TEMA D

Dati e previsioni

La statistica è una disciplina all'ordine del giorno. Basta sfogliare un giornale per trovare dati statistici sui prezzi dei prodotti in commercio, sugli apprezzamenti dei programmi televisivi e anche su risultati politici o economici. In medicina, questa disciplina è uno strumento di fondamentale importanza per classificare e analizzare la diffusione delle patologie e i loro legami con i fattori che le determinano. Inoltre, tutti i settori della scienza impiegano i metodi della statistica per ordinare e analizzare i dati numerici ottenuti negli esperimenti.

Studiando i metodi di presentazione dei dati della statistica scopriremo come, mediante opportuni grafici, sia possibile rendere intuitivi dati che altrimenti non sarebbero altro che una sterile successione di numeri. Però proprio su questo punto bisogna prestare attenzione: infatti, in alcuni casi chi trae delle conclusioni con grafici o percentuali lo fa in maniera non corretta, scegliendo la rappresentazione dei dati in base al risultato che vuole suggerire. Studiando questa Unità capiremo perciò quanto sia importante saper leggere i dati con metodi corretti, senza dover dipendere dall'interpretazione di altri.

PREREQUISITI

- Numeri reali e intervalli
- Le quattro operazioni
- Le percentuali
- Estrazione di radice
- Ordini di grandezza

COMPETENZE

- Utilizzare il linguaggio e i metodi propri della matematica per organizzare e valutare adeguatamente informazioni qualitative e quantitative

Tutte le nazioni sviluppate del mondo dispongono di un istituto nazionale di statistica, che studia per esempio lo sviluppo demografico della popolazione, l'andamento della produzione agricola e industriale e tanti altri aspetti fondamentali per la gestione di un Paese. In Italia l'istituto nazionale di statistica, l'ISTAT, è presente dal 1926 ed è la principale fonte di statistica ufficiale a supporto dei cittadini e delle istituzioni.

Unità 10
Richiami e complementi di statistica

Unità 10

Richiami e complementi di statistica

1. Introduzione alla statistica

Il linguaggio della statistica

Attenzione!
In questo e nel prossimo paragrafo richiamiamo sinteticamente i principali concetti di statistica che hai appreso nel primo biennio.

Il primo passo per iniziare lo studio della statistica è conoscere il significato di alcuni termini specifici.

POPOLAZIONE E UNITÀ STATISTICA
L'insieme degli individui oggetto di un'indagine statistica si chiama **popolazione** (o **universo** o **collettivo**); ciascun individuo facente parte della popolazione viene chiamato anche **unità statistica**.

Alcune indagini statistiche consentono di interpellare *tutti* i membri della popolazione; in altri casi, per motivi di costi e di tempi, bisogna limitarsi a sottoporre le domande o le richieste di informazioni solo a *una parte* della popolazione, che viene chiamata **campione**.

CARATTERE
Si chiama **carattere** la proprietà che è oggetto di studio in un'indagine statistica.

Per esempio, il *peso* di una persona può essere un possibile carattere di un'indagine. In corrispondenza di ogni individuo della popolazione, il carattere oggetto di studio assume una determinata *modalità*; per esempio, il carattere «peso» può assumere in corrispondenza di un dato individuo la modalità 72 kg, in corrispondenza di un altro la modalità 80 kg e così via.

MODALITÀ
Si chiama **modalità** ciascuna delle varianti con cui un carattere può presentarsi; le modalità osservate si chiamano **dati.**

I caratteri si classificano secondo le seguenti definizioni.

Modi di dire
«Variabile» e «mutabile» sono, in questo contesto, sostantivi. Si usa talvolta parlare di **valori** di una variabile, anziché di **modalità** di una variabile.

CARATTERI QUANTITATIVI E CARATTERI QUALITATIVI
Un carattere le cui modalità sono espresse da numeri è detto **carattere quantitativo** (o **variabile**), un carattere le cui modalità **non** sono espresse da numeri è detto **carattere qualitativo** (o **mutabile**).

ESEMPI

Caratteri quantitativi	Caratteri qualitativi
• I giorni di assenza di uno studente in un anno scolastico • La quantità, in kilogrammi, di mele vendute da un negozio in un giorno • Il numero di multe dato in un giorno dai vigili di una città	• Il gusto di gelato preferito • Il tipo di alimentazione (benzina o diesel) delle macchine che ci sono in un parcheggio • Il colore delle automobili vendute da un'agenzia in un giorno

Le *variabili* (ovvero i caratteri *quantitativi*) si classificano ulteriormente a seconda del tipo di valori che possono assumere.

VARIABILI DISCRETE E VARIABILI CONTINUE

Una variabile si dice **discreta** quando può assumere soltanto un numero **finito** di valori (o un insieme di valori che può essere posto in corrispondenza biunivoca con l'insieme dei numeri naturali); si dice **continua** quando può assumere (almeno teoricamente) tutti i valori reali di un determinato intervallo.

Tipicamente, le variabili discrete sono quelle che si rilevano *contando* (per esempio il numero dei figli in una famiglia o il numero dei dipendenti di un'azienda), mentre le variabili continue sono quelle che si rilevano mediante *misurazioni* (per esempio il peso di un bambino a una certa età o la temperatura massima giornaliera registrata a Milano in un dato giorno).

ESEMPI RIASSUNTIVI

Fenomeno studiato	Popolazione	Carattere	Modalità	Tipo di carattere
Il colore degli occhi degli italiani	Tutti gli italiani	Il colore degli occhi	Verdi, azzurri, marroni ecc.	Qualitativo
Altezza (misurata in metri) degli studenti di una classe	Gli studenti della classe	La misura dell'altezza	1,72 m; 1,85 m; 1,78 m	Quantitativo continuo
L'anno di nascita degli iscritti a una palestra	Tutti gli iscritti alla palestra	L'anno di nascita	..., 1970, ..., 1981, 1965, ...	Quantitativo discreto

Distribuzioni di frequenze

Ricordiamo anzitutto alcune definizioni.

Termine	Significato
Frequenza (assoluta) di una modalità	Il numero di volte in cui una modalità è stata osservata
Frequenza relativa di una modalità	Il rapporto tra la frequenza assoluta della modalità e il numero di individui della popolazione
Frequenza percentuale di una modalità	La rappresentazione in percentuale della frequenza relativa
Frequenza cumulata di una modalità di un carattere quantitativo	La somma delle frequenze di tutte le modalità minori o uguali a quella considerata

Una prima forma di elaborazione dei dati, volta a ottenere una maggiore *sintesi*, consiste nel costruire una tabella in cui riportare, per ciascuna delle *modalità* differenti x_1, x_2, ..., x_k osservate, la rispettiva frequenza assoluta; questa tabella, in cui si aggiunge di solito un'ultima riga in cui si riporta il totale n delle frequenze (*vedi* la tabella qui a fianco), è detta **distribuzione di frequenze** del carattere esaminato. Il numero n rappresenta il numero complessivo di unità della popolazione. Nella costruzione della tabella bisogna ricordare, se il carattere è quantitativo, di ordinare le modalità in senso crescente.

In modo del tutto analogo si possono costruire le **distribuzioni di frequenze relative** o **percentuali**, sostituendo le frequenze assolute con quelle relative o percentuali.

X	Frequenze
x_1	f_1
x_2	f_2
.....	.....
x_k	f_k
Totale	n

ESEMPIO **Distribuzione di frequenze**

Supponiamo di avere rilevato, in una classe di una scuola, il colore degli occhi degli allievi. Il risultato della rilevazione fornisce i così detti **dati grezzi**, che abbiamo raccolto nella seguente tabella in cui a ogni *unità statistica* (cioè a ogni studente) è stata associata la *modalità* del carattere osservata, cioè il colore nero (N), marrone (M) azzurro (A) o verde (V).

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Colore degli occhi	N	M	A	N	N	M	A	V	V	M	M	N	N	V	V	N	M	N

Associando a ogni modalità la sua *frequenza* assoluta, la tabella dei dati grezzi si sintetizza nella seguente, che rappresenta la distribuzione di frequenze del carattere esaminato.

Colore degli occhi	**Numero di studenti**
Nero	7
Marrone	5
Azzurro	2
Verde	4
Totale	18

In alcuni casi, prima di costruire la tabella che rappresenta la distribuzione di frequenze, è utile *accorpare* le modalità in *intervalli* tra loro *disgiunti*, detti **classi**.

ESEMPIO **Distribuzione di frequenze suddivisa per classi**

Nella stessa classe in cui prima abbiamo rilevato il *colore degli occhi* degli studenti, rileviamo ora la loro *statura*. Il risultato della rilevazione fornisce i dati grezzi riassunti nella seguente tabella.

Studente	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Altezza (in cm)	173	164	174	180	182	176	184	185	170	172	186	167	188	183	168	176	184	178

Suddividendo le possibili altezze degli studenti (misurate in cm) nei seguenti intervalli:

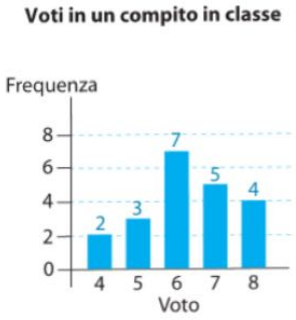
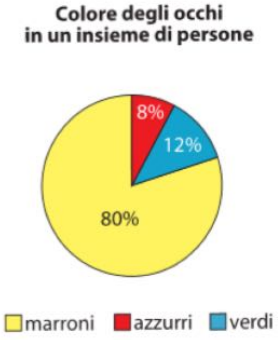
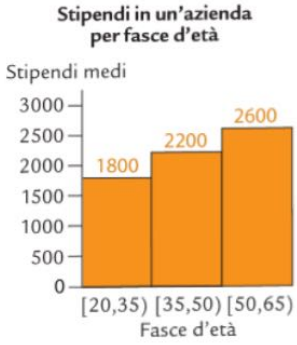
(160, 165]; (165, 170]; (170, 175]; (175, 180]; (180, 185]; (185, 190]

otteniamo la distribuzione di frequenze rappresentata nella tabella qui a fianco.

Altezza (in cm)	**Frequenza**
(160, 165]	1
(165, 170]	3
(170, 175]	3
(175, 180]	4
(180, 185]	5
(185, 190]	2

Principali rappresentazioni grafiche

Esiste una grande varietà di grafici utilizzati in statistica, ma i più importanti si possono ricondurre alle quattro categorie di cui puoi vedere alcuni esempi nella tabella seguente: diagrammi a barre, diagrammi circolari, diagrammi cartesiani e istogrammi.

Diagramma a barre (rettangoli distanziati)	Diagramma circolare	Diagramma cartesiano	Istogramma (rettangoli affiancati)
			

Prova tu **ESERCIZI** a p. 389

a. Nell'insieme degli studenti che hanno superato l'esame di Stato in una data scuola si esegue un'indagine statistica che ha per oggetto il voto conseguito. Indica, per questa indagine statistica: la popolazione, il carattere, alcune possibili modalità del carattere, e stabilisci se il carattere è quantitativo o qualitativo; in caso sia quantitativo, specifica se è discreto o continuo.

b. Considera la tabella dei dati grezzi dell'ultimo esempio di questo paragrafo (relativa alla rilevazione delle altezze in una classe di studenti). Suddividi i dati nelle classi (160, 170], (170, 180] , (180, 190] e costruisci una tabella che rappresenti la distribuzione delle frequenze assolute e relative delle classi.

2. Indici di posizione e di variabilità

Richiamiamo in questo paragrafo le formule e i metodi per determinare i principali indici di *posizione* (media, moda, mediana) e di *variabilità* (varianza e deviazione standard). Gli indici di posizione e di variabilità consentono di sintetizzare in pochi numeri significativi le principali caratteristiche del fenomeno indagato. In particolare, gli indici di *variabilità* forniscono informazioni sull'attitudine di un fenomeno a manifestarsi sulle varie unità statistiche con modalità diverse e distanti tra loro.

Il caso in cui è data una distribuzione di dati grezzi

Ci riferiamo a un carattere *quantitativo* X, di cui sono stati osservati i valori x_1, x_2, ..., x_n.

Termine	Definizione	Esempio
Media aritmetica (indicata con $\overline{x}$ o con μ)	$\mu = \overline{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$	La media dei tre numeri 2, 4 e 6 è: $\overline{x} = \dfrac{2+4+6}{3} = 4$

Termine	Definizione	Esempio
Mediana	Ordinati i numeri $x_1, x_2, \ldots, x_n$ in senso crescente (o decrescente) la loro mediana è: – il numero che occupa la posizione centrale, se n è *dispari*; – la media aritmetica dei due numeri che occupano le posizioni centrali, se n è *pari*.	La mediana dei tre numeri: 4, 5, 6 $\quad n = 3$ (dispari) è il numero 5. La mediana dei quattro numeri: 4, 5, 6, 7 $\quad n = 4$ (pari) è la media aritmetica dei due numeri che occupano le posizioni centrali, quindi è $\frac{5+6}{2} = \frac{11}{2}$.
Moda	Il dato (o i dati) che hanno la massima frequenza.	Dati i numeri: 1, 2, 3, 4, 3 la moda è il numero 3, che compare con frequenza massima, uguale a 2.
Varianza (indicata con V o con σ^2)	$\sigma^2 = V = \frac{(x_1 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n}$ oppure $\sigma^2 = V = \frac{x_1^2 + \ldots + x_n^2}{n} - \overline{x}^2$	Consideriamo i due numeri 4 e 8. La loro media aritmetica è: $\overline{x} = \frac{4+8}{2} = 6$ La loro varianza, in base alla seconda formula, è: $\sigma^2 = V = \frac{4^2 + 8^2}{2} - 6^2 = 4$
Deviazione standard	$\sigma = \sqrt{V}$	La deviazione standard di 4 e 8, in base alla varianza poc'anzi calcolata, è $\sigma = \sqrt{4} = 2$.

Il caso in cui è data una distribuzione di frequenze

Nel caso che sia assegnata una distribuzione di frequenze, occorre tenere presente che:

- per il calcolo della *media* e della *varianza* si utilizzano le formule seguenti:

$$\overline{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_k f_k}{f_1 + f_2 + \ldots + f_k} \qquad \sigma^2 = \frac{x_1^2 \cdot f_1 + \ldots + x_k^2 \cdot f_k}{f_1 + \ldots + f_k} - \overline{x}^2$$

  dove $x_1, x_2, \ldots, x_k$ sono i valori osservati, rispettivamente con frequenze $f_1, f_2, \ldots, f_k$, del carattere (quantitativo) X in esame;
- per il calcolo della *mediana* conviene ricorrere al calcolo delle frequenze cumulate, come spiegato nel prossimo esempio.

ESEMPIO **Valori medi nel caso di una distribuzione di frequenze**

Un'indagine effettuata su un campione di famiglie ha prodotto la distribuzione di frequenze rappresentata nella tabella qui sotto. Determiniamo la media, la mediana e la moda della distribuzione.

Numero di figli per famiglia	Frequenza
0	9
1	27
2	40
3	20
4	3
5	1

- Il numero medio di figli per famiglia è dato dalla formula:

$$\bar{x} = \frac{0 \cdot 9 + 1 \cdot 27 + 2 \cdot 40 + 3 \cdot 20 + 4 \cdot 3 + 5 \cdot 1}{9 + 27 + 40 + 20 + 3 + 1} = \frac{184}{100} = 1{,}84$$

- Per il calcolo della mediana determiniamo preliminarmente le frequenze cumulate:

Numero di figli per famiglia	Frequenza	Frequenza cumulata
0	9	9
1	27	$9 + 27 = 36$
2	40	$36 + 40 = 76$
3	20	$76 + 20 = 96$
4	3	$96 + 3 = 99$
5	1	$99 + 1 = 100$

Il numero complessivo di famiglie intervistate è $n = 100$ (ultima frequenza cumulata). Le due famiglie che occupano le posizioni centrali sono la cinquantesima e la cinquantunesima. Dalla colonna delle frequenze cumulate deduciamo che le famiglie dalla numero 37 alla numero 76 hanno 2 figli, quindi in particolare hanno 2 figli le famiglie corrispondenti alle due posizioni centrali. La mediana è, per definizione, la media fra i figli di queste due famiglie, dunque è 2.

- La moda della distribuzione è chiaramente «2», che corrisponde alla massima frequenza (uguale a 40).

Rifletti

1. Il significato della mediana è il seguente: almeno il 50% delle famiglie hanno un numero di figli maggiore o uguale a 2 e almeno il 50% delle famiglie hanno un numero di figli minore o uguale a 2.
2. Nell'esempio qui a fianco la mediana e la moda coincidono. In generale non è detto che ciò avvenga.

Il caso in cui è data una distribuzione suddivisa per classi

Se è data una distribuzione di frequenze **suddivisa per classi**:

a. si assume come *media* della distribuzione il valore che si ottiene sostituendo ciascuna classe con il suo *valore centrale* (cioè con la semisomma degli estremi della classe) e calcolando la media della distribuzione di frequenze così ottenuta;
b. si assume come *mediana* il valore centrale della classe che contiene la mediana;
c. si assume come *classe modale* quella che ha maggiore frequenza se le classi hanno la stessa ampiezza, e quella che ha maggiore *densità di frequenza* in caso contrario.

Ricorda

Si chiama **densità di frequenza** il rapporto tra la frequenza della classe e la sua ampiezza.

ESEMPIO Valori medi nel caso di una distribuzione suddivisa in classi

Un'indagine effettuata su un campione di individui ha prodotto la seguente distribuzione di frequenze. Determiniamo media, mediana e moda della distribuzione.

Peso (in kg)	Frequenza
$40 \leq p < 50$	16
$50 \leq p < 60$	48
$60 \leq p < 70$	45
$80 \leq p < 90$	8
$90 \leq p < 100$	3

→

a. Sostituendo ogni classe con il suo valore centrale otteniamo la seguente distribuzione di frequenze.

Per esempio, il valore centrale della classe $40 \leq p < 50$ è:
$\frac{40+50}{2} = 45$

Peso (in kg)	Frequenza
45	16
55	48
65	45
85	8
95	3

A questo punto il peso medio $\overline{p}$ può essere ricavato con una media aritmetica ponderata:

$$\overline{p} = \frac{45 \cdot 16 + 55 \cdot 48 + 65 \cdot 45 + 85 \cdot 8 + 95 \cdot 3}{16 + 48 + 45 + 8 + 3} = \frac{7250}{120} \simeq 60{,}4 \text{ kg}$$

b. Per individuare la classe che contiene la mediana della distribuzione è utile calcolare le *frequenze cumulate*:

Peso (in kg)	Frequenza	Frequenza cumulata
$40 \leq p < 50$	16	16
$50 \leq p < 60$	48	64
$60 \leq p < 70$	45	109
$80 \leq p < 90$	8	117
$90 \leq p < 100$	3	120

Il collettivo è composto complessivamente da 120 individui (*pari*), la mediana è data perciò dalla media fra il sessantesimo e il sessantunesimo peso osservato. Dalla colonna delle frequenze cumulate si deduce che i pesi osservati dal numero 17 al numero 64 appartengono alla classe $50 \leq p < 60$. Pertanto anche la mediana appartiene a tale classe. Come approssimazione della mediana prendiamo il valore centrale di tale classe: $\frac{50+60}{2} = 55$. Il peso mediano è quindi 55 kg.

c. Dal momento che le classi hanno la stessa ampiezza (uguale a 10), la classe modale è quella che ha maggiore frequenza, ossia $50 \leq p < 60$.

Prova tu **ESERCIZI** a p. 392

Considera il seguente insieme di dati:

1, 4, 7, 2, 3, 4, 4, 8, 8, 3, 1, 1, 2, 5, 4

Determina la media, la mediana e la moda. [Media = 3,8; mediana = moda = 4]

3. Tabelle a doppia entrata

Nei paragrafi precedenti abbiamo richiamato le nozioni fondamentali relative alla statistica **univariata**, ossia a quella parte della statistica che si occupa dell'analisi dei dati provenienti dalla rilevazione di *un solo* carattere su una data popola-

zione. In questo e nei prossimi paragrafi vogliamo invece introdurre le nozioni di base relative alla statistica **bivariata**: vogliamo cioè vedere come si estendono le nozioni della statistica univariata quando vengono rilevati congiuntamente *due* caratteri, diciamo X e Y. L'obiettivo ulteriore che ci porremo, in questo nuovo contesto, sarà quello di scoprire e mettere in luce eventuali *relazioni* tra X e Y.

Distribuzioni congiunte e marginali

Supponiamo dunque che due caratteri X e Y siano osservati (insieme) su ciascuna delle n unità che compongono la popolazione in esame.
Il risultato della rilevazione è un insieme di coppie ordinate (x, y) che possono essere rappresentate in una tabella come quella qui sotto, detta **tabella dei dati grezzi**:

Unità statistiche	Modalità di X rilevata	Modalità di Y rilevata
1	x_1	y_1
2	x_2	y_2
.....	.....	.....
n	x_n	y_n

ESEMPIO **Tabella dei dati grezzi**

Su una popolazione formata da cinque amici si sono rilevati due caratteri: l'età (X) e la città di nascita (Y). I dati grezzi possono essere organizzati nella tabella sottostante.

Nome	Età	Città di nascita
Alberto	30	Milano
Maria	35	Torino
Giovanni	32	Milano
Paola	30	Milano
Alessandro	32	Roma

Nell'ambito della statistica univariata abbiamo visto che per rendere i dati grezzi meglio leggibili è utile costruire la tabella che ne rappresenta la distribuzione di frequenze. Similmente si procede nell'ambito della statistica bivariata, costruendo una **tabella a doppia entrata**, che riporti le frequenze con cui si manifestano le varie *coppie* di modalità osservate.
Più precisamente, supponiamo che il carattere X abbia manifestato le modalità distinte:

$$x_1, x_2, \ldots, x_k$$

e che il carattere Y abbia manifestato le modalità distinte:

$$y_1, y_2, \ldots, y_h$$

Costruiamo una tabella di $k+1$ righe e $h+1$ colonne convenendo di riportare nella prima *colonna* le modalità $x_1, x_2, \ldots, x_k$ di X e nella prima *riga* le modalità $y_1, y_2, \ldots, y_h$ di Y. Nella casella all'incrocio tra una riga, diciamo la i, e una colonna, diciamo la j, riporteremo la frequenza assoluta della coppia (x_i, y_j), che nel seguito indicheremo con il simbolo $f(x_i, y_j)$. Le frequenze assolute di queste coppie sono dette **frequenze congiunte** e la tabella così costruita viene detta **distribuzione doppia di frequenze**.

Altre notazioni

La frequenza congiunta della coppia (x_i, y_j) viene spesso indicata anche con il simbolo f_{ij}.

ESEMPIO **Distribuzione doppia di frequenze**

In riferimento all'esempio precedente, il carattere X (*età*) manifesta tre modalità distinte: 30, 32, 35; così pure il carattere Y (*città di nascita*) manifesta tre modalità distinte: Milano, Torino, Roma. I dati grezzi possono essere organizzati nella seguente tabella a doppia entrata, costruita secondo le modalità poc'anzi descritte.

X \ Y	Milano	Torino	Roma
30	2	0	0
32	1	0	1
35	0	1	0

La tabella a doppia entrata che rappresenta le frequenze congiunte di X e Y si completa di solito con un'ultima riga, dove vengono riportate le *somme* delle frequenze di ciascuna colonna, e un'ultima colonna, dove vengono riportate le *somme* delle frequenze di ciascuna riga. L'ultima riga e l'ultima colonna rappresentano le cosiddette **distribuzioni marginali** dei due caratteri, cioè le distribuzioni di X e Y che si avrebbero se ciascuno di essi fosse stato rilevato *singolarmente*.
All'incrocio dell'ultima riga e dell'ultima colonna si pone il numero complessivo di unità della popolazione.

ESEMPIO **Distribuzioni marginali**

In riferimento alla tabella dell'esempio precedente, ne risultano le distribuzioni marginali messe in evidenza sull'ultima riga e sull'ultima colonna:

X \ Y	Milano	Torino	Roma	Totale
30	2	0	0	2
32	1	0	1	2
35	0	1	0	1
Totale	3	1	1	5

Distribuzione marginale di X

Distribuzione marginale di Y

Numero complessivo di unità del collettivo

Osserva

Il numero complessivo di unità del collettivo è uguale sia alla somma delle frequenze marginali di X, sia alla somma delle frequenze marginali di Y.

Altre notazioni

La frequenza marginale $f(x_i)$ viene indicata anche con il simbolo $f_{i\cdot}$ e la frequenza marginale $f(y_j)$ con il simbolo $f_{\cdot j}$. Il punto posto prima o dopo l'indice indica il fenomeno che è stato trascurato, per ricordare che si sta lavorando con frequenze marginali.

Nel seguito indicheremo le frequenze (dette **frequenze marginali**) che costituiscono le distribuzioni marginali di X e Y rispettivamente con i simboli $f(x_1)$, ..., $f(x_k)$ e $f(y_1)$, ..., $f(y_h)$. È inoltre possibile ottenere le **distribuzioni marginali relative** di X e Y costruendo i rapporti tra le frequenze marginali e il numero complessivo di unità del collettivo.

Distribuzioni condizionate

Facciamo ancora riferimento alla tabella dell'ultimo esempio. Se per esempio fissiamo l'attenzione sulla riga corrispondente alla prima modalità di X, $x_1 = 30$, leggiamo come si distribuisce il carattere Y *tra le unità della popolazione che manifestano la modalità* x_1 *di* X. Per questo motivo si dice che tale riga rappresenta la **distribuzione condizionata** di Y rispetto alla modalità x_1 di X.

Distribuzione condizionata di Y rispetto alla modalità $x_1 = 30$ di X

X \ Y	Milano	Torino	Roma	Totale
30	2	0	0	2
32	1	0	1	2
35	0	1	0	1
Totale	3	1	1	5

In simboli

La distribuzione di X condizionata a una modalità y_j di Y si indica con il simbolo $X|y_j$.
Analogamente va interpretato il simbolo $Y|x_i$.

Più in generale, fissare l'attenzione su una sola riga (o colonna) della tabella (escludendo quelle delle modalità e dei totali) significa restringersi alla sottopopolazione che presenta una data modalità di X (o di Y): ciascuna di queste righe o colonne, singolarmente presa, rappresenta perciò una particolare **distribuzione condizionata**.
È anche possibile costruire le **distribuzioni condizionate relative**, ponendo a rapporto le frequenze congiunte che appartengono alla distribuzione condizionata considerata con i corrispondenti totali di riga o di colonna. Per esempio, in riferimento alla tabella sopra, la distribuzione condizionata di X rispetto a y_1 e la corrispondente distribuzione condizionata relativa sono rappresentate qui sotto.

X	Distribuzione condizionata $X\|y_1$	Distribuzione condizionata relativa
30	2	2/3
32	1	1/3
35	0	0/3
Totale	**3**	**1**

SINTESI

- I dati grezzi raccolti in seguito alla rilevazione congiunta di due caratteri X e Y possono essere organizzati in una tabella a doppia entrata del tipo seguente, che rappresenta la distribuzione doppia di frequenze di X e Y:

Frequenza assoluta con cui la modalità x_i si presenta congiuntamente alla modalità y_j

X \ Y	y_1	....	y_j	.....	y_h	Totale
x_1	....	....	....	....	....	$f(x_1)$
....	....	....	....	....	....	....
x_i	....	....	$f(x_i, y_j)$	....	....	....
....	....	....	....	....	....	....
x_k	....	....	....	....	....	$f(x_k)$
Totale	$f(y_1)$	....	....	....	$f(y_h)$	n

Somme per riga: forniscono la distribuzione marginale di X

Somme per colonna: forniscono la distribuzione marginale di Y

Numero complessivo di unità del collettivo

- L'organizzazione dei dati in una tabella a doppia entrata permette di riassumere molti tipi di informazioni:
  1. il comportamento *congiunto* di X e di Y: è rappresentato nelle caselle della tabella non appartenenti ai bordi, dove sono riportate le frequenze *congiunte* di X e Y;

2. il comportamento di X e Y, considerati *singolarmente*: è rappresentato sull'ultima colonna e sull'ultima riga della tabella, dove sono riportate le frequenze *marginali* di X e Y;
3. il comportamento di un carattere (X o Y) condizionatamente a una modalità dell'altro: è rappresentato sulle righe e sulle colonne *interne* alla tabella, considerate singolarmente.

Prova tu **ESERCIZI** a p. 395

Sono stati intervistati i 20 studenti di una classe e su ciascuno sono stati rilevati congiuntamente i due caratteri X: il sesso (M = maschio, F = femmina) e Y: il numero di ore che dedicano mediamente allo studio in una giornata. Si sono ottenuti i dati nella seguente tabella:

X	M	M	F	M	M	F	M	F	M	F	F	M	F	M	F	M	F	M	M	F
Y	1	2	2	4	3	1	2	3	3	4	2	3	4	2	1	2	3	3	3	4

a. Costruisci una tabella a doppia entrata che organizzi i dati grezzi e individua le distribuzioni marginali dei due caratteri X e Y.
b. Determina la distribuzione di X, condizionata alla modalità «3 ore di studio al giorno» e la corrispondente distribuzione condizionata relativa.
c. Determina la distribuzione di Y, condizionata alla modalità «femmina» e la corrispondente distribuzione condizionata relativa.

4. Dipendenza e indipendenza statistica

Come già anticipato, lo studio statistico di due caratteri X e Y, rilevati congiuntamente su una data popolazione, si pone tra i vari obiettivi anche quello di stabilire se sussiste qualche *relazione di dipendenza* tra X e Y.
I metodi che esporremo in questo paragrafo per valutare l'eventuale dipendenza tra due caratteri possono essere applicati sia a caratteri quantitativi sia a caratteri qualitativi, perché faranno riferimento solo alle *frequenze*; tuttavia, nella pratica si utilizzano prevalentemente per caratteri di tipo *qualitativo*, poiché per caratteri quantitativi esistono strumenti statistici più adeguati (che presenteremo nel prossimo paragrafo).

Dipendenza e indipendenza

Dati due caratteri X e Y, per stabilire se X dipende o meno da Y viene naturale l'idea di confrontare le distribuzioni di X *condizionate* alle modalità di Y con la distribuzione *marginale* di X (che esprime il comportamento di X considerato singolarmente). Se c'è indipendenza, c'è da aspettarsi che il condizionamento di X alle modalità di Y non abbia alcun effetto, ossia che le distribuzioni condizionate si mantengano uguali a quella marginale. Occorre però prestare attenzione a un aspetto: le frequenze *marginali* si riferiscono all'*intera popolazione*, mentre le frequenze *condizionate* si riferiscono soltanto alla *sottopopolazione* che presenta la modalità rispetto cui stiamo condizionando. Non sarebbe perciò corretto eseguire il confronto tra le frequenze assolute: il confronto deve essere fatto tra frequenze *relative*. Queste considerazioni portano alla seguente definizione.

INDIPENDENZA
Il carattere X si dice **indipendente** da Y se le distribuzioni **condizionate relative** di X rispetto alle modalità di Y sono uguali alla distribuzione **marginale relativa** di X.