7

Elementi di statistica descrittiva

7.0 Scopi del capitolo

Il termine *statistica* venne introdotto nel diciassettesimo secolo col significato di *scienza dello stato*, volta a raccogliere e ordinare informazioni utili all'amministrazione pubblica: entità e composizione della popolazione, movimenti migratori, mutamenti anagrafici, tavole di natalità e mortalità, dati sui commerci, sui raccolti, sulla distribuzione della ricchezza, sull'istruzione e la sanità. Oggigiorno la statistica si applica a tutte le scienze sperimentali.

Il primo passo dell'attività statistica è la raccolta di dati che, se ben organizzata, permette la corretta impostazione del lavoro di analisi. Si dice *unità statistica* la minima unità della quale si raccolgono i dati. Si dice *popolazione* l'insieme delle unità statistiche oggetto di studio. Si dicono *caratteri* o *variabili* ciò che si rileva sulla popolazione. Chiamiamo *modalità* i possibili valori che può assumere un dato carattere. Per esempio, se il carattere indica il colore degli occhi, le modalità sono : castani, chiari,

neri etc. I caratteri possono essere *qualitativi* o *quantitativi*. Per esempio, sono caratteri qualitativi lo stato civile (celibe o nubile, coniugato/a, etc.) o il sesso (maschio o femmina). I caratteri quantitativi sono esprimibili numericamente e si dividono in *discreti* e *continui*. I caratteri discreti, come il numero degli alunni di una classe, o di reti segnate in una partita di calcio, possono assumere solo determinati valori, quasi sempre numeri interi. I caratteri continui, quali i pesi, le stature e più in generale le grandezze che possono essere misurate, possono assumere qualsiasi valore reale in un dato intervallo (anche se usualmente si impiegano numeri decimali finiti).
In questo capitolo introdurremo soltanto alcuni elementi di statistica descrittiva, il cui compito è organizzare in modo efficace i dati raccolti sull'intera popolazione in esame. Più precisamente, ci concentreremo sulle rappresentazioni grafiche dei dati e su alcuni parametri con i quali si riassumono i dati rilevati, ossia le *medie* e gli *indici di dispersione*.

7.1 Rappresentazione dei dati

Prima di procedere all'analisi dei dati ricordiamo brevemente i vari tipi di scale di misura comunemente utilizzati.

Una *scala nominale* è usata per classificare le unità statistiche in termini di uguaglianza di certi loro attributi o proprietà fissati. Ad esempio, è nominale la scala usata nella sistematica di Linneo[1]. O, ancora, fra gli individui di una popolazione esposta ad una malattia epidemica si usa la scala nominale: S - individui sani non infettati, I - individui infetti, R - individui non infettabili.

Una *scala ordinale* è usata quando gli oggetti possono essere classificati secondo un ordine rispetto ad una data proprietà. I numeri assegnati alle classi ordinali seguono l'ordine naturale $0, 1, 2, 3, 4$, etc. Esempi tipici sono la scala di Mohs di durezza dei minerali (da 1 a 10), la scala Mercalli di intensità dei terremoti (da 1 a 11), la scala Beaufort dell'intensità del

[1]Nome latinizzato del botanico svedese Carl von Linné (1707-1778), che descrisse tutte le specie viventi all'epoca conosciute assegnando a ciascuna di esse un doppio nome (nomenclatura binomia).

vento (da 0 a 12), la scala di Welzenbach delle difficoltà alpinistiche (da 1 a 6). Si noti che in una scala ordinale l'ampiezza degli intervalli fra i vari valori non ha significato.

Quando invece le ampiezze degli intervalli diventano significative, si parla di *scala intervallare*. Un tipico esempio di scala intervallare è la misura del tempo in secondi o in altre unità. Un altro esempio è la misura empirica della temperatura, dove viene misurata la differenza di temperatura in base alla variazione di volume nel tubo del termometro. Non vi è nelle scale intervallari un punto *zero* di riferimento, e non vi sono quindi misurazioni *assolute*, ma solo misure di differenze.
Nelle scale intervallari non ha senso considerare i rapporti: se prendiamo due corpi, uno a 80 ^oC (gradi Celsius) e l'altro a 40 ^oC, otteniamo $80/40 = 2$, ma in gradi Fahrenheit[2], i due corpi hanno rispettive temperature di 176 ^oF e 104 ^oF, con rapporto $176/104 \approx 1,69 \neq 2$.

Viene usata la *scala a rapporti* quando è possibile parlare di rapporti fra diverse quantità della grandezza esaminata. Una scala a rapporti deve necessariamente avere uno zero assoluto di riferimento. La lunghezza e la massa sono tipici esempi di grandezze misurate con una scala a rapporti. La misura della temperatura in gradi Kelvin è un altro esempio: esiste infatti in tale scala, per definizione, lo zero assoluto corrispondente a $-273,14$ gradi Celsius.

Iniziamo ora l'analisi dei dati. Come descritto nel paragrafo introduttivo gli elementi della popolazione in esame costituiscono le *unità statistiche* di osservazione. Indicheremo con N il numero di unità statistiche che costituiscono la popolazione e denoteremo le unità statistiche con: $U_1, U_2, \ldots, U_N$.

Definizione 7.1. Chiamiamo *variabile* o *carattere* ciò che si misura o osserva sulle unità statistiche di una popolazione.

[2]La formula di trasformazione da gradi Celsius a gradi Fahrenheit è

$$^oF = 32 + \frac{9}{5}{}^oC \ .$$

Siano

$$X = \{x_1, x_2, \ldots, x_N\}$$

i valori di una variabile X misurati su una popolazione, dove si è indicato con x_k il valore della variabile X relativo all'unità statistica U_k.
Il numero che rappresenta quante unità statistiche presentano una certa modalità del carattere X prende il nome di *frequenza assoluta* di quella modalità. L'insieme delle coppie ordinate [modalità, frequenza assoluta] si chiama *distribuzione di frequenze*.
Vediamo ora, con alcuni esempi guida, come procedere per la rappresentazione grafica dei dati a seconda della natura del carattere.

♢ **Esempio 7.1.** Simulazione del lancio di due dadi non truccati a sei facce (250 lanci)

dadi = {6, 6, 5, 6, 3, 4, 8, 7, 7, 6, 9, 5, 10, 6, 6, 7, 10, 10, 3, 3, 8, 5, 7, 6, 10, 7, 6, 7, 9, 4, 7, 2, 5, 11, 6, 6, 8, 6, 4, 7, 7, 9, 7, 7, 8, 10, 9, 5, 8, 6, 6, 7, 5, 5, 11, 4, 10, 7, 9, 9, 7, 4, 9, 5, 10, 8, 5, 6, 9, 7, 6, 4, 7, 7, 6, 3, 2, 8, 9, 4, 8, 11, 2, 8, 9, 7, 11, 6, 9, 4, 8, 7, 6, 3, 6, 7, 4, 2, 6, 3, 4, 6, 3, 5, 4, 10, 6, 9, 9, 9, 3, 7, 6, 9, 9, 4, 6, 7, 7, 5, 11, 8, 10, 3, 10, 8, 8, 4, 4, 5, 9, 7, 5, 11, 8, 9, 11, 3, 9, 6, 7, 8, 5, 2, 8, 4, 6, 7, 9, 5, 6, 6, 5, 5, 8, 10, 12, 7, 8, 6, 6, 8, 6, 3, 8, 5, 7, 3, 2, 8, 8, 9, 9, 8, 4, 5, 8, 7, 8, 5, 7, 10, 7, 8, 5, 7, 2, 10, 7, 3, 5, 5, 6, 8, 11, 7, 8, 7, 6, 11, 12, 8, 7, 5, 9, 4, 10, 4, 10, 4, 5, 7, 6, 12, 6, 9, 4, 6, 7, 3, 11, 12, 6, 5, 7, 6, 2, 4, 11, 5, 4, 9, 7, 10, 8, 7, 8, 2, 7, 11, 3, 7, 12, 11, 6, 8, 5, 8, 10, 8} .

In questo caso il carattere (numero ottenuto ad ogni lancio) presenta un numero discreto di modalità: i numeri interi compresi tra 2 e 12. Si può quindi calcolare la frequenza assoluta di ciascuna modalità.
Così facendo si ottiene la Tabella 7.1 dalla quale si evince che il 7 è il numero più frequente.

Tabella 7.1 – Frequenza dei numeri da 2 a 12 nell'Esempio 7.1.

numero uscito	2	3	4	5	6	7	8	9	10	11	12
frequenza	9	15	22	28	39	43	34	25	17	13	5

Le frequenze si possono anche rappresentare con l'ausilio degli *istogrammi* come mostrato nella Figura 7.1.

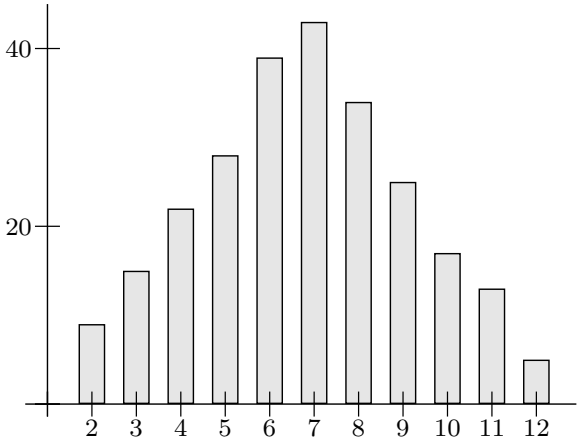


Figura 7.1 – Istogramma delle frequenze assolute dell'Esempio 7.1 calcolate nella Tabella 7.1. Le altezze dei rettangoli sono pari al numero di volte che è uscito il corrispondente numero alla base.

♢ **Esempio 7.2.** Il peso in grammi di 300 spigole:

peso spigole = {217, 250, 297, 212, 380, 344, 259, 269, 303, 327, 285, 341, 326, 233, 217, 379, 284, 307, 377, 369, 382, 253, 311, 342, 309, 409, 287, 341, 259, 392, 250, 296, 336, 239, 301, 235, 368, 264, 288, 269, 255, 254, 391, 311, 363, 251, 294, 287, 287, 328, 227, 158, 303, 371, 312, 306, 341, 347, 314, 342, 283, 345, 347, 250, 328, 213, 284, 269, 240, 193, 260, 282, 344, 316, 405, 269, 355, 356, 253, 299, 395, 293, 283, 394, 291, 296, 277, 353, 287, 314, 322, 274, 340, 394, 236, 448, 258, 269, 358, 323, 268, 327, 338, 332, 334, 344, 292, 337, 373, 244, 334, 276, 296, 297, 227, 259, 244, 193, 301, 274, 286, 378, 288, 267, 369, 215, 232, 350, 333, 240, 349, 320, 277, 311, 296, 360, 316, 265, 249, 270, 222, 380, 249, 291, 320, 249, 273, 251, 239, 254, 325, 345, 244, 334, 315, 245, 345, 323, 241, 307, 314, 363, 256, 339, 304, 320, 409, 265, 301, 271, 333, 287, 367, 220, 268, 239, 276, 282, 288, 285, 317, 304, 313, 251, 363, 330, 271, 247, 279, 351, 340, 278, 332, 316, 291, 276, 225, 330, 317, 254, 244, 179, 263, 334, 285, 359, 343, 275, 269, 256, 244, 302, 364, 290, 303, 320, 247, 348, 290, 318, 257, 221, 418, 218, 395, 325, 332, 348, 283, 339, 243, 351, 305, 234, 300, 399, 320, 310, 309, 320, 322, 331, 258, 384, 329, 277, 339, 271, 308, 270, 255, 303, 269, 315, 304, 337, 334, 267, 355, 356, 242, 239, 319, 323, 305, 323, 346, 357, 316, 250, 293, 228, 270, 374, 278, 375, 299, 364, 258, 357, 238, 300, 298, 321, 202, 368, 371, 422, 212, 349, 306, 344, 303, 328, 339, 363, 264, 305, 295, 256} .

In questo caso la frequenza di ciascun valore non è indicativa in quanto, essendo la variabile peso di tipo continuo, quasi ogni valore potrebbe avere frequenza 1.

Quando la variabile è di tipo continuo per sintetizzare i dati è più indicativa la frequenza che i valori $X = \{x_1, \ldots, x_N\}$ hanno in un dato intervallo (classe) piuttosto che la frequenza di ogni modalità. Per far questo si suddivide l'*ampiezza* $I = x_{\max} - x_{\min}$, dove $x_{\max}$ e $x_{\min}$ rappresentano rispettivamente il massimo e il minimo degli $\{x_1, \ldots, x_N\}$, in sotto intervalli $I_1, \ldots, I_k$ di ampiezza ℓ e si definiscono le *frequenze assolute* relative alla classe I_k, indicate con n_k, come il numero di unità statistiche con un valore della variabile X nella classe I_k. Le frequenze assolute sono generalmente rapportate al numero totale di unità statistiche della popolazione, definendo in tal modo le *frequenze relative* $f_k = n_k/N$ che, comunemente, vengono espresse in percentuale.

Nel caso del'Esempio 7.2, suddividendo l'ampiezza $I = 448 - 158 = 290$ in sotto intervalli di ampiezza 50, si trova la Tabella 7.2.

Tabella 7.2 – Frequenze relative e assolute dell'Esempio 7.2. Si noti l'uso delle parentesi: il primo estremo è incluso mentre il secondo è escluso.

Classi	n_k	f_k
$I_1 = [150, 200)$	4	1.33 %
$I_2 = [200, 250)$	41	13.66 %
$I_3 = [250, 300)$	98	32.66 %
$I_4 = [300, 350)$	108	36 %
$I_5 = [350, 400)$	43	14.33 %
$I_6 = [400, 450)$	6	2 %

Le frequenze assolute o quelle relative si possono rappresentare tramite gli *istogrammi* come mostrato in Figura 7.2.

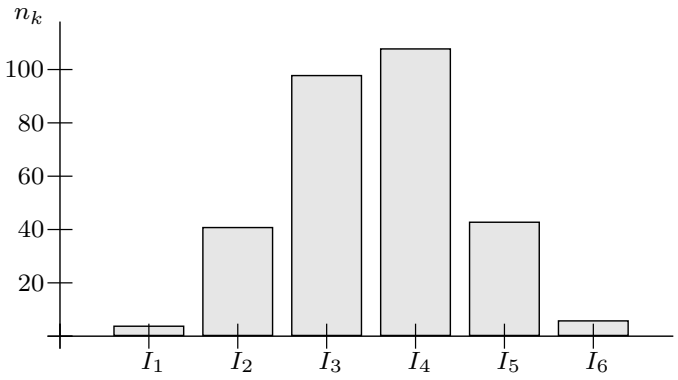


Figura 7.2 – Istogramma delle frequenze assolute dell'Esempio 7.2, calcolate nella Tabella 7.2. Le altezze dei rettangoli sono pari alle frequenze assolute dei corrispondenti intervalli.

♢ **Esempio 7.3.** I valori della pressione sistolica e diastolica del sangue di una persona adulta sono classificati nel modo seguente:

classificazione	**sistolica**	**diastolica**
ottimale	< 120	< 80
normale	120-129	80-84
normale alta	130-139	85-89
alta lieve	140-159	90-99
alta moderata	160-179	100-109
alta grave	> 180	> 110

In uno studio medico viene misurata la pressione sistolica in 100 pazienti ipertesi ottenendo i seguenti valori (ordinati in ordine crescente):

$$\mathbf{P} = \{110, 110, 110, 110, 111, 112, 113, 114, 115, 115, 115, 115, 117, 118, 119, 119, 119, 119, 120, 120, 121, 121, 121, 121, 121, 122, 122, 122, 124, 124, 125, 125, 125, 125, 126, 126, 126, 128, 128, 129, 129, 130, 131, 132, 134, 134, 135, 135, 136, 136, 136, 137, 137, 137, 137, 137, 137, 137, 138, 138, 138, 138, 138, 139, 139, 139, 139, 140, 140, 140, 143, 144, 145, 146, 146, 149, 153, 153, 153, 156, 157, 157, 158, 162, 163, 163, 165, 165, 166, 168, 170, 170, 173, 175, 178, 185, 185, 186, 186, 187\} .$$

Se si vuole procedere al calcolo delle frequenze assolute e relative risulta naturale utilizzare la suddivisione dell'ampiezza di P nelle classi indicate nella classificazione della pressione sistolica. Un'analisi attenta mostra che questa suddivisione non è omogenea: le classi non hanno tutte la stessa ampiezza. Per procedere alla rappresentazione dell'istogramma può essere utile, in modo da tener conto delle diverse ampiezze, il seguente metodo. Si costruiscono dei rettangoli, relativi ad ogni classe, di base pari alla lunghezza della relativa classe e area pari alla frequenza relativa. Dalla formula dell'area di un rettangolo segue che l'altezza dei rettangoli è:

$$h_k = \frac{f_k}{\text{lunghezza di } I_k} \ .$$

Applicando tale metodo si ottengono i valori riportati nella Tabella 7.3. In Figura 7.3 sono mostrati i due tipi di istogrammi relativi alla Tabella 7.3: quello non omogeneo, dove l'area dei rettangoli è pari alla frequenza relativa; quello omogeneo dove l'altezza dei rettangoli è pari alla frequenza relativa.

♢ **Esempio 7.4.** Nel caso di una variabile qualitativa si utilizza molto spesso, per la rappresentazione delle frequenze relative o assolute, il *diagramma circolare* (comunemente denominato *grafico a torta* o *diagramma*

Tabella 7.3 – Frequenze relative e altezze h_k dell'Esempio 7.3.

Intervalli	f_k	h_k
< 120	18 %	$18/10 = 1.8$
$[120, 130)$	23 %	$23/10 = 2.3$
$[130, 140)$	26 %	$26/10 = 2.6$
$[140, 160)$	16 %	$16/20 = 0.8$
$[160, 180)$	12 %	$12/20 = 0.6$
> 180	5 %	$5/7 = 0.7$

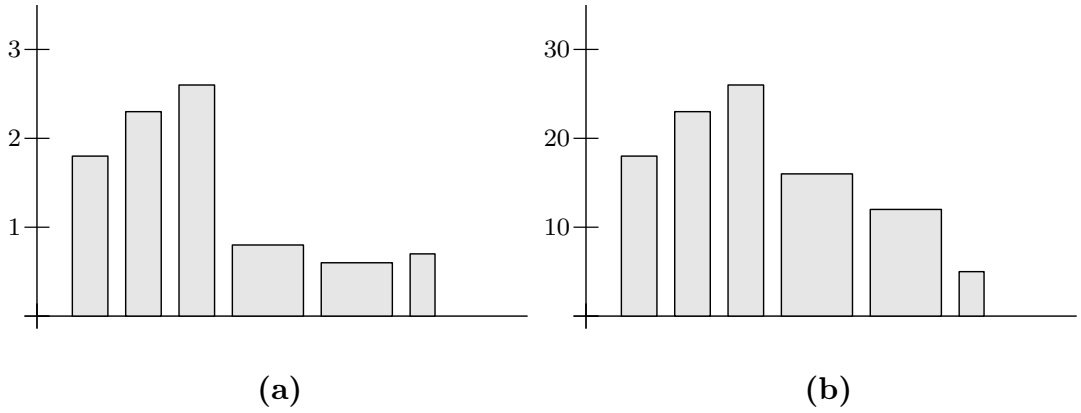


Figura 7.3 – (a) Istogramma a barre non omogeneo dell'Esempio 7.3 dove le aree dei rettangoli sono pari alle frequenze relative. (b) Diagramma a barre dell'Esempio 7.3 dove le altezze dei rettangoli sono pari alle frequenze relative.

a torta), al fine di evitare di stabilire, anche involontariamente, un ordine che non esiste nella variabile (cosa che accadrebbe utilizzando un istogramma).

Un diagramma circolare viene costruito dividendo un cerchio in spicchi le cui ampiezze angolari sono proporzionali alle classi di frequenza. Come per l'istogramma, le aree sono proporzionali alle frequenze.

Per esempio, in Figura 7.4, è mostrato il diagramma circolare della distribuzione delle fonti utilizzate nel mondo per la produzione di energia primaria.

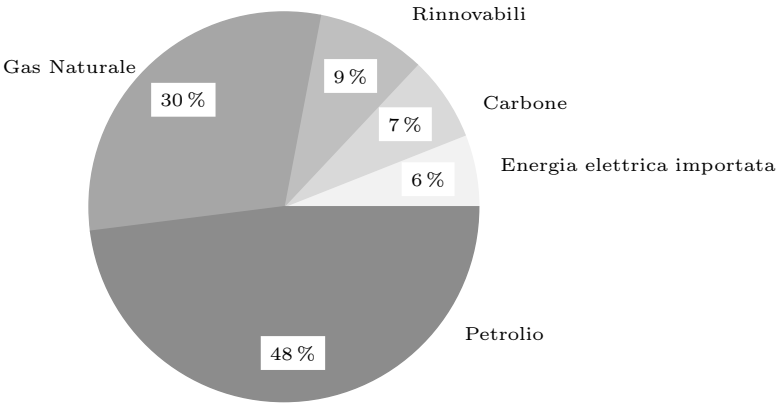


Figura 7.4 – Fonte dati: ENEA *Rapporto Energia e Ambiente 2003. Le fonti rinnovabili*. Roma, 2003. pp. 15-16.

7.2 Indicatori di centralità (medie)

Gli *indicatori di centralità* indicano un *centro* dei dati e si utilizzano per rappresentare i valori $X = \{x_1, x_2, \ldots, x_N\}$ tramite un unico numero che, in qualche modo, dovrebbe rappresentare la totalità dei dati.
In sostanza, un indicatore di centralità è un valore opportunamente scelto e compreso fra il minimo e il massimo dei dati. In tutti i casi, è un numero che ne sintetizza molti, e consente di averne una visione unitaria, ovviamente nascondendo la molteplicità dei dati da cui è ottenuto. Così, il reddito medio delle famiglie italiane è un valore unico, utile per fare confronti con altre nazioni o con periodi passati, ma non evidenzia che i redditi sono molto diversi e molte famiglie sono al di sotto della soglia della sopravvivenza, mentre altre possiedono beni in grande quantità; la statura media ci consente di dire che gli svedesi sono, in media, più alti degli italiani, ma non evidenzia che molti italiani sono più alti di parecchi svedesi.
Prenderemo in esame i seguenti indici di centralità: moda, mediana, media aritmetica, media quadratica, media geometrica e media armonica.

Definizione 7.2. Si dice *moda* la modalità a cui corrisponde la massima frequenza.

♢ **Esempio 7.5.** La sequenza di numeri $X = \{5, 6, 8, 8, 8, 12, 12, 14\}$ ha moda 8. La sequenza di numeri $X = \{5, 6, 8, 8, 8, 12, 14, 14, 14\}$ ha due mode: 8 e 14. Nella sequenza di numeri $X = \{1, 2, 3, 4, 5, 6\}$ si potrebbe

anche dire, a stretto rigore, che vi sono sei mode; ma è più ragionevole concludere che in questo caso la moda non esiste.

▷ **Esercizio 7.1.** In un condominio ci sono venti appartamenti suddivisi in: otto quadrivani, sei trivani, due monolocali, tre bivani e un pentavano. Quali sono gli appartamenti più di moda nel condominio?

Soluzione. In questo caso la variabile che misura il numero di stanze di un appartamento è $X = \{1, 1, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5\}$ da cui si evince che la moda è 4. ◁

Definizione 7.3. Si dice *mediana* di una serie di dati $X = \{x_1, \ldots, x_N\}$, e si denota con $X_{0.5}$, un numero tale che almeno il 50 % dei valori di X sono minori o uguali di $X_{0.5}$ e almeno il 50 % dei valori di X sono maggiori o uguali di $X_{0.5}$.

Definizione 7.4. Si dice *media aritmetica* (o semplicemente *media*) di una serie di dati $X = \{x_1, \ldots, x_N\}$ il numero

$$\overline{X} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{1}{N} \sum_{j=1}^{N} x_j \ . \tag{7.2.1}$$

▷ **Esercizio 7.2.** Calcolare media e mediana delle due serie di dati:

$$X = \{1, 2, 4, 0, 2, 4, 1, 3, 6, 2, 1\}$$
$$Y = \{20, 2, 4, 0, 2, 4, 0, 3, 6, 2\} \ .$$

Soluzione. Un calcolo diretto mostra che $\overline{X} = 2.364$ e $\overline{Y} = 4.3$. Per la mediana si procede nel modo seguente. Riordinando i valori di X e Y si trova

$$X^* = \{0, 1, 1, 1, 2, 2, 2, 3, 4, 4, 6\}$$
$$Y^* = \{0, 0, 2, 2, 2, 3, 4, 4, 6, 20\} \ .$$

Essendo X^* composto da un numero dispari di elementi esiste un numero centrale che rappresenta la mediana, quindi $X_{05} = 2$. Nel caso di Y^*, essendo il numero dei dati pari, si trovano i due valori centrali 2 e 3 e si sceglie, per convenzione, come mediana la media aritmetica di questi due, quindi $Y_{05} = 2.5$. Si osservi che media e mediana di X non si discostano di molto, mentre per Y la differenza è evidente. Tale fenomeno è dovuto alla disomogeneità dei valori in Y rispetto a quelli in X.

♢ **Esempio 7.6** (Il voto di laurea)**.** Il voto di laurea è quasi sempre calcolato sulla base della *media pesata* dei voti riportati nei singoli esami. La media pesata è una media aritmetica che tiene conto del peso di ciascun esame, dove il peso è calcolato in CFU (credito formativo universitario). Indicato con $(v_k, p_k), k = 1, \dots N$, il voto e i corrispondenti CFU di N esami, si definisce media pesata il numero

$$\overline{X}_p = \frac{\sum_{j=1}^{N} v_j p_j}{\sum_{j=1}^{N} p_j} \ .$$

A titolo di esempio, la media pesata dei seguenti voti

$$X = \{(26, 6), (30, 8), (18, 4), (29, 6)\}$$

è $\overline{X}_p = 26.75$, mentre la media dei voti, senza contare i CFU, è $\overline{X} = 25.75$. In questo caso la mediana vale $X_{05} = 27.5$.

La media aritmetica è di gran lunga la più nota e usata delle medie. Il suo uso acritico e indiscriminato va però evitato: non è vero che, se io ho due polli e tu nessuno, è come se avessimo un pollo a testa; che per due amiche sia indifferente andare a passeggio con due ragazzi alti 170 cm, o con uno alto 140 cm e l'altro alto 200 cm; e così via. È invece indifferente se su un ascensore, di portata massima 240 Kg, salgono tre persone il cui peso è 60 Kg, 70 Kg e 110 Kg rispettivamente, o tre persone tutte del peso di 80 Kg. In generale, ogniqualvolta ha senso sommare i dati, l'uso della media aritmetica è appropriato. In tal caso essa esprime quale sarebbe l'intensità costante del carattere in esame, se fosse ripartita in parti uguali.

Definizione 7.5. Si dice *media quadratica* di una serie di dati

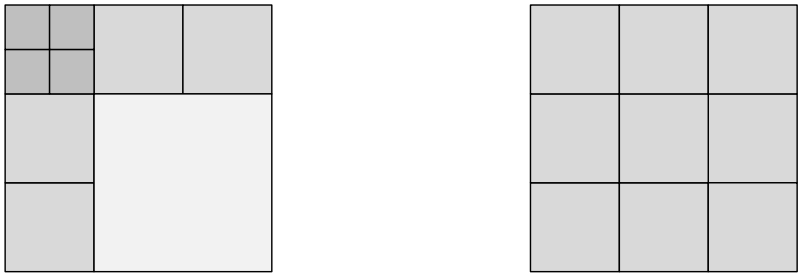
$$X = \{x_1, \dots, x_N\} \ ,$$

il numero

$$\overline{X}_q = \sqrt{\frac{x_1^2 + x_2^2 + \ldots + x_N^2}{N}} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} x_j^2} \ . \tag{7.2.2}$$

♢ **Esempio 7.7.** La media quadratica della serie di numeri

$$X = \{1, 1, 1, 1, 2, 2, 2, 2, 4\}$$



(a) **(b)**

Figura 7.5 – Significato geometrico della media quadratica.

è $\overline{X}_q = 2$. Geometricamente ciò si può interpretare dicendo che quattro quadrati di lato 1, quattro quadrati di lato 2 e un quadrato di lato 4 equivalgono a nove quadrati di lato 2 come mostra la Figura 7.5.

▷ **Esercizio 7.3.** Si vogliono sostituire tre tubi di raggio rispettivamente 2 cm, 3 cm e 4 cm con tre tubi di uguale raggio in modo che la portata complessiva resti inalterata. Quale deve essere il loro raggio?

Soluzione. La portata di un tubo dipende dall'area della sezione dello stesso. Detta x la misura in cm del raggio incognito, deve essere,

$$3\pi\, x^2 = \pi\, 2^2 + \pi\, 3^2 + \pi\, 4^2 \ ,$$

quindi

$$x = \sqrt{\frac{2^2 + 3^2 + 4^2}{3}} = 3.11 \ .$$

Il raggio richiesto è la media quadratica dei raggi dei tre tubi dati. ◁

Definizione 7.6. Si dice *media geometrica* di una serie di dati positivi

$$X = \{x_1, \ldots, x_N\} \ ,$$

il numero

$$\overline{X}_g = \sqrt[N]{x_1 \cdot x_2 \cdots x_N} = (x_1 \cdot x_2 \cdots x_N)^{\frac{1}{N}} \ . \qquad (7.2.3)$$

Evidentemente l'uso della media geometrica è appropriato quando il carattere in esame è moltiplicativo, cioè quando ha significato moltiplicare i dati.

Osservazione 7.1. Dalle proprietà dei logaritmi segue che

$$\log \overline{X}_g = \log \left((x_1 \cdot x_2 \cdots x_N)^{\frac{1}{N}} \right) = \frac{1}{N} \sum_{j=1}^{N} \log x_j \ ,$$

cioè il logaritmo della media geometrica corrisponde alla media aritmetica del logaritmo dei dati.

▷ **Esercizio 7.4.** Una ditta fattura 100 nel 2007, 112 nel 2008, 140 nel 2009 e 168 nel 2010. Qual è il tasso di incremento medio del fatturato?

Soluzione. Gli incrementi nei tre periodi sono

$$\frac{112}{100} = 1,12 \qquad \frac{140}{112} = 1,25 \qquad \frac{168}{140} = 1,2 \ .$$

Sia x il tasso di incremento medio, allora si deve avere $100 \cdot 1,12 \cdot 1,25 \cdot 1,2 = 100 \cdot x \cdot x \cdot x = 100 \cdot x^3$, per cui

$$x = \sqrt[3]{1,12 \cdot 1,25 \cdot 1,2} = 1,18878$$

rappresenta la media geometrica degli incrementi nei tre periodi. ◁

▷ **Esercizio 7.5.** Un trasformatore rende l'81 %, un altro il 64 %. Se si applicano in serie il rendimento complessivo è pari al prodotto dei due rendimenti. Se volessi usare due trasformatori uguali ed avere lo stesso rendimento dei due sopra, quanto dovrebbe essere il rendimento dei nuovi trasformatori?

Soluzione. Sia x il rendimento incognito, allora si deve avere

$$x \cdot x = x^2 = 0,81 \cdot 0,64 \ .$$

Segue che

$$x = \sqrt{0,81 \cdot 0,64} = 0,72 = 72\,\%$$

rappresenta proprio la media geometrica dei due valori 0,81 e 0,64. ◁

▷ **Esercizio 7.6.** Se il cambio tra una valuta A ed una B è 16/1 nel 2010 e diventa 25/1 nel 2011, qual è il cambio medio tra le due valute?

Soluzione. Si potrebbe pensare che la media aritmetica $(16 + 25)/2 = 20,5$ fornisca la risposta corretta. Se però consideriamo il cambio tra la valuta B e la valuta A si trova che nel 2010 era 1/16 mentre nel 2011 è 1/25. La media di questi ultimi due valori

$$\frac{\frac{1}{16} + \frac{1}{25}}{2} = 0,05125$$

non è però pari al reciproco della media tra 16 e 20, infatti $1/20,5 = 0,04878$. Un buon metodo di calcolo non dovrebbe dipendere dall'ordine con cui si confrontano le due valute. Per risolvere il problema basta osservare che

$$\frac{1}{\sqrt[N]{x_1 \cdot x_2 \cdots x_N}} = \sqrt[N]{\frac{1}{x_1} \cdot \frac{1}{x_2} \cdots \frac{1}{x_N}} \, .$$

In parole, il reciproco della media geometrica è uguale alla media geometrica dei reciproci dei singoli valori. Quindi come cambio medio è corretto utilizzare la media geometrica fra i due, ottenendo $\sqrt{16 \cdot 25} = 20$. ◁

Definizione 7.7. Si dice *media armonica* di una serie di dati positivi

$$X = \{x_1, \ldots, x_N\} \, ,$$

il reciproco della media aritmetica dei loro reciproci. In formula:

$$\overline{X}_a = \frac{1}{\frac{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}}{N}} = \frac{N}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_N}} \, . \tag{7.2.4}$$

▷ **Esercizio 7.7.** Percorro 21 Km alla velocità di 30 Km/h e altri 21 Km alla velocità di 70 Km/h. Qual è la velocità media?

Soluzione. Risolviamo il problema in generale. Dette s la lunghezza comune dei due tratti e v_1 e v_2 le due velocità, il tempo t_1 impiegato nel primo tratto è $t_1 = s/v_1$. Analogamente il tempo t_2 impiegato nel secondo tratto è $t_2 = s/v_2$ Il tempo complessivo è $t = t_1 + t_2 = s/v_1 + s/v_2$, per cui la velocità media v_m risulta:

$$v_m = \frac{2s}{t} = \frac{2s}{t_1 + t_2} = \frac{2s}{\frac{s}{v_1} + \frac{s}{v_2}} = \frac{2}{\frac{1}{v_1} + \frac{1}{v_2}}$$

che corrisponde alla media armonica delle due velocità. Nel caso esplicito del problema si ottiene la velocità media di

$$\frac{2}{\frac{1}{30}+\frac{1}{70}} = 42\,\mathrm{Km/h}\ .$$

$\triangleleft$

7.3 Indicatori di dispersione

Si considerino i seguenti dati:

$$X = \{0, 1, 10, 10, 19, 20\}$$
$$Y = \{0, 9, 10, 10, 11, 20\}\ .$$

Un calcolo diretto mostra che X ed Y hanno stessa moda, media, mediana ed ampiezza. Nonostante ciò è evidente, a colpo d'occhio, che i dati in X e Y non sono uguali. Si vuole quindi introdurre un nuovo indice che permetta di distinguere X da Y. Sempre dall'osservazione diretta risulta evidente che i valori in X sono *mediamente* più distanti dalla media (più dispersi) rispetto ai corrispondenti valori di Y.

Detto ciò si definisce lo *scarto medio* come la media delle distanze dei valori dalla media. In formula:

$$\frac{1}{N}\sum_{j=1}^{N} |\, x_j - \overline{X}\,|\ . \tag{7.3.1}$$

In statistica si preferisce utilizzare un altro coefficiente di dispersione considerando la media quadratica delle distanze dei valori dalla media. Più precisamente si ha la seguente:

Definizione 7.8. Si chiama *scarto quadratico medio* (o *deviazione standard*) di un insieme $X = \{x_1, \ldots, x_N\}$ di N osservazioni il numero

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N}(x_j - \overline{X})^2}{N}}\ . \tag{7.3.2}$$

La quantità σ^2 è detta *varianza* (denotata anche $\mathrm{Var}(X)$).

▷ **Esercizio 7.8.** Dimostrare la seguente *Formula di König* :

$$\mathrm{Var}(X) = \overline{(X^2)} \, - \, (\overline{X})^2 \ ,$$

dove, se $X = \{x_1, \ldots, x_N\}$, si definisce $X^2 = \{x_1^2, \ldots, x_N^2\}$.

Soluzione. Tenendo conto delle proprietà (4.1.5), si ha

$$\begin{aligned} \mathrm{Var(X)} &= \frac{1}{N}\sum_{j=1}^{N}(x_j \, - \, \overline{X})^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j^2 \, - \, 2x_j\,\overline{X} \, + \, \overline{X}^2) \\ &= \frac{1}{N}\sum_{j=1}^{N} x_j^2 \, - \, 2\overline{X}\,\frac{1}{N}\sum_{j=1}^{N} x_j + \frac{1}{N}\sum_{j=1}^{N}\overline{X}^2 \\ &= \overline{(X^2)} \, - \, 2\,\overline{X}\,\overline{X} \, + \, (\overline{X})^2 = \overline{(X^2)} \, - \, (\overline{X})^2 \ . \end{aligned}$$

◁

▷ **Esercizio 7.9** (*)**.** Dato un insieme $X = \{x_1, \ldots, x_N\}$ di N osservazioni, si consideri la funzione $f : \mathbb{R} \to \mathbb{R}$ definita da $f(x) = \sum_{j=1}^{N}(x_j \, - \, x)^2$. Dimostrare che la funzione f assume valore minimo quando $x = \overline{X}$.

Soluzione. Posto $d = \overline{X} - x$, da cui $x = \overline{X} - d$, si ha

$$\begin{aligned} \sum_{j=1}^{N}[x_j \, - \, x]^2 &= \sum_{j=1}^{N}[(x_j - \overline{X}) \, + \, d]^2 \\ &= \sum_{j=1}^{N}(x_j - \overline{X})^2 + 2d\sum_{j=1}^{N}(x_j - \overline{X}) + Nd^2 \\ &= \sum_{j=1}^{N}(x_j - \overline{X})^2 + Nd^2 \ , \end{aligned}$$

dove si è utilizzato che $\sum_{j=1}^{N}(x_j - \overline{X}) = \sum_{j=1}^{N} x_j - N\,\overline{X} = N\,\overline{X} - N\,\overline{X} = 0$. In conclusione

$$f(x) = f(\overline{X}) + Nd^2 \geq f(\overline{X}) \ , \qquad \forall\, x \in \mathbb{R}$$

e l'uguale vale se e solo se $d = 0$, cioè $x = \overline{X}$. ◁

Osservazione 7.2. In teoria, nella definizione di deviazione standard, si potrebbero considerare le distanze da uno degli altri indici di centralità piuttosto che dalla media aritmetica. L'Esercizio 7.9 mostra però che, detto M uno dei possibili indici di centralità, la quantità

$$\sqrt{\frac{\sum_{j=1}^{N}(x_j - M)^2}{N}}$$

assume valore minimo quando $M = \overline{X}$.

Ai fini descrittivi è molto utile un'altra misura di dispersione, cioè il cosiddetto *coefficiente di variazione*, definito come la deviazione standard espressa come percentuale della media aritmetica. In formula:

$$\mathrm{CV} = \frac{\sigma}{\overline{X}} \cdot 100 \ .$$

A differenza della deviazione standard, che è espressa nella stessa unità di misura della variabile originale, il CV è un numero puro, svincolato da ogni scala di misura, ed è quindi un indice diretto della variabilità.

▷ **Esercizio 7.10.** Si consideri il peso di ventisei maschi e di ventisei femmine

$$\begin{aligned}\mathrm{pesoM} = \mathrm{M} = (&75, 75, 65, 90, 90, 60, 60, 72, 73, 61, 73, 72, 68, 67, 56, 65,\\ &80, 67, 78, 64, 58, 56, 65, 61, 86, 63)\end{aligned}$$

$$\begin{aligned}\mathrm{pesoF} = \mathrm{F} = (&59, 59, 40, 52, 52, 45, 65, 53, 46, 42, 58, 55, 75, 74, 48, 52,\\ &48, 47, 58, 42, 48, 52, 53, 50, 50, 67) \ .\end{aligned}$$

È corretto affermare che il peso dei maschi varia di più rispetto a quello delle femmine?

Soluzione. Il calcolo della deviazione standard per le due serie di dati fornisce:

$$\sigma_M = 9,53\,, \qquad \sigma_F = 8,86 \ .$$

Sarebbe quindi naturale affermare che il peso dei maschi varia di più rispetto a quello delle femmine. Tuttavia bisogna tener conto che il peso medio di un maschio è maggiore di quello di una femmina e che quindi stesse variazioni assolute di peso incidono più su una femmina che su un maschio. Per un confronto equo

bisogna ricorrere al coefficiente di variazione che tiene conto del peso medio. Con un calcolo diretto si trova

$$\mathrm{CV(M)} = 13,7\,\%\,, \qquad \mathrm{CV(F)} = 16,5\,\%$$

che porta a concludere che è vera l'ipotesi opposta, cioè il peso delle femmine varia di più rispetto a quello dei maschi[3]. ◁

Osservazione 7.3. In molte circostanze si verifica che le frequenze di un dato carattere hanno una distribuzione normale, ossia si distribuiscono in modo simmetrico e decrescente rispetto a un valore (in corrispondenza della moda) al quale spetta la massima frequenza. L'andamento delle frequenze è allora rappresentato da una curva a campana, detta *distribuzione di Gauss* (o, anche, *distribuzione normale*), come mostra la Figura 7.6. Il lettore dovrebbe rivedere gli istogrammi mostrati nelle Figura 7.2 e Figura 7.1 e riconoscerne l'andamento a campana.

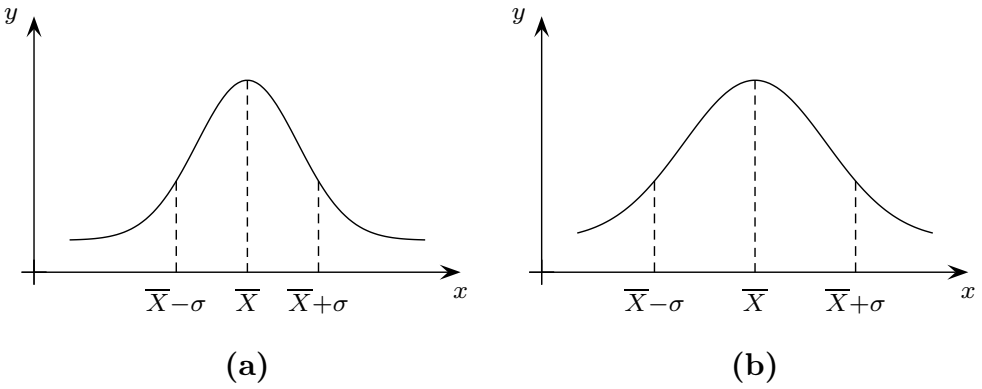


Figura 7.6 – La curva a campana della distribuzione di Gauss.

Ad esempio, hanno una distribuzione normale le stature, i pesi, le misure toraciche delle persone, i valori ottenuti con misurazioni ripetute di una stessa grandezza (se esse sono soggette solo ad errori accidentali), i valori dei pezzi lavorati dalle macchine (soggetti ad errori di lavorazione e di misurazione).
Nelle distribuzioni normali la media aritmetica, la moda e la mediana coincidono ed in corrispondenza di tale valore la curva raggiunge il suo valore massimo. Lo scarto quadratico medio determina la forma della curva di

[3]Tale risultato vale per le 52 persone delle quali sono riportati i pesi in questo esercizio e non si intende qui estenderlo a tutta la popolazione!

Gauss. Nella Figura 7.6 sono rappresentate due distribuzioni normali che hanno stesso valore medio, ma diversa ampiezza dovuta a differenti scarti quadratici medi. Per completezza, segnaliamo che l'espressione analitica che definisce la distribuzione normale è la seguente:

$$N(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\; e^{-\frac{(x-\overline{X})^2}{2\,\sigma^2}} \;.$$

In generale, quando un carattere ha distribuzione normale, si può dimostrare che:

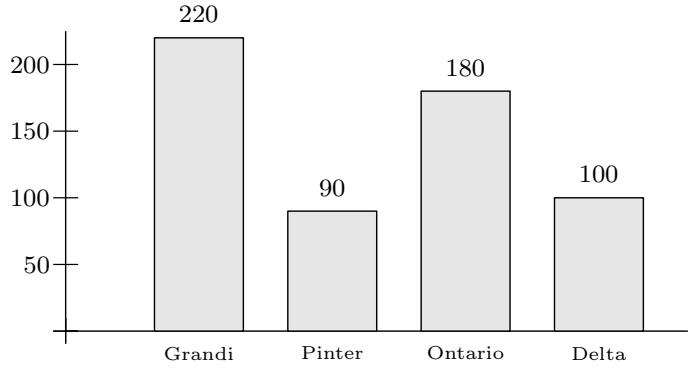
(a) il 68.27 % dei dati è compreso fra $\overline{X} - \sigma$ e $\overline{X} + \sigma$;

(b) il 95.45 % dei dati è compreso fra $\overline{X} - 2\sigma$ e $\overline{X} + 2\sigma$;

(c) il 99.73 % dei dati è compreso fra $\overline{X} - 3\sigma$ e $\overline{X} + 3\sigma$.

7.4 Esercizi di riepilogo

▷ **Esercizio 7.11.** In figura sono riportati i dati sulle vendite nell'anno 2008 di quattro ditte. Nel 2009 le vendite delle ditte Delta, Pinter e Grandi aumentano di 10 unità, mentre la ditta Ontario incrementa le sue vendite di 80 unità. Nel 2009 quale è la percentuale di vendite sul totale della ditta Ontario?



Soluzione. Nel 2009 le vendite sono

Grandi	230
Pinter	100
Ontario	260
Delta	110 ,

per un totale di 700. Segue che la percentuale di vendite della ditta Ontario sul totale è

$$\frac{260}{700} = 0.37\overline{142857} \approx 37\,\% \ .$$

◁

▷ **Esercizio 7.12.** Supponiamo che nel corso dell'anno il pane sia aumentato del 18 %, il prosciutto del 42 % e il burro del 30 %. Appare naturale dare un peso maggiore all'aumento del pane che non a quello del prosciutto o del burro. Se il costo del pane incide otto volte di più rispetto a quello del prosciutto e due volte di più rispetto a quello del burro, stabilire l'aumento percentuale medio del costo della vita.

Soluzione. L'aumento percentuale medio è dato dalla media pesata dei singoli aumenti. Pertanto, dando al prosciutto peso 1, si trova

$$\frac{18 \cdot 8 + 30 \cdot 4 + 42 \cdot 1}{8 + 4 + 1} \approx 23.5\,\% \ .$$

◁

▷ **Esercizio 7.13.** Un risparmiatore impiega, in ciascuno di due acquisti successivi, 2.100 euro per comperare monete d'oro la cui quotazione è una volta di 70 euro e l'altra volta di 30 euro. Qual è il prezzo medio di acquisto?

Soluzione. Il risparmiatore acquista la prima volta $2.100/70 = 30$ monete e la seconda volta $2.100/30 = 70$ monete. Complessivamente spende 4.200 euro per procurarsi 100 monete, ognuna delle quali gli è costata mediamente 42 euro. Tale prezzo è la media armonica dei due prezzi d'acquisto:

$$\frac{2}{\dfrac{1}{70} + \dfrac{1}{30}} = 42 \ .$$

◁

▷ **Esercizio 7.14** (*)**.** Data la serie di dati positivi

$$X = \{x_1, \ldots, x_N\} \ ,$$

dimostrare che

$$\overline{X}_a \leq \overline{X}_g \leq \overline{X} \leq \overline{X}_q \ .$$

Soluzione. Per semplicità espositiva dimostriamo le diseguaglianze solo per il caso in cui X è composto da soli due dati a e b. In questo caso bisogna dimostrare che

$$\frac{2}{\frac{1}{a}+\frac{1}{b}} \leq \sqrt{ab} \leq \frac{a+b}{2} \leq \sqrt{\frac{a^2+b^2}{2}} \ .$$

Da

$$0 \leq (a-b)^2 = (a+b)^2 - 4ab \ ,$$

si trova

$$\frac{4(ab)^2}{(a+b)^2} \leq ab \ .$$

Estraendo la radice quadrata si perviene alla prima disuguaglianza. Le altre disuguaglianze si riconducono sempre alla veridicità della disuguaglianza $0 \leq (a-b)^2$. Tale ragionamento porta anche alla conclusione che gli uguali valgono se e solo se $a = b$. Il lettore è invitato a provare che le disuguaglianze rimangono valide qualunque sia il numero di dati in considerazione. ◁

▷ **Esercizio 7.15.** Se tra mille persone si osserva un peso medio di 73 Kg con uno scarto quadratico medio di 5 Kg, quante persone (circa) hanno un peso tra 68 e 78 Kg?

Soluzione. Dall'Osservazione 7.3, avendo il carattere peso una distribuzione normale, segue che il $68,27\,\%$ delle persone ha un peso tra 73 ± 5 Kg, cioè tra 68 e 78 Kg. Possiamo quindi affermare che circa 683 persone hanno un peso tra 68 e 78 Kg. ◁

7.5 Esercizi proposti

▷ **Esercizio 7.16.** In un'azienda gli stipendi annui, in migliaia di euro, sono così distribuiti:

2 direttori	50
4 capi ufficio	40
10 impiegati	20
30 operai	15 .

Calcolare la media aritmetica, la mediana e la moda degli stipendi.

▷ **Esercizio 7.17.** Un contadino possiede cinque campi di forma quadrata di lato 20 m, 30 m, 60 m, 80 m e 100 m rispettivamente. Gli si propone lo scambio con cinque campi quadrati uguali, dei quali si chiede di determinare il lato affinché lo scambio sia equo.

▷ **Esercizio 7.18.** Calcolare la media armonica della serie di numeri $X = \{2, 4, 5, 8, 10\}$.

▷ **Esercizio 7.19.** Un ciclista percorre due tappe di 200 Km ciascuna, la prima ad una velocità media di 40 Km/h, la seconda ad una velocità media di 20 Km/h. Determinare la velocità media complessiva nelle due tappe.

▷ **Esercizio 7.20.** Per ciascuna delle due serie di dati, peso e altezza, si calcolino la media, la mediana e la deviazione standard.

$$\mathrm{P} = \{51, 44, 59, 48, 62, 40, 51, 46, 50, 57, 41, 60, 48, 49, 45, 45, 51, 47, 54, 46\}$$

$$\mathrm{A} = \{167, 163, 162, 160, 171, 155, 161, 163, 163, 170, 152, 165, 160, 165, 155, 156, 165, 173, 162, 158\} \ .$$

▷ **Esercizio 7.21.** I dati seguenti forniscono il numero di cuccioli partoriti in un anno da un gruppo di venticinque gatte adulte:

$$\{1, 5, 3, 1, 3, 2, 2, 1, 2, 5, 3, 0, 1, 4, 3, 7, 1, 3, 1, 7, 2, 1, 2, 4, 8\} \ .$$

(i) Costruire la tabella con le frequenze assolute e relative di queste osservazioni.

(ii) Disegnare l'istogramma delle frequenze relative.

(iii) Dire se la distribuzione è unimodale o plurimodale e determinare la moda.

(iv) Calcolare media, mediana, varianza e deviazione standard.

▷ **Esercizio 7.22.** Quattro amici hanno sostenuto finora tre esami universitari, con i seguenti voti:

Mario	30	29	28
Giovanni	24	19	23
Francesca	22	26	27
Cinzia	18	25	26 .

Solo uno di essi, dopo aver superato un quarto esame, potrà avere esattamente la media aritmetica del 25: chi? (L'esame si ritiene superato se il voto è un numero intero compreso tra 18 e 30.)

7.6 Commenti e note bibliografiche

Come abbiamo visto la statistica descrittiva si occupa dell'analisi dei dati osservati prescindendo dal fatto che l'insieme dei dati sia un campione estratto da una popolazione più vasta o sia invece l'intera popolazione.

La branca della statistica che studia le probabili conclusioni che si possono trarre sulla popolazione complessiva, a partire dall'indagine su un campione, prende il nome di *statistica inferenziale*. Le conclusioni della statistica inferenziale non sono certezze, ma asserzioni formulate con i metodi del calcolo delle probabilità. Al fine di chiarire il compito dell'inferenza statistica consideriamo la seguente situazione. Sia data un'urna con dieci palline, di cui sei sono bianche e quattro rosse. Utilizzando il calcolo delle probabilità, possiamo dire che, se estraiamo dall'urna una pallina a caso, la probabilità che essa sia bianca è $0,6$. Si ha invece un problema di inferenza statistica se abbiamo un'urna con palline di cui non conosciamo la composizione, estraiamo n palline e ne osserviamo il colore e, a partire da queste osservazioni, vogliamo indurre la composizione dell'urna.

Per un approccio operativo alla statistica, una referenza bibliografica molto buona, e che non richiede conoscenze matematiche troppo avanzate, è [16]. Per ulteriori approfondimenti ed un approccio avanzato alla statistica si può consultare [1] .