il minimo e il massimo valore da essa assunto, il che, per definizione, implica considerare, rispettivamente, le modalità x_1 e x_k.
L'intervallo reale $[x_1; x_k]$ viene comunemente detto *intervallo di escursione*, o *range* in letteratura anglosassone, e rappresenta appunto il codominio della v.s. X.
Ovviamente la conoscenza del range ci informa unicamente circa la posizione dei "valori estremi" che la v.s. assume sul nostro collettivo; si tratta di una misura per così dire grossolana di posizione, ma tuttavia non è priva di utilità. Ad esempio abbiamo già visto come i valori dell'intervallo di escursione possano tornare utili ai fini del raccoglimento dei dati in classi. Su tale argomento torneremo nel prossimo capitolo.
Una misura che sintetizza i due precedenti parametri ci è offerta dalla loro semisomma, in simboli

$$\frac{x_1 + x_k}{2} \tag{5.7}$$

che come vedremo può essere utile se confrontata con altre misure di posizione quali la media aritmetica, la mediana, ...
È appena il caso di osservare che tali misure di posizione:

- ⋆ sono altamente influenzate dalla presenza di dati "anomali", e ciò per definizione;
- ⋆ sono calcolate, a differenza delle medie algebriche, tenendo conto di sole due determinazioni assunte dalla v.s. in esame;
- ⋆ soddisfano la definizione di media secondo Cauchy.

5.2.2 I QUANTILI

Dei quantili ci siamo già occupati nel corso del capitolo precedente, al quale rimandiamo per i particolari. Dal momento che per definizione essi soddisfano la condizione del Cauchy, i quantili possono, pertanto, essere considerati valori medi.
Chiaramente, i quantili rappresentano di una variabile statistica altrettante misure di posizione che, a differenza di quelle presentate sino ad ora, risultano in generale meno sensibili alla presenza di valori "anomali" nei dati. Misure di sintesi che godono di tale caratteristica vengono abitualmente dette *robuste*.
Il fatto che tali misure possano essere dette robuste, discende dal fatto che i quantili sono calcolati non già mediante un'operazione algebrica sulle singole determinazioni, bensì mediante un'operazione di conteggio delle unità statistiche.

▷ ESEMPIO 5.6

Si immagini che l'insieme dei dati individuali di una v.s. X sia $\{1, 2, 3, 4, 5\}$. Chiaramente si ha $\mu_X = 3$ e $x_{0.5} = 3$.
Si supponga ora di modificare gli ultimi due elementi dell'insieme dei dati individuali sì da avere $\{1, 2, 3, 100, 1000\}$. In questo caso la mediana non muta, cioè $x_{0.5} = 3$, mentre la media aritmetica che ne risulta è $\mu_X = 221.2$.

◁

Nello studio di una variabile statistica è abitudine, e potremmo dire buona norma, fornire i valori corrispondenti ai primi tre quartili che rappresentano le soglie non superate, rispettivamente, dal 25%, 50% e 75% delle unità statistiche.
Una misura di posizione basata sui quartili che può a volte tornare utile proporre a fianco di quelle sin'ora proposte sempre al fine della descrizione della distribuzione di una variabile statistica X è rappresentata dalla *media interquartile*, la semisomma cioè tra primo e terzo quartile

$$\frac{x_{0.25} + x_{0.75}}{2} \tag{5.8}$$

che non necessariamente coincide con la mediana.
In alcune situazioni può essere di interesse il calcolo di altri particolari quantili che possono fornire indicazioni circa le "code" della distribuzione in esame. Ad esempio il quantile di ordine $\alpha = 0.05$ corrisponde al valore che lascia alla sua sinistra il 5% delle unità statistiche, mentre il quantile di ordine $\alpha = 0.95$ rappresenta il valore che lascia il 5% delle stesse alla sua destra.

▷ ESEMPIO 5.7

Si immagini che la rilevazione di un carattere quantitativo su due popolazioni di ugual numerosità abbia dato luogo alle v.s. X e Y con distribuzioni di frequenze:

	v.s. X	v.s. Y
Classi	n_i	n_i
$100 \dashv 120$	5	15
$120 \dashv 140$	10	20
$140 \dashv 160$	20	10
$160 \dashv 180$	10	3
$180 \dashv 200$	5	2

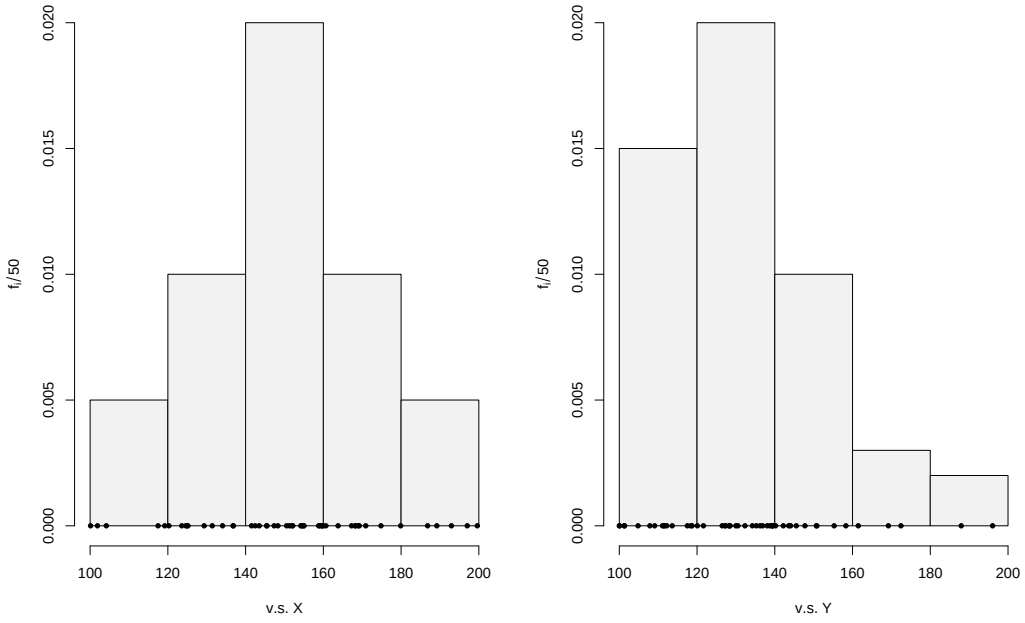


Figura 5.4 Due distribuzioni a confronto, esempio 5.7.

Nonostante esse abbiano uguale range, corrispondente all'intervallo $[100; 200]$, la *forma* della loro distribuzione è assai diversa, come peraltro evidenziato in figura (5.4). In particolare:

- ⋆ la v.s. X ha *distribuzione simmetrica*; in questo caso, come si può facilmente verificare numericamente, media e mediana coincidono (150), così come la media interquartilica (150) e la semisomma del range (150).
- ⋆ la v.s. Y ha *distribuzione asimmetrica.* In tale situazione le precedenti misure di posizione non vengono più a coincidere, infatti risulta $\mu_Y = 132.800$, $y_{0.50} = 130.000$ e media interquartile pari a 130.833 (dal momento che $y_{0.25} = 116.667$ e $y_{0.75} = 145.000$), mentre rimane immutata la semisomma del range (150).

Si osservi che per una distribuzione simmetrica varrà sempre l'ugualgianza tra media, mediana, media interquartilica e semisomma del range, ma non necessariamente è vera la proposizione inversa. L'esempio (5.8) può essere spunto di riflessione a tal proposito.

◁

5.2.3 LA MODA

Un'ultima misura di posizione è rappresentata dalla *moda* che può essere definita come la modalità $x^* \in \{x_i\}_{i=1,\dots,k}$ che si presenta con maggior frequenza.
Non desiderando trattare più in dettaglio tale argomento ci limitiamo ad alcune osservazioni:

- ⋆ se tutte le frequenze di una distribuzione sono uguali, si dice che la variabile statistica è priva di moda, potremmo dire che la distribuzione è uniforme;
- ⋆ la moda, a differenza degli altri valori medi considerati, può non essere unica; esistono v.s., dette *plurimodali*, che hanno più valori di moda, posseggono cioè modalità con la stessa frequenza che è la più "alta";
- ⋆ se la v.s. è continua con dati raccolti in classi si parla di *classe modale* come quella classe che possiede il rettangolo di massima area nell'istogramma;
- ⋆ il concetto di moda viene a volte esteso anche alle mutabili statistiche definendo per esse la moda come la modalità tra quelle osservate che si ripete più frequentemente.

▷ ESEMPIO 5.8

Da un'indagine condotta su 50 famiglie risulta la seguente distribuzione di frequenze del reddito netto mensile (v.s. X):

Classi di reddito	n_i	f_i	F_i
$900 \dashv 1000$	3	0.06	0.06
$1000 \dashv 1100$	3	0.06	0.12
$1100 \dashv 1200$	4	0.08	0.20
$1200 \dashv 1300$	9	0.18	0.38
$1300 \dashv 1400$	5	0.10	0.48
$1400 \dashv 1500$	1	0.02	0.50
$1500 \dashv 1600$	1	0.02	0.52
$1600 \dashv 1700$	3	0.06	0.58
$1700 \dashv 1800$	5	0.10	0.68
$1800 \dashv 1900$	9	0.18	0.86
$1900 \dashv 2000$	7	0.14	1.00

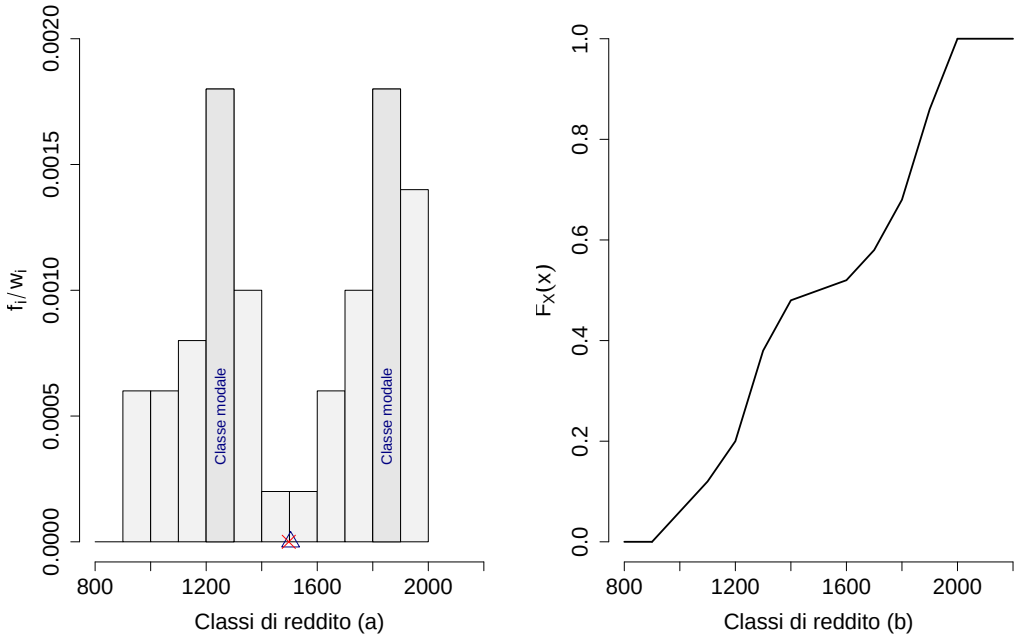


Figura 5.5 Distribuzione bimodale, esempio 5.8.

Per tale v.s. valor medio ($\mu_X = 1504.21$) e mediana ($x_{0.5} = 1500$) sono pressoché uguali. Tuttavia (cfr. figura 5.5, a) siamo in presenza di una distribuzione chiara-

mente *bimodale*, dal momento che le classi di reddito $]1200; 1300]$ e $]1800; 1900]$ presentano entrambe la massima frequenza ($f_i = 0.18$).
Da un punto di vista operativo possiamo affermare di trovarci in presenza di una "mistura" di due distribuzioni. Ciò accade sovente qualora la distribuzione di una v.s. presenti (anche se non accentuate come nel caso in esame) due o più mode. Nel nostro caso, volutamente, si sono "mescolati" i redditi di famiglie monoreddito con quelli di famiglie plurireddeito.
Si noti che a questa conclusione saremmo giunti anche analizzando l'andamento della funzione di ripartizione (cfr. figura 5.5, b), che cresce in modo uniforme tranne che nell'intervallo $[1400; 1600]$ dove è pressoché costante.
È appena il caso di osservare che nonostante il valor medio μ_X e la mediana $x_{0.50}$ siano prossimi tra loro, la distribuzione della v.s. X non può certo dirsi simmetrica.

$\triangleleft$

Fino ad ora abbiamo visto come le misure di posizione, congiuntamente a corrette rappresentazioni grafiche, consentono di cogliere alcuni aspetti delle v.s. oggetto di studio. Tuttavia esse non esauriscono l'insieme delle misure di sintesi di una distribuzione, non riuscendo da sole ad evidenziare altri aspetti assai importanti delle variabili statistiche in esame.
L'esempio che segue è introduttivo agli argomenti che verranno affrontati nel prossimo capitolo.

$\triangleright$ ESEMPIO 5.9

Si immagini che la rilevazione di un carattere quantitativo su due popolazioni di ugual numerosità abbia dato luogo alle v.s. X e Y con distribuzioni di frequenze:

	v.s. X	v.s. Y
Classi	n_i	n_i
$100 \dashv 120$	5	1
$120 \dashv 140$	10	14
$140 \dashv 160$	20	20
$160 \dashv 180$	10	14
$180 \dashv 200$	5	1

Osservando la tabella di frequenze proposta e alla luce degli istogrammi riportati in figura (5.6) che la sintetizzano, appare evidente che le due distrubuzioni, pur entrambe simmetriche, hanno "in qualche modo forma diversa".
Tale aspetto, tuttavia, non viene colto mediante le misure di posizione sino ad ora introdotte, coincidendo tutte tra loro entro la classe modale, come peraltro il Lettore può verificare per via numerica.

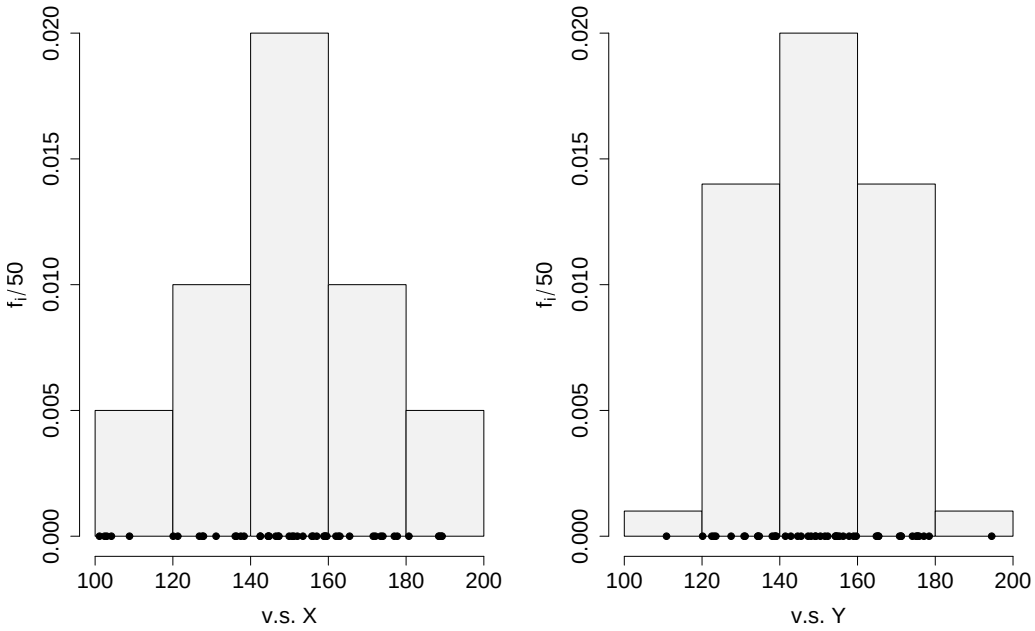


Figura 5.6 Due distribuzioni a confronto, esempio 5.9.

In sostanza le misure di posizione, pur sintetizzando alcuni aspetti delle v.s. in esame ne tralasciano altri, ai quali dedicheremo il capitolo successivo.

◁

5.3. IL FOGLIO ELETTRONICO

Consideriamo le due variabili statistiche presenti nell'ormai noto file `university.sxc` e vediamo come poter calcolare la loro media aritmetica con il foglio elettronico.
In figura (5.7) nelle celle `K9` e `K10` osserviamo la media aritmetica della variabile statistica *anni dalla laurea* calcolata rispettivamente a partire dalla distribuzione di frequenze e dai dati individuali. Il valore 2.63, che compare nella cella `K9`, è il risultato della funzione in essa inserita `=K8/H8`, dove in `K8` abbiamo inserito la funzione `=SOMMA(K4:K7)` e in `H8` vi è la funzione `=SOMMA(H4:H7)`. L'intervallo di celle `K4:K7` contiene i prodotti di ciascuna modalità per la frequenza assoluta associata, ottenuti, ad esempio per la cella `K4` inserendo la funzione `=G4*H4`. D'altro canto il valore 2.63, che compare nella cella `K10`

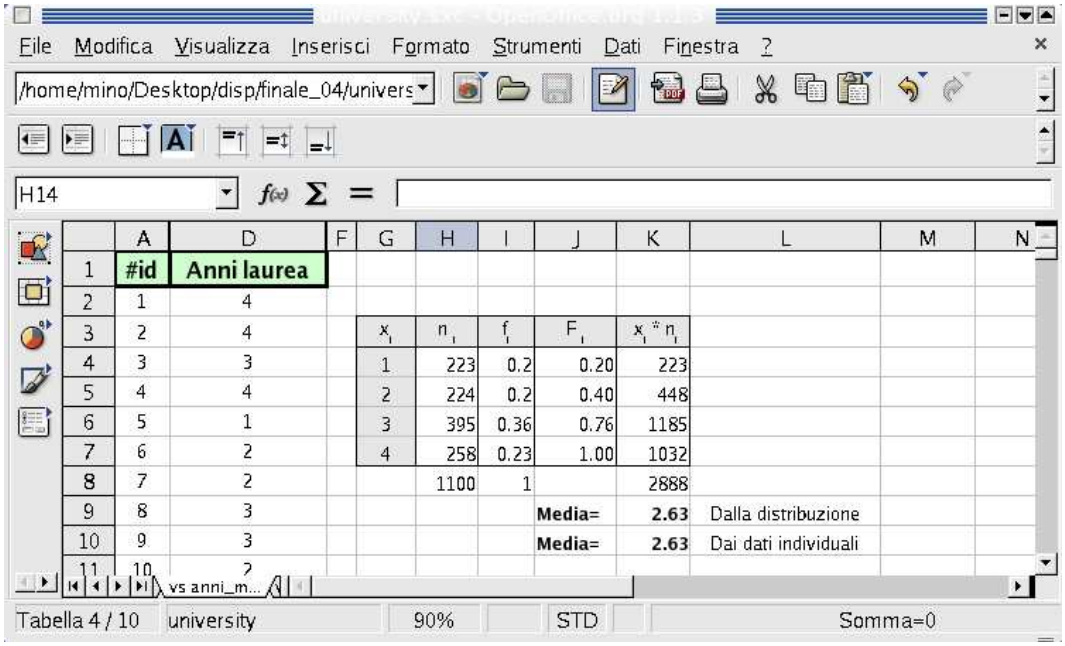


Figura 5.7 Calcolo della media aritmetica per la v.s. *anni dalla laurea.*

è il risultato della funzione interna di OpenOffice `=MEDIA(D2:D1101)`. In questo caso, ovviamente, i due risultati coincidono e la media aritmetica può essere indifferentemente calcolata in entrambi i modi.

Per quanto riguarda la variabile statistica *stipendio* proponiamo in figura (5.8) la videata del foglio elettronico che presenta la distribuzione dei dati raccolti in quattro classi e nelle celle `L9` ed `L10` i valori della media aritmetica calcolati nuovamente nei due casi di distribuzione di frequenza e di dati individuali.

Analogamente al caso della v.s. *anni dalla laurea* abbiamo inserito nel'intervallo di celle `L3:L6` i prodotti del centro di classe per la frequenza assoluta, infatti nella cella `L3` è stata inserita la funzione `=J3*K3` e così per le successive. Nelle celle `L7` e `K7` abbiamo inserito rispettivamente le funzioni `=SOMMA(L3;L6)` e `=SOMMA(K3;K6)` così il valore 1396.14 di cella `L9` è il risultato della formula `=L7/K7`. Quest'ultimo valore differisce, in questo caso, dal contenuto della cella `L10` nella quale è stata inserita la funzione `=MEDIA(E2:E1101)` poiché nel raggruppare i dati in classi, come è ormai noto, si perdono informazioni. Per il calcolo della media aritmetica in casi simili a quest'ultimo sarà quindi sempre bene ricorrere alla funzione predefinita `=MEDIA()` applicata all'intervallo di celle contenente i dati indivuduali.

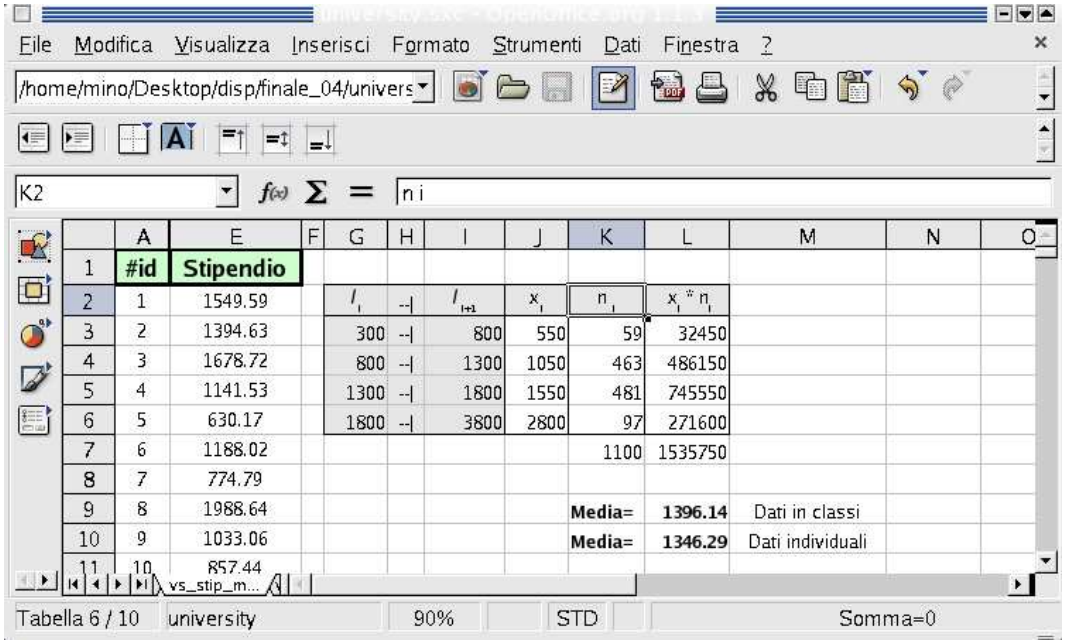


Figura 5.8 Calcolo della media aritmetica per la v.s. *stipendio.*

5.4. ESERCIZI

▷ ESERCIZIO 5.1

Si supponga che per la v.s. $X = \{\textit{numero di pagine prodotte al minuto}\}$, rilevata su un collettivo costituito da 12 stamparti laser, si abbia il seguente insieme di dati individuali

$$\{\tilde{x}_\alpha\}_{\alpha=1,\dots,12} = \{4, 5, 4, 6, 8, 8, 10, 8, 6, 5, 4, 7\}$$

Calcolare la media aritmetica, armonica, quadratica e geometrica.

◁

▷ ESERCIZIO 5.2

I valori che seguono, rappresentano le temperature (espresse in gradi Celsius) registrate alle ore 8.00 in alcuni comuni dell'Italia Nord-occidentale:

-2	2	3	-1	2	1	2	-2

Si calcolino la media aritmetica e la media geometrica.

◁

▷ ESERCIZIO 5.3

Un revisore contabile esamina le pratiche di pagamenti sinistri effettuati da una compagnia di assicurazione nel settore R.C. auto, e sottopone al responsabile dell'agenzia a cui fanno capo i clienti rimborsati, la seguente distribuzione di frequenze per la v.s. $X = \{\textit{importo liquidato}\}$, in euro:

Classi	di	Importo	n_i
0	$\dashv$	400	1200
400	$\dashv$	600	2500
600	$\dashv$	1000	1200
1000	$\dashv$	2000	980

Con riferimento alla v.s. X, calcolarne il valor medio nonché i primi tre quartili.

◁

▷ ESERCIZIO 5.4

Si supponga che per la v.s. $X = \{\textit{litri di miscela erogati settimanalmente}\}$, rilevata su un collettivo costituito dai 12 distributori di un comune, si abbia il seguente insieme di dati individuali

$$\{\tilde{x}_\alpha\}_{\alpha=1,\ldots,12} = \{92, 105, 87, 102, 110, 97, 85, 100, 115, 101, 80, 72\}$$

Calcolare la somma dei quadrati degli scarti dalla media aritmetica, nonché la mediana.

◁

▷ ESERCIZIO 5.5

I valori che seguono, rappresentano le temperature (espresse in gradi Farenheit) registrate alle ore 10.00 nel centro di Denver in una settimana di gennaio:

20	19	18	15	17	16	11

Si calcoli la media aritmetica di tali temperature. Si esprima, infine, la media delle temperature registrate espressa *in gradi Celsius*. Per inciso si ricorda la relazione $C = \frac{1}{5}(F - 32)$.

◁

▷ ESERCIZIO 5.6

Uno studente ha sotenuto in tutto 12 esami e la sua media aritmetica è 24.58, mentre la mediana dei voti è 25.50. Sapendo che il voto di un nuovo esame, il tredicesimo, è pari a 30, calcolare la nuova media e mediana dei voti.

◁

▷ ESERCIZIO 5.7

Si considerino le distribuzioni di frequenza del salario annuo lordo di 43300 operai e di 16600 operaie di una grande azienda:

Salario (classi)	*Operai* n_i	*Operaie* n_i
$30 \dashv 35$	1045	7664
$35 \dashv 40$	2465	5240
$40 \dashv 45$	4675	1066
$45 \dashv 50$	9180	1008
$50 \dashv 55$	11220	926
$55 \dashv 60$	8560	516
$60 \dashv 65$	6155	180

Per ciascuna delle distribuzioni si calcolino la media aritmetica e la mediana. Costruiti i corrispondenti istogrammi, si tenti un'iterpretazione del fenomeno in esame.

◁

CAPITOLO 6

MISURE DI VARIABILITÀ

In questo capitolo dedicato al concetto di variabilità ne verranno definte ed interpretate le più comuni sue misure. Dagli intervalli di variazione si passerà alle differenze medie fino a giungere alla varianza, alla quale sarà dedicato ampio spazio per l'importante ruolo che essa giuoca in ambito staistico. Quale strumento globale di informazione riassuntiva della distribuzione di una variabile statistica introdurremo la diseguaglianza di Tchebycev evidenziandone le potenzialità di applicazione con l'ausilio di esempi di specie. In ultimo verranno presentati alcuni tra i più comuni indici relativi di variabilità.

6.1. LA VARIABILITÀ

Le medie trattate nel modulo precedente sono parametri di sintesi per una distribuzione di frequenze, esse forniscono informazioni circa la posizione delle unità statistiche rispetto al carattere esaminato. Pur essendo parametri fondamentali nella descrizione del fenomeno di studio esse non forniscono un'informazione esaustiva dello stesso, il quale deve essere studiato anche dal punto di vista della *variabilità* dei valori che la variabile statistica assume sul collettivo considerato. Per meglio comprendere l'importanza dello studio della variabilità, valga l'esempio che segue.

▷ ESEMPIO 6.1

Una azienda fornitrice di ricambi per macchine a controllo numerico opera direttamente nelle ditte clienti attraverso due agenti rappresentanti. Nel primo semestre dell'anno in corso ciascun agente ha stipulato 45 contratti di vendita, ed entrambi hanno prodotto lo stesso fatturato complessivo di 45000 euro. Poiché il sistema provvigionale prevede incrementi proporzionali al fatturato del singolo contratto, studiamo la distribuzione degli importi fatturati per contratto dei due agenti. Considerate le v.s.

$X = \{$*fatturato dell'agente A*$\}$ e $Y = \{$*fatturato dell'agente B*$\}$ le due distribuzioni risultano:

$$X \equiv \begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\ldots,5} = \begin{Bmatrix} 600 & 800 & 1000 & 1200 & 1400 \\ 5 & 10 & 15 & 10 & 5 \end{Bmatrix}$$

$$Y \equiv \begin{Bmatrix} y_i \\ n_i \end{Bmatrix}_{i=1,\ldots,5} = \begin{Bmatrix} 200 & 600 & 1000 & 1400 & 1800 \\ 5 & 10 & 15 & 10 & 5 \end{Bmatrix}$$

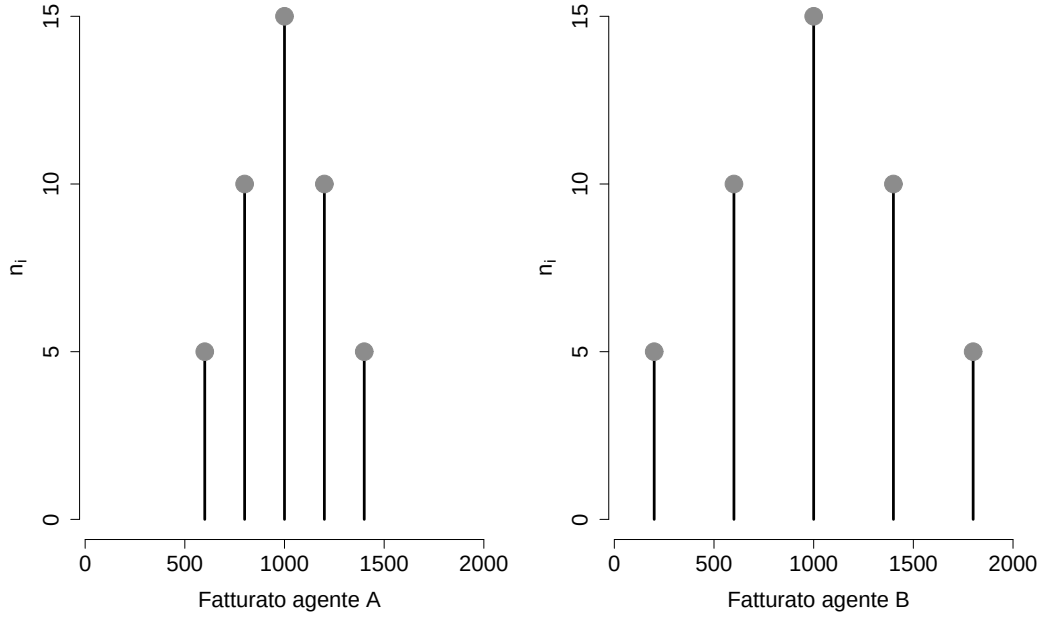


Figura 6.1 Distribuzioni del fatturato semestrale, esempio 6.1.

È immediato osservare che il fatturato medio per contratto è lo stesso per i due agenti ($\mu_X = \mu_Y = 1000$), entrambi hanno il maggior numero di contratti con importo pari a 1000 euro (le due distribuzioni hanno egual moda) ed il 50% degli ordini stipulati supera i 1000 euro (le mediane delle due distribuzioni coincidono). Potrebbe dunque sembrare che le due variabili statistiche siano del tutto equivalenti, tuttavia, se osserviamo la rappresentazione grafica delle due distribuzioni, così come evidenziate in figura (6.1), ci accorgiamo della loro diversità e comprendiamo la necessità di definire un indice che sia "in un qualche modo" discriminante delle due situazioni.

$\triangleleft$

Sorge dunque la necessità di sintetizzare la distribuzione di frequenze di una variabile statistica oltre che con misure di posizione con qualche parametro che fornisca una misura della dispersione delle unità statistiche rispetto al carattere considerato.
La variabilità di una variabile statistica può essere considerata sotto diversi aspetti e quindi valutata a mezzo di indicatori di misura differenti.
A grandi linee possiamo dire che la misura di variabilità può essere calcolata in base a tre diversi aspetti, e precisamente considerando:

- ⋆ gli *intervalli di variazione*, cioè intervalli i cui estremi corrispondono a particolari misure di posizione;
- ⋆ la *distanza che ciascun dato individuale ha con tutti gli altri*;
- ⋆ gli *scostamenti dei dati individuali da un valore medio*, scelto quale misura di posizione.

6.2. GLI INTERVALLI DI VARIAZIONE

Per quanto attiene agli intervalli di variazione, data una v.s. X con distribuzione di frequenze $\{x_i; n_i\}_{i=1,\dots,k}$, una misura di variabilità di immediata interpretazione è quella data dall'*intervallo di escursione* (o range) $[x_1; x_k]$, concetto già introdotto al capitolo precedente. Desiderando sintetizzare le informazioni fornite dalle due misure di posizione x_1 e x_k, potremmo considerare la loro differenza, cioè $x_k - x_1$, che rappresenta appunto l'ampiezza del range. Si osservi che la differenza $x_k - x_1$ testè proposta viene spesso detta "campo di escursione" e pertanto confusa con il range.
Seppur di semplice interpretazione, l'intervallo di escursione può tuttavia essere una misura poco rappresentativa della variabilità essendo pesantemente influenzato dalla presenza dei valori anomali.
Per sopperire a tali inconvenienti, si può ricorrere a misure robuste, quali ad esempio quelle basate sui quantili. Proprio in tale ottica un intervallo di variazione che meno risente del problema costituito dagli outliers è quello i cui estremi sono il primo ed il terzo quartile, in simboli $[x_{0.25}; x_{0.75}]$. Una misura di variabilità è rappresentata dalla *differenza interquartile* (d.i.), che possiamo definire quale differenza, appunto, fra il terzo ed il primo quartile, in simboli $x_{0.75} - x_{0.25}$.
Evidentemente, essa fornisce l'ampiezza dell'intervallo nel quale cade il 50% delle unità statistiche considerate eliminando così dalla misura di variabilità fornita l'influenza dei valori estremi.

Va infine ricordato che la presentazione delle sole differenze $x_k - x_1$ e $x_{0.75} - x_{0.25}$ di per sè poco dice circa la variabilità della v.s. in esame; è buona norma che siano affiancate dagli estremi degli intervalli a cui esse si riferiscono.

▷ ESEMPIO 6.2

Con riferimento all'esempio (6.1) gli intervalli di escursione per le due v.s. X e Y risultano rispettivamente $[600; 1400]$ e $[200; 1800]$. Una misura della variabilità delle due v.s. è offerta dalle loro diverse ampiezze, 800 per X e 1600 per Y.
Essendo inoltre $x_{0.25} = 800$ e $x_{075} = 1200$ per la v.s. X e $y_{0.25} = 600$ e $y_{075} = 1400$ per la v.s. Y, le differenze interquartile saranno rispettivamente d.i.$_X = 400$ e d.i.$_Y = 800$. La differenza colta ad occhio tra le due distribuzioni rappresentate graficamente in figura (6.1), imputabile alla loro diversa variabilità, viene ora riassunta in un numero, la differenza interquartile appunto. Possiamo affermare che l'agente A presenta minor variabilità nel fatturato situandosi il 50% dei contratti da lui stipulati in un intervallo di 400 euro e precisamente in $[800; 1200]$.

◁

Uno strumento grafico di semplice costruzione e tuttavia assai utile al fine di ottenere informazioni circa la variabilità e l'eventuale simmetria o asimmetria di una distribuzione è rappresentato dal cosiddetto *diagramma a scatola e baffi*, o "boxplot" in letteratura anglosassone.
Per costruire un diagramma a scatola e baffi è sufficiente rappresentare sul piano cartesiano di riferimento un rettangolo, cioè la "scatola", di altezza arbitraria, avente per base il segmento di retta i cui estremi corrispondono rispettivamente al primo ed al terzo quartile; successivamente si suddivide il rettangolo con un segmento in corrispondenza al valore della mediana. Costruita la scatola si tracciano ai suoi lati due segmenti di retta, i "baffi" appunto, con estremi rispettivamente il valore minimo ed il valore massimo dei dati.
L'interpretazione di tale grafico è piuttosto semplice, infatti solo nel caso in cui la "scatola" è perfettamente centrata sui "baffi" e la mediana poggia sul punto medio della "scatola" avremo a che fare con una distribuzione simmetrica; la minore o maggiore lunghezza dei "baffi" avverte se ci si trova in presenza di distribuzioni più o meno addensate rispetto alla mediana.

▷ ESEMPIO 6.3

Si immagini che la rilevazione su 50 clienti all'uscita di due negozi di abbigliamento (A e B) circa l'ammontare della spesa effettuata abbia dato luogo alle v.s. X e Y con distribuzione di frequenze con dati raccolti in classi:

	Negozio A	Negozio B
Classi di spesa	n_i	n_i
$30 \dashv 60$	5	10
$60 \dashv 90$	10	20
$90 \dashv 120$	20	15
$120 \dashv 150$	10	3
$150 \dashv 180$	5	2

In figura (6.2), per ciascuna v.s. sono riportati l'istogramma e il corrispondente diagramma a scatola e baffi, quest'ultimo costruito in base alle seguenti informazioni ottenute sui dati individuali, che per motivi di spazio non sono qui riportati:

	Minimo	*I Quartile*	*II Quartile*	*III Quartile*	*Massimo*
v.s. X	32.28	72.92	104.50	132.86	178.52
v.s. Y	30.42	65.99	82.66	104.57	178.86

A commento della figura (6.2) osserviamo che:

- $\star$ il diagramma a scatola e baffi conferma l'idea di asimmetria della distribuzione della v.s. Y, infatti la mediana non è centrata sulla scatola, così come lo è quella della v.s. X;
- $\star$ dal momento che la lunghezza delle scatole è diversa, saremmo propensi ad affermare che, sulla base della differenza interquartilica, la v.s. Y presenta una minor variabilità. Tale affermazione viene meno se, osservando i baffi, consideriamo l'ampiezza del range.

Si invita il Lettore a riprodurre i diagrammi a scatola e baffi calcolati sulla base della precedente tabella di frequenze.

$\triangleleft$

Osserviamo che al fine di voler evidenziare l'eventuale presenza di outliers o mitigare l'influenza di quelle che vengono comunemente dette *code* della distribuzione, si potrebbero modificare gli estremi dei due "baffi" sì che essi corrispondano, ad esempio, alle coppie di percentili di ordine, rispettivamente, 5% e 95%.
Tale è la proposta di default prevista dalla maggior parte dei software statistici.

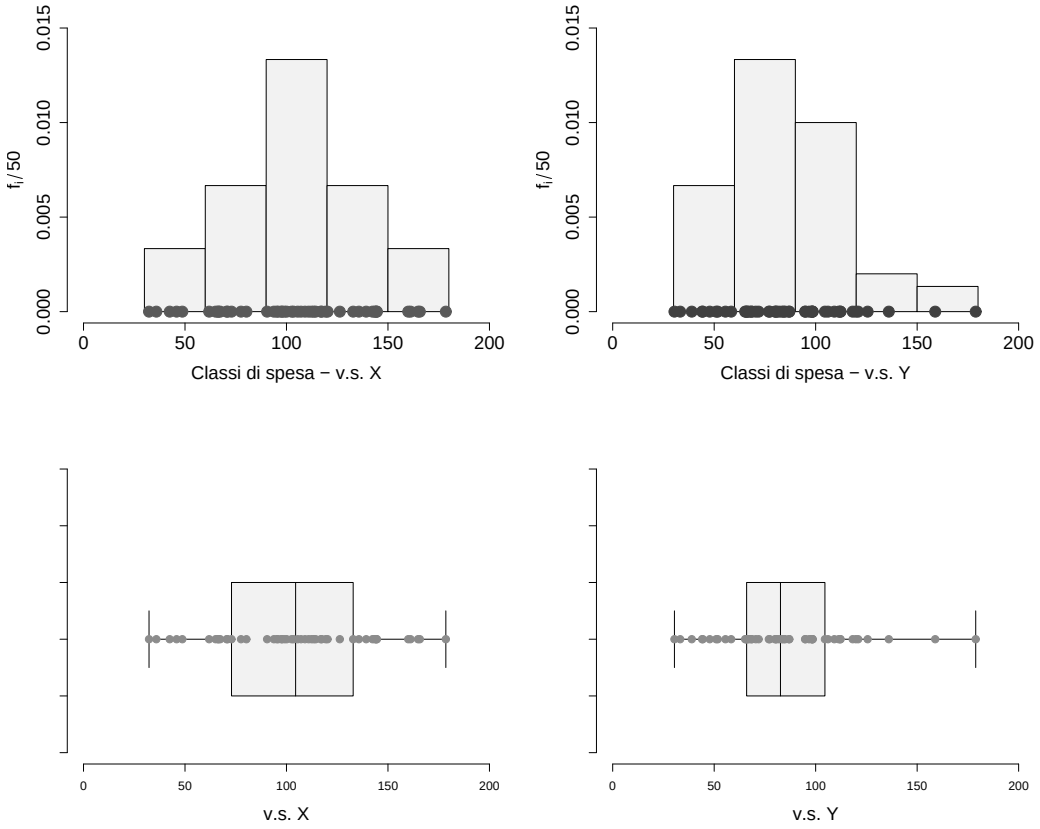


Figura 6.2 Distribuzioni della spesa per abbigliamento, esempio 6.3.

▷ ESEMPIO 6.4

Riprendendo la situazione descritta all'esempio (6.3), il grafico proposto in figura (6.3, a) riporta i diagrammi a scatola e baffi con estremi i percentili di ordine, rispettivamente, $\alpha = 0.05$ e $\alpha = 0.95$, calcolati, come si disse in base ai dati individuali, per cui:

	Percentile 5%	*Percentile* 95%
v.s. X	43.99	163.17
v.s. Y	41.25	131.35

Evidentemente le informazioni offerte dai due diagrammi a scatola e baffi sono le stesse di quelle quelle fornite dai grafici di figura (6.2). Si noti come la differenza tra

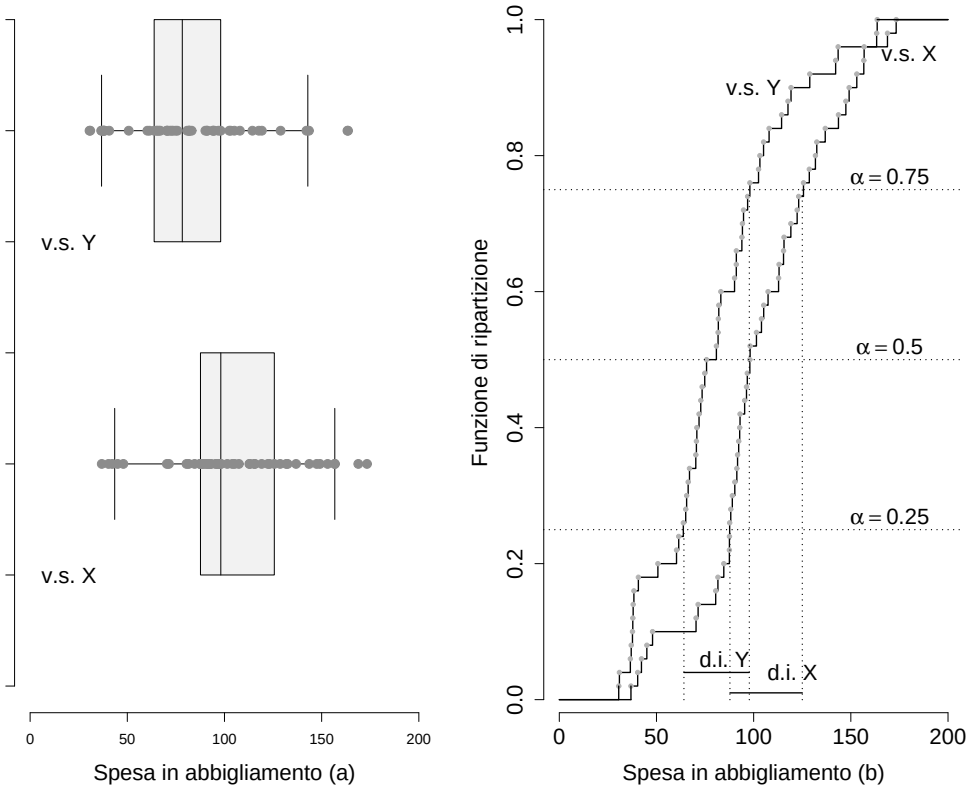


Figura 6.3 Boxplot e funzioni di ripartizione, esempio 6.4.

i due percentili sia diversa per le due variabili a conferma della maggior variabilità di X. Il baffo destro così calcolato risultando più corto evidenzia come la v.s. Y presenti una coda destra assai accentuata, a prova della sua asimmetria.

Sempre in figura (6.3, b) abbiamo riportato il grafico delle funzioni di riaprtizione dei dati individuali delle v.s. X e Y. Interessante è notare come le informazioni fornite dai precedenti boxplot possono essere tratte da un'attenta lettura del grafico delle corrispondenti funzioni di ripartizione.

◁

6.3. LE DIFFERENZE MEDIE ASSOLUTE

La variabilità di una v.s. X può essere valutata sulla base delle differenze, in valore assoluto, che ciascun dato ha rispetto a tutti gli altri sintetizzando tali informazioni mediante la loro media aritmetica.
Più precisamente, posto $\{\tilde{x}_\alpha\}_{\alpha=1,\dots,n}$ l'insieme dei dati individuali, se organizziamo le differenze $|\tilde{x}_\alpha - \tilde{x}_\beta|$, con ovviamente $\alpha, \beta = 1, \dots, n$, nella forma tabellare

	$\tilde{x}_1$	$\tilde{x}_2$	$\dots$	$\tilde{x}_\alpha$	$\dots$	$\tilde{x}_n$
$\tilde{x}_1$	$\lvert\tilde{x}_1 - \tilde{x}_1\rvert$	$\lvert\tilde{x}_1 - \tilde{x}_2\rvert$	$\dots$	$\lvert\tilde{x}_1 - \tilde{x}_\alpha\rvert$	$\dots$	$\lvert\tilde{x}_1 - \tilde{x}_n\rvert$
$\tilde{x}_2$	$\lvert\tilde{x}_2 - \tilde{x}_1\rvert$	$\lvert\tilde{x}_2 - \tilde{x}_2\rvert$	$\dots$	$\lvert\tilde{x}_2 - \tilde{x}_\alpha\rvert$	$\dots$	$\lvert\tilde{x}_2 - \tilde{x}_n\rvert$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\tilde{x}_\alpha$	$\lvert\tilde{x}_\alpha - \tilde{x}_1\rvert$	$\lvert\tilde{x}_\alpha - \tilde{x}_2\rvert$	$\dots$	$\lvert\tilde{x}_\alpha - \tilde{x}_\alpha\rvert$	$\dots$	$\lvert\tilde{x}_\alpha - \tilde{x}_n\rvert$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
$\tilde{x}_n$	$\lvert\tilde{x}_n - \tilde{x}_1\rvert$	$\lvert\tilde{x}_n - \tilde{x}_2\rvert$	$\dots$	$\lvert\tilde{x}_n - \tilde{x}_\alpha\rvert$	$\dots$	$\lvert\tilde{x}_n - \tilde{x}_n\rvert$

ricaviamo da essa una misura di variabilità semplicemente calcolando la media aritmetica delle n^2 differenze assolute che vi compaiono. Pertanto offriamo la seguente

Definizione 6.1 (Differenza media assoluta con ripetizione)
definiamo differenza media assoluta con ripetizione della v.s. X la media aritmetica delle n^2 differenze in valore assoluto tra ciascun dato individuale e gli altri, cioè

$$\Delta_R = \frac{1}{n^2} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} |\tilde{x}_\alpha - \tilde{x}_\beta| \tag{6.1}$$

□

▷ ESEMPIO 6.5

I dati che seguono si riferiscono al numero di pratiche espletate in un determinato giorno dai nove dipendenti di un Ufficio del Pubblico Registro

20	10	30	20	10	10	30	20	20

Al fine del calcolo della differenza media assoluta con ripetizione, consideriamo la tabella delle differenze:

	20	10	30	20	10	10	30	20	20
20	0	10	10	0	10	10	10	0	0
10	10	0	20	10	0	0	20	10	10
30	10	20	0	10	20	20	0	10	10
20	0	10	10	0	10	10	10	0	0
10	10	0	20	10	0	0	20	10	10
10	10	0	20	10	0	0	20	10	10
30	10	20	0	10	20	20	0	10	10
20	0	10	10	0	10	10	10	0	0
20	0	10	10	0	10	10	10	0	0

Dal momento che $\sum_{\alpha=1}^{9} \sum_{\beta=1}^{9} |\tilde{x}_\alpha - \tilde{x}_\beta| = 640$, applicando la (6.1) otteniamo $\Delta_R = 7.9012$.

◁

Se disponessimo, in luogo della successione dei dati individuali, della distribuzione di frequenze assolute della v.s., ciascun elemento della matrice delle differenze $|x_i - x_j|$, con $i, j = 1, \ldots, k$, verrebbe ad essere moltiplicato per il prodotto delle rispettive frequenze $n_i\, n_j$ ed in accordo con la definizione posta in (6.1) avremmo:

$$\Delta_R = \frac{1}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{k} |x_i - x_j|\, n_i\, n_j \tag{6.2}$$

A ben vedere, l'espressione per Δ_R può essere scritta in una forma alternativa. Se osserviamo, infatti, che la tabella delle differenze $|x_i - x_j|$ presenta tutti zeri sulla diagonale principale e che gli elementi del triangolo inferiore sono uguali a quelli del triangolo superiore, abbiamo allora:

$$\Delta_R = \frac{2}{n^2} \sum_{i=1}^{k} \sum_{j=1}^{i} (x_i - x_j)\; n_i\, n_j$$

A tal proposito valga l'esempio che segue.

▷ ESEMPIO 6.6

Se osserviamo che le nove osservazioni individuali di cui all'esempio (6.5) possono dare luogo alla seguente distribuzione di frequenze assolute:

$$\begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,2,3} = \begin{Bmatrix} 10 & 20 & 30 \\ 3 & 4 & 2 \end{Bmatrix} \tag{6.3}$$

al fine del calcolo della differenza media assoluta con ripetizione, consideriamo:

x_i ; n_i	x_i	10	20	30
	n_i	3	4	2
10 ; 3		0	10	20
20 ; 4		10	0	10
30 ; 2		20	10	0

e calcoliamo le differenze medie con ripetizione come:

$$\Delta_R = \frac{0 + 120 + 120 + 120 + 0 + 80 + 120 + 80 + 0}{9^2} = 7.9012$$

oppure

$$\Delta_R = \frac{2}{9^2}\,(10 \cdot 3 \cdot 4 + 20 \cdot 3 \cdot 2 + 10 \cdot 4 \cdot 2) = 7.9012$$

◁

La differenza media assoluta con ripetizione considera, fra le altre, le n differenze che ciascun termine ha con se stesso; desiderando derivare una misura di variabilità che non tenga conto di tali differenze, peraltro nulle, sarà sufficiente considerare quale denominatore della (6.1) o della (6.2) la quantità $n\,(n-1)$.
Definiamo pertanto per la v.s. X *differenza media assoluta senza ripetizione* la media aritmetica delle $n\,(n-1)$ differenze in modulo tra ciascuna osservazione individuale e le restanti, cioè:

$$\Delta = \frac{1}{n\,(n-1)} \sum_{\alpha=1}^{n} \sum_{\beta=1}^{n} |\tilde{x}_\alpha - \tilde{x}_\beta| \tag{6.4}$$

Si osservi che Δ e Δ_R differiscono tra loro unicamente per la diversa quantità che compare a denominatore; come il Lettore può facilmente verificare, risulta inoltre l'uguaglianza:

$$\Delta = \frac{n}{n-1}\,\Delta_R$$

6.4. La variabilità rispetto ad un valore medio

La variabilità di una variabile statistica può essere intesa come dispersione dei dati attorno ad un valore medio assunto come posizione centrale. In tale contesto, scelto il valore medio rispetto al quale si vuole misurare la dispersione, pare logico individuare una media degli scostamenti dei singoli dati dal valore medio di riferimento.

Così, ad esempio, potremmo scegliere quale misura di posizione la *mediana* e, individuata la distribuzione degli scarti in valore assoluto tra ciascuna modalità e la mediana stessa, calcolarne la media aritmetica, in simboli $n^{-1}\sum_{i=1}^{k}|x_i - x_{0.5}|\,n_i$. Perverremmo in tal modo ad una misura di variabilità detta *scostamento medio semplice dalla mediana*. Trattasi di una misura di variabilità che esprime quanto in media le modalità si discostano dalla mediana.
Tuttavia, per quanto riguarda la dispersione dei dati attorno ad un valore medio, la misura di variabilità di gran lunga più impiegata, per diversi motivi che appariranno chiari nel seguito, è senza dubbio la *varianza*. Data l'importanza di tale parametro, dedichiamo ad esso i successivi due paragrafi.

6.4.1 La varianza e lo scarto quadratico medio

Un'interessante misura di variabilità la si può ottenere considerando la dispersione dei dati attorno alla media aritmetica. Tale dispersione potrebbe essere individuata calcolando gli scarti dalla media aritmetica delle diverse modalità osservate e sintetizzando la distribuzione degli scarti mediante la loro media aritmetica. Tuttavia per una nota proprietà della media aritmetica, la somma degli scarti è nulla e di conseguenza nulla sarà la loro media. Un modo per ovviare a tale inconveniente è quello di considerare gli scarti al quadrato. In tal modo non solo la loro media sarà in generale diversa da zero ma si viene in qualche modo a "penalizzare" le modalità più distanti dal baricentro della distribuzione.
Ciò premesso, diamo la seguente

Definizione 6.2 (Varianza)
si dice varianza di una variabile statistica X la media aritmetica del quadrato degli scarti di ogni singola modalità dalla media aritmetica della v.s. Essa corrisponde pertanto al valore numerico risultante dall'operazione

$$\begin{aligned} V[X] &= \frac{1}{n}\sum_{\alpha=1}^{n}(\tilde{x}_\alpha - \mu_X)^2 = \frac{1}{n}\sum_{i=1}^{k}(x_i - \mu_X)^2\, n_i \\ &= \sum_{i=1}^{k}(x_i - \mu_X)^2\, f_i = \sigma_X^2 \end{aligned} \tag{6.5}$$

□

Anche in questo caso, come in quello della media aritmetica, abbiamo introdotto un operatore, $V[\,\cdot\,]$, che applicato ad una qualunque v.s. fornisce uno ed uno solo numero reale, la varianza, appunto, che viene solitamente indicata con σ_X^2, o più semplicemente con σ^2 qualora non sussistano dubbi di ambiguità con altre variabili statistiche.

Osserviamo che, dalla definizione(6.2), l'operatore $V[\,\cdot\,]$, che applicato ad una qualunque v.s. X ne fornisce la varianza, può essere espresso in termini di operatore $E[\,\cdot\,]$ come segue:

$$V[X] = E[(X - E[X])^2] \tag{6.6}$$

La varianza si ottiene pertanto facendo la media aritmetica di una trasformata della v.s., più precisamente, posto $Y = (X - E[X])^2$, si ha $V[X] = E[Y]$.

OSSERVAZIONE: è bene tenere a mente che la grandezza:

$$\sum_{\alpha=1}^{n}(\tilde{x}_\alpha - \mu_X)^2 = \sum_{i=1}^{k}(x_i - \mu_X)^2 n_i$$

viene detta *Devianza* della v.s. X, in simboli $Dev[X]$, e che per essa, evidentemente, si ha $Dev[X] = n\,\sigma_X^2$.

$\star$

Una misura della variabilità, strettamente legata alla varianza ed espressa nella stessa unità di misura delle modalità della v.s. in esame, è data dallo *scarto quadratico medio.* Potremmo dire che esso corrisponde alla media quadratica degli scarti tra ogni modalità osservata e la corrispondente media aritmetica. Più precisamente:

Definizione 6.3 (Scarto Quadratico Medio)
si dice scarto quadratico medio la radice quadrata della varianza, cioè

$$\sigma_X = \sqrt{\sigma_X^2}$$

□

▷ ESEMPIO 6.7

Si immagini che la v.s. $X = \{$*km percorsi con un litro di benzina*$\}$, definita a partire dalla rilevazione della percorrenza di 45 autovetture di piccola cilindrata alimentate a benzina, possegga distribuzione di frequenze assolute:

$$\begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\ldots,5} = \begin{Bmatrix} 18.5 & 19.5 & 20.5 & 21.5 & 22.5 \\ 5 & 10 & 15 & 10 & 5 \end{Bmatrix}$$

e quindi valor medio $\mu_X = 20.5$ km. Dalla definizione di varianza, calcoliamo:

$$V[X] = \frac{1}{n}\sum_{i=1}^{k}(x_i - \mu_X)^2\, n_i = \\ = \frac{(18.5-20.5)^2\,5 + (19.5-20.5)^2\,10 + \ldots + (22.5-20.5)^2\,5}{45} = \\ = \frac{60}{45} = 1.33\bar{3} = \sigma_X^2$$

Naturalmente, nel caso in esame, mentre le singole modalità e dunque la loro media sono espresse in km, la varianza è espressa in km^2 e pertanto non è direttamante paragonabile, ad esempio, con la media. Se consideriamo lo scarto quadratico medio $\sigma_X = \sqrt{\sigma_X^2} = 1.1547$, espresso in km, possiamo affermare che in media le autovetture hanno una percorrenza di 20.5 km con un litro di carburante ma che esse da tale valore si discostano in media di 1.1547 km.

◁

6.4.2 PRINCIPALI PROPRIETÀ DELLA VARIANZA

Passiamo in rassegna alcune proprietà della varianza immaginando che la variabile statistica X abbia distribuzione di frequenze assolute $\{x_i; n_i\}_{i=1,\ldots,k}$ ovvero distribuzione di frequenze relative $\{x_i; f_i\}_{i=1,\ldots,k}$.

Proprietà 6.1 La varianza è una quantità positiva o nulla, cioè $V[X] = \sigma_X^2 \geq 0$.

◁

Dimostrazione: è sufficiente osservare che in quanto media aritmetica di quadrati non può assumere valori negativi ($\sigma_X^2 > 0$), mentre è uguale a zero qualora la v.s. X assumesse modalità tutte uguali tra loro e quindi eguali alla media aritmetica.

□

Proprietà 6.2 La varianza può essere espressa come differenza tra due valori medi al quadrato, infatti vale la relazione $V[X] = E[X^2] - (E[X])^2$.

◁

Dimostrazione: dalla definizione di varianza e ricordando le proprietà dell'operatore $E[\,\cdot\,]$, si ha:

$$V[X] = E[(X - E[X])^2] = E[X^2 - 2\,E[X]\,X + (E[X])^2] = \\ = E[X^2] - 2\,E[X]\,E[X] + (E[X])^2 = E[X^2] - (E[X])^2 \qquad (6.7)$$

Trattasi, come vedremo, di una proprietà di utile impiego anche nelle operazioni di calcolo numerico.

□

Proprietà 6.3 La varianza è il valore assunto dalla funzione $\varphi(a) = \sum_{i=1}^{k}(x_i - a)^2 f_i$ nel suo punto di minimo.

◁

Dimostrazione: la funzione $\varphi(a)$ ha un minimo in corrispondenza ad $a = \mu_X$ (cfr. la proprietà 5.2 del capitolo precedente) e pertanto è dimostrato che essa assume valore pari alla varianza in tale punto.

□

▷ ESEMPIO 6.8

Con riferimento alla v.s. X di cui all'esempio (6.7) con distribuzione di frequenze assolute:

$$\begin{Bmatrix} x_i \\ n_i \end{Bmatrix}_{i=1,\ldots,5} = \begin{Bmatrix} 18.5 & 19.5 & 20.5 & 21.5 & 22.5 \\ 5 & 10 & 15 & 10 & 5 \end{Bmatrix}$$

e valor medio $\mu_X = 20.5$ km, sfruttando la proprietà (6.2) abbiamo:

$$\begin{aligned} V[X] = E[X^2] - (E[X])^2 &= \frac{1}{n}\sum_{i=1}^{k} x_i^2\, n_i - \mu_X^2 = \\ &= \frac{18.5^2\,5 + 19.5^2\,10 + 20.5^2\,15 + 21.5^2\,10 + 22.5^2\,5}{45} - 20.5^2 = \\ &= 421.5833 - 420.2500 = 1.33\bar{3} = \sigma_X^2 \end{aligned}$$

Tale è il modo in cui conviene calcolare la varianza di una v.s., poiché oltre a snellire i calcoli, esso riduce il problema delle approssimazioni successive.

◁

Come abbiamo già detto a proposito della media aritmetica, a volte ci si trova a dover lavorare con una trasformata della variabile statistica X originaria e sotto condizioni assai generali tale trasformata costituisce una nuova variabile statistica.
Se si considera la trasformata lineare $Y = a + b\,X$, con $a, b \in \mathbb{R}$, per proprietà ormai note si ha $E[Y] = a + b\,E[X]$ e ci si potrebbe chiedere se non sussita una proprietà analoga per la varianza della nuova v.s. Y. A ciò sopperisce la seguente:

Proprietà 6.4 data la v.s. X con media μ_X e varianza σ_X^2, la varianza della trasformata $Y = a + b\,X$, con $a, b \in \mathbb{R}$ è data da $V[Y] = b^2\,V[X]$.

◁

Dimostrazione: sfruttando la proprietà (6.2), data la trasformata $Y = a + b\,X$ e ricorrendo alle proprietà dell'operatore $E[\,\cdot\,]$ si ha:

$$\begin{aligned} V[Y] &= E[Y^2] - E[Y]^2 = E[(a + b\,X)^2] - (E[a + b\,X])^2 \\ &= E[a^2] + E[2\,a\,b\,X] + E[b^2\,X^2] - (a + b\,E[X])^2 = \\ &= a^2 + 2\,a\,b\,E[X] + b^2\,E[X^2] - a^2 - 2\,a\,b\,E[X] - b^2\,(E[X])^2 = \\ &= b^2\,E[X^2] - b^2\,(E[X])^2 = b^2\left(E[X^2] - (E[X])^2\right) = b^2\,V[X] \end{aligned} \tag{6.8}$$

□

▷ ESEMPIO 6.9

Si immagini che l'ufficio amministrativo di un'azienda meccanica disponga della v.s. $X = \{\textit{numero di ore di straordinario}\}$, effettuate nell'ultimo mese dai suoi 24 addetti al settore montaggio, con insieme dei dati individuali

$$\begin{aligned} \{\tilde{x}_\alpha\}_{\alpha=1,\ldots,24} = \{&10, 10, 11, 3, 4, 8, 7, 9, 9, 16, 19, 18, \\ &9, 2, 3, 2, 12, 13, 15, 12, 14, 8, 6, 7\} \end{aligned}$$

Volendo determinare l'importo medio mensile e il rispettivo scarto quadratico medio della paga dei 24 dipendenti, sarà sufficiente, sapendo che la paga base è di 800 euro mensili e che un'ora di straodinario viene pagata 25 euro, calcolare media e varianza della trasformata $Y = 800 + 25\,X$. Ricordando che $E[Y] = 800 + 25\,E[X]$ e, per la proprietà appena dimostrata, $V[Y] = 25^2\,V[X]$, avremo:

$$\begin{aligned} E[Y] &= 800 + 25\,9.458 = 1036.458 \\ V[Y] &= 25^2\,22.498 = 14672.78 \end{aligned}$$

e dunque lo scarto quadratico medio ricercato sarà pari a 121.13 euro.

◁

Concludiamo questo paragrafo sulla varianza di una v.s. definendo il *procedimento di standardizzazione*, che può talvolta tornare utile nello studio di variabili statistiche. A tal proposito, valga la seguente

Proprietà 6.5 data una v.s. X con valor medio μ_X e varianza σ_X^2, la trasformazione lineare:

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{6.9}$$

fornisce una variabile statistica con media nulla e varianza unitaria, per cui $E[Z] = 0$ e $V[Z] = 1$.

◁

Dimostrazione: è sufficiente applicare le proprietà degli operatori $E[\cdot]$ e $V[\cdot]$ alla trasformata linerare $Z = \frac{1}{\sigma_X} X - \frac{\mu_X}{\sigma_X}$.

□

▷ ESEMPIO 6.10

La tabella che segue riporta la distribuzione di frequenze assolute, con dati raccolti in classi di modulo costante, della variabile statistica X:

Classi v.s. X	n_i
$1 \dashv 3$	5
$3 \dashv 5$	10
$5 \dashv 7$	20
$7 \dashv 9$	5

Se consideriamo la v.s. standardizzata $Z = \sigma_X^{-1}(X - \mu_X)$, per essa, in base alla proprietà (6.5), si ha $E[Z] = 0$ e $V[Z] = 1$.
Se si osservano gli istogrammi delle v.s. X e Z riportati in figura (6.4), appare evidente come la standardizzazione non solo effettui una traslazione di locazione della v.s. X ($\mu_X = 5.25 \rightarrow 0 = \mu_Z$), ma altresì una trasformazione di scala ($\sigma_X^2 = 2.9375 \rightarrow 1 = \sigma_Z^2$).
Ciò è dovuto all'effetto che la standardizzazione ha sulle classi. Queste, pur mantenendosi a modulo costante, mutano di ampiezza, ad esempio la seconda classe in termini di trasformata Z corrisponde all'intervallo:

$$\left]\frac{3 - 5.25}{1.7139}; \frac{5 - 5.25}{1.7139}\right] = [-1.3128; -0.1459]$$

◁

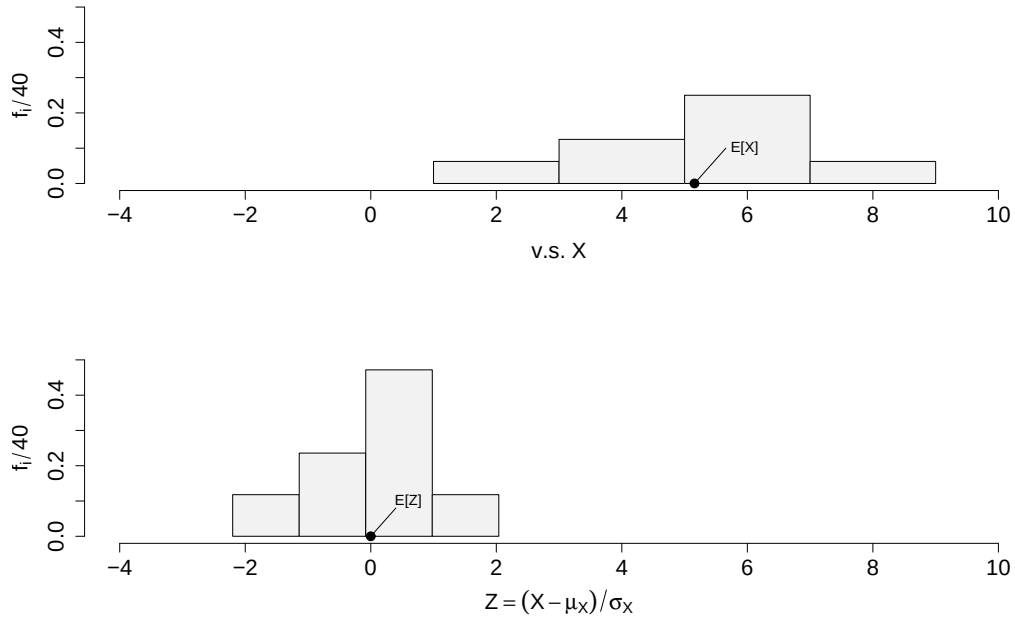


Figura 6.4 Istogrammi delle v.s. X e $Z = \sigma_X^{-1}\,(X - \mu_X)$, esempio 6.10.

6.5. LA DISEGUAGLIANZA DI TCHEBYCHEV

Nello studio di una variabile statistica, il ricorso al valore medio ed alla varianza per sintetizzarne in due soli numeri il comportamento può essere in qualche modo giustificato dall'esistenza di una celebre diseguaglianza, appunto la *diseguaglianza di Tchebychev*.
Prima di proporre l'enunciato del teorema, consideriamo un semplice esempio. Si immagini che, data una v.s. X definita a partire da un collettivo Ω, ci si ponga il problema di calcolare quante sono le unità statistiche ω_α in corrispondenza alle quali la v.s. in esame assume valori appartenenti ad un intervallo di interesse, ad esempio l'intervallo simmetrico attorno al valor medio $[\mu_X - k\,\sigma_X; \mu_X + k\,\sigma_X]$, con $k > 0$. Possedendo l'insieme dei dati individuali, o perlomeno la distribuzione di frequenze, della v.s. in esame il problema sarebbe di facile soluzione risolvendosi in un semplice conteggio e in ciò potrebbe essere di ausilio il ricorso alla funzione di ripartizione. Viceversa, non disponendo, come a volte capita, che dei due valori di sintesi μ_X e σ_X, il problema parrebbe insolubile. Con la diseguaglianza di Tchebychev, tuttavia, è possibile individuare una soglia inferiore per la proporzione di unità statistiche che appartengono all'intervallo e ciò indipendentemente

dalla conoscenza della distribuzione in esame.

Teorema 6.1 (Diseguaglianza di Tchebychev)
Data una generica v.s. X con valore medio μ_X e varianza σ_X^2, qualunque sia la forma della sua distribuzione, la proporzione di unità statistiche per cui $X(\omega_\alpha)$ risulta esterno all'intervallo $[\mu_X - k\,\sigma_X; \mu_X + k\,\sigma_X]$ è non maggiore di k^{-2}, con $k \in \mathbb{R}^+$. In simboli:

$$\frac{Nu\left\{\omega_\alpha : |X(\omega_\alpha) - \mu_X| > k \cdot \sigma_X\right\}}{Nu(\Omega)} \leq \frac{1}{k^2} \tag{6.10}$$

♢

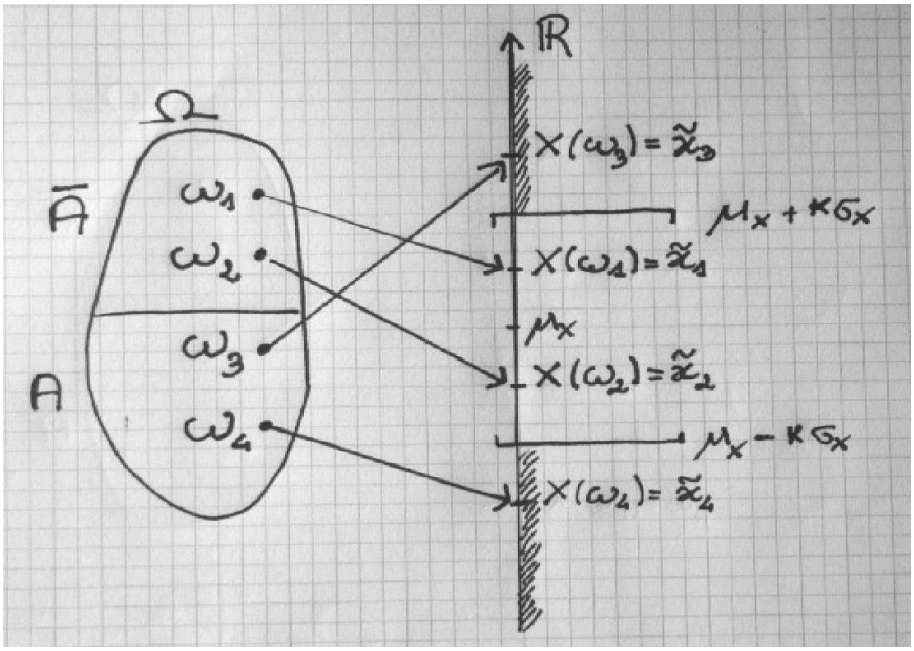


Figura 6.5 Partizione di Ω per la dimostrazione della diseguaglianza di Tchebycev.

Dimostrazione: per la dimostrazione il trucco consiste nel maggiorare la varianza della v.s. X. A tal fine si individua una partizione del collettivo Ω in due sottoinsiemi e si nota che la varianza è sempre maggiore (o al più uguale) alla media degli scarti al quadrato dei valori assunti dalle unità statistiche che appartengono ad uno solo dei due insiemi della

partizione. Dal un punto di vista formale, definiamo i sottoinsiemi:

$$A = \{\omega_\alpha : |X(\omega_\alpha) - \mu_X| > k \cdot \sigma_X\}$$
$$\bar{A} = \{\omega_\alpha : |X(\omega_\alpha) - \mu_X| \leq k \cdot \sigma_X\}$$

e osserviamo (cfr. figura 6.5) che A è costituito da tutti gli elementi ω_α in corrispondenza ai quali la v.s. X assume determinazioni esterne all'intervallo $[\mu_X - k\,\sigma_X; \mu_X + k\,\sigma_X]$. Vista la partizione indotta su Ω dagli insiemi A e $\bar{A}$, la varianza di X può essere scritta quale somma di due addendi e precisamente:

$$\begin{aligned}\sigma_X^2 &= n^{-1} \sum_{\omega_\alpha \in \Omega} \left(X\left(\omega_\alpha\right) - \mu_X\right)^2 = \\ &= n^{-1} \sum_{\omega_\alpha \in A} \left(X\left(\omega_\alpha\right) - \mu_X\right)^2 + n^{-1} \sum_{\omega_\alpha \in \bar{A}} \left(X\left(\omega_\alpha\right) - \mu_X\right)^2 \end{aligned} \tag{6.11}$$

Essendo i due addenti che compongono la (6.11) quantità non negative, la varianza sarà certamente maggiore (o al più uguale) ad uno solo dei due e ricordando la definizione dell'insieme A, segue pertanto che:

$$\sigma_X^2 \geq n^{-1} \sum_{\omega_\alpha \in A} \left(X\left(\omega_\alpha\right) - \mu_X\right)^2 \geq n^{-1} \sum_{\omega_\alpha \in A} k^2 \cdot \sigma_X^2 \tag{6.12}$$

Prima di procedere, osserviamo che, con la notazione ora introdotta, la tesi (6.10) può essere riscritta come:

$$\frac{Nu(A)}{n} \leq \frac{1}{k^2}$$

Riprendendo ora la (6.12) ed osservando che:

$$\sum_{\omega_\alpha \in A} k^2 \cdot \sigma_X^2 = k^2 \cdot \sigma_X^2 \sum_{\omega_\alpha \in A} 1 = k^2 \cdot \sigma_X^2 \cdot Nu(A)$$

risulterà, a prova della (6.10):

$$\sigma_X^2 \geq k^2 \cdot \sigma_X^2 \frac{Nu(A)}{n} \qquad \rightarrow \qquad \frac{Nu(A)}{n} \leq \frac{1}{k^2}$$

□

Da un punto di vista applicativo, la sola conoscenza di μ_X e di σ_X ci consente, ad esempio, di affermare che all'esterno dell'intervallo $[\mu_X - k\,\sigma_X; \mu_X + k\,\sigma_X]$:

- ⋆ cade non più del 25% delle frequenze se poniamo $k = 2$;
- ⋆ cade non più del 11% delle frequenze se poniamo $k = 3$.

Se la diseguaglianza di Tchebychev è stata dimostrata per qualunque k reale positivo, da un punto di vista operativo solo alcuni valori di k hanno utilizzo. Così ad esempio ponendo $k = 1$ la diseguaglianza porterà ad affermare che non più del 100% delle unità statistiche assume valori esterni all'intervallo $[\mu_X - \sigma_X; \mu_X + \sigma_X]$; tale affermazione, pur essendo corretta, non apporta alcuna utile informazione.

Prima di passare ad alcune applicazioni, osserviamo che la diseguaglianza di Tchebychev porge informazioni anche circa le unità che assumono valori all'interno dell'intervallo. Infatti dalla (6.10) segue che, qualunque sia la forma della distribuzione di una v.s, la proporzione di unità statistiche appartenenti all'intervallo $[\mu_X - k\,\sigma_X; \mu_X + k\,\sigma_X]$ è maggiore di $1 - k^{-2}$.

▷ ESEMPIO 6.11

Sia X una variabile statistica con $\mu_X = 50$ e $\sigma_X = 1$, definita su un collettivo di $n = 72$ unità. Se poniamo ad esempio $k = 1.2$, dalla diseguaglianza di Tchebychev segue che non più di $72 \cdot 1.2^{-2} = 50$ unità cadono all'esterno dell'intervallo $[48.8; 51.2]$, cioè non più del 69.44% delle unità del colletivo assumo valori esterni a tale intervallo. Ne consegue che per più di 22 unità osserveremo valori interni.
Con i soli due parametri, media e scarto quadratico medio, riusciamo a ottenere informazioni sulla distribuzione delle unità statistiche rispetto alla variabile in esame. Tuttavia l'informazione fornitaci dalla diseguaglianza di Tchebychev non ci consente di conoscere come la proporzione di unità statistiche esterne all'intervallo sia in realtà ripartita nelle code della distribuizione. In figura (6.6) sono proposte quattro diverse ipotetiche situazioni che si accordano tutte con i parametri $\mu_X = 50$ e $\sigma_X = 1$ e per le quali notiamo che l'area dell'istogramma esterna all'intervallo $[48.8; 51.2]$ è non maggiore del 69.44% dell'area totale. Si può notare che solo nei casi (a) e (d) l'area esterna all'intervallo di Tchebychev è equamente ripartita nelle due code; ciò non accade nelle situazioni (b) e (c) dove la distribuzione è ipotizzata non essere simmetrica.

◁

▷ ESEMPIO 6.12

Sia X una v.s. con $\mu_X = 34.42$ e $\sigma_X = 18.54$, definita su un collettivo di $n = 24$ unità statistiche. Desiderando conoscere qual'è il numero delle unità statistiche che

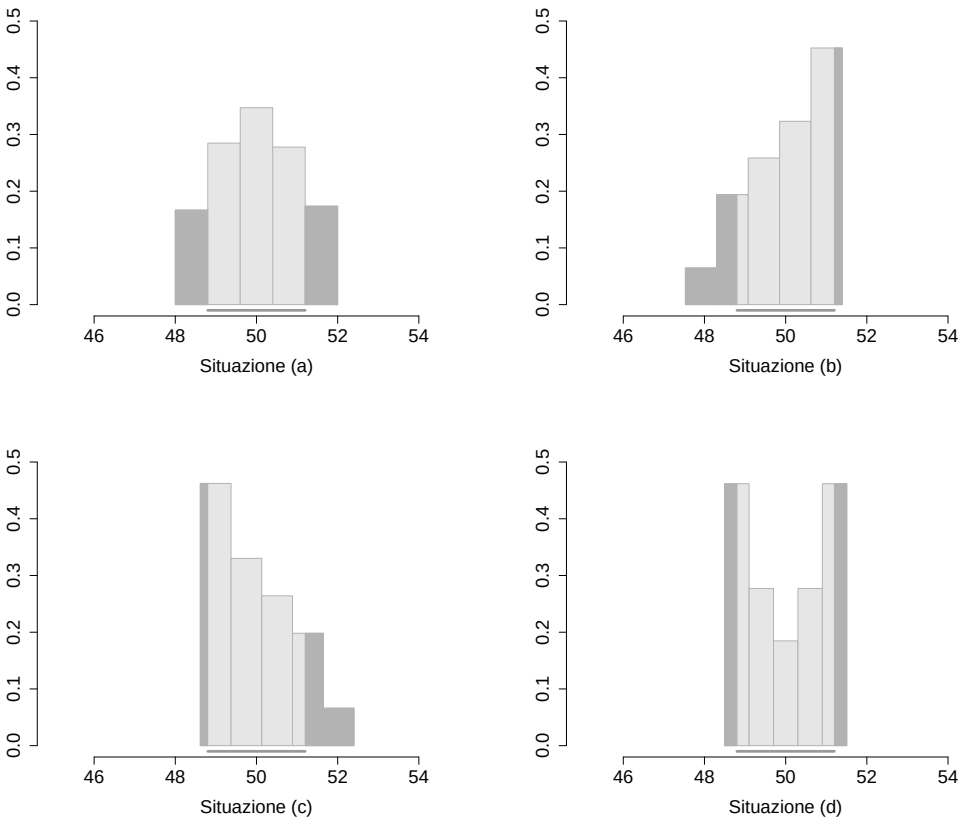


Figura 6.6 Intervallo di Tchebychev e distribuzioni a confronto, esempio 6.11.

assumono valori interni all'intervallo simmetrico rispetto alla media $[11.25; 57.59]$ possiamo, non disponendo dei dati individuali, ricorrere alla diseguaglianza di Tchebychev. Nel caso in esame, essendo $\mu_X + k\,\sigma_X = 57.59$, otteniamo $k = 1.25$; il numero ricercato sarà maggiore di $(1 - k^{-2})\,n = 8.64$, ovvero 9 unità.
Si immagini, ora, che i dati individuali della v.s. X in esame siano i seguenti:

$$\begin{aligned}\{\tilde{x}_\alpha\}_{\alpha=1,\ldots,24} = \{&34.05, 39.69, 38.21, 32.20, 33.72, 40.52, 47.69, 47.35,\\ &55.00, 1.36, 14.56, 13.29, 10.43, 20.50, 22.71, 26.37,\\ &59.31, 56.83, 65.46, 68.89, 22.53, 31.00, 41.89, 2.59\}\end{aligned}$$

Un semplice conteggio ci consente di asserire che le unità statistice che assumo valori entro l'intervallo $[11.25; 57.59]$ sono esattamente 18 e ciò è in accordo con quanto affermato dalla diseguaglianza di Tchebychev.

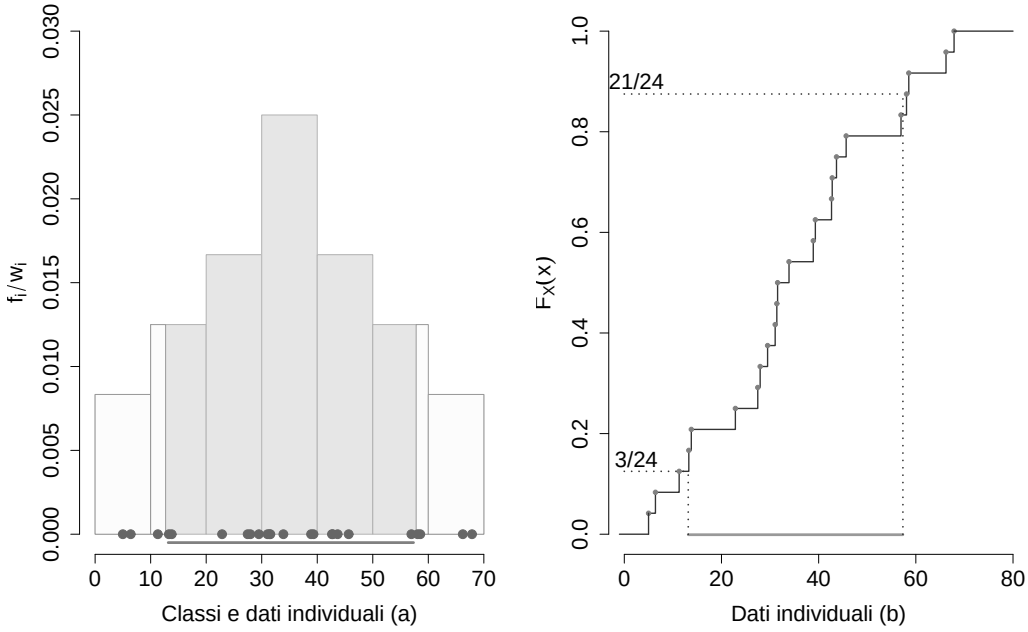


Figura 6.7 Intervallo di Tchebychev e dati individuali, esempio 6.12.

Alla stessa soluzione saremmo giunti analizzando i grafici proposti in figura (6.7).

◁

6.6. INDICI DI VARIABILITÀ RELATIVI

Le misure di variabilità presentate nei precedenti paragrafi hanno la prerogativa di assumere valore nullo qualora il fenomeno in esame non presenti variabilità e valori via via maggiori al crescere di questa.
A volte può interessare effettuare il confronto fra misure di variabilità:

- ⋆ riferentesi a due diversi caratteri che sono stati rilevati mediante differenti unità di misura. Si pensi, ad esempio, di voler stabilire, con riferimento ad un medesimo collettivo statistico, se la variabile *peso* presenti maggiore variabilità della variabile *statura*; la non omogeneità delle due unità di misura adottate impedisce un confronto diretto;

- ⋆ di uno stesso fenomeno rilevato su due o più distinti collettivi statistici. Si immagini, ad esempio, di voler verificare se i *il reddito dei lavoratori autonomi* presenti la stessa variabilità dei *redditi dei lavoratori dipendenti*; l'unità di misura è in questo caso la stessa ma la misura di variabilità, essendo legata all'ordine di grandezza dei valori assunti dalle due variabili, potrebbe non essere paragonabile in ermini assoluti.

Per tali motivi le misure di variabilità precedentemente introdotte si rivelano a volte inadeguate, donde la necessità di ricorrere a particolari numeri adimensionali detti abitualmente *indici relativi di variabilità.*
Con riferimento ad una v.s. X con distribuzione di frequenze $\{x_i, n_i\}_{i=1,..,k}$, dal punto di vista applicativo gli indici relativi di variabilità più frequentemente impiegati sono:

- ⋆ l'ampiezza dell'intervallo di escursione rapportato al valore minimo oppure al massimo od ancora alla media aritmetica, cioè:

$$\frac{x_k - x_1}{x_1} \qquad \frac{x_k - x_1}{x_k} \qquad \frac{x_k - x_1}{\mu_X}$$

- ⋆ la differenza interquartile rapportata al primo quartile oppure al terzo quartile od ancora alla mediana, cioè:

$$\frac{x_{0.75} - x_{0.25}}{x_{0.25}} \qquad \frac{x_{0.75} - x_{0.25}}{x_{0.75}} \qquad \frac{x_{0.75} - x_{0.25}}{x_{0.50}}$$

- ⋆ lo scarto quadratico medio rapportato alla media aritmetica, cioè:

$$\frac{\sigma}{\mu_X} \tag{6.13}$$

  L'indice relativo proposto in (6.13), di impiego frequente, viene generalmente detto *coefficiente di variazione.*

- ⋆ la differenza media assoluta, con o senza ripetizione, rapportata alla media aritmetica, cioè:

$$\frac{\Delta_R}{\mu_X} \qquad \frac{\Delta}{\mu_X}$$

▷ ESEMPIO 6.13

Si immagini di avere rilevato su di un collettivo si $n = 20$ individui di sesso maschile e di età compresa tra 30 e 35 anni i caratteri *statura* in centimetri e *peso* in chilogrammi, e che tale operazione abbia dato luogo ad altrettante v.s. X e Y rispettivamente, con valori individuali:

$\tilde{x}_\alpha$	$\tilde{y}_\alpha$	$\tilde{x}_\alpha$	$\tilde{y}_\alpha$	$\tilde{x}_\alpha$	$\tilde{y}_\alpha$
175.4	76.63	154.7	65.14	156.8	73.03
160.3	55.50	170.4	63.42	186.0	90.41
168.9	76.71	174.5	75.25	176.0	64.74
172.7	90.91	170.7	75.99	177.4	78.95
171.8	73.87	159.8	62.76	153.6	85.45
165.7	76.34	175.3	71.37	166.5	72.35
180.5	83.76	182.3	78.12		

Effettuati alcuni semplici calcoli, per esse otteniamo i seguenti risultati di sintesi:

$\mu_X =$	169.9	$\mu_Y =$	74.53
$x_{max} - x_{min} =$	32.0	$y_{max} - y_{min} =$	35.40
$\sigma_X =$	8.8	$\sigma_Y =$	8.94
$x_{0.75} - x_{0.25} =$	10.8	$y_{0.75} - y_{0.25} =$	8.49
$\Delta_{R,X} =$	10.1	$\Delta_{R,Y} =$	10.00

Il confronto diretto tra le misure proposte, che indurrebbe erroneamente a ritenere che le v.s. X e Y abbiano pressapoco ugual varibilità, non è corretto essendo queste espresse in unità di misura diverse. Se ricorriamo agli indici relativi:

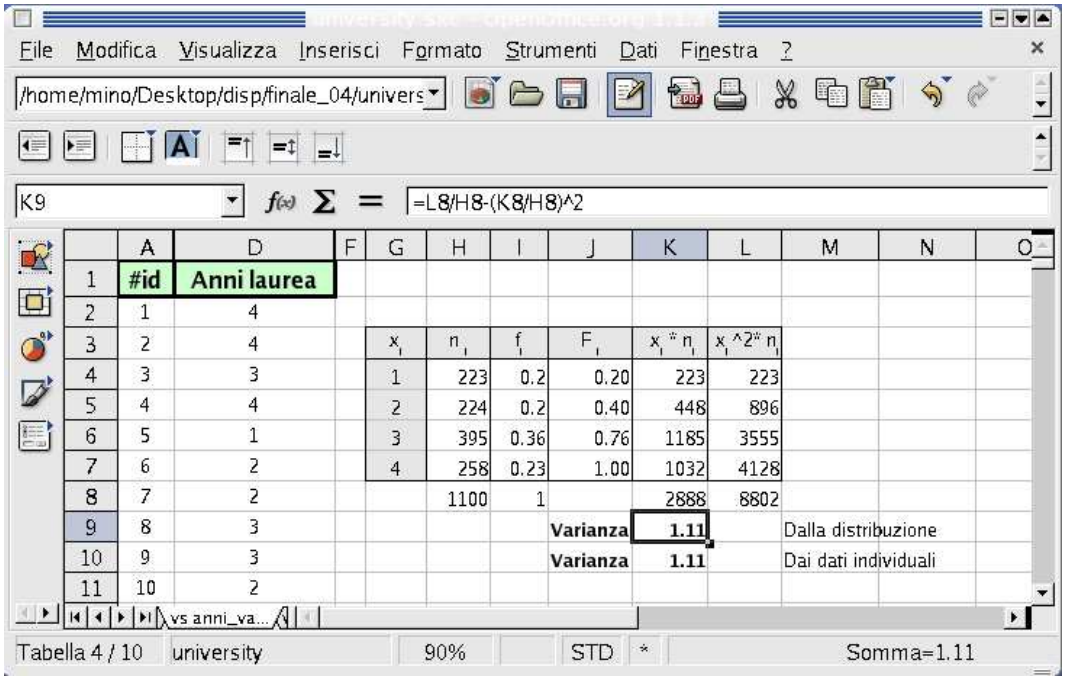
$(x_{max} - x_{min})/x_{min} =$	0.21	$(y_{max} - y_{min})/y_{min} =$	0.64
$\sigma_X/\mu_X =$	0.05	$\sigma_Y/\mu_Y =$	0.12
$x_{0.75} - x_{0.25}/x_{0.50} =$	0.06	$y_{0.75} - y_{0.25}/y_{0.50} =$	0.11
$\Delta_{R,X}/\mu_X =$	0.06	$\Delta_{R,Y}\mu_Y =$	0.13

possiamo affermare che la distribuzione del peso degli individui censiti presenta variabilità all'incirca doppia rispetto a quella della statura.

◁

6.7. IL FOGLIO ELETTRONICO

Consideriamo la variabile statistica *anni dalla laurea* del solito file `university.sxc` e vediamo come calcolare la varianza con il foglio elettronico.



	A	D	G	H	I	J	K	L	M	N
1	#id	Anni laurea								
2	1	4								
3	2	4	x_i	n_i	f_i	F_i	$x_i * n_i$	$x_i \hat{} 2 * n_i$		
4	3	3	1	223	0.2	0.20	223	223		
5	4	4	2	224	0.2	0.40	448	896		
6	5	1	3	395	0.36	0.76	1185	3555		
7	6	2	4	258	0.23	1.00	1032	4128		
8	7	2		1100	1		2888	8802		
9	8	3				**Varianza**	**1.11**		Dalla distribuzione	
10	9	3				**Varianza**	**1.11**		Dai dati individuali	
11	10	2								

Figura 6.8 Calcolo della varianza per la v.s. *anni dalla laurea.*

In figura (6.8) nelle celle `K9` e `K10` osserviamo la varianza calcolata rispettivamente a partire dalla distribuzione di frequenze e dai dati individuali.
Il valore 1.11, che compare nella cella `K9` è stato calcolato usando la forma della varianza

$$V[X] = E[X^2] - (E[X])^2$$

ed è il risultato della formula in essa inserita `=L8/H8-(K8/H8)^2`.
Nella cella `L8` abbiamo inserito la funzione `=SOMMA(L4:L7)`, nella cella `K8` abbiamo inserito la funzione `=SOMMA(K4:K7)` e in `H8` vi è la funzione `=SOMMA(H4:H7)`.
L'intervallo di celle `L4:L7`, ci è utile per calcolare la $E[X^2]$ e contiene i prodotti del quadrato di ciscuna modalità per la frequenza assoluta associata, ottenuti, ad esempio per la cella `L4` inserendo la formula `=G4^2*H4`. Mentre l'intervallo di celle `K4:K7` contiene i prodotti di ciscuna modalità per la frequenza assoluta associata, ottenuti, ad esempio per la cella `K4` inserendo la faormula `=G4*H4` già utilizzatia per il calcolo della media aritmetica nel precedente capitolo.
Il valore 1.11, che compare nella cella `K10` è il valore della varianza calcolata a partire dai dati individuali ed è il risultato della funzione di OpenOffice `=VAR.POP(D2:D1101)`.

In questo caso, ovviamente, i due risultati coincidono e la varianza può essere indifferentemente calcolata in entrambi i modi.
Si noti che OpenOffice fornisce anche la funzione VAR( ) che calcola la *varianza campionaria* e non deve essere confusa con VAR.POP( ), da noi utilizzata.

6.8. ESERCIZI

▷ ESERCIZIO 6.1

Siano X e Y due variabili statistiche con insiemi dei dati individuali rispettivamente

$$\{\tilde{x}_\alpha\}_{\alpha=1,\dots,20} = \{8,6,5,4,9,6,3,4,5,7,8,9,6,5,8,7,6,6,5,6\}$$
$$\{\tilde{y}_\alpha\}_{\alpha=1,\dots,20} = \{8,5,4,5,2,3,4,7,5,4,3,6,7,8,5,4,7,6,5,5\}$$

Costruiti i diagrammi a scatola e baffi per ciascuna variabile statistica, se ne valuti la variabilità.

◁

▷ ESERCIZIO 6.2

La tabella che segue riporta la distribuzione di frequenze, con dati opportunamente raccolti in classi, delle variabili statistiche X e Y che rappresentano rispettivamente gli importi in euro delle fatture emesse da due società nell'ultima settimana di ottobre 2004

			v.s. X	v.s. Y
Classi	di	Importo	n_i	n_i
0	⊣	200	15	25
200	⊣	300	35	50
300	⊣	400	50	15
400	⊣	1000	60	40

Valutare la variabilità di ciascuna variabile statistica con la misura che si ritiene più opportuna.
Quale variabile statistica presenta maggiore variabilità?

◁

▷ ESERCIZIO 6.3

Sia X una variabile statistica con insieme dei dati individuali

$$\{\tilde{x}_\alpha\}_{\alpha=1,\dots,12} = \{120, 125, 115, 112, 110, 116, 123, 121, 127, 114, 118, 121\}$$

Calcolare media e varianza delle trasformate lineari

$$Y = X - 118.5 \qquad\qquad Z = \frac{X - 118.5}{\sqrt{25.25}}$$

◁

▷ ESERCIZIO 6.4

Si supponga che per la v.s. $X = \{$*litri di miscela erogati settimanalmente*$\}$, rilevata su un collettivo costituito dai 12 distributori di un comune, si abbia il seguente insieme dei dati individuali

$$\{\tilde{x}_\alpha\}_{\alpha=1,\dots,12} = \{92, 105, 87, 102, 110, 97, 85, 100, 115, 101, 80, 72\}$$

Calcolare la differenza interquartile, la differenza media assoluta e lo scarto quadratico medio.

◁

▷ ESERCIZIO 6.5

Un revisore contabile esamina le pratiche di pagamento sinistri effettuati da una compagnia di assicurazione nel settore R.C. auto, e sottopone al responsabile dell'agenzia a cui fanno capo i clienti rimborsati, la seguente distribuzione di frequenze per la v.s. $X = \{$*importo liquidato*$\}$, in migliaia di euro:

Classi	di	Importo	n_i
0	⊣	4	1850
4	⊣	6	2500
6	⊣	10	1200
10	⊣	40	950

Calcolare la differenza interquartile nonché la varianza e lo scarto quadratico medio. Indicare, infine, il numero di pratiche di pagamento il cui importo rientra nell'intervallo $[\mu_X - \sigma_X; \mu_X + \sigma_X]$.

◁

▷ ESERCIZIO 6.6

Con riferimento alla variabile statistica X, di cui all'esercizio 6.5, calcolarne la differenza media assoluta senza ripetizione.

◁

▷ ESERCIZIO 6.7

Di una variabile statistica X, rilevata su un collettivo di $n = 1200$ unità, sono noti $\mu_X = 200$ e $\sigma_X^2 = 169$. Indicare il limite inferiore per il numero di unità statistiche che assumono valori interni all'intervallo $[180.5; 219.5]$.

◁

▷ ESERCIZIO 6.8

Con riferimento alla variabile statistica X di cui all'esercizio 6.7, indicare qual'è l'intervallo, simmetrico rispetto al valor medio, all'esterno del quale rientra al più il 64% delle unità del collettivo.

◁

CAPITOLO 7

STUDIO CONGIUNTO DI DUE CARATTERI

In questo capitolo, che introduce allo studio congiunto di due caratteri, sono inizialmente definte le mutabili e le variabili statistiche bivariate e sono descritti i passi necessari alla determinazione delle loro distribuzioni congiunte e marginali. Delle variabili statistiche condizionate si individuano le distribuzioni di frequenze e si definiscono media e varianza. Per la varibile statistica doppia si analizzano le proprietà della covarianza e si introduce la combinazione lineare delle sue componenti.

7.1. MUTABILI E VARIABILI STATISTICHE BIVARIATE

I metodi di analisi sino ad ora proposti hanno coinvolto mutabili o variabili statistiche univariate. Ora ci occuperemo dello studio congiunto di due caratteri simultaneamente rilevati sulle unità di un generico colletivo statistico.
Con riferimento, dunque, ad un collettivo statistico Ω consideriamo due caratteri a cui, una volta definite le opportune scale di misura, vengono associati gli insiemi delle modalità M_1 ed M_2 rispettivamente. Per lo studio congiunto dei due caratteri sarà necessario considerare il prodotto cartesiano $M_1 \times M_2$ formato da tutte le possibili coppie di elementi di M_1 ed M_2.

▷ ESEMPIO 7.1

Volendo indagare circa gli infortuni sul lavoro accaduti l'ultimo anno in una azienda metalmeccanica si pensi di voler indagare circa i legami esitenti tra il turno in cui si è verificato l'infortunio e la natura della lesione. Con riferimento al collettivo statistico formato dall'insieme di tutti gli infortuni avvenuti consideriamo il carattere *turno* e

il carattere *natura della lesione* con insiemi delle modalità rispettivamente

$$M_1 = \{\text{Mattino, Pomeriggio, Notte}\} = \{1^\circ, 2^\circ, 3^\circ\}$$
$$M_2 = \{\text{Ferita, Distorsione, Frattura, Corpo Estraneo}\}$$
$$= \{\text{FE, DI, FR, CE}\}$$

Il prodotto cartesiano di questi due insiemi sarà un insieme costituito dalle 12 possibili coppie di attributi, e più precisamnete:

$$M_1 \times M_2 = \{(1^\circ; \text{FE}), (1^\circ; \text{DI}), (1^\circ; \text{FR}), (1^\circ; \text{CE}),$$
$$(2^\circ; \text{FE}), (2^\circ; \text{DI}), (2^\circ; \text{FR}), (2^\circ; \text{CE}),$$
$$(3^\circ; \text{FE}), (3^\circ; \text{DI}), (3^\circ; \text{FR}), (3^\circ; \text{CE})\}$$

◁

L'esempio precedente riguarda la costruzione dell'insieme prodotto cartesiano dei due insiemi di modalità di caratteri entrambi qualitativi che, come abbiamo visto, risulta un insieme finito formato da coppie di attributi. Nel caso in cui uno dei due caratteri sia di tipo quantitativo il prodotto cartesiano sarà formato da coppie miste di attributo e numero e potrà avere numerosità finita o infinita in accordo con la natura dell'insieme delle modalità del carattere quantitativo considerato. Qualora entrambi i caratteri siano di tipo quantitativo il prodotto cartesiano dei loro insiemi di modalità sarà formato da coppie di numeri e potrà essere finito o infinito.

▷ ESEMPIO 7.2

Si consideri il collettivo statistico formato dalle bolle di accompagnamento emesse nel mese di dicembre da una ditta produttrice di eletrodomestici sul quale si vogliano rilevare il *luogo di spedizione*, il *numero di colli* spediti e il *peso* della merce spedita. Se gli insiemi di modalità dei precedenti caratteri sono:

- $M_1 = \{\text{Italia, Europa, Extra Europa}\} = \{\text{It, Eu, EE}\}$
- $M_2 = \{1, 2, \ldots, 40\}$
- $M_3 = [10; 500]$ kg.

l'insieme $M_1 \times M_2$ è finito ed è formato da $3 \cdot 40 = 120$ coppie miste di attributo e numero, e cioè $M_1 \times M_2 = \{(\text{It}; 1), (\text{It}; 2), \ldots, (\text{EE}; 40)\}$. Mentre gli insiemi $M_1 \times M_3$ e $M_2 \times M_3$, formati rispettivamente da coppie miste attributo e numero e da coppie di numeri, hanno entrambi numerosità infinita.

◁

Procedendo alla rilevazione congiunta dei due caratteri su ciascuna unità statistica, si individua una corrispondenza tra collettivo statistico Ω e l'insieme $M_1 \times M_2$ delle coppie di modalità. In analogia con quanto esposto nel caso univariato, siamo in grado di dare la seguente

Definizione 7.1 (Mutabile e variabile statistica bivarata)
l'applicazione che associa a ciascuna unità ω_α del collettivo Ω uno ed uno solo elemento dell'insieme $M_1 \times M_2$ viene detta:

- *mutabile statistica bivariata, e indicata con (A, B), se gli elementi di M_1 e M_2 sono attributi;*
- *variabile statistica mista, e indicata con (A, Y), se gli elementi di M_1 sono attributi e M_2 è costituito da elementi di $\mathbb{R}$;*
- *variabile statistica bivariata, e indicata con (X, Y), se gli insiemi M_1 e M_2 sono costituiti da elementi di $\mathbb{R}$.*

□

La rilevazione congiunta dei due caratteri sulle unità del collettivo statistico darà luogo ad un insieme di coppie di modalità che, analogamente al caso univariato, verrà detto *insieme dei dati individuali.* Più precisamente tale insieme verrà indicato in simboli:

⋆ per la mutabile statistica bivariata

$$\{(A, B)(\omega_\alpha)\}_{\alpha=1,\dots,n} = \{(\tilde{a}_\alpha; \tilde{b}_\alpha)\}_{\alpha=1,\dots,n} \tag{7.1}$$

⋆ per la variabile statistica mista

$$\{(A, Y)(\omega_\alpha)\}_{\alpha=1,\dots,n} = \{(\tilde{a}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,n} \tag{7.2}$$

⋆ per la variabile statistica bivariata

$$\{(X, Y)(\omega_\alpha)\}_{\alpha=1,\dots,n} = \{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,n} \tag{7.3}$$

▷ ESEMPIO 7.3

Con riferimento ai caratteri definiti nell'esempio (7.1), supponiamo che gli infortuni nell'anno siano stati $n = 10$. Rilevando per ciascun infortunio il *turno* in cui esso

è occorso e la *natura della lesione* si disporrà di una mutabile statististica bivariata (A, B) con insieme dei dati individuali:

$$\{(\tilde{a}_\alpha; \tilde{b}_\alpha)\}_{\alpha=1,\dots,10} = \{(2^\circ; \text{CE}), (1^\circ; \text{FE}), (3^\circ; \text{CE}), (2^\circ; \text{DI}), (3^\circ; \text{FE}), \\ (2^\circ; \text{DI}), (3^\circ; \text{CE}), (1^\circ; \text{FR}), (3^\circ; \text{CE}), (1^\circ; \text{FE})\}$$

Ipotizzando che il collettivo statistico delle bolle di accompagnamento dell'esempio (7.2) sia anche esso formato da $n = 10$ unità statistiche, rilevando congiuntamente il *luogo di spedizione* e il *numero di colli* spediti si disporrà di una variabile statistica mista (A, Y) con insieme dei dati individuali:

$$\{(\tilde{a}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,10} = \{(\text{Eu}; 35), (\text{It}; 10), (\text{Eu}; 23), (\text{It}; 18), (\text{Eu}; 18), \\ (\text{Eu}; 23), (\text{It}; 10), (\text{Eu}; 23), (\text{Eu}; 18), (\text{Eu}; 23)\}$$

Sempre sul medesimo collettivo statistico delle 10 bolle di accompagnamento la rilevazione congiunta dei caratteri *numero di colli* e *peso* della merce spedita darà luogo alla variabile statistica bivariata (X, Y) con insieme dei dati individuali:

$$\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,10} = \{(35; 450), (10; 103), (23; 252), (18; 124), (18; 287), \\ (23; 345), (10; 80), (23; 462), (18; 235), (23; 425)\}$$

Si noti che gli insiemi dei dati individuali delle mutabili e delle variabili doppie qui definite non necessariamente contengono tutti gli elementi dei rispettivi insiemi prodotto cartesiano e sicuramente così sarà sempre per le variabili statistiche miste o le variabili doppie con codominio un insieme infinito.

◁

Come sempre, gli insiemi dei dati individuali contengono tutte le informazioni circa i due caratteri congiuntamente rilevati. Per visualizzare sinteticamente l'informazione in essi contenuta, ricorriamo alla *distribuzione di frequenze congiunte*.

7.2. DISTRIBUZIONE DI FREQUENZE CONGIUNTE

In questo paragrafo ripercorriamo i quattro passi necessari all'individuazione della distribuzione di frequenze congiunte.
Per semplicità di esposizione, faremo riferimento unicamente ad una variabile statistica mista (A, Y), ma la procedura può essere estesa senza alcuna difficoltà alle mutabili ed alle variabili statistiche doppie.

Con riferimento, dunque, ad una variabile statistica mista (A, Y) con insieme dei dati individuali $\{(\tilde{a}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,n}$, il primo passo sarà quello di individuare il corrispondente insieme delle modalità distinte. A tal fine consideriamo i due insiemi:

$$\{a_i\}_{i=1,\dots,r} = \{a_1, \dots, a_r\}$$
$$\{y_j\}_{j=1,\dots,s} = \{y_1, \dots, y_s\}$$

costituiti rispettivamente dagli $r \leq n$ e dagli $s \leq n$ *elementi distinti* (posti in ordine crescente se possibile) presenti fra le $\tilde{a}_\alpha$ e $\tilde{y}_\alpha$ che compaiono nell'insieme dei dati individuali.
Il loro prodotto cartesiano fornisce l'insieme

$$\{(a_i; y_j)\}_{\substack{i=1,\dots,r \\ j=1,\dots,s}} \tag{7.4}$$

che costituisce l'insieme delle *modalità distinte* della variabile statistica mista (A, Y) e fornisce informazioni su quali e quanti siano i valori distinti che essa assume nell'insieme $M_1 \times M_2$.

▷ ESEMPIO 7.4

Con riferimento all'esempio (7.3) consideriamo l'insieme dei dati individuali della variabile statistica mista $(A, Y) = \{$*luogo di spedizione, numero di colli*$\}$, che ricordiamo essere

$$\{(\tilde{a}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,10} = \{(\text{Eu}; 35), (\text{It}; 10), (\text{Eu}; 23), (\text{It}; 18), (\text{Eu}; 18), \\ (\text{Eu}; 23), (\text{It}; 10), (\text{Eu}; 23), (\text{Eu}; 18), (\text{Eu}; 23)\}$$

scorrendo le componenti delle coppie che lo compongono individuiamo gli insiemi:

$$\{a_i\}_{i=1,2} = \{\text{It, Eu}\} \qquad \{y_j\}_{j=1,\dots,4} = \{10, 18, 23, 35\}$$

e costruiamo l'insieme delle *modalità distinte* della variabile statistica mista (A, Y) facendo il loro prodotto cartesiano, cioè:

$$\{(a_i; y_j)\}_{\substack{i=1,2 \\ j=1,\dots,4}} = \{(\text{It}; 10), (\text{It}; 18), (\text{It}; 23), (\text{It}; 35), \\ (\text{Eu}; 10), (\text{Eu}; 18), (\text{Eu}; 23), (\text{Eu}; 35)\}$$

◁

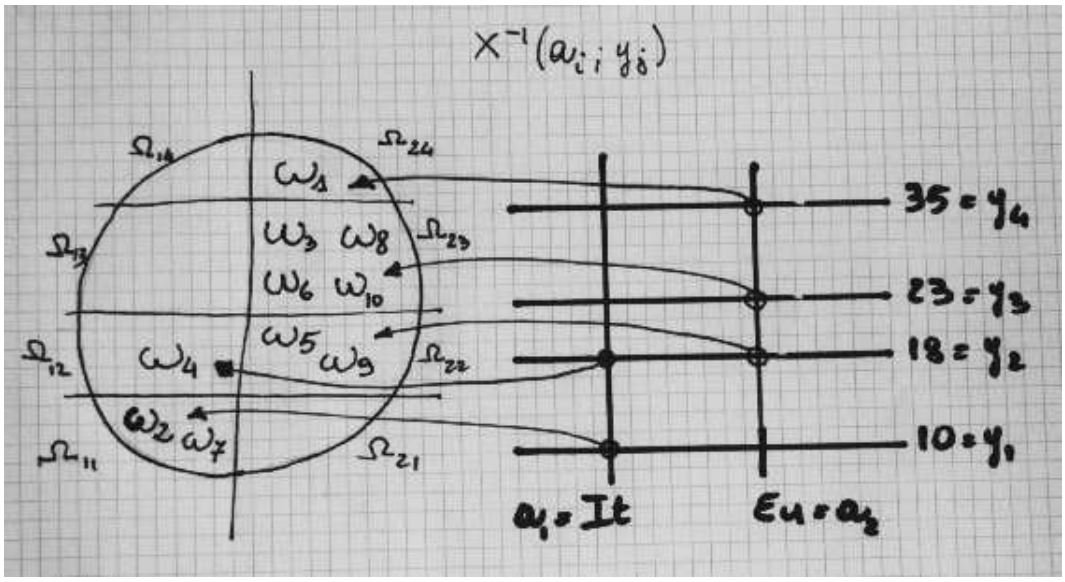


Figura 7.1 I sottoinsiemi Ω_{ij}, esempio 7.4.

Il secondo passo consiste nell'individuare nel collettivo statistico i sottoinsiemi i cui elementi sono associati dall'applicazione (A, Y) alla medesima modalità $(a_i; y_j)$. A tal fine consideriamo gli insiemi Ω_{ij} (cfr. figura 7.1) costituiti dalle controimmagini di $(a_i; y_j)$ nell'applicazione $(A, Y)(\,\cdot\,)$ e cioè $\forall i = 1, \ldots, r$ e $\forall j = 1, \ldots, s$

$$\Omega_{ij} = \{\omega_\alpha : (A, Y)(\omega_\alpha) = (a_i; y_j)\} \tag{7.5}$$

tali insiemi sono formati dalle unità ω_α del collettivo che presentano contemporaneamente la modalità a_i della mutabile A e la modalità y_j della variabile Y e formano, palesemente, una partizione di Ω.

▷ ESEMPIO 7.5

Riprendiamo la v.s. mista $(A, Y) = \{$*luogo di spedizione, numero di colli*$\}$ il cui insieme dei dati individuali è stato individuato nell'esempio (7.4) ed è formato da $2 \cdot 4 = 8$ coppie miste. Determiniamo dunque gli 8 sottoinsiemi Ω_{ij} che risultano

essere:

$$\begin{aligned}
\Omega_{11} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_1; y_1)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{It}; 10)\} = \\
&= \{\omega_2, \omega_7\} \\
\Omega_{12} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_1; y_2)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{It}; 18)\} = \\
&= \{\omega_4\} \\
\Omega_{13} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_1; y_3)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{It}; 23)\} = \\
&= \emptyset \\
\Omega_{14} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_1; y_4)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{It}; 35)\} = \\
&= \emptyset \\
\Omega_{21} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_2; y_1)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{Eu}; 10)\} = \\
&= \emptyset \\
\Omega_{22} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_2; y_2)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{Eu}; 18)\} = \\
&= \{\omega_5, \omega_9\} \\
\Omega_{23} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_2; y_3)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{Eu}; 23)\} = \\
&= \{\omega_3, \omega_6, \omega_8, \omega_{10}\} \\
\Omega_{24} &= \{\omega_\alpha : (A,Y)(\omega_\alpha) = (a_2; y_4)\} = \{\omega_\alpha : (A,Y)(\omega_\alpha) = (\text{Eu}; 35)\} = \\
&= \{\omega_1\}
\end{aligned}$$

Come possiamo notare, anche osservando la figura (7.1) i sottoinsiemi ora definiti sono disgiunti, formano una partizione di Ω ed alcuni di essi sono vuoti.

$\triangleleft$

Osservando che gli elementi di Ω_{ij} sono tutte e sole le unità statistiche che presentano la modalità $(a_i; y_j)$, compiamo il terzo passo definendo la sua numerosità come frequenza assoluta congiunta.

Definizione 7.2 (Frequenza congiunta assoluta)
Qualsiasi siano $i = 1, \ldots, r$ e $j = 1, \ldots, s$ definiamo frequenza assoluta congiunta associata alla modalità $(a_i; y_j)$ della variabile statistica mista (A, Y) il numero n_{ij} di elementi dell'insieme Ω_{ij}, cioè $n_{ij} = Nu\{\Omega_{ij}\}$.

$\square$

Come di consueto, il rapporto $f_{ij} = \frac{n_{ij}}{n}$ verrà detto *frequenza relativa congiunta* associata alla modalità $(a_i; y_j)$ della variabile statistica mista (A, Y).

Il Quarto ed ultimo passo consiste nel sintetizzare l'informazione contenuta nell'insieme dei dati individuali per mezzo della distribuzione di frequenze congiunte definita come segue.

Definizione 7.3 (Distribuzione di frequenze congiunte)
Con riferimento ad una generica variabile statistica mista (A, Y)*, definiamo distribuzione di frequenze assolute congiunte l'insieme di coppie*

$$\{((a_i; y_j); n_{ij})\}_{\substack{i=1,\dots,r \\ j=1,\dots,s}} \tag{7.6}$$

e,in modo del tutto equivalente, distribuzione di frequenze relative congiunte l'insieme di coppie

$$\{((a_i; y_j); f_{ij})\}_{\substack{i=1,\dots,r \\ j=1,\dots,s}} \tag{7.7}$$

□

È abitudine presentare la distribuzione di frequenze congiunte con la seguente tabella, anche detta *tabella a doppia entrata.* Ad esempio nel caso di frequenze assolute congiunte:

$A \downarrow Y \rightarrow$	y_1	$\cdots$	y_j	$\cdots$	y_s
a_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}

▷ ESEMPIO 7.6

Contiamo gli elemeni degli insiemi Ω_{ij} individuati nell'esempio (7.5) per la variabile statistica mista $(A, Y) = \{$*luogo di spedizione, numero di colli*$\}$ e determiniamo le frequenze assolute congiunte associate alle sue 8 modalità distinte, ciè:

$$\begin{aligned} &n_{11} = 2, \quad &&n_{12} = 1, \quad &&n_{13} = 0, \quad &&n_{14} = 0, \\ &n_{21} = 0, \quad &&n_{22} = 2, \quad &&n_{23} = 4, \quad &&n_{24} = 1. \end{aligned}$$

Associando a queste ultime le rispettive modalità otteniamo la distribuzione di frequenze assolute congiunte della variabile statistica mista (A, Y) che rappresentiamo in forma tabellare come segue

$A \downarrow Y \rightarrow$	$y_1 = 10$	$y_2 = 18$	$y_3 = 23$	$y_4 = 35$
$a_1 =$ It	2	1	0	0
$a_1 =$ Eu	0	2	4	1

Dalla tabella ricaviamo, ad esempio, che sono due le bolle di accompagnamento che riguardano dieci colli spediti in Italia mentre quelle spedite in Europa riguardanti 23 colli sono quattro.

$\triangleleft$

Si invita il Lettore a determinare la distribuzione di frequenze congiunte della mutabile statistica doppia $(A, B) = \{$*turno, natura della lesione*$\}$ e della variabile statistica doppia $(X, Y) = \{$*numero di colli, peso*$\}$ i cui insiemi di dati individuali sono riportati nell'esempio (7.3).

7.3. DISTRIBUZIONI MARGINALI E CONDIZIONATE

Disponendo della disribuzione di frequenze congiunte di una variabile statistica mista (A, Y) è possibile ricavare la distribuzione di frequenze della mutabile statistica univariata A e quella della variabile statistica univariata Y, così come sono state definite nel terzo capitolo.
L'insieme delle modalità distinte assunte dalla componente m.s. A è, per quanto già visto, l'insieme $\{a_i\}_{i=1,\ldots,r}$ mentre quello della componente v.s. Y è $\{y_j\}_{j=1,\ldots,s}$. Quanto alle frequenze assolute associabili a ciascuna modalità delle due componenti, ricordandone la definizione, dovremmo individuare il numero di unità statistiche che posseggono tali modalità. In altri termini per la mutabile statistica A dobbiamo individuare la numerosità di ciascuno degli insiemi Ω_i (con $i = 1, \ldots, r$) contenenti le unità statistiche che assumono modalità a_i, cioè:

$$\Omega_i = \{\omega_\alpha : A(\omega_\alpha) = a_i\}$$

Tali insiemi sono ricavabili, qualunque sia i, dall'unione, al variare di j da 1 a s, dei sottoisiemi Ω_{ij} definiti dall'equazione (7.5), infatti si ha:

$$\Omega_i = \bigcup_{j=1}^{s} \Omega_{ij} \tag{7.8}$$

Per visualizzare il significato dell'unione qui sopra definita si consideri nuovamente la figura (7.1) e si osservi che, per l'esempio a cui essa si riferisce, le modalità distinte della

mutabile A sono due e i sottoinsiemi $\Omega_1 = \cup_{j=1}^{4}\Omega_{1j}$ e $\Omega_2 = \cup_{j=1}^{4}\Omega_{2j}$ contengono tutte e sole le unità statistiche che assumono modalità rispettivamente $a_1 =$ It e $a_2 =$ Eu.
La frequenza assoluta associabile a ciscuna modalità a_i della componente A corrisponde per definizione al numero di elementi di Ω_i, pertanto dalla relazione (7.8) ricaviamo che:

$$n_i = Nu\{\Omega_i\} = \sum_{j=1}^{s} Nu\{\Omega_{ij}\} = \sum_{j=1}^{s} n_{ij}$$

Ciò premesso, possiamo ricavare le frequenze assolute n_i associate alle modalità a_i della componente A di una variabile statistica mista (A, Y) dalla sua distribuzione di frequenze congiunte sommando rispetto all'indice j le frequenze n_{ij}.
Ragionando in modo analogo, le frequenze assolute n_j associate alle modalità y_j della componente Y possono essere ricavate dalla distribuzione di frequenze congiunte sommando rispetto all'indice i le frequenze n_{ij}.
È prassi comune aggiungere un'ultima colonna a margine della tabella della distribuzione di frequenze congiunte contenente la somma delle frequenze di ciascuna riga e un'ultima riga a margine con la somma delle frequenze di ciascuna colonna, così da avere:

$A \downarrow Y \rightarrow$	y_1	$\cdots$	y_j	$\cdots$	y_s	
a_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

ove

$$n_i = n_{i\cdot} = \sum_{j=1}^{s} n_{ij}$$

$$n_j = n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

$$n = n_{\cdot\cdot} = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{s} n_{\cdot j}$$

La notazione $n_{i\cdot}$ sta ad indicare che tale frequenza è stata ottenuta sommando rispetto all'indice j tutte le frequenze congiunte n_{ij} corrispondenti alla i-esima modalità di A. Così

per $n_{\cdot j}$ la somma è stata fatta sommando rispetto all'indice i tutte le frequenze congiunte n_{ij} corrispondenti alla j-esima modalità di Y. Naturalmente $n_{\cdot\cdot}$ corrisponde alla somma di tutte le frequenze congiunte n_{ij} e rappresenta dunque la numerosità del collettivo statistico.
In tal modo, leggendo congiuntamente la prima e l'ultima colonna a margine della tabella si ha la distribuzione di frequenze assolute della m.s. univariata A mentre dalla prima e dall'ultima riga a margine della tabella si ha la distribuzione di frequenze assolute della v.s. univariata Y; che risultano essere rispettivamente:

$$A \equiv \left\{ \begin{matrix} a_i \\ n_{i\cdot} \end{matrix} \right\}_{i=1,\dots,r} = \left\{ \begin{matrix} a_1 & \dots & a_i & \dots & a_r \\ n_{1\cdot} & \dots & n_{i\cdot} & \dots & n_{r\cdot} \end{matrix} \right\}$$

$$Y \equiv \left\{ \begin{matrix} y_j \\ n_{\cdot j} \end{matrix} \right\}_{j=1,\dots,s} = \left\{ \begin{matrix} y_1 & \dots & y_j & \dots & y_s \\ n_{\cdot 1} & \dots & n_{\cdot j} & \dots & n_{\cdot s} \end{matrix} \right\}$$

Le distribuzioni univariate delle due componenti vengono anche dette *distribuzioni marginali* proprio perché scritte a margine della tabella.

▷ ESEMPIO 7.7

La variabile statistica mista $(A, Y) = \{$*tipo di pagamento, numero di ordini annui* $\}$ rilevata su un collettivo statisitico formato da 244 clienti di una azienda che vende per corrispondenza possiede la seguente distribuzione di frequenze congiunte:

$A \downarrow Y \rightarrow$	$y_1 = 3$	$y_2 = 4$	$y_3 = 5$	$y_4 = 6$
$a_1 =$ contrassegno	44	25	18	6
$a_2 =$ carta di credito	12	21	57	61

Osservando che

$$n_{1\cdot} = \sum_{j=1}^{4} n_{1j} = 44 + 25 + 18 + 6 = 93$$

$$n_{2\cdot} = \sum_{j=1}^{4} n_{2j} = 12 + 21 + 57 + 61 = 151$$

e che

$$n_{\cdot 1} = \sum_{i=1}^{2} n_{i1} = 44 + 12 = 56, \qquad n_{\cdot 2} = \sum_{i=1}^{2} n_{i2} = 25 + 21 = 46$$
$$n_{\cdot 3} = \sum_{i=1}^{2} n_{i3} = 18 + 57 = 75, \qquad n_{\cdot 4} = \sum_{i=1}^{2} n_{i4} = 6 + 61 = 67$$

aggiungiamo queste frequenze marginali alla tabella della distribuzione di frequenze congiunte

$A \downarrow Y \rightarrow$	$y_1 = 3$	$y_2 = 4$	$y_3 = 5$	$y_4 = 6$	
$a_1 =$ contrassegno	44	25	18	6	93
$a_2 =$ carta di credito	12	21	57	61	151
	56	46	75	67	244

dalla quale ricaviamo le distribuzioni marginali delle due componenti, cioè:

$$A \equiv \left\{ \begin{matrix} a_i \\ n_{i\cdot} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{contrassegno} & \text{carta di credito} \\ 93 & 151 \end{matrix} \right\}$$

$$Y \equiv \left\{ \begin{matrix} y_j \\ n_{\cdot j} \end{matrix} \right\}_{j=1,\ldots,4} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ 56 & 46 & 75 & 67 \end{matrix} \right\}$$

◁

A ben vedere, disponendo della distribuzione di frequenze congiunte di una variabile statistica mista, è possibile studiare il comportamento delle sue due componenti A e Y rispetto a sottoinsiemi del collettivo Ω. Può infatti essere interessante verificare il comportamento di una delle due componenti univariate, ad esempio la v.s. Y, in riferimento alle sole unità statistiche che assumono una determinata modalità dell'altra componente. Si tratta, più precisamente, di individuare la *distribuzione di frequenze della v.s. Y condizionata alla modalità a_i della m.s. A*.

Definizione 7.4 (Variabile statistica condizionata)
data una variabile statistica mista (A, Y), definita da Ω a $M_1 \times M_2$, e scelta una qualunque modalità distinta a_i di A, definiamo variabile statistica Y condizionata a $A = a_i$, in simboli $Y|a_i$, l'applicazione che associa a ciascuna unità statistica ω_α per cui $A(\omega_\alpha) = a_i$ uno ed uno solo elemento dell'insieme M_2

□

Si tratta di una variabile statistica univariata la cui definizione coincide con quella già data all'inizio di questo libro a meno del dominio di definizione. Per la v.s. $Y|a_i$ il dominio non è tutto l'insieme Ω bensì il suo sottoinsieme Ω_i formato da tutte e sole le unità statistiche che assumono modalità a_i per la m.s. A.
Ovviamente le v.s. $Y|a_i$ individuabili dalla distribuzione di frequenze congiunte di una variabile statistica mista (A, Y) sono tante quante sono le modalità distinte della componente A; se l'indice i varia da 1 a r avremo dunque r variabili statistiche condizionate $Y|a_i$. Nello studio delle v.s. condizionate si stratifica il collettivo statistico rispetto alla distribuzione di frequenze di una delle due componenti e si analizza il comportamento dell'altra nei diversi strati.
La *distribuzione di frequenze assolute* della v.s. condizionata $Y|a_i$ è ricavabile, ormai per cose ben note, dalla forma tabellare della distribizione di frequenze congiunte della v.s. mista (A, Y) leggendo congiuntamente la prima riga a margine e la i-esima riga, e cioè qualunque sia $i = 1, \ldots, r$ si ha:

$$Y|a_i \equiv \left\{ \begin{matrix} y_j \\ n_{ij} \end{matrix} \right\}_{j=1,\ldots,s} = \left\{ \begin{matrix} y_1 & \ldots & y_j & \ldots & y_s \\ n_{i1} & \ldots & n_{ij} & \ldots & n_{is} \end{matrix} \right\} \tag{7.9}$$

OSSERVAZIONE: poiché la distribuzione di frequenze della v.s. condizionata $Y|a_i$ si ricava dalla distribuzione di frequenze congiunte, le sue modalità distinte vengono ad essere tutte e sole quelle della v.s. Y. Pertanto potrà accadere che in corrispondenza ad una modalità y_j della v.s. $Y|a_i$ si abbia frequenza nulla.

$\star$

▷ ESEMPIO 7.8

Con riferimento alla v.s. mista dell'esempio (7.7) la cui distribuzione di frequenze congiunte ricordiamo essere

$A \downarrow Y \rightarrow$	$y_1 = 3$	$y_2 = 4$	$y_3 = 5$	$y_4 = 6$
$a_1 =$ contrassegno	44	25	18	6
$a_2 =$ carta di credito	12	21	57	61

ricaviamo le distribuzioni di frequenze assolute delle due variabili statistiche $Y|a_i$, cioè

$$Y|a_1 \equiv \left\{ \begin{matrix} y_j \\ n_{1j} \end{matrix} \right\}_{j=1,\ldots,4} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ 44 & 25 & 18 & 6 \end{matrix} \right\}$$

$$Y|a_2 \equiv \left\{ \begin{matrix} y_j \\ n_{1j} \end{matrix} \right\}_{j=1,\ldots,4} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ 12 & 21 & 57 & 61 \end{matrix} \right\}$$

La distribuzione di frequenze assolute della v.s. condizionata $Y|a_1$ ci informa circa il numero di ordini annui fatti dai soli clienti che pagano in contrassegno e ovviamente quella della v.s. $Y|a_2$ rappresenta la distribuzione del numero di ordini fatti dai clienti che pagano con carta di credito.

$\triangleleft$

Come vedremo nel seguito, e come si può dedurre dall'esempio precedente, è il confronto tra le distribuzioni condizionate che può essere di un qualche interesse nello studio congiunto dei due caratteri. Per tale motivo è bene considerare sempre le distribuzioni delle v.s. condizionate in termini di frequenze relative che, come abbiamo già visto, non essendo influenzate dalla numerosità dello strato, consentono il confronto.
Dalla distribuzione di frequenze assolute (7.9) della generica v.s. condizionata $Y|a_i$ ricaviamo, come di consueto, la *distribuzione di frequenze relative* dividendo ciascuna frequenza assoluta per la numerosità $n_{i\cdot}$ dello strato, cioè

$$Y|a_i \equiv \left\{ \begin{matrix} y_j \\ \frac{n_{ij}}{n_{i\cdot}} \end{matrix} \right\}_{j=1,\ldots,s} = \left\{ \begin{matrix} y_1 & \cdots & y_j & \cdots & y_s \\ \frac{n_{i1}}{n_{i\cdot}} & \cdots & \frac{n_{ij}}{n_{i\cdot}} & \cdots & \frac{n_{is}}{n_{i\cdot}} \end{matrix} \right\} \tag{7.10}$$

$\triangleright$ ESEMPIO 7.9

Se riprendiamo la tabella della distribuzione di frequenze congiunte dell'esempio (7.7) alla quale abbiamo aggiunto le frequenze marginali

$A \downarrow Y \rightarrow$	$y_1 = 3$	$y_2 = 4$	$y_3 = 5$	$y_4 = 6$	
$a_1 =$ contrassegno	44	25	18	6	93
$a_2 =$ carta di credito	12	21	57	61	151
	56	46	75	67	244

ricaviamo le distribuzioni di frequenze relative delle due variabili statistche $Y|a_i$ dividendo per le frequenze a margine di ciscuna riga, cioè:

$$Y|a_1 \equiv \left\{ \begin{matrix} y_j \\ \frac{n_{1j}}{n_{1\cdot}} \end{matrix} \right\}_{j=1,\ldots,4} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ \frac{44}{93} & \frac{25}{93} & \frac{18}{93} & \frac{6}{93} \end{matrix} \right\} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ 0.47 & 0.27 & 0.19 & 0.07 \end{matrix} \right\}$$

$$Y|a_2 \equiv \left\{ \begin{matrix} y_j \\ \frac{n_{2j}}{n_{2\cdot}} \end{matrix} \right\}_{j=1,\ldots,4} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ \frac{12}{151} & \frac{21}{151} & \frac{57}{151} & \frac{61}{151} \end{matrix} \right\} = \left\{ \begin{matrix} 3 & 4 & 5 & 6 \\ 0.08 & 0.14 & 0.38 & 0.40 \end{matrix} \right\}$$

Solo confrontando queste distribuzioni di frequenze relative siamo in grado di affermare (crf. figura 7.2) che i clienti che pagano in contrassegno tendono ad essere

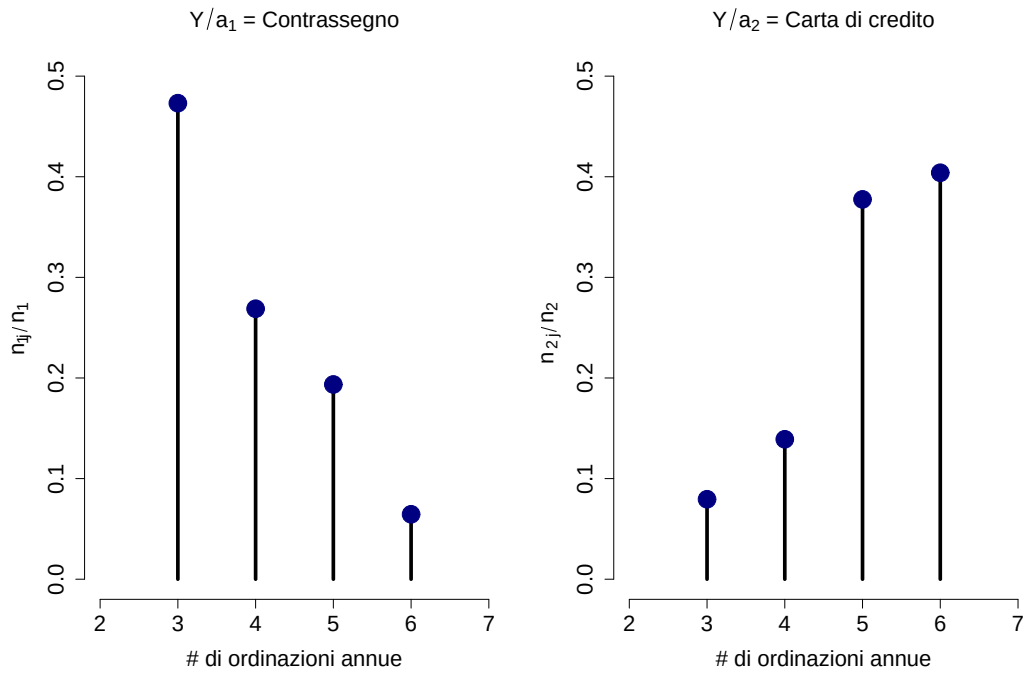


Figura 7.2 Distribuzioni di frequenze delle v.s. condizionate $Y|a_i$, esempio 7.9.

quelli che effettuano pochi ordini e viceversa quelli che pagano con carta di credito sono propensi a fare più ordini durante l'anno.

◁

Prima di concludere questo paragrafo osserviamo che dalla distribuzione di frequenze congiunte di una variabile statistica mista (A, Y) è possibile individuare anche le s distribuzioni di frequenze relative delle mutabili statistiche condizionate $A|y_j$ che risultano, qualunque sia $j = 1, \ldots, s$, essere:

$$A|y_j \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{ij}}{n_{\cdot j}} \end{matrix} \right\}_{i=1,\ldots,r} = \left\{ \begin{matrix} a_1 & \ldots & a_i & \ldots & a_r \\ \frac{n_{1j}}{n_{\cdot j}} & \ldots & \frac{n_{ij}}{n_{\cdot j}} & \ldots & \frac{n_{rj}}{n_{\cdot j}} \end{matrix} \right\}$$

Abbiamo volutamente tralasciato di definire la mutabile statistica condizionata $A|y_j$ poiché è del tutto analoga, mutatis mutandis, a quella data per la variabile statistica condizionata $Y|a_i$.

▷ ESEMPIO 7.10

Se riprendiamo la tabella della distribuzione di frequenze congiunte dell'esempio (7.7) alla quale abbiamo aggiunto le frequenze marginali

$A \downarrow Y \rightarrow$	$y_1 = 3$	$y_2 = 4$	$y_3 = 5$	$y_4 = 6$	
$a_1 =$ contrassegno	44	25	18	6	93
$a_2 =$ carta di credito	12	21	57	61	151
	56	46	75	67	244

ricaviamo le distribuzioni di frequenze relative delle quattro mutabili statistche $A|y_j$ dividendo per le frequenze a margine di ciascuna colonna, cioè:

$$A|y_1 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i1}}{n_{\cdot 1}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ \frac{44}{56} & \frac{12}{56} \end{matrix} \right\} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ 0.79 & 0.21 \end{matrix} \right\}$$

$$A|y_2 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i2}}{n_{\cdot 2}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ \frac{25}{46} & \frac{21}{46} \end{matrix} \right\} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ 0.54 & 0.46 \end{matrix} \right\}$$

$$A|y_3 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i3}}{n_{\cdot 3}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ \frac{18}{75} & \frac{57}{75} \end{matrix} \right\} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ 0.24 & 0.76 \end{matrix} \right\}$$

$$A|y_4 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i4}}{n_{\cdot 4}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ \frac{6}{67} & \frac{61}{67} \end{matrix} \right\} = \left\{ \begin{matrix} \text{contr.} & \text{carta} \\ 0.09 & 0.91 \end{matrix} \right\}$$

◁

In definitiva, da una tabella a doppia entrata si possono individuare $3 + r + s$ distribuzioni di frequenze, e precisamente:

- ⋆ la distribuzione congiunta della v.s. mista (A, Y);
- ⋆ la distribuzione della m.s. univariata A (colonne a margine);
- ⋆ la distribuzione della v.s. univariata Y (righe a margine);
- ⋆ s distribuzioni della m.s. univariata A condizionata alle modalità di Y;
- ⋆ r distribuzioni della v.s. univariata Y condizionata alle modalità di A.

▷ ESEMPIO 7.11

Vediamo in questo esempio come quanto detto sia immediatamente trasferibile al caso di una mutabile statistica bivariata. A tal proposito consideriamo il collettivo statistico costituito da 1200 lavoratori dipendenti residenti in un comune piemontese e studiamo congiuntamente i due caratteri *sesso* e *settore di attività*. Poiché gli insiemi delle modalità dei due caratteri sono rispetivamente

$$M_1 = \{\text{Maschio, Femmina}\} = \{\text{M, F}\}$$

$$M_2 = \{\text{Agricolo, Industriale, Terziario, Servizi}\} = \{\text{A, I, T, S}\}$$

il loro prodotto cartesiano sarà formato dalle 8 coppie di attributi come segue:

$$M_1 \times M_2 = \{(M;\text{A}), (M;\text{I}), (M;\text{T}), (M;\text{S}),\\ (F;\text{A}), (F;\text{I}), (F;\text{T}), (F;\text{S})\}$$

A rilevazione avvenuta disporremo di una mutabile statistica bivariata (A, B) dal cui insieme delle modalità distinte $\{(A, B)(\omega_\alpha)\}_{\alpha=1,\dots,n} = \{(\tilde{a}_\alpha; \tilde{b}_\alpha)\}_{\alpha=1,\dots,n}$, costituito da 1200 coppie di attributi, ricaviamo la distribuzione di frequenze congiunte alla quale aggiungiamo già da ora le frequenze marginali

$A \downarrow B \rightarrow$	$b_1 = \text{A}$	$b_2 = \text{I}$	$b_3 = \text{T}$	
$a_1 = \text{M}$	200	400	100	700
$a_2 = \text{F}$	150	100	250	500
	350	500	350	1200

Ricaviamo da questa tabella le due distribuzioni univariate delle mutabili statistiche A e B, cioè

$$A \equiv \left\{ \begin{matrix} a_i \\ n_{i\cdot} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{M} & \text{F} \\ 700 & 500 \end{matrix} \right\}$$

$$B \equiv \left\{ \begin{matrix} b_j \\ n_{\cdot j} \end{matrix} \right\}_{j=1,\dots,3} = \left\{ \begin{matrix} \text{A} & \text{I} & \text{T} \\ 350 & 500 & 350 \end{matrix} \right\}$$

le due distribuzioni della mutabile statistica B condizionate alle modalità di A, cioè

$$B|a_1 \equiv \left\{ \begin{matrix} b_j \\ \frac{n_{1j}}{n_{1\cdot}} \end{matrix} \right\}_{j=1,\dots,3} = \left\{ \begin{matrix} \text{A} & \text{I} & \text{T} \\ \frac{200}{700} & \frac{400}{700} & \frac{100}{700} \end{matrix} \right\} = \left\{ \begin{matrix} \text{A} & \text{I} & \text{T} \\ 0.29 & 0.57 & 0.14 \end{matrix} \right\}$$

$$B|a_2 \equiv \left\{ \begin{matrix} b_j \\ \frac{n_{2j}}{n_{2\cdot}} \end{matrix} \right\}_{j=1,\dots,3} = \left\{ \begin{matrix} \text{A} & \text{I} & \text{T} \\ \frac{150}{500} & \frac{100}{500} & \frac{250}{500} \end{matrix} \right\} = \left\{ \begin{matrix} \text{A} & \text{I} & \text{T} \\ 0.30 & 0.20 & 0.50 \end{matrix} \right\}$$

ed infine le tre distribuzioni della mutabile statistica A condizionate alle modalità di B, cioè

$$A|b_1 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i1}}{n_{\cdot 1}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{M} & \text{F} \\ \frac{200}{350} & \frac{150}{350} \end{matrix} \right\} = \left\{ \begin{matrix} \text{M} & \text{F} \\ 0.57 & 0.43 \end{matrix} \right\}$$

$$A|b_2 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i2}}{n_{\cdot 2}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{M} & \text{F} \\ \frac{400}{500} & \frac{100}{500} \end{matrix} \right\} = \left\{ \begin{matrix} \text{M} & \text{F} \\ 0.80 & 0.20 \end{matrix} \right\}$$

$$A|b_3 \equiv \left\{ \begin{matrix} a_i \\ \frac{n_{i3}}{n_{\cdot 3}} \end{matrix} \right\}_{i=1,2} = \left\{ \begin{matrix} \text{M} & \text{F} \\ \frac{100}{350} & \frac{250}{350} \end{matrix} \right\} = \left\{ \begin{matrix} \text{M} & \text{F} \\ 0.29 & 0.71 \end{matrix} \right\}$$

◁

7.3.1 MEDIE E VARIANZE CONDIZIONATE

Se fino ad ora non si è riscontrata una sostanziale differenza tra le m.s. doppie e le v.s. miste, vediamo ora che nel caso di v.s. mista sarà possibile, per la componente variabile Y, calcolare tutti i parametri caratteristici che abbiamo definito per le v.s. quando trattavamo il caso univariato. Con riferimento, dunque, alla distribuzione di frequenze congiunte di una variabile mista (A, Y)

$A \downarrow Y \rightarrow$	y_1	$\cdots$	y_j	$\cdots$	y_s	
a_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1s}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{is}	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	n_{r1}	$\cdots$	n_{rj}	$\cdots$	n_{rs}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

consideriamo la componente variabile Y e osserviamo che la sua media aritmetica e la sua varianza possono essere calcolate a partire sia dalla distribuzione di frequenze congiunte che dalla distribuzione marginale della v.s. Y cioè:

$$E[Y] = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} y_j n_{ij} = \frac{1}{n} \sum_{j=1}^{s} y_j n_{\cdot j} = \mu_Y \tag{7.11}$$

$$V[Y] = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} (y_j - E[Y])^2 n_{ij} = \frac{1}{n} \sum_{j=1}^{s} (y_j - E[Y])^2 n_{\cdot j} = \sigma_Y^2 \tag{7.12}$$

Inoltre ciascuna delle r variabili statistiche condizionate $Y|a_i$ è dotata di un proprio valore medio (detto in breve *media condizionata*) e di una propria varianza (detta *varianza condizionata*), per cui, qualunque sia $i = 1, \ldots, r$, si ha:

$$E[Y|a_i] = \frac{1}{n_{i\cdot}} \sum_{j=1}^{s} y_j n_{ij} = \mu_{Y|a_i} \tag{7.13}$$

$$V[Y|a_i] = \frac{1}{n_{i\cdot}} \sum_{j=1}^{s} (y_j - E[Y|a_i])^2 n_{ij} = \sigma^2_{Y|a_i} \tag{7.14}$$

OSSERVAZIONE: ovviamente per media e varianza valgono tutte le proprietà dimostrate nel caso univariato in particolare per il calcolo della varianza di Y utilizzeremo anche in questo caso la proprietà:

$$V[Y] = E[Y^2] - (E[Y])^2 = \frac{1}{n} \sum_{j=1}^{s} y_j^2 n_{\cdot j} - \left(\frac{1}{n} \sum_{j=1}^{s} y_j n_{\cdot j} \right)^2$$

che per la variabile condizionata sarà

$$V[Y|a_i] = E[Y^2|a_i] - (E[Y|a_i])^2 = \frac{1}{n_{i\cdot}} \sum_{j=1}^{s} y_j^2 n_{ij} - \left(\frac{1}{n_{i\cdot}} \sum_{j=1}^{s} y_j n_{ij} \right)^2$$

$\star$

▷ ESEMPIO 7.12

Riprendiamo dall'esempio (7.7) la distribuzione di frequenze congiunte della variabile statistica mista $(A, Y) = \{$*tipo di pagamento, numero di ordini annui* $\}$ rilevata su un collettivo statisitico formato da 244 clienti di una azienda che vende per corrispondenza

$A \downarrow Y \rightarrow$	$y_1 = 3$	$y_2 = 4$	$y_3 = 5$	$y_4 = 6$	
$a_1 =$ contrassegno	44	25	18	6	93
$a_2 =$ carta di credito	12	21	57	61	151
	56	46	75	67	244

Calcoliamo, a titolo esemplificativo, la media della v.s. Y a partire dalla distribuzione congiunta

$$E[Y] = \frac{1}{244} \sum_{i=1}^{2} \sum_{j=1}^{4} y_j n_{ij} = \frac{1}{244}(3 \cdot 44 + 4 \cdot 25 + \ldots + 6 \cdot 61) = 4.627$$

nonché dalla distribuzione marginale:

$$E[Y] = \frac{1}{244}\sum_{j=1}^{4} y_j n_{\cdot j} = \frac{1}{244}(3 \cdot 56 + 4 \cdot 46 + 5 \cdot 75 + 6 \cdot 67) = 4.627$$

Quanto alla varianza, lavorando sulla distribuzione marginale, applicando la proprietà sopra citata sarà:

$$\begin{aligned} V[Y] &= \frac{1}{244}\sum_{j=1}^{4} y_j^2 n_{\cdot j} - (E[Y])^2 = \\ &= \frac{1}{244}(9 \cdot 56 + 16 \cdot 46 + 25 \cdot 75 + 36 \cdot 67) - (4.627)^2 = 1.242 \end{aligned}$$

Se consideriamo le distribuzioni condizionate avremo per la v.s. $Y|a_1$

$$E[Y|a_1] = \frac{1}{93}\sum_{j=1}^{4} y_j n_{1j} = \frac{1}{93}(3 \cdot 44 + 4 \cdot 25 + 5 \cdot 18 + 6 \cdot 6) = 3.849$$

$$\begin{aligned} V[Y|a_1] &= \frac{1}{93}\sum_{j=1}^{4} y_j^2 n_{1j} - (E[Y|a_1])^2 \\ &= \frac{1}{93}(9 \cdot 44 + 16 \cdot 25 + 25 \cdot 18 + 36 \cdot 6) - (3.849)^2 = 0.902 \end{aligned}$$

e per la v.s. $Y|a_2$

$$E[Y|a_2] = \frac{1}{151}\sum_{j=1}^{4} y_j n_{2j} = \frac{1}{151}(3 \cdot 12 + 4 \cdot 21 + 5 \cdot 57 + 6 \cdot 61) = 5.106$$

$$\begin{aligned} V[Y|a_2] &= \frac{1}{151}\sum_{j=1}^{4} y_j^2 n_{2j} - (E[Y|a_2])^2 \\ &= \frac{1}{151}(9 \cdot 12 + 16 \cdot 21 + 25 \cdot 57 + 36 \cdot 61) - (5.106)^2 = 0.850 \end{aligned}$$

◁

7.4. OSSERVAZIONI SULLA VARIABILE STATISTICA DOPPIA

Nella definizione di variabile statistica mista si è lasciato intendere, senza dirlo esplicitamente, che le modalità distinte della componente Y fossero in numero ridotto per cui la distribuzione di frequenze congiunte, facilmente individuabile, fosse formata da un esiguo numero di coppie di modalità distinte. Tuttavia nella pratica, non sempre ciò accade poiché l'insieme delle modalità distinte di Y è sovente assai numeroso; si pensi al caso di una variabile statistica continua che può assumere valori tutti distinti tra loro. Tale problema si può verificare per entrambe le componenti se l'oggetto di studio è una variabile statistica doppia (X, Y).

Come nel caso univariato, ai fini della costruzione della distribuzione di frequenze congiunte è giocoforza ricorrere al raccoglimento dei dati individuali in classi. Ciò facendo, come di consueto, ci esponiamo al rischio di perdere parte dell'informazione contenuta nell'insieme dei dati individuali della v.s. bivariata guadagnando tuttavia in chiarezza di sintesi. L'esempio che segue vuole illustrare per via grafica, con l'ausilio del *diagramma a dispersione*, il meccanismo sottostante al raccoglimento in classi.

▷ ESEMPIO 7.13

La v.s. bivariata (X, Y), rilevata su di un collettivo di 10 unità statisitche possiede il seguente insieme di dati individuali:

$$\begin{aligned}\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,10} = \{&(12.2; 76.4), (12.2; 77.6),\\ &(11.4; 78.9), (11.8; 79.0), (12.9; 79.2), (11.1; 82.6),\\ &(11.8; 84.7), (11.8; 86.3), (12.5; 88.3), (12.3; 89.3)\}\end{aligned}$$

Nel diagramma a dispersione di figura (7.3, pannello a) abbiamo rappresentato le coppie $(\tilde{x}_\alpha; \tilde{y}_\alpha)$ dell'insieme dei dati individuali così da avere una percezione della distribuzione della v.s. doppia.

Scegliendo di raccogliere per la X i dati individuali in quattro classi di ampiezza 0.5 nell'intervallo $]11; 13]$ e sostituendo a ciascuna $\tilde{x}_\alpha$ il valore del centro di classe a cui essa appartiene il diagramma a dispersione diventa quello proposto in figura (7.3, pannello b). Confrontando i due diagrammi si nota come i punti, pur mantenendo lo stesso valore di ordinata, vengono centrati sulla corrispondente classe in ascissa.

Il pannello (c) della stessa figura rappresenta la situazione a cui si perverrebbe raccogliendo i dati individuali della sola componente Y in tre classi di ampiezza 5 nell'intervallo $]75; 90]$. In questo caso le ascisse dei punti rimangono quelle del pannello (a) e le ordinate vengono a coincidere con i corrispondenti centri di classe.

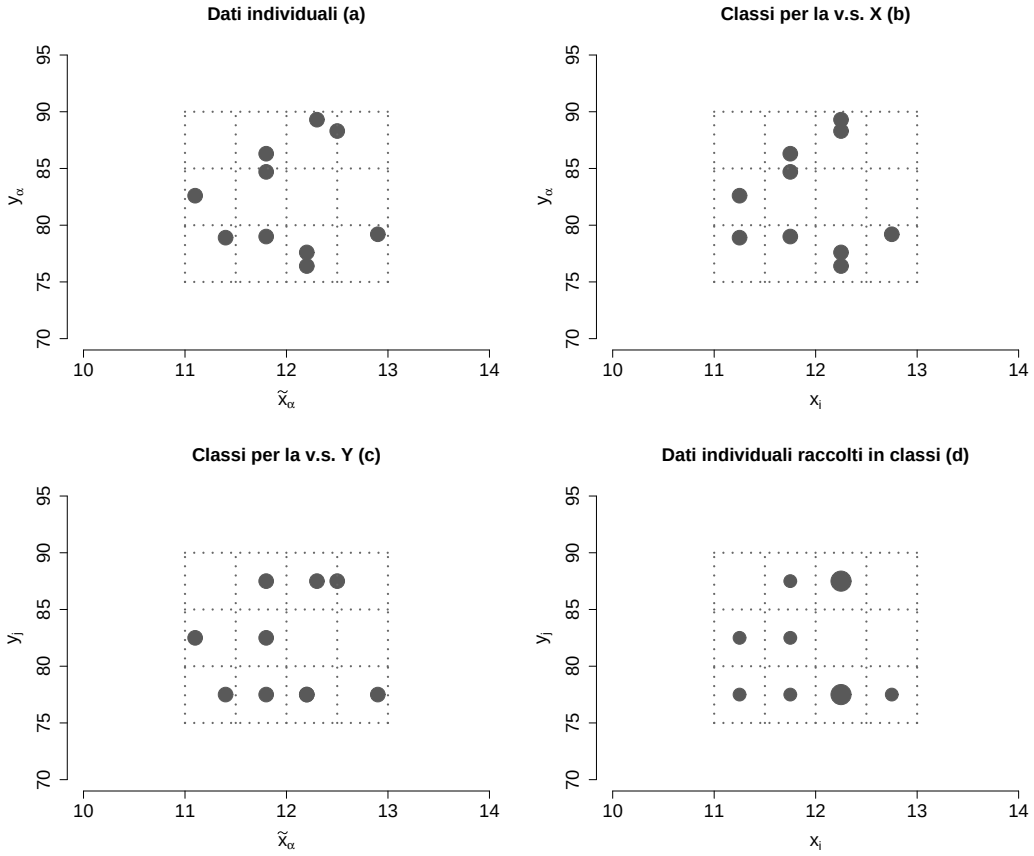


Figura 7.3 Dati inidviduali e raccoglimento in classi, esempio 7.13.

Nel pannello (d) è presentato il *diagramma a bolle* della distribuzione di frequenze congiunte della v.s. con dati raccolti, per entrambe le componenti, secondo le precedenti classi che è la seguente:

$X \downarrow Y \rightarrow$	$75 \dashv 80$	$80 \dashv 85$	$85 \dashv 90$
$11.0 \dashv 11.5$	1	1	0
$11.5 \dashv 12.0$	1	1	1
$12.0 \dashv 12.5$	2	0	2
$12.5 \dashv 13.0$	1	0	0

Si ricorda brevemente che:

- ⋆ con un *diagramma a dispersione* si rappresentano nel piano cartesiano i punti individuati dalle coppie di dati individuali di una v.s. doppia;

⋆ per *diagramma a bolle* intendiamo un diagramma a dispersione per le modalità distinte di una v.s. doppia in cui la grandezza dei simboli utilizzati per rappresentare i punti è proporzionale alla frequenza congiunta associata alla coppia di modalità corrispondente.

◁

Osserviamo ancora che nel caso di variabile statistica bivariata è possibile calcolare media e varianza per entrambe le componenti oltre che per tutte le $r+s$ variabili condizionate. In aggiunta a quanto già detto per la componente Y di una variabile statistica mista abbiamo in questo caso per la componente X

$$E[X] = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} x_i n_{ij} = \frac{1}{n}\sum_{i=1}^{r} x_i n_{i\cdot} = \mu_X \tag{7.15}$$

$$V[X] = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} (x_i - E[X])^2 n_{ij} = \frac{1}{n}\sum_{i=1}^{r} (x_i - E[X])^2 n_{i\cdot} = \sigma_X^2 \tag{7.16}$$

Inoltre ciascuna delle s variabili statistiche condizionate $X|y_j$ è dotata di un proprio valore medio e di una propria varianza per cui, qualunque sia $j = 1, \ldots, s$, si ha:

$$E[X|y_j] = \frac{1}{n_{\cdot j}}\sum_{i=1}^{r} x_i n_{ij} = \mu_{X|y_j} \tag{7.17}$$

$$V[X|y_j] = \frac{1}{n_{\cdot j}}\sum_{i=1}^{r} (x_i - E[X|y_j])^2 n_{ij} = \sigma^2_{X|y_j} \tag{7.18}$$

▷ ESEMPIO 7.14

Aggiungendo le frequenze marginali alla distribuzione della v.s. (X, Y) con dati raccolti in classi dell'esempio (7.13) otteniamo la tabella:

$X\downarrow\ Y\rightarrow$	$75 \dashv 80$ $(y_1 = 77.5)$	$80 \dashv 85$ $(y_2 = 82.5)$	$85 \dashv 90$ $(y_3 = 87.5)$	
$11.0 \dashv 11.5$ $(x_1 = 11.25)$	1	1	0	2
$11.5 \dashv 12.0$ $(x_2 = 11.75)$	1	1	1	3
$12.0 \dashv 12.5$ $(x_3 = 12.25)$	2	0	2	4
$12.5 \dashv 13.0$ $(x_4 = 12.75)$	1	0	0	1
	5	2	3	10

Per quanto riguarda la componente X si hanno, media e varianza

$$E[X] = \frac{1}{10}(11.25 \cdot 2 + 11.75 \cdot 3 + 12.25 \cdot 4 + 12.75 \cdot 1) = 11.950$$

$$V[X] = \frac{1}{10}(11.25^2 \cdot 2 + 11.75^2 \cdot 3 + 12.25^2 \cdot 4 + 12.75^2 \cdot 1) - (11.950)^2 = \\ = 0.210$$

quanto alle distribuzioni condizionate si hanno:

$$E[X|y_1] = \frac{1}{5}(11.25 \cdot 1 + 11.75 \cdot 1 + 12.25 \cdot 2 + 12.75 \cdot 1) = 12.050$$

$$V[X|y_1] = \frac{1}{5}(11.25^2 \cdot 1 + 11.75^2 \cdot 1 + 12.25^2 \cdot 2 + 12.75^2 \cdot 1) - (12.050)^2 = \\ = 0.260$$

$$E[X|y_2] = \frac{1}{2}(11.25 \cdot 1 + 11.75 \cdot 1) = 11.500$$

$$V[X|y_2] = \frac{1}{2}(11.25^2 \cdot 1 + 11.75^2 \cdot 1) - (11.500)^2 = 0.0625$$

$$E[X|y_3] = \frac{1}{3}(11.75 \cdot 1 + 12.25 \cdot 2) = 12.083$$

$$V[X|y_3] = \frac{1}{3}(11.75^2 \cdot 1 + 12.25^2 \cdot 2) - (12.083)^2 = 0.0636$$

Per quanto riguarda la componente Y si lascia al Lettore verificare che:

$$\begin{aligned} E[Y] &= 81.5 & V[Y] &= 19.0 \\ E[Y|x_1] &= 80.0 & V[Y|x_1] &= 6.25 \\ E[Y|x_2] &= 82.5 & V[Y|x_2] &= 16.67 \\ E[Y|x_3] &= 82.5 & V[Y|x_3] &= 25.00 \\ E[Y|x_4] &= 77.5 & V[Y|x_4] &= 0 \end{aligned}$$

In figura (7.4) mediante un diagramma a bolle abbiamo riportato le medie delle v.s. condizionate $Y|x_i$ e $X|y_j$ utilizzando simboli di grandezza proporzionale alle numerosità dello strato a cui esse si riferiscono. Sullo stesso diagramma, per completezza, abbiamo rappresentato anche la distribuzione di frequenze congiunte della v.s.

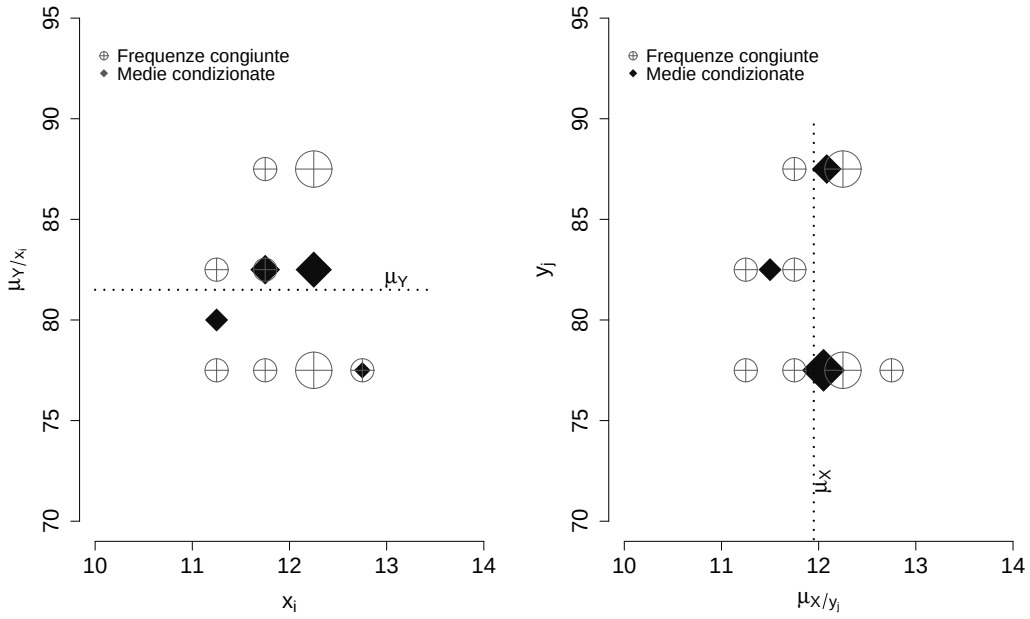


Figura 7.4 Medie delle v.s. condizionate $Y|x_i$ e $X|y_j$, esempio 7.14.

(X, Y) nonché una linea tratteggiata in corrispondenza della media di Y e di X. Si osservi che μ_Y e μ_X sono comprese negli intervalli di variazione delle rispettive medie condizionate.
Si noti infine che le misure di posizione e di variabilità per le componenti X e Y testè calcolate possono, per effetto del duplice raccoglimento dei dati in classi, differire da quelle calcolate a partire di dati individuali.

◁

7.4.1 La covarianza

Nel caso si operi su una variabile statistica bivariata (X, Y) è possibile definire una nuova misura di sintesi detta *covarianza.*

Definizione 7.5 (Covarianza)
si dice covarianza di una v.s. bivariata (X, Y) la media aritmetica del prodotto fra gli scarti delle $\tilde{x}_\alpha$ dalla media di X e gli scarti delle $\tilde{y}_\alpha$ dalla media di Y. Essa corrisponde

pertanto al valore numerico risultante dall'operazione

$$Cov[X,Y] = \frac{1}{n}\sum_{\alpha=1}^{n}(\tilde{x}_\alpha - \mu_X)\,(\tilde{y}_\alpha - \mu_Y) = \sigma_{X,Y} \tag{7.19}$$

□

La covarianza può essere interpretata come misura della *variabilità congiunta* delle due v.s. X e Y rispetto al centro di coordinate $(\mu_X;\mu_Y)$. Con il seguente esempio tentiamo, attraverso il significato geometrico della covarianza, di comprendere il concetto di misura di variabilità congiunta.

▷ ESEMPIO 7.15

A puro scopo interpretativo si consideri l'insieme di dati individuali della v.s. doppia (X,Y) rilevata su un collettivo di 4 unità

$$\{(\tilde{x}_\alpha;\tilde{y}_\alpha)\}_{\alpha=1,\dots,4} = \{(50;157),\ (55;170),\ (62;159),\ (68;175)\}$$

Osservando che le medie delle componenti sono $E[X] = 59.75$ e $E[Y] = 165.25$, valutiamo per ogni punto il suo apporto alla sommatoria nella definizione di covarianza

$$\begin{aligned}
(\tilde{x}_1 - \mu_X)(\tilde{y}_1 - \mu_Y) &= (50 - 59.75)(157 - 165.25) = (-9.75)(-8.25) = \\
&= +80.4375 \\
(\tilde{x}_2 - \mu_X)(\tilde{y}_2 - \mu_Y) &= (55 - 59.75)(170 - 165.25) = (-4.75)(+4.75) = \\
&= -22.5625 \\
(\tilde{x}_3 - \mu_X)(\tilde{y}_3 - \mu_Y) &= (62 - 59.75)(159 - 165.25) = (+2.25)(-6.25) = \\
&= -14.0625 \\
(\tilde{x}_4 - \mu_X)(\tilde{y}_4 - \mu_Y) &= (68 - 59.75)(175 - 165.25) = (+8.25)(+9.75) = \\
&= +80.4375
\end{aligned}$$

Per ogni punto di coordinate $(\tilde{x}_\alpha;\tilde{x}_\alpha)$ gli scarti dalle rispettive medie corrispondono alla lunghezza (con segno) dei lati dei rettangoli evidenziati in figura (7.5) ed il prodotto sarà dunque l'area (con segno) del rettangolo. La covarianza consiste pertanto nella media aritmetica delle aree, con segno, dei rettangoli individuati dalle proiezioni delle coordinate dei punti sugli assi corrispondenti alle medie delle due componenti della variabile statistica. La posizione di ciascun punto rispetto al baricentro viene, per questa misura di variabilità, valutata per mezzo dell'area del rettangolo corrispondente.

◁

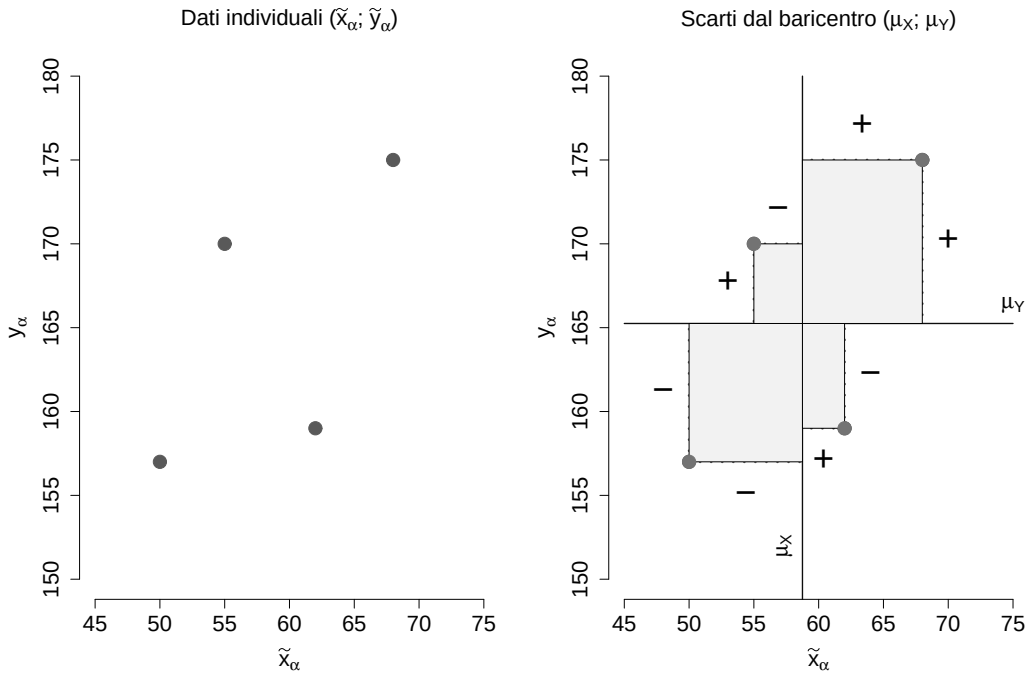


Figura 7.5 Interpretazione geometrica della covarianza, esempio 7.15.

Se le coppie di dati individuali variano nel piano in modo che i punti del loro diagramma a dispersione si dispongono prevalentemente nel primo e terzo quadrante individuato dagli assi delle medie (si veda il primo diagramma a dispersione di figura 7.6), le due componenti della bivariata tendono ad avere un legame direttamente proporzionale e la covarianza assume valore positivo. Viceversa, se le coppie di dati individuali variano nel piano in modo che i punti del loro diagramma a dispersione si dispongono prevalentemente nel secondo e quarto quadrante individuato dagli assi delle medie, le due componenti della bivariata tendono a ad avere un legame inversamente proporzionale e la covarianza assume valore negativo. Nel caso in cui i punti si dispongano in tutti e quattro i quadranti in modo sparso o simmetrico, tra le le due componenti non esiste alcun legame proporzionale e la covarianza assume valori prossimi allo zero.

Manifestamente la definizione di covarianza è stata data in termini di dati individuali; qualora si disponesse della distribuzione di frequenze congiunte della v.s. (X, Y)

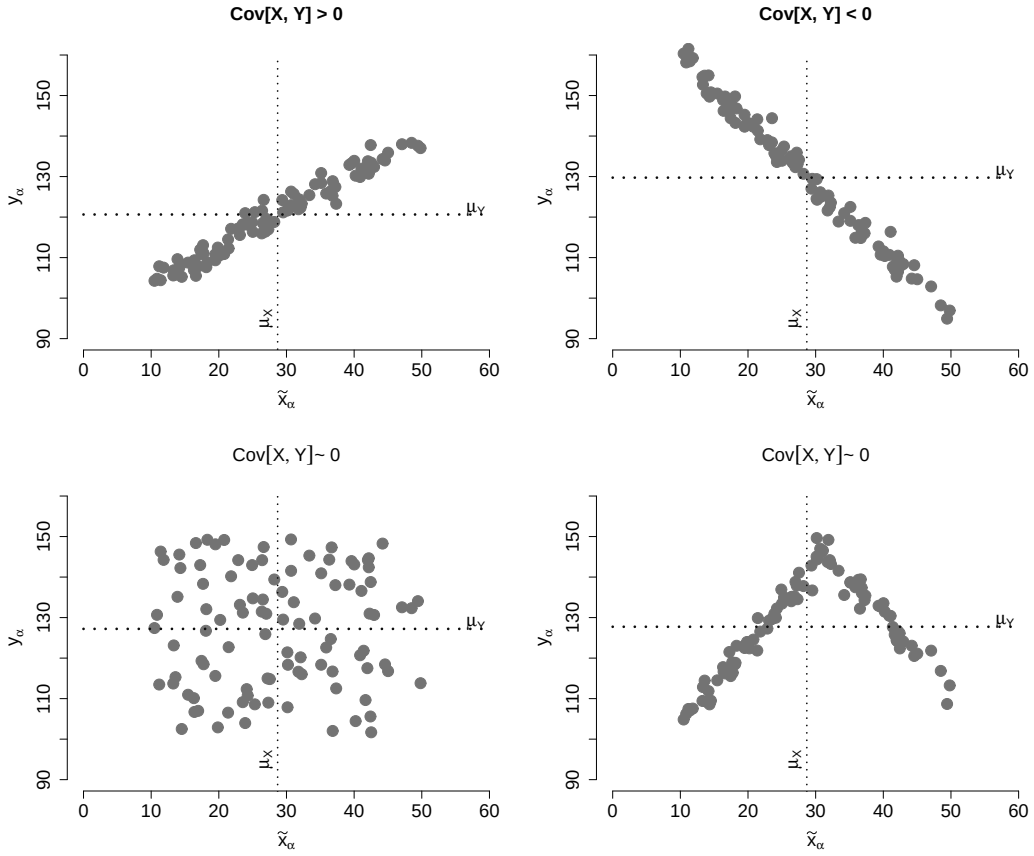


Figura 7.6 La covarianza di particolari variabili statistiche doppie

l'espressione (7.19) diverrebbe:

$$Cov[X,Y] = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i - \mu_X)\,(y_j - \mu_Y)\,n_{ij} \tag{7.20}$$

▷ ESEMPIO 7.16

La tabella che segue riporta la distribuzione di frequenze congiunte della variabile statistica $(X, Y) = \{$*età, anzianità di ruolo*$\}$ rilevata sul collettivo statistico formato dagli insegnanti in ruolo l'anno scorso nelle quattro scuole elementari della città di Catanzaro

$X\downarrow\ Y\rightarrow$	$5 \dashv 15$ $(y_1 = 10)$	$15 \dashv 25$ $(y_2 = 20)$	$25 \dashv 35$ $(y_3 = 30)$	
$25 \dashv 35$ $(x_1 = 30)$	6	2	0	8
$35 \dashv 45$ $(x_2 = 40)$	0	10	10	20
$45 \dashv 65$ $(x_3 = 55)$	0	3	19	22
	6	15	29	50

Osservando che $E[X] = 45$ e $E[Y] = 24.6$, calcoliamo la covarianza utilizzando la formula definita in (7.20), cioè:

$$\begin{aligned} Cov[X,Y] &= \frac{1}{50}\sum_{i=1}^{3}\sum_{j=1}^{3}(x_i - 45)\,(y_j - 24.6)\,n_{ij} = \\ &= \frac{1}{50}((30-45)(10-24.6)\cdot 6 + (30-45)(20-24.6)\cdot 2 + \\ &+ (40-45)(20-24.6)\cdot 10 + (40-45)(30-24.6)\cdot 10 + \\ &+ (55-45)(20-24.6)\cdot 3 + (55-45)(30-24.6)\cdot 19) = 46 \end{aligned}$$

Come era nelle aspettative, la covarianza è positiva il che denota un legame direttamente proporzionale tra l'età e l'anzianità di ruolo degli insegnanti.

$\triangleleft$

Avendo già osservato che la covarianza è la media aritmetica del prodotto degli scarti dalla media delle due componenti della v.s. doppia, possiamo ora ricorrere all'operatore $E[\,\cdot\,]$ e scrivere la (7.19) nella forma:

$$Cov[X,Y] = E\left[(X - E[X])(Y - E[Y])\right] \tag{7.21}$$

Esprimere la covarianza in termini di operatore $E[\,\cdot\,]$ come nella (7.21) è di utile impiego nella dimostrazione di alcune proprietà della covarianza che diamo nel seguito.

Proprietà 7.1 La covarianza è simmetrica negli argomenti, cioè $Cov[X,Y] = Cov[Y,X]$ La dimostrazione segue direttamente dalla definizione.

$\triangleleft$

Proprietà 7.2 La covarianza può essere espressa come differenza tra due valori medi, infatti vale la relazione: $Cov[X,Y] = E[XY] - E[X]E[Y]$

$\triangleleft$

Dimostrazione: dalla (7.21) sviluppando il prodotto $(X-E[X])(Y-E[Y])$ e ricordando le proprietà dell'operatore $E[\,\cdot\,]$ si ha:

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] = \\ &= E[XY - XE[Y] - YE[X] + E[X]E[Y]] = \\ &= E[XY] - E[X]E[Y] - E[Y]E[X] + E[X]E[Y] = \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

□

Proprietà 7.3 La covarianza ammette sempre la maggiorazione:

$$|Cov(X,Y)| \leq \sqrt{V[X]V[Y]} \tag{7.22}$$

◁

Dimostrazione:
Il trucco consiste nel considerare la funzione $\varphi(t)$

$$\varphi(t) = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left[(x_i-\mu_X) + t(y_j-\mu_Y)\right]^2 n_{ij}$$

essa risulta essere maggiore o uguale a zero, per ogni $t \in \mathbb{R}$, in quanto somma di quantità positive. Inoltre sviluppando il quadrato all'interno della sommatoria ci rendiamo conto che $\varphi(t)$ è un polinomio di secondo grado in t, infatti:

$$\begin{aligned} \varphi(t) = t^2\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(y_j-\mu_Y)^2 n_{ij} + 2\,t\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i-\mu_X)(y_j-\mu_Y)n_{ij} + \\ + \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(x_i-\mu_X)^2 n_{ij} \geq 0 \end{aligned}$$

ovvero

$$V[Y]\,t^2 + 2Cov[X,Y]\,t + V[X] \geq 0 \tag{7.23}$$

La funzione $\varphi(t)$ è dunque una parabola a valori sempre maggiori o uguali a zero, con la concavità rivolta verso l'alto e non ammette pertanto radici reali distinte. Tutto ciò premesso imponendo la condizione di non positività del discriminante della (7.23), cioè $\Delta \leq 0$ otteniamo:

$$4\,Cov[X,Y]^2 - 4\,V[X]\,V[Y] \leq 0 \quad \Rightarrow \quad Cov[X,Y]^2 \leq V[X]\,V[Y]$$

da cui la tesi, cioè $|Cov[X,Y]| \leq \sqrt{V[X]V[Y]}$.

□

Prendendo spunto dalla proprietà precedente e dato il particolare ruolo che verrà ad assumere la disuguaglianza (7.22), concludiamo questo paragrafo dando la seguente:

Definizione 7.6 (Coefficiente di correlazione lineare)
chiamiamo coefficiente di correlazione lineare (in simboli ρ), il rapporto

$$\rho = \frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y} \tag{7.24}$$

□

Rimandando al capitolo dedicato alla regressione lineare per l'interpretazione e l'impiego di questo indice, ci limitiamo ad osservare che esso, per costruzione, assume valori compresi tra -1 e $+1$.

▷ ESEMPIO 7.17

Riprendendo la distribuzione di frequenze congiunte della v.s. (X,Y) definita nell'esempio (7.16)

$X\downarrow\ Y\rightarrow$	$5 \dashv 15$ $(y_1 = 10)$	$15 \dashv 25$ $(y_2 = 20)$	$25 \dashv 35$ $(y_3 = 30)$	
$25 \dashv 35$ $(x_1 = 30)$	6	2	0	8
$35 \dashv 45$ $(x_2 = 40)$	0	10	10	20
$45 \dashv 65$ $(x_3 = 55)$	0	3	19	22
	6	15	29	50

ci proponiamo di calcolare la covarianza con l'ausilio della proprietà (7.2), poiché

$$E[X \cdot Y] = \frac{1}{50}(30 \cdot 10 \cdot 6 + 30 \cdot 20 \cdot 2 + 40 \cdot 20 \cdot 10 + 40 \cdot 30 \cdot 10 + \\ + 55 \cdot 20 \cdot 3 + 55 \cdot 30 \cdot 19) = 1153$$

e ricordando che $E[X] = 45$ e $E[Y] = 24.6$ la covarianza sarà:

$$Cov[X,Y] = E[X \cdot Y] - E[X]E[Y] = 11530 - 45 \cdot 24.6 = 46$$

Osservando in ultimo che $V[X] = 90$ e $V[Y] = 48.84$ siamo in grado di calcolare il coefficiente di correlazione lineare ρ cioè:

$$\rho = \frac{Cov[X,Y]}{\sqrt{V[X]V[Y]}} = \frac{46}{\sqrt{90 \cdot 48.84}} = 0.694$$

◁

7.4.2 COMBINAZIONI LINEARI DI VARIABILI STATISTICHE

Avendo rilevato congiuntamente due caratteri quantitativi sulle unità di un collettivo, può essere interessante condurre un'analisi statistica su una particolare combinazione lineare delle due componenti della variabile bivariata corrispondente.
Data una v.s. bivarata (X, Y) si definisce *combinazione lineare* delle sue componenti la v.s. univariata

$$Z = a + b\,X + c\,Y \tag{7.25}$$

con $a, b, c \in \mathbb{R}$.
La (7.25) definisce effettivamente una variabile statistica; Z è, infatti, una applicazione che associa a ciascun elemento del collettivo Ω uno ed un solo numero reale, in altri termini, $\forall \alpha = 1, 2..., n$:

$$\tilde{z}_\alpha = a + b\,\tilde{x}_\alpha + c\,\tilde{y}_\alpha = a + b\,X(\omega_\alpha) + c\,Y(\omega_\alpha) = Z(\omega_\alpha)$$

Se da un punto di vista puramente algebrico la v.s. Z può essere definita qualunque siano i caratteri rilevati, per quanto concerne l'analisi statistica la nuova v.s. Z sarà oggetto di studio se e solo se possiede un senso compiuto la combinazione lineare delle componenti X e Y. Così ad esempio:

- ⋆ rilevando sul collettivo costituito dalle aziende italiane operanti nel settore elettronico la v.s. bivariata $(X, Y) = \{$*fatturato in Italia, fatturato all'estero*$\}$, espressi rispettivamente in euro e dollari, ha senso compiuto la combinazione lineare $Z = X + cY$, dove il parametro c è tale da rendere omogenee le grandezze in gioco trasformando i dollari in euro. Manifestamente Z consiste nel *fatturato totale* espresso in euro;

- ⋆ rilevando sul collettivo costituito dalle famiglie residenti in italia la v.s doppia $(X, Y) = \{$*numero di cellulari, numero di televisori*$\}$ non ha alcun senso considerare la combinazione lineare $Z = X + Y$.

▷ ESEMPIO 7.18

Rilevando su un collettivo statistico formato da 20 automobilisti il *numero di contravvenzioni ricevute per eccesso di velocità* e il *numero di contravvenzioni ricevute per cintura di sicurezza non allacciata* si è ottenuta la v.s. (X, Y) il cui iniseme di

dati individuali risulta il seguente:

$$\{(\tilde{x}_\alpha;\tilde{y}_\alpha)\}_{\alpha=1,\dots,20} = \{(0;1),(0;2),(0;0),(0;0),(0;1),\\ (0;2),(1;0),(1;1),(1;0),(1;2),\\ (1;0),(1;1),(1;0),(2;2),(2;0),\\ (2;0),(2;1),(3;2),(3;0),(3;1)\}$$

Volendo indagare circa il numero di punti patente ancora a disposizione degli automobilisti e ponendo che una contravvenzione per eccesso di velocità comporti la perdita di 5 punti e una contravvenzione per cintura non allacciata 2 punti, introduciamo la combinazione lineare $Z = 20 - 5\,X - 2\,Y$, il cui insieme di dati individuali risulta essere:

$$\{\tilde{z}_\alpha\}_{\alpha=1,\dots,20} = \{18,16,20,20,18,16,15,13,15,\\ 11,15,13,15,6,10,10,8,1,5,3\}$$

Calcolando media e varianza di questa v.s. univariata otteniamo $E[Z] = 12.40$ e $V[Z] = 28.94$.

◁

A ben vedere, qualora si sia interessati ai soli valor medio e varianza della v.s. Z, non è necessario determinare i valori individuali da essa assunti poiché tali parametri sono ricavabili a partire dai corrispondenti parametri delle componenti X e Y che la definiscono. A tal proposito:

- ⋆ per la media $E[Z] = a + b\,E[X] + c\,E[Y]$.

  Per sincerarsene:

$$E[Z] = E[a + b\,X + c\,Y] = \frac{1}{n}\sum_{\alpha=1}^{n}(a + b\,\tilde{x}_\alpha + c\,\tilde{y}_\alpha) = \\ = \frac{1}{n}\sum_{\alpha=1}^{n} a + \frac{1}{n}\sum_{\alpha=1}^{n} b\,\tilde{x}_\alpha + \frac{1}{n}\sum_{\alpha=1}^{n} c\,\tilde{y}_\alpha = a + b\,E[X] + c\,E[Y]$$

- ⋆ per la varianza $V[Z] = b^2\,V[X] + c^2\,V[Y] + 2\,b\,c\,Cov[X,Y]$.

Per sincerarsene, osserviamo che $V[Z] = E[Z^2] - (E[Z])^2$ e applichamo le proprietà dell'operatore $E[\,\cdot\,]$:

$$\begin{aligned} V[Z] &= E[(a + b\,X + c\,Y)^2] - (a + b\,E[X] + c\,E[Y])^2 = \\ &= a^2 + b^2 E[X^2] + c^2 E[Y^2] + 2abE[X] + 2acE[Y] + 2bcE[X \cdot Y] - \\ &- a^2 - b^2 E[X]^2 - c^2 E[Y]^2 - 2abE[X] - 2acE[Y] - 2bcE[X]E[Y] = \\ &= b^2\,V[X] + c^2\,V[Y] + 2\,b\,c\,Cov[X,Y] \end{aligned}$$

▷ ESEMPIO 7.19

Riprendendo la v.s. (X, Y) dell'esempio (7.18), dalla conoscenza di media e varianza delle sue componenti nonché della covarianza, ricaviamo media e varianza della combinazione lineare $Z = 20 - 5\,X - 2\,Y$.
Essendo

$$E[X] = 1.20 \quad E[Y] = 0.80 \quad V[X] = 1.06 \quad V[Y] = 0.66 \quad Cov[X,Y] = -0.01$$

si ha immediatamente

$$\begin{aligned} E[Z] &= 20 - 5\,E[X] - 2\,E[Y] = 12.40 \\ V[Z] &= 5^2\,V[X] + (-2)^2\,V[Y] + 2 \cdot (-5) \cdot (-2)\,Cov[X,Y] = 28.94 \end{aligned}$$

◁

▷ ESEMPIO 7.20

Un'indagine condotta su di un collettivo statistico costituito da 204 coppie di sposi, circa il reddito mensile lordo della moglie (X) e il reddito mensile lordo del marito (Y), espressi in in migliaia di euro, ha dato luogo alla variabile statistica doppia (X, Y) con distribuzione di frequenze congiunte

X ↓ Y →	$1 \dashv 2$ $(y_1 = 1.5)$	$2 \dashv 3$ $(y_2 = 2.5)$	$3 \dashv 4$ $(y_3 = 3.5)$	
$0 \dashv 1$ $(x_1 = 0.5)$	12	42	49	103
$1 \dashv 2$ $(x_2 = 1.5)$	25	54	22	101
	37	96	71	204

Le medie e varianze delle v.s. X e Y risultano essere rispettivamente:

$$E[X] = 0.995 \quad V[X] = 0.2500 \quad E[Y] = 2.667 \quad V[Y] = 0.5016$$

Volendo indagare circa il reddito mensile lordo familiare, definiamo la combinazione lineare $Z = X + Y$ e, senza individuarne la distribuzione, ne determiniamo valor medio e varianza.
Quanto alla media, si ha:

$$E[Z] = E[X] + E[Y] = 0.995 + 2.667 = 3.662$$

Per la varianza, dal momento che $Cov[X, Y] = -0.0972$, si ha:

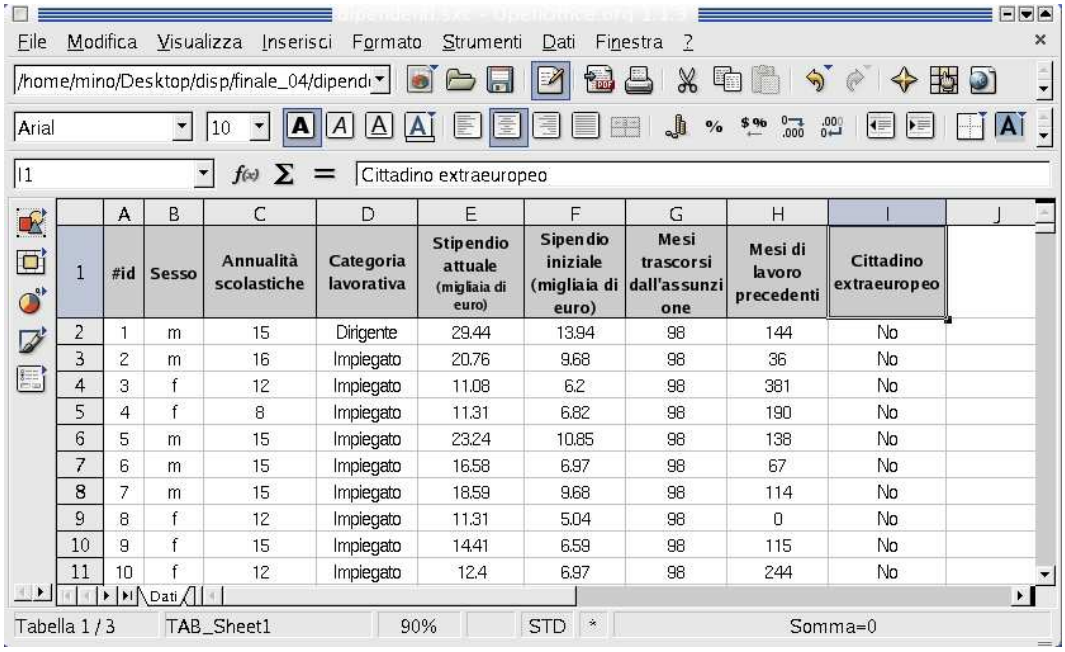
$$\begin{aligned} V[Z] &= V[X] + V[Y] + 2Cov[X, Y] = \\ &= 0.2500 + 0.5016 - 2 \cdot 0.0972 = 0.5572 \end{aligned}$$

◁

7.5. IL FOGLIO ELETTRONICO

In questo paragrafo introduciamo un nuovo file di dati che utilizzeremo per illustrare alcuni possibili modi di operare con il foglio elettronico *OpenOffice 1.1.2* per lo studio di congiunto di due caratteri.
Il file `dipendenti.sxc` contiene i dati relativi a 473 dipendenti dell'amministrazione di una regione italiana.
La Figura (7.7) riporta la videata OpenOffice delle prime righe del foglio di lavoro. Si noti che sono stati rilevati i seguenti caratteri:

- ⋆ il *sesso*, con modalità codificate in *f* se Femmina, *m* se Maschio;
- ⋆ il *numero di anni di scuola*;
- ⋆ la *categoria lavorativa*, con modalità *Impiegato, Funzionario, Dirigente*;
- ⋆ lo *stipendio annuo lordo attuale* in migliaia di euro;
- ⋆ lo *stipendio annuo lordo iniziale* in migliaia di euro;
- ⋆ il *numero di mesi trascorsi dall'assunzione nella categoria lavorativa attuale*
- ⋆ il *numero di mesi lavorati prima dell'assunzione nella categoria lavorativa attuale*
- ⋆ il *cittadinanza extraeuropea*, con modalità *no* se cittadino europeo, *sì* se extrauropeo



	A	B	C	D	E	F	G	H	I	J
1	#id	Sesso	Annualità scolastiche	Categoria lavorativa	Stipendio attuale (migliaia di euro)	Sipendio iniziale (migliaia di euro)	Mesi trascorsi dall'assunzione	Mesi di lavoro precedenti	Cittadino extraeuropeo	
2	1	m	15	Dirigente	29.44	13.94	98	144	No	
3	2	m	16	Impiegato	20.76	9.68	98	36	No	
4	3	f	12	Impiegato	11.08	6.2	98	381	No	
5	4	f	8	Impiegato	11.31	6.82	98	190	No	
6	5	m	15	Impiegato	23.24	10.85	98	138	No	
7	6	m	15	Impiegato	16.58	6.97	98	67	No	
8	7	m	15	Impiegato	18.59	9.68	98	114	No	
9	8	f	12	Impiegato	11.31	5.04	98	0	No	
10	9	f	15	Impiegato	14.41	6.59	98	115	No	
11	10	f	12	Impiegato	12.4	6.97	98	244	No	

Figura 7.7 Videata OpenOffice, file `dipendenti.sxc`

Ci proponiamo ora di determinare la distribuzione di frequenze congiunte della mutabile statistica doppia $(A, B) = \{$*Cittadinanza, Categoria lavorativa*$\}$.

Riportiamo in una nuova cartella le due colonne corrispondenti alle mutabili in oggetto e affianchiamo a queste una colonna con intestazione `unità` contenente in tutte le celle il numero 1. Dopo aver selezionato le tre colonne scegliamo dal menù `Dati` l'opzione `DataPilot` e `Avvia` così come evidenziato in figura (7.8).

Verrà, a questo punto, proposta la maschera di figura (7.9) nella quale compaiono a destra della tabella tre "bottoni" rettangolari nominati come l'intestazione rispettivamente delle tre colonne selezionate. Cliccando sul bottone `Categoria lavorativa` lo abbiamo trascinato sulla maschera della tabella nella posizione `Colonna` per indicare che le modalità di tale mutabile dovranno essere poste nelle colonne; mentre il bottone `Cittadino extraeuropeo` è stato trascinato nella posizione `Riga`. Infine il bottone `unità` è stato trascinato nella posizione `Dati` determinando così le frequenze congiunte.

Si noti che nella parte riguardante il `Risultato` abbiamo selezionato `nuova tabella` così che la tabella della distribuzione di frequenze restituita dalla procedura verrà posta in un nuovo foglio di lavoro così come si vede in figura (7.10).

La procedura `DataPilot` sarà anche utile per la determinazione della distribuzione di

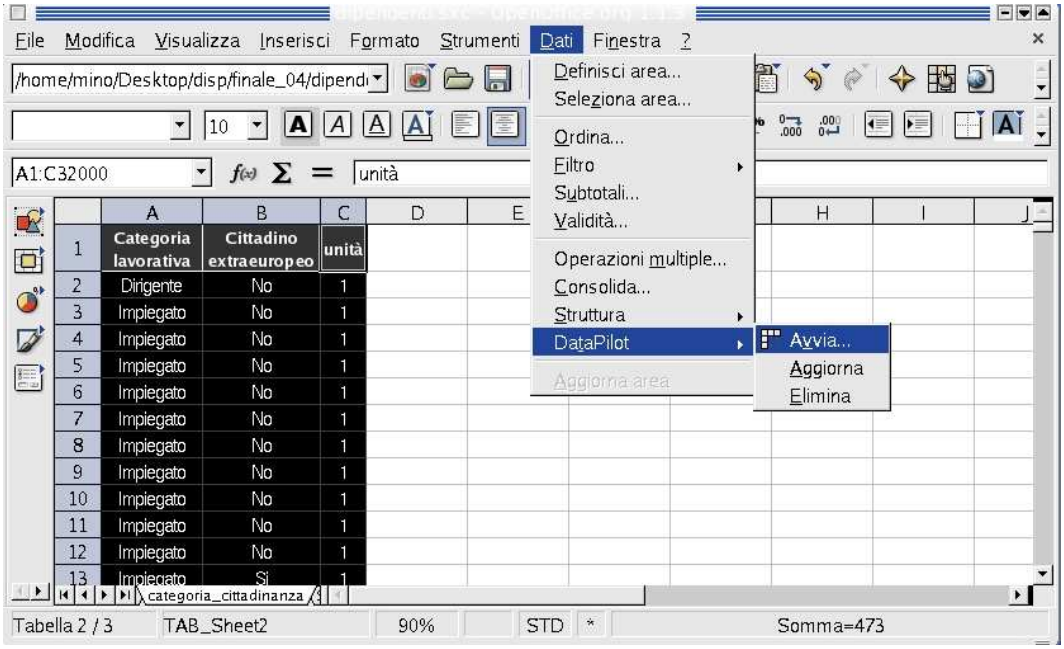


Figura 7.8 Opzione DataPilot per la distribuzione di frequenze congiunte

frequenze congiunte di variabili statistiche miste e di variabili doppie. Tuttavia, se si desidera raccogliere in classi i dati individuali di una delle componenti della variabile e successivamente ottenere la corrispondente distribuzione di frequenze congiunte è necessario, prima di utilizzare `DataPilot`, "ricodificare" la componente con i centri di classe.
Immaginiamo di volere la distribuzione di frequenze congiunte della variabile statistica mista $(A, Y) = \{$*Sesso, Stipendio attuale*$\}$, e di scegliere per la componente Y le seguenti sei classi di stipendio:

$$]5;10] \quad]10;15] \quad]15;20] \quad]20;30] \quad]30;40] \quad]40;60]$$

A tal fine si consideri la colonna `C` della videata proposta in figura (7.11) nella quale si trova la variabile *stipendio ricodificato*, al variare della riga compare nella cella il valore del centro di classe a cui appartiene il valore corrispondente della variabile *Stipendio iniziale*. Per creare tale colonna abbiamo utilizzato la funzione `SE()` nidificata cinque volte, così, ad esempio nella cella `C2`, abbiamo inserito la funzione

=SE(B2<=10;7.5;SE(B2<=15;12.5;SE(B2<=20;17.5;SE (B2<=30;25;SE(B2<=40;35;50)))))

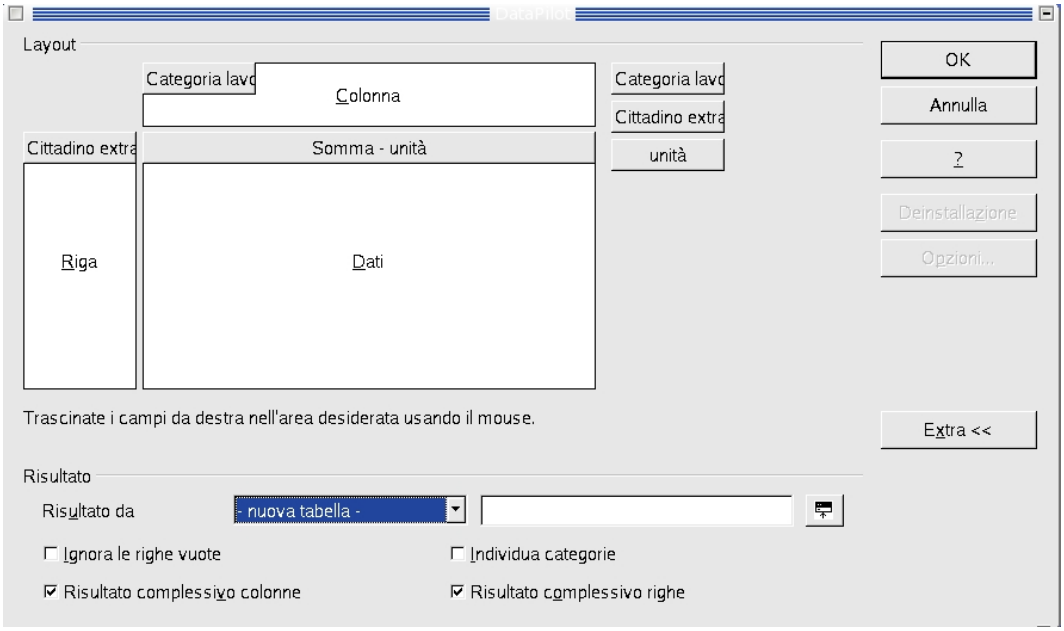


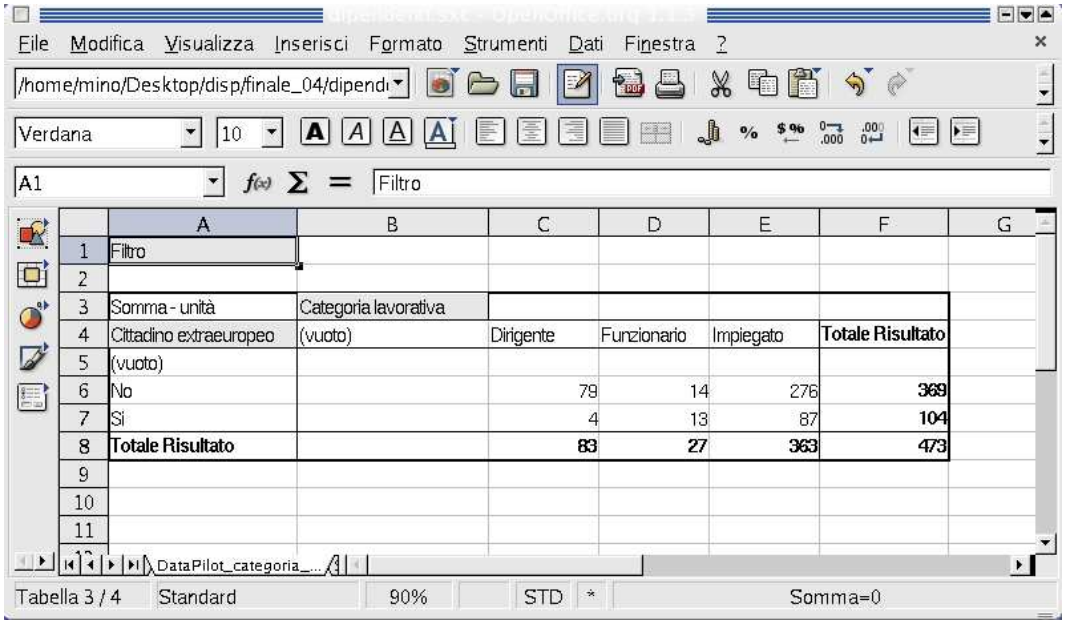
Figura 7.9 Maschera di creazione della tabella

Con la condizione `B2<=10` del primo `SE` verifichiamo se lo Stipendio iniziale in `B2` appartiene alla prima classe, se ciò è vero attribuiamo 7.5 alla cella `C2` che è il valore centrale della prima classe, se la condizione non è verificata nidifichiamo un altro `SE()` all'interno del primo per determinare se il contenuto di `B2` appartiene alla seconda classe e così via. Si osservi che il `SE(B2<=40;35;50)`, che è il più interno, verifica se il valore di `B2` appartiene alla penultima classe e che se ciò non è verificato si attribuisce, per esclusione, alla cella `C2` il valore centrale dell'ultima classe.
Solo dopo aver creato la colonna `C` sarà possibile, con l'impiego della procedura `DataPilot` così come descritto per la mutabile bivariata $(A, B) = \{$*Cittadinanza, Categoria lavorativa*$\}$, ottenere la distribuzione di frequenze congiunte della variabile statistica mista $(A, Y) = \{$*Sesso, Stipendio attuale*$\}$, che si ha in figura (7.11) nell'intervallo di celle `F5:N10`.

7.6. ESERCIZI

▷ ESERCIZIO 7.1

Si immagini che la rilevazione congiunta dei due caratteri mutabili A=$\{$*sesso*$\}$ e B=$\{$*nazionalità*$\}$, condotta sui 54 ospiti di un villaggio turistico, abbia dato luogo



	A	B	C	D	E	F	G
1	Filtro						
2							
3	Somma - unità	Categoria lavorativa					
4	Cittadino extraeuropeo	(vuoto)	Dirigente	Funzionario	Impiegato	**Totale Risultato**	
5	(vuoto)						
6	No		79	14	276	**369**	
7	Si		4	13	87	**104**	
8	**Totale Risultato**		**83**	**27**	**363**	**473**	
9							
10							
11							

Figura 7.10 Distribuzione di frequenze congiunte m.s. *(Categoria, Cittadinanza)*

alle seguenti coppie di valori individuali $(\tilde{a}_\alpha; \tilde{b}_\alpha)$:

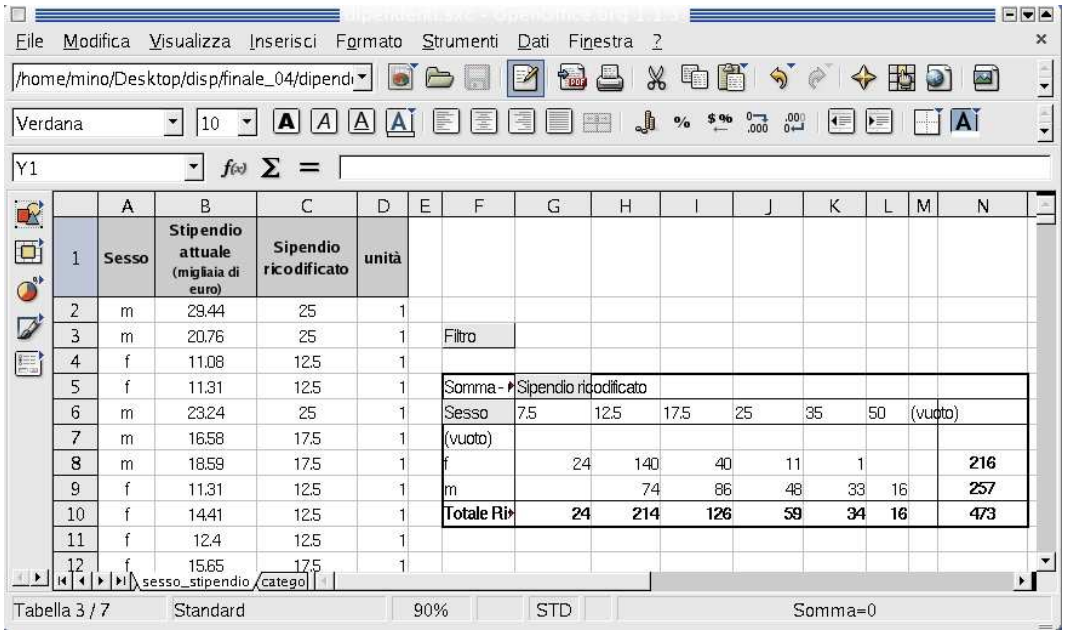
(M, I)	(M, I)	(M, S)	(M, I)	(M, S)	(M, I)	(M, I)	(M, I)	(M, I)
(M, S)	(M, I)	(M, S)	(M, S)	(M, S)	(M, I)	(F, I)	(M, I)	(F, S)
(F, I)	(M, I)	(M, I)	(M, I)	(F, I)	(M, I)	(F, I)	(F, I)	(F, S)
(F, S)	(F, I)	(M, I)	(M, I)	(M, I)	(F, I)	(F, I)	(F, S)	(F, S)
(F, I)	(F, I)	(F, S)	(F, S)	(M, I)	(M, I)	(M, I)	(M, I)	(M, S)
(F, I)	(M, I)	(M, I)	(M, I)	(F, I)	(M, I)	(F, I)	(F, I)	(F, S)

dove, per praticità, per le mutabili statistiche in esame si è scelto di porre:

$$A = \begin{cases} \text{F} & \text{femmina} \\ \text{M} & \text{maschio} \end{cases} \qquad B = \begin{cases} \text{I} & \text{italiano/a} \\ \text{S} & \text{straniero/a} \end{cases}$$

Si presenti in forma tabellare la distribuzione delle frequenze assolute congiunte della m.s. doppia (A, B).

◁



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Sesso	Stipendio attuale (migliaia di euro)	Sipendio ricodificato	unità										
2	m	29.44	25	1										
3	m	20.76	25	1		Filtro								
4	f	11.08	12.5	1										
5	f	11.31	12.5	1		Somma -	Sipendio ricodificato							
6	m	23.24	25	1		Sesso	7.5	12.5	17.5	25	35	50	(vuoto)	
7	m	16.58	17.5	1		(vuoto)								
8	m	18.59	17.5	1		f	24	140	40	11	1			216
9	f	11.31	12.5	1		m		74	86	48	33	16		257
10	f	14.41	12.5	1		Totale Ri	24	214	126	59	34	16		473
11	f	12.4	12.5	1										
12	f	15.65	17.5	1										

Figura 7.11 Distribuzione di frequenze congiunte v.s. mista *(Sesso, Stipendio attuale)*

▷ ESERCIZIO 7.2

Con riferimento alla situazione di cui all'esercizio 7.1, si individuino e si confrontino tra loro la distribuzione marginale e le distribuzioni condizionate della componente B=$\{$*nazionalità*$\}$.

◁

▷ ESERCIZIO 7.3

Da un'indagine condotta sui 200 prodotti commercializzati in un grande magazzino, con riferimento ai caratteri:

$$A = \{\textit{prodotto presente sul volantino pubblicitario inviato ai clienti}\}$$
$$B = \{\textit{reparto di vendita}\}$$

è emersa la seguente distribuzione di frequenze relative congiunte:

$B \rightarrow$ $A \downarrow$	*Casalinghi*	*Abbigliamento*	*Alimentari*	*Detersivi*
Sì	0.02	0.05	0.50	0.03
No	0.08	0.10	0.15	0.07

Si individuino:

- $\star$ la distribuzione di frequenze assolute congiunte;
- $\star$ la distribuzione condizionata, espressa in frequenze relative, del reparto di vendita dei prodotti presenti sul volantino pubblicitario;
- $\star$ la distribuzione condizionata, espressa in frequenze relative, della presenza sul volantino dei prodotti venduti nel reparto casalinghi.

$\triangleleft$

$\triangleright$ ESERCIZIO 7.4

Un'indagine compiuta sugli Hotel della regione Toscana ha comportato la rilevazione dei seguenti caratteri: *categoria dell'hotel* e *numero di stanze.* La variabile mista (A, Y) risultante ha distribuzione di frequenze congiunte:

Y $\rightarrow$ A $\downarrow$	30	50	80	120
$\star\,\star$	60	44	20	6
$\star\,\star\,\star$	44	36	40	50
$\star\,\star\,\star\,\star$	26	38	56	74

Si proceda a:

- $\star$ individuare le distribuzioni marginali della m.s. A e della v.s. Y;
- $\star$ individuare le $s = 4$ distribuzioni delle mutabili condizionate $A|y_j$;
- $\star$ individuare le $r = 3$ distribuzioni delle variabili condizionate $Y|a_i$ e calcolare per ciascuna di esse media e varianza;
- $\star$ calcolare la mediana della v.s. Y nonché quella della v.s. condizionata $Y|a_2$.

$\triangleleft$

▷ ESERCIZIO 7.5

Un'indagine compiuta sui turisti sbarcati dai voli internazionali in arrivo all'aeroporto di Malpensa 2000 il 9 gennaio 2005 ha comportato la rilevazione dei caratteri: *giudizio sui servizi fruiti durante il soggiorno* e *spesa individuale giornaliera per il soggiorno.* La variabile mista (A, Y) risultante ha distribuzione di frequenze congiunte:

Y $\rightarrow$ A $\downarrow$	$0 \dashv 50$	$50 \dashv 75$	$75 \dashv 100$	$100 \dashv 200$
ottimo	2	9	15	10
buono	7	10	25	14
suff.	12	30	13	8
insuff.	8	7	2	0

Si proceda a:

- $\star$ individuare le distribuzioni della m.s. A e della v.s. Y,
- $\star$ individuare le $r = 4$ distribuzioni delle variabili condizionate $Y|a_i$ e calcolare per ciascuna media e varianza;

$\triangleleft$

▷ ESERCIZIO 7.6

Data la distribuzione di frequenze congiunte della variabile statistica mista (A, Y):

Y $\rightarrow$ A $\downarrow$	0	1	2	3	4	5
a_1	2	10	9	5	1	0
a_1	0	3	15	31	10	8

rappresentare i diagrammi a scatole e baffi per le distribuzioni delle v.s. condizionate $Y|a_1$ e $Y|a_2$ tentandone un confronto.

$\triangleleft$

▷ ESERCIZIO 7.7

Da un censimento di imprese artigiane del settore manifatturiero si è rilevata la seguente v.s. (X, Y) dove:

$X = \{$*tributi versati nell'anno 2004*$\}$ in migliaia di euro
$Y = \{$*volume degli affari nell'anno 2004*$\}$ in migliaia di euro.

con distribuzione congiunta di frequenze assolute:

Y $\rightarrow$ X $\downarrow$	$0 \dashv 100$	$100 \dashv 200$	$200 \dashv 300$	$300 \dashv 500$
$0 \dashv 30$	8	2	0	0
$30 \dashv 60$	3	20	9	1
$60 \dashv 90$	1	25	55	37
$90 \dashv 150$	0	3	7	10

Si proceda a:

- $\star$ individuare le distribuzioni marginali delle v.s. X e Y;
- $\star$ calcolare la covarianza tra X e Y e successivamente ρ;
- $\star$ individuare una trasformazione lineare delle due v.s. X e Y che abbia un senso compiuto ed individuarne media, varianza e distribuzione di frequenze.

◁

▷ ESERCIZIO 7.8

Si sono eseguite 500 misurazioni della temperatura (Y) dell'acqua di un lago percorso da correnti a diverse profondità (X) ottenendo la seguente distribuzione di frequenze congiunte:

Temperatura Y $\rightarrow$ *Profondità* X $\downarrow$	$5 \dashv 10$	$10 \dashv 20$	$20 \dashv 25$
$5 \dashv 10$	40	10	50
$10 \dashv 15$	80	100	20
$15 \dashv 20$	80	20	100

Si proceda a:

- $\star$ calcolare media e varianza della temperatura del lago;

- ⋆ individuare la distribuzione della profondità condizionata alla temperatura massima;
- ⋆ calcolare la covarianza di X e Y nonché il corrispondente coefficiente di correlazione lineare.

◁

▷ ESERCIZIO 7.9

Si immagini che la rilevazione di due caratteri quantitativi su una popolazione di 10 unità statistiche abia dato luogo alla variabile statistica bivariata (X, Y) con insieme dei dati individuali:

$$\begin{aligned}\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\ldots,10} = \{&(12.5; 10.5), (10.7; 10.8), (13.3; 12.5), (18.0; 20.3),\\ &(18.6; 19.6), (13.4; 12.4), (17.5; 16.7), (15.0; 14.1),\\ &(16.1; 17.2), (15.7; 16.8)\}\end{aligned}$$

Calcolata la covarianza tra X e Y, determinare media e varianza della v.s. $W = X + Y$.

◁

▷ ESERCIZIO 7.10

Di una variabile staistica bivariata (X, Y) è noto che

$$E[X] = \frac{1}{2}\cdot E[Y] = 6 \qquad E[X^2] = \frac{1}{4}\cdot E[Y^2] = 52 \qquad E[X\cdot Y] = \frac{6}{5}\cdot E[X]\cdot E[Y]$$

Calcolare $Cov[X, Y]$ e $\rho_{X,Y}$.
Introdotte, ora, le trasformate $Z = \dfrac{X - \mu_X}{\sigma_X}$ e $W = \dfrac{Y - \mu_Y}{\sigma_Y}$, calcolare $Cov[Z, W]$ e $\rho_{Z,W}$.

◁

▷ ESERCIZIO 7.11

Con riferimento ai dati dell'esercizio 7.10, calcolare media e varianza delle combinazioni lineari $U = 7 \cdot X - 5 \cdot Y + 2$ e $T = X - Y$.

◁

CAPITOLO 8

L'INDIPENDENZA

Introdurremo i concetti di indipendenza statistica e di indipendenza in media e, per ciascuno di essi, perverremo alla costruzione di un indice assoluto e di un indice normalizzato di dipendenza. Esula dallo scopo di questo capitolo la ricerca degli eventuali legami funzionali tra le variabili in esame, che verrà considerata in seguito. Definizioni e proprietà sono qui date per una variabile statistica mista; là dove possibile, gli stessi concetti verranno estesi alle muabili e variabili statistiche bivariate con l'ausilio di esempi di specie.

8.1. INDIPENDENZA STATISTICA

Se nel capitolo precedente si sono poste le basi per un'analisi bivariata dei dati statistici, nel seguito esamineremo in dettaglio come sia possibile evidenziare la presenza di legami di dipendenza tra variabili o le mutabili statistiche congiuntamente rilevate su un medesimo collettivo e in caso affermativo misurarne l'intensità mediante opportuni indici di dipendenza.
Anche in questa occasione, per semplicità di esposizione, faremo riferimento unicamente ad una variabile statistica mista (A, Y) per la quale sia definita la distribuzione di frequenze congiunte, ricordando al Lettore che i concetti via, via esposti possono venire estesi, senza alcuna difficoltà, alle mutabili ed alle variabili statistiche bivariate.
Data, dunque, una variabile statistica mista (A, Y) può essere interessante chiedersi se, ad esempio, i valori assunti da Y sul collettivo statistico sono in qualche modo influenzati dalle modalità assunte dalla mutabile statistica A.

Definizione 8.1 (Indipendenza di Y rispetto ad A)
data una variabile statistica mista (A, Y), diremo che la v.s Y è indipendente dalla m.s. A se e solo se risultano essere identiche tra loro le r distribuzioni di frequenze delle v.s. condizionate $Y|a_i$.

□

Dalla definizione appena proposta, segue che la v.s. Y è indipendente da A se e solo se le r distribuzioni di frequenze di Y condizionata a ciascuna modalità di A, e precisamente:

$$Y|a_1 \equiv \left\{ \begin{matrix} y_1 & \dots & y_j & \dots & y_s \\ \frac{n_{11}}{n_{1\cdot}} & \dots & \frac{n_{1j}}{n_{1\cdot}} & \dots & \frac{n_{1s}}{n_{1\cdot}} \end{matrix} \right\}$$

$$\dots\dots$$

$$Y|a_i \equiv \left\{ \begin{matrix} y_1 & \dots & y_j & \dots & y_s \\ \frac{n_{i1}}{n_{i\cdot}} & \dots & \frac{n_{ij}}{n_{i\cdot}} & \dots & \frac{n_{is}}{n_{i\cdot}} \end{matrix} \right\}$$

$$\dots\dots$$

$$Y|a_r \equiv \left\{ \begin{matrix} y_1 & \dots & y_j & \dots & y_s \\ \frac{n_{r1}}{n_{r\cdot}} & \dots & \frac{n_{rj}}{n_{r\cdot}} & \dots & \frac{n_{rs}}{n_{r\cdot}} \end{matrix} \right\}$$

risultano uguali tra loro. Nel caso di indipendenza sarà, qualunque sia $i = 1, \dots, r$:

$$Y|a_i \equiv \left\{ \begin{matrix} y_1 & \dots & y_j & \dots & y_s \\ \delta_1 & \dots & \delta_j & \dots & \delta_s \end{matrix} \right\}$$

con, $\forall j = 1, \dots, s$:

$$\delta_j = \frac{n_{1j}}{n_{1\cdot}} = \dots = \frac{n_{ij}}{n_{i\cdot}} = \dots = \frac{n_{rj}}{n_{r\cdot}} \tag{8.1}$$

▷ ESEMPIO 8.1

Si immagini che un'indagine condotta sui 570 sottoscrittori di polizze RC auto di un'agenzia assicurativa, circa il *numero annuo di sinistri denunciati* (Y) ed il *tipo di veicolo assicurato* (A), abbia dato luogo alla variabile statistica mista (A, Y) con distribuzione di frequenze congiunte assolute:

$A \downarrow Y \rightarrow$	0	1	2
Motocicli	124	44	12
Autovetture	225	55	37
Autocarri	64	7	2

Per verificare l'indipendenza della v.s. Y dalla m.s A, in accordo con la definizione (8.1), è sufficiente verificare se le $r = 3$ distribuzioni di frequenze delle v.s.

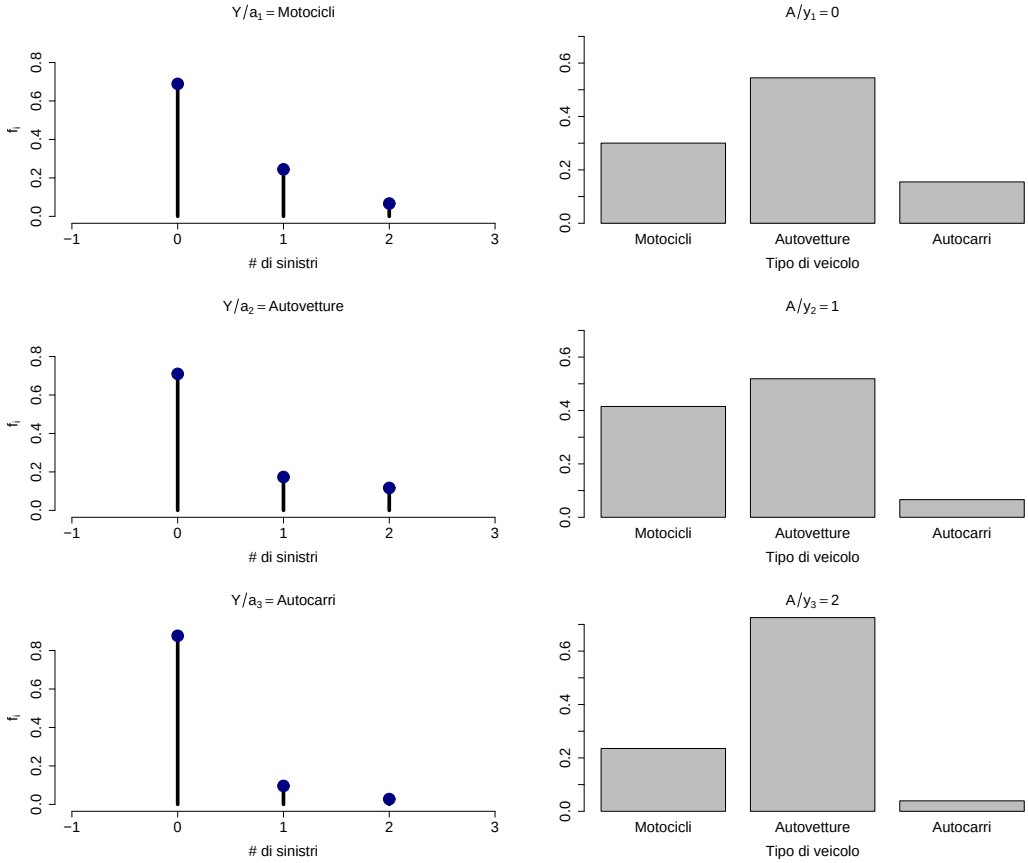


Figura 8.1 Distribuzioni di frequenze di $Y|a_i$ e $A|y_j$, esempio 8.1.

condizionate $Y|a_i$ sono tra loro uguali. Dal momento che (cfr. figura 8.1):

$$Y|a_1 \equiv \left\{ \begin{matrix} 0 & 1 & 2 \\ \frac{124}{180} & \frac{44}{180} & \frac{12}{180} \end{matrix} \right\} = \left\{ \begin{matrix} 0 & 1 & 2 \\ 0.69 & 0.24 & 0.07 \end{matrix} \right\}$$

$$Y|a_2 \equiv \left\{ \begin{matrix} 0 & 1 & 2 \\ \frac{225}{317} & \frac{55}{317} & \frac{37}{317} \end{matrix} \right\} = \left\{ \begin{matrix} 0 & 1 & 2 \\ 0.71 & 0.17 & 0.12 \end{matrix} \right\}$$

$$Y|a_3 \equiv \left\{ \begin{matrix} 0 & 1 & 2 \\ \frac{64}{73} & \frac{7}{73} & \frac{2}{73} \end{matrix} \right\} = \left\{ \begin{matrix} 0 & 1 & 2 \\ 0.88 & 0.10 & 0.03 \end{matrix} \right\}$$

possiamo affermare che la v.s. Y non è indipendente, dal punto di vista statistico, dalla m.s. A; in altri termini potremmo dire che il numero annuo di sinistri denunciati

presso l'agenzia assicuratrice viene in qualche modo a dipendere dal tipo di veicolo assicurato.

◁

Evidentemente un'analoga definizione di indipendenza può essere data per la m.s. A rispetto alla v.s. Y. Si dirà che A è indipendente da Y qualora le s distribuzioni di frequenze delle mutabili statistiche condizionate $A|y_j$ siano tutte uguali tra loro.

▷ ESEMPIO 8.2

Con riferimento alla situazione di cui all'esempio 8.1, la verifica dell'indipendenza della m.s. A dalla v.s Y avviene verificando se le $s = 3$ distribuzioni di frequenze delle m.s. condizionate $A|y_j$ sono uguali tra loro. Dal momento che (cfr. figura 8.1):

$$A|y_1 \equiv \begin{Bmatrix} Motocicli & Autovetture & Autocarri \\ \frac{124}{413} & \frac{225}{413} & \frac{64}{413} \end{Bmatrix} = \begin{Bmatrix} M. & A. & C. \\ 0.30 & 0.55 & 0.15 \end{Bmatrix}$$

$$A|y_2 \equiv \begin{Bmatrix} Motocicli & Autovetture & Autocarri \\ \frac{44}{106} & \frac{55}{106} & \frac{7}{106} \end{Bmatrix} = \begin{Bmatrix} M. & A. & C. \\ 0.41 & 0.52 & 0.07 \end{Bmatrix}$$

$$A|y_3 \equiv \begin{Bmatrix} Motocicli & Autovetture & Autocarri \\ \frac{12}{51} & \frac{73}{51} & \frac{2}{51} \end{Bmatrix} = \begin{Bmatrix} M. & A. & C. \\ 0.23 & 0.73 & 0.04 \end{Bmatrix}$$

possiamo affermare che la m.s. A non è statisticamente indipendente dalla v.s. Y.
In conclusione, la variabile statistica mista (A, Y) ha *componenti statisticamente dipendenti.*

◁

I precedenti due esempi suggeriscono che l'indipendenza della v.s. Y rispetto alla m.s. A ha qualche legame con l'indipendenza della m.s. A rispetto alla v.s. Y. Tale intuizione è corroborata dalla seguente

Proprietà 8.1 Se la v.s. Y è indipendente dalla m.s. A, allora

$$n_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n} \tag{8.2}$$

e ciò $\forall i = 1, \ldots, r$ e $\forall j = 1, \ldots, s$.

◁

Dimostrazione: dall'uguaglianza posta in (8.1), $\forall i = 1, ..., r$, si ha $\delta_j = \dfrac{n_{ij}}{n_{i\cdot}}$, per cui $n_{i\cdot}\delta_j = n_{ij}$, che, sommando rispetto all'indice i, porge

$$\sum_{i=1}^{r} n_{i\cdot}\delta_j = \sum_{i=1}^{r} n_{ij} \qquad \rightarrow \qquad \delta_j\, n = n_{\cdot j}$$

Si ha la tesi sostituendo in quest'ultima l'espressione di δ_j data in (8.1), cioè

$$\frac{n_{ij}}{n_{i\cdot}}\, n = n_{\cdot j} \qquad \rightarrow \qquad n_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

□

La precedente proprietà ci consente di affermare che l'indipendenza di Y da A implica l'indipendenza di A da Y. Per dimostrare tale affermazione è sufficiente ricordare che A è indipendente da Y se e solo se $\forall j = 1, \ldots, s$:

$$A|y_j \equiv \begin{Bmatrix} a_1 & \ldots & a_i & \ldots & a_r \\ \delta_1 & \ldots & \delta_i & \ldots & \delta_r \end{Bmatrix}$$

con, $\forall i = 1, \ldots, r$:

$$\delta_i = \frac{n_{1j}}{n_{\cdot 1}} = \ldots = \frac{n_{ij}}{n_{\cdot j}} = \ldots = \frac{n_{rj}}{n_{\cdot s}}$$

Occorre dunque dimostrare che la frequenza $\delta_i = \frac{n_{ij}}{n_{\cdot j}}$ associata alla i-esima modalità della mutabile condizionata $A|y_j$ non dipende dall'indice j. Sfruttando l'equazione (8.2) si ha infatti:

$$\delta_i = \frac{n_{ij}}{n_{\cdot j}} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}\, \frac{1}{n_{\cdot j}} = \frac{n_{i\cdot}}{n}$$

Nel seguito, alla luce di quanto esposto, parleremo di variabili statistiche miste a componenti indipendenti qualora si sia verificata indifferentemente l'indipendenza di una componente rispetto all'altra.

▷ ESEMPIO 8.3

Un'indagine condotta su 200 possessori di carta di credito circa il *sesso* (A) ed il *tipo di carta* (B), ha dato luogo alla mutabile statistica doppia (A, B) con distribuzione di frequenze congiunte:

$A \downarrow B \rightarrow$	American Express	Visa
Maschio	48	72
Femmina	32	48

è immediato accertarsi che la m.s. doppia in esame ha componenti statisticamente indipendenti risultando (cfr. figura 8.2) uguali tra loro le distribuzioni di frequenze delle mutabili statistiche condizionate $A|b_j$ e $B|a_i$, con $i = 1, 2$ e $j = 1, 2$:

$$A|b_1 \equiv \begin{Bmatrix} Maschio & Femmina \\ 0.6 & 0.4 \end{Bmatrix} \quad A|b_2 \equiv \begin{Bmatrix} Maschio & Femmina \\ 0.6 & 0.4 \end{Bmatrix}$$

$$B|a_1 \equiv \begin{Bmatrix} Am.Express & Visa \\ 0.4 & 0.6 \end{Bmatrix} \quad B|a_2 \equiv \begin{Bmatrix} Am.Express & Visa \\ 0.4 & 0.6 \end{Bmatrix}$$

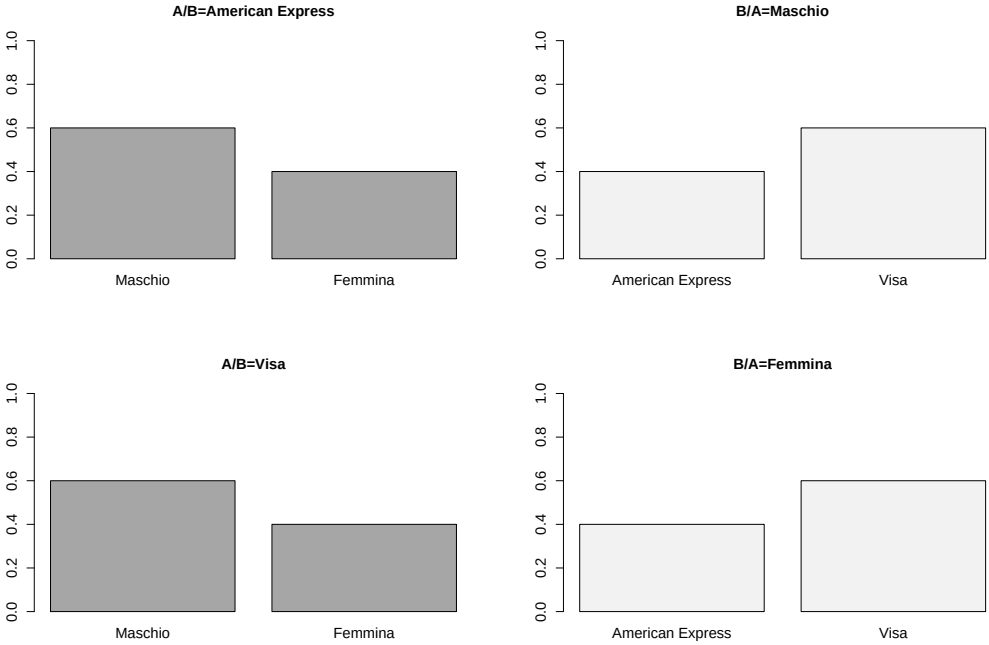


Figura 8.2 Distribuzioni di frequenze di $A|b_j$ e $B|a_i$, esempio 8.3.

Potremmo dunque affermare che la scelta tra i due tipi di carta di credito pare non influenzata dalla mutabile sesso ed anche che il sesso sembra non dipendere dalla carta di credito posseduta, anche se tale ultima affermazione non ha alcun senso!

Per curiosità verifichiamo empiricamente la proprietà (8.1), cioè che tutte le quattro frequenze congiunte osservate n_{ij} soddisfano l'uguaglianza posta in (8.2):

$$n_{11} = \frac{120 \cdot 80}{200} = 48 \qquad n_{12} = \frac{120 \cdot 120}{200} = 72$$
$$n_{21} = \frac{80 \cdot 80}{200} = 32 \qquad n_{22} = \frac{80 \cdot 120}{200} = 48$$

$\triangleleft$

La proprietà (8.1) risulta utile anche dal punto di vista operativo. Per verificare infatti l'indipendenza, piuttosto che confrontare le distribuzioni condizionate, potremmo verificare se vale l'uguaglianza (8.2) per tutte le $r \cdot s$ frequenze congiunte della distribuzione doppia.

$\triangleright$ ESEMPIO 8.4

Riprendendo in esame la variabile statistica mista (A, Y) di cui all'esempio (8.1), per la quale si propose la seguente distribuzione di frequenze congiunte:

$A \downarrow Y \rightarrow$	0	1	2	
Motocicli	124	44	12	180
Autovetture	225	55	37	...
Autocarri	64	7	2	...
	413	...	...	570

È immediato verificare che essa è a *componenti dipendenti*, infatti, considerando ad esempio le frequenze congiunte osservate n_{11} corrispondenti alla coppia di modalità (a_1, y_1), essendo $n_{1\cdot} = 180$, $n_{\cdot 1} = 413$ e $n = 570$, sfruttando la (8.2) abbiamo:

$$\frac{n_{1\cdot} \cdot n_{\cdot 1}}{n} = \frac{180 \cdot 413}{570} = 130.4211 \neq 124 = n_{11}$$

$\triangleleft$

OSSERVAZIONE: è appena il caso di notare che, qualora la distribuzione di frequenze congiunta di una variabile statistica mista presentasse una o più frequenze nulle in corrispondenza ad una particolare coppia di modalità (a_i, y_j), per essa l'uguaglianza (8.2) non risulterebbe soddisfatta e immediatamente si potrebbe affermare che la v.s. mista in questione ha componenti dipendenti.

$\star$

8.1.1 MISURE DELLA DIPENDENZA STATISTICA

Qualora le componenti una variabile statistica mista risultino tra loro dipendenti dal punto di vista statistico, è del tutto lecito chiedersi qual'è il grado di tale dipendenza. Dedichiamo, pertanto, questo paragrafo alla costruzione di un indice di dipendenza statistica.
A tal fine osserviamo che data una variabile statistica mista (A, Y) con distribuzione di frequenze congiunte

$$\{((a_i; y_j)\,; n_{ij})\}_{\substack{i=1,\dots,r \\ j=1,\dots,s}}$$

è sempre possibile, $\forall\, i = 1, \dots, r$ e $\forall\, j = 1, \dots, s$, calcolare le quantità

$$\hat{n}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n} \tag{8.3}$$

che vengono abitualmente dette *frequenze teoriche* o anche *frequenze attese*, nel senso che esse sarebbero le frequenze congiunte della v.s. mista nel caso in cui le due sue componenti A e Y fossero indipendenti, e ciò in accordo con la proprietà (8.1).
Esse danno luogo alla *distribuzione delle frequenze teoriche*

$$\{((a_i; y_j)\,; \hat{n}_{ij})\}_{\substack{i=1,\dots,r \\ j=1,\dots,s}}$$

che può essere posta nella consueta forma tabellare:

$A \downarrow Y \rightarrow$	y_1	$\cdots$	y_j	$\cdots$	y_s	
a_1	$\hat{n}_{11}$	$\cdots$	$\hat{n}_{1j}$	$\cdots$	$\hat{n}_{1s}$	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_i	$\hat{n}_{i1}$	$\cdots$	$\hat{n}_{ij}$	$\cdots$	$\hat{n}_{is}$	$n_{i\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
a_r	$\hat{n}_{r1}$	$\cdots$	$\hat{n}_{rj}$	$\cdots$	$\hat{n}_{rs}$	$n_{1\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot j}$	$\cdots$	$n_{\cdot s}$	$n_{\cdot\cdot}$

Le frequenze teoriche $\hat{n}_{ij}$ calcolate in accordo alla (8.3) sono tali da non modificare le distribuzioni di frequenze marginali di A e di Y, infatti:

$$\sum_{i=1}^{r} \hat{n}_{ij} = \sum_{i=1}^{r} \frac{n_{i\cdot} n_{\cdot j}}{n} = \frac{n_{\cdot j}}{n} \sum_{i=1}^{r} n_{i\cdot} = n_{\cdot j}$$

$$\sum_{j=1}^{s} \hat{n}_{ij} = \sum_{j=1}^{s} \frac{n_{i\cdot} n_{\cdot j}}{n} = \frac{n_{i\cdot}}{n} \sum_{j=1}^{s} n_{\cdot j} = n_{i\cdot}$$

né ovviamente la numerosità del collettivo statistico, infatti:

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\hat{n}_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{i\cdot}n_{\cdot j}}{n} = \frac{1}{n}\sum_{i=1}^{r}n_{i\cdot}\sum_{j=1}^{s}n_{\cdot j} = \frac{n\,n}{n} = n$$

In altre parole potremmo dire che applicando la (8.3) redistribuiamo le n unità del collettivo statistico, mantenendo inalterate le numerosità dei sottoinsiemi Ω_i e Ω_j generati, rispettivamente, dalle applicazioni A e Y, in modo da verificare la condizione di indipendenza statistica tra le componenti la v.s. mista (A, Y).

▷ ESEMPIO 8.5

Sempre con riferimento alla variabile statistica mista (A, Y) introdotta all'esempio (8.1), il calcolo delle frequenze teoriche n_{ij} in accordo alla (8.3), porge, come il Lettore può facilmente verificare, la seguente distribuzione congiunta:

$A \downarrow Y \rightarrow$	0	1	2	
Motocicli	130.4211	33.4737	16.1053	...
Autovetture	229.6860	58.9509	28.3632	317
Autocarri	52.8930	13.5754	6.5316	...
	413	...	...	570

per la quale, come detto, risulta, ad esempio:

$$\sum_{i=1}^{3}\hat{n}_{i1} = 130.4211 + 229.6860 + 52.8930 = 413 = n_{\cdot 1}$$

$$\sum_{j=1}^{3}\hat{n}_{2j} = 229.6860 + 58.9509 + 28.3632 = 317 = n_{2\cdot}$$

e ancora:

$$\sum_{i=1}^{3}\sum_{j=1}^{3}\hat{n}_{ij} = 130.4211 + 33.4737 + \ldots + 13.5754 + 6.5316 = 570 = n$$

Si noti che per ottenere la distribuzione di frequenze teoriche congiunte, sarà sufficiente calcolare, applicando la (8.3), solo $(r-1)(s-1)$ frequenze teoriche $\hat{n}_{ij}$, dal momento che le restanti possono essere ottenute per differenza.

◁

Una misura della discrepanza tra le frequenze osservate e le frequenze teoriche ci è offerta dal calcolo delle *contingenze*, definite, $\forall\, i = 1, \ldots, r$ e $\forall\, j = 1, \ldots, s$, quali differenza tra le frequenze osservate e le frequenze teoriche, in simboli:

$$c_{ij} = n_{ij} - \hat{n}_{ij} \tag{8.4}$$

per le quali si ha:

$$\sum_{i=1}^{r} c_{ij} = \sum_{i=1}^{r} (n_{ij} - \hat{n}_{ij}) = n_{\cdot j} - n_{\cdot j} = 0$$
$$\sum_{j=1}^{s} c_{ij} = \sum_{j=1}^{s} (n_{ij} - \hat{n}_{ij}) = n_{i\cdot} - n_{i\cdot} = 0$$

e di conseguenza $\sum_{i=1}^{r} \sum_{j=1}^{s} c_{ij} = 0$.

Evidentemente, in caso di indipendenza le frequenze osservate n_{ij} coincidono con le frequenze teoriche $\hat{n}_{ij}$ per cui segue che l'indipendenza tra le componenti la variabile statistica mista (A, Y) implica la nullità di tutte le contingenze c_{ij}.
A questo punto, per misurare il grado di dipendenza tra le componenti la variabile statistica mista (A, Y) appare del tutto lecito e naturale utilizzare le contingenze e tentare di quantificare la loro complessiva vicinanza allo zero.
Tuttavia, come si è visto, la somma delle contingenze è sempre nulla e, dunque, queste non possono essere impiegate in modo diretto per pervenire ad una misura del grado di dipendenza statistica. Un'alternativa è considerare le contingenze in valore assoluto oppure al quadrato. Seguendo tale ultimo approccio è abitudine rapportare il quadrato delle contingenze alla corrispondente frequenza teorica, cioè

$$\frac{c_{ij}^2}{\hat{n}_{ij}} = \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}}$$

sì da pervenire alla misura di dipendenza introdotta da K. Pearson e definita come segue.

Definizione 8.2 (Chi-quadrato di Pearson)
Data una variabile statistica mista (A, Y), si definisce chi-quadrato di Pearson la somma dei rapporti tra le contingenze al quadrato e le corrispondenti frequenze teoriche, cioè:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{c_{ij}^2}{\hat{n}_{ij}} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} \tag{8.5}$$

□

Osserviamo che al fine del calcolo di χ^2 non è necessario calcolare le contingenze c_{ij} e si possono così evitare errori di approssimazioni successive. La (8.5) può essere, infatti, scritta nella forma:

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \tag{8.6}$$

Per sincerarsene è sufficiente sviluppare il quadrato a numeratore della (8.5), infatti:

$$\begin{aligned}\chi^2 &= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{(n_{ij}-\hat{n}_{ij})^2}{\hat{n}_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2 - 2\,\hat{n}_{ij}\,n_{ij} + \hat{n}_{ij}^2}{\hat{n}_{ij}} = \\ &= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{\hat{n}_{ij}} - 2\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\hat{n}_{ij}\,n_{ij}}{\hat{n}_{ij}} + \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\hat{n}_{ij}^2}{\hat{n}_{ij}} = \\ &= \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{\hat{n}_{ij}} - n = \sum_{i=1}^{r}\sum_{j=1}^{s}n_{ij}^2\,\frac{n}{n_{i\cdot}\cdot n_{\cdot j}} - n = n\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}\cdot n_{\cdot j}} - n\end{aligned}$$

▷ ESEMPIO 8.6

Con riferimento, ancora, alla variabile statistica mista (A, Y) introdotta all'esempio (8.1), ci poponiamo di calcolare il valore dell'indice χ^2 ricorrendo alla (8.6). Riprendendo la distribuzione di frequenze congiunte della v.s. mista (A, Y):

$A \downarrow Y \rightarrow$	0	1	2	
Motocicli	124	44	12	180
Autovetture	225	55	37	317
Autocarri	64	7	2	73
	413	106	51	570

calcoliamo dapprima:

$$\begin{aligned}\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} &= \frac{124^2}{180\cdot 413} + \frac{44^2}{180\cdot 106} + \frac{12^2}{180\cdot 51} + \\ &+ \frac{225^2}{317\cdot 413} + \frac{55^2}{317\cdot 106} + \frac{37^2}{317\cdot 51} + \\ &+ \frac{64^2}{73\cdot 413} + \frac{7^2}{73\cdot 106} + \frac{2^2}{73\cdot 51} = 1.0287\end{aligned}$$

A questo punto sarà:

$$\chi^2 = 570\,(1.028639 - 1) = 16.359$$

◁

Nel caso di dipendenza, il chi-quadrato di Pearson, così come definito in (8.5), assume valori reali positivi e poco dice circa l'intesità del grado di dipendenza tra le componenti la variabile statistica mista in esame. Esso, inoltre, mal si presta qualora si debba confrontare la dipendenza tra le componenti di variabili statistiche rilevate su collettivi diversi.
Tuttavia, è possibile individuare una soglia superiore per il valore di χ^2, che viene riassunta nella seguente

Proprietà 8.2 Qualunque sia la variabile statistica mista in osservazione, si ha la maggiorazione

$$\chi^2 \leq min\,\{n(s-1); n(r-1)\}$$

(per la dimostrazione, cfr. paragrafo 8.3).

◁

Tale proprietà ci consente di definire, a partire dal χ^2, un indice di dipendenza normalizzato come segue:

Definizione 8.3 (Indice V di Cramér)
si definisce indice V di Cramér la radice quadrata del rapporto tra il χ^2 ed il suo valore massimo, cioè, in virtù della proprietà (8.2)

$$V = \sqrt{\frac{\chi^2}{max(\chi^2)}} = \sqrt{\frac{\chi^2}{min\,\{n(s-1); n(r-1)\}}} \tag{8.7}$$

□

Evidentemente, per le condizioni poste in essere, $0 \leq V \leq 1$.

▷ ESEMPIO 8.7

Con riferimento alla situazione descritta all'esempio (8.1), essendo $\chi^2 = 16.359$ e $max(\chi^2) = min\,\{570\,(3-1); 570\,(3-1)\} = 1140$, l'indice normalizzato V di Cramér risulta:

$$V = \sqrt{\frac{\chi^2}{max(\chi^2)}} = \sqrt{\frac{16.359}{1140}} = 0.11979$$

mostrando una debole dipendenza tra le componenti la variabile statistica mista in esame.

◁

▷ ESEMPIO 8.8

Si immagini di avere rilevato, su un collettivo formato da 725 pratiche di liquidazione sinistri, i caratteri *agenzia assicuratrice di provenienza* e *giudizio di conformità della pratica* e che tale indagine abbia dato luogo alla mutabile statistica doppia (A, B) con distribuzione di frequenze congiunte:

Giudizio $\rightarrow$ *Agenzia* $\downarrow$	Conforme	Non conforme
Ag. A	180	26
Ag. B	144	65
Ag. C	205	105

Desiderando indagare circa l'indipendenza statistica tra le m.s. $A =$ *Agenzia* e $B =$ *Giudizio*, è sufficiente verificare se, per qualsiasi indice $i = 1, 2, 3$ e $j = 1, 2, 3$, vale la relazione $n_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$. Dal momento che:

$$\frac{n_{1\cdot}\, n_{\cdot 1}}{n} = \frac{206 \cdot 529}{725} = 150.31 \neq 180 = n_{11}$$

possiamo affermare che la mutabile statistica doppia presenta componenti dipendenti. Stando così le cose, calcoliamo il grado di dipendenza tra A e B ricorrendo all'indice normalizzato di Cramér. Osservando che applicando la (8.6):

$$\chi^2 = n\left(\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) = 725\left(\frac{180^2}{206 \cdot 529} + \frac{26^2}{206 \cdot 196} + \right.$$
$$\left. + \frac{144^2}{209 \cdot 529} + \frac{65^2}{206 \cdot 196} + \frac{205^2}{310 \cdot 529} + \frac{65^2}{310 \cdot 196} - 1\right) = 30.7910$$

e che $max(\chi^2) = min\left\{725\,(3-1); 725\,(2-1)\right\} = 725$, sarà:

$$V = \sqrt{\frac{\chi^2}{max(\chi^2)}} = \sqrt{\frac{30.791}{725}} = 0.2061$$

Anche in questo caso la dipendenza tra le componenti la mutabile statistica bivariata (A, B) è piuttosto debole.

◁

Nel caso si operi su una variabile statistica bivariata (X, Y) per la quali l'insieme delle modalità osservate è costituito da un elevato numero di coppie di modalità distinte, la verifica dell'indipendenza statistica tra le sue componenti e l'eventuale calcolo degli indici χ^2 e V in base alle informazioni contenute nella distribuzione di frequenze congiunte, implica il raccoglimento dei dati in classi di modalità.
Cruciale sarà la scelta del numero e dell'ampiezza delle classi per ciascuna componente la v.s. bivariata, poiché può produrre distorsioni, più o meno marcate, nelle conclusioni dell'analisi.

▷ ESEMPIO 8.9

Un'indagine condotta su 192 single a proposito dell'ammontare della spesa mensile per generi alimentari e per abbigliamento, ha dato luogo alla variabile statistica bivariata $(X, Y) =$ *spesa per generi alimentari, spesa per abbigliamento* con distribuzione di frequenze assolute congiunte:

$X \downarrow \; Y \rightarrow$	$100 \dashv 200$	$200 \dashv 300$	$300 \dashv 400$
$100 \dashv 200$	24	18	12
$200 \dashv 300$	16	26	34
$300 \dashv 400$	4	14	44

Se osserviamo, ad esempio, che:

$$n_{11} = 24 \neq 12.375 = \frac{54 \cdot 44}{192} = \frac{n_{1\cdot} \cdot n_{\cdot 1}}{n}$$

possiamo senz'altro affermare che la v.s. doppia (X, Y) ha componenti statisticamente dipendenti. La diversità delle v.s. condizionate $Y|x_i$ è evidenziata dai primi tre istogrammi di figura (8.3). L'ultimo istogramma riportato in figura è quello della v.s. Y, che appare come "mistura" delle distribuzioni condizionate.
Ai fini del calcolo dell'indice normalizzato V di Cramér, ricorrendo alla (8.6), abbiamo innanzitutto:

$$\begin{aligned}\chi^2 &= n\left(\sum_{i=1}^{3}\sum_{j=1}^{3}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) = \\ &= 45\left(\frac{24^2}{54\cdot 44} + \frac{18^2}{54\cdot 58} + \frac{12^2}{54\cdot 58} + \frac{16^2}{76\cdot 44} + \frac{26^2}{76\cdot 58} + \frac{34^2}{76\cdot 58} + \right. \\ &\left. + \frac{4^2}{62\cdot 44} + \frac{14^2}{62\cdot 58} + \frac{44^2}{62\cdot 58} - 1\right) = 34.8949\end{aligned}$$

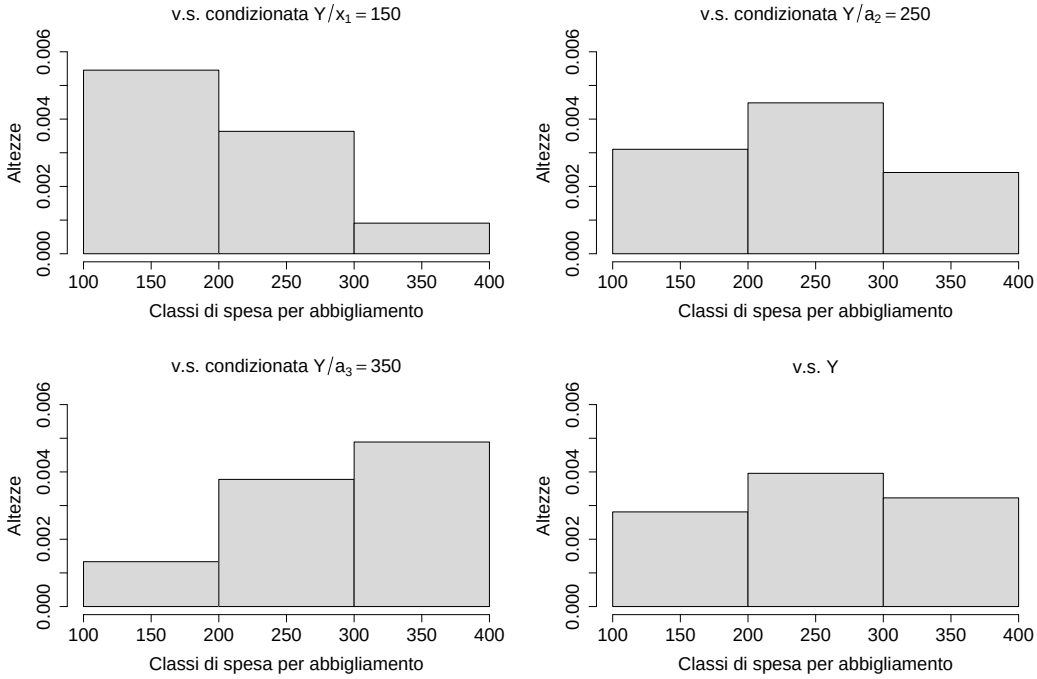


Figura 8.3 Istogrammi delle v.s. condizionate $Y|x_i$ e della v.s. Y, esempio 8.9.

Risultando poi $max(\chi^2) = 384$, si ha:

$$V = \sqrt{\frac{\chi^2}{max(\chi^2)}} = \sqrt{\frac{34.8949}{384}} = 0.30145$$

In questo caso il grado di dipendenza statistica tra le componenti la variabile bivariata (X, Y) è piuttosto alta.
Si tenga ben presente che il raccoglimento dei dati individuali in classi, agendo sulla distribuzione marginale e quindi sulle frequenze congiunte, può produrre distorsioni, più o meno marcate, nelle conclusioni circa l'intensità della dipendenza tra le componenti della v.s. doppia in esame.
A tal proposito, si immagini di accorpare le prime due classi della v.s. Y nella sola classe $100 \dashv 300$. La distribuzione di frequenze congiunte diverrebbe:

$X \downarrow \; Y \rightarrow$	$100 \dashv 300$	$300 \dashv 400$
$100 \dashv 200$	42	12
$200 \dashv 300$	42	34
$300 \dashv 400$	18	44

Per essa, come il Lettore può verificare numericamente, si avvrebbe $\chi^2 = 27.7705$ e di conseguenza $V = 0.38031$.

◁

Nel capitolo precedente, a proposito di una variabile statistica bivariata (X, Y), è stato introdotto il concetto di covarianza quale misura della variabilità congiunta tra le sue componenti X e Y. Sebbene su tale argomento torneremo più oltre allorché affronteremo il problema della regressione lineare, è importante qui sottolineare la relazione esistente tra i concetti di *indipendenza statistica* e *covarianza*, riassunta dalla seguente proprietà.

Proprietà 8.3 Se (X, Y) è una variabile statistica a componenti statisticamente indipendenti la covarianza tra X e Y è nulla.

◁

Dimostrazione: ricordando che la covarianza tra X e Y può essere nella forma

$$Cov[X, Y] = E[X \cdot Y] - E[X] \cdot E[Y]$$

e che nel caso di indipendenza statistica, per qualsiasi indice $i = 1, \ldots, r$ e $j = 1, \ldots, s$, si ha

$$n_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

allora

$$E[X \cdot Y] = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} x_i \, y_j \, n_{ij} = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} x_i \, y_j \, \frac{n_{i\cdot} n_{\cdot j}}{n} =$$

$$= \frac{1}{n} \sum_{i=1}^{r} x_i \, n_{i\cdot} \, \frac{1}{n} \sum_{j=1}^{s} y_j \, n_{\cdot j} = E[X] \cdot E[Y]$$

per cui la tesi $Cov[X, Y] = E[X] \cdot E[Y] - E[X] \cdot E[Y] = 0$.

□

Si osservi, tuttavia, che *non vale la proposizione inversa*; in altri termini se per le componenti la v.s. doppia (X, Y) risulta $Cov[X, Y] = 0$, non necessariamente X e Y sono indipendenti dal punto di vista statistico.

▷ ESEMPIO 8.10

Desiderando illustrare quanto detto a proposito della relazione tra i concetti di indipendenza statistica e di covarianza, così come riassunto dalla proprietà (8.3), si considerino le due situazioni:

⋆ *caso (a)*: sia (X, Y) una la v.s. bivariata con distribuzione di frequenze congiunte:

$X \downarrow Y \rightarrow$	$y_1 = 1$	$y_2 = 2$	$y_3 = 3$
$x_1 = 1$	0	10	0
$x_2 = 2$	10	0	10

Manifestamente essa ha componenti statisticamente dipendenti, infatti alcune frequenze congiunte sono nulle. Per essa, inoltre si ha:

$$\chi^2 = 30 \left(\frac{10^2}{10 \cdot 10} + \frac{10^2}{10 \cdot 20} + \frac{10^2}{10 \cdot 20} - 1 \right) = 30$$

Quanto alla covarianza, risultando $E[X] = 5/3$ e $E[Y] = 2$, abbiamo:

$$Cov[X, Y] = \frac{1}{30} (1 \cdot 2 \cdot 10 + 2 \cdot 1 \cdot 10 + 2 \cdot 3 \cdot 10) - \frac{10}{3} = 0$$

⋆ *caso (b)*: sia (X, Y) una la v.s. bivariata con distribuzione di frequenze congiunte:

$X \downarrow Y \rightarrow$	$y_1 = 1$	$2_2 = 2$	$y_3 = 3$
$x_1 = 1$	5	5	5
$x_2 = 2$	10	10	10

Manifestamente essa ha componenti statisticamente indipendenti, infatti le distribuzioni delle v.s. condizionate sono uguali. Pertanto $\chi^2 = 0$.
Quanto alla covarianza, risultando $E[X] = 5/3$ e $E[Y] = 2$, abbiamo:

$$\begin{aligned} Cov[X, Y] = \frac{1}{45} (&1 \cdot 1 \cdot 5 + 1 \cdot 2 \cdot 5 + 1 \cdot 3 \cdot 5 + \\ &+ 2 \cdot 1 \cdot 10 + 2 \cdot 2 \cdot 10 + 2 \cdot 3 \cdot 10) - \frac{10}{3} = \frac{10}{3} - \frac{10}{3} = 0 \end{aligned}$$

La figura (8.4) riporta i diagrammi a bolle delle due distribuzioni di frequenze congiunte. Nonostante in entrambe le situazioni contemplate si abbia un valore nullo della covarianza, nel primo caso possiamo individuare un legame tra le due componenti, mentre nel secondo non emerge alcun legame tra le componenti.

◁

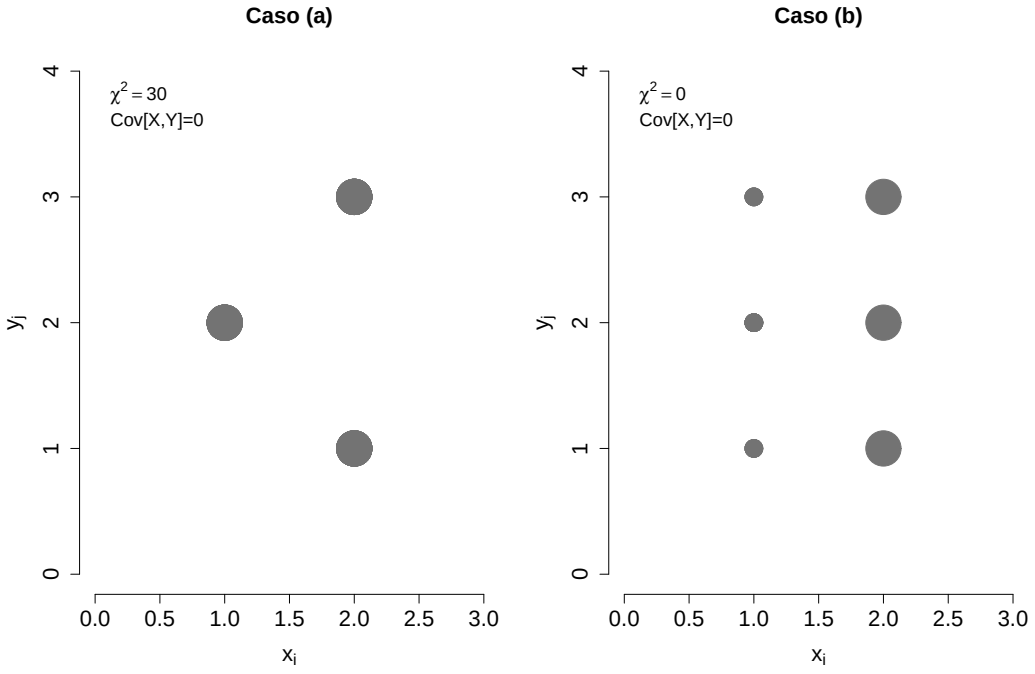


Figura 8.4 Diagrammi a bolle, esempio 8.10.

8.2. INDIPENDENZA IN MEDIA

Quando si studia una variabile statistica mista oppure una variabile statistica bivariata, è possibile indagare circa un ulteriore tipo di indipendenza, verificare cioè se è soddisfatta la seguente:

Definizione 8.4 (Indipendenza in media di Y dalla m.s. A)
data una variabile statistica mista (A, Y) diremo che Y è indipendente in media da A se le r medie delle v.s. condizionate $Y|a_i$ sono uguali tra loro, cioè se

$$E[Y|a_1] = \ldots = E[Y|a_i] = \ldots = E[Y|a_r] \tag{8.8}$$

□

L'indipendenza in media è senza dubbio una condizione meno restrittiva rispetto all'indipendenza statistica, se quest'ultima comporta l'uguaglianza delle distribuzioni condizionate l'indipendenza in media richiede più semplicemente l'uguaglianza delle medie di tali

distribuzioni. Evidentemente l'indipendenza statistica implica l'indipendenza in media, infatti, se le v.s $Y|a_i$ hanno distribuzioni uguali necessariamente sono uguali anche le loro medie, non è vero tuttavia il viceversa.
Può accadere che si presentino situazioni del tipo di quelle evidenziate in figura (8.5) nella quale vediamo che le forme degli istogrammi sono differenti ma le tre distribuzioni hanno medie coincidenti, per cui risulta verificata l'indipendenza in media e non quella statistica.

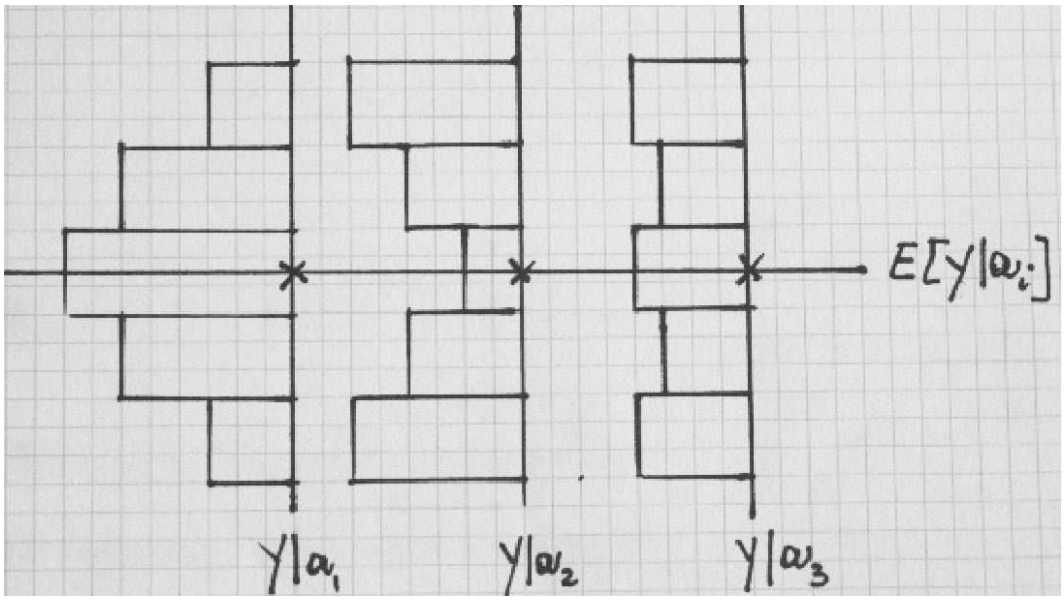


Figura 8.5 Variabili condizionate con medie uguali e distribuzioni differenti.

▷ ESEMPIO 8.11

Da un'indagine condotta su 175 utenti di tre società di telefonia mobile circa il numero di chiamate mensili al centro servizi clienti, risulta che la variabile statistica mista $(A, Y) = \{$*società, # di chiamate*$\}$, ha la seguente distribuzione di frequenze assolute congiunte:

$A \downarrow Y \rightarrow$	0	1	2
Vodafone	10	20	10
Tim	20	5	20
Wind	30	30	30

Dal momento che, ad esempio:

$$\frac{n_{1\cdot}\, n_{\cdot 1}}{n} = \frac{40 \cdot 60}{175} = 13.714 \neq 10 = n_{11}$$

possiamo affermare che le componenti A e Y della v.s. mista in esame sono tra loro statisticamente dipendenti. Si lascia al Lettore la verifica che $\chi^2 = 15.1726$ e $V = 0.20821$ nonché la rappresentazione grafica delle distribuzioni delle variabili condizionate.
Il fatto che Y sia una variabile statistica, ci autorizza ad indagare circa l'eventuale sua dipendenza in media da A. A tal fine, in accordo alla condizione posta dalla (8.8), occorre calcolare e confrontare tra loro i valori medi delle tre variabili condizionate $Y|a_i$.
Facilmente abbiamo:

$$E[Y|a_1] = \frac{10 \cdot 0 + 20 \cdot 1 + 10 \cdot 2}{40} = 1$$
$$E[Y|a_2] = \frac{20 \cdot 0 + 5 \cdot 1 + 20 \cdot 2}{45} = 1$$
$$E[Y|a_3] = \frac{30 \cdot 0 + 30 \cdot 1 + 30 \cdot 2}{90} = 1$$

Manifestamente la condizione (8.8) è soddisfatta e pertanto possiamo affermare che la variabile statistica Y è indipendente in media da A.
Per inciso, si osservi che il valor medio di Y è anch'esso pari a 1, infatti:

$$E[Y] = \frac{20 \cdot 1 + 10 \cdot 2 + 5 \cdot 1 + 20 \cdot 2 + 30 \cdot 1 + 30 \cdot 2}{175} = 1$$

◁

▷ ESEMPIO 8.12

Un'indagine condotta su 220 individui circa la *condizione professionale* (A) e l'ammontare della *spesa annua per spettacoli culturali* (Y), ha dato luogo alla variabile statistica mista (A, Y) con distribuzione di frequenze assolute congiunte:

$A \downarrow Y \rightarrow$	$0 \dashv 50$	$50 \dashv 100$	$100 \dashv 150$	$150 \dashv 200$
Operaio	44	28	3	5
Impiegato	24	16	14	6
Quadro	8	12	21	9
Dirigente	4	4	12	10

Se si osserva, ad esempio, che:

$$\frac{n_{1\cdot}\, n_{\cdot 1}}{n} = \frac{80 \cdot 80}{220} = 29.091 \neq 44 = n_{11}$$

siamo in grado di affermare che le componenti A e Y della variabile statistica mista in esame non sono tra loro indipendenti. In figura (8.6) sono riportati gli istogrammi delle distribuzioni di frequenze delle v.s. condizionate $Y|a_i$. Lasciamo al Lettore la verifica numerica che $\chi^2 = 59.8424$ e $V = 0.30112$.

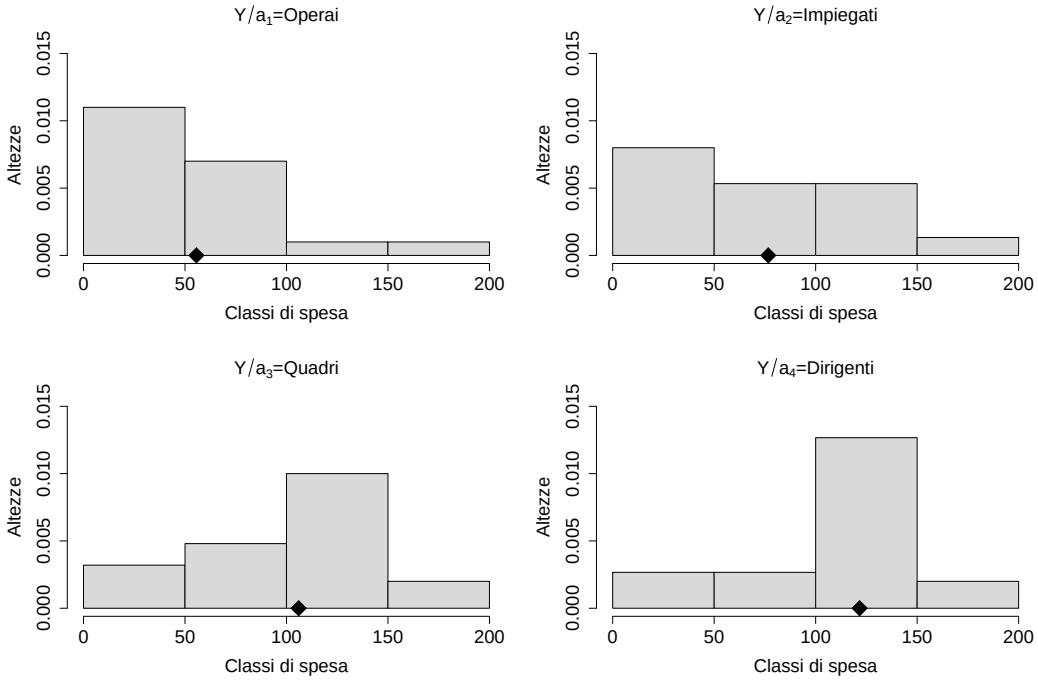


Figura 8.6 Istogrammi delle v.s. condizionate $Y|a_i$, esempio 8.12.

Data la natura di Y, ci pare del tutto lecito indagare circa la sua eventuale dipendenza in media da A. A tal fine, in accordo alla condizione posta dalla (8.8), occorre calcolare e confrontare tra loro i valori medi delle variabili condizionate $Y|a_i$.

Arricchita la precedente tabella con le distribuzioni marginali di A e individuati i centri di classe che rappresenteranno le modalità di Y:

$A \downarrow Y \rightarrow$	$0 \dashv 50$ $y_1 = 25$	$50 \dashv 100$ $y_2 = 75$	$100 \dashv 150$ $y_3 = 125$	$150 \dashv 200$ $y_4 = 175$	
Operaio	44	28	3	5	80
Impiegato	24	16	14	6	60
Quadro	8	12	21	9	50
Dirigente	4	4	12	10	30

siamo in grado di ottenere i valori medi delle v.s. $Y|a_i$:

$$E[Y|a_1] = \frac{44 \cdot 25 + 28 \cdot 75 + 3 \cdot 125 + 5 \cdot 175}{80} = 55.625$$
$$E[Y|a_2] = \frac{24 \cdot 25 + 16 \cdot 75 + 14 \cdot 125 + 6 \cdot 175}{60} = 76.667$$
$$E[Y|a_3] = \frac{8 \cdot 25 + 12 \cdot 75 + 21 \cdot 125 + 9 \cdot 175}{50} = 106.000$$
$$E[Y|a_4] = \frac{4 \cdot 25 + 4 \cdot 75 + 12 \cdot 125 + 10 \cdot 175}{30} = 121.667$$

Palesemente la condizione (8.8) non è soddisfatta come si evince anche dalla figura (8.6) e, pertanto, possiamo affermare che la variabile statistica Y è dipendente in media da A.

$\triangleleft$

8.2.1 MISURE DELLA DIPENDENZA IN MEDIA

Qualora la variabile statistica Y non fosse indipendente in media dalla mutabile statistica A, potrebbe essere interessante misurare il grado della sua dipendenza in media da A.
Al fine di pervenire ad un indice di dipendenza in media, appare del tutto ovvio prendere le mosse dalle medie condizionate $E[Y|a_i]$ e misurarne la variabilità. Infatti, in base alla (8.8) nel caso di indipendenza in media di Y da A, la variabilità delle medie condizionate sarebbe nulla.
Se consideriamo le r variabili statistiche condizionate $Y|a_i$, per ciascuna di esse, come già si è detto, è possibile calcolarne valor medio $E[Y|a_i]$ e varianza $V[Y|a_i]$. Tali misure, al variare delle modalità di A, descrivono altrettante variabili statistiche dette brevemente *medie condizionate* e *varianze condizionate* ed indicate, rispettivamente, con $E_{Y|A}$ e $V_{Y|A}$. In particolare, poi:

- $\star$ la v.s. *medie condizionate* ha distribuzione di frequenze:

$$E_{Y|A} \equiv \left\{ \begin{matrix} E[Y|a_i] \\ n_{i\cdot} \end{matrix} \right\}_{i=1,\ldots,r}$$

Per essa appare del tutto evidente calcolarne valor medio e varianza. A tal propostio abbiamo:

$$E[E_{Y|A}] = \frac{1}{n}\sum_{i=1}^{r} E\left[Y|a_i\right]\, n_{i\cdot} = \frac{1}{n}\sum_{i=1}^{r}\frac{1}{n_{i\cdot}}\sum_{j=1}^{s} y_j n_{ij}\, n_{i\cdot} =$$
$$= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s} y_j\, n_{ij} = E[Y] \tag{8.9}$$

il che significa che il valor medio delle r medie delle variabili statistiche condizionate $Y|a_i$ viene a coincidere con la media della v.s. Y.

Forti di tale risultato, osserviamo che se è verificata la condizione di indipendenza in media, allora la condizione posta in (8.8) implica,$\forall i = 1, \ldots, r$, l'uguaglianza $E[Y|a_i] = E[Y]$.

Quanto alla varianza di $E_{Y|A}$, risulta:

$$V[E_{Y|A}] = \frac{1}{n}\sum_{i=1}^{r}(E[Y|a_i] - E[Y])^2\, n_{i\cdot} \tag{8.10}$$

Evidentemente se la v.s. Y risultasse indipendente in media dalla m.s. A, allora la varianza di $E_{Y|A}$ sarebbe nulla.

$\star$ la v.s. *varianze condizionate* ha distribuzione di frequenze:

$$V_{Y|A} \equiv \left\{\begin{matrix} V[Y|a_i] \\ n_{i\cdot} \end{matrix}\right\}_{i=1,\ldots,r}$$

e per essa ci limitiamo a calcolarne il valor medio, che risulta:

$$E[V_{Y|A}] = \frac{1}{n}\sum_{i=1}^{r} V\left[Y|a_i\right]\, n_{i\cdot} = \frac{1}{n}\sum_{i=1}^{r}\frac{1}{n_{i\cdot}}\sum_{j=1}^{s}(y_j - E[Y|a_i])^2 n_{ij}\, n_{i\cdot} =$$
$$= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}(y_j - E[Y|a_i])^2 n_{ij} \tag{8.11}$$

Si noti che in generale la media della varianza condizionata non è uguale alla varianza di Y, risultando, come vedremo tra poco, $E[V_{Y|A}] \leq V[Y]$.

Da quanto esposto, risulta evidente che una misura *assoluta* del grado di dipendenza in media di Y rispetto ad A ci è offerta dalla varianza delle medie condizionate, $V[E_{Y|A}]$, che, come si disse, è nulla solo nel caso di indipendenza in media.

▷ ESEMPIO 8.13

Riprendendo la situazione di cui all'esempio 8.12, è immediato accertarsi che la distribuzione della v.s. medie condizionate risulta:

$$E_{Y|A} \equiv \left\{ \begin{matrix} E[Y|a_i] \\ n_{i\cdot} \end{matrix} \right\}_{i=1,\ldots,4} = \left\{ \begin{matrix} 55.625 & 76.667 & 106.000 & 121.667 \\ 80 & 60 & 50 & 30 \end{matrix} \right\}$$

Per tale variabile statistica si ha valor medio:

$$\begin{aligned} E[E_{Y|A}] &= \frac{1}{n} \sum_{i=1}^{4} E\left[Y|a_i\right] \, n_{i\cdot} = \\ &= \frac{(55.625 \cdot 80 + 76.667 \cdot 60 + 106.000 \cdot 50 + 121.667 \cdot 30)}{220} = \\ &= 81.818 = E[Y] \end{aligned}$$

Quanto alla varianza di $E_{Y|A}$, questa può essere calcolata come di consueto cioè $V[E_{Y|A}] = E[E^2_{Y|A}] - \left(E[E_{Y|A}]\right)^2$. Pertanto, risultando:

$$\begin{aligned} E[E^2_{Y|A}] &= \frac{1}{n} \sum_{i=1}^{4} E\left[Y|a_i\right]^2 \, n_{i\cdot} = \\ &= \frac{\left(55.625^2 \cdot 80 + \ldots + 121.667^2 \cdot 30\right)}{220} = 7300.37 \end{aligned}$$

sarà $V[E_{Y|A}] = 7300.370 - 6694.185 = 606.185$.

◁

Analogamente all'indice di dipendenza statistica χ^2, la varianza delle medie condizionate poco dice circa l'intesità del legame di dipendenza in media. È possibile giungere ad un *indice normalizzato*, in grado quindi di misurare l'intensità della dipendenza in media, semplicemente sfruttando la seguente proprietà della "varianza totale" della v.s. Y, detta *di scissione della varianza*.

Proprietà 8.4 Data la variabile statistica mista (A, Y), la varianza della sua componente Y è data dalla somma del valor medio delle varianze condizionate più la varianza delle medie condizionate, in simboli vale cioè l'uguaglianza:

$$V[Y] = E[V_{Y|A}] + V[E_{Y|A}] \tag{8.12}$$

(per la dimostrazione, cfr. paragrafo 8.3).

◁

A commento osserviamo che:

- $\star$ il valor medio delle varianze condizionate è uguale alla varianza di Y se e solo se Y è indipendente in media da A, in tal caso infatti risulta $V[E_{Y|A}] = 0$ e dunque solo in questo caso $E[V_{Y|A}] = V[Y]$;
- $\star$ la varianza delle medie condizionate non può essere maggiore della varianza totale di Y, cioè $V[E_{Y|A}] \leq V[Y]$.

Tale ultima osservazione ci permette di costruire un indice normalizzato di dipendenza in media, detto *rapporto di correlazione eta quadro di Pearson*, definito come segue:

Definizione 8.5 (Rapporto di correlazione η^2 di Pearson)
data una variabile mista (A, Y) il rapporto tra la varianza delle r medie delle variabili statistiche condizionate $Y|a_i$ e la varianza totale della variabile statistica Y viene detto rapporto di correlazione Eta quadro di Pearson, in simboli:

$$\eta^2_{Y|A} = \frac{V[E_{Y|A}]}{V[Y]} \tag{8.13}$$

□

▷ ESEMPIO 8.14

Con riferimento alla variabile statistica (A, Y) introdotta all'esempio 8.12, per la quale $V[E_{Y|A}] = 606.185$ (cfr. esempio 8.13) e, come il Lettore può facilmente verificare, $V[Y] = 2794.421$, il rapporto di correlazione η^2 di Pearson risulta:

$$\eta^2_{Y|A} = \frac{V[E_{Y|A}]}{V[Y]} = \frac{606.185}{2794.421} = 0.217$$

il che denota una debole dipendenza in media di Y da A.

◁

Ricordando la proprietà della scissione della varianza è possibile verificare che il rapporto di correlazione di Pearson può essere posto nella forma:

$$\eta^2_{Y|A} = 1 - \frac{E[V_{Y|A}]}{V[Y]} \tag{8.14}$$

Inoltre, è immediato verificare che $0 \leq \eta^2_{Y|A} \leq 1$, infatti:

- $\star$ $\eta^2_{Y|A} > 0$ perché rapporto fra due quantità sicuramente positive ed $\eta^2_{Y|A} = 0$ quando le medie condizionate hanno varianza nulla,

- $\star$ $\eta^2_{Y|A} \leq 1$ perché il numeratore è minore, o al più uguale, al denominatore.

Osserviamo, infine, che nel caso di indipendenza in media $\eta^2_{Y|A} = 0$, mentre la situazione di massima dipendenza in media si ha quando $\eta^2_{Y|A} = 1$, ciò comporta infatti che la variabilità della v.s. Y sia interamente dovuta alla variabilità delle medie condizionate dovendo essere $E[V_{Y|A}] = 0$ e di conseguenza $V[E_{Y|A}] = V[Y]$.

È interessante notare che nel caso si operi su una variabile statistica bivariata (X, Y), a componenti statisticamente dipendenti, è possibile indagare, sulla base della distribuzione di frequenze congiunte, circa l'eventuale dipendenza in media di Y da X e parimenti di X da Y.
Evidentemente, ricordando la definizione (8.4), in tale situazione diremo che:

- $\star$ Y è indipendente in media da X se risulta soddisfatta, per qualsiasi $i = 1, \ldots, r$, la condizione

$$E[Y|x_1] = \ldots = E[Y|x_i] = \ldots = E[Y|x_r] \tag{8.15}$$

- $\star$ X è indipendente in media da Y se risulta soddisfatta, per qualsiasi $j = 1, \ldots, s$, la condizione

$$E[X|y_1] = \ldots = E[X|y_j] = \ldots = E[X|y_s] \tag{8.16}$$

Manifestamente, se entrambe le condizioni (8.15) e (8.16) non fossero verificate, potremmo misurare l'intensità del grado di dipendenza in media di Y da X e di X da Y ricorrendo, rispettivamente, al rapporto di correlazione η^2 di Pearson:

$$\eta^2_{Y|X} = \frac{V[E_{Y|X}]}{V[Y]} \qquad \eta^2_{X|Y} = \frac{V[E_{X|Y}]}{V[X]}$$

Va da sè che i due rapporti di correlazione non necessariamente coincidono, così come la dipendenza in media di Y da X non esclude che X sia indipendente in media da Y.

È bene tenere a mente che nel caso si operi su due variabili statistiche, siano esse discrete o continue, la costruzione della distribuzione di frequenze congiunte non è sempre immediata e spesso ci si trova a dover raccogliere i dati individuali in classi. Tale compromesso è dovuto non solo all'esigenza di sintetizzare l'insieme dei dati individuali della v.s. (X, Y) bensì a quella di poter disporre delle v.s. condizionate $Y|x_i$ o $X|y_j$.

▷ ESEMPIO 8.15

La rilevazione della *statura* (in cm) e del *peso corporeo* (in kg) su un gruppo di 100 individui, ha dato luogo all'insieme di 100 coppie di dati individuali $(\tilde{x}_\alpha, \tilde{y}_\alpha)$, il cui diagramma a dispersione è proposto in figura (8.8, pannello a).

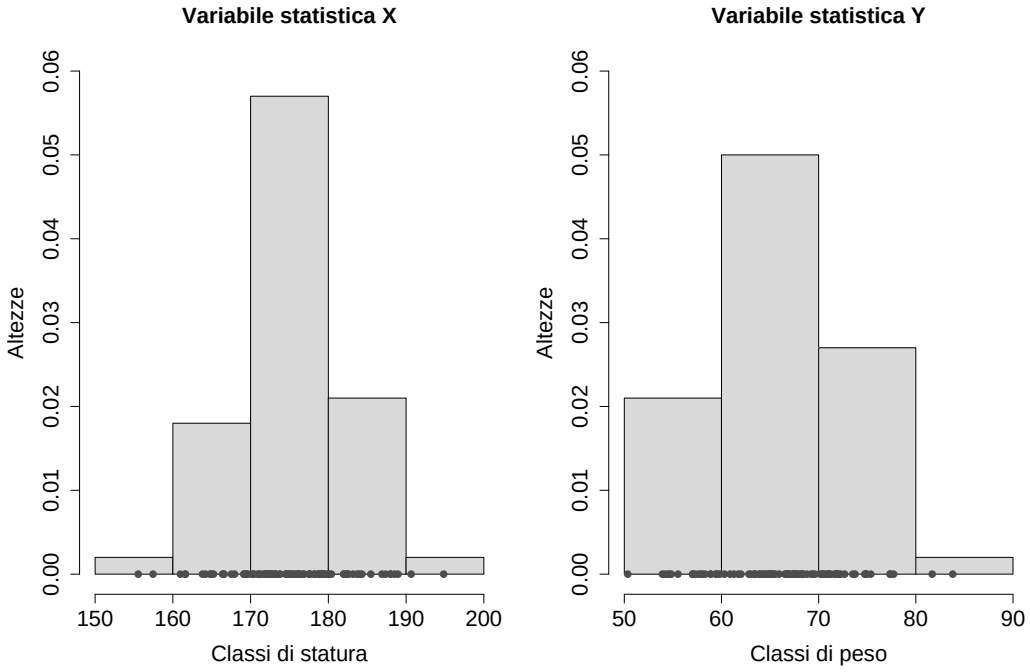


Figura 8.7 Distribuzioni di frequenze marginali di X e Y, esempio 8.15.

Desiderando studiare la dipendenza tra le v.s. $X =$ *statura* e $Y =$ *peso corporeo*, introduciamo la variabile statistica bivariata (X, Y) dopo aver raccolto, per ciascuna sua componente, i dati individuali in classi, sì che le distribuzioni marginali di X e Y presentino l'aspetto di figura (8.7). Consci che tale modo di procedere è a scapito di una perdita di informazione, otteniamo la seguente distribuzione di frequenze congiunte:

$X \downarrow Y \rightarrow$	$50 \dashv 60$ $(y_1 = 55)$	$60 \dashv 70$ $(y_2 = 65)$	$70 \dashv 80$ $(y_3 = 75)$	$80 \dashv 90$ $(y_4 = 85)$
$150 \dashv 160$ $(x_1 = 155)$	2	0	0	0
$160 \dashv 170$ $(x_2 = 165)$	9	8	1	0
$170 \dashv 180$ $(x_3 = 175)$	9	34	14	0
$180 \dashv 190$ $(x_4 = 185)$	1	8	10	2
$190 \dashv 200$ $(x_5 = 195)$	0	0	2	0

la cui rappresentazione grafica è proposta in figura (8.8, pannello b). Si noti che:

$$E[Y] = 66.00 \qquad V[Y] = 55.00 \qquad E[X] = 175.30 \qquad V[X] = 54.91$$

Manifestamente le componenti la v.s. doppia (X, Y) non sono statisticamente indipendenti, e ciò per la presenza di frequenze congiunte nulle, ad esempio $n_{13} = 0$. Come il Lettore può facilmente verificare per via numerica, si ha $\chi^2 = 39.223$ e $V = 0.313$.
Tale considerazione ci permette di affermare che esiste, peraltro cosa del tutto ovvia, un legame direttamente proporzionale tra il peso corporeo e la statura degli individui osservati.
Una lettura del diagramma a dispersione di figura (8.8, pannello b), suggerisce l'esistenza di una dipendenza in media di Y da X, ed infatti le medie condizionate $E[Y|x_i]$ risultano:

$$E[Y|x_1 = 155] = 55.000 \qquad E[Y|x_2 = 165] = 60.556$$
$$E[Y|x_3 = 175] = 65.878 \qquad E[Y|x_4 = 185] = 71.190$$
$$E[Y|x_5 = 195] = 75.000$$

Dal momento, poi, che $V[E_{Y|X}] = 15.042$, si avrà $\eta^2_{Y|X} = 0.2735$, il che conferma l'esistenza di una discreta dipendenza in media di Y da X.
In modo del tutto analogo, sempre in base al diagramma a dispersione di figura (8.8, pannello b), intuiamo l'esistenza di una dipendenza in media di X da Y, ed infatti:

$$E[X|y_1 = 55] = 169.286 \qquad E[X|y_2 = 65] = 175.000$$
$$E[X|y_3 = 75] = 179.815 \qquad E[X|y_4 = 85] = 185.000$$

Essendo $V[E_{X|Y}] = 15.026$, si avrà $\eta^2_{X|Y} = 0.2737$, risultato che conferma la discreta dipendenza in media tra la statura e il peso corporeo. Ovviamente da un punto di vista pratico tale relazione di dipendenza ha scarso interesse. Si noti che la discrepanza tra $\eta^2_{Y|X}$ e $\eta^2_{X|Y}$ tende a diminuire se le medie condizionate tendono ad allinearsi lungo una retta. La figura (8.9) riporta i diagrammi a bolle delle medie

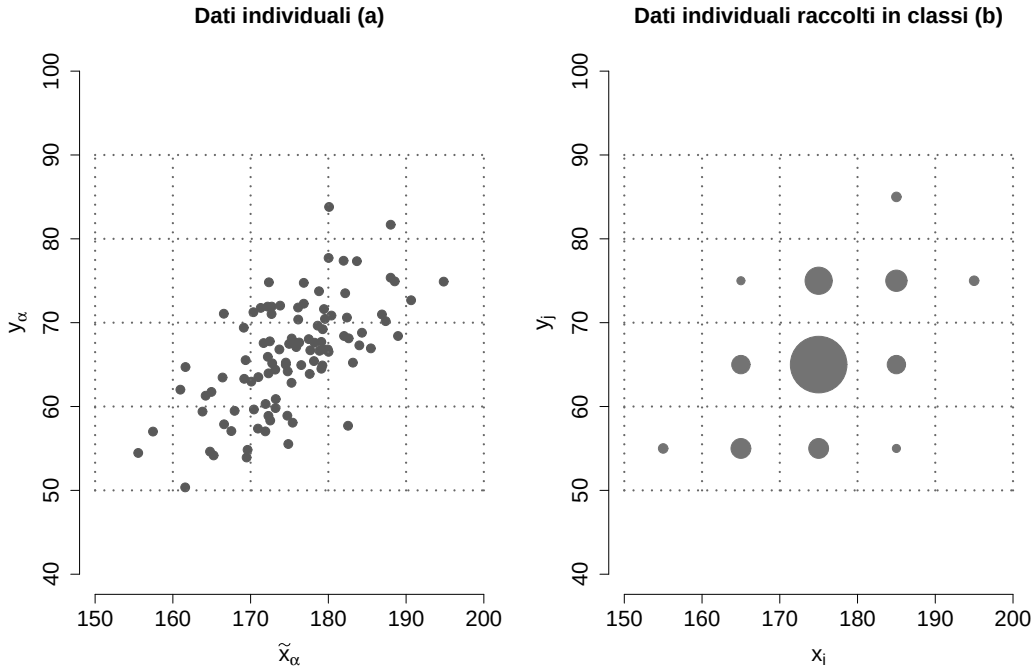


Figura 8.8 Diagramma a dispersione delle coppie $(\tilde{x}_\alpha, \tilde{y}_\alpha)$ e (x_i, y_i), esempio 8.15.

delle distribuzioni condizionate $Y|x_i$ e $X|y_j$. La grandezza dei simboli per i punti di coordinate $(x_i; \mu_{Y|x_i})$ e $(\mu_{X|y_j}; y_j)$ è proporzionale alle numerosità, rispettivamente, $n_{i\cdot}$ e $n_{\cdot j}$.
Osserviamo, infine che la scelta del numero e dell'ampiezza delle classi per ciascuna variabile statistica, passo necessario ai fini della costruzione della distribuzione di frequenze congiunte e ciò ai fini della verifica della indipendenza statistica e in media tra le componenti la v.s. bivariata (X, Y), si rivela ancora una volta cruciale poiché può produrre distorsioni, più o meno marcate, nelle conclusioni dell'analisi. Possiamo affermare, che qualora si scegliesse un numero esiguo di classi, il valore di $\eta^2_{Y|X}$ (o di $\eta^2_{X|Y}$) tenderebbe ad annullarsi, mentre all'aumentare delle classi esso tenderebbe all'unità. Tale certamente è il caso qualora esso venisse calcolato sui dati individuali.

◁

Concludiamo osservando che qualora la componente, ad esempio, X di una variabile statistica sia di tipo discreto ed interessi indagare circa la dipendenza in media di Y dalle modalità x_i di X, non è necessario procedere al raccoglimento dei dati individuali in

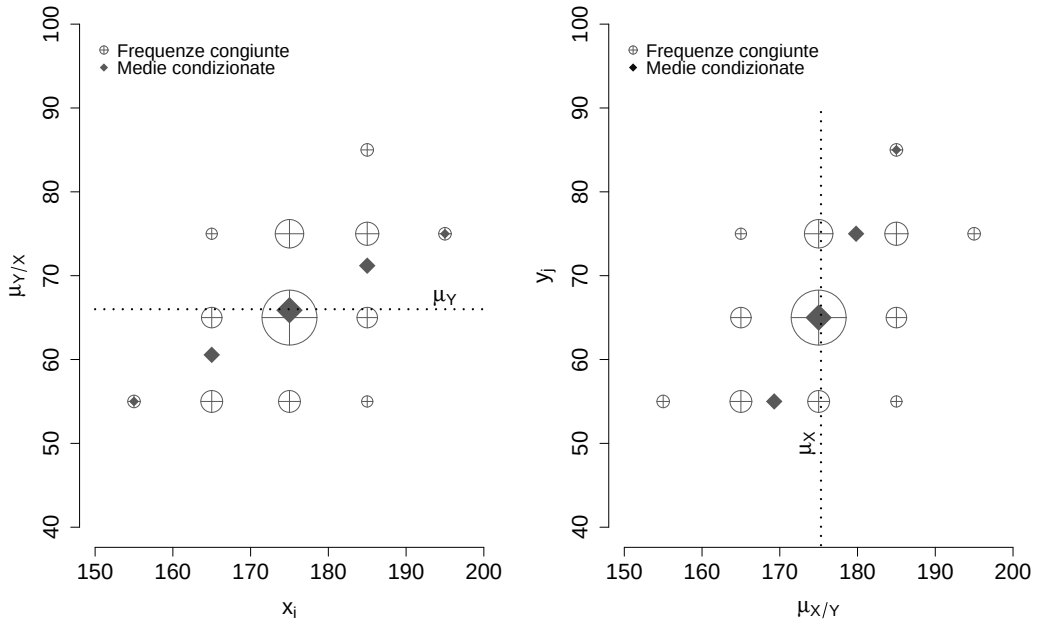


Figura 8.9 Medie delle v.s. condizionate $Y|x_i$ e $X|y_j$, esempio 8.15.

classi. Tale è il caso di molte situazioni sperimentali, emergenti in campo fisico, chimico, medico, ecc…, in cui interessa indagare circa i valori assunti da una variabile continua Y, abitualmente detta *variabile risposta,* in corrispondenza a particolari modalità, generalmente dette *livelli,* di una variabile di controllo, detta *fattore.*

▷ ESEMPIO 8.16

Supponiamo di avere rilevato su di un collettivo statistico costutito da 20 mutuati di uno studio medico il *numero di visite annue* e *l'età* e ottenendo la v.s. doppia $(X, Y) = \{$*numero di visite annue, età*$\}$ il cui insieme dei dati individuali sia:

$$\begin{aligned}\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\ldots,20} = \{\ & (1;30), (2;36), (6;72), (5;65), (3;55),\\ & (1;27), (3;48), (1;25), (6;74), (2;37),\\ & (5;67), (6;60), (5;63), (6;75), (5;56),\\ & (1;27), (5;62), (2;39), (3;58), (1;35)\}\end{aligned}$$

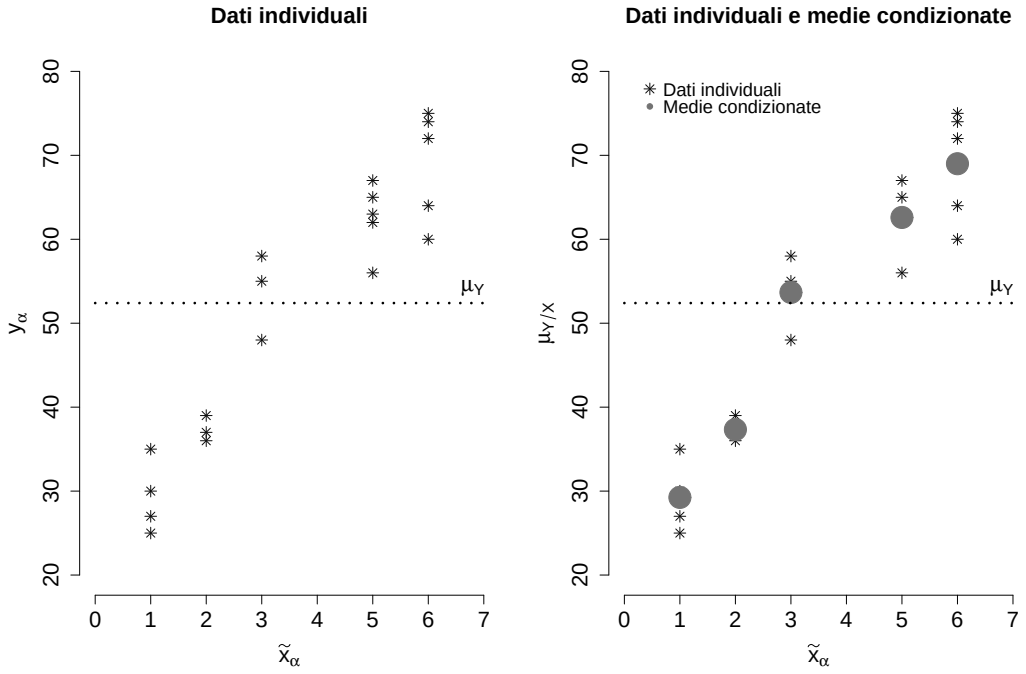


Figura 8.10 Dati individuali e valori medi delle v.s. condizionate $Y|x_\alpha$, esempio 8.16.

Desiderando verificare se esiste una relazione tra l'età dei mutuati ed il numero annuo di viste ambulatoriali, possiamo ricorrere ad un semplice diagramma a dispersione dei dati individuali, così come evidenziato in figura (8.10). Dalla sua lettura, appare evidente l'esistenza di una dipendenza statistica nonché in media tra le componenti la v.s. doppia (X, Y). La stessa figura riporta il diagramma a dispersione dei valori medi delle v.s. condizionate $Y|x_\alpha$, chiaramente tutti diversi tra loro. Infatti, lasciando al Lettore l'onere delle verifica numerica:

$$E[Y|x_1 = 155] = 29.250 \qquad E[Y|x_2 = 165] = 37.334$$
$$E[Y|x_3 = 175] = 53.667 \qquad E[Y|x_4 = 185] = 62.6$$
$$E[Y|x_5 = 195] = 69.000$$

Volendo quantificare il legame di dipendenza in media, essendo $V[Y] = 254.340$ e $V[E_{Y|X}] = 236.376$, abbiamo $\eta^2_{Y|X} = 0.9294$ a conferma della forte dipendenza in media tra Y e X. Come ci si poteva aspettare, l'età media dei mutuati pare crescere all'aumentare del numero di vistite ambulatoriali.
Saremmo giunti a risultati analoghi, senza perdita di informazione, se avessimo costruito la distribuzione di frequenze congiunte della v.s. (X, Y), scegliendo di racco-

gliere le modalità osservate di Y in queste tre particolari classi di età $]20;40]$, $]40;60]$ e $]60;80]$:

$X \downarrow Y \rightarrow$	$20 \dashv 40$	$40 \dashv 60$	$60 \dashv 80$
1	4	0	0
2	3	0	0
3	0	3	0
5	0	1	4
6	0	1	4

Lasciamo al Lettore la verifica numerica dei risultati $V = 0.8602$ e $\eta^2_{Y|X} = 0.9197$.

$\triangleleft$

8.3. ALCUNE UTILI DIMOSTRAZIONI

Nel seguito presenteremo le dimostrazioni di due proprietà che abbiamo sfruttato ai fini della costruzione degli indici normalizzati V di Cramér e $\eta^2_{Y|A}$ di Pearson.

Il massimo di χ^2

Ci proponiamo di dimostrare la proprietà (8.2) che afferma che qualunque sia la variabile statistica mista in osservazione, risulta:

$$\chi^2 \leq min\{\ n(s-1); n(r-1)\ \}$$

Occorre, palesemente, trovare una maggiorazione per χ^2. Se consideriamo l'espressione corrispondente alla (8.6), e cioè

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right) \tag{8.17}$$

si tratterà di trovare una maggiorazione per la quantità

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}\,n_{\cdot j}} \tag{8.18}$$

Ricordando che per definizione e per qualsiasi indice $i = 1,\ldots,r$ e $j = 1,\ldots,s$ sarà $n_{ij} \leq n_{i\cdot}$, possiamo affermare che

$$n_{ij}^2 \leq n_{ij}n_{i\cdot}$$

A questo punto, sommando su ciascun indice i e j per la (8.18) varrà la maggiorazione:

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} \leq \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{n_{i\cdot}}\frac{n_{i\cdot}}{n_{\cdot j}} = \sum_{j=1}^{s}\frac{1}{n_{\cdot j}}\sum_{i=1}^{r}n_{ij} = \sum_{j=1}^{s}\frac{n_{\cdot j}}{n_{\cdot j}} = s$$

Tornando, quindi, alla (8.17), abbiamo:

$$\chi^2 \leq n(s-1)$$

In modo del tutto analogo, osservando che per qualsiasi indice $j = 1, \ldots, s$ per definizione si ha $n_{ij} \leq n_{\cdot j}$, sarà

$$n_{ij}^2 \leq n_{ij}n_{\cdot j}$$

per cui sommando su ciascun indice i e j per la (8.18) varrà la maggiorazione:

$$\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} \leq \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}}{n_{i\cdot}}\frac{n_{\cdot j}}{n_{\cdot j}} = \sum_{i=1}^{r}\frac{1}{n_{i\cdot}}\sum_{j=1}^{s}n_{ij} = \sum_{i=1}^{r}\frac{n_{i\cdot}}{n_{i\cdot}} = r$$

donde il risultato

$$\chi^2 \leq n(r-1)$$

Dovendo essere verificate entrambe le diseguaglianze, si ha la tesi, cioè:

$$\chi^2 \leq min\{\ n(s-1)\ ;\ n(r-1)\ \}$$

Scissione della varianza

Ci proponiamo, ora, di dimostrare la proprietà (8.4), che va sotto il nome di *scissione della varianza* e afferma che, data la variabile statistica mista (A, Y), la varianza della sua componente Y è uguale dalla somma del valor medio delle varianze condizionate più la varianza delle medie condizionate, cioè in simboli vale l'uguaglianza:

$$V[Y] = E[V_{Y|A}] + V[E_{Y|A}]$$

Ai fini della dimostrazione, ricordando la definizione di varianza di Y

$$V[Y] = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - E[Y]\right)^2 n_{ij} \tag{8.19}$$

il trucco consiste nel sommare e sottrarre alla componente quadratica della (8.19) la quantità $E[Y|a_i]$, cioè:

$$V[Y] = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - E[Y]\right)^2 n_{ij} =$$
$$= \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(\left(y_j - E[Y|a_i]\right) + \left(E[Y|a_i] - E[Y]\right)\right)^2 n_{ij}$$

Sviluppando il quadrato e distribuendo le sommatorie, la (8.19) viene ad essere espressa quale somma di tre addendi:

$$V[Y] = \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - E[Y|a_i]\right)^2 n_{ij} +$$
$$+ \frac{2}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - E[Y|a_i]\right)\cdot\left(E[Y|a_i] - E[Y]\right) n_{ij} +$$
$$+ \frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(E[Y|a_i] - E[Y]\right)^2 n_{ij}$$

Se consideriamo ciascun addendo separatamente, abbiamo:

- ⋆ il primo addendo, ricordando la (8.11), coincide con il valor medio della distribuzione delle varianze condizionate, infatti:

$$\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - E[Y|a_i]\right)^2 n_{ij} = \frac{1}{n}\sum_{i=1}^{r} V[Y|a_i]\, n_{i\cdot} = E[V_{Y|A}]$$

- ⋆ il secondo addendo è nullo, infatti:

$$\frac{2}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(y_j - E[Y|a_i]\right)\cdot\left(E[Y|a_i] - E[Y]\right) n_{ij} =$$
$$= \frac{2}{n}\sum_{i=1}^{r}\left(E[Y|a_i] - E[Y]\right)\, n_{i\cdot}\, \frac{1}{n_{i\cdot}}\sum_{j=1}^{s}\left(y_j - E[Y|a_i]\right)\, n_{ij} = 0$$

  dal momento che

$$\frac{1}{n_{i\cdot}}\sum_{j=1}^{s}\left(y_j - E[Y|a_i]\right)\, n_{ij}$$

è nullo per qualsiasi $i = 1, \ldots, r$ in quanto media degli scarti dal valor medio della v.s. condizionata $Y|a_i$.

- ⋆ infine il terzo addendo viene a coincidere con la varianza della distribuzione delle medie condizionate cioè $V[E_{Y|A}]$, infatti:

$$\frac{1}{n}\sum_{i=1}^{r}\sum_{j=1}^{s}\left(E[Y|a_i]-E[Y]\right)^2 n_{ij} = \frac{1}{n}\sum_{i=1}^{r}\left(E[Y|a_i]-E[Y]\right)^2\sum_{j=1}^{s} n_{ij} =$$
$$= \frac{1}{n}\sum_{i=1}^{r}\left(E[Y|a_i]-E[Y]\right)^2 n_{i\cdot} = V[E_{Y|A}]$$

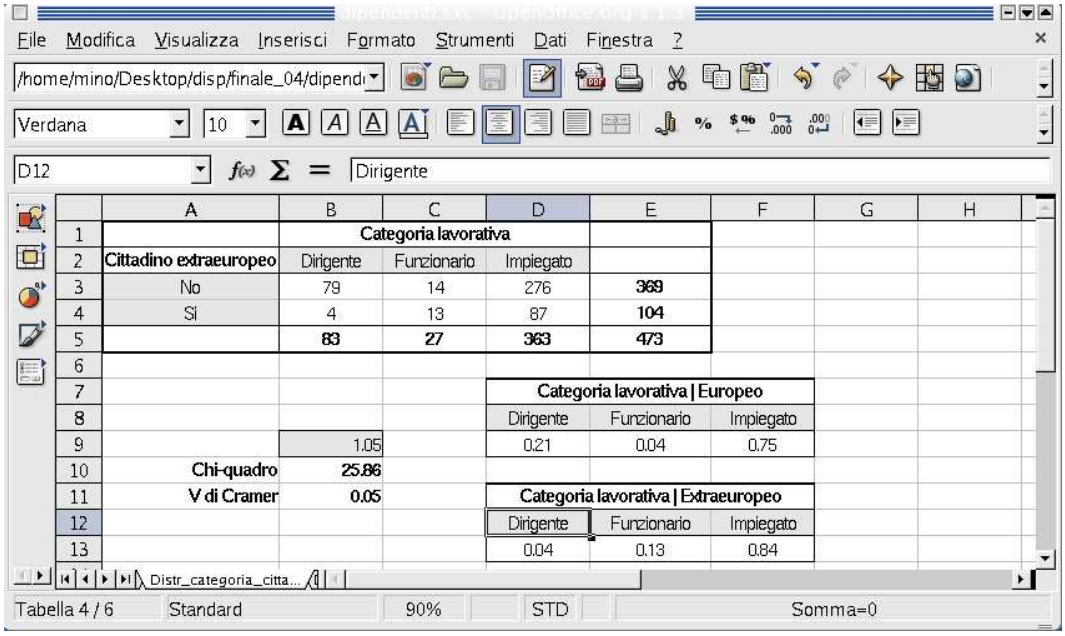
8.4. IL FOGLIO ELETTRONICO

Riprendiamo la distribuzione di frequenze congiunte della mutabile statistica bivariata $(A, B) = \{$*Cittadinanza, Categoria lavorativa*$\}$ del file `dipendenti.sxc` già individuata nel capitolo precedente e determiniamo le distribuzioni di frequenza delle due m.s. $A|b_j$ nonché i valori di χ^2 e V di Cramér.
Inizialmente copiamo e *incolliamo come valori* in una nuova cartella la tabella della distribuzione di frequenze congiunte restituita dalla procedura `Data Pilot` così che le celle che la compongono contengano numeri e non formule (cfr. figura 8.11).
Per ottenere la distribuzione di frequenze relative della m.s. *Categoria lavorativa* condizionata alla modalità $a_1 =$ *no* della m.s. *Cittadinanza* abbiamo posto nell'intervallo di celle `D8:F8` le modalità distinte della m.s. *Categoria lavorativa* e nell'intervallo `D9:F9` abbiamo posto le frequenze ad esse associate inserendo, ad esempio, nelle cella `D9` la formula `=B3/$E$3`. In modo del tutto analogo abbiamo posto nell'intervallo di celle `D12:F13` la distribuzione di frequenze relative m.s. *Categoria lavorativa* condizionata alla modalità $a_2 =$ *sì*.
Per il calcolo di χ^2 ci siamo basati sulla, ormai ben nota formula

$$\chi^2 = n\left(\sum_{i=1}^{r}\sum_{j=1}^{s}\frac{n_{ij}^2}{n_{i\cdot}n_{\cdot j}} - 1\right)$$

al fine di ottenere il valore numerico della doppia sommatoria che in essa compare, abbiamo inserito nella cella `B9`

=B3^2/(E3*B5)+C3^2/(E3*C5)+D3^2/(E3*D5)+B4^2/(E4*B5)+ C4^2/(E4*C5)+D4^2/(E4*D5)



	A	B	C	D	E
1		Categoria lavorativa			
2	Cittadino extraeuropeo	Dirigente	Funzionario	Impiegato	
3	No	79	14	276	369
4	Si	4	13	87	104
5		83	27	363	473

	A	B
9		1.05
10	Chi-quadro	25.86
11	V di Cramer	0.05

Categoria lavorativa | Europeo		
Dirigente	Funzionario	Impiegato
0.21	0.04	0.75

Categoria lavorativa | Extraeuropeo		
Dirigente	Funzionario	Impiegato
0.04	0.13	0.84

Figura 8.11 Misure di dipendenza statistica per la m.s. *(Cittadinanza, Categoria)*

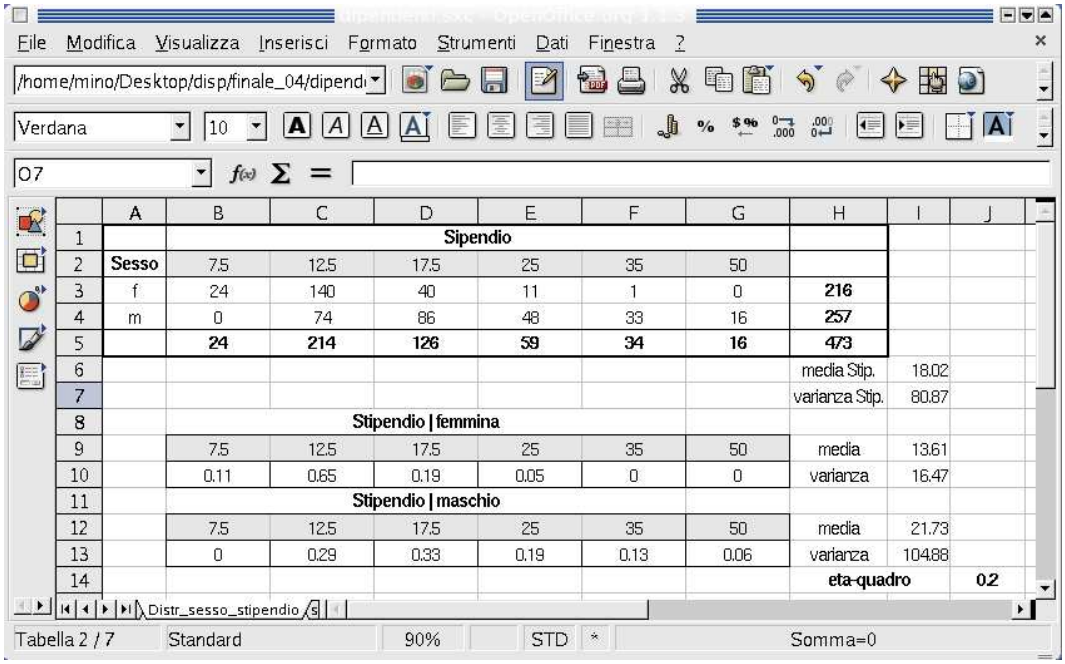
così, per χ^2, abbiamo inserito nella cella `B10` la funzione `=E5*(B9-1)`. Osservando che il massimo di χ^2 è, in questo caso, pari a n il valore di V di Cramér nella cella `B11` è dato dalla funzione `=B10/E5`.
Riprendendo ora la variabile statistica mista $(A, Y) = \{$*Sesso, Stipendio attuale*$\}$, individuiamo dalla distribuzione di frequenze congiunte le distribuzioni delle due variabili condizionate $Y|x_1$ e $Y|x_2$ (cfr. figura 8.12), procedendo analogamente a quanto fatto per le mutabili precedenti.
Al fine di calcolare $\eta^2_{Y|A}$ usando la consueta formula

$$\eta^2_{Y|A} = \frac{V[E_{Y|A}]}{V[Y]}$$

determiniamo inizialmente media e varianza della v.s. Y inserendo nelle celle `I6` e `I7` le formule rispettivamente

=(B2*B5+C2*C5+D2*D5+E2*E5+F2*F5+G2*G5)/H5

=(B2^2*B5+C2^2*C5+D2^2*D5+E2^ 2*E5+F2^2*F5+G2^2*G5)/H5-I6^2



	A	B	C	D	E	F	G	H	I	J
1		Sipendio								
2	Sesso	7.5	12.5	17.5	25	35	50			
3	f	24	140	40	11	1	0	216		
4	m	0	74	86	48	33	16	257		
5		24	214	126	59	34	16	473		
6								media Stip.	18.02	
7								varianza Stip.	80.87	
8		Stipendio | femmina							
9		7.5	12.5	17.5	25	35	50	media	13.61	
10		0.11	0.65	0.19	0.05	0	0	varianza	16.47	
11		Stipendio | maschio							
12		7.5	12.5	17.5	25	35	50	media	21.73	
13		0	0.29	0.33	0.19	0.13	0.06	varianza	104.88	
14								eta-quadro		0.2

Figura 8.12 Misura di dipendenza in media per la v.s. *(Sesso, Stipendio attuale)*

Per quanto riguarda la media delle variabili condizionata $Y|x_i$ abbiamo inserito nelle celle `I9` e `I12` le funzioni

=B9*B10+C9*C10+D9*D10+E9*E10+F9*F10+G9*G10

=B12*B13+C12*C13+D12*D13+E12*E13+F12*F13+G12*G13

Infine la formula `=((I9^2*H3+I12^2*H4)/H5-I6^2)/I7` inserita nella cella `J14` restituisce il valore di $\eta^2_{Y|A}$.
Si invita il Lettore a ripetere il calcolo di $\eta^2_{Y|A}$, a partire dalle varianze delle variabili condizionate, poste nelle celle `I10` e `I13`, utilizzando la formula opportuna.

8.5. Esercizi

▷ Esercizio 8.1

La rilevazione del numero di dipendenti (X) e del fatturato giornaliero (Y), su un collettvo statistico costituito da 60 esercizi pubblici ha dato luogo alla seguente

distribuzione di frequenze congiunte:

$X \downarrow$ $Y \rightarrow$	$200 \dashv 400$	$400 \dashv 800$	$800 \dashv 1000$	$1000 \dashv 2000$
1	10	5	2	0
2	4	12	2	1
3	1	2	11	3
4	0	1	6	10

verificare se le componenti X e Y possono ritenersi statisticamente indipendenti. In caso contrario calcolare l'indice normalizzato V di Cramér.

◁

▷ ESERCIZIO 8.2

Con riferimento alla situazione descritta al punto precedente (esercizio 8.1), calcolare $\eta^2_{Y|X}$, $Cov[X, Y]$ e ρ e se ne dia un'interpretazione.

◁

▷ ESERCIZIO 8.3

Da un'indagine condotta su 690 studenti universitari, residenti nella cintura torinese, circa il sesso e il mezzo di trasporto abitualmente impiegato per recarsi in Facoltà si è ottenuta la m.s. bivariata $(A, B) = \{$*sesso, mezzo di trasporto*$\}$ con distribuzione di frequenze congiunte:

$A \downarrow$ $B \rightarrow$	Pubblico	Auto	Scooter
Femmina	120	60	44
Maschio	275	103	88

Si verifichi l'indipendenza statistica tra le componenti A e B. In caso di dipendenza, si proceda al calcolo dell'indice normalizzato V di Cramér e si tenti un'interpretazione del fenomeno.

◁

▷ ESERCIZIO 8.4

Calcolare media e varianza della componente Y di una v.s. mista (A, Y) sapendo che:

$$\begin{array}{lll} \mu_{Y|a_1} = 4/9 & \sigma^2_{Y|a_1} = 20/81 & n_{1\cdot} = 9 \\ \mu_{Y|a_2} = 1/3 & \sigma^2_{Y|a_2} = 2/9 & n_{2\cdot} = 3 \\ \mu_{Y|a_3} = 0 & \sigma^2_{Y|a_3} = 0 & n_{3\cdot} = 3 \end{array}$$

◁

▷ ESERCIZIO 8.5

Posto che la v.s. bivariata (X, Y) abbia distribuzione di frequenze congiunte:

$X \downarrow \quad Y \rightarrow$	-1	0	1
-1	4	1	0
0	2	6	2
1	0	1	4

si individui la distribuzione di frequenze congiunte della v.s. bivariata (W, Z), dove si è posto $W = X - E[X]$ e $Z = Y - E[Y]$. Verificare, inoltre se (W, Z) ha componenti indipendenti dal punto di vista statistico e in media.
Si calcoli, infine $Cov[W, Z]$

◁

▷ ESERCIZIO 8.6

Con riferimento alla variabile statistica bivariata di cui all'esercizio (8.5), individuare la distribuzione di frequenze congiunte della v.s. doppia (U, W), dove

$$U = \left(\frac{X - \mu_X}{\sigma_X}\right) \qquad W = \left(\frac{Y - \mu_Y}{\sigma_Y}\right)$$

e si verifichi se essa ha componenti indipendenti dal punto di vista statistico e in media.
Si calcoli, infine $Cov[U, W]$ e ρ^2.

◁

▷ ESERCIZIO 8.7

Posto che la v.s. bivariata (X, Y) abbia distribuzione di frequenze congiunte:

$X \downarrow \quad Y \rightarrow$	0	1
-1	3	0
0	1	4
1	0	5

calcolare $\eta^2_{Y|X}$, $Cov[X, Y]$ e ρ.

◁

▷ ESERCIZIO 8.8

Da un censimento di imprese artigiane del settore manifatturiero si è rilevata la seguente v.s. (X, Y) dove:

$X = \{$*tributi versati nell'anno 2004*$\}$ in euro
$Y = \{$*volume degli affari nell'anno 2004*$\}$ in migliaia di euro.

con distribuzione congiunta di frequenze assolute:

$X \downarrow Y \rightarrow$	$0 \dashv 100$	$100 \dashv 200$	$200 \dashv 300$	$300 \dashv 500$
$0 \dashv 30$	8	2	0	0
$30 \dashv 60$	3	20	9	1
$60 \dashv 90$	1	25	55	37
$90 \dashv 150$	0	3	7	10

Si proceda a:

- ⋆ verificare l'indipendenza statistica e calcolare il χ^2;
- ⋆ verificare l'indipendenza in media di Y da X e calcolare $\eta^2_{Y|X}$;
- ⋆ calcolare la covarianza tra X e Y e successivamente ρ;

◁

▷ ESERCIZIO 8.9

Con riferimento alla situazione descritta all'esercizio 8.8, si considerino le variabili statistiche condizionate $Y|x_1$ e $Y|x_4$. Per ciascuna di esse si costruisca il rispettivo istogramma e successivamente il diagramma a scatola e baffi.

◁

CAPITOLO 9

REGRESSIONE LINEARE

In questo che è l'ultimo capitolo dedicato alla statistisca descrittiva daremo alcuni concetti di regressione lineare. Verranno determinati i parametri della retta interpolante secondo il metodo dei minimi quadrati e il concetto di regressione sarà introdotto con l'ausilio di un esempio. Al fine di definire il coefficiente di determinazione R^2 quale misura di bontà di adattamento del modello ai dati saranno enunciate e dimostrate alcune proprietà dei residui di regressione. Si perverrà infine all'interpretazione del coefficiente di correlazione lineare, il cui quadrato, nel caso di retta di regressione, coincide con il coefficiente di determinazione.

9.1. INTRODUZIONE

Alla luce di quanto proposto nei capitoli precedenti, tentiamo ora di tradurre in forma funzionale il legame di dipendenza che può intercorrere tra le componenti di una variabile statistica bivariata. A tale fine opereremo sull'insieme delle coppie di dati individuali $(\tilde{x}_\alpha, \tilde{y}_\alpha)_{\alpha=1,\dots,n}$ della variabile statistica bivariata (X, Y) per cui risulti evidente una relazione tra le sue componenti. Più precisamente supporremo che modalità assunte sul collettivo statistico dalla componente Y possano essere spiegate e riassunte mediante un modello analitico dalle modalità assunte dalla componente X.
Si immagini che le coppie di dati individuali $(\tilde{x}_\alpha; \tilde{y}_\alpha)_{\alpha=1,\dots,n}$ di una v.s. bivariata possano essere rappresentati graficamente sul piano cartesiano mediante un diagramma a dispersione. Il problema consiste nella:

- ⋆ scelta del modello matematico, generalmente un modello lineare nei parametri $\hat{Y} = f(\tilde{x}_\alpha; a_1, \dots, a_m)$, ritenuto più idoneo alla descrizione sintetica del legame funzionale tra le componenti Y e X della v.s. bivariata (X, Y);
- ⋆ determinazione dei valori dei parametri del modello in accordo ad un prescelto criterio.

Quanto al primo punto, osserviamo che esso deve essere di semplice struttura e al contempo deve evidenziare la "tendenza" della relazione tra le grandezze in gioco.
Quanto al secondo problema, ovviamente potremmo scegliere di individuare i parametri del modello prescelto semplicemente ipotizzando che esso, anziché passare per tutti gli n punti di coordinate $(\tilde{y}_\alpha; \tilde{x}_\alpha)$, passi solo per alcuni punti del diagramma a dispersione.
A tal proposito, la figura (9.1) riporta il diagramma a dispersione di una possibile situazione. Se si osserva la "nuvola di punti" di coordinate $(\tilde{y}_\alpha; \tilde{x}_\alpha)$, appare del tutto naturale supporre che il legame funzionale tra le v.s. Y e X possa ben essere rappresentato da una retta, cioè si suppone valga la relazione $\hat{Y} = a_0 + a_1\, X$.

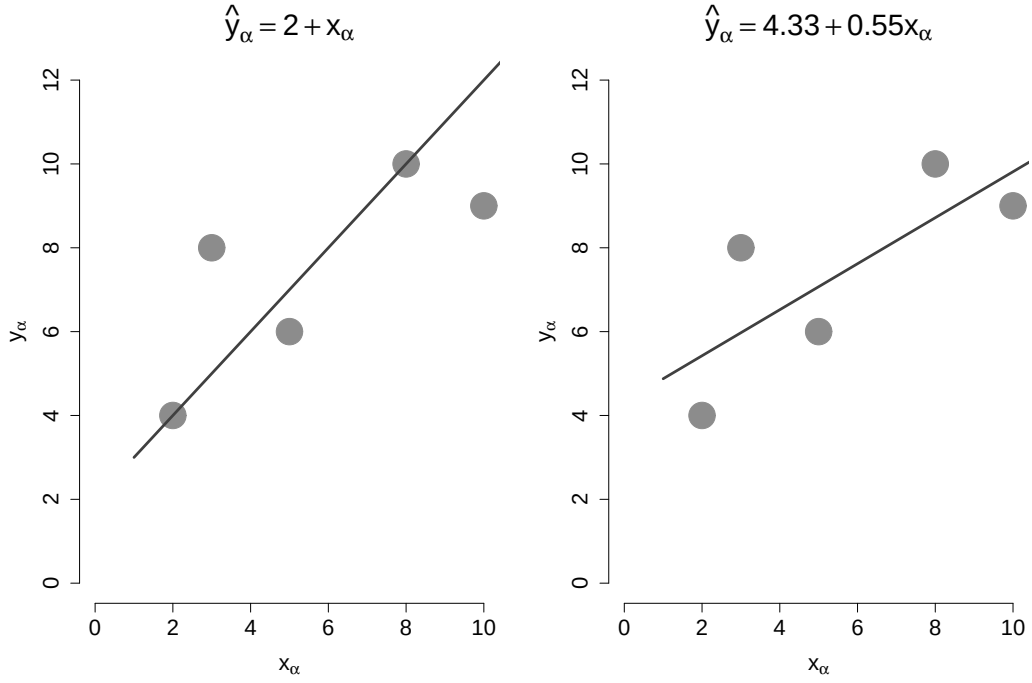


Figura 9.1 Diagramma a dispersione e possibili rette interpolanti.

Il problema è ora quello di determinare il valore dei parametri a_0 e a_1 del modello. Sempre con riferimento alla figura (9.1), osserviamo che:

- ⋆ il pannello (a) propone quale modello quello della retta passante per i punti di coordinate $(\tilde{x}_1 = 2; \tilde{y}_1 = 4)$ e $(\tilde{x}_4 = 8; \tilde{x}_4 = 10)$; si ipotizza cioè che il modello sia $\hat{Y} = 2 + X$;

- ⋆ il pannello (b) propone quale modello quella della retta che globalmente passa "tra" tutti i punti $(\tilde{x}_\alpha; \tilde{y}_\alpha)$; si ipotizza cioè che il modello sia $\hat{Y} = 4.33 + 0.55\, X$.

Definito modello $\hat{Y} = f(X; a_0, \ldots, a_m)$ lineare nei parametri a_h, con $h = 0, \ldots, m \leq n$, abitualmente si ricorre al criterio, detto dei *minimi quadrati*, che permette di individuare il valore dei parametri in modo che essi minimizzino la somma del quadrato delle distanze tra i punti di coordinate $(\tilde{x}_\alpha; \tilde{y}_\alpha)$ e $(\tilde{x}_\alpha; \hat{y}_\alpha)$.

9.2. IL METODO DEI MINIMI QUADRATI E LA RETTA DI REGRESSIONE

Scelto il modello funzionale, il criterio a cui si fa abitualmente ricorso in ambito statistico è quello che che va sotto il nome di *metodo dei minimi quadrati.*
Esso consente, per un dato modello $f\,(\tilde{x}_\alpha; a_0, \ldots, a_m)$, di determinarne l'insieme dei parametri $\{a_h\}_{h=0,\ldots,m}$ in modo che essi minimizzino, nell'ambito di quella famiglia di curve, il quadrato degli scarti tra i valori osservati ($\tilde{y}_\alpha$) e i corrispondenti valori teorici ($\hat{y}_\alpha$), cioè desunti dal modello stesso.

Definizione 9.1 (Metodo dei minimi quadrati)
scelto quale legame funzionale tra le componenti di una variabile statistica doppia (X, Y) *un modello* $\hat{Y} = f\,(X; a_0, \ldots, a_m)$*, il metodo dei minimi quadrati individua, sulla base dell'insieme dei dati individuali* $\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\ldots,n}$*, i valori dei parametri in modo che essi rendano minima la funzione:*

$$\varphi\,(a_0, \ldots, a_m) = \sum_{\alpha=1}^{n} \left(\tilde{y}_\alpha - f\,(\tilde{x}_\alpha; a_0, \ldots, a_m)\right)^2 \tag{9.1}$$

□

Se quale legame funzionale tra le componenti di una variabile statistica bivariata (X, Y) si scegliesse il modello lineare

$$\hat{Y} = a_0 + a_1\, X \tag{9.2}$$

applicando la definizione (9.1), i valori dei parametri a_0 e a_1 secondo il metodo dei minimi quadrati sono individuati minimizzando la funzione

$$\varphi\,(a_0, a_1) = \sum_{\alpha=1}^{n} \left(\tilde{y}_\alpha - a_0 - a_1\, \tilde{x}_\alpha\right)^2 \tag{9.3}$$

La minimizzazione della (9.3) comporta la risoluzione, rispetto ai due parametri a_0 e a_1, del seguente sistema di equazioni (per la ricerca dei punti stazionari):

$$\begin{cases} \dfrac{\partial\,\varphi\,(a_0, a_1)}{\partial\, a_0} = 0 \\ \dfrac{\partial\,\varphi\,(a_0, a_1)}{\partial\, a_1} = 0 \end{cases} \tag{9.4}$$

Dal momento che le due derivate prime parziali di $\varphi\,(a_0, a_1)$ risultano

$$\frac{\partial\,\varphi\,(a_0, a_1)}{\partial\, a_0} = -2 \sum_{\alpha=1}^{n} (\tilde{y}_\alpha - a_0 - a_1\,\tilde{x}_\alpha)$$
$$\frac{\partial\,\varphi\,(a_0, a_1)}{\partial\, a_1} = -2 \sum_{\alpha=1}^{n} (\tilde{y}_\alpha - a_0 - a_1\,\tilde{x}_\alpha)\;\tilde{x}_\alpha$$

il sistema di equazioni (9.4) assume la forma

$$\begin{cases} \displaystyle\sum_{\alpha=1}^{n} (\tilde{y}_\alpha - a_0 - a_1\,\tilde{x}_\alpha) & = 0 \\ \displaystyle\sum_{\alpha=1}^{n} (\tilde{y}_\alpha - a_0 - a_1\,\tilde{x}_\alpha)\;\tilde{x}_\alpha & = 0 \end{cases} \tag{9.5}$$

e viene detto *sistema delle equazioni normali.*
Dividendo, ora, ambo i membri delle equazioni per n e distribuendo le sommatorie che vi compaiono, il sistema (9.5) diviene

$$\begin{cases} \dfrac{1}{n}\left(\displaystyle\sum_{\alpha=1}^{n} \tilde{y}_\alpha - n\,a_0 - a_1 \sum_{\alpha=1}^{n} \tilde{x}_\alpha\right) & = 0 \\ \dfrac{1}{n}\left(\displaystyle\sum_{\alpha=1}^{n} \tilde{y}_\alpha\,\tilde{x}_\alpha - a_0 \sum_{\alpha=1}^{n} \tilde{x}_\alpha - a_1 \sum_{\alpha=1}^{n} \tilde{x}_\alpha^2\right) & = 0 \end{cases}$$

per cui, ricorrendo all'operatore $E[\,\cdot\,]$

$$\begin{cases} a_0 + a_1\,E[X] = E[Y] \\ a_0\,E[X] + a_1\,E[X^2] = E[X \cdot Y] \end{cases} \tag{9.6}$$

Ai fini della risoluzione rispetto ad a_0 e a_1 del sistema di equazioni (9.6), conviene porre il sistema stesso nella forma matriciale

$$\mathbf{X}\,\mathbf{a} = \mathbf{y}$$

dove $\mathbf{a} = [a_0, a_1]^T$ è il vettore delle incognite, $\mathbf{y} = [E[Y], E[X \cdot Y]]^T$ il vettore dei termini noti ed infine

$$\mathbf{X} = \begin{bmatrix} 1 & E[X] \\ E[X] & E[X^2] \end{bmatrix}$$

rappresenta la matrice dei coefficienti del sistema stesso.
Se si osserva che $det(\mathbf{X}) = E[X^2] - (E[X])^2 = V[X]$, sfruttando la regola di Cramer, il coefficiente angolare della retta risulta

$$a_1 = \frac{det \begin{bmatrix} 1 & E[Y] \\ E[X] & E[X \cdot Y] \end{bmatrix}}{V[X]} = \frac{E[X \cdot Y] - E[X] \cdot E[Y]}{V[X]} = \\ = \frac{Cov\,[X, Y]}{V[X]} \tag{9.7}$$

Quanto al parametro a_0, esso può essere ottenuto in funzione di a_1, infatti dalla prima equazione del sistema (9.6) si ha

$$a_0 = E[Y] - a_1\, E[X] \tag{9.8}$$

A commento delle soluzioni ottenute per il modello interpolante $Y = a_0 + a_1\, X$, osserviamo che:

- ⋆ il segno del coefficiente angolare della retta interpolante ai minimi quadrati è determinato da quello della covarianza tra le componenti la v.s. bivariata (X, Y);
- ⋆ qualora la covarianza tra X e Y risultasse nulla, il che implicherebbe $a_1 = 0$, dalla (9.8) risulterebbe $\hat{Y} = E[Y]$;
- ⋆ la retta interpolante ai minimi quadrati passa sempre per il punto di coordinate $(\mu_X; \mu_Y)$. Infatti, posto $x = \mu_X$, risulta $\hat{Y} = E[Y] - a_1\, \mu_X + a_1\, \mu_X = \mu_Y$;
- ⋆ il valor medio della v.s. $\hat{Y}$ coincide con il valor medio di Y, infatti

$$E[\hat{Y}] = E[a_0 + a_1\, X] = a_0 + a_1\, E[X] = \\ = E[Y] - a_1\, E[X] + a_1\, E[X] = E[Y] \tag{9.9}$$

▷ ESEMPIO 9.1

Si immagini che la rilevazione su 13 autovetture di piccola cilindrata ad alimentazione a benzina, circa la cilindrata in (cm^3) ed il consumo di carburante (litri $\times$ 100 km) abbia dato luogo alla v.s. bivariata $(X, Y) = \{\textit{cilindrata, consumo}\}$ con insieme dei dati individuali:

$$\begin{aligned}\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\dots,13} =& (1050; 4.6), (1055; 5.1), (1075; 4.9), (1075; 5.0),\\ & (1095; 4.8), (1100; 5.2), (1125; 5.4), (1165; 4.9),\\ & (1170; 5.3), (1175; 5.4), (1225; 4.9), (1245; 5.2),\\ & (1275; 5.3)\end{aligned}$$

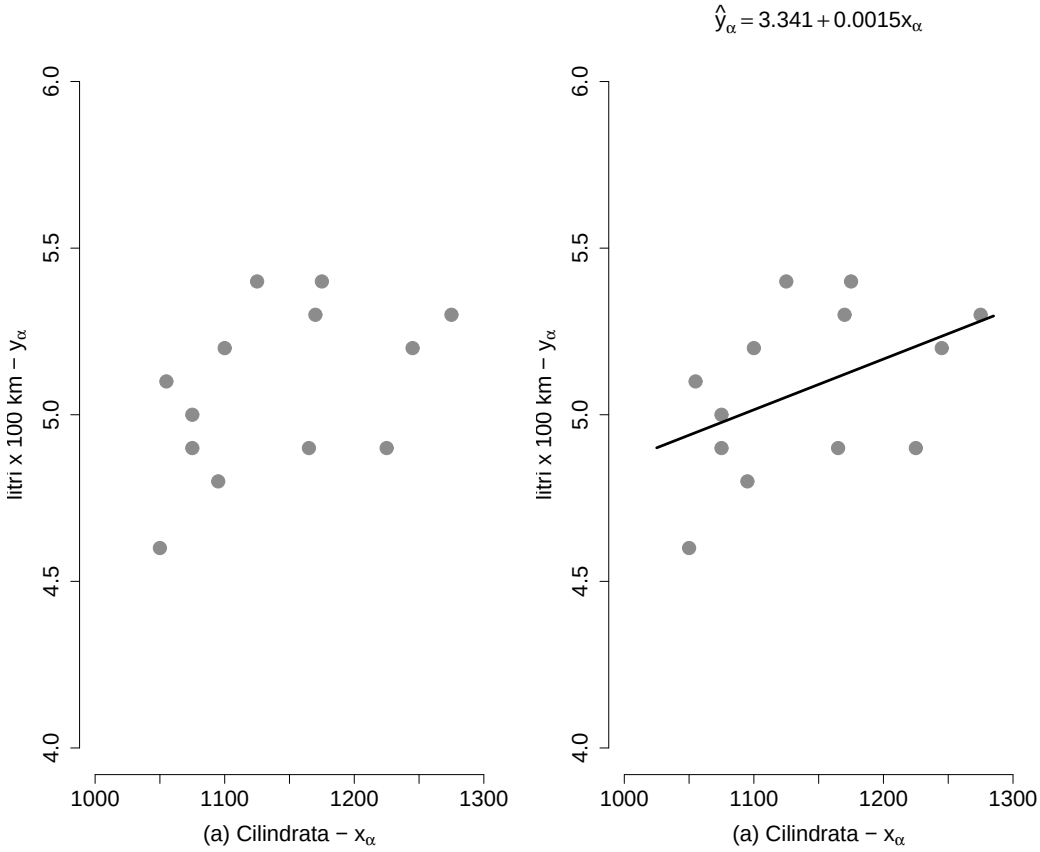


Figura 9.2 Diagrammi a dispersione e retta interpolante a minimi quadrati, esempio 9.1.

Il diagramma a dispersione dei dati individuali, proposto in figura (9.2, pannello a), suggerisce come la relazione che lega il consumo di carburante e la cilindrata possa essere riassumibile dal modello lineare $\hat{Y} = a_0 + a_1\, X$.
Scelto pertanto di ricorrere a tale modello, il valore dei parametri a_0 e a_1 viene ad essere determinato, in accordo al metodo dei minimi quadrati, come

$$a_1 = \frac{Cov\,[X,Y]}{V[X]} = \frac{7.825444}{5141.716} = 0.0015$$
$$a_0 = E[Y] - a_1\, E[X] = 5.077 - 0.0015 \cdot 1140.769 = 3.366$$

In definitiva il modello, il cui grafico è riportato in figura (9.2, pannello b), diviene $\hat{Y} = 3.366 + 0.0015\, X$.

◁

Qualora la componente X della variabile statistica bivariata in esame presentasse un esiguo numero di modalità distinte x_i, allora in corrispondenza a ciascuna di esse si avrebbe una variabile statistica condizionata $Y|x_i$ dotata ovviamente di valor medio $\mu_{Y|x_i}$.
Tale situazione è rappresentata graficamente dai diagrammi e dispersione di figura (9.3). Il pannello (a) propone il diagramma a dispersione dell'insieme delle coppie di dati individuali $\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\ldots,n}$, mentre nel pannello (b) è stato aggiunto l'insieme delle coppie $\{(x_i; \mu_{Y|x_i})\}_{i=1,\ldots,r}$.
Se tale è la situazione, allora, scelto il modello $\hat{Y} = a_0 + a_1\, X$, i parametri individuati secondo il metodo dei minimi quadrati operando sui dati individuali vengono a coincidere con quelli ottenibili minimizzando il quadrato degli scarti tra le medie condizionate ($E[Y|x_i]$) e i corrispondenti valori teorici ($\hat{y}_i$), cioè minimizzando la funzione

$$\varphi\,(a_0, a_1) = \sum_{i=1}^{r} \left(E[Y|x_i] - a_0 - a_1\, x_i\right)^2\, n_{i\cdot} \tag{9.10}$$

Infatti, il sistema delle derivate prime parziali rispetto ad a_0 e a_1 diviene

$$\begin{cases} \displaystyle\sum_{i=1}^{r} \left(E[Y|x_i] - a_0 - a_1\, x_i\right)\, n_{i\cdot} & = 0 \\ \displaystyle\sum_{i=1}^{r} \left(E[Y|x_i] - a_0 - a_1\, x_i\right)\, x_i\, n_{i\cdot} & = 0 \end{cases}$$

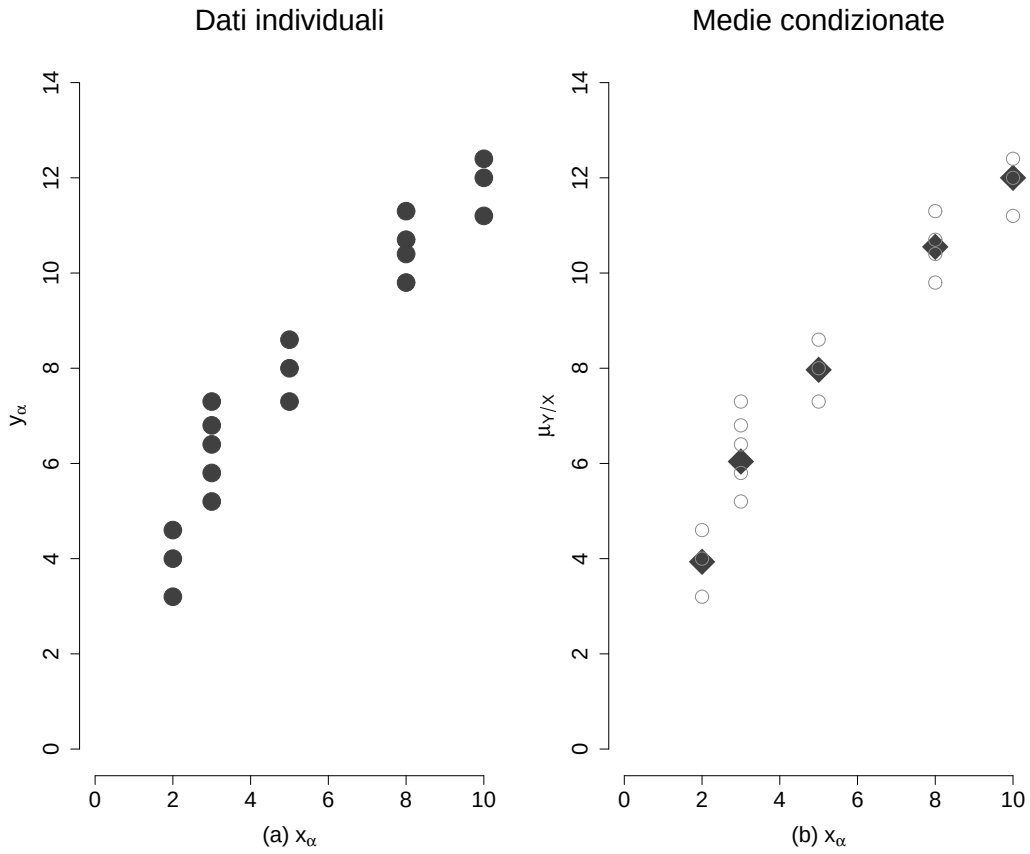


Figura 9.3 Diagramma a dispersione: dati individuali e medie condizionate.

che, dividendo ambo i membri delle equazioni per n e distribuendo le sommatorie che vi compaiono, può essere scritto

$$\begin{cases} \dfrac{1}{n}\left(\displaystyle\sum_{i=1}^{r} E[Y|x_i]\, n_{i\cdot} - a_0 \sum_{i=1}^{r} n_{i\cdot} - a_1 \sum_{i=1}^{r} x_i\, n_{i\cdot}\right) & = 0 \\ \dfrac{1}{n}\left(\displaystyle\sum_{i=1}^{r} E[Y|x_i]\, x_i\, n_{i\cdot} - a_0 \sum_{i=1}^{r} x_i\, n_{i\cdot} - a_1 \sum_{i=1}^{r} x_i^2\, n_{i\cdot}\right) & = 0 \end{cases}$$

Ricorrendo in ultimo all'operatore $E[\,\cdot\,]$, otteniamo il sistema nella forma

$$\begin{cases} a_0 + a_1\, E[X] = E[E_{Y|X}] \\ a_0\, E[X] + a_1\, E[X^2] = E[X \cdot E_{Y|X}] \end{cases}$$

da cui le soluzioni

$$\begin{aligned} a_1 &= \frac{Cov\left[X, E_{Y|X}\right]}{V[X]} \\ a_0 &= E[E_{Y|X}] - a_1\, E[X] \end{aligned}$$

Ricordando che $E[E_{Y|X}] = E[Y]$, osservando che la covarianza tra X e $E_{Y|X}$ può essere vista come

$$Cov[X, E_{Y|X}] = E[X \cdot E_{Y|X}] - E[X]\, E[E_{Y|X}] = E[X \cdot E_{Y|X}] - E[X]\, E[Y]$$

dove il termine $E[X \cdot E_{Y|X}]$, tenendo a mente la definizione di medie condizionate, viene ad essere

$$\begin{aligned} E[X \cdot E_{Y|X}] = \frac{1}{n} \sum_{i=1}^{r} x_i\, E[Y|x_i]\, n_{i\cdot} = \frac{1}{n} \sum_{i=1}^{r} x_i \frac{1}{n_{i\cdot}} \sum_{j=1}^{s} y_j\, n_{ij}\, n_{i\cdot} = \\ = \frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{s} x_i\, y_j\, n_{ij} = E[X \cdot Y] \end{aligned}$$

le soluzioni, come già detto, coincidono con quelle ottenibili con il metodo dei minimi quadrati operando sui dati individuali, cioè

$$\begin{aligned} a_1 &= \frac{Cov\,[X, Y]}{V[X]} \\ a_0 &= E[Y] - a_1\, E[X] \end{aligned}$$

Tali considerazioni, ci consentono di capire perché la retta interpolante ai minimi quadrati viene abitualmente detta *retta di regressione*. In statistica matematica si chiama *funzione di regressione* l'insieme di coppie costituite dalle modalità distinte x_i e dalle corrispondenti medie condizionate $E[Y|x_i]$; visto che la retta interpolante ai minimi quadrati è quella che "più si avvicina alla funzione di regressione", per analogia viene detta *retta di regressione*.

▷ ESEMPIO 9.2

Si immagini che la rilevazione del voto della prova scritta di Statistica ed il voto finale conseguito da 20 studenti in una sessione di esame abbia dato luogo alla variabile statistica bivariata $(X, Y) = \{$*voto della prova scritta, voto finale*$\}$ con insieme dei dati individuali

$$\begin{aligned}\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\ldots,25} =& (18,18), (18,21), (18,23), (18,24), (18,19),\\ & (20,19), (20,23), (20,25), (23,18), (23,23),\\ & (23,26), (23,21), (23,25), (24,24), (24,25),\\ & (24,27), (24,30), (25,25), (25,27), (25,30)\end{aligned}$$

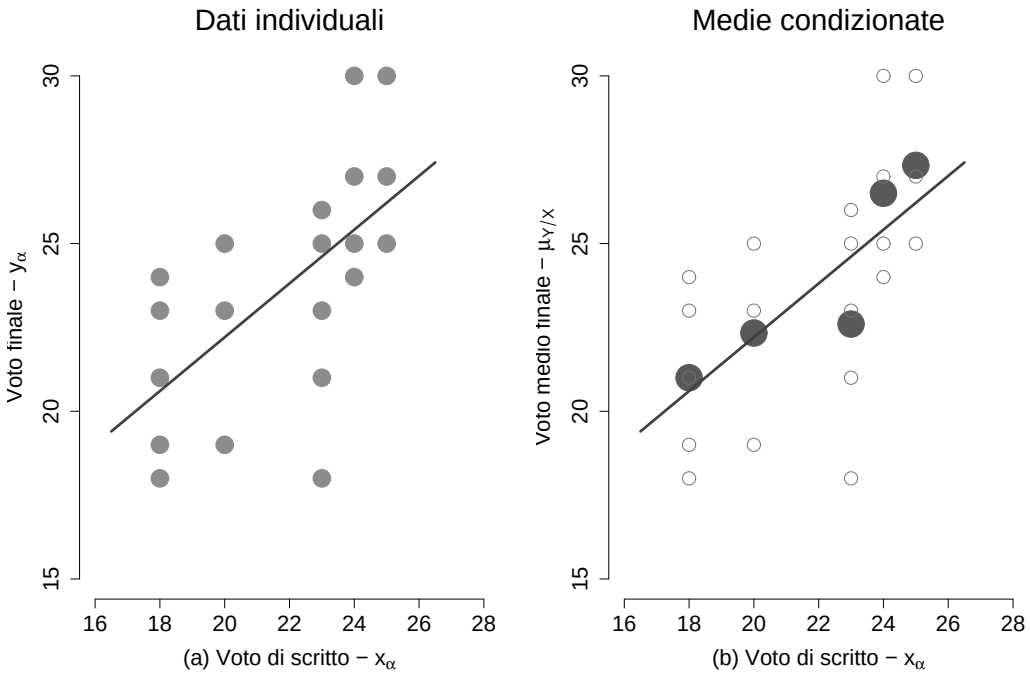


Figura 9.4 Medie condizionate $Y|\tilde{x}_\alpha$ e retta di regressione, esempio 9.2.

Il diagramma a dispersione dei dati individuali, proposto in figura (9.4, pannello a), suggerisce come la relazione tra il voto della prova scritta ed il voto finale possa essere riassunta dal modello $\hat{Y} = a_0 + a_1 X$.

Pertanto il valore dei parametri a_0 e a_1 viene ad essere determinato, in accordo al metodo dei minimi quadrati, come

$$a_1 = \frac{Cov\,[X,Y]}{V[X]} = \frac{5.58}{6.96} = 0.802$$
$$a_0 = E[Y] - a_1\,E[X] = 23.65 - 0.802 \cdot 21.80 = 6.166$$

L'andamento della retta di regressione $\hat{Y} = 6.166 + 0.802\,X$ è riportato in figura (9.4, pannello a).
Se si osserva che in tale situazione, l'insieme delle modalità distinte assunte dalla v.s. discreta X corrisponde all'insieme $\{x_i\}_{i=1,..,5} = \{18, 20, 23, 24, 25\}$, allora è possibile, senza alcuna perdita di informazione, costruire la seguente distribuzione di frequenze congiunte per la v.s. bivariata (X, Y)

$X \downarrow Y \rightarrow$	18	19	21	23	24	25	26	27	30	
18	1	1	1	1	1	0	0	0	0	5
20	0	1	0	1	0	1	0	0	0	3
23	1	0	1	1	0	1	1	0	0	5
24	0	0	0	0	1	1	0	1	1	4
25	0	0	0	0	0	1	0	1	1	3

La distribuzione delle medie delle cinque v.s. condizionate $Y|x_i$ risulta

$$E_{Y|X} \equiv \left\{ \begin{matrix} E[Y|x_i] \\ n_{i\cdot} \end{matrix} \right\}_{i=1,..,5} = \left\{ \begin{matrix} 21.00 & 22.33 & 22.60 & 26.50 & 27.33 \\ 5 & 3 & 5 & 4 & 3 \end{matrix} \right\}$$

e dal momento che per essa

$$Cov\left[X, E_{Y|X}\right] = Cov\,[X,Y] = 5.58 \qquad E[E_{Y|X}] = E[Y] = 23.65$$

si ottiene

$$a_1 = \frac{Cov\left[X, E_{Y|X}\right]}{V[X]} = \frac{5.58}{6.96} = 0.802$$
$$a_0 = E[E_{Y|X}] - a_1\,E[X] = 23.65 - 0.802 \cdot 21.80 = 6.166$$

◁

Concludiamo il paragrafo osservando che data una v.s. bivariata le cui componenti siano v.s. continue per le quali non sia possibile pervenire alla distribuzione di frequenze congiunte se non raccogliendo i dati individuali in classi, il calcolo dei valori dei parametri della retta di regressione dovrà essere effettuato sulla base delle coppie di dati individuali $(\tilde{x}_\alpha; \tilde{y}_\alpha)$. Anche in questo caso, infatti, la scelta del numero e dell'ampiezza delle classi per ciascuna variabile statistica condurrebbe inevitabilmente ad una distorsione nelle conclusioni dell'analisi qualora i parametri della retta fossero determinati a partire dalla distribuzione di frequenze congiunte. A tal riguardo, valga l'esempio che segue.

▷ ESEMPIO 9.3

Si supponga che la rilevazione della *statura* (in cm) e del *peso corporeo* (in kg) di 20 persone abbia dato luogo alla v.s. bivariata $(X, Y) = \{$*statura, peso*$\}$ con insieme di dati individuali

$$\begin{aligned}\{(\tilde{x}_\alpha; \tilde{y}_\alpha)\}_{\alpha=1,\ldots,20} =& (160.5; 57.0), (162.2; 58.2), (164.8; 59.5), (166.0; 60.3),\\ & (168.4; 60.7), (188.5; 84.7), (168.6; 62.3), (169.2; 65.1),\\ & (169.8; 66.2), (171.1; 70.7), (173.8; 70.8), (176.0; 73.1),\\ & (178.8; 73.1), (180.8; 73.2), (181.2; 74.1), (183.3; 75.6),\\ & (183.5; 78.8), (184.0; 80.3), (184.8; 84.5), (190.4; 89.0)\end{aligned}$$

il cui diagramma a dispersione è proposto in figura (9.5, pannello a), dal quale si evince, come del resto ci si aspetta, l'esistenza di un legame lineare tra le variabili statistiche Y ed X. Scelto pertanto il modello $\hat{Y} = a_0 + a_1\, X$, posto di ricorrere al metoto dei minimi quadrati, i valori dei parametri della retta di regressione risultano

$$\begin{aligned} a_1 &= \frac{Cov\,[X, Y]}{V[X]} = \frac{79.509}{77.223} = 1.0296\\ a_0 &= E[Y] - a_1\, E[X] = 70.860 - 1.0296 \cdot 175.285 = -109.613\end{aligned}$$

Il grafico della retta di regressione $\hat{Y} = -109.613 + 1.0296\, X$ è riportato sempre in figura (9.5, pannello a).
Si immagini, ora, di raccogliere i dati individuali di ciascuna v.s. in 4 classi di ampiezza costante sì da ottenere la seguente distribuzione di frequenze congiunte

$X \downarrow Y \rightarrow$	$50 \dashv 60$	$60 \dashv 70$	$70 \dashv 80$	$80 \dashv 90$
$160 \dashv 170$	3	5	0	0
$170 \dashv 180$	0	0	4	0
$180 \dashv 190$	0	0	4	3
$190 \dashv 200$	0	0	0	1

il cui corrispondente diagramma a bolle è riportato in figura (9.5, pannello b), da cui appare ancora lecito supporre una relazione di tipo lineare tra le variabili Y ed X. Scelto pertanto il modello di regressione $\hat{Y} = a_0 + a_1\, X$, individuati i centri di classe per entrambe le variabili statistiche, calcolate le medie, le varianze nonché la covarianza sulla base della distribuzione di frequenze congiunte, i valori dei parametri risultano

$$\begin{aligned} a_1 &= \frac{Cov\,[X, Y]}{V[X]} = \frac{81.750}{94.750} = 0.8628\\ a_0 &= E[Y] - a_1\, E[X] = 71.500 - 0.8628 \cdot 175.500 = -79.921\end{aligned}$$

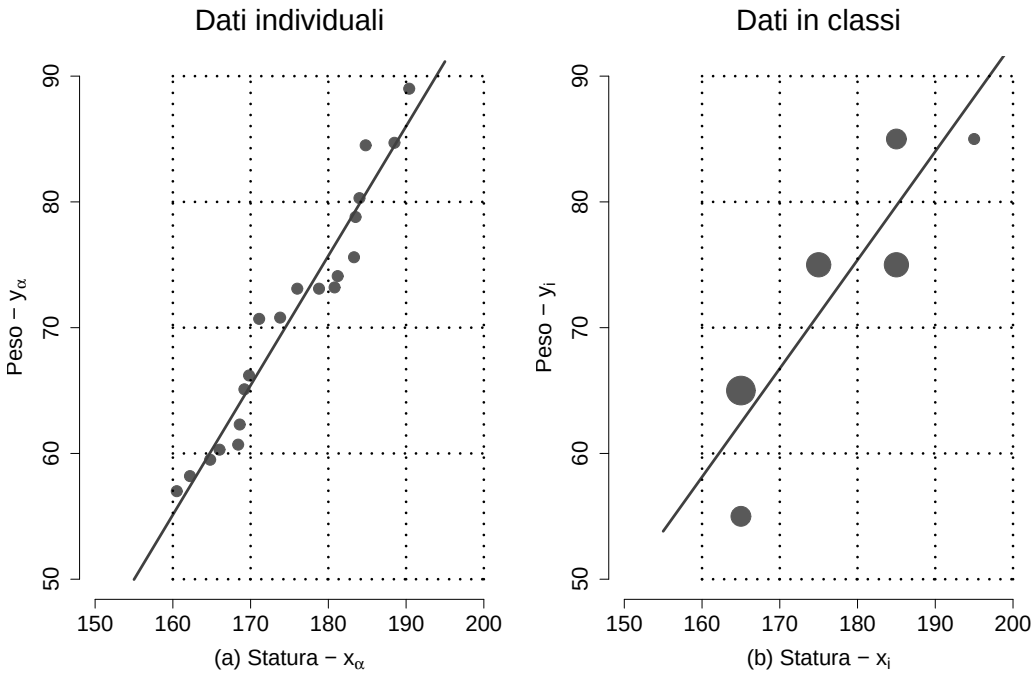


Figura 9.5 Rette di regressione a confronto, esempio 9.3.

Il grafico della retta di regressione $\hat{Y} = -79.921 + 0.8628\,X$ è riportato sempre in figura (9.5, pannello b).

Manifestamente i due modelli di regressione differiscono nei parametri poiché si riferiscono a due diverse situazioni. Ciò è unicamente dovuto all'effetto del raggruppamento, peraltro artificioso ed arbitrario, dei dati individuali in classi. Tale effetto può essere colto ad occhio osservando la distribuzione dei punti entro le griglie di figura (9.5).

◁

9.3. BONTÀ DI ADATTAMENTO

Individuati i parametri della retta di regressione, disponiamo della nuova variabile statistica $\hat{Y}$ il cui insieme di dati individuali sarà determinabile a partire da quello della v.s. X tramite la trasformata lineare $\hat{y}_\alpha = a_0 + a_1\,\tilde{x}_\alpha$. Ovviamente, per costruzione, le

coppie $(\tilde{x}_\alpha; \hat{y}_\alpha)$ corrisponderanno nel piano cartesiano a punti appartenenti alla retta di regressione.

Per avere un'idea di quanto la retta ben si "adegui" ai dati osservati, ci pare del tutto naturale considerare la variabile statistica, detta *residui di regressione*, definita quale differenza tra la v.s. osservata Y e la v.s. teorica $\hat{Y}$, cioè

$$Y - \hat{Y} \tag{9.11}$$

La costruzione ed interpretazione di un diagramma a dispersione delle coppie $(\tilde{x}_\alpha; \tilde{y}_\alpha - \hat{y}_\alpha)$ è un modo semplice ed intuitivo per cogliere l'adeguatezza o meno del modello ai dati. Se il modello ben si adatta ai dati, allora i punti di coordinate $(\tilde{x}_\alpha; \tilde{y}_\alpha - \hat{y}_\alpha)$ si dispongono nel piano in modo sparso, non evidenziando alcuna tendenza di fondo, così come nel grafico proposto in figura (9.6, pannello a). Allo scemare della bontà del modello, il diagramma tende ad evidenziare regolarità di comportamento, come ad esempio illustrato in figura (9.6, pannello b).

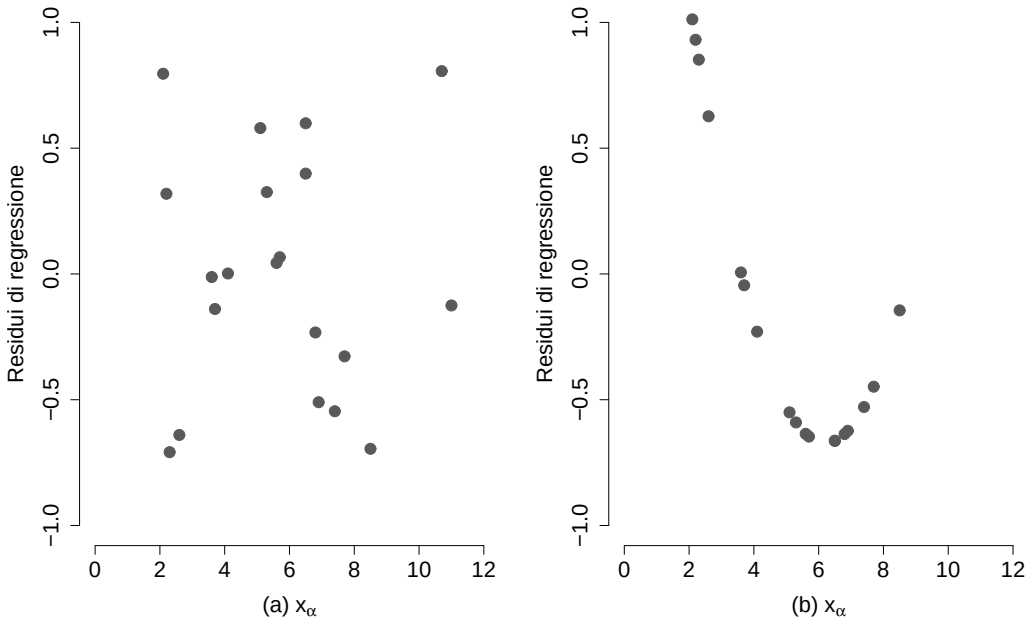


Figura 9.6 Residui di due diversi modelli di regressione.

Ovviamente, qualora i valori osservati appartenessero tutti alla retta di regressione, il modello scelto si adatterebbe perfettamente ai nostri dati e la variabile residui di regressione sarebbe costantemente nulla. Va da sè che al crescere ed al variare dei valori assunti dai residui di regressione, l'adeguatezza del modello scema via, via. Una misura, dunque, di bontà di adattamento della retta di regressione ai dati potrebbe essere offerta dal valor medio dei residui, tuttavia esso è nullo; ricordando infatti la (9.9) risulta $E[Y - \hat{Y}] = E[Y] - E[\hat{Y}] = 0$. L'alternativa è rappresentata dal ricorso alla varianza dei residui di regressione, data da

$$V[Y - \hat{Y}] = \frac{1}{n} \sum_{\alpha=1}^{n} (\tilde{y}_\alpha - \hat{y}_\alpha)^2$$

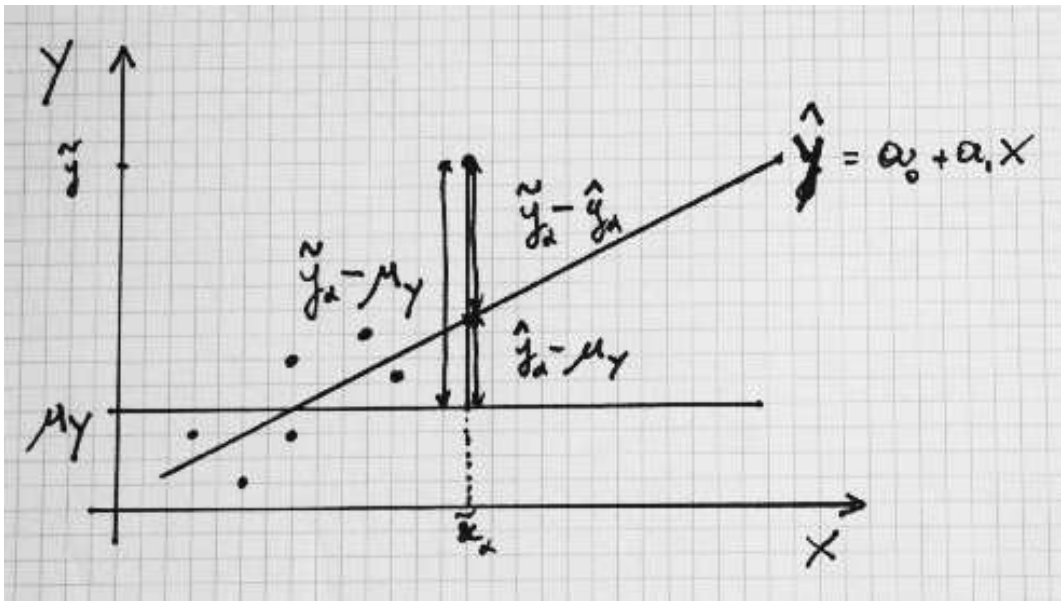


Figura 9.7 Uguaglianza $\tilde{y}_\alpha - \mu_Y = (\tilde{y}_\alpha - \hat{y}_\alpha) + (\hat{y}_\alpha - \mu_Y)$.

Desiderando pervenire ad un indice normalizzato di misura della bontà di adattamento del modello ai dati, possiamo, analogamente a quanto fatto per gli indici di dipendenza, cercare una maggiorazione della varianza dei residui. A tal fine possiamo sfruttare la seguente proprietà della varianza della v.s. Y, che è vera per qualsiasi modello lineare nei parametri e che si basa, così come evidenziato in figura (9.7), sulla uguaglianza

$$\tilde{y}_\alpha - \mu_Y = (\tilde{y}_\alpha - \hat{y}_\alpha) + (\hat{y}_\alpha - \mu_Y)$$

valida qualunque sia $\alpha = 1, \ldots, n$.

Proprietà 9.1 La varianza della componente Y di una v.s. doppia (X, Y) può essere scomposta nella somma della varianza dei residui di regressione e della varianza spiegata dal modello di regressione, cioè

$$V[Y] = V\left[Y - \hat{Y}\right] + V\left[\hat{Y} - E[Y]\right] \tag{9.12}$$

(per la dimostrazione, cfr. paragrafo 9.5).

◁

Evidentemente dall'equazione (9.12) risulta che la varianza dei residui di regressione sarà sempre non maggiore della varianza di Y, per cui il rapporto

$$\frac{V\left[Y - \hat{Y}\right]}{V[Y]}$$

è sempre compreso tra zero ed uno e potrebbe essere assunto quale indice normalizzato di bontà di adattamento. Poiché in tal modo un valore nullo del rapporto implicherebbe un perfetto adattamento del modello ai dati e viceversa un valore pari all'unità sarebbe indice di pessima adeguatezza, quale misura di bontà si preferisce utilizzare il complemento ad uno di tale rapporto.

Definizione 9.2 (Coefficiente di determinazione $\mathbf{R^2}$)
Definiamo misura normalizzata di bontà di adattamento di un modello di regressione ai dati la quantità

$$R^2 = 1 - \frac{V\left[Y - \hat{Y}\right]}{V[Y]} \tag{9.13}$$

detta coefficiente di determinazione e per la quale vale $0 \leq R^2 \leq 1$.

□

Ricordando la (9.12), il coefficiente di determinazione R^2 può anche essere scritto nella forma

$$R^2 = \frac{V\left[\hat{Y} - E[Y]\right]}{V[Y]}$$

Valori di R^2 prossimi all'unità consentiranno di ritenere soddisfacente il modello scelto, mentre bassi valori condurranno ad una sua eventuale riformulazione.

▷ ESEMPIO 9.4

A commento di quanto esposto, riprendiamo la situazione di cui all'esempio (9.3). Come si disse le coppie di dati individuali $(\tilde{x}_\alpha; \tilde{y}_\alpha)$ evidenziano di per sè che la relazione tra il peso e la statura degli individui censiti possa essere riassunta dal modello di regressione $\hat{Y} = -109.613 + 1.0296\, X$, così come evidenziato dal grafico proposto in figura (9.8, pannello a).

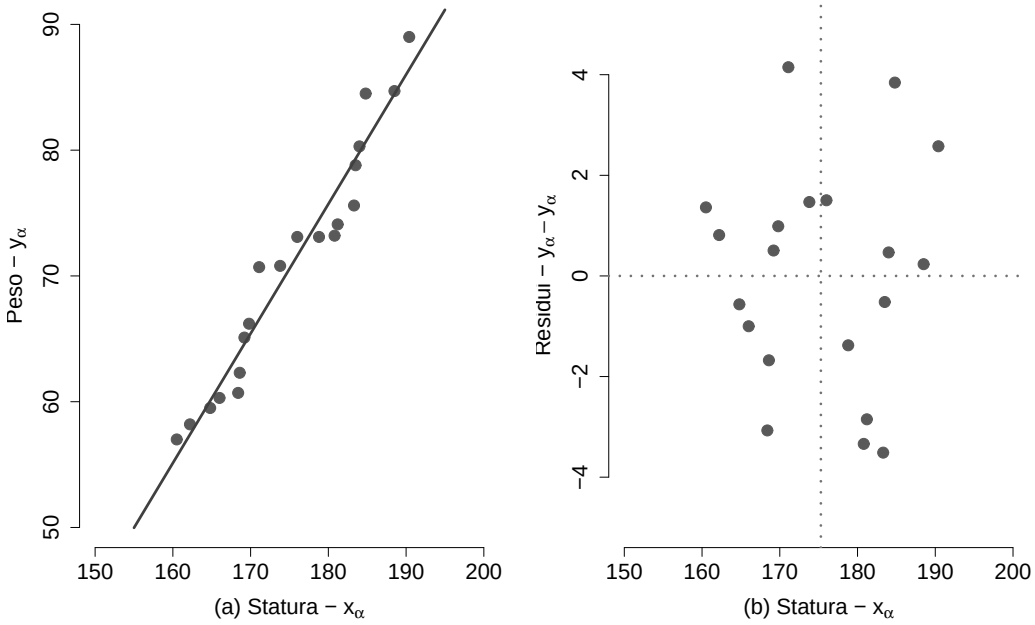


Figura 9.8 Retta di regressione e residui di regressione, esempio 9.4.

Ricorrendo a tale modello, in forma tabellare abbiamo

$\tilde{x}_\alpha$	$\tilde{y}_\alpha$	$\hat{y}_\alpha$	$\tilde{y}_\alpha - \hat{y}_\alpha$		$\tilde{x}_\alpha$	$\tilde{y}_\alpha$	$\hat{y}_\alpha$	$\tilde{y}_\alpha - \hat{y}_\alpha$
160.5	57.0	55.637	1.363		173.8	70.8	69.331	1.469
162.2	58.2	57.388	0.812		176.0	73.1	71.596	1.504
164.8	59.5	60.065	-0.565		178.8	73.1	74.479	-1.379
166.0	60.3	61.300	-1.000		180.8	73.2	76.538	-3.338
168.4	60.7	63.771	-3.071		181.2	74.1	76.950	-2.850
188.5	84.7	84.466	0.234		183.3	75.6	79.112	-3.512
168.6	62.3	63.977	-1.677		183.5	78.8	79.318	-0.518
169.2	65.1	64.595	0.505		184.0	80.3	79.833	0.467
169.8	66.2	65.213	0.987		184.8	84.5	80.657	3.843
171.1	70.7	66.551	4.149		190.4	89.0	86.422	2.578

Come ci si poteva attendere, l'adattamento del modello di regressione ai dati osservati è buono, come emerge dal diagramma a dispersione dei residui proposto in figura (9.8, pannello b) ove il punti di coordinate $(\tilde{x}_\alpha; \tilde{y}_\alpha - \hat{y}_\alpha)$ sono "casualmente" disposti nel piano.
Desiderando misurare il grado di adattamento del modello di regressione ricorrendo al coefficiente di determinazione R^2, sarà sufficiente calcolare la varianza di Y nonché quella dei residui. Dal momento che

$$V[Y] = 86.582 \qquad V[Y - \hat{Y}] = 4.720$$

abbiamo

$$R^2 = 1 - \frac{V[Y - \hat{Y}]}{V[Y]} = 1 - \frac{4.720}{86.582} = 0.945$$

il che conferma quanto già intuito dalla lettura del diagramma a dispersione dei residui.

◁

Il coefficiente di determinazione R^2 è una misura di bontà di adattamento ai dati qualsiasi sia il modello a cui si fa ricorso. Solo nel caso si adotti quale modello la retta esso coincide con il quadrato del coefficiente di correlazione lineare, cioè ρ^2.
Tale affermazione si basa sulla seguente proprietà della varianza dei residui dalla retta di regressione.

Proprietà 9.2 La varianza dei residui dalla retta di regressione è proporzionale alla varianza di Y tramite il fattore $1 - \rho^2$, cioè $V[Y - \hat{Y}] = V[Y]\ (1 - \rho^2)$.

(per la dimostrazione, cfr. paragrafo 9.5).

◁

Sostituendo nella (9.13) la varianza dei residui dalla retta di regressione, otteniamo che $R^2 = \rho^2$, infatti

$$R^2 = 1 - \frac{V\,[Y]\;(1-\rho^2)}{V\,[Y]} = \rho^2$$

Qualora il coefficiente di correlazione lineare ρ fosse, in valore assoluto, pari all'unità la variabilità di Y verrebbe ad essere totalmente spiegata dalla retta di regressione, infatti si avrebbe $V[Y] = V\left[\hat{Y} - E[Y]\right]$. In tal caso i dati individuali $(\tilde{x}_\alpha; \tilde{y}_\alpha)$ corrisponderebbero a punti del piano cartesiano allineati su una retta, appunto quella di regressione.
Viceversa, qualora il coefficiente di correlazione lineare ρ fosse nullo, il che implicherebbe una covarianza nulla tra le componenti la variabile statistica doppia e quindi un coefficiente angolare nullo della retta di regressione con l'ovvia conseguenza che questa risulterebbe $\hat{Y} = E[Y]$, il modello nulla spiegherebbe della variabilità di Y, infatti si avrebbe $V\,[Y] = V[Y - \hat{Y}]$.
Va da sè che nella realtà ci si troverà sempre in situazioni intermedie, nel senso che generalmente $-1 < \rho < 1$.

9.4. MODELLI LINEARIZZABILI

A compendio di quanto esposto sulla regressione lineare, dedichiamo questo paragrafo alla presentazione di alcuni casi particolari che possono tornare utili in pratica.
In molti casi è la natura stessa del problema, di cui le coppie $(\tilde{x}_\alpha; \tilde{y}_\alpha)$ ne sono manifestazione, che suggerisce il tipo di modello interpolante, in altri casi sono l'occhio e la perspicacia del ricercatore a guidare nella scelta di un modello che bene si adatti ai punti assegnati.
Esistono alcuni modelli non lineari nei parametri che tuttavia possono essere ricondotti alla forma lineare. Semplici considerazioni matematiche consentono infatti di trasformare modelli del tipo $Y = a \cdot b^X$, $Y = a \cdot X^b$ e $Y = (a + b\,X)^{-1}$ in forma lineare. Infatti:

- per $Y = a \cdot b^X$ passando al logaritmo naturale si ha

  $$ln(Y) = ln(a \cdot b^X) = ln(a) + ln(b^X) = ln(a) + ln(b)\,X$$

  cioè $Z = a_0 + a_1\,W$, se poniamo $Z = ln(Y)$, $W = X$, $a_0 = ln(a)$ e $a_1 = ln(b)$.

- per $Y = a \cdot X^b$ passando al logaritmo naturale si ha

  $$ln(Y) = ln(a \cdot X^b) = ln(a) + ln(X^b) = ln(a) + b\,ln(X)$$

  cioè $Z = a_0 + a_1\,W$, se poniamo $Z = ln(Y)$, $W = ln(X)$, $a_0 = ln(a)$ e $a_1 = b$.

- per $Y = (a + b\,X)^{-1}$ considerandone il reciproco

  $$\frac{1}{Y} = a + b\,X$$

  cioè $Z = a_0 + a_1\,W$, se poniamo $Z = 1/Y$, $W = X$, $a_0 = a$ e $a_1 = b$.

Dal punto di vista statistico per scegliere il modello adeguato si può procedere in maniera empirica. Se il diagramma a dispersione suggerisce che il modello lineare non è adeguato a descrivere la relazione tra le due componenti la v.s. doppia (X, Y), si può ricercare se il legame intercorrente è descrivibile da un modello linearizzabile opportunamente trasformando le variabili X e Y e cogliendo dal nuovo diagramma a dispersione l'eventuale legame lineare.
Tra le possibili trasformazioni, citiamo le seguenti:

- *trasformazione semi-logaritmica*: $W = X, Z = ln(Y)$.

  In tale caso

  $$(\tilde{x}_\alpha; \tilde{y}_\alpha) \quad \rightarrow \quad (\tilde{w}_\alpha = \tilde{x}_\alpha; \tilde{z}_\alpha = ln(\tilde{y}_\alpha))$$

  Se la retta pare ben adattarsi all'insieme dei dati individuali della nuova v.s. (W, Z), cioè appare lecito porre $\hat{z}_\alpha = a_0 + a_1\,\tilde{w}_\alpha$, allora il modello interpolante per la v.s. bivariata (X, Y) sarà $\hat{y}_\alpha = a \cdot b^{\tilde{x}_\alpha}$, con $a = exp(a_0)$ e $b = exp(a_1)$.

- *trasformazione doppia-logaritmica*: $W = ln(X), Z = ln(Y)$.

  $$(\tilde{x}_\alpha; \tilde{y}_\alpha) \quad \rightarrow \quad (\tilde{w}_\alpha = ln(\tilde{x}_\alpha); \tilde{z}_\alpha = ln(\tilde{y}_\alpha))$$

  Se la retta pare ben adattarsi all'insieme dei dati individuali della nuova v.s. (W, Z), cioè appare lecito porre $\hat{z}_\alpha = a_0 + a_1\,\tilde{w}_\alpha$, allora il modello interpolante i punti del piano (x, y) sarà $\hat{y}_\alpha = a \cdot \tilde{x}_\alpha^b$, con $a = exp(a_0)$ e $b = a_1$.

- *trasformazione semi-reciproca*: $W = X, Z = 1/Y$.

  $$(\tilde{x}_\alpha; \tilde{y}_\alpha) \quad \rightarrow \quad (\tilde{w}_\alpha = \tilde{x}_\alpha; \tilde{z}_\alpha = \tilde{y}_\alpha^{-1})$$

  Se la retta pare ben adattarsi all'insieme dei dati individuali della nuova v.s. (W, Z), cioè appare lecito porre $\hat{z}_\alpha = a_0 + a_1\,\tilde{x}_\alpha$, allora il modello interpolante i punti del piano (x, y) sarà $\hat{y}_\alpha = (a + b\,\tilde{x}_\alpha)^{-1}$, con $a = a_0$ e $b = a_1$.

▷ ESEMPIO 9.5

A commento si immagini di disporre delle seguenti coppie di dati individulali

$$\{(\tilde{x}_\alpha;\tilde{y}_\alpha)\}_{\alpha=1,\ldots,10} = (1;3),(2;12),(3;9),(4;20),(5;37),\\ (6;45),(7;67),(8;80),(9;130),(10;210)$$

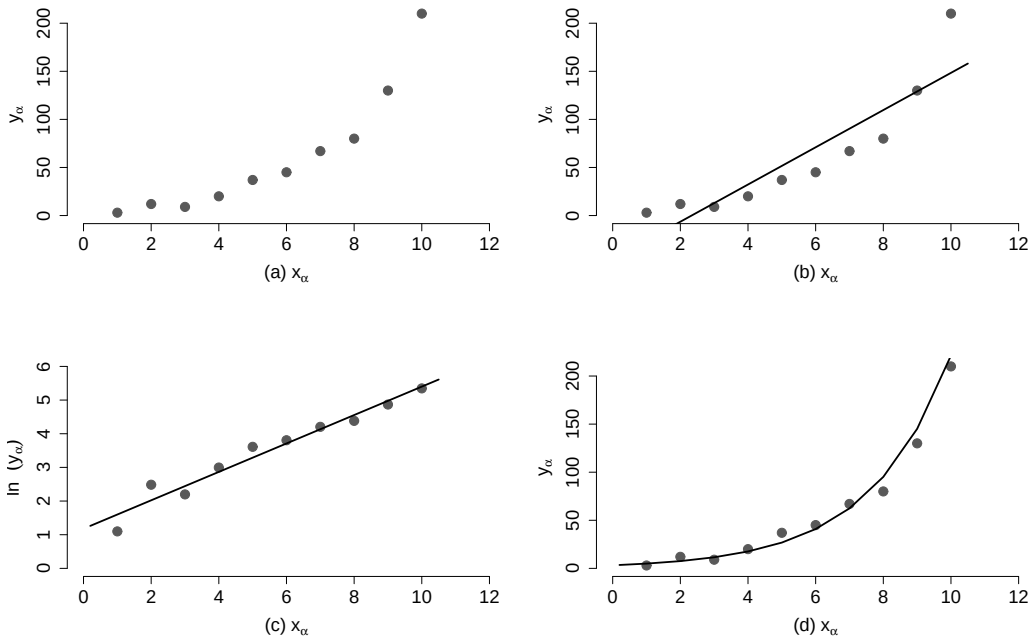


Figura 9.9 Modelli interpolanti a confronto, esempio 9.5.

il cui diagramma a dispersione è riportato in figura (9.9, pannello a). Da una sua lettura appare chiaro come l'interpolazione dei dieci punti mediante una retta non riesca a cogliere il legame funzionale tra Y ed X. A tal riguardo si osservi il grafico di figura (9.9, pannello b) che riporta la retta interpolante ai minimi quadrati. Appare lecito supporre un legame funzionale di tipo esponenziale, cioè considerare il modello $\hat{Y} = a\, b^X$. Al fine di individuare, ricorrendo al metodo dei minimi quadrati, i valori dei parametri a e b del modello prescelto ricorriamo alla trasformata semilogratitmica, cioè consideriamo il modello lineare $Z = a_0 + a_1\, X$, dove $Z = log(Y)$, $a_0 = log(a)$ e $a_1 = log(b)$.

A questo punto sarà

$$a_1 = \frac{Cov\,[X,Z]}{V[X]} = \frac{3.483}{8.250} = 0.422$$
$$a_0 = E[Z] - a_1\,E[X] = 3.500 - 0.422 \cdot 5.500 = 1.177$$

per cui $\hat{Z} = 1.177 + 0.422\,X$, il cui grafico è illustrato in figura (9.9, pannello c).
In definitiva avremo $a = exp(a_0) = 3.246$ e $b = exp(a_1) = 1.525$, sicché $\hat{Y} = 3.246 \cdot 1.525^X$, il cui grafico è riportato in figura (9.9, pannello d).

◁

9.5. ALCUNE UTILI DIMOSTRAZIONI

Dedichiamo questo paragrafo alla dimostrazione delle due proprietà che abbiamo sfruttato ai fini della costruzione del coefficiente di determinazione R^2.
A tal fine premettiamo alcune proprietà dei residui di regressione che ci consentiranno di alleggerire i successivi passaggi.
Riprendendo il sistema delle equazioni normali dato in (9.5) e dividendo ambo i membri di entrambe le equazioni per n, otteniamo il sistema equivalente

$$\begin{cases} \dfrac{1}{n}\displaystyle\sum_{\alpha=1}^{n} (\tilde{y}_\alpha - \hat{y}_\alpha) & = 0 \\ \dfrac{1}{n}\displaystyle\sum_{\alpha=1}^{n} (\tilde{y}_\alpha - \hat{y}_\alpha)\;\tilde{x}_\alpha & = 0 \end{cases} \tag{9.14}$$

Da questo possiamo ora dimostrare senza fatica le seguenti proprietà dei residui dalla retta di regressione:

- ⋆ il valor medio dei residui di regressione è nullo, cioè $E[Y - \hat{Y}] = 0$ e ciò in virtù della prima equazione del sistema (9.14).
- ⋆ i residui risultano ortogonali ai valori teorici, cioè $E[(Y - \hat{Y})\,\hat{Y}] = 0$. La dimostrazione segue direttamente dal sistema di equazioni (9.14), osservando infatti che

$$\begin{aligned} E[(Y - \hat{Y})\,\hat{Y}] &= E[(Y - \hat{Y})\,(a_0 + a_1\,X)] = \\ &= E[(Y - \hat{Y})\,a_0] + E[(Y - \hat{Y})\,a_1\,X)] = \\ &= a_0\,E[Y - \hat{Y}] + a_1\,E[(Y - \hat{Y})\,X] \end{aligned}$$

sarà $E[(Y - \hat{Y})\,\hat{Y}] = 0$ in quanto somma delle due equazioni del sistema di equazioni normali moltiplicate rispettivamente per le costanti a_0 e per a_1.

- $\star$ la covarianza tra i residui e i valori teorici è nulla, cioè $Cov[(Y - \hat{Y}), \hat{Y}] = 0$, e ciò in virtù delle due precedenti proprietà.

Scomposizione della varianza

Ci proponiamo di dimostrare la proprietà (9.1) introdotta al paragrafo precedente, ovvero che la varianza della componente Y di una v.s. doppia (X, Y) può essere scomposta nella somma della varianza dei residui di regressione e della varianza spiegata dal modello di regressione, cioè

$$V\,[Y] = V\left[Y - \hat{Y}\right] + V\left[\hat{Y} - E[Y]\right] \tag{9.15}$$

Ricordando che $Y - E[Y] = (Y - \hat{Y}) + (\hat{Y} - E[Y])$, e che $V[Y] = E\,[(Y - E[Y])^2]$ sarà

$$\begin{aligned} V[Y] &= E\left[\left((Y - \hat{Y}) + (\hat{Y} - E[Y])\right)^2\right] = \\ &= E\left[\left(Y - \hat{Y}\right)^2\right] + E\left[\left(\hat{Y} - E[Y]\right)^2\right] + 2E\left[\left(Y - \hat{Y}\right)\,\left(\hat{Y} - E[Y]\right)\right] = \\ &= V\left[Y - \hat{Y}\right] + V\left[\hat{Y} - E[Y]\right] \end{aligned}$$

poiché il doppio prodotto è nullo, infatti

$$E\left[(Y - \hat{Y})\,(\hat{Y} - E[Y])\right] = E\left[(Y - \hat{Y})\,\hat{Y}\right] - E\left[(Y - \hat{Y})\,E[Y]\right] = 0$$

in virtù delle precedenti proprietà.

Varianza dei residui dalla retta di regressione

Ci proponiamo di dimostrare la proprietà (9.2) del paragrafo precedente, ovvero che la varianza dei residui dalla retta di regressione è proporzionale alla varianza di Y tramite il fattore $1 - \rho^2$, cioè

$$V[Y - \hat{Y}] = V[Y]\,(1 - \rho^2)$$

A tal fine osserviamo inizialmente che si tratta della varianza di una combinazione lineare di v.s., pertanto

$$V[Y - \hat{Y}] = V[Y] + V[\hat{Y}] - 2\,Cov[Y, \hat{Y}]$$

Poiché si ha

$$V[\hat{Y}] = V[a_0 + a_1\,X] = a_1^2\,V[X]$$

e ancora

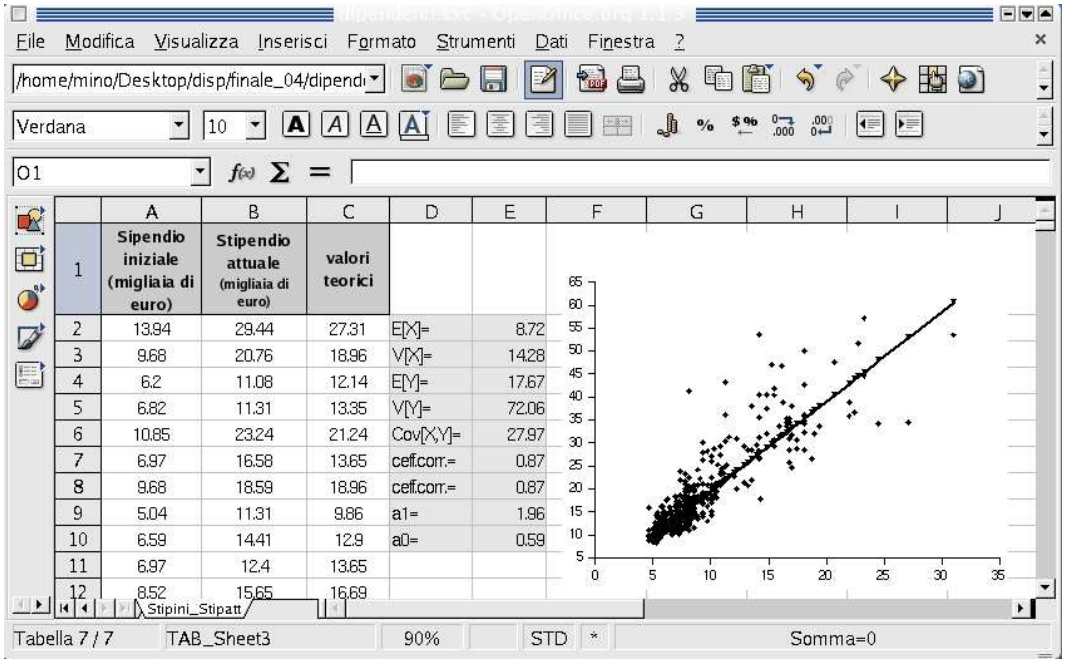
$$\begin{aligned} Cov[Y, \hat{Y}] &= E[Y \cdot \hat{Y}] - E[Y]\,E[\hat{Y}] = E[Y\,(a_0 + a_1\,X)] - E[Y]^2 = \\ &= a_0\,E[Y] + a_1\,E[X \cdot Y] - E[Y]^2 = \\ &= E[Y]^2 - a_1\,E[X]\,E[Y] + a_1\,E[X \cdot Y] - E[Y]^2 = \\ &= a_1\,Cov[X, Y] \end{aligned}$$

allora

$$\begin{aligned} V[Y - \hat{Y}] &= V[Y] + a_1^2\,V[X] - 2\,a_1\,Cov[X, Y] = \\ &= V[Y] + \frac{Cov[X, Y]^2}{V[X]} - 2\,\frac{Cov[X, Y]^2}{V[X]} = \\ &= V[Y] - \frac{Cov[X, Y]^2}{V[X]}\,\frac{V[Y]}{V[Y]} = V[Y]\,(1 - \rho^2) \end{aligned}$$

9.6. IL FOGLIO ELETTRONICO

Consideriamo la v.s. doppia $(X, Y) = \{$*Stipendio iniziale, Stipendio attuale*$\}$ del consueto file `dipendenti.sxc` e costruiamo il diagramma a dispersione così come appare in figura (9.10).
Per determinare i parametri della retta di regressione calcoliamo: $E[X]$, $V[X]$, $E[Y]$ nonché $Cov[X, Y]$ inserendo nella cella `E2` la funzione `=MEDIA(A2:A474)`, nella cella `E3` la funzione `=VAR.POP(A2:A474)`, nella cella `E4` la funzione `=MEDIA(B2:B474)` e nella cella `E6` la funzione `COV(A2:A474;B2:B474)` che restituisce il valore della covarianza.
Per a_1 sara sufficiente, a questo punto, inserire nella cella `E9` la formula `=E6/E3`, mentre per a_0 poniamo nella cella `E10` la formula `=E4-E9*E2`.
Volendo aggiungere al diagramma a dispersione la retta di regressione calcoliamo i valori teorici che poniamo nell'intervallo di celle `c2:c474` inserendo ad esempio in `C2` la



	A	B	C	D	E
1	Sipendio iniziale (migliaia di euro)	Stipendio attuale (migliaia di euro)	valori teorici		
2	13.94	29.44	27.31	E[X]=	8.72
3	9.68	20.76	18.96	V[X]=	14.28
4	6.2	11.08	12.14	E[Y]=	17.67
5	6.82	11.31	13.35	V[Y]=	72.06
6	10.85	23.24	21.24	Cov[X,Y]=	27.97
7	6.97	16.58	13.65	ceff.corr.=	0.87
8	9.68	18.59	18.96	ceff.corr.=	0.87
9	5.04	11.31	9.86	a1=	1.96
10	6.59	14.41	12.9	a0=	0.59
11	6.97	12.4	13.65		
12	8.52	15.65	16.69		

Figura 9.10 Retta di regressione residui di regressione, esempio 9.4.

formula `=$E$10+$E$9*A2`. Potremmo così aggiungere al grafico a dispersione `la serie di valori` nelle celle `c2:c474` scegliendo di congiungerli con una linea. Concludiamo il paragrafo osservando che nelle celle `E7` ed `E8` compare il valore del coefficiente di correlazione lineare ρ. In `E7` abbiamo inserito la formula `=E6/radq(E3*E5)` calcolando ρ a partire dalla sua definizione

$$\rho = \frac{Cov[X,Y]}{\sqrt{V[X]\,V[Y]}}$$

Mentre nella cella `E8` abbiamo usato la funzione predefinita di OpenOffice che restituisce il coefficiente di correlazione lineare `=CORRELAZIONE(A2:A474;B2:B474)`.

9.7. Esercizi

▷ Esercizio 9.1

Com'è noto la concentrazione di ozono (Y), espressa in mg per m^3, nelle grandi aree metropolitane dipende, oltre che da altri fattori, in gran misura dalla temperatura

ambientale (X). I dati che seguono riguardano $n = 25$ misurazioni registrate da una centralina di controllo in altrettanti giorni feriali dei mesi di giugno e luglio alle ore 12.00 nei pressi di un'importante crocevia

$$\begin{aligned}\{(\tilde{x}_\alpha;\tilde{y}_\alpha)\}_{\alpha=1,\ldots,25} =\{&(22;9),(24;23),(25;21),(25;22),(26;31),\\ &(26;44),(26;45),(26;59),(27;9),(27;39),\\ &(27;65),(27;88),(28;16),(28;28),(29;35)\\ &(30;44),(30;73),(30;78),(31;66),(32;89)\\ &(32;110),(32;122),(34;85),(34;118),(36;96)\}\end{aligned}$$

Si proceda a:

- ⋆ costruire un diagramma a dispersione per le coppie di valori osservati;
- ⋆ calcolare i parametri della retta di regressione $\hat{Y} = a_0 + a_1\, X$;
- ⋆ calcolare la varianza dei residui di regressione nonché il coefficiente di determinazione del modello.

◁

▷ ESERCIZIO 9.2

La rilevazione del numero di dipendenti (X) e del fatturato giornaliero (Y), su un collettvo statistico costituito da 70 esercizi pubblici ha dato luogo alla seguente distribuzione di frequenze congiunte:

$X\downarrow$ $Y\rightarrow$	$200 \dashv 400$	$400 \dashv 800$	$800 \dashv 1000$	$1000 \dashv 2000$
1	10	5	2	0
2	4	12	2	1
3	1	2	11	3
4	0	1	6	10

Si proceda a:

- ⋆ calcolare i parametri della retta di regressione $\hat{Y} = a_0 + a_1\, X$;
- ⋆ calcolare la varianza dei residui di regressione nonché il coefficiente di determinazione del modello.

◁

▷ ESERCIZIO 9.3

Si sono eseguite 24 misurazioni della temperatura (Y) dell'acqua di un lago percorso da correnti a diverse profondità (X) ottenendo la seguente distribuzione di frequenze congiunte:

$$\begin{aligned}\{(\tilde{x}_\alpha;\tilde{y}_\alpha)\}_{\alpha=1,\dots,24} =&\{(5;21.5),(5;20.5),(5;20.7),(10;20.3),(10;21.1),\\ &(10;20.44),(15;19.2),(15;18.7),(15;19.5),(20;17.5),\\ &(20;16.5),(20;17.2),(25;16.5),(25;16.2),(25;15.5)\\ &(30;14.8),(30;14.5),(30;13.8),(35;13.5),(35;13.1)\\ &(35;13.7),(40;12.2),(40;11.5),(40;11.8)\}\end{aligned}$$

Si proceda a:

- ⋆ costruire un diagramma a dispersione per le coppie di valori osservati;
- ⋆ calcolare i parametri della retta di regressione $\hat{Y} = a_0 + a_1\, X$;
- ⋆ calcolare la varianza dei residui di regressione nonché il coefficiente di determinazione del modello.

◁

▷ ESERCIZIO 9.4

Rispondere agli stessi quesiti di cui all'esercizio 9.3, dopo ave costruito la distribuzione di frequenze congiunte della variabile statistica (X, Y).

◁

▷ ESERCIZIO 9.5

Sia (X, Y) una variabile statistica bivariata con insieme di dati individuali

$$\{\tilde{x}_\alpha;\tilde{y}_\alpha\}_{\alpha=1,\dots,5} = \{(10;0.32),(11;0.39),(15;0.70),(16;0.90),(18;1.29)\}$$

Individuare i parametri di un modello ai minimi quadrati che ben si adatti ai dati individuali.

◁

▷ ESERCIZIO 9.6

Sia (X, Y) una variabile statistica bivariata con insieme di dati individuali

$$\{\tilde{x}_\alpha; \tilde{y}_\alpha\}_{\alpha=1,\dots,8} = \{(1; 6.9), (2; 8.0), (3; 8.6), (4; 11.0), (7; 18.5), \\ (8; 25.0), (10; 39.0), (12; 60.0)\}$$

individuare i parametri di un modello interpolante a minimi quadrati che ben si adatti ai dati individuali.

◁

▷ ESERCIZIO 9.7

Da un censimento di imprese artigiane del settore manifatturiero si è rilevata la seguente v.s. (X, Y) dove:

$X = \{$*tributi versati nell'anno 2004*$\}$ in euro
$Y = \{$*volume degli affari nell'anno 2004*$\}$ in migliaia di euro.

con distribuzione congiunta di frequenze assolute:

$X \downarrow \quad Y \rightarrow$	$0 \dashv 100$	$100 \dashv 200$	$200 \dashv 300$	$300 \dashv 500$
$0 \dashv 30$	8	2	0	0
$30 \dashv 60$	3	20	9	1
$60 \dashv 90$	1	25	55	37
$90 \dashv 150$	0	3	7	10

Si proceda a:

- ⋆ calcolare i parametri della retta di regressione $\hat{Y} = a_0 + a_1 X$;
- ⋆ calcolare la varianza dei residui di regressione nonché il coefficiente di determinazione del modello.

◁

Bibliografia

Fabbris L. (1997) *Statistica Multivariata: Analisi Esplorativa dei Dati*, MacGraw Hill, Milano.

Frosini B. (1990) *Lezioni di Statistica*, Vita e Pensiero, Milano.

Frosini B. (1995) *Introduzione alla Statistica*, Nuova Italia Scientifica, Roma.

Jalla E. (1989) *Principi di Statistica Teorica*, Giappichelli Editore, Torino.

Jalla E. (1991) *Appunti di Statistica*, Giappichelli Editore, Torino.

Landenna G. (1997) *Fondamenti di Statistica Descrittiva*, Il Mulino, Bologna.

Leti G. (1983) *Statistica Descrittiva*, Il Mulino, Bologna.

Mignani S. and Montanari A. (2001) *Appunti di Analisi Statistica Multivariata*, Società Editrice Esculapio, Bologna.

Naddeo A. (1972) *Appunti delle Lezioni di Statistica*, CLEUC, Venezia.

Parpinel F. and Provasi C. (2004) *Elementi di Probabilità e Statistica per le Scienze Economiche*, Giappichelli Editore, Torino.

Piccolo D. (2000) *Statistica*, Il Mulino, Bologna.

Vitali O. (1997) *Statistica per le Scienze Applicate*, volume 2, Cacucci, Bari.