ro debitori e li ringraziamo pubblicamente. Vanno inoltre ringraziati gli studenti che hanno sperimentato gli esercizi ed i testi che sono poi diventati questo volume. A loro va tutta la nostra riconoscenza.

Pur essendo il testo frutto di una lunga collaborazione e dell'impegno comune il dott. Francesco Paolo Borazzo ha elaborato la parte relativa agli esercizi con particolare attenzione all'utilizzo delle funzioni Excel mentre la dott.ssa Paola Perchinunno ha svolto la parte teorica e metodologica.

Gli autori saranno lieti di ricevere consigli e suggerimenti, critiche e correzioni agli indirizzi e-mail: francesco.borazzo@unito.it, p.perchinunno@dss.uniba.it

1 Statistica descrittiva univariata

1.1 Introduzione alla statistica

La *statistica* è un insieme di metodologie e tecniche per la raccolta, la presentazione e la definizione di informazioni allo scopo di agevolare l'analisi dei dati nei processi decisionali e più in generale nei processi interpretativi della realtà.

Usualmente la statistica si distingue in *statistica descrittiva*, costituita dall'insieme dei metodi deduttivi utilizzati per la raccolta, la definizione e la presentazione di un insieme di dati che rappresentano la totalità dei casi presi in considerazione allo scopo di descriverne le caratteristiche principali, e *statistica inferenziale*, costituita dall'insieme dei metodi induttivi che consentono di ottenere informazioni sulla popolazione a partire dai dati riguardanti solo una parte della popolazione.

Nella prima definizione si è impiegato il termine *deduttivo* per indicare il fatto che i dati vengono descritti e ricavati a partire dall'insieme scelto e cioè vengono dedotti dall'insieme. Per contro nella seconda si è utilizzato il termine *induttivo* per indicare come le informazioni si ottengono per induzione, ovvero con un processo di generalizzazione delle conclusioni, dall'osservazione di un gruppo ristretto si estendono all'insieme di cui il gruppo fa parte.

La prima parte di questo testo riguarda la statistica descrittiva, ovvero analizza i principali metodi e le tecniche in grado di riassumere le informazioni contenute in una tabella di dati mediante semplici indicatori. Utilizzando percentuali, medie, scarti quadratici medi e altri coefficienti si è in grado di esprimere l'andamento e la variabilità del fenomeno osservato.

La seconda parte del testo, viceversa, analizza l'insieme dei metodi di statistica inferenziale che consentono di derivare la stima di una caratteristica della popolazione basandosi esclusivamente sui dati di un campione estratto da essa. In questo caso viene estratto dalla popolazione, secondo alcune tecniche di campionamento, un sottoinsieme di elementi che costituiranno il campione oggetto di studio. Le conclusioni che si potranno trarre sul campione, verranno poi estese, con una certa probabilità, all'intera popolazione.

1.2 Rilevazioni statistiche

La statistica, al pari di qualunque altra disciplina scientifica, utilizza un glossario specialistico basato su un ristretto numero di termini le cui definizioni verranno fornite nel resto del paragrafo.

La *rilevazione statistica* è costituita dal complesso di operazioni rivolte ad acquisire una o più informazioni su un fenomeno empirico oggetto di studio. Una rilevazione statistica può essere:

globale: se si riferisce ad indagini riguardanti una popolazione presa nel suo complesso, come avviene nel caso dei censimenti che ogni dieci anni fotografano la situazione demografica dell'intera nazione.
parziale: se si riferisce a indagini effettuate soltanto su una piccola parte estratta dalla popolazione; si parla in questo caso di *indagini campionarie*.

Per *popolazione* o *collettivo statistico* intendiamo la totalità dei casi su cui si manifesta il fenomeno oggetto di analisi (per esempio la popolazione residente a Roma, gli esercizi commerciali di Bologna, i passeggeri di un aereo, il numero di lupi presenti in un parco alpino, ecc.).

L'*unità statistica* è il singolo elemento della popolazione: l'individuo nel caso di popolazioni umane, il negozio se si stanno considerando gli esercizi commerciali e così via.

Per *carattere* intendiamo un particolare aspetto, rilevato o misurato sulle unità statistiche, che sintetizza il fenomeno oggetto di studio. A seconda del fenomeno osservato i caratteri potranno assumere nature differenti: statura, sesso, età, titolo di studio, fatturato, colore dell'auto, concentrazione di un inquinante, ecc.

Per *modalità* intendiamo il modo in cui i caratteri si manifestano nelle singole unità statistiche, ovvero il numero (per caratteri quantitativi) o l'attributo (per caratteri qualitativi) che l'unità statistica manifesta (per esempio maschio o femmina, 20, 21 o 22 anni, ecc.).

Chiariamo con alcuni esempi queste prime definizioni.

Esempio 1.1

Poiché la statistica trova applicazione in quasi tutte le discipline esistenti, le popolazioni oggetto di studio possono essere di tipo più vario: di seguito vengono presentati alcuni esempi di popolazioni:

- i residenti di un comune o di un'area geografica
- gli abitanti di una nazione o di un continente
- le donne con figli residenti in un quartiere
- i giovani tra i dodici ed i diciotto anni di un comune
- gli anziani oltre i 75 anni di età
- i batteri di una coltura

- i malati affetti da una particolare patologia ricoverati in ospedale
- le mucche da latte presenti in una zona geografica
- le aziende che nascono o falliscono in un quadrimestre
- le fotocopiatrici presenti in una azienda
- i motori che escono dalla linea di produzione alla settimana
- le bottiglie di vino prodotte in un anno
- i brani musicali trasmessi in un mese da MTV
- gli appartamenti sfitti in un comune
- gli evasori fiscali

□

Esempio 1.2

La tabella seguente (Tabella 1.1) elenca oltre ad alcuni esempi di popolazioni oggetto di indagine, i caratteri e le modalità rilevate:

Tabella 1.1 Caratteri e modalità per alcuni tipi di popolazioni.

Popolazione	Caratteri misurati	Modalità con cui possono presentarsi
Alberghi della riviera ligure		
	Categoria	Numero di stelle (1, 2, 3, 4, ...)
	Posti letto	Numero (20, 50, 100, ...)
	Presenza ristorante	Si/No
Clienti di un ipermercato		
	Età	Numero (15, ..., 20, ..., 25, ...)
	Sesso	M o F
	Titolo di studio	1: nessuno 2: licenza elementare o media 3: diploma 4: laurea
	Reddito	1: meno di 15.000 euro 2: tra 15.001 e 20.000 3: tra 20.001 e 25.000 4: oltre 25.000
	Gradimento di un prodotto	1: non gradito 2: sufficiente 3: gradito

(continua)

Tabella 1.1 *(segue)*

Popolazione	Caratteri misurati	Modalità con cui si presentano
Azioni Wal Mart alla borsa di New York dal 1 gennaio al 31 dicembre 2006		
	Prezzo iniziale (open)	Numero (dollari)
	Prezzo più alto (high)	Numero (dollari)
	Prezzo più basso (low)	Numero (dollari)
	Prezzo finale (close)	Numero (dollari)

□

Un qualsiasi carattere può assumere modalità differenti a seconda delle diverse unità statistiche del collettivo. Un carattere statistico si definisce:

- *quantitativo* se le modalità assunte hanno forma numerica, come conteggio o misura: ne sono un esempio l'età, il peso o la statura degli individui, ma anche il fatturato in euro di un'azienda, il numero di telefonate effettuate, il numero di posti letto, ecc.
- *qualitativo* se le modalità assunte sono espresse come attributo come nel caso del sesso, nazionalità, titolo di studio, giudizio.

Un carattere *quantitativo* può essere distinto in:

- *discreto* se può assumere un numero intero di possibili modalità
- *continuo* se può prendere qualunque valore in un intervallo reale: si avranno così infinite possibili modalità.

I dati quantitativi sono espressi in termini numerici discreti (cioè sono espressi solo da numeri naturali, 0, 1, 2, 3, 4, ecc) quando derivano da un processo di *conteggio* o di *enumerazione*. Il numero dei nati vivi in un certo mese, il numero di figli, il numero di presenti, il numero di globuli rossi per mm^3 di sangue sono tutti esempi di conteggi che hanno come risposta un numero intero, maggiore o tutt'al più pari a zero.

I dati quantitativi sono espressi in termini numerici continui quando derivano, in maniera diretta o indiretta, da un processo di *misurazione*, cioè sono letti con l'ausilio di uno strumento di misura. L'altezza di una persona è ricavata leggendo il valore su un metro, il peso utilizzando una bilancia, la corrente elettrica che percorre un filo con un voltmetro, la durata delle telefonate in uscita da un cellulare con un cronometro. Sono tutti esempi di misure dirette. D'altra parte, l'area di una superficie, il volume di un corpo, una densità, si ricavano in maniera indiretta a partire sempre da una misurazione: l'area di un rettangolo si ricava misurando i lati e moltiplicandoli tra loro.

Osserviamo che enumerare o contare sono un processo di calcolo, mentre utilizzare uno strumento di misura implica un processo di misurazione. Il valore di una misura dipende dallo strumento utilizzato e può essere espresso con maggiore o minore precisione. Se si misura l'altezza di una persona con un metro a nastro come quello usato dai sarti, si ottiene una misura approssimata rispetto a quella ottenuta con uno strumento utilizzato in un ambulatorio medico.

Un carattere *qualitativo* (o *mutabile statistica*) può essere distinto in:

- *sconnesso* se le modalità che assume non presentano un ordine naturale di successione
- *rettilineo* se le modalità che assume sono ordinabili in base ad un ordine naturale e logico, con una modalità iniziale ed una finale
- *ciclico* se le modalità che assume sono ordinabili in base ad un ordine logico ma non è possibile definire una modalità iniziale ad una finale.

Sono qualitativi sconnessi i dati espressi attraverso attributi o categorie dove la presenza o meno di un ordine non è rilevante. Per esempio, la provincia di nascita, il tipo di diploma di maturità, il colore dell'auto, sono esempi di categorie le cui modalità sono prive di alcun ordine prefissato. Questo tipo di caratteri vengono detti anche *nominali*.

Per contro sono qualitativi rettilinei gli attributi che possiedono un ordine naturale come le categorie di un albergo (bed & breakfast, lusso, extralusso, ecc.), il livello di soddisfazione del cliente (buono, sufficiente, ottimo), il titolo di studio (elementare, media, superiore, laurea). Questo tipo di caratteri sono detti anche *ordinali*.

Infine, sono qualitativi ciclici quei fenomeni che si ripetono in maniera continua come i giorni della settimana, i mesi dell'anno, le direzioni dei venti, ecc.

Esempio 1.3

Vengono elencate di seguito alcune modalità suddivise secondo la classificazione appena introdotta:

- *Carattere quantitativo discreto*: numero di figli, voto d'esame o di laurea, numero di pezzi venduti o prodotti, numero di passeggeri, numero di personal computer, stanze di un appartamento, abitanti di un luogo, costo di un bene;
- *Carattere quantitativo continuo*: peso, altezza, temperatura, distanza, durata;
- *Carattere qualitativo sconnesso:* sesso, stato civile, colore degli occhi;
- *Carattere qualitativo rettilineo:* gradi militari, titolo di studio, categorie di un albergo, voto nella scuola media, gradimento di un servizio;
- *Carattere qualitativo ciclico:* mesi dell'anno, giorni della settimana, stagioni, direzione del vento.

□

Concludiamo dicendo che le misure descrittive di un campione prendono il nome di *statistiche* mentre le misure descrittive di un popolazione prendono il nome di *parametri*.

La tabella seguente riassume le definizioni introdotte:

Tabella 1.2 Quadro riassuntivo dei termini e delle definizioni introdotte.

Rilevazione statistica	Le operazioni per acquisire informazioni su un fenomeno empirico. Si distingue tra globale e parziale.
Indagine campionaria	Rilevazione statistica parziale, eseguita cioè su campione estratto dalla popolazione.
Popolazione, Universo, Collettivo statistico	Totalità degli elementi presi in esame
Campione	Parte della popolazione o universo selezionata per l'analisi
Unità statistica	Ogni singolo elemento appartenente al campione o alla popolazione
Carattere	Caratteristica misurata o rilevata sull'unità statistica
Modalità	Modo in cui si manifesta un carattere. Può essere numerico (per caratteri numerici) o un attributo (per caratteri qualitativi)
Variabile statistica	Un carattere che assume per modalità dei numeri reali
Mutabile statistica	Un carattere che assume per modalità degli attributi
Parametro	Misura riassuntiva che descrive una caratteristica dell'intera popolazione
Statistica	Misura riassuntiva che descrive una caratteristica del campione

Esempio 1.4

In una indagine condotta sui 200 studenti iscritti al II anno della Facoltà di Economia dell'Università degli Studi di Cosenza si vuole analizzare la votazione ottenuta all'esame di Statistica. Intendiamo, quindi, per:

Popolazione: i 200 studenti iscritti al II anno
Unità statistica: ogni singolo studente
Carattere: la votazione ottenuta dallo studente all'esame di statistica
Modalità: il valore numerico ottenuto all'esame di statistica (18, 19…..30) □

1.3 Fonti statistiche

Elemento indispensabile per effettuare una indagine statistica sono le informazioni derivanti dalla osservazione e rilevazione dei fenomeni che ci circondano. La rilevazione dei dati può avvenire attraverso:

- indagini condotte a titolo personale;
- materiale raccolto ed elaborato da strutture di ricerca pubbliche o private.

Nel primo caso sarà necessario effettuare una indagine attraverso l'impiego di tecniche di rilevazione dei dati in genere somministrando un questionario. Nel secondo caso è possibile ricorrere a collezioni di dati già predisposti da enti o società esterne che si occupano della raccolta di informazioni a vario livello. Diventa però di indispensabile interesse la conoscenza della fonte di provenienza, poiché l'attendibilità dei dati è direttamente riconducibile all'autorevolezza e alla competenza dell'ente che li ha elaborati. Le *principali fonti* di dati presenti nel nostro paese sono rilevate dall'*Istituto Nazionale di Statistica (ISTAT,* www.istat.it) che istituzionalmente accentra tutta la rilevazione statistica ufficiale italiana e provvede alla diffusione del materiale raccolto sotto forma di pubblicazioni periodiche (annuali, mensili, ecc) od occasionali. Le principali indagini svolte periodicamente dall'Istat sono definite *Censimenti* e riguardano aspetti diversi della situazione del paese.

- *Censimento Generale della Popolazione e delle Abitazioni*: elabora, con cadenza decennale, dati riguardanti le caratteristiche strutturali della popolazione (popolazione residente, stranieri, famiglie, persone che vivono in convivenze, grado di istruzione e condizione professionale dei cittadini, età, sesso, stato civile, ecc.) e delle abitazioni (caratteristiche strutturali di edifici e abitazioni).
- *Censimento Generale dell'Industria e dei Servizi*: sempre con cadenza decennale, raccoglie ed elabora i dati riguardanti le caratteristiche strutturali del sistema economico del paese (Imprese, istituzioni pubbliche e non profit, unità locali e addetti, suddivisi per attività economica, classe di addetti e forma giuridica).
- *Censimento Generale dell'Agricoltura*: ogni dieci anni rileva i dati riguardanti le caratteristiche strutturali delle aziende agricole (nuove attività, colture biologiche, agriturismo, artigianato, nuove tecnologie, utilizzazione dei terreni, irrigazione, allevamenti, mezzi meccanici, forza lavoro e approcci al mercato).

Per ognuno di questi argomenti l'Istat ha preparato un apposito sito web dove è possibile scaricare in formato Excel tabelle di dati aggregati. Per esempio sul sito web del censimento della popolazione 2001 http://dawinci.istat.it/daWinci/jsp/MD/dawinciMD.jsp si possono scaricare una grande quantità di tavole numeriche classificate per tema scelto, per indicatori e suddivisioni territoriali.

A queste indagini censuarie si aggiungono alcune *indagini campionarie* a carattere periodico, tra le quali possiamo ricordare:

- *Indagine sulle forze di lavoro* che su base trimestrale rileva i dati riguardanti la condizione occupazionale del paese (occupati, disoccupati, in cerca di occupazione, ecc)
- *Indagine sui consumi delle famiglie*: ogni trimestre raccoglie i dati riguardanti il consumo, il reddito e il risparmio familiare

- *Indagine multiscopo*: periodicamente raccoglie notizie dettagliate su tematiche di interesse per le famiglie (salute, tempo libero, anziani, ecc).

Altro organismo pubblico presente nel nostro paese è il *Sistema Statistico Nazionale* (*SISTAN*, www.sistan.it) cui afferiscono le istituzioni, pubbliche e private, deputate alla raccolta, elaborazione e diffusione di dati di interesse per la collettività. Tra le *fonti ufficiali* hanno, poi, particolare rilevanza le fonti amministrative, ovvero i dati rilevati dagli uffici statistici interni agli organi amministrativi, quali le anagrafi, gli uffici statistici dei Comuni, Ministeri, Banca d'Italia, Unità Sanitarie Locali, Camere di Commercio e numerosi altri enti, che in modo istituzionale raccolgono e conservano dati. Tali uffici possiedono informazioni di rilevante interesse, consultabili su richiesta (gratuitamente oppure a pagamento), spesso però riservate ad utenti istituzionali per questioni di riservatezza. Altre fonti sono costituite dalle banche dati di *società private di ricerca* che elaborano indagini, sulla base di campioni selezionati, su settori specifici di interesse nazionale e che di norma vengono venduti a clienti privati.

A livello internazionale troviamo organismi pubblici quali l'*EUROSTAT* (www.eurostat.eu), organismo dell'Unione Europea che ha il compito di armonizzare la raccolta di dati ufficiali riguardanti le nazioni aderenti, le Nazioni Unite (*UN*, United Nation, www.un.org oppure www.onuitalia.it), l'*UNESCO* (United Nations Educational, Scientific and Cultural Organization), la *FAO* (Food and Agricolture Organization).

Ovviamente con l'avvento della società dell'informazione e di Internet è ormai possibile reperire in rete quasi qualunque tipo di dato, con l'avvertenza che la bontà dell'analisi che si vuole condurre dipende direttamente dall'affidabilità e dalla veridicità delle informazioni recuperate dalla Rete.

Esempio 1.5

Scaricare dal sito dell'Istat l'elenco delle province e dei comuni d'Italia.

Con un browser connettersi al sito Istat al seguente url:

http://www.istat.it/strumenti/definizioni/comuni/

Nell'elenco dei dati che compare sulla destra dello schermo cliccare su "Ripartizioni geografiche, province e regioni (codici e denominazioni)" come mostrato in Figura 1.1

Comparirà la videata di download simile a quella di Figura 1.2 (in realtà dipende dal browser utilizzato, gli autori hanno scaricato i dati con MS Explorer).

Dare un click su bottone "Salva sul disco" e indicare una cartella o directory di destinazione sul proprio computer. Al termine del download (durerà uno o due secondi con una linea ADSL, meno di un minuto con un modem tradizionale) si potrà caricare il file ricevuto in Excel ottenendo qualcosa di simile a quanto appare in Figura 1.3. Dopo aver scaricato la tabella della suddivisione in regioni e province proseguire l'operazione scaricando con lo stesso meccanismo la tabella "Elenco dei comuni italiani al 1° gen-

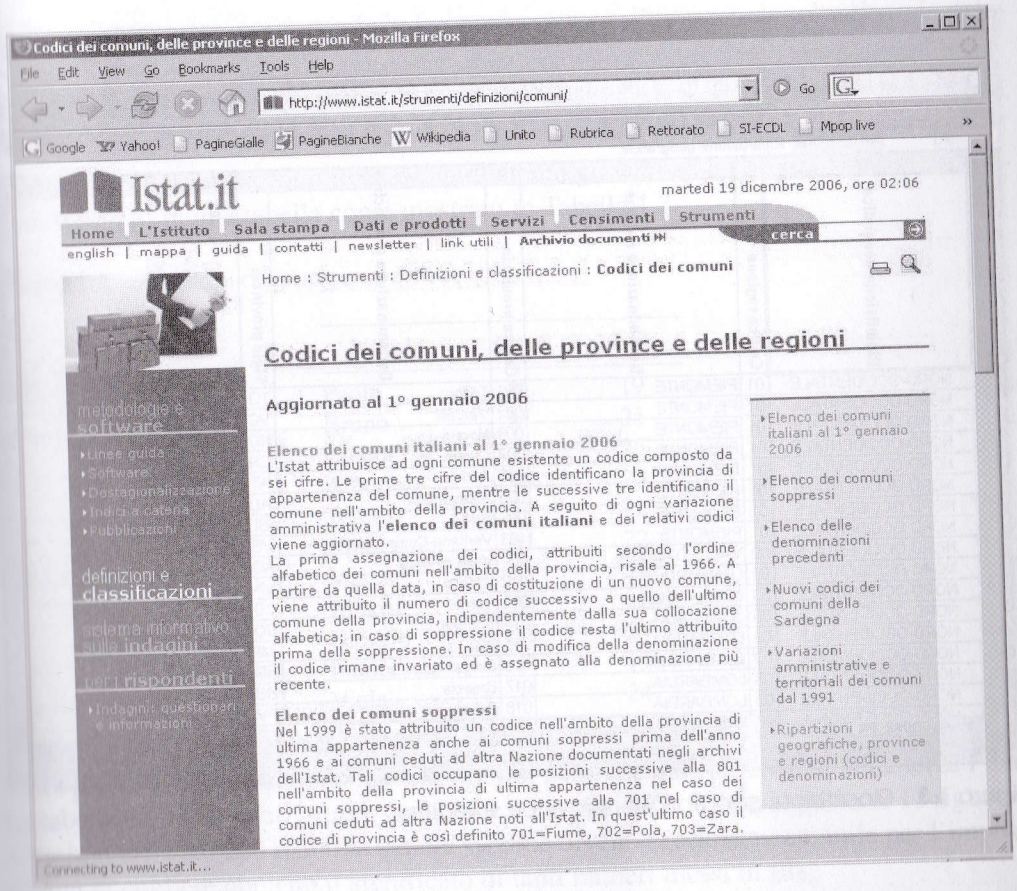
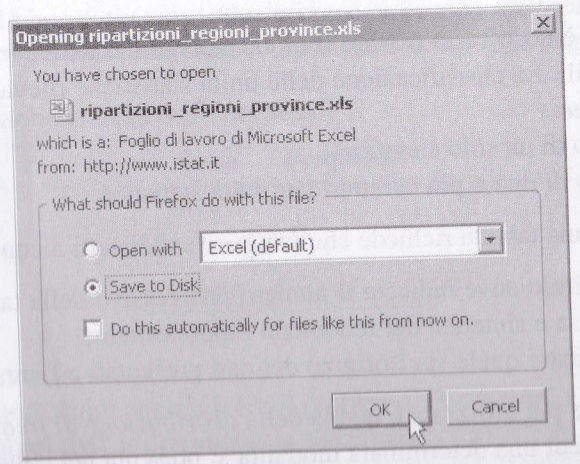
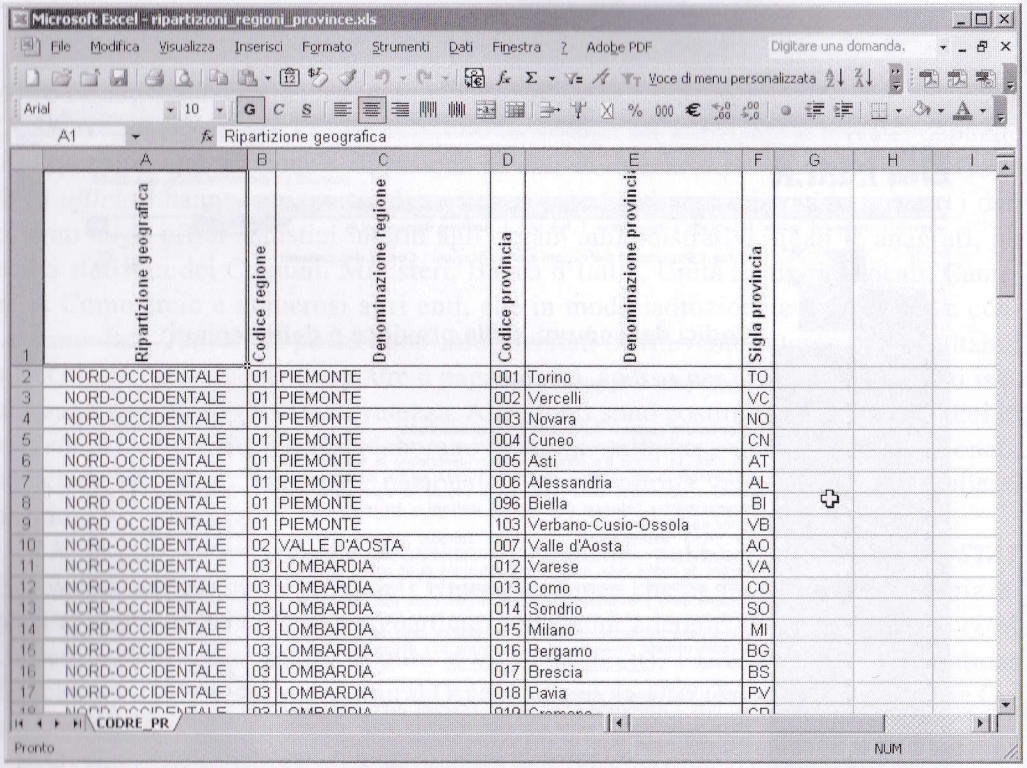


Figura 1.1 Una pagina del sito Istat.





	A	B	C	D	E	F
1	Ripartizione geografica	Codice regione	Denominazione regione	Codice provincia	Denominazione provincia	Sigla provincia
2	NORD-OCCIDENTALE	01	PIEMONTE	001	Torino	TO
3	NORD-OCCIDENTALE	01	PIEMONTE	002	Vercelli	VC
4	NORD-OCCIDENTALE	01	PIEMONTE	003	Novara	NO
5	NORD-OCCIDENTALE	01	PIEMONTE	004	Cuneo	CN
6	NORD-OCCIDENTALE	01	PIEMONTE	005	Asti	AT
7	NORD-OCCIDENTALE	01	PIEMONTE	006	Alessandria	AL
8	NORD OCCIDENTALE	01	PIEMONTE	096	Biella	BI
9	NORD OCCIDENTALE	01	PIEMONTE	103	Verbano-Cusio-Ossola	VB
10	NORD-OCCIDENTALE	02	VALLE D'AOSTA	007	Valle d'Aosta	AO
11	NORD-OCCIDENTALE	03	LOMBARDIA	012	Varese	VA
12	NORD-OCCIDENTALE	03	LOMBARDIA	013	Como	CO
13	NORD-OCCIDENTALE	03	LOMBARDIA	014	Sondrio	SO
14	NORD-OCCIDENTALE	03	LOMBARDIA	015	Milano	MI
15	NORD-OCCIDENTALE	03	LOMBARDIA	016	Bergamo	BG
16	NORD-OCCIDENTALE	03	LOMBARDIA	017	Brescia	BS
17	NORD-OCCIDENTALE	03	LOMBARDIA	018	Pavia	PV

Figura 1.3 Ripartizioni geografiche, province e regioni.

1.4 Organizzazione dei dati

Una volta raccolti è necessario trovare il modo di organizzare e convertire i dati grezzi in informazioni utili. La classificazione delle unità statistiche rilevate avviene in forma di *tabelle statistiche*:

- *semplici* nel caso di un solo carattere,
- *multiple* nel caso di due o più caratteri per unità statistica.

La costruzione di una tabella richiede che vengano soddisfatti alcuni requisiti:

- l'*intestazione* o titolo deve indicare il numero progressivo della tavola e contenere la descrizione precisa e sintetica del fenomeno rappresentato;
- la *fonte* deve indicare quale sia l'origine dei dati presentati e l'anno di riferimento.

La rappresentazione più utilizzata è quella della *distribuzione di frequenza* che indica il numero di volte in cui una determinata modalità si presenta nel collettivo in esame. La distribuzione si presenta sotto la forma di una tabella dove ad ogni valore che una variabile può assumere, corrisponde la *frequenza*, ovvero il numero di volte che tale valore si presenta.

Esempio 1.6

Supponiamo di aver raccolto le adesioni di alcuni giovani ad un gruppo sportivo e scriviamone i dati in una tabella come mostrato in Tabella 1.3.

Tabella 1.3 Adesioni al gruppo sportivo.

Nomi	Sesso	Età
Paolo	M	19
Bruno	M	25
Chiara	F	25
Alberto	M	23
Teresa	F	22
Andrea	M	22
Marco	M	22
Alberto	M	22
Giorgio	M	24
Francesca	F	19
Viola	F	24

Nella prima colonna vengono raccolti i nomi, nella seconda il genere e nella terza l'età. La tabella contiene i dati grezzi, così come sono stati raccolti. Se al posto di 11 righe ne avessimo elencate 200 o 300 la tabella sarebbe diventata scarsamente utile per l'impossibilità di cogliere appieno il significato di tanti numeri messi in fila.

Di norma si opera aggregando i dati, calcolandone le frequenze ovvero il numero di volte con cui si presentano le modalità. Nella Tabella 1.4 è stata calcolata la frequenza con cui si presentano gli 11 iscritti per sesso. Quella che si osserva è la *distribuzione di frequenza* che fornisce una certa informazione sulla composizione del gruppo di giovani.

Tabella 1.4 Distribuzione di frequenza per sesso.

Sesso	Frequenza
Maschi	7
Femmine	4
Totale:	*11*

Proposta di soluzione in Excel

In Excel esistono molti modi per risolvere il problema. Per esempio, per contare tutte le caselle che contengono "M" utilizzeremo la funzione:

```
=CONTA.SE(<intervallo>; <criteri>)
```

	A	B	C	D	E	F	G	H	I	J
1	Nomi	Sesso	Età							
2	Paolo	M	19							
3	Bruno	M	25		Sesso	Frequenza				
4	Chiara	F	25		Maschi	7	→	=CONTA.SE(B2:B12;"=M")		
5	Alberto	M	23		Femmine	4	→	=CONTA.SE(B2:B12;"=F")		
6	Teresa	F	22		*Totale:*	*11*				
7	Andrea	M	22							
8	Marco	M	22							
9	Alberto	M	22							
10	Giorgio	M	24							
11	Francesca	F	19							
12	Viola	F	24							
13										

Figura 1.4 Calcolo delle frequenze per sesso.

dove al posto del parametro `<intervallo>` andranno inserite le celle che contengono i dati, ovvero il range `B2:B12`, mentre al posto di `<criteri>` si dovrà inserire `"=M"` per ottenere il conteggio dei maschi e `"=F"` per il conteggio delle femmine.

In Figura 1.4 si può osservare il risultato con indicazione delle formule scritte nelle corrispondenti celle della colonna. □

Prima di illustrare con esempi più significativi quanto finora detto dobbiamo introdurre una rappresentazione più formale.

In una popolazione costituita da N elementi, una variabile statistica X può assumere k modalità distinte, diciamo $x_1, x_2, x_3, \ldots x_k$, in modo tale che la modalità x_1 si presenta n_1 volte, la modalità x_2 si presenta n_2 volte, fino alla modalità x_k che si presenta n_k volte. La frequenza *i-esima* è indicata con n_i e prende il nome di *frequenza assoluta*. Tali frequenze sono sempre numeri interi caratterizzati dal fatto che[1]:

$$n_1 + n_2 + \ldots + n_i + \ldots + n_k = \sum_{i=1}^{k} n_i = N$$

La *distribuzione di frequenza* associa, quindi, alle modalità che può assumere un carattere X, qualitativo o quantitativo, le corrispondenti frequenze assolute. La frequenza totale, indicata con N, è pari al totale delle osservazioni:

$$\sum_{i=1}^{k} n_i = N$$

Sotto forma di tabella, la distribuzione di frequenza della *variabile statistica X*, utilizzata nel caso di caratteri quantitativi, avrà il seguente aspetto:

[1] Si utilizza l'indice i che varia da 1 a k per indicare ognuno degli elementi del collettivo, così che x_1 sarà il valore della variabile riferito al primo elemento, x_2 sarà il valore del secondo elemento e così via.

Modalità della variabile X	Frequenze assolute n_i
x_1	n_1
x_2	n_2
....	....
x_i	n_i
....	....
x_k	n_k
	N

Nei casi reali, soprattutto quando il numero delle osservazioni è cospicuo, le modalità con cui la variabile di distribuisce possono essere numerose, creando così la necessità di ottenere tabelle ben più grandi di quella del precedente esempio. In questi casi si preferisce raggruppare i dati in *classi di frequenza* in modo da ridurne le modalità. Chiariamo questo aspetto con il seguente esempio.

Esempio 1.7

L'ufficio marketing di una catena di supermercati svolgendo un'indagine sulla soddisfazione del cliente, intervista in un dato giorno un campione di 152 clienti adulti. La Tabella 1.5 mostra le prime righe della tabella con i dati grezzi dei soggetti intervistati: difficilmente si potrebbero ricavare rapidamente utili considerazioni esaminando due o tre pagine di dati siffatti.

Tabella 1.5 Prime righe della tabella con dati grezzi.

Numero cliente	Età
1	39
2	30
3	25
4	33
5	21
6	26
7	25
8	23
9	22
10	45
il	29
12	26
13	47
.....	
152	30

Raggruppando invece le età degli intervistati in classi di ampiezza pari a 10 anni, si ottiene la Tabella 1.6. Questo tipo di rappresentazione è sicuramente più comoda rispetto ad un lungo elenco di dati difficile da leggere e interpretare a colpo d'occhio.

Tabella 1.6 Età degli intervistati in classi di ampiezza 10.

Età in classi	Età
18-25	3 2
26-35	49
36-45	21
46-55	28
56-65	17
66-75	5
Totale	152

Proposta di soluzione in Excel

Poiché si tratta di dati numerici discreti per lo svolgimento si può utilizzare la funzione `=FREQUENZA( )` che possiede il vantaggio di essere più flessibile e diretta di quella utilizzata nell'esempio precedente in cambio di qualche rigidità in più nel suo funzionamento. I valori oscillano tra 18 e 75 raggruppando i dati nelle sei classi.

Nella tabella bisogna inserire una colonna per indicare il valore massimo (estremo superiore) che la classe può assumere (ovvero 25, 35….75) come fatto nella colonna "D".

In seguito è necessario selezionare contemporaneamente le celle contigue `F2:F7` e scrivere la seguente formula badando di mantenere attiva la selezione:

```
=FREQUENZA(B2:B153;D2:D7)
```

Al termine non dare `Invio`, ma premere contemporaneamente[2] i tasti `Ctrl+Maiusc+Invio`. Il risultato dovrebbe apparire come in Figura 1.5. □

	A	B	C	D	E	F	G	H	I	J
1	**Numero cliente**	**Età**			**Età in classi**	**Età**				
2	1	39		25	18-25	32		<=FREQUENZA(B2:B153; D2:D7)		
3	2	30		35	26-35	49				
4	3	25		45	36-45	21				
5	4	33		55	46-55	28				
6	5	21		65	56-65	17				
7	6	26		75	66-75	5				
8	7	25			**Totale**	**152**				
9	8	23								
10	9	22								
11	10	45								
12	11	29								
13	12	26								
14	13	47								

Figura 1.5 L'inserimento della funzione Excel `=FREQUENZA( )`.

[2] Il termine contemporaneamente si usa per indicare che vanno premuti tre tasti anche in successione: mentre si tiene premuto il primo, premere gli altri due.

Data la distribuzione di dati grezzi occorre individuare il numero e l'ampiezza delle classi per poi contare le osservazioni che ricadono in ciascuna classe. La scelta del numero e dell'ampiezza delle classi è totalmente arbitraria ed in genere dipende dal tipo di problema che si sta analizzando. Premesso questo si possono individuare alcuni criteri utili per trovare rapidamente una soluzione.

Una classe è definita come l'insieme delle osservazioni che ricadono tra il limite inferiore e quello superiore; necessariamente uno dei due limiti, l'inferiore o il superiore, deve essere incluso nell'intervallo, in quanto le classi non si possono sovrapporre e nel contempo devono contenere tutti gli elementi della popolazione o del campione che si sta analizzando. Usando un linguaggio formale, dati i limiti a e b con la scritta

$$a \vdash b$$

si indica che il limite sinistro o inferiore, è compreso nell'intervallo[3] mentre quello destro o superiore, ne è escluso, ovvero che il valore x sarà:

$$a \leq x < b$$

In modo speculare con la scritta

$$a \dashv b$$

si indica questa volta che il limite destro è compreso nell'intervallo mentre quello sinistro ne è escluso[4], ovvero che il valore x sarà:

$$a < x \leq b$$

Il trattamento degli estremi così descritto ci permette di affermare che ogni elemento del collettivo preso in esame ricadrà in una ed una sola classe di frequenza[5]. Dopo aver suddiviso i dati nelle rispettive classi si può costruire la tabella indicando in una colonna i limiti inferiore e superiore delle classi e nella colonna adiacente il numero di casi che ricadono in ciascuna classe come mostrato nella seguente tabella:

Classi di modalità di X	Frequenze assolute n_i
$x_1 \vdash x_2$	n_1
$x_2 \vdash x_3$	n_2
….	….
$x_i \vdash x_{i+1}$	n_i
….	….
$x_{k-1} \vdash x_k$	n_k
	N

[3] Si può anche dire che l'intervallo è chiuso a sinistra e aperto a destra.

[4] Si può dire che l'intervallo è chiuso a destra e aperto a sinistra.

[5] Capita sovente leggendo i periodici, di osservare tabelle o grafici con diciture del tipo "da 26 a 35", "da 36 a 45". Tali forme sono tipiche del linguaggio comune, decisamente più approssimativo di quello formale. Infatti potremmo essere assaliti dal dubbio di dove collocare un'età di 35,5 anni. Il linguaggio comune ha il vantaggio di essere più semplice ed immediato ma può risultare ambiguo come in questo caso: in realtà lascia intendere che il valore 35,5 debba ricadere nella classe "da 26 a 35" poiché 35,5 anni non sono 36.

Anche nel caso di *mutabili statistiche X*, dove le modalità sono caratteri qualitativi o attributi, i dati possono essere raggruppati in classi. Si pensi di rappresentare il titolo di studio suddividendo i dati in sole tre classi: da nessun titolo a media inferiore compreso, diploma superiore, laurea o titoli superiori (master post laurea, specialità, dottorato, ecc.). In generale indicando le modalità con a_1, a_2, ...a_k e le frequenze con n_1, n_2, ... n_k si otterrà una tabella di distribuzione di frequenza analoga alla precedente.

Esempio 1.8

In una indagine per la valutazione del servizio vengono rilevati età, genere e tipo di servizio richiesto dei clienti che entrano nell'arco di 30 minuti in un ufficio postale. Nella Tabella 1.7 viene mostrata la tabella dei dati grezzi raccolti, dove la sigla B si riferisce al servizio relativo ai prodotti di Banco posta mentre con P si indicano i prodotti postali.

Tabella 1.7 Tabella di dati grezzi raccolti in un ufficio postale.

Cliente	Età	Genere	Servizio
1	24	M	B
2	18	M	B
3	65	F	B
4	21	F	P
5	67	F	B
6	65	M	B
7	34	F	B
8	71	F	B
9	23	M	P
10	45	F	P
11	71	F	B
12	65	M	B
13	52	F	B
14	63	F	P
15	64	F	B
16	35	M	P
17	30	M	P
18	22	F	B
19	21	F	B
20	56	M	B

Supponiamo di raggruppare le età in sole tre classi che rappresentano tre gruppi sociali: i *giovani* ovvero coloro che possiedono un età inferiore a 25 anni, gli *adulti* da 25 anni fino a 64 compreso ed infine gli *anziani* quelli di età superiore. In questo esempio vediamo come raggruppare i dati dell'età con una semplice procedura manuale ponendo dei simboli differenti a seconda della classe considerata. Sarà quindi necessario contare, contrassegnando tutte le caselle della tabella originale, le età che ricadono nella prima classe (giovani), poi quelle della classe successiva utilizzando un altro segno (barra, croce, cerchio, ecc.). Otterremo così una tabella come la seguente:

Tabella 1.8 Tabella per classi di età.

Classi di Età	Frequenza
Da 18 a 24	6
Da 25 a 65	8
Oltre 65	6
Totale	**20**

Ripetendo il medesimo procedimento anche per la mutabile statistica "Servizio richiesto" si otterrà facilmente la Tabella 1.9.

Tabella 1.9 Tabella per "Tipo di servizio richiesto".

Servizio richiesto	Frequenza
Banco posta	14
Prodotti postali	6
Totale	**20**

Proposta di soluzione in Excel

Per calcolare la frequenza delle età divise in classi è possibile usare la funzione `=FREQUENZA()` inserendo una colonna per indicare il valore massimo che la classe può assumere (24, 65, 71) come fatto nella colonna `F`. In seguito è necessario selezionare poi le celle contigue `H2:H4` e scrivere la seguente formula badando di mantenere attiva la selezione:

```
=FREQUENZA(B2:B21;F2:F4)
```

Al termine non dare `Invio`, ma premere contemporaneamente[6] i tasti `Ctrl+Maiusc+Invio`.

Viceversa per contare tutte le caselle che contengono "B" o"P" utilizzeremo la funzione Excel:

```
=CONTA.SE(<intervallo>; <criteri>)
```

dove al posto del parametro `<intervallo>` andranno inserite le celle che contengono i dati, ovvero il range `D2:D21`, mentre al posto di `<criteri>` si dovrà inserire `"=B"` per ottenere il conteggio dei banco posta e `"=P"` per il conteggio dei prodotti postali. □

Il risultato dovrebbe apparire come in Figura 1.6.

Occupiamoci ora del problema di determinare in quante classi suddividere le modalità ovvero di determinare l'ampiezza delle classi, fissando i limiti inferiori e superiori per

[6] Il termine contemporaneamente si usa per indicare che vanno premuti tre tasti anche in successione: mentre si tiene premuto il primo, premere gli altri due.

	A	B	C	D	E	F	G	H	I	J	K
1	Cliente	Età	Genere	Servizio			Classi di età	Frequenza			
2	1	24	M	B		24	Giovani (fino a 24 anni compresi)	6	<=FREQUENZA(B2:B21; F2:F4)		
3	2	18	M	B		65	Adulti (da 25 a 65 anni	11			
4	3	65	F	B		71	Anziani (oltre 65 anni)	3			
5	4	21	F	P			Totale:	20			
6	5	67	F	B							
7	6	65	M	B			Servizio richiesto	Frequenza			
8	7	34	F	B			Banco posta	14	<=CONTA.SE(D2:D21;"=B")		
9	8	71	F	B			Prodotti postali	6	<=CONTA.SE(D2:D21;"=P")		
10	9	23	M	P			Totale	20			
11	10	45	F	P							
12	11	71	F	B							
13	12	65	M	B							
14	13	52	F	B							
15	14	63	F	P							
16	15	64	F	B							
17	16	35	M	P							
18	17	30	M	P							
19	18	22	F	B							
20	19	21	F	B							
21	20	56	M	B							
22											

Figura 1.6 L'inserimento della funzione Excel FREQUENZA ( ) e CONTA.SE.

ognuna di esse. Come detto all'inizio del paragrafo, la scelta del numero e dell'ampiezza delle classi è totalmente arbitraria ed in genere dipende dal tipo di problema che si sta analizzando. Individuiamo alcuni criteri utili per trovare rapidamente una soluzione.

Un primo criterio è quello di utilizzare di norma non meno di 5 e non più di 15 classi. Se, per esempio si stanno analizzando le età di una cinquantina di soggetti che vanno da 20 a 65 anni può essere utile considerare 5 classi di ampiezza pari a 10 anni (da 20 compreso a 30 escluso, da 30 compreso a 40 escluso, ecc.). Ma se si stessero analizzando le età dei giovani frequentatori di un fast food, classi di 10 anni rischierebbero di risultare vuote. Occorre sottolineare che il numero delle classi e la loro dimensione dipende in primo luogo dal tipo di dati che si stanno analizzando ma anche dall'obiettivo dell'analisi. Un procedimento un po' meno arbitrario viene svolto dai software specializzati in calcoli statistici. Chiamando k il numero di classi da costruire si utilizza la seguente formula:

$$k = 1 + \log_2(N) \tag{1.1}$$

nella quale il numero delle classi è la parte intera del numero che si ricava dal logaritmo[7] in base 2 del numero di elementi presi in esame. Se il totale delle osservazioni fosse pari a 50 come nel caso precedente, avremo:

[7] Attenzione al fatto che il logaritmo in base 2 non è presente sulle comuni calcolatrici dove di norma sono più utilizzati il logaritmo naturale (simbolo Ln) oppure il logaritmo decimale (simbolo Log). Il problema non si pone per le calcolatrici scientifiche.

$$k = 1 + \log_2(50) = 1 + 5{,}643 = 6{,}643$$

da cui, prendendo soltanto la parte intera si ricava $k = 6$. Fissato in questo modo il numero delle classi si può calcolare l'ampiezza delle classi utilizzando la formula:

$$Ampiezza = \frac{x_{\max} - x_{\min}}{k} \tag{1.2}$$

dove x_{max} ed x_{min} rappresentano rispettivamente il valore più grande e quello più piccolo della distribuzione e k è il numero delle classi appena calcolato.

Esempio 1.9

Continuando l'esempio precedente se i valori estremi delle età fossero 20 e 65 avremmo

$$Ampiezza = \frac{65 - 20}{6} = 7{,}5$$

valore oltremodo scomodo da trattare. In genere conviene arrotondare il valore così ottenuto ad un valore significativo. Nel caso in esame, classi di 10 anni (o di 5 o di 20) sono più semplici da trattare di classi ampie 7,5 anni (o 6 o 9).

Occorre infine prestare attenzione ai valori estremi delle classi: poiché ogni osservazione deve cadere in una ed una sola classe il limite superiore della classe non deve coincidere con quello inferiore della successiva.

Proposta di soluzione in Excel

Per effettuare il calcolo di K (numero di classi) e di A (ampiezza delle classi) basterà riportare le formule in Excel nel seguente modo:

considerata una popolazione con N=50 clienti bisogna prima calcolare l'estremo inferiore e quello superiore con le seguenti formule:

```
=MIN(B2:B51); =MAX(B2:B51)
```

È, inoltre, possibile calcolare la numerosità N con la seguente formula:

```
=CONTA.VALORI(B2:B51)
```

il numero di classi come da Formula (1.1) ovvero:

```
=INT(1+LOG(E5;2))
```

e l'ampiezza delle classi come da Formula (1.2) ovvero:

```
=(E4-E3)/E6
```

Infine, definita una ampiezza pari a 10, in quanto più significativa, basterà applicare nuovamente la funzione:

```
=FREQUENZA(B2:B51;D11:D15)
```

	A	B	C	D	E	F	G	H	I
1	Cliente	Età		Minimo:	20	<=MIN(B2 : B51)			
2	1	24		Massimo:	65	<=MAX(B2 : B51)			
3	2	20		N:	50	<=CONTA.VALORI(B2 : B51)			
4	3	65		k:	6	<=INT(1 + LOG(E5 ; 2))			
5	4	21		ampiezza:	7,5	<=(E4 - E3) / E6			
6	5	65							
7	6	65		Anzianità in classi		Frequenza			
8	7	34		fino a 30	30	18	<=FREQUENZA(B2 : B51; F8 : F12)		
9	8	65		31-40	40	5			
10	9	23		41-50	50	3			
11	10	45		51-60	60	5			
12	11	65		oltre 60	71	19			
13	12	65			*Totale:*	*50*			
14	13	52							
15	14	63							
16	15	64							

Figura 1.7 Calcolo della numerosità e ampiezza delle classi.

e premere contemporaneamente[8] i tasti `Ctrl+Maiusc+Invio`. Il risultato dovrebbe apparire come in Figura 1.7. □

Esempio 1.10

Viene condotta una indagine sulla modalità con cui si distribuiscono i ritardi che affliggono i passeggeri di alcuni treni a lunga percorrenza in arrivo in una stazione importante del Nord Italia nell'arco di una settimana. I dati seguenti rappresentano i minuti di ritardo:

49, 5, 68, 51, 35, 16, 56, 5, 90, 6, 7, 0, 101, 31, 40, 119, 115, 30, 89, 5, 116, 8, 6, 6, 0, 15, 6, 96, 137, 47, 158, 7, 13, 25, 141, 144, 21, 21, 17, 9, 21, 33, 5, 127, 28, 18, 7, 0, 158, 151, 17, 8, 5, 5.

Per calcolare numero e ampiezza delle classi applichiamo lo schema proposto nel paragrafo:

1. calcoliamo la numerosità delle classi applicando la Formula (1.1)
2. determiniamo il ritardo più piccolo e quello più grande
3. calcoliamo l'ampiezza delle classi applicando la Formula (1.2)
4. ricaviamo la tabella di frequenza

Il numerosità del collettivo è pari a 54. Applicando la Formula (1.1) avremo:

$$k = 1 + \log_2(54) = 1 + 5,755 = 6,755$$

[8] Il termine contemporaneamente si usa per indicare che vanno premuti tre tasti anche in successione: mentre si tiene premuto il primo, premere gli altri due.

da cui, estraendo la parte intera, si ottiene $k = 6$. Dando uno sguardo alla tabella osserviamo che il valore minore è 5 mentre il massimo è 158.
Applicando quindi la Formula (1.2) otterremo:

$$Ampiezza = \frac{158 - 5}{6} = 25,5$$

valore che può agevolmente essere arrotondato a 30 minuti. Quindi si può procedere a costruire la tabella mostrata in Tabella 1.10 in cui si può notare che l'estremo inferiore di ogni classe viene incluso nell'intervallo.

Tabella 1.10 Frequenza dei ritardi per classi di 30 minuti.

Mnuti ritardo	Numero treni
5 ⊢ 30	31
30 ⊢ 60	8
60 ⊢ 90	2
90 ⊢ 120	6
120 ⊢ 150	4
150 ⊢ 180	3
Totale:	*54*

Proposta di soluzione in Excel
Per calcolare le frequenze dei ritardi bisogna raggruppare i dati in classi. L'esempio si svolge in modo simile ai precedenti. Bisognerà quindi procedere con il calcolo del:

```
=MIN(A2:F10); =MAX(A2:F10)
```

È inoltre, possibile calcolare la numerosità N con la seguente formula:

```
=CONTA.VALORI(A2:F10)
```

il numero di classi come da Formula (1.1) ovvero:

```
=INT(1+LOG(J11;2))
```

e l'ampiezza delle classi come da Formula (1.2) ovvero:

```
=(J10-J9)/J12
```

Infine, definita una ampiezza pari a 10, in quanto più significativa, basterà applicare nuovamente la funzione

```
=FREQUENZA(A2:F10;H2:H7)
```

e premere contemporaneamente[9] i tasti `Ctrl+Maiusc+Invio`. □

[9] Il termine contemporaneamente si usa per indicare che vanno premuti tre tasti anche in successione: mentre si tiene premuto il primo, premere gli altri due.

	A	B	C	D	E	F	G	H	I	J	K
1	Minuti di ritardo accumulati dai treni in arrivo							Minuti ritardo		Numero treni	
2	49	5	68	51	35	16		5 ⊢ 30	30	31	<=FREQUENZA(A2:F10; H2:H7)
3	5	56	5	90	6	7		30 ⊢ 60	60	8	
4	101	31	40	119	115	30		60 ⊢ 90	90	3	
5	6	89	5	116	8	6		90 ⊢ 120	120	5	
6	7	15	6	96	137	47		120 ⊢ 150	150	4	
7	144	158	7	13	25	141		150 ⊢ 180	180	3	
8	21	21	17	9	21	33		*Totale:*		*54*	
9	7	5	127	28	18	7			Minimo:	5	<=MIN(A2 : F10)
10	158	151	17	8	5	5			Massimo:	158	<=MAX(A2 : F10)
11									N:	54	<=CONTA.VALORI(A2 : F10)
12									k:	6	<=INT(1 + LOG(J11 ; 2))
13									Ampiezza:	25,5	<=(J10 - J9) / J12
14											

Figura 1.8 Calcolo della numerosità e ampiezza delle classi.

1.5 Calcolo di frequenze relative e cumulate

Nel paragrafo precedente abbiamo visto come passare dalla tabella dei dati grezzi alla tabella di frequenze per ricavare rapidamente informazioni utili. Ma le frequenze calcolate, essendo assolute, dipendono direttamente dalla dimensione del campione della popolazione considerata; se si dovessero confrontare le frequenze ottenute da un campione collettivo di grande dimensione con uno molto più piccolo sarebbe impossibile realizzare un utile riscontro.

Per ovviare a tale problema si usa calcolare le *frequenze relative* ottenute dividendo ciascuna singola frequenza con la frequenza totale. In formula avremo

$$f_i = \frac{n_i}{N} \tag{1.3}$$

Ogni frequenza f_i avrà un valore $0 \le f_i \le 1$ e la somma di tutte le frequenze sarà pari a 1. Sovente al posto delle frequenze relative si preferisce utilizzare le frequenze relative percentuali che si ottengono semplicemente moltiplicando le frequenze relative per 100.

$$f_i = \left(\frac{n_i}{N}\right) \times 100 \tag{1.4}$$

La somma di tutte le frequenze percentuali è pari a 100. Sotto forma di tabella la distribuzione di frequenza della variabile X sarà la seguente:

Modalità della variabile X	Frequenze assolute n_i	Frequenze relative f_i	Frequenze relative percentuali $f_i \times 100$
x_1	n_1	$\frac{n_1}{N}$	$f_1 \times 100$
x_2	n_2	$\frac{n_2}{N}$	$f_2 \times 100$
....	....	....	....
x_k	n_k	$\frac{n_k}{N}$	$f_k \times 100$
	N	1	100

Esempio 1.11

Il nuovo direttore del personale di un'azienda di servizi ritiene che vi sia qualche differenza tra addetti maschi e femmine nel salario percepito. I dati delle retribuzioni sono raccolti nelle tabella seguente:

Tabella 1.11 Ripartizione della retribuzione lorda per genere e classe di salario.

Classi di salario	Maschi		Femmine	
	n_i	$f_i \times 100$	n_i	$f_i \times 100$
800 ⊢ 1000	31	20,0	33	42,9
1000 ⊢ 1200	56	36,1	26	33,8
1200 ⊢ 1500	42	27,1	12	15,6
1500 ⊢ 2000	26	16,8	6	7,8
Totale:	**155**	**100**	**77**	**100**

Poiché il numero dei dipendenti maschi presenti è circa il doppio di quello delle femmine, le frequenze assolute (colonna n_i) non sono tra loro agevolmente confrontabili mentre le frequenze relative percentuali (colonna $f_i \times 100$) mostrano con evidenza maggiore come le dipendenti donna occupino categorie inferiori rispetto a quelle dei colleghi maschi. Le frequenze relative, riducendo i dati alla comune base 100, ci permettono un più agevole confronto.

Proposta di soluzione in Excel

Per calcolare le frequenze relative percentuali nelle celle della colonna C occorre dividere il valore della frequenza assoluta (ad esempio 31 nella cella B3) con la somma totale delle frequenze posta in B7. Nel copiare (trascinare) questa formula dalla C3 nelle

restanti celle in basso, bisogna assicurarsi che il riferimento al denominatore B7 rimanga fisso e quindi andrà utilizzato un riferimento assoluto (ottenuto anche premendo il tasto f4). La formula da scrivere in C3 sarà:

```
=B3/$B$7
```

per trascinarla poi nelle celle in basso. Mantenendo le celle selezionate cliccare prima sul formato percentuale (%) poi su Aumenta decimali (entrambi i comandi sono contenuti nel riquadro Numero).
Ripetere lo stesso procedimento sulla cella E3 scrivendo la seguente formula:

```
=D3/$D$7
```

Si otterrà la seguente tabella:

	A	B	C	D	E	F	G
1	Classi di salario	Maschi		Femmine			
2		n_i	$f_i \times 100$	n_i	$f_i \times 100$		
3	800 ⊢ 1000	31	20,0%	33	42,9%		
4	1000 ⊢ 1200	56	36,1%	26	33,8%		
5	1200 ⊢ 1500	42	27,1%	12	15,6%		
6	1500 ⊢ 2000	26	16,8%	6	7,8%		
7	Totale dipendenti:	155		77			
8			=B3/$B$7		=D3/$D$7		
9							
10							

Figura 1.9 Calcolo della frequenza relativa percentuale.

Nelle tabelle Excel le percentuali vengono formattate e mostrate con il segno % a differenza di quanto è mostrato nelle tabelle del testo in cui si preferisce utilizzare una notazione più formale. □

I caratteri quantitativi sono in genere ordinati in maniera crescente e può risultare quindi interessante calcolare le *frequenze cumulate* ossia le frequenze ottenute sommando progressivamente le frequenze (sia *assolute* che le *relative*). Sotto forma di tabella la distribuzione di frequenza cumulata della variabile X avrà il seguente aspetto:

Modalità della variabile X	Frequenze assolute n_i	Frequenze assolute cumulate	Frequenze relative cumulate
x_1	n_1	n_1	f_1
x_2	n_2	$n_1 + n_2$	$f_1 + f_2$
....	....	....	....
x_k	n_k	$n_1 + n_2 + ... + n_k$	$f_1 + f_2 + ... + f_k$

Esempio 1.12

Riprendiamo i dati dell'Esempio 1.10 e costruiamo la tabella con le frequenze assolute, cumulate, relative in percentuale e relative cumulate:

Tabella 1.12 Tabella riassuntiva della distribuzione di frequenza dei ritardi.

Minuti ritardo	Frequenza assoluta	Frequenza assoluta cumulata	Frequenza relativa in %	Frequenza relativa % cumulata
	n_i	$n_1 + n_2 + ... + n_k$	$f_i \times 100$	$f_1 + f_2 + ... + f_k$
5 ⊢ 10	19	19	35,2	35,2
10 ⊢ 20	6	25	11,1	46,3
20 ⊢ 30	6	31	11,1	57,4
30 ⊢ 60	8	39	14,8	72,2
60 ⊢ 120	8	47	14,8	87,0
120 ⊢ 180	7	54	13,0	100,0
Totale	*54*		*100,0*	

Osservando la colonna delle frequenze relative cumulate posta all'estrema destra si può verificare che il 60% circa (il 57,4% per la precisione) delle osservazioni ricade nei primi 30 minuti di ritardo e quasi i tre quarti di esse (il 72,2%) stia all'interno dei 60 minuti.

Proposta di soluzione in Excel

Per calcolare le frequenze cumulate bisognerà porre nella cella D3 il valore della frequenza C3 mentre nella cella D4 la formula:

```
=C4+D3
```

Tale cella va copiata (trascinata) nelle successive.

Le frequenze relative percentuali nelle celle della colonna E vanno invece calcolate indicando la seguente formula nella cella E3 :

```
=C3/$C$9
```

per trascinarla poi nelle celle in basso. Mantenendo le celle selezionate cliccare prima sul formato percentuale poi su Aumenta decimali (entrambi i comandi sono contenuti nel riquadro Numero).

Le frequenze relative % cumulate verranno calcolate come le frequenze cumulate ponendo nella cella F3 il valore della frequenza E3 mentre nella cella F4 la formula:

```
=E4+F3
```

Tale cella va copiata (trascinata) nelle successive. □

	B	C	D	E	F	G
1	Minuti ritardo	Frequenza assoluta	Frequenza assoluta cumulata	Frequenza relativa in %	Frequenza relativa % cumulata	
2		n_i	$n_1 + n_2 + \ldots + n_k$	$f_i \times 100$	$f_1 + f_2 + \ldots + f_k$	
3	5 ├ 10	19	19	35,2%	35,2%	
4	10 ├ 20	6	25	11,1%	46,3%	
5	20 ├ 30	6	31	11,1%	57,4%	
6	30 ├ 60	8	39	14,8%	72,2%	
7	60 ├ 120	8	47	14,8%	87,0%	
8	120 ├ 180	7	54	13,0%	100,0%	
9	Totale	54		100,0%		
10		=C3				
11			=C4+D3	=C3/$C$9	=E3	=E4+F3
12						

Figura 1.10 Calcolo delle frequenze relative e cumulate.

1.6 Esercizi

Esercizio 1.1

Si vogliono realizzare alcune indagini statistiche per descrivere e comprendere dei fenomeni sociali. Di ciascuno di quelli elencati di seguito individuare:

a) la popolazione, indicando esplicitamente da quali elementi è composta,
b) caratteri oggetto di studio, classificandone il tipo,
c) modalità con cui ciascun carattere si manifesta,
d) possibili fonti dei dati.

Gli argomenti da indagare sono:

1. La capacità ricettiva alberghiera della propria regione.
2. Numero, estensione e tipologia dei parchi naturali e delle aree protette del Piemonte.
3. Incidenti stradali mortali nel 2005. Ampiezza del fenomeno e cause di morte.
4. Origine demografica del personale sanitario della regione Emilia e Romagna.
5. Le imprese commerciali fallite nel comune di Bari nel 2006. Ampiezza del fenomeno.
6. Un istituto di credito intende conoscere l'atteggiamento della propria clientela *retail* verso l'impiego via web delle operazioni bancarie.

7. Un pastificio industriale intende indagare la qualità dei propri prodotti forniti al mercato della grande distribuzione, iniziando dalle oscillazioni di peso nelle confezioni di spaghetti da 500 gr.
8. La condizione lavorativa nel 2006, a cinque anni dalla laurea, dei laureati in Scienze Politiche ed Economia dell'Ateneo locale.
9. La diffusione sul territorio regionale degli istituti di credito.

Risposte alle domande dell'esercizio 1

1. Risposte:
   a. tutti gli alberghi della regione
   b. numero dei posti letto, variabile quantitativa discreta
   c. da 0 a *n* da suddividere in classi di frequenza
   d. Istat, Camera di commercio, Federalberghi, Enit.
2. Risposte:
   a. tutte le aree protette: Parchi nazionali, regionali, riserve, aree protette
   b. kmq o ettari, variabili quantitative continue, tipologia, variabile qualitativa
   c. per l'estensione, da 0 a *n* da suddividere in classi di frequenza, per la tipologia, elenco.
   d. Istat, regione Piemonte, siti web dei parchi.
3. Risposte:
   a. Tutti gli incidenti stradali avvenuti in Italia nel corso del 2005
   b. Numero feriti e numero morti per incidente, variabili quantitative discrete. Causa di morte, variabile qualitativa
   c. per la numerosità, da 0 a *n* da suddividere in classi di frequenza, per la tipologia, elenco.
   d. Istat, ACI, Ministero interno, polizia.
4. Risposte:
   a. Tutti i dipendenti del settore sanitario della regione
   b. Luogo di nascita o cittadinanza, variabili qualitative
   c. Elenco, attributi
   d. Istat, Ministero sanità, Regione Emilia Romagna, Aziende sanitarie.
5. Risposte:
   a. Tutte le aziende commerciali presenti nel territorio del comune di Bari
   b. Superficie esercizio, variabile quantitativa continua, tipologia esercizio, forma giuridica, variabili qualitative, fallimento, dicotomica (Si/No)
   c. per l'estensione, da 0 a *n* in m^2, da suddividere in classi di frequenza, per la tipologia e la forma giuridica, elenco, attributi, per il fallimento, attributi Si/no.
   d. Istat, Camera di commercio (Registro delle imprese), Cerved, Associazioni categoria commercianti (Confcommercio, Confesercenti, Ascom, Fipe).
6. Risposte:
   a. Tutti i propri clienti nel settore *retail*
   b. Età (D), genere (C), titolo di studio (C), dimensione economica (D), numero ope-

razioni effettuate lo scorso anno (D), atteggiamento verso operazioni via web (C)
c. Tutte le variabili andranno suddivise in classi.
d. Banca dati interna.

7. Risposte:
a. Tutte le confezioni di spaghetti da 500 gr. prodotte in un determinato arco di tempo
b. Peso in gr, variabile quantitativa continua
c. Peso da 0 a *n* da suddividere in classi di frequenza

8. Risposte:
a. Tutti gli studenti della Facoltà di Scienze politiche ed Economia che si sono laureati nel 2001
b. Tipo di laurea (C), voto di laura (D), condizione lavorativa (C)
c. Archivio Alma Laurea, dati segreteria studenti Ateneo.

9. Risposte:
a. Le sedi centrali e periferiche degli istituti di credito presenti nella regione
b. Localizzazione (C), tipologia (C), numero clienti (D), ammontare risorse amministrate (D)
c. Istat, ABI, Istituti di credito.

Esercizio 1.2

Calcolo di frequenze di tipo categoriale

Il foglio Excel "Esercizio 1.2" contiene alcuni dati dei clienti di una piccola impresa che vende su Internet. Creare una tabella di frequenza per la variabile "Genere".

I dati si presentano come in Figura 1.11.

	A	B	C	D
1	ID cliente	Genere	Prov	
2	1	M	VR	
3	2	F	PD	
4	3	F	FE	
5	4	M	PD	
6	5	M	VR	
7	6	F	FE	
8	7	F	VR	
9	8	F	VI	
10	9	F	PD	
11	10	F	VR	
12	11	F	VI	
13	12	F	PD	

	A	B	C	D	E	F	G
1	ID cliente	Genere	Prov		Tabella di frequenza		
2	1	M	VR		Maschi:	25	< =CONTA.SE(B2:B73;"=M")
3	2	F	PD		Femmine:	47	< =CONTA.SE(B2:B73;"=F")
4	3	F	FE		*Totale:*	72	< =SOMMA(F2:F3)
5	4	M	PD				
6	5	M	VR				

Figura 1.12 La tabella al termine dell'esercizio.

La prima colonna contiene un semplice numero identificativo del cliente che sostituisce il nome per non incorrere nei vincoli imposti dalle norme sulla privacy. La seconda colonna contiene il genere mentre la terza verrà utilizzata nei successivi esercizi. Per risolvere l'esercizio si deve calcolare la frequenza dei maschi e delle femmine. Esistono in Excel molti modi per risolvere il problema. Negli esercizi di questo capitolo ne vedremo alcuni, altri seguiranno nei successivi.

Per contare tutte le caselle che contengono "M" utilizzeremo la funzione Excel

```
=CONTA.SE(<intervallo>; <criteri>)
```

dove al posto del parametro `<intervallo>` andranno inserite le celle che contengono i dati, ovvero il range `A2:A73`, mentre al posto di `<criteri>` si dovrà inserire `"=M"` per ottenere il conteggio dei maschi e `"=F"` per il conteggio delle femmine. Nell'ambito delle celle descritte in `<intervallo>`, la funzione conta le celle che soddisfano il criterio `"=M"`. In *Figura 1.12* si può osservare il risultato. Nelle celle `G2`, `G3` e `G4`, sono mostrate le formule che in realtà sono scritte nelle corrispondenti celle della colonna `F`.

Esercizio 1.3

Calcolo di frequenze numeriche discrete semplici

Una azienda produttrice di cereali vende ai supermercati i propri prodotti in scatole da 20 confezioni ciascuna. Poiché dai clienti sono giunti numerosi rilievi sullo stato delle confezioni (i pacchetti si presentano rotti o non saldati correttamente o stampati male), il direttore della produzione effettua un controllo su un campione di 100 scatole all'uscita della linea di imballaggio. Ogni scatola viene riaperta, controllato il contenuto e registrato su un foglio il numero di confezioni trovate con difetti. Il foglio "Esercizio 1.3" contiene il risultato dell'indagine. Predisporre una tabella delle frequenze trovate.

Osservando la Figura 1.13 vediamo come le 100 osservazioni siano raccolte nelle prime cinque colonne. Poiché si tratta di dati numerici discreti si può utilizzare la funzione

	A	B	C	D	E
1	Numero di confezioni che presentano difetti per ciascuna delle 100 scatole esaminate[1]				
2	0	0	2	0	0
3	0	0	2	0	0
4	1	0	0	1	2
5	1	0	0	0	0
6	0	0	0	1	4
7	0	0	0	0	0
8	0	0	0	2	0
9	0	0	0	0	1
10	1	0	3	1	0
11	0	0	0	0	0
12	1	0	0	2	0
13	0	0	1	3	0
14	0	0	0	0	0
15	0	0	0	0	0
16	0	0	0	0	1
17	0	2	0	0	0
18	0	1	0	0	0
19	0	0	0	0	0
20	0	0	0	2	0
21	1	0	1	0	0
22	Nota 1: vengono esaminate 100 scatole contenenti 20 confezioni ciascuna. Ogni numero della tabella indica quante confezioni difettose vengono trovate in ogni scatola.				

Figura 1.13 La tabella del foglio Esercizio 1.3.

`=FREQUENZA()` che possiede il vantaggio di essere più flessibile e diretta di quella utilizzata nell'esercizio precedente in cambio di qualche rigidità in più nel suo funzionamento. I valori oscillano tra lo 0 e 4. A partire dalla cella `G2` scrivere i valori `0`, `1`, `2`, `3`, `4`. Selezionare le celle contigue `H2:H6` e, come mostrato in Figura 1.14, scrivere la seguente formula badando di mantenere attiva la selezione:

```
=FREQUENZA(A1: A100; G2:G6)
```

Al termine non dare `Invio`, ma premere contemporaneamente[10] i tasti `Ctrl+Maiusc+Invio`. Il risultato dovrebbe apparire come in Figura 1.15.

Note sulla funzione `FREQUENZA()`

La funzione `FREQUENZA()` è una funzione matrice, così chiamate quando restituiscono più valori insieme. Nella colonna `G` di Figura 1.14, sono stati inseriti tutti i possibili valori di cui si vuole ottenere la frequenza: le celle a destra vengono selezionate per scrivere la funzione. Premendo i tasti `Ctrl+Maiusc+Invio` la funzione riempirà

[10] Il termine contemporaneamente si usa per indicare che vanno premuti tre tasti anche in successione: mentre si tiene premuto il primo, premere gli altri due.

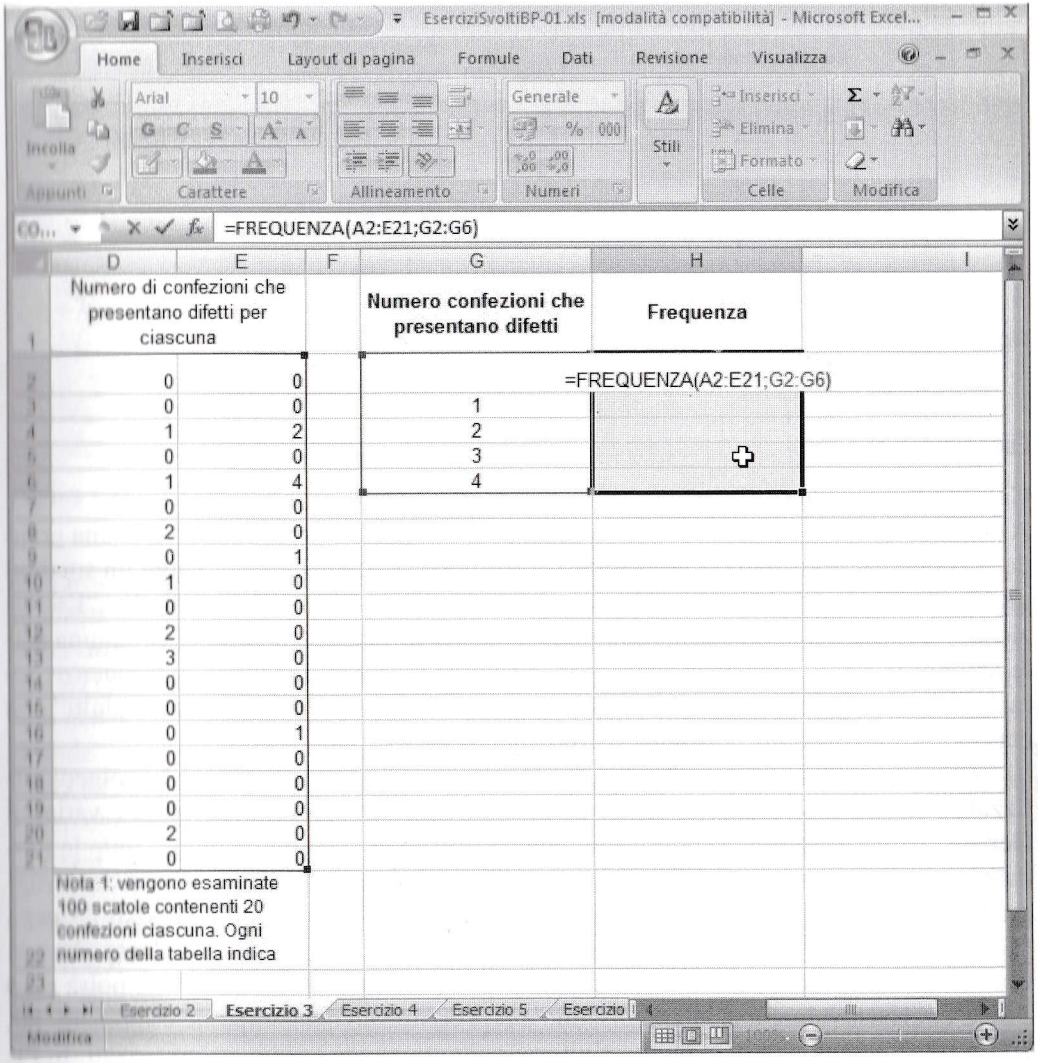


Figura 1.14 L'inserimento della funzione Excel `=FREQUENZA()`.

tutte le celle selezionate (da `H2` a `H6`) con la frequenza calcolata. Nella colonna `C` si sarebbero potuto porre i valori in altro modo: digitando ad esempio soltanto 1 e 4, la funzione avrebbe restituito nelle due celle contigue i valori 90 e 10, ossia la frequenza di tutti valori fino a 1 compreso e da 1 (escluso) a 4 compreso. Con un po' di astuzia è possibile ottenere qualunque raggruppamento in classi da questa funzione, purché i valori siano espressi da numeri e non da testo. La funzione è molto flessibile e verrà impiega-

	F	G	H	I
1		Numero confezioni che presentano difetti	Frequenza	
2		0	77	< =FREQUENZA(A2:E21; G2:G6)
3		1	13	
4		2	7	
5		3	2	
6		4	1	
7				
8				

Figura 1.15 Il risultato ottenuto utilizzando la funzione `=FREQUENZA()`.

Esercizio 1.4

Calcolo di frequenze di tipo categoriale

A causa delle forti lamentele dei passeggeri dei treni per pendolari in arrivo in una grande stazione del nord Italia, l'assessorato ai trasporti del comune fa svolgere un sondaggio tra i passeggeri dei treni coinvolti dalla polemica. La tabella nel foglio "Esercizio 1.4" contiene alcune risposte degli intervistati di uno di tali convogli. Per ognuna delle colonne (Genere, Età, Titolo di studio, Reddito) preparare la tabelle di frequenze. Le età andranno raggruppate in tre categorie: "giovani" con età fino a 25 anni, "adulti" con età tra i 26 ed i 59 anni, "anziani" dai 60 anni in su.

Dato che tutte le modalità sono espresse con numeri, l'esercizio si può svolgere utilizzando la funzione `FREQUENZA()` come visto nell'esercizio precedente. Le variabili Genere, Titolo di studio, Reddito, sono già codificate in categorie e quindi tutte le modalità andranno elencate. Diverso è il caso della variabile Età dato che si chiede di raggruppare i dati in tre classi soltanto. I limiti superiori delle prime due classi, 25 per i "Giovani" e 59 per gli "Adulti" sono esplicitamente indicati, per la terza classe, dove si usa il termine "oltre", occorre invece fissare lo stesso un tetto massimo che potrà però essere qualsiasi valore anche grandissimo e non compreso nel campione esaminato, 80, 100 o 1000: Excel includerà in quest'ultimo raggruppamento tutti i valori trovati. In Figura 1.16 si può osservare il risultato finale con le formule impiegate.

Esercizio 1.5

Calcolo di frequenze numeriche continue per classi di frequenza

Un'azienda metalmeccanica produce parti di carrozzeria dell'auto. L'azienda committente richiede che alcune parti della lamiera abbiano uno spessore di 0,75 mm con la tolleranza massima di 5 centesimi di millimetro, ossia che lo spessore non possa essere inferiore 0,70 mm. o superiore a 0,80 mm. Nel foglio "Esercizio 1.5" sono contenute 1000 misurazioni effettuate: ricavare una tabella con la distribuzione delle frequenze dopo aver calcolato con l'apposita formula il numero delle classi e l'ampiezza

	G	H	I	J	K
1			Minuti di ritardo suddivisi in classi	Frequenza	
2		10	Fino a 10 minuti	19	< =FREQUENZA(A2:F10; H2:H7)
3		20	10 ├– 20	6	
4		30	20 ├– 30	6	
5		60	30 ├– 60	8	
6		120	60 ├– 120	8	
7		180	120 ├– 180	7	
8					

Figura 1.16 La tabella delle distribuzioni di frequenza delle variabili dell'esercizio.

In questo esercizio occorre determinare prima di tutto la numerosità e l'ampiezza delle classi applicando le formule (1.1) ed (1.2) del paragrafo 1.4. Il numero totale delle osservazioni presenti nella tabella si calcola utilizzando la funzione `=CONTA.NUMERI(A2:D251)` che restituisce nella cella `G5` di Figura 1.17 il conteggio dei numeri trovati nell'intervallo `A2:D51` (le celle contenenti i dati). Il valore più piccolo e quello più grande si ottengono rispettivamente con le funzioni `=MIN(A2:D251)` e `=MAX(A2:D251)` poste nelle caselle `G3` e `G4`. La numerosità delle classi si ottiene con l'impiego della formula

```
=INT(1+LOG(G5;2))
```

dove la funzione `LOG(G5;2)` calcola il logaritmo in base 2 del numero contenuto nella cella `G5`, mentre la funzione `INT()` prende solo la parte intera del logaritmo a cui è stato aggiunto 1.

	A	B	C	D	E	F	G	H
1	Spessore in mm della lamiera							
2	0,7623	0,7584	0,7242	0,7660				
3	0,7430	0,7720	0,7450	0,7656		Valore minimo:	0,7044	< =MIN(A2:D251)
4	0,7628	0,7720	0,7506	0,7482		Valore massimo:	0,7962	< =MAX(A2:D251)
5	0,7364	0,7722	0,7501	0,7476		Numero elementi:	1000	< =CONTA.NUMERI(A2:D251)
6	0,7633	0,7551	0,7555	0,7515				
7	0,7511	0,7585	0,7530	0,7653		k:	10	< =INT(1+LOG(G5;2))
8	0,7374	0,7430	0,7453	0,7483		ampiezza:	0,0092	< =(G4-G3)/G7
9	0,7562	0,7799	0,7570	0,7607				
10	0,7727	0,7569	0,7446	0,7626				

Figura 1.17 Impiego delle funzioni per determinare numerosità ed ampiezza delle classi.

	I	J	K	L	M	N	O
1	**Spessore in mm. suddiviso in classi**		**Frequenza**				
2	0,700 — 0,710	0,710	5	< =FREQUENZA(A2:D251; J2:J11)			
3	0,710 — 0,720	0,720	21				
4	0,720 — 0,730	0,730	72				
5	0,730 — 0,740	0,740	171				
6	0,740 — 0,750	0,750	245				
7	0,750 — 0,760	0,760	229				
8	0,760 — 0,770	0,770	163				
9	0,770 — 0,780	0,780	65				
10	0,780 — 0,790	0,790	25				
11	0,790 — 0,800	0,800	4				
12	*Totale osservazioni:*		1000				
13							

Figura 1.18 Le 10 classi (colonna I) e la corrispondente distribuzione di frequenza (colonna K).

Il risultato 10 è un valore adatto al nostro caso[11] e lo si può inserire nella formula in G8 ottenendo così 0,0092. L'intervallo dei valori osservati va da 0,700 mm. a 0,800 mm. circa da suddividere in 10 classi di ampiezza 0,010 (un centesimo) molto vicino[12] a quanto ricavato con la formula in `G8`.

Nell'intervallo di celle `J2:J11` si devono porre i limiti superiori delle classi, nelle corrispondenti celle a sinistra (colonna I) si potranno inserire gli estremi della classe (a scopo descrittivo) poi, selezionando l'intervallo `K2:K11`, si scriverà la funzione come descritto nella Figura 1.18

Esercizio 1.6

Calcolo di frequenze assolute, relative e cumulate.
Un corriere cittadino utilizza quattro autisti per le consegne quotidiane. Il responsabile del magazzino vuole esaminare il lavoro svolto dagli autisti nell'ultimo mese per ridistribuire con equità il carico di lavoro. Nel foglio "Esercizio 1.6", sono presentati i dati suddivisi per giorno e per autista. Si calcolino le frequenze assolute raggruppate per classi di ampiezza 10. Si calcolino inoltre le frequenze relative percentuali e le frequenze relative cumulate.

Analogamente a quanto fatto nei precedenti esercizi, per calcolare la frequenza si inizia ponendo in una colonna (nel nostro caso la `J` come si può osservare dalla Figura 1.19) i limiti superiori delle classi di frequenza. Nella colonna `I` la descrizione. Si seleziona poi l'intervallo `K2:K7` e si scrive la funzione

```
=FREQUENZA(C2:F31; J2:J7)
```

[11] Non c'è nulla che impedisca di prendere un qualsiasi valore compreso tra 5 e 15, si veda quanto scritto nel paragrafo 1.4.

Al termine si premono i tasti `Ctrl+Maiusc+Invio` e la funzione riempirà tutte le celle selezionate con la frequenza calcolata.

Per calcolare le frequenze relative percentuali nelle celle della colonna L occorre dividere il valore della frequenza assoluta (ad esempio 11 nella cella `K2`) con la somma totale delle frequenze posta in `K8`. Nel copiare questa formula dalla `L2` nelle restanti celle in basso, bisogna assicurarsi che il riferimento al denominatore `K8` rimanga fisso e quindi andrà utilizzato un riferimento assoluto. La formula da scrivere in `L2` sarà:

```
=K2/$K$8
```

per trascinarla poi nelle celle in basso. Mantenendo le celle selezionate cliccare prima sul formato percentuale poi su `Aumenta decimali` (entrambi i comandi sono contenuti nel riquadro `Numero`). Le frequenze cumulate verranno calcolate a partire dalla cella M3 sommando la frequenza percentuale in `L3` con quanto presente in `L2` e ripetendo lo stesso calcolo nelle celle in basso come si può osservare in Figura 1.19.

Esercizio 1.7

Il direttore commerciale di un ipermercato intende svolgere un veloce sondaggio per verificare il gradimento che riceve dai giovani clienti l'ambientazione innovativa del settore dedicato alla musica. Vengono intervistati 68 giovani di età compresa tra i 16 ed i 25 anni ed i dati raccolti sul foglio Excel "Esercizio 1.7". Per ognuna delle quattro variabili statistiche (Età, Genere, Titolo di studio, Gradimento) costruire una tabella di frequenza senza raggruppare i dati in classi. Nella tabella del gradimento aggiungere le frequenze relative e quelle cumulate e rispondere alle seguenti domande:

- Quale percentuale di giovani manifesta un atteggiamento indifferente o negativo?
- Quale percentuale manifesta un atteggiamento positivo?

Il calcolo delle distribuzioni di frequenza per Genere, Età e Titolo di studio è del tutto simile a quello svolto nei precedenti esercizi ed il risultato è mostrato in Figura 1.20.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Giorno		Andrea	Bruno	Carlo	Davide			Consegne		Frequenza	Frequenza %	Frequenza % cumulata	
2	Gi	1	33	6	18	31			fino a 10	10	11	12,5%	12,5%	< =L2
3	Ve	2	15	37	32	42			Da 11 a 20	20	23	26,1%	38,6%	< =M2+L3
4	Sa	3							Da 21 a 30	30	15	17,0%	55,7%	< =M3+L4
5	Do	4							Da 31 a 40	40	17	19,3%	75,0%	< =M4+L5
6	Lu	5	13	19	11	46			Da 41 a 50	50	17	19,3%	94,3%	< =M5+L6
7	Ma	6	34	16	12	26			Da 51 a 60	60	5	5,7%	100,0%	< =M6+L7
8	Me	7	24	39	9	27			*Totale:*		88	< =SOMMA(K2:K7)		
9	Gi	8	36	23	43	46							=K2/$K$8	
10	Ve	9	12	34	13	6			Minimo:	5	< =MIN(C2:F31)			
11	Sa	10							Massimo:	59	< =MAX(C2:F31)			
12	Do	11												
13	Lu	12	33	10	21	34								
14	Ma	13	49	19	7	45								
15	Me	14	33	29	55	59								
16	Gi	15	21	31	22	44								
17	Ve	16	41	7	36	5								
18	Sa	17												

=FREQUENZA(C2:F31;J2:J7)

Figura 1.19 La tabella con tutte le formule impiegate.

	F	G	H	I	J	K	L
1	Età	Freq.		Genere		Frequenza	
2	16	3	< =FREQUENZA(B2:B69; G2:G11)	Maschio	1	40	< =FREQUENZA(C2:C69; J2:J3)
3	17	9		Femmina	2	28	
4	18	9					
5	19	7		Titolo di studio		Frequenza	
6	20	11		Scuola media o inf.	1	27	< =FREQUENZA(D2:D69; J6:J8)
7	21	6		Diploma superiore	2	35	
8	22	6		Laurea	3	6	
9	23	8					
10	24	7					
11	25	2					
12							

Figura 1.20 Il calcolo delle frequenze per Genere, Età e Titolo di studio.

In Figura 1.21 si possono osservare i calcoli delle frequenze assolute (colonna Q), relative (colonna R) e cumulate (colonna T) con le formule impiegate inserite nelle colonne adiacenti.

Osservando la colonna delle frequenze cumulate diventa agevole rispondere ai due quesiti posti dal problema: il 47,1% dei giovani intervistati manifesta un atteggiamento negativo, comprendendo con questo termine anche gli indifferenti a cui viene sovente data una valenza negativa. Per contro il 53% circa (100-47) giudica positivamente l'ambientazione appena realizzata.

	O	P	Q	R	S	T	U
1	Gradimento		Freq.	Freq %		Freq % cumul.	
2	Non piace affatto	1	5	7,4%	< =Q2/$Q$7	7,4%	< =R2
3	Non piace	2	12	17,6%	< =Q3/$Q$7	25,0%	< =T2+R3
4	Né si né no (indifferente)	3	15	22,1%	< =Q4/$Q$7	47,1%	< =T3+R4
5	Piace	4	23	33,8%	< =Q5/$Q$7	80,9%	< =T4+R5
6	Piace molto	5	13	19,1%	< =Q6/$Q$7	100,0%	< =T5+R6
7	*Totale risposte:*		*68*	< =SOMMA(Q2:Q6)			
8							

Figura 1.21 La tabella di frequenza per il gradimento ed il calcolo delle frequenze relative e cumulate.

Esercizio 1.8

Un'azienda di ricerche di mercato deve svolgere una vasta indagine telefonica tra gli abitanti del Piemonte. A questo scopo scarica da un sito web l'elenco di tutti i comuni della regione così come sono raccolti nel foglio Excel "Esercizio 1.8". Creare una tabella di frequenza raggruppando i comuni per provincia e calcolare le frequenze relative del numero di comuni per provincia.

La tabella mostrata in Figura 1.22 è stata ricavata direttamente dal sito web dell'Annuario statistico della regione Piemonte e in origine conteneva i principali dati relativi a tutti i comuni della regione omessi in questo esercizio per non appesantire il file Excel.

	A	B	C	D	E	F
1	Comuni del Piemonte					
2	Prov	Comune	Codice Istat	Superficie (Km²)	Residenti (Istat 2004)	
3	CN	Acceglio	004001	151,94	167	
4	AL	Acqui Terme	006001	33,42	20.142	
5	AT	Agliano Terme	005001	15,38	1.658	
6	TO	Agliè	001001	13,28	2.645	
7	NO	Agrate Conturbia	003001	14,51	1.351	
8	BI	Ailoche	096001	10,26	321	
9	TO	Airasca	001002	15,7	3.652	
10	CN	Aisone	004002	36,87	269	
11	TO	Ala di Stura	001003	46,09	469	
12	VC	Alagna Valsesia	002002	72,8	451	
13	CN	Alba	004003	54,01	30.083	
14	VC	Albano Vercellese	002003	13,8	330	
15	CN	Albaretto della Torre	004004	4,35	249	
16	AL	Albera Ligure	006002	21,34	349	

Figura 1.22 La tabella contenente l'elenco di tutti i comuni del Piemonte come è stata scaricata da Internet.

Una siffatta tabella può essere difficile da gestire per il rilevante numero di righe (i comuni sono oltre 1200) e di colonne (più di 40 nella versione originale). Excel offre alcuni strumenti efficaci e potenti per ottenere informazioni sintetiche utilizzando pochi comandi, a volte solo un paio di click. Per classificare e contare i comuni per provincia si utilizza il comando `Dati/Struttura/Subtotale` che crea una struttura nella tabella suddividendola in livelli. Come vedremo tra breve, condizione indispensabile per poter operare è che i dati siano ordinati per provincia e non in ordine alfabetico come appaiono dalla tabella mostrata in Figura 1.22. La tabella può essere agevolmente ordinata selezionando una qualsiasi cella della colonna della provincia, cliccando quindi sul comando `Home/Modifica/Ordina` come mostrato in Figura 1.23.

Dopo aver ordinato l'intera tabella suddividendo i comuni nelle rispettive province di appartenenza si può procedere nel creare la struttura selezionando una cella della colonna A delle province e dando il comando `Dati/Struttura/Subtotale`. Comparirà la finestra mostrata in Figura 1.24 da utilizzare per indicare le impostazioni della struttura.

Il comando agisce inserendo una riga di subtotali ogni volta che interviene un cambiamento di valore nella colonna dove è posizionato il cursore. Il subtotale può essere calcolato come conteggio (è il nostro caso) oppure come somma, media ed altro ancora. I valori così ottenuti possono essere posti sotto ogni colonna della tabella senza limitazioni. Nella figura sono mostrate le impostazioni da dare per la soluzione dell'esercizio. Dopo aver dato `Invio` compariranno alla sinistra della tabella le linee verticali nere che definiscono i livelli della struttura. Cliccando sul pulsante 2 come mostrato nella Figura 1.25 si otterrà la

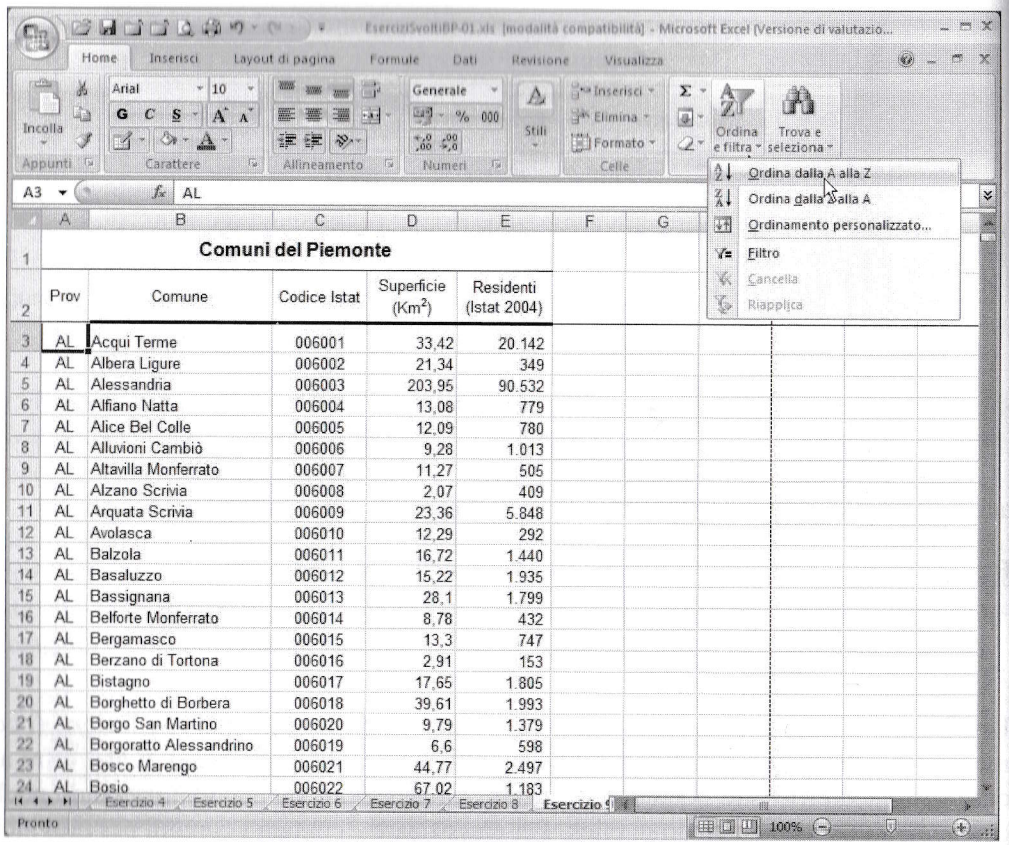


Figura 1.23 Il comando per ordinare le righe della tabella rispetto a una colonna.

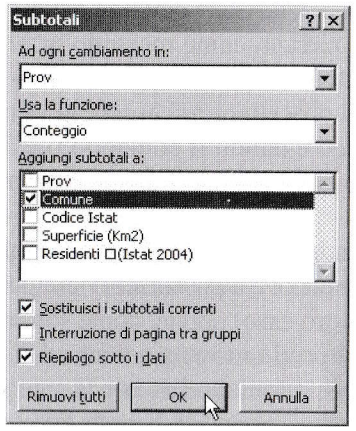


Figura 1.24 Le impostazioni per inserire subtotali ed effettuarne il conteggio.

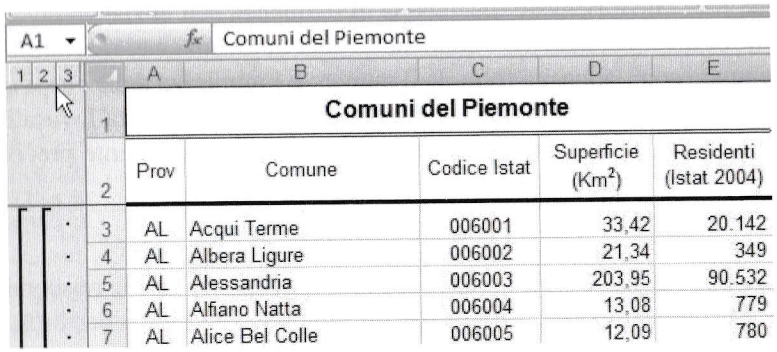


Figura 1.25 Cliccando su uno dei livelli in cui Excel ha suddiviso le righe della tabella si ottengono i subtotali.

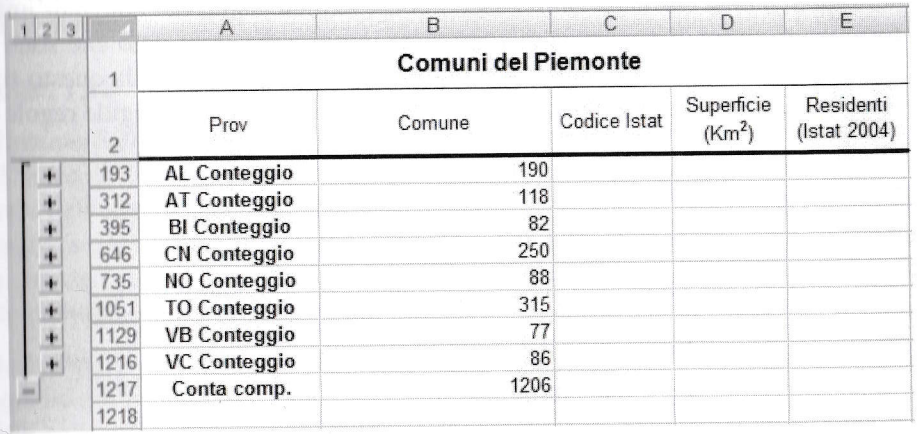


Figura 1.26 La tabella è stata raggruppata in modo automatico in base alle province.

contrazione di ogni raggruppamento creato. Cliccando sul pulsante 3 si ritornerà alla situazione di partenza, mentre cliccando sul pulsante 1 l'intera tabella si ridurrà ad una sola riga contenente il conteggio di tutti i comuni della regione. Excel ha creato una struttura che, utilizzando le province, suddivide in tre possibili livelli la tabella: al livello 1 (pulsante 1) si ottiene una sola riga con il riepilogo di tutta la tabella. Al livello due (pulsante 2) avremo 8 righe come mostrato in Figura 1.26 (la regione Piemonte si suddivide in 8 province) ed infine al livello 3 la tabella riappare in tutta la sua estensione. Osservando la tabella si noti che utilizzando altre funzioni si può ricavare il totale degli abitanti della provincia o la popolazione media. Le varianti ottenibili sono molto numerose.

Esercizio 1.9

Calcolare l'anzianità in anni e rappresentarla con una tabella di frequenze per classi.

L'ufficio personale vuole ricavare una tabella di frequenza dell'anzianità lavorativa dei 62 dipendenti. Il foglio "Esercizio 1.9" contiene la data di assunzione, ricavare l'anzianità in anni e determinare con l'apposita formula il numero e l'ampiezza delle classi di frequenza.

Questo esempio contiene come novità il calcolo dell'anzianità lavorativa a partire dalla data di assunzione. La si può ottenere in modo sufficientemente preciso utilizzando la formula mostrata in Figura 1.27:

```
=INT((OGGI()-B2)/365,25)
```

dove il termine `OGGI()-B2` serve a calcolare il numero di giorni trascorsi tra la data odierna e la data di assunzione. Dividendo questa durata in giorni per `365,25` si ottiene la durata in anni (un anno è composto da `365,25` giorni) e di questa durata la funzione `INT()`ne prende solo la parte intera escludendo la parte decimale.

Ottenuta così l'anzianità dei dipendenti si può procedere secondo quanto visto negli esercizi precedenti calcolando i valori necessari a determinare il numero delle classi k e la loro ampiezza (Figura 1.28). Si osservi come le indicazioni ottenute dall'impiego delle formule, 6 classi di ampiezza 5, sono state in parte disattese dato che sono state utilizzate 7 classi. In genere è corretto interpretare i risultati ottenuti da questo tipo di calcoli come indicazioni attendibili, ma pur sempre indicazioni e non rigide regole a cui

	A	B	C	D	E	F
1	Dipendente	Data assunzione	Anzianità			
2	1	23/03/1984	22	< =INT((OGGI()-B2)/365,25)		
3	2	24/11/1996	10			
4	3	23/11/2003	3			
5	4	04/05/2004	2			
6	5	14/04/1985	21			
7	6	15/06/1998	8			
8	7	13/09/1993	13			

Figura 1.27 La formula per il calcolo di una durata impiega la funzione OGGI() che restituisce la data corrente.

	E	F	G	H	I
1					
2	Minimo:	1	< =MIN(C2:C63)		
3	Massimo:	31	< =MAX(C2:C63)		
4	N:	62	< =CONTA.VALORI(C2:C63)		
5	k:	6	< =INT(1+LOG(F4;2))		
6	ampiezza:	5,0	< =(F3-F2)/F5		
7					

Figura 1.28 Il calcolo dei valori necessari per determinare numerosità e ampiezza delle classi

	E	F	G	H	I	J	K
8	Anzianità in classi		Frequenza	Freq %		Freq % cumul	
9	Meno di 5	5	13	21,0%	< =G9/$G$16	21,0%	< =H9
10	Da 5 a 10	10	12	19,4%		40,3%	< =H10+J9
11	Da 10 a 15	15	11	17,7%		58,1%	< =H11+J10
12	Da 15 a 20	20	13	21,0%		79,0%	< =H12+J11
13	Da 20 a 25	25	7	11,3%		90,3%	< =H13+J12
14	Da 25 a 30	30	4	6,5%		96,8%	< =H14+J13
15	Da 30 a 35	35	2	3,2%		100,0%	< =H15+J14
16		Totale:	62				

Figura 1.29 La tabella delle frequenze assolute, relative e cumulate per la variabile Anzianità.

sottostare. Considerare una classe in più ci ha consentito di mantenere l'ampiezza di 5 anni e di includere tutti i casi osservati senza dover "tirare" la classe posta all'estremità, estendendone l'ampiezza per far rientrare tutti i casi.

Il risultato finale con le formule da impiegare è mostrato in Figura 1.29.

Esercizio 1.10

Gradimento della mensa universitaria, calcolo di frequenze relative e cumulate

L'Ente diritto allo studio della regione vuole valutare il servizio mensa offerto agli studenti dell'ateneo del capoluogo e per questo fa girare tra gli utenti un questionario. I dati sono raccolti nel foglio Excel "Esercizio 1.10". Creare una tabella delle distribuzioni di frequenza per le cinque risposte possibili. Dopo aver calcolato le frequenze relative e quelle cumulate rispondere ai quesiti seguenti:

- Quale frequenza percentuale ha il punteggio più alto?
- E quello più basso?
- Sulla base delle risposte come risulta l'apprezzamento del cibo della mensa? Motivare la risposta.

Suggerimenti per lo svolgimento:
la tabella con le frequenze assolute, relative e cumulate è mostrata in Figura 1.30. L'esercizio non presenta alcuna novità rispetto a quelli finora visti.

	E	F	G	H	I	J	K
8	Anzianità in classi		Frequenza	Freq %		Freq % cumul	
9	Meno di 5	5	13	21,0%	< =G9/$G$16	21,0%	< =H9
10	Da 5 a 10	10	12	19,4%		40,3%	< =H10+J9
11	Da 10 a 15	15	11	17,7%		58,1%	< =H11+J10
12	Da 15 a 20	20	13	21,0%		79,0%	< =H12+J11
13	Da 20 a 25	25	7	11,3%		90,3%	< =H13+J12
14	Da 25 a 30	30	4	6,5%		96,8%	< =H14+J13
15	Da 30 a 35	35	2	3,2%		100,0%	< =H15+J14
16		Totale:	62				

Figura 1.30 Tabella delle frequenze assolute, relative e percentuali.