# 4.1 Caratteristiche e descrizione numerica di una distribuzione di dati

Cominciamo quindi a definire cosa è una statistica.

**Definizione.** *Statistica* è ogni funzione dei dati, che può avere anche valori vettoriali.

I dati quantitativi possono essere divisi in :

- **Dati discreti** che hanno valori in un insieme numerabile, come per esempio il numero di successi, il numero di arrivi, ....
- **Dati continui** che hanno valori in un intervallo di numeri, come per esempio tempo, la lunghezza ....

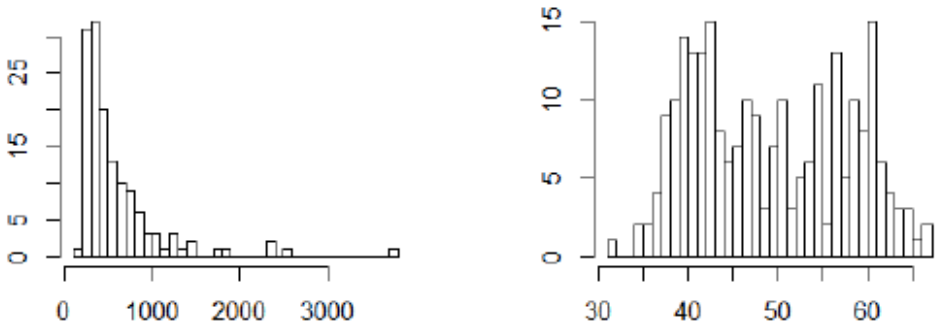
Supponiamo nel seguito di avere $N$ dati denominati $x_1, x_2, \ldots x_N$.

In questo paragrafo consideriamo, a titolo esemplificativo, i dati contenuti del file `discoveries` del pacchetto `datasets` di R, costituito da un vettore che memorizza il numero delle grandi invenzioni scientifiche anno per anno, dal 1860 al 1959.

## 4.1.1 Caratteristiche di un insieme di dati

1. **Centro.** È una delle caratteristiche più importanti dell'insieme dei dati. È associato ad un numero che rappresenta la tendenza centrale dei dati. Questo numero può essere calcolato in diversi modi, come vedremo in seguito.
2. **Diffusione.** Rappresenta l'estensione dei dati nel range dei valori. Un insieme con diffusione grande ha un range ampio di valori.
3. **Forma.** Riferita di solito alla forma di un grafico, tipo istogramma. Può servire per decidere come trattare i dati da un punto di vista statistico. Ci sono prevalentemente due aspetti della forma da considerare:
   - **Simmetria (skewness).** Una distribuzione si dice *simmetrica* quando i dati sono simmetrici rispetto al centro. Si dice **asimmetrica a destra (right skewed)** se la coda destra della distribuzione appare allungata rispetto al centro dei dati, **asimmetrica a sinistra (left skewed)** se lo è la coda di sinistra. In **Figura 4.1a** un esempio di distribuzione asimmetrica a destra (right skewed) e in **Figura 4.1b** un esempio di distribuzione simmetrica.
   - **Curtosi.** La curtosi indica la somiglianza rispetto alla distribuzione normale per quanto riguarda la forma del 'picco'. In particolare, la distribuzione si dice **platicurtica** se è più appiattita rispetto alla normale, **leptocurtica** se è più allungata rispetto alla normale. Si dice **mesocurtica** se non è né platicurtica, né mesocurtica, quindi non si discosta troppo dalla normale.
4. **Clusters e gaps.** I dati possono essere raggruppati intorno ad alcuni valori, formando quindi **ammassi (cluster)**, oppure ci possono essere intervalli in cui i dati sono quasi totalmente assenti, **vuoti (gap)**.

5. **Outliers.** Ci possono essere osservazioni estreme che provocano in statistica dei problemi, perché influenzano stime e parametri in modo eccessivo. Tali osservazioni sono chiamate *outliers*. Il bin all'estrema destra della **Figura 4.1a** mostra un esempio di outlier. Possono essere diverse le cause di un outlier e diverso il modo in cui vengono trattati nell'analisi di dati:
   - Non appartengono all'insieme dei dati che si vogliono analizzare, perché derivano per esempio da un errore di misura, oppure sono effettivamente dati dell'esperimento ma causati da un dispositivo malfunzionante, da un oggetto costruito in modo sbagliato ecc. In questo caso è opportuno eliminare gli outliers prima di procedere all'analisi dei dati.
   - Possono essere errori tipografici o di scrittura in questo caso di solito si individuano dopo un riesame o rilettura del file di dati e quindi l'errore viene corretto e il dato inserito nell'insieme da analizzare.
   - Possono indicare una tendenza non prevista del fenomeno da analizzare ed essere quindi spia di un comportamento nuovo, a volte anche molto interessante. in questo caso ovviamente l'outlier viene considerato nell'insieme di dati da analizzare e anzi il suo contributo diventa molto importante.




**(a)** Distribuzione asimmetrica a destra **(b)** Distribuzione simmetrica

**Figura 4.1:** Esempi di alcune distribuzioni di dati

## 4.1.2 Misure delle caratteristiche dei dati

### 4.1.2.1 Misure del centro

- **Media semplice.**

$$\bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \tag{4.1}$$

+: facile da calcolare
-: molto sensibile ai valori estremi.

```
> mean(discoveries)

## [1] 3.1
```

- **Mediana semplice.** Si devono ordinare i dati in maniera crescente. Se N è dispari, la mediana è il valore centrale, se N è pari è la media fra i due valori centrali.
  +: resistente ai valori estremi
  -: non ha le proprietà matematiche della media e necessita dell'ordinamento dei dati per essere calcolata.

```
> median(discoveries)

## [1] 3
```

- **Media troncata.**
  Nella media troncata (**trimmed mean**) si elimina una frazione $0 < t < 0.5$ delle osservazioni dagli estremi della lista ordinata e poi si calcola la media sui dati che rimangono.
  +: resistente ai valori estremi, buone proprietà statistiche
  -: richiede ordinamento dei valori

```
> mean(discoveries, trim=0.01)

## [1] 3.040816
```

#### 4.1.2.2 Statistica ordinata

Per statistica ordinata si intende una statistica fatta dopo avere ordinato i dati in ordine non decrescente: $x_1 \leq x_2 \leq x_N$. La statistica ordinata dà informazioni sulla distribuzione dei dati e su dove essi sono più o meno concentrati.

Lo strumento più utilizzato è il *quantile semplice*. Il quantile semplice di ordine $p$, $0 < p < 1$, indicato con $q_p$, è un valore che indica che circa il $100p\%$ dei dati è minore di $q_p$.

Innanzitutto i dati sono ordinati in modo non decrescente: $x_1 \leq x_2 \leq x_N$.

Non c'è una definizione univoca su come calcolare il quantile. Per esempio R li calcola in nove modi differenti (selezionabili con il parametri di input `type`). Uno dei modi utilizzati è il seguente, per trovare il quantile di ordine $p$, denotato $q_p$:

1. Calcolare $(n-1)p+1$ e scriverlo nella forma mista $k.d$.
2. $q_p = x_k + d(x_{k+1} - x_k)$

I quantili più utilizzati sono il $25\%$, $50\%$, e $75\%$.

```
> quantile(discoveries)

##   0%  25%  50%  75% 100%
##    0    2    3    4   12
```

#### 4.1.2.3 Misure della diffusione dei dati

Anche per avere informazioni sulla diffusione dei dati ci sono diverse misure che possono essere utilizzate. Ne presentiamo alcune.

- **Varianza e deviazione standard.**
  La *varianza campionaria* è definita come:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

  La *deviazione standard* è $s = \sqrt{s^2}$.
  +: buone proprietà statistiche
  -: molto sensibile ai valori estremi.

```
> var(discoveries)

## [1] 5.080808

> sd(discoveries)

## [1] 2.254065
```

- **Range interquartile (IQR).**
  Il *range interquartile* è definito come $IQR = q_{0.75} - q_{0.25}$.
  +: poco sendibile ai valori estremi
  -: necessario l'ordinamento dei dati; -: coinvolge solo il $50\%$ dei dati.

```
> IQR(discoveries)

## [1] 2
```

- **Deviazione assoluta dalla media (MAD).**
  Per calcolare la *Deviazione assoluta dalla media (MAD)* si deve calcolare innanzitutto la mediana $\bar{x}$, quindi le deviazioni assolute dalla mediana $|x_1 - \bar{x}|, |x_2 - \bar{x}|, \ldots, |x_n - \bar{x}|$. Quindi si calcola la mediana di questi valori, moltiplicata per un coefficiente $c$:

$$MAD = c \cdot median(|x_1 - \bar{x}|, |x_2 - \bar{x}|, \ldots, |x_n - \bar{x}|$$

+: molto robusto, anche più di IQR
-: poco conosciuto, più difficile da spiegare di IQR.

```
> mad(discoveries)

## [1] 1.4826
```

#### 4.1.2.4 Misure della forma

- **Misura della simmetria.**

$$g_1 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}. \tag{4.2}$$

Il valore di $g_1$ è nell'intervallo $(-\infty, +\infty)$. Se $g_1 > 0$ la distribuzione è considerata *asimmetrica a destra*, se $g_1 < 0$ è *asimmetrica a sinistra*, se $|g_1| < \epsilon$ è *simmetrica*. Anche se è difficile dare delle informazioni di tipo quantitativo, si definisce in generale *asimmetrica* una distribuzione che abbia un valore

$$|g_1| > 2\sqrt{6/n}.$$

- **Misura della curtosi.**

$$g_2 = \frac{1}{n}\frac{\sum_{i=1}^{n}(x_i - \overline{x})^4}{s^4} - 3. \tag{4.3}$$

Il valore di $g_2$ è nell'intervallo $(-2, +\infty)$. Se $g_2 > 0$ la distribuzione è considerata *leptocurtica*, se $g_2 < 0$ è *platicurtica*, se $|g_1| < \epsilon$ è *mesocurtica*. Anche in questo caso è difficile dare delle informazioni di tipo quantitativo, ma si dice che una distribuzione ha un *eccesso di curtosi* se

$$|g_2| > 4\sqrt{6/n}.$$

#### 4.1.2.5 I cardini e i cinque numeri di sintesi

**I cardini** di un insieme sono definiti nel modo seguente, dopo avere ordinato in modo crescente gli elementi $x_1 < x_2 < \ldots < x_N$:

- *cardine inferiore* $h_L$ è l'elemento in posizione:

$$L = \lfloor (n+3)/2 \rfloor / 2,$$

dove $\lfloor x \rfloor$ indica l'intero più grande minore o uguale ad x. Se L non è intero, allora il cardine inferiore è la media dei due valori adiacenti ad L.
- *cardine superiore* $h_U$ è l'elemento in posizione $n + 1 - L$.

Dati i cardini, i **cinque numeri di sintesi (5NS)** sono definiti come:

$$5NS = (x_1, h_L, \tilde{x}, h_U, x_N),$$

dove $\tilde{x}$ è la mediana.

In R il comando `summary` fornisce i cinque numeri di sintesi:

```
> summary(discoveries)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     2.0     3.0     3.1     4.0    12.0
```

I 5NS sono utili in quanto descrivono con pochi valori diverse caratteristiche dell'insieme dei dati, il centro, la diffusione e la forma. La rappresentazione grafica dei 5NS è il `boxplot` che vedremo nella sezione successiva.

#### 4.1.2.6 Outliers

Si definisce **outlier potenziale** un'osservazione che si trova ad una distanza dal centro superiore a 1.5 volte rispetto all'ampiezza dell'intervallo $[h_L, h_U]$, cioè che appartiene all'intervallo $[h_L - 1.5(h_U - h_L), h_U + 1.5(h_U - h_L)]$.

Si definisce **outlier sospetto** un'osservazione che si trova ad una distanza dal centro superiore a 3 volte rispetto all'ampiezza dell'intervallo $[h_L, h_U]$, cioè che appartiene all'intervallo $[h_L - 3(h_U - h_L), h_U + 3(h_U - h_L)]$.

Nel momento in cui si identificano degli outliers, è necessario innanzitutto cercare di capire quale è la causa, per poi agire di conseguenza. Per esempio, se si verifica essere un errore di trascrizione della misura, può essere corretto; se invece dipende da qualche malfunzionamento, si elimina; se non si trova la causa, bisogna considerare la possibilità che sia effettivamente un valore dei dati e quindi tenere conto della modifica delle caratteristiche dell'insieme stesso dei dati.

# 4.2 Visualizzazione dei dati con la libreria di base

Come evidenziato nella Sezione 3.5 la visualizzazione dei dati ha un ruolo importante nell'analisi dei dati. Citando Wickham [2009a], la libreria di base è stata sviluppata non sulla base di una grammatica della grafica, quanto piuttosto sulla base di un modello *carta e penna*. Nonostante questo, la libreria di base può essere utile perché per la sua facilità d'uso può servire in una fase iniziale per capire come *sono fatti* i dati che si vogliono rappresentare ed analizzare. Inoltre, molte funzioni di R la utilizzano.

Nelle sezioni seguenti si presenteranno prima, nella Sezione 4.2.1, le funzioni di basso livello, che, per rimanere nel modello *carta e penna*, permettono di definire l'area della *carta* su cui si traccerà il grafico e le proprietà della *penna* che traccerà il grafico; nella Sezione 4.2.2 le funzioni di alto livello, in particolare quelle per costruire istogrammi, diagrammi di dispersione e le cosiddette *scatole coi baffi* (*box-and-whisker*), i `boxplot`. Queste funzioni sono state utilizzate nelle sezioni precedenti e saranno utilizzate in modo prevalente nei capitoli che riguardano la probabilità, Capitolo 5, e la statistica inferenziale, Capitolo 6.

Per questa sezione sono necessari, i pacchetti di sistema `dataset`, `graphics` e `stats`. Per illustrare il comportamento delle funzioni si utilizzeranno in prevalenza i dati installati con i pacchetti di sistema, che verranno indicati quando li si usa.

## 4.2.1 Le funzioni di basso livello

La definizione della pagina può essere effettuata tramite la funzione che definisce i parametri grafici `par`. Questa funzione ha argomenti che permettono di definire, per esempio,

- la dimensione del grafico, con parametro `pin`;
- i margini da lasciare ai bordi, parametro `mai` che definisce i margini sotto, a sinistra, sopra e a destra in pollici oppure `mar` che li definisce in linee;
- quanti grafici disporre sulla pagina con parametro `mfrow` oppure `mfcol`.

Per vedere i parametri possibili si può dare il comando `?par` nella finestra di console di R. I parametri hanno dei valori predefiniti, che sono visualizzabili con il comando `par()`, sempre nella console.

Per realizzare un grafico si può usare la funzione `plot`. Le funzioni `points` e `lines` aggiungono punti e linee ad un `plot`. Punti e linee a loro volta possono essere rappresentati in modo diverso da quello predefinito, tramite opportuni parametri grafici. I colori delle diverse componenti possono a loro volta essere definiti o ridefiniti.

Se si vogliono introdurre i colori ed altri effetti "estetici" si consulti il manuale relativo ai comandi ed il sito `http://research.stowers-institute.org/efg/R/Color/Chart/`. Seguendo il link `PDFofChartofRColors` del

sito si può anche scaricare un file che documenta i colori che è possibile utilizzare in R.

---

***Esempio 4.1*** *Esempi di grafici*

***Soluzione.***

- Grafico, con margini, due funzioni tracciate per punti e per linee di diverso tipo e colore, due assi in riferimento alle diverse funzioni.
  Come al solito si inizia pulendo lo spazio di lavoro:

```
> source('~/RStudio/Script/Util/clean.R')
```

  In via preliminare, definiamo alcune funzioni e costruiamo dati da visualizzare

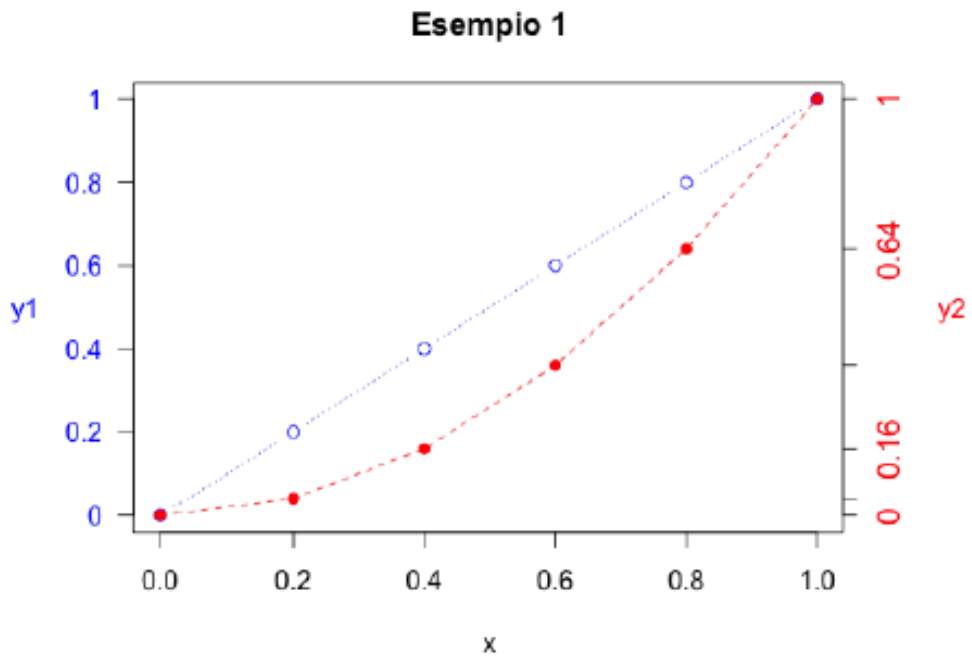
```
> f1 <- function(x) x
> f2 <- function(x) x * x
> x <- seq(0, 1, 0.2)
> y1 <- sapply(x, f1)
> y2 <- sapply(x, f2)
```

  Si osservi che per non creare confusione all'inizio viene salvata la configurazione predefinita degli argomenti grafici di `par` (`parPrima <- par(no.readonly = TRUE)`). Questa configurazione viene ripristinata alla fine con `par(parPrima)`. Il primo grafico è generato specificando in sequenza (`c(sotto, sinistra, sopra, destra)`) i margini della figura con parametro `mar`. Poi si genera il grafico con `plot` specificando prima i valori per l'asse `x` e per l'asse `y`, `y1`, e poi con i seguenti parametri grafici:
  - con `type = "b"` si indica che ci sono punti e linee,
  - con `pch = 21` si seleziona il simbolo per i punti (in questo caso un cerchio vuoto),
  - con `col = "blue"` si definisce il colore dei punti e delle righe,
  - con `yaxt = "n"` si indica che non si vuole l'asse `y` (sarà definito dopo),
  - con `lty = 3` il tipo di linea (in questo caso una linea punteggiata),
  - con `ann = FALSE` per rifiutare titoli ed etichette predefinite in alcune funzioni di R.

  Con `lines` si aggiunge il grafico in `x` dell'altra funzione, `y2`, utilizzando linee tratteggiate (`type = "l"` e `lty = 2`). Con `points` si aggiungono punti al grafico generato da `lines`, questa volta cerchietti pieni (`pch = 16`). Linee e punti di colore rosso (`col = "red"`).
  Con la prima funzione `axis` si definisce l'asse a sinistra (`side = 2`) con tacche in corrispondenza dei valori di `y(x)`, con valori a due cifre (`labels = round(y1, digits = 2)`), in colore corrispondente alla curva e ai punti (`col = "blue"`) e perpendicolari all'asse (`las = 2`). Con la funzione

`mtext` si mettono etichette sull'asse specificando l'etichetta (`"y1"`), il lato (`side = 2`), la distanza in linee dall'asse (`line = 3`), la dimensione uguale a quella predefinita (`cex.lab = 1`), la perpendicolarità alla direzione dell'asse (`las = 2`) ed il colore (`col = "blue"`).
Si ripete per l'altro asse sul lato destro (`side = 4`) con la differenza che ora valori ed etichette sono rossi e i valori sono paralleli all'asse (`las = 0`).
Infine, con la funzione `title` si definisce il titolo del grafico, `"Esempio 1"`, e l'etichetta dell'asse `x`, `xlab = "x"`.
Può non essere superfluo sottolineare che il grafico creato serve solo come dimostrazione del modo in cui si definiscono le varie componenti di un grafico.

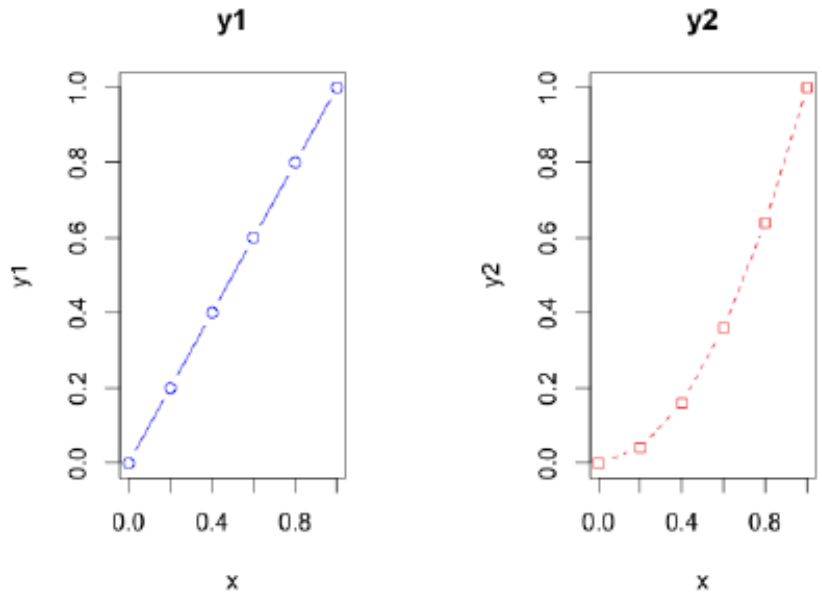
```
> # Salva valori predefiniti par
> parPrima <- par(no.readonly = TRUE)
> # definisce margini
> # c(sotto, sinistra, sopra, destra)
> par(mar = c(5, 6, 4, 5) + .1)
> # genera il grafico con linee e punti
> # specificando che non si vuole asse
> # yaxt = ""n
> plot(x, y1, type = "b", pch  = 21, col = "blue",
+      yaxt = "n", lty = 3, ann = FALSE)
> # aggiunge linee y2 al grafico
> lines(x, y2, type = "l",
+       col = "red", lty = 2)
> # aggiunge punti y2 al grafico
> points(x, y2, pch = 16, col = "red")
> # tacche e valori a sinistra
> axis(side = 2, at = x,
+      labels = round(y1, digits = 2),
+      col.axis = "blue", las = 2)
> # etichette asse sinistra
> mtext("y1", side = 2, line = 3, cex.lab = 1,
+       las = 2, col = "blue")
> # tacche e valori a destra
> axis(side = 4, at = y2,
+      labels   = round(y2, digits = 2),
+      col.axis = "red", las = 0,
+      cex.axis = 1.2, tck = -0.03)
> # testi a destra
> mtext("y2", side = 4, line = 3, cex.lab = 1,
+       las = 2, col = "red")
> # titolo grafico ed etichetta asse x
> title("Esempio 1", xlab = "x")
```




```
> # ripristina valori predefiniti par
> par(parPrima)
```

- Due grafici affiancati.
  Anche qui si proteggono i valori dei parametri grafici predefiniti. I due grafici vengono realizzati con due invocazioni della funzione `plot`. L'istruzione `par(mfrow = c(1,2))` definisce come dovranno essere impaginati i due grafici (si provi a modificare l'istruzione in `par(mfrow = c(2,1))`).
  Ci sono dei parametri grafici nuovi rispetto a quelli visti nell'**Esempio 4.1** per `plot`:
  - `xlab` che definisce l'etichetta per l'asse `x`,
  - `ylab` che definisce l'etichetta per l'asse `y`,
  - `main` che definisce il titolo del grafico.

```
> parPrima <- par(no.readonly = TRUE)
> # due grafici affiancati
> par(mfrow = c(1,2))
> # genera il grafico con linee e punti
> plot(x, y1, type = "b", pch  = 21, col = "blue",
+      xlab = "x", ylab = "y1",
+      main = "y1", lty = 1)
> # genera altro grafico con linee e punti
> plot(x, y2, type = "b", pch = 22, col = "red",
+      xlab = "x", ylab = "y2",
+      main = "y2", lty = 2)
```




```
> par(parPrima)
```

## 4.2.2 Le funzioni di alto livello

Delle funzioni di alto livello si considereranno quelle che permettono di realizzare istogrammi (`hist`), diagrammi di dispersione (`plot`) e grafici per rappresentare i cinque numeri di sintesi (`boxplot`). Le funzioni verranno presentate con le opzioni predefinite; per modificarle valgono gli stessi criteri visti nella sezione precedente. informazioni sulle funzioni e sui loro parametri grafici possono essere ottenute con i comandi `?hist`, `?plot` e `?boxplot`, nella console.

Gli istogrammi sono usati tipicamente per rappresentare dati quantitativi continui. Si devono stabilire delle classi, o `bin`, che dividono l'intervallo dei dati in un insieme di sotto–intervalli che contengono i valori dei dati. Quindi vengono disegnati dei rettangoli di altezza proporzionale al numero di osservazioni che cadono nel sotto–intervallo che costituisce la base del rettangolo.

*Esempio 4.2* *Consideriamo come esempio i dati contenuti nel file* `AirPassengers` *del pacchetto* `datasets`. *Il file contiene* 144 *dati, in forma di numeri interi, che rappresentano il numero di passeggeri mensili dal* 1949 *al* 1960 *della compagnia Box & Jenkins airline.*

*Soluzione.* Il comando (`main=""` elimina il titolo della figura):

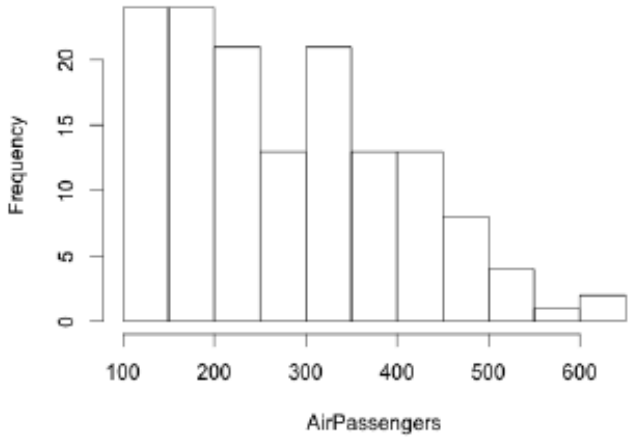



**Figura 4.2:** Istogramma con bin predefiniti di `AirPassengers`.

```
> data(AirPassengers, package = "datasets")
> hist(AirPassengers, main = "")
```

importa il file di dati nell'ambiente di lavoro e produce la **Figura 4.2** con i bin predefiniti della funzione `hist`.

Cambiando il numero di suddivisioni con il parametro `breaks`

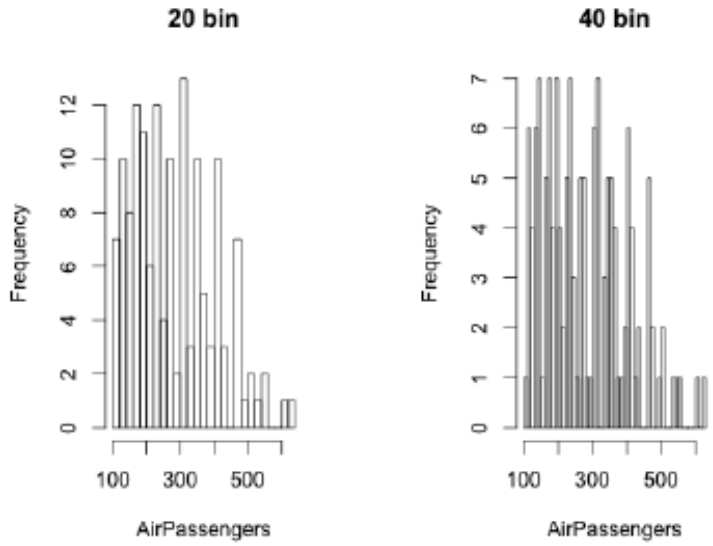



**Figura 4.3:** Istogramma del dataset *AirPassengers* con diversi numeri di bin.

```
> hist(AirPassengers, main = "", breaks = 20)
> hist(AirPassengers, main = "", breaks = 40)
```

si ottengono i grafici in **Figura 4.3** che mostrano, a sinistra, l'istogramma con 20 bin e, a destra, quello con 40 bin.

È evidente dalle figure che il risultato di un istogramma dipende molto dal numero di bin utilizzati. Ci sono algoritmi in R che *scelgono* il numero dei bin, ma nessuno di essi è ottimale in tutte le situazioni. Si dovrebbero sempre fare varie prove con diverse suddivisioni, per scegliere quello che si ritiene migliore.

---

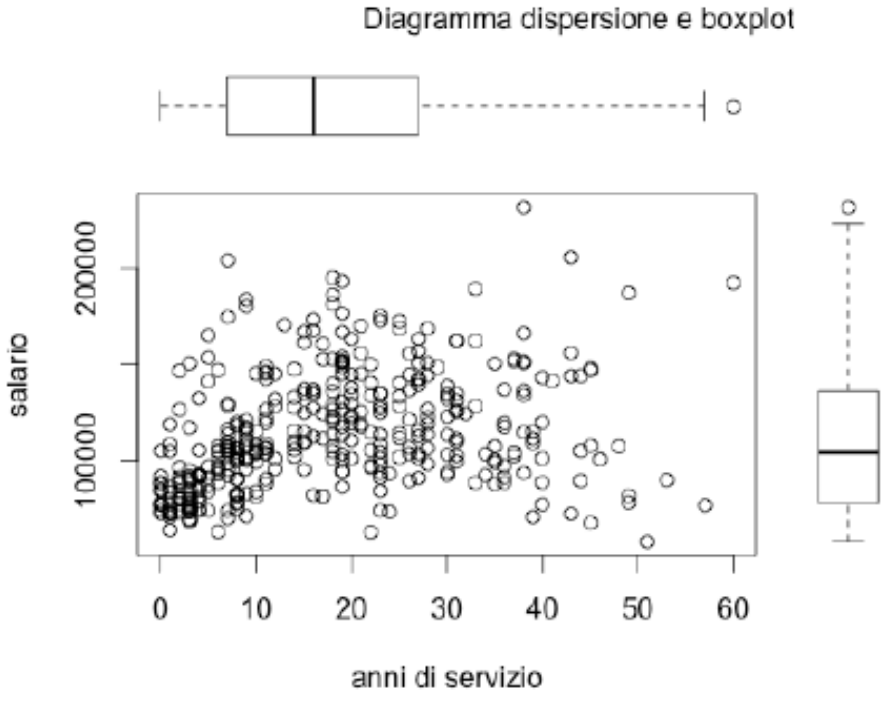
Con l'esempio seguente mostriamo come usare le funzioni `plot` e `boxplot` insieme per visualizzare sia un diagramma di dispersione di due variabili che i cinque numeri di sintesi che caratterizzano le due variabili.

---

***Esempio 4.3*** *Esempio di utilizzo di diagrammi di dispersione e della funzione che permette di visualizzare graficamente i cinque numeri di sintesi che caratterizzano le variabili.*

***Soluzione.***

Per questo esempio utilizziamo i dati `Salaries` del pacchetto `car` che devono essere importati nell'ambiente di lavoro.



```
> parPrima <- par(no.readonly = TRUE)
> par(fig = c(0., 0.8, 0., 0.8))
> plot(Salaries$yrs.service, Salaries$salary,
+      xlab = "anni di servizio", ylab = "salario")
> par(fig = c(0., 0.8, 0.45, 1), new = TRUE)
> boxplot(Salaries$yrs.service, horizontal = TRUE,
+         axes = FALSE)
> par(fig = c(0.55, 0.95, 0., 0.8), new = TRUE)
> boxplot(Salaries$yrs.service, axes = FALSE)
> mtext("Diagramma dispersione e boxplot",
+       side = 3, outer = TRUE, line = -4)
> par(parPrima)
```

Le figure sopra e a destra del diagramma di dispersione riportano i cinque numeri di sintesi del salario e degli anni di servizio. I numeri sono identificati nelle barrette agli estremi dei *baffi*, `min` e `max`, nei lati del parallelogramma alla base dei *baffi*, $h_L$ ed $h_U$, e nella linea in grassetto al centro della scatola, $\tilde{x}$.