## 31.1.4 Sintesi dei dati

**Sintesi grafica mediante istogrammi e diagrammi a gradini.** Quando si ha a che fare con una serie numerosa di dati, per stabilire quanto siano dispersi e con quale frequenza si presenti ciascuno di essi, si può costruire un istogramma oppure un diagramma a gradini.

Consideriamo per esempio il semplice caso dell'analisi di un aceto commerciale in cui un gruppo d'analisti ha ottenuto i dati riportati in tabella 31.1.

**Tabella 31.1** Determinazione dell'acido acetico in un aceto commerciale

| *n* | $CH_3COOH$ (g/100 mL) | *n* | $CH_3COOH$ (g/100 mL) | *n* | $CH_3COOH$ (g/100 mL) |
|---|---|---|---|---|---|
| 1 | 5,87 | 11 | 6,02 | 21 | 6,05 |
| 2 | 5,90 | 12 | 6,03 | 22 | 6,06 |
| 3 | 5,95 | 13 | 6,03 | 23 | 6,07 |
| 4 | 5,99 | 14 | 6,04 | 24 | 6,07 |
| 5 | 6,00 | 15 | 6,05 | 25 | 6,10 |
| 6 | 6,01 | 16 | 6,05 | 26 | 6,12 |
| 7 | 6,01 | 17 | 6,05 | 27 | 6,14 |
| 8 | 6,01 | 18 | 6,05 | 28 | 6,15 |
| 9 | 6,02 | 19 | 6,05 | 29 | 6,17 |
| 10 | 6,02 | 20 | 6,05 | 30 | 6,20 |
| | | | | Media | 6,04 |

Per stabilire il numero di intervalli (**classi**) in cui suddividere una serie di dati, si può usare, a titolo indicativo, la formula di **Dixon** e **Kronmal**, valida per $n > 100$:

$$r = 10 \cdot \log n \tag{31.2}$$

oppure la formula di **Valleman**, valida per $n < 100$:

$$r = 2\sqrt{n}$$

Nel nostro esempio, usando la seconda formula si ottiene:

$$r = 2\sqrt{30} = 11$$

Tuttavia, poiché il *range* (intervallo), ovvero la differenza tra il valore più alto e quello più basso delle misure è 0,33 (= 6,20 − 5,87) è più conveniente suddividere l'insieme dei valori solo in 9 classi (►tab. 31.2), invece di 11.
Diagrammando le frequenze relative in funzione delle classi si ottiene un istogramma (►fig. 31.1*a*); se invece si diagrammano le frequenze cumulate si ottiene un grafico a gradini (►fig. 31.1*b*).

**Tabella 31.2** **Suddivisione in classi dei valori di tabella 31.1**

| Classe | Intervallo | nr. di dati | Frequenza relativa* | Frequenza cumulata** |
|---|---|---|---|---|
| 1 | 5,85-5,88 | 1 | 0,03 | 0,03 |
| 2 | 5,89-5,92 | 1 | 0,03 | 0,06 |
| 3 | 5,93-5,96 | 1 | 0,03 | 0,09 |
| 4 | 5,97-6,00 | 2 | 0,07 | 0,16 |
| 5 | 6,01-6,04 | 9 | 0,30 | 0,46 |
| 6 | 6,05-6,08 | 10 | 0,33 | 0,79 |
| 7 | 6,09-6,12 | 2 | 0,07 | 0,86 |
| 8 | 6,13-6,16 | 2 | 0,07 | 0,93 |
| 9 | 6,17-6,20 | 2 | 0,07 | 1,00 |

* La frequenza relativa è pari a: $n_{classe}/n_{tot}$, dove $n_{classe}$ è il numero di dati in ciascuna classe.
** La frequenza cumulata è pari a: $\sum n_{classe}/n_{tot}$.

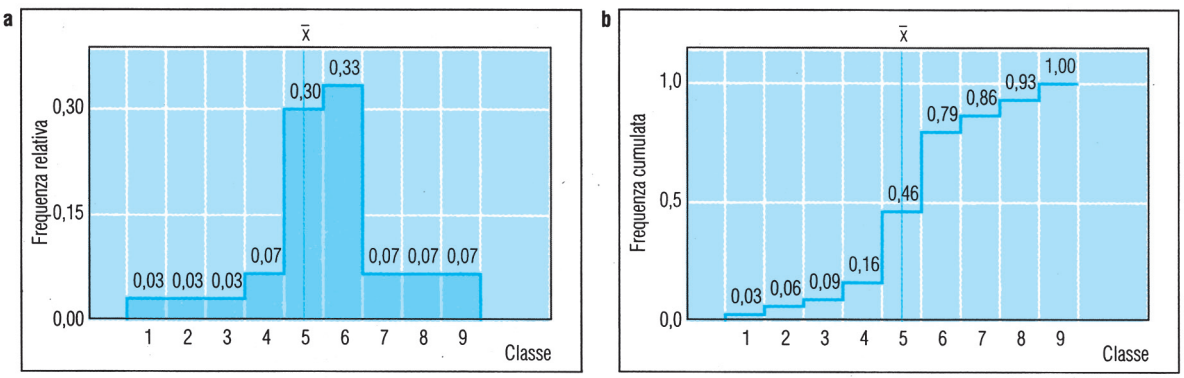


**Figura 31.1**
Determinazione dell'acido acetico (*v.* tabella 31.1); (**a**) istogramma delle frequenze relative; (**b**) diagramma a gradini delle frequenze cumulate. La media ($\bar{x}$) cade nella quinta classe.

**Sintesi grafica mediante box plot.** Un altro metodo per visualizzare le informazioni contenute in una serie di dati e confrontarle (con l'immediatezza del colpo d'occhio) con altre serie, è quello dei **box plot** (detti anche *box and whiskers plot*).

Riprendiamo l'esempio dell'analisi dell'aceto (*v.* tabella 31.2). Per applicare questo metodo, anzitutto si devono individuare: la mediana ($m$), il quarto superiore ($Q_s$), il quarto inferiore ($Q_i$) e i valori adiacenti, inferiore ($A_i$) e superiore ($A_s$).

La mediana corrisponde al valore centrale della serie, che occupa dunque la seguente posizione:

$$n_{med} = \frac{n+1}{2} = \frac{30+1}{2} = 15,5$$

Poiché i dati sono 30, la mediana è uguale alla media fra il quindicesimo e a sedicesimo dato e dal momento che $x_{15} = x_{16} = 6,05$ anche la mediana è uguale a questo valore:

$$m = 6,05$$

Il quarto inferiore e superiore corrispondono, rispettivamente, al valore centrale della prima e della seconda metà dei dati; la loro posizione è data dalle seguenti relazioni:

$$n_{Qi} = \frac{n_{med} + 1}{2}$$

$$n_{Qs} = n_{med} + \frac{n_{med} + 1}{2}$$

dove • $n_{med}$ indica la posizione della mediana, espressa mediante un valore intero non approssimato.

Nell'esempio, $n_{med} = 15$ e quindi le posizioni del quarto inferiore e del quarto superiore sono:

$$n_{Qi} = 8 \qquad n_{Qs} = 15 + 8 = 23$$

per cui:

$$Q_i = x_8 = 6{,}01 \qquad Q_s = x_{23} = 6{,}07$$

A questo punto, per determinare i valori adiacenti, prima si calcola la differenza fra i valori dei due quarti (detta **intervallo interquartile**, $\Delta_q$):

$$\Delta_q = 6{,}07 - 6{,}01 = 0{,}06$$

e poi si calcolano i valori:

$$A_i = Q_i - 1{,}5 \cdot \Delta_q \qquad \text{(nell'esempio: } 6{,}01 - 1{,}5 \cdot 0{,}06 = 5{,}92\text{)}$$

$$A_s = Q_s + 1{,}5 \cdot \Delta_q \qquad \text{(nell'esempio: } 6{,}07 + 1{,}5 \cdot 0{,}06 = 6{,}16\text{)}$$

Infine si traccia il grafico (►fig. 31.2*a*), ponendo in ordinata i seguenti valori della serie:

- il quarto inferiore e superiore ($Q_i$, $Q_s$), che rappresentano i limiti del rettangolo (box);
- la mediana ($m$), che interseca il rettangolo nella posizione corrispondente;

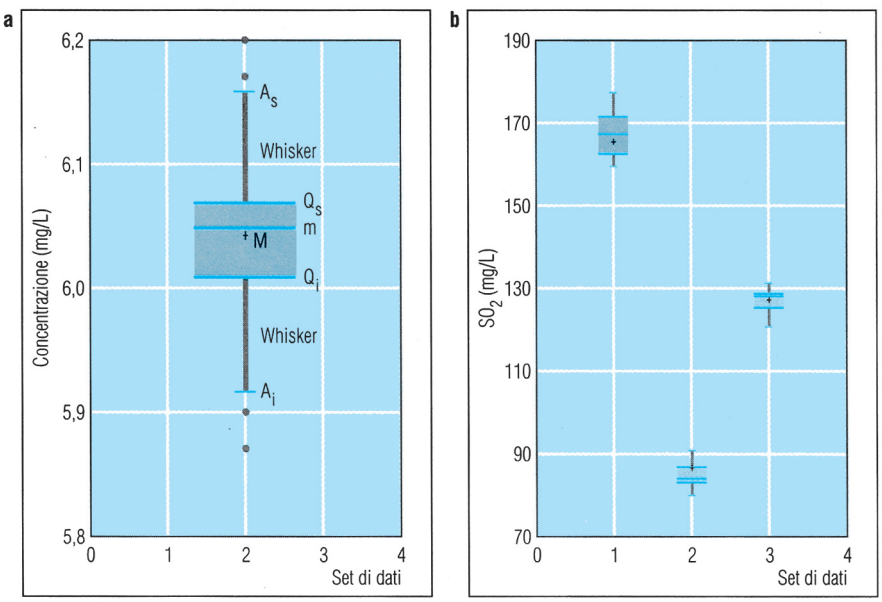


**Figura 31.2**
(**a**) Box plot relativo alla determinazione dell'acido acetico nell'aceto (*v.* tabella 31.1). I quattro valori rappresentati con un punto grigio sono da considerarsi aberranti (*M* indica il valore medio). (**b**) Box plot relativi alla determinazione di $SO_2$ in tre campioni diversi provenienti da una stessa partita di vino bianco. Come si può notare, il contenuto di $SO_2$ è decisamente diverso da campione a campione.

- i valori adiacenti ($A_i$, $A_s$), dai quali partono due segmenti (detti *whiskers*) che raggiungono il rettangolo.

I dati maggiori di $A_s$ o minori di $A_i$ vengono considerati aberranti e vengono marcati con un asterisco o comunque differenziati dagli altri.

**Diagrammi *steam* and *leaf*.** Si tratta di un metodo molto semplice, in parte numerico e in parte grafico, per rappresentare una serie di dati dividendoli in intervalli, in modo da costruire una sorta di istogramma di tipo numerico.

Consideriamo la seguente serie di dati, riportati in ordine crescente solo per comodità (anche se non è necessario, ai fini della stesura del diagramma):

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9**0** | 9**9** | 10**2** | 11**1** | 11**5** | 11**7** | 12**0** | 12**9** | 13**1** |
| 13**3** | 13**3** | 14**1** | 14**3** | 14**4** | 14**4** | 14**4** | 14**5** | 15**2** |
| 15**8** | 15**8** | 15**9** | 16**0** | 16**3** | 16**4** | 17**2** | 18**1** | 18**6** |
| 19**5** | | | | | | | | |

Per costruire il diagramma, si preparano due colonne. Nella prima colonna si pongono, in ordine crescente, tutte le cifre di ogni valore della serie, tranne l'ultima (che abbiamo evidenziato in neretto). Nella seconda colonna si pongono tutte le ultime cifre riscontrate nella serie, l'una di fianco all'altra, in corrispondenza delle cifre che le precedono.

Per esempio **0** e **9** (ultime cifre di 9**0** e 9**9**) vengono posti uno di seguito all'altro nella colonna a destra, mentre a sinistra c'è il numero 9 (cioè la cifra che li precede in 9**0** e 9**9**); **2** (colonna a destra) sta di fianco a 10 (colonna a sinistra); **1**, **5** e **7** stanno di fianco a 11; e così via. Si ottiene così un **diagramma *steam and leaf*** (dall'inglese «picciolo e foglia»), che dà un'idea della distribuzione dei valori in sottoclassi.

In questo esempio, come si può notare dal diagramma, la classe di dati più popolata è quella che va da 140 a 150, seguita dalla classe successiva.

| *Steam* | *Leaf* |
|---|---|
| 9 | 09 |
| 10 | 2 |
| 11 | 157 |
| 12 | 09 |
| 13 | 133 |
| 14 | 134445 |
| 15 | 2889 |
| 16 | 034 |
| 17 | 2 |
| 18 | 16 |
| 19 | 5 |