Data Analysis

Prime rappresentazioni grafiche e altre sintesi dei dati - una singola variabile

Nicola Torelli

18/3/2021

Contents

Primi grafici e misure sintetiche

Riassumere i dati: indici e grafici

Gran parte dell'analisi (statistica) dei dati si basa sull'idea di riassumere efficacemente i dati così da riuscire a estrarre al meglio le informazioni che essi contengono su un determinato collettivo.

Vi sono varie strategie per ottenere **riassunti** dei dati in ambito descrittivo/esplorativo (nell'ambito dell'inferenza statistica il concetto di riassunto dei dati può essere impostato in modo più formalizzato).

In estrema sintesi tuttavia gli strumenti che si utilizzano in ambito esplorativo sono di fatto due e sono complementari:

- utilizzo di **misure sintetiche** di specifiche caratteristiche dei dati (esempi banali sono la media e la varianza come misure, rispettivamente, di centralità e dispersione dei dati)
- utilizzo di **sintesi grafiche** e di tecniche di visualizzazione.

L'idea di fondo è che l'elaborazione dei dati grezzi (pensiamo alla matrice dei dati) sia necessaria così da comprimere e rendere comprensibile l'informazione in essi contenuta oltre a renderla più semplice da comunicare e conservare.

Uno dei principi fondamentali è che la elaborazione dei dati e la loro sintesi (con indici o grafici) comporti inevitabilmente una perdita di informazione: è il prezzo da pagare per poter *leggere* i dati e per trarre da essi

informazione utile. Ci si aspetta tuttavia che l'informazione che si perde sia non rilevante per interpretare il fenomeno e consente di ottenere una visione di insieme delle caratteristiche salienti del collettivo che si sta esaminando.

La semplice elaborazione con tabelle di frequenza, che abbiamo già introdotto, è un primo esempio di sintesi dei dati e può essere usata per illustrare questo aspetto. Se i miei dati contengono informazioni su una variabile (ad esempio il ricorso all'avvocato nei dati `AutoBi`) e il mio obiettivo è sapere se in quel collettivo si tenda a ricorrere spesso all'avvocato, la tabella di frequenza riassume tali informazioni in modo efficace. Se conservo solo la tabella perdo l'informazione sulla scelta del singolo caso, contenuta nella relativa riga della matrice dei dati, ma questa è irrilevante per rispondere al mio obiettivo conoscitivo. Il pezzo di informazione che ho perso è in questo trascurabile.

Si consideri ora la variabile `LOSS`. Anche per essa avevamo costruito una tabella di frequenza. Tuttavia per fare questa semplice elaborazione abbiamo costruito le classi di valori e abbiamo trasformato la variabile in fattore dove invece del singolo dato (che in questo caso era verosimilmente diverso per ogni unità) conservavo solo la categoria ovvero la classe cui quel valore apparteneva.

```
library(insuranceData)
data("AutoBi")
AutoBi$LOSSclass<-cut(AutoBi$LOSS,breaks=c(0,0.5,2,4,8,1100))
rbind(table(AutoBi$LOSSclass), prop.table(table(AutoBi$LOSSclass), ))
```

```
##          (0,0.5]    (0.5,2]       (2,4]       (4,8] (8,1.1e+03]
## [1,] 288.0000000 324.000000 396.0000000 195.0000000 137.0000000
## [2,]   0.2149254   0.241791   0.2955224   0.1455224   0.1022388
```

La tabella fornisce una sintesi del fenomeno in esame: ad esempio ora sappiamo che circa il 45% dei casi ha subito un danno sotto i 2000 dollari.

La sintesi è molto più leggibile della lista di 1340 valori riportati nel data frame. Però nel fare la tabella abbiamo perso informazioni. Non sappiamo, se disponessimo solo della tabella, se vi sono dei casi che hanno un danno fra 1500 e 2000 dollari e quanti sono. Ho perso quindi alcuni dettagli e la sintesi potrebbe rivelarsi eccessiva.

Il compromesso fra l'esigenza di sintesi e quella di mantenere il dettaglio informativo è un motivo ricorrente dell'analisi dei dati e in generale dell'elaborazione statistica.

Nella scelta di una opportuna sintesi numerica o di una tecnica grafica questo sarà quindi un tema sempre presente. In molti casi sono proprio le visualizzazioni grafiche che risultano più flessibili e maggiormente in grado di conservare buona parte dell'informazione originale.

Primi semplici grafici con `R`

Vedremo ora prime semplici rappresentazioni, immediatamente comprensibili e ben note, che ci consentiranno di iniziare a esplorare le potenzialità grafiche di `R`, che come vedremo sono enormi.

Abbiamo visto come le tabelle di frequenza rappresentino la più elementare forma di elaborazione e sintesi dei dati. Ad esse si possono associare alcune rappresentazioni grafiche che quindi utilizzeranno direttamente i dati riassunti nell'oggetto generato dal comando `table()`.

Il diagramma a torta

Consideriamo ancora data set `Cars93`

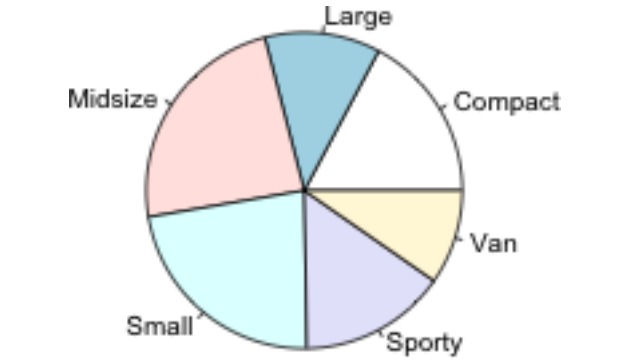
```
library(MASS)
data("Cars93")
#str(Cars93)
```

La variabile "Type"' è il tipo di auto ed è qualitativa per cui possiamo ottenere la tabella di frequenza

```
tabtipo<-table(Cars93$Type)
```

Possiamo associare a tale tabella il cosidetto "pie chart" o, in italiano, **diagramma a torta** con la funzione "'pie()""

```
pie(tabtipo)
```



La fetta di torta corrispondente a ciascuna modalità ha un'area proporzionale alla frequenza della stessa.

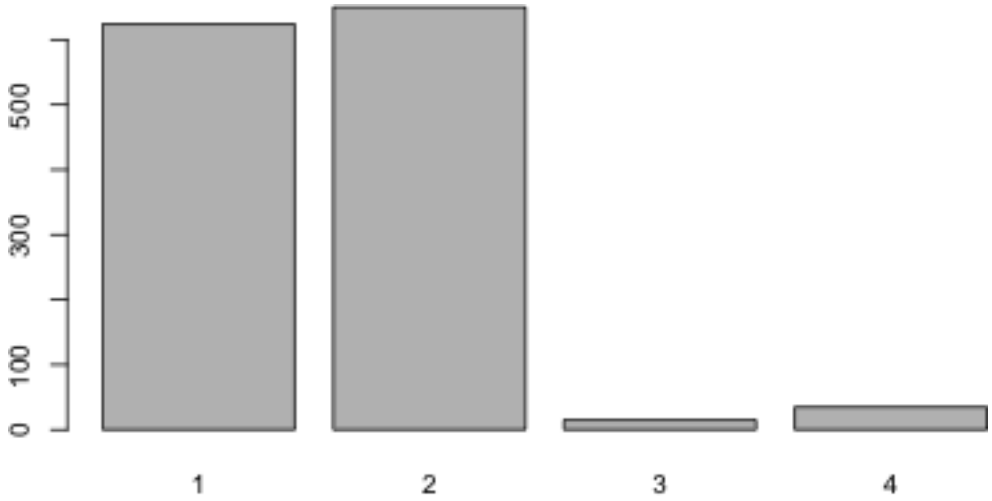
Il diagramma a torta è tanto noto quanto sconsigliato. Se andate a leggere la nota in fondo all'help su `pie()` in `R` trovate: "Pie charts are a very bad way of displaying information. The eye is good at judging linear measures and bad at judging relative areas. A bar chart or dot chart is a preferable way of displaying this type of data".

Il diagramma a barre: `barplot()`

Un diagramma adatto a rappresentare variabili qualitative, o meglio le tabelle di ferquenza ottenute con fattori qualitativi è costituito dal **diagramma a barre**.

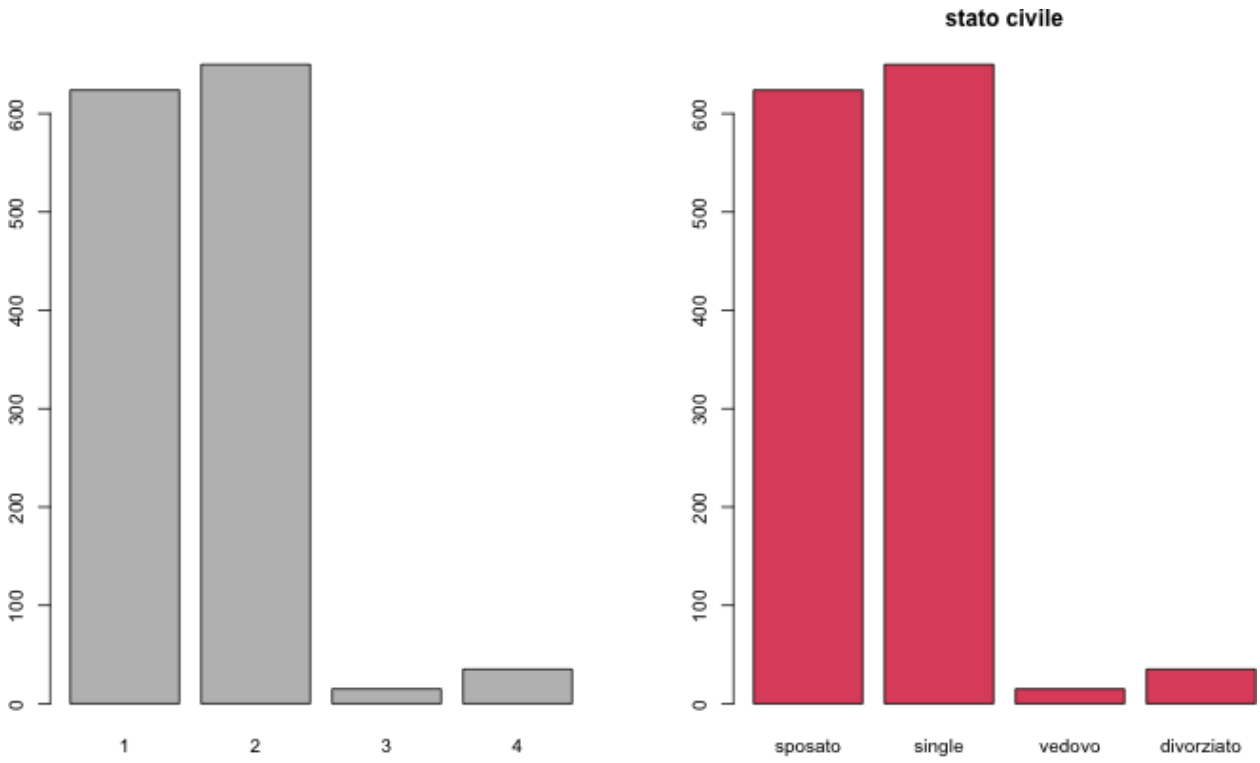
Anche in questo caso la funzione `barplot()` si applica non al vettore di dati originale ma alla tabella di frequenze ricavata da esso. Vediamo un esempio con la variabile `MARITAL` del data frame `AutoBi`:

```
maritab<-table(AutoBi$MARITAL)
barplot(maritab)
```



Stavolta le frequenze sono proporzionali alle altezze dei rettangoli ciascuno dei quali rappresenta una modalità. La funzione contiene alcuni parametri che consentono di personalizzare il grafico. ad esempio:

```
par(mfrow=c(1,2))
# questa funzione serve per settare alcuni parametri della finestra grafica:
# in questo caso divide la finestra grafica in due colonne e una riga. Per cui
# affiancherà i due grafici successivi. si noti che nel secondo grafico
# sono anche utilizzati alcuni parametri per mettere il titolo al grafico
# (main="titolo" o per cambiare il colore col="2")
barplot(maritab)
barplot(maritab, main="stato civile", names.arg=c("sposato", "single",
                                    "vedovo", "divorziato"), col=2)
```



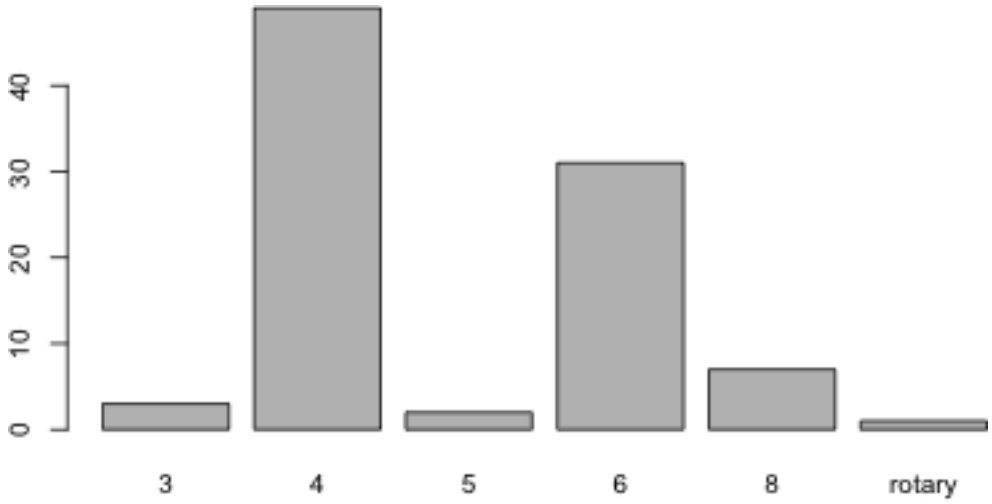

```
par(mfrow=c(1,1))
# così torniamo ad avere un solo grafico per ciascuna finestra grafica
```

Si vede bene che ogni rettangolo è separato, la sua larghezza e la posizione sull'asse orizzontale è arbitraria. Questo deve far riflettere sul fatto che tale grafico possa essere poco appropriato se la tabella di frequenze riguarda una variabile quantitativa trasformata in fattore con il metodo delle classi di valori.

Infatti in quest'ultimo caso i valori numerici postrebbero essere rappresentati più apportunamente sull'asse orizzontale. In particolare, ma non solo, se si tratta di una variabile quantitativa continua.

Il messaggio è quindi: **non** utilizzare il barplot per rapfresentare una variabile quantitativa continua (in classi). Talvolta esso si può tuttavia utilizzare con una variabile discreta con un numero di valori molto basso. ad esempio nel caso del numero di cilindri nel data set delle auto.

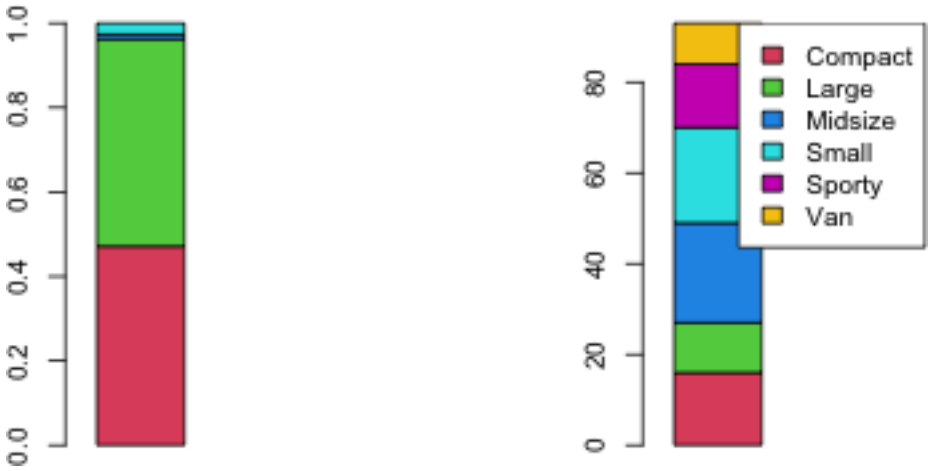
```
barplot(table(Cars93$Cylinders))
```



Il diagramma a barre sovrapposte

Un altro metodo per rappresentare una distribuzione di frequenza è quello di ricorrere al diagramma a barre "impilato". Invece di avere una barra separata per ogni frequenza si rappresentano in un'unica barra le porzioni relative a ciascna modalità con le frequenze rappresentate dalla lunghezza di ciascuna porzione. Si suddividono le modalità con dei segmenti di un unico rettangolo la cui lunghezza è proporzionale alla frequenza (versione 'stacked'). Occorre che i dati della tabella di frequenza siano trasformati in un vettore colonna (una matrice con tante righe quante le modalità e una solo colonna).

```
# esempio con stato civile
par(mfrow=c(1,2))
freq<-matrix(prop.table(table(AutoBi$MARITAL)),
             nrow=length(table(AutoBi$MARITAL)), ncol=1)
barplot(freq, xlim=c(0,4), col=(2:5), )
# esempio con auto
nrig=length(table(Cars93$Type))
autof<-matrix(table(Cars93$Type), nrig, ncol=1)
colr=(2:(nrig+1))
barplot(autof, xlim=c(0,4), col=(2:(nrig+1)))
legend(x="top", legend=levels(Cars93$Type), fill=colr)
```



```
par(mfrow=c(1,1))
```

Rappresentare graficamente e riassumere variabili quantitative

Nel caso dell'analisi di variabili quantitative abbiamo già osservato che occorre maggiore attenzione poichè l'eventuale riassunto grafico (o numerico) comporta spesso la perdita di alcune informazioni. Occorre sempre chiedersi se la perdita di informazione è ben bilanciata dalla qualità informativa del riassunto stesso.

Nel seguito introdurremo elementi per analizzare variabili quantitative mediante rappresentazioni grafiche e attraverso riassunti numerici. Vale la pena di segnalare che gli strumenti che verranno introdotti verranno in genere riferiti all'analisi di variabili quantitative continue.

Tuttavia essi sono in genere del tutto adeguati anche nel caso di variabili discrete e specialmente nel caso di variabili discrete che assumono valori elevati e distinti: ad esempio, se le unità statistiche fossero i comuni italiani e la variabile la popolazione degli stessi (che è una variabile discreta), mi aspetto che la variabile assuma valori dell'ordine delle centinaia e che verosimilmente osserverei numerosi valori distinti.

Se si tratta di variabili discrete (spesso di conteggio) che assumono solo valori bassi, quindi con pochi valori distinti (esempio sono la variabile numero di fratelli, o il numero di sinistri con l'auto che l'assicurato denuncia in un anno) l'uso dei semplici grafici visti per le variabili categoriali si rivela spesso del tutto adeguato. I riassunti numerici come media, varianza, etc., sono comunque appropriati e possono essere anche in questo caso utilizzati.

Si ricorda, invece, che i riassunti numerici e le rappresentazioni grafiche che vedremo, **non** devono essere utilizzate **mai** per variabili qualitative/fattori/categoriali (errore che spesso si fa se le variabili sono rappresentate invece che con la modalità con un valorie numerico). E' per tale motivo che risulta sempre opportuno, nell'analisi con `R`, la trasformazione di tali variabili in fattori.

Il diagramma ramo e foglie (stem and leaf)

Un semplice grafico che ha il vantaggio peraltro di non perdere informazione realtivamente alla variabile quantitativa è il **diagramma ramo e foglie**. In `R` viene evocato con la funzione `stem()`. Ad esempio consideriamo la lunghezza in pollici delle auto:

```
Cars93$Length
stem(Cars93$Length)
```

```
##  [1] 177 195 180 193 186 189 200 216 198 206 204 182 184 193 198 178 194 214 179
## [20] 203 183 203 174 172 181 175 192 180 174 202 141 171 177 180 179 176 192 212
## [39] 151 164 175 173 185 168 172 166 184 200 188 191 205 219 164 172 184 190 169
## [58] 175 187 166 199 172 190 170 181 190 188 188 190 194 201 173 177 181 196 195
## [77] 177 184 176 146 175 179 161 162 174 188 187 163 187 180 159 190 184
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   14 | 16
##   15 | 19
##   16 | 123446689
##   17 | 012222334445555666777778999
##   18 | 000011123444445677788889
##   19 | 000001223344556889
##   20 | 001233456
##   21 | 2469
```

Come si vede il grafico è costituito da valori a destra di una linea verticale che corrispondono alla prime due cifre della variabile. A sinistra della linea viene messa la terza cifra (ordinando dalla più bassa alla più alta).

Tali cifre si impilano e la lunghezza della pila (la foglia) rappresenta direttamente la frequenza assoluta.

Il grafico viene utilizzato per piccoli insiemi di dati (di solito non superiori al centinaio di casi). Se i casi sono più numerosi o se vi sono variabili molto disperse (per cui risulta difficile trovare efficacemente i valori da usare per il ramo) la rappresentazione risulta meno efficace e sarà necessario perdere qualche dettaglio: ad esempio se provo a usarla con `AutoBi$LOSS`

```
stem(AutoBi$LOSS)
```

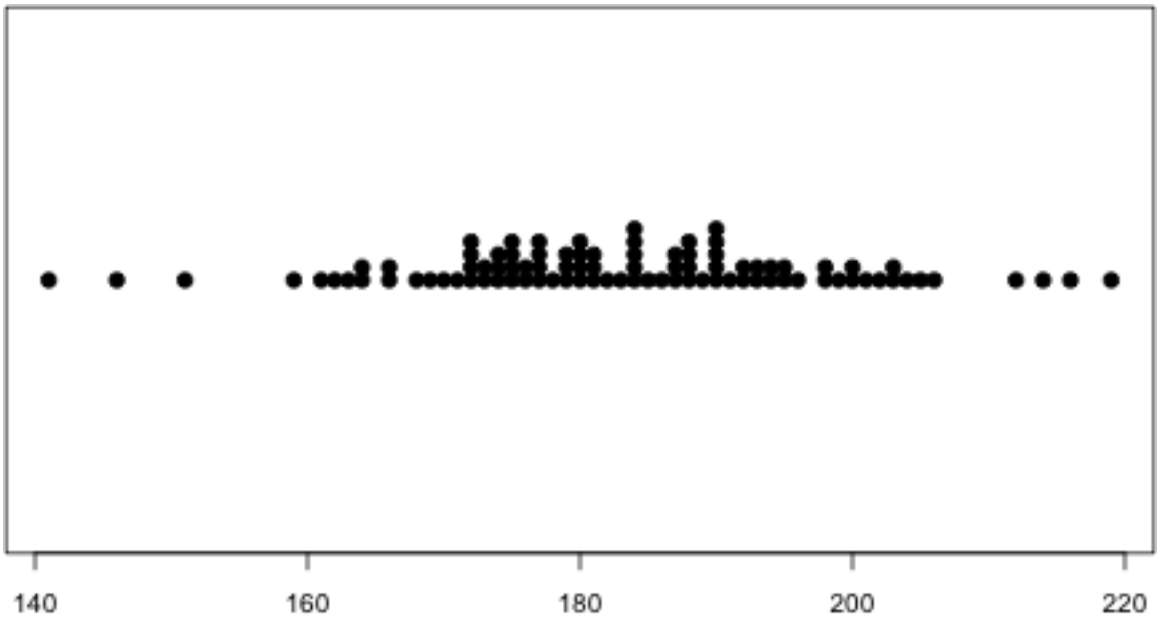
```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##    0 | 00000000000000000000000000000000000000000000000000000000000000000000000000000000+1251
##    1 | 015699
##    2 | 27
##    3 |
##    4 |
##    5 |
##    6 |
##    7 |
##    8 |
##    9 |
##   10 | 7
```

è evidente che, in questo caso, non risulta una rappresentazione grafica efficace.

Il diagramma a punti

Il diagramma a punti di fatto rappresenta su un grafico i singoli valori osservati. Per cui anche in questo caso non si perde alcun dettaglio. La funzione `stripchart()` del package `graphics` può essere usata a tale scopo.

```
stripchart(Cars93$Length, pch=19, method="stack", cex=1.2, ylim=c(0,2))
```




```
# Si noti come nella funzione vengano fissati alcuni parametri. Si tratta di
# parametri grafici che potranno essere utilizzati anche in altri casi Ad esempio
```

```
# pch permette di scegliere quale simbolo utilizzare per rappresentare un punto
# mentre cex permette di variare la grandezza del simbolo stesso
```

Il parametro `method=stack` consente di sovrapporre dati osservati che risultano spesso avere il medesimo valore. Il default è invece quello di mostrare il valore osservato senza dare conto della eventuale molteplicità.

Si considerino ad esempio i dati `Ozone` nel package `mlbench` e in particolare la variabile `V4` che contiene il livello di concentrazione dell'ozono osservato (che appare misurato come una variabile che assume valori interi discreti) in California in tutti i giorni dell'anno

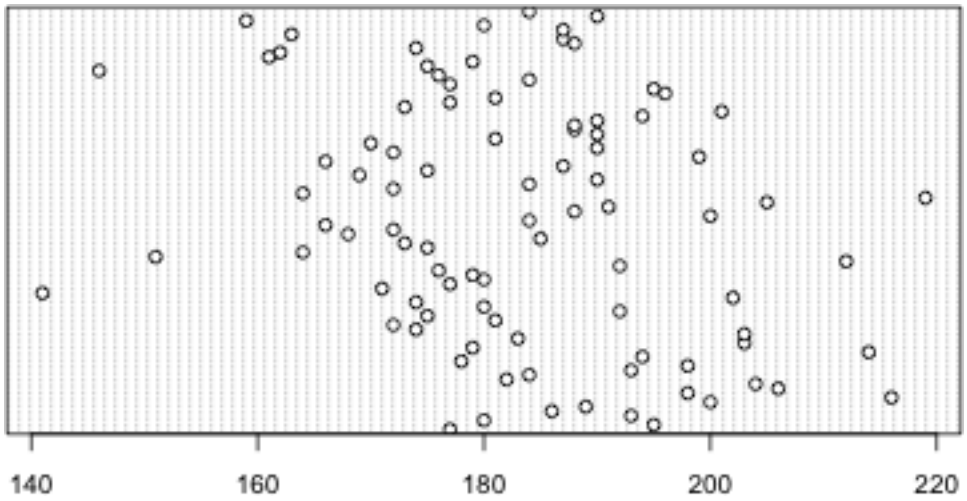
Anche in questo caso sono visibili i singoli valori.

Il diagramma a punti e la funzione `plot()`

Il diagramma a punti fornisce semplicemente una rappresentazione in un grafico di ogni singolo punto osservato. Non vi è quindi alcuna sintesi dei dati. Tuttavia il suo utilizzo è certamente di interesse se si dispone di un numero limitato di casi da analizzare.

La funzione dotchart fa essenzialmente questo: rappresenta su un sistema di assi cartesiano i singoli punti.

```
dotchart(Cars93$Length)
```

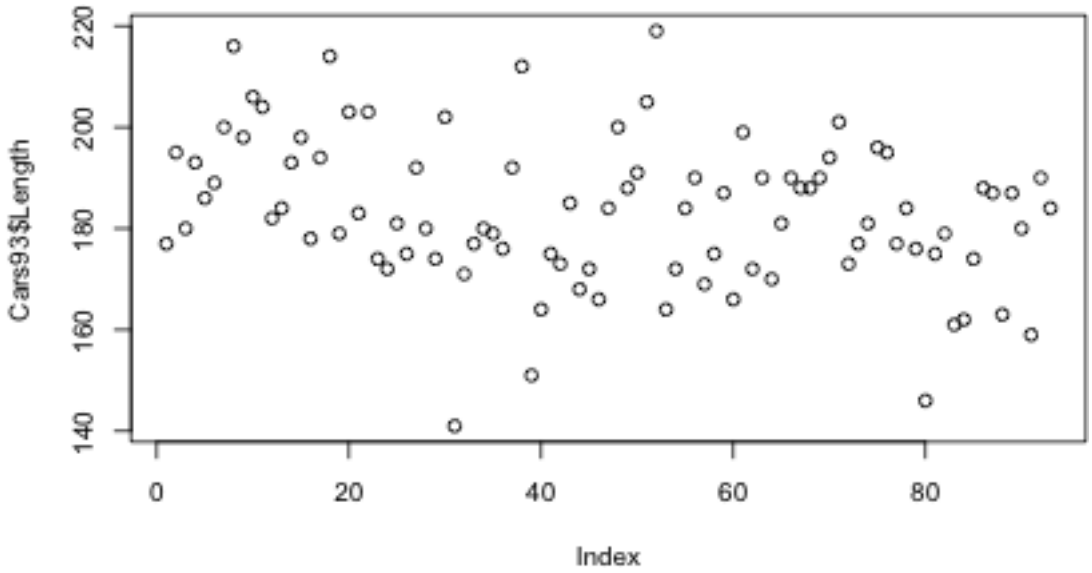


Le coordinate sull'asse verticale riflettono solo l'ordine con cui appaiono nella variabile le singole osservazioni.

E' importante ossservare che tale grafico è del tutto equivalente (nel senso che si sono solo invertiti gli assi) alla funzione **`plot()`** che introduciamo ora ma tratteremo con maggior dettaglio in seguito. Essa consente di riprodurre su un sistema di assi un'insieme di punti fornendo le loro coordinate. Di base quindi si dovrebbe fornire una coppia di vettori che rappresentano le coordinate. Tale funzione è estremamente generale e consentirà di ottenere facilmente molte rappresentazioni grafiche di insiemi di dati bivariati (oltre che essere utilizzato per rappresentare graficamente, ad esempio, curve o funzioni di una variabile).

In effetti posso usare semplicemente

```
plot(Cars93$Length)
```



Il sistema si aspetta una coppia di variabili ma se ne indico solo una usa come coordinata sull'asse delle ascisse semplicemente l'ordine con cui il valore appare nel vettore (che come si vede è chiamato `Index` nel grafico).

Riassunti numerici per variabili quantitative

Prima di procedere a illustrare altri metodi grafici per presentare un insieme di dati relativi a una variabile quantitativa converrà richiamare alcuni concetti di base che riguardano le misure di sintesi per riassumere le caratteristiche salienti di una variabile quantitiva (discreta o continua che sia).

Come si è già anticipato, quando si dispone dei dati per una variabile quantitativa è necessario sintetizzare e elaborare i dati così da fare emergere i fatti e le informazioni salienti. Spesso è necessario rinunciare a dettagli minuti che sono relativi a singole unità per tentare di dare una lettura di insieme del fenomeno con riferimento al collettivo da cui i dati sono tratti.

È una operazione essenziale per trasformare i dati grezzi in informazione.

L'attenzione va quindi concentrata su quali caratteristiche del collettivo sono di interesse e decidere se e come queste possono essere riassunte attraverso opportune sintesi numeriche.

I due aspetti dell'insieme di dati cui si guarda in via preliminare sono

- **la tendenza centrale**: ovvero posso con un singolo valore riassumere l'ordine di grandezza del fenomeno di interesse?
- **la dispersione (o variabilità)**: cioè un singolo valore può aiutare ad avere un'idea di quanto i valori numerici della variabile siano diversi tra loro o viceversa si somiglino?

Al fine di trattare tale argomento, e anche per sviluppi successivi, vale forse la pena di introdurre una notazione generale.

Rappresenteremo i valori assunti da una variabile quantitiva X per un insieme di n unità con la notazione $x_1, x_2, \ldots, x_n$.

Indici di tendenza centrale (un solo valore al posto di tanti)

Gli indici di tendenza centrale sono anche noti come **medie**. Si suole distingure far *medie analitiche* e *medie di posizione*.

Medie analitiche

1. **La media aritmetica**

La media aritmetica M è di gran lunga la media analitica più nota e si calcola come

$$M = \sum_{i=1}^{n} \frac{x_i}{n}$$

È già noto che in `R` esiste la funzione `mean` che calcola la media aritmetica degli elementi di un vettore di dati quantitativi.

La media aritmetica ha alcune importanti proprietà:

a. assume un valore compreso fra $min(x_i)$ e $max(x_i)$
b. la somma degli scarti dalla media e nulla: $\sum_{i=1}^{n}(x_i - M) = 0$
c. ha la seguente proprietà di minimo $M = \arg\min_{a \in R} \sum_{i=1}^{n}(x_i - a)^2$

La terza proprietà si può leggere come segue: se decido che invece di ogni singolo dato x_i uso una sintesi che è M potrei avere una perdita di informazione che decido di misurare con la differenza (scarto dalla media) al quadrato $(x_i - a)^2$: la perdita di informazione totale è quindi $\sum_{i=1}^{n}(x_i - a)^2$. La media M è quel valore che rende minima tale perdita.

Inoltre dalla seconda proprietà consegue che $\sum_{i=1}^{n} x_i = nM$. Per cui si può dire che la media M è quel valore che sostituito a ogni singolo dato lascia invariata la somma.

2. **Altre medie analitiche**

Se si parte da questa ultima notazione, si potrebbe cercare un valore che analogamente alla media aritmetica, lascia invariata una certa funzione dei dati.

1. Ad esempio se invece della somma si considera il prodotto dei dati e si cerca la costante M_g che lascia invariata questa funzione, cioè: $\prod_{i=1}^{n} x_i = M_g^n$ si ottiene la **media geometrica**.

$$M_g = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$$

2. Se il valore cercato lascia invariata la somma dei reciproci, si ottiene la media armonica

$$M_{ar} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

3. Analogamente si ottengono le medie di potenze r-esima (con r intero)

$$M_r = \left[\sum_{i=1}^{n} \frac{x_i^r}{n}\right]^{\frac{1}{r}}$$

   che hanno la proprietà di esser funziona monotona non decrescente di r. In virtù di ciò, e osservando che si potrebbe dimostrare che consoiderando il limite per r che tende a 0 si ottiene la media geometrica, si ha quindi il seguente ordinamento $M_{ar} \leq M_g \leq M \leq M_2$ (ove l'uguaglianza vale se tutti i dati hanno lo stesso valore. Da questo segue anche che $M_2^2 \geq M^2$.
4. Si noti che la potenza r-esima della media potenziata di ordine r fornisce il **momento empirico r-esimo**.

Quantili e medie di posizione

Le **medie (e gli indici) di posizione** sono particolari valori che hanno una posizione definita rispetto la sequenza (ordinata) dei dati (e che non è detto appartengano alla sequenza stessa).

A tal fine è rilevante il concetto di **quantile empirico**, ovvero fissata una proporzione p (con $0 \leq p \leq 1$) si definisce x_p **quantile empirico** di ordine p, con $p \in [0,1]$, come quel valore $x \in \mathbb{R}$ tale che

$$x_p : \frac{\#(x_i \leq x_p)}{n} = p$$

.

Si noti che dato un valore p:

1. non è detto che esista un unico valore x_p che realizzi la condizione data;
2. potrebbe non esistere alcun valore che realizzi la condizione data, ma si potranno ritrovare valori di quantili $x_{p'}$ con p' molto prossimo a p.

Si noti quindi che si può considerare la sequenza ordinata $x_1, x_2, \ldots, x_n$

$x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n-1)} \leq x_{(n)}$

e che i valori della sequenza ordinata definiscono i quantili di ordine $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}$.

Pertanto, il generico valore della sequenza ordinata $x_{(i)}$ rappresenta il quantile di ordine p, x_p con $p = \frac{i}{n}$.

Si noti inoltre il rilievo che assume in questo contesto la definizione che specifica x_p come il valore per cui sia p la proporzione di valori **minori o uguali** a esso. Se disponiamo di un insieme di dati non particolarmente numeroso chiedere di contare i dati **strettamente minori** di x_p o quelli minori o uguali può fare differenza. In tal caso infatti la sequenza ordinata fornirebbe i quantili $\frac{0}{n}, \frac{1}{n}, \ldots, \frac{n-2}{n}, \frac{n-1}{n}$.

Per tale motivo a volte si usa una diversa convenzione per cui i dati ordinati $x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n-1)} \leq x_{(n)}$ rappresentano i quantili di ordine $\frac{0.5}{n}, \frac{1.5}{n}, \ldots, \frac{n-1.5}{n}, \frac{n-0.5}{n}$. Cioè secondo tale convenzione $x_{(i)}$ rappresenta il quantile di ordine $x_{\frac{i-0.5}{n}}$ di ordine $p = \frac{i-0.5}{n}$.

Si noti che se n è grande ci si attende che le differenze derivanti dalle diverse convenzioni si annullino.

La mediana e i quartili

Come detto alcuni particolari quantili hanno uno speciale rilievo. In particolare si definiscono i tre **quartili** come $x_{0.25}, x_{0.5}, x_{0.75}$, denominati ripettivamente primo, secondo e terzo quartile.

Particolare rilievo ha poi il secondo quartile che è anche detto **mediana**.

La mediana è quindi definita come quel valore che è tale per cui il 50% dei dati è inferiore (o uguale) a esso e una analoga percentuale è superiore.

Come si è osservato sopra, se cerchiamo un valore $Me = x_{0.5}$ (con la condizione che siano il 50% quelli minori o uguali di Me) potremmo:

- trovare infiniti valori che realizzino esattamente la condizione data (se ad esempio n è pari)
- non trovare alcun valore che realizzi la condizione data (ad esempio se n è dispari).

La cosa si risolve convenzionalmente ponendo:

- $Me = x_{\left(\frac{n+1}{2}\right)}$ se n è dispari
- $Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$ se n è pari.

Si noti tuttavia che qualsiasi valore $x_{\left(\frac{n}{2}\right)} \leq x \leq x_{\left(\frac{n}{2}+1\right)}$ avrebbe la proprietà della mediana.

La **mediana** è usata spesso in coppia con la media aritmetica per caratterizzare il valore centrale di una distribuzione. Si noti che la mediana ha una proprietà analoga a quella della media aritmetica. In particolare

$$Me = \arg\min_{a \in \mathbb{R}} \sum_{i=1}^{n} |x_i - a|$$

Come vedremo meglio con esempi su vari insiemi di dati, la mediana è molto meno sensibile alla presenza di valori estremi (e anomali rispetto alla massa dei dati). In tal senso essa è più **resistente**. La media arimetica invece tende a essere molto influenzata da pochi valori estremi molto diversi dalla massa dei dati.

È quindi una buona regola guardare a entrambi i valori nel riassumere un insieme di dati.

Come per la media, esiste una funzione in `R` per calcolare la mediana usando la convenzione esposta.

```
median(Cars93$Length)
median(AutoBi$LOSS)
```

```
## [1] 183
## [1] 2.331
```

Anche il primo e il terzo quartile sono molto informativi sulla tendenza centrale della distribuzione: si noti infatti che il 50% dei valori centrali sono compresi fra di essi.

Per quanto riguarda la determinazione di tali due quartili vi sono difficoltà analoghe a quanto visto per la mediana. In generale per i quantili quelli usati in `R` si basano sull'interpolazione lineare fra i due quantili successivi osservati x_{p_1} e x_{p^2} tali che $p_1 \leq p \leq p_2$. (se si vuole approfondire si veda l'help della funzione `quantile`)

Va infine ricordato che per caratterizzare la distribuzione dei dati oltre ai quartili (che dividono i dati ordinati in quattro porzioni ugualmente numerose) spesso si definiscono i **decili** (cioè i quantili $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$) o i percentili (cioè i quantili $x_{0.01}, x_{0.02}, \ldots, x_{0.98}, x_{0.99}$).

In `R` la funzione `quantile()` consente di calcolare i quantili per diversi valori di p. La convenzione in questo caso è che la più piccola osservazione sia il quantile di ordine 0 e la più grande sia quello di ordine 1.

```
# se non si indica nulla vengono calcolati i quartili e i quantili di
# ordine 0 e 1, cioè il massimo e il minimo
quantile(AutoBi$LOSS)
```

```
##         0%        25%        50%        75%       100%
##    0.00500    0.64000    2.33100    3.99475 1067.69700
```

```
# calcoliamo i decili per la variabile LOSS
quantile(AutoBi$LOSS, probs=seq(0,1,.1))
```

```
##         0%        10%        20%        30%        40%        50%        60%        70%
##     0.0050     0.2380     0.4842     0.9970     1.6746     2.3310     2.9700     3.6555
##        80%        90%       100%
##     4.5148     8.0765 1067.6970
```

```
quantile(AutoBi$LOSS)
```

```
##         0%        25%        50%        75%       100%
##    0.00500    0.64000    2.33100    3.99475 1067.69700
```

```
# Verifichiamo infine banalmente che il quantile di ordine 0.5 è la mediana
median(AutoBi$LOSS)==quantile(AutoBi$LOSS, probs=0.5)
```

```
##  50%
## TRUE
```

Riassunto dei 5 valori

Si è visto che i 3 quartili danno un riassunto interessante sulla distribuzione.

La mediana è un indice di tendenza centrale e i due quartili esterni danno un'idea di dove sia collocato il 50% centrale della distribuzione. Aggiungendo a questi il valore massimo e il valore minimo osservati nella sequenza di dati abbiamo un'idea di come i dati siano distribuiti sulle code.

I 5 valori <minimo, I quartile. mediana, III quartile, massimo> sono detti riassunto dei 5 valori (in inglese **five numbers summary**) e esiste una funzione `fivenum()`che li restituisce se applicata a un vettore di dati (numerici), inoltre gli stessi vengono restituiti anche nella funzione `summary()`, che fornice però anche la media aritmetica, quando applicata a un vettore numerico. In entrambi i casi i valori mancanti vengono rimossi automaticamente. Ad esempio:

```
# otteniamo  per la variabile LOSS  il riassunto dei 5 valori e il summary
fivenum(AutoBi$LOSS)
summary(AutoBi$LOSS)
```

```
## [1]    0.0050    0.6400    2.3310    3.9955 1067.6970
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.005    0.640    2.331    5.954    3.995 1067.697
```

Le medie sfrondate (trimmed)

Come si è detto, il principale indice di posizione, cioè la media aritmetica, rischia di essere molto influenzata dalla presenza di (pochi) valori difformi rispetto alla gran parte dei dati.

Un tema rilevante è se tali dati anomali (outliers) debbano essere esclusi dalle successive analisi o essere mantenuti. Come vedremo su tale aspetto non esiste una regola generale perchè dipende dalla natura dei dati e in alcuni casi i fenomeni che osserviamo sono tali da poter ammettere che vi siano osservazioni estreme e quindi eliminare tali osservazioni potrebbe essere del tutto arbitrario e inappropriato.

Il tema che si pone è quindi quello comunque di trovare riassunti della tendenza centrale dei dati che non siano eccessivamente influenzati da tali dati anomali. A tal fine si possono calcolare delle medie dei dati disponibili sfrondando l'insieme dei dati dei valori troppo grandi e/o troppo piccoli.

Questo conduce a calcolare una **media sfrondata** fissando una proporzione α di valori da escludere tra quelli estremi.

La funzione `R` per calcolare la media di un vettore ha fra i parametri la possibilità di indicare se si vuole una media sfrondata. Ad esempio:

```
# calcoliamo la media sfrondata al 5% per la variabile LOSS
mean(AutoBi$LOSS, na.rm=TRUE, trim=0.05)
mean(AutoBi$LOSS, na.rm=TRUE)
```

```
## [1] 2.858701
## [1] 5.953461
```

Indici di dispersione (o variabilità)

L'altro importante aspetto di una distribuzione di dati è costituito dalla dispersione dei dati, ovvero quanto i dati sono diversi l'uno dall'altro. Tale aspetto è anche detto **variabilità**.

Anche in questo caso si possono calcolare degli indici numerici che riassumano tale caratteristica e anche in questo caso possiamo scegliere versioni più resistenti ai valori anomali.

La varianza, scarto quadratico medio, coefficiente di variazione

La più nota misura di dispersione per un insieme di n dati è la varianza. Essa è basata sulla quantità

$$DEV = \sum_{i=1}^{n} (x_i - M)^2$$

ovvero la somma degli scarti dalla media aritmetica che è detta **devianza**.

La **varianza** V è la devianza media per cui è definita come

$$V = \frac{\text{devianza}}{n} = \sum_{i=1}^{n} \frac{(x_i - M)^2}{n}$$

Qunado tuttavia la varianza viene calcolata nell'ambito di un approccio inferenziale, ovvero essa non è semplicemente un indice di dispersione ma assume il ruolo di stimatore della varianza di una popolazione infinita da cui è tratto un campione casuale, si conviene usare la divisione per $n-1$. Poichè questo è un contesto piuttosto comune non deve stupire che in `R` la varianza, calcolata con la funzione `var()` è definita come $\frac{\text{devianza}}{n-1}$.

È evidente che se n è grande non vi è grande differenza fra i valori che si ottengono nei due casi e l'intepretazione in termini descrittivi non cambia nella sostanza. Tuttavia è bene essere consapevoli di quale versione si sta usando.

La devianza, e quindi la varianza, è tanto più grande quanto più i dati sono distanti da un valore che è la media aritmetica che è candidato a rappresentarli tutti.

La varianza è espressa in termini di distanze al quadrato e quindi è non è espressa nella stessa unità di misura della variabile. Per tale motivo è frequente ricorrere alla radice quadrata dell'indice, che è detto **scarto quadratico medio** (SQM) o **deviazione standard** (in inglese *standard deviation*).

$$\text{SQM} = \sqrt{V}$$

che è nella stessa unità di misura della variabile. In `R` la funzione che calcola lo scarto quadratico medio è `sd()` (basata sulla varianza divisa per n-1).

Per definizione la varianza (e lo SQM) assumono solo valori positivi e sono pari a 0 solo nel caso che tutti i valori sono uguali tra loro (variabilità nulla).

L'interpretazione dell'indice non è sempre agevole non essendo definito un limite superiore. L'interesse di solito risiede nell'utilizzarla se ad esempio si vuole comparare la variabilità in due diversi insiemi di dati.

Vediamo un esempio in cui calcoliamo con `R` la varianza in due insiemi di dati.

```
# calcoliamo la varianza  e la deviazione standard per la variabile LOSS
# per i maschi e per le femmine
"Maschi"
var(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
sd(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
"Femmine"
var(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
sd(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
##
```

```
## [1] "Maschi"
## [1] 301.0511
## [1] 17.35082
## [1] "Femmine"
## [1] 1745.729
## [1] 41.78192
```

La comparazione è più agevole se le medie nei due gruppi sono simili

```
# calcoliamo la media per la variabile LOSS per i maschi e per le femmine
mean(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
mean(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
##
```

```
## [1] 5.652647
## [1] 6.213765
```

In questo caso i due gruppi hanno media non molto diversa e qui il confronto darebbe un'indicazione che vi è più variabilità nel caso delle femmine. Tuttavia è facile notare che tale maggiore variabilità è fortemente condizionata dalla presenza di valori anomali. A tal fine si consideri il summary dei due insiemi:

```
# calcoliamo la varianza per la variabile LOSS per i maschi e per le femmine
summary(AutoBi$LOSS[AutoBi$CLMSEX==1])
summary(AutoBi$LOSS[AutoBi$CLMSEX==2])
##
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.030   0.628   2.372   5.653   3.901 222.405      12
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##    0.0050    0.6905    2.2270    6.2138    4.0252 1067.6970        12
```

Si vede chiaramente che nel caso delle femmine vi sia un valore estremamente elevato. Esso ha impatto sulla media (non sulla mediana) e quindi sulla varianza. La varianza è, come la media un indice poco "resistente".

Altre misure di variabilità: coefficiente di variazione, scarto interquartile, MAD

Come per la tendenza centrale anche per la variabilità esistono altre misure che hanno caratteristiche che li rendono in alcuni casi preferibili alla varianza o quanto meno da affiancare ad essa.

1. **lo scarto interquartile**: è semplicemente definito come la differenza fra il III quartile e il I quartile. Ovviamente risente meno dei valori estremi, ma tiene conto solo della dispersione nella parte centrale della distribuzione. Ovviamente è anche calcolabile direttamente dai dati che fornisce `summary`
2. **il MAD**: Scarto assoluto mediano dalla mediana. esso è definito come

$$MAD = \text{Mediana}(|x_i - Me|)$$

   È un indice resistente, cioè non risente di eventuali valori anomali. Si calcola con la funzione `mad()`.
3. il **coefficiente di variazione**: spesso il confronto fra la variabilità in due collettivi ha senso se nei due collettivi la variabile ha circa lo stesso ordine di grandezza (i due gruppi hanno media non eccessivamente diversa). Per ovviare a questo problema si introduce il coefficiente di variazione che è quindi un indice di variabilità relativo. Esso è semplicemente definito come il rapporto fra scarto quadratico medio e media aritmetica.

Vediamo ancora l'esempio precedente con i dati di `LOSS` separatamente per maschi e femmine.

```
# calcoliamo la varianza e il MAD per la variabile LOSS per i maschi e
# per le femmine
"varianza"
var(AutoBi$LOSS[AutoBi$CLMSEX==1], na.rm=TRUE)
var(AutoBi$LOSS[AutoBi$CLMSEX==2], na.rm=TRUE)
"MAD"
mad(AutoBi$LOSS[AutoBi$CLMSEX==1], constant=1, na.rm=TRUE)
mad(AutoBi$LOSS[AutoBi$CLMSEX==2], constant=1, na.rm=TRUE)
##
```

```
## [1] "varianza"
## [1] 301.0511
## [1] 1745.729
## [1] "MAD"
## [1] 1.6635
## [1] 1.6755
```

Si noti come la variabilità misurata dal MAD è molto simile nei due gruppi a testimonianza dell'impatto dell'unico valore estremo nel gruppo delle femmine sul calcolo della varianza.

Simmetria e curtosi

Sintetizzare solo la tendenza centrale o la dispersione di una variabile quantitativa lascia in ombra molte altre caratteristiche della distribuzione dei dati. È pertanto buona regola guardare a altre specificità e cercare dei valori che li sintetizzino.

Una rilevante caratteristica di un insieme di dati è legata alla tendenza degli stessi a distribuirsi in modo molto diverso nella parte sinistra della distribuzione (i valori bassi quindi) rispetto alla parte destra.

Se si fa riferimento a un indice di tendenza centrale, ad esempio la mediana, potrebbe accadere che i dati a destra siano molto più dispersi rispetto a quelli a sinistra ovvero la distribuione presenta una coda a destra più lunga. Si dirà che la distribuzione ha una asimmetria positiva. Viceversa, se vi è una coda a sinistra più lunga si parla di asimmetria negativa.

Se accade questo i dati evidenziano una **asimmetria** (*skewness* in inglese) nella distribuzione della variabile. Il concetto complementare è quindi quello di **simmetria** che riguarda la situazione in cui i dati a destra e a sinistra esibiscono un comportamento analogo.

È facile apprezzare la simmetria (o l'asimmetria) guardando ad alcuni percentili della distribuzione. Vediamo due esempi in cui consideriamo i decili di due variabili nei dati di `AutoBI`:

```
# otteniamo i decili per le variabile LOSS e CLMAGE
"LOSS"
quantile(AutoBi$LOSS, probs = seq(0,1,0.1), na.rm=T, digits=2)

"CLMAGE"
quantile(AutoBi$CLMAGE, probs = seq(0,1,0.1), na.rm=T)
```

```
## [1] "LOSS"
##      0%     10%     20%     30%     40%     50%     60%     70%
##  0.0050  0.2380  0.4842  0.9970  1.6746  2.3310  2.9700  3.6555
##     80%     90%    100%
##  4.5148  8.0765 1067.6970
## [1] "CLMAGE"
##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    0   13   18   21   26   31   36   41   47   55   95
```

Confrontando i quantili a destra e a sinistra della mediana è evidente che la distribuzione della variabile età è meno asimmetrica della variabile `LOSS`.

Si tratta quindi di proporre degli indici sintetici che riassumano tale caratteristica (ovvero l'asimmetria).

Indice basati su particolari quantili

Come abbiamo visto un modo per valutare la simmetria è quello di guardare la posizione di alcuni quantili rispetto alla mediana. Avendo già discusso del rilievo che si dà ai tre quartili, un indice può esser costruito guardando alla posizione del secondo quartile $Q2 = x_{.5}$, ovvero la mediana, rispetto agli altri quartili $Q1 = x_{0.25}$ e $Q3 = x_{0.75}$.

- in assenza di asimmetria sarà $(Q3 - Q2) = (Q2 - Q1)$
- se vi è asimmetria positiva (coda a destra lunga) $(Q3 - Q2) > (Q2 - Q1)$
- se vi è asimmetria negativa (coda a destra lunga) $(Q3 - Q2) > (Q2 - Q1)$

Allora un indice (di Galton) basato sui quartili potrebbe allora essere il seguente:

$$G = \frac{Q3 + Q1 - 2Q2}{Q3 - Q1}$$

La stessa idea potrebbe applicarsi a quantili più estremi, ad esempio, il primo e il nono decile $x_{0.1}$ e $x_{0.9}$ confrontati con la mediana.

$$K = \frac{x_{0.9} + x_{0.1} - 2x_{0.5}}{x_{0.9} - x_{0.1}}$$

Indici basati sul confronto fra media e mediana

Abbiamo già notato che il diverso comportamento di media e mediana è legato alle quantità che minimizzano e alla tendenza della media a essere influenzata da valori sulla coda della distribuzione. Quindi:

- se la media è circa uguale alla mediana è un sintomo di assenza di asimmetria
- se la media è superiore (inferiore) alla mediana è un sintomo di asimmetria positiva (negativa) in quanto avrò prevalenza di valori (alcuni molto distanti dalla media)sulla coda destra (sinistra)

Pertanto si può costruire un indice (coefficiente di asimmetria) come segue

$$\frac{3(M - Me)}{sd}$$

Indice di asimmetria

L'altra possibilità è quella di costruire un indice basato ancora sugli scarti dalla media elevati al cubo $(x_i - M)^3$. Ovviamente è rilevante in questo caso se uno scarto è positivo o negativo.

L'**indice di asimmetria** è definito come:

$$\gamma = \frac{\sum_{i=1}^{n} \frac{(x_i - M)^3}{n}}{sd^3}$$

Tale indice è nullo in assenza di asimmetria e assume valori positivi nel caso di asimmetria positiva (coda destra lunga) e negativi nel caso di asimmetria negativa (coda sinistra lunga).

In `R` esso può essere calcolato con la funzione `skewness()` che però si trova nel pacchetto `moments`

```
library(moments)
# otteniamo varie misure di asimmetria per le variabili LOSS e CLMAGE
"LOSS"
"Indice di Galton"
Qloss<-fivenum(AutoBi$LOSS, na.rm=T)
g<-(Qloss[4]+Qloss[2]-2*Qloss[3])/(Qloss[4]-Qloss[2])
"Indice di asimmetria"
skewness(AutoBi$LOSS)
"CLMAGE"
```

```
## [1] "LOSS"
## [1] "Indice di Galton"
## [1] "Indice di asimmetria"
## [1] 25.68795
## [1] "CLMAGE"
```

Indice di curtosi

La curtosi è un'ulteriore caratteristica della distribuzione di una varaibile quantitativa. Essa misura la tendenza, in distribuzioni simmetriche, di mostrare code corte che scendono velocemente allontanandosi dal centro della distribuzione oppure code più 'pesanti' che tendono a rimanere significative anche allontanandosi dal centro.

Il punto di riferimento è il modello della gaussiana e in genere si valuta la curtosi proprio andando a comparare il comportamento delle code con quello della gaussiana.

L'indice di **curtosi** δ è definito come segue

$$\delta = \frac{\sum_{i=1}^{n} \frac{(x_i - M)^4}{n}}{sd^2}$$

Anch'esso è disponibile nel pacchetto sopracitato (`moments`) ed è pari a 3 se la curtosi è pari a quella della gaussiana. Se è minore di 3 allora la distribuzione ha code piu corte (rispetto alla gaussiana, leptocurtosi) nel caso contrario più lunghe (platicurtosi).

```
# otteniamo l'indice di curtosi per la variabile CLMAGE
"CLMAGE"
kurtosis(AutoBi$CLMAGE, na.rm=TRUE)
```

```
## [1] "CLMAGE"
## [1] 3.042604
```

Non ha una curtosi molto diversa da quella della gaussiana.

Trasformazione delle variabili

Le variabili quantitative osservate sono di solito il risultato di misurazioni e valori ottenuti in relazione a un preciso sistema di riferimento. Questo implica che si possano effettuare trasformazioni delle variabili per riportarle a un'unità di misura convenzionale diversa. L'esempio che viene subito in mente è la temperatura per cui potrei avere ottenuto le misure in gradi centigradi e volere poi trasformare gli stessi in gradi farenait. In questo caso si tratta di una trasformazione lineare. O ancora si può pensare a esprimere misure monetarie utilizzando una diversa unità (ad es., conversione da dollari a euro).

I altri casi si ricorre a trasformazioni delle variabili per agevolare la rappresentazione grafica o per rendere più agevole il confronto in situazioni diverse. Citiamo alcune trasformazioni il cui uso è consueto proprio per agevolare la rappresentazione grafica o l'analisi di taluni aspetti della variabile

Trasformazioni lineari e standardizzazione

Spesso si procede a operazioni molto semplici come aggiungere o togliere una costante o dividere/moltiplicare i dati per un valore. Si noti che in generale tali operazioni hanno un impatto noto sulle misure di centralità edispersione principali.

In particolare se otteniamo una varabile Z a partire dai valori osservati su Y, ad esempio $Z_i = bX_i + a$ dove a e b sono costanti reali note, si noti che:

- $M_Z = bM_X + a$ dove M_Z e M_X sono rispettivamente la media della variabile trasformata e di quella originale
- $V_Z = b^2 V_X$ dove V_Z e V_X sono rispettivamente la varianza della variabile trasformata e di quella originale.

Se io considero la seguente trasformazione: $Z_i = \frac{Y_i - M_Y}{\sqrt{V_Y}} = \frac{Y_i}{\sqrt{V_Y}} - \frac{M_Y}{\sqrt{V_Y}}$ che è detta standardizzazione si verifica immediatamente che è

- $M_Z = 0$
- $V_Z = 1$.

Questo premette di confrontare varaibili in relazione a altri aspetti (l'asimmetria ad esempio) eliminando l'impatto di media e varianza.

Trasformazioni non lineari

L'uso di trasformazioni non lineari è piuttosto consueto quando una variabile assume valori molto difformi per cui diventa difficile la sua rappresentazione grafica. L'esempio più tipico è quello in cui una variabile ha

code lunghissime e valori estremi. In tal caso è utile ricorrere a trasformazioni come $Z_i = Y_i^{1/k}$ con $k > 1$ (ad esempio la radice quadrata o il cubo) o ancora meglio alla trasformazione logaritmica $Z_i = log(Y_i)$.

Si noti che in tal caso se si conoscesse la media della variabile originale non potremmo ottenere media della variabile originale semplicemnte applicando ad esse la medesima trasformazione (una conseguenza della disuguaglianza di Jensen).

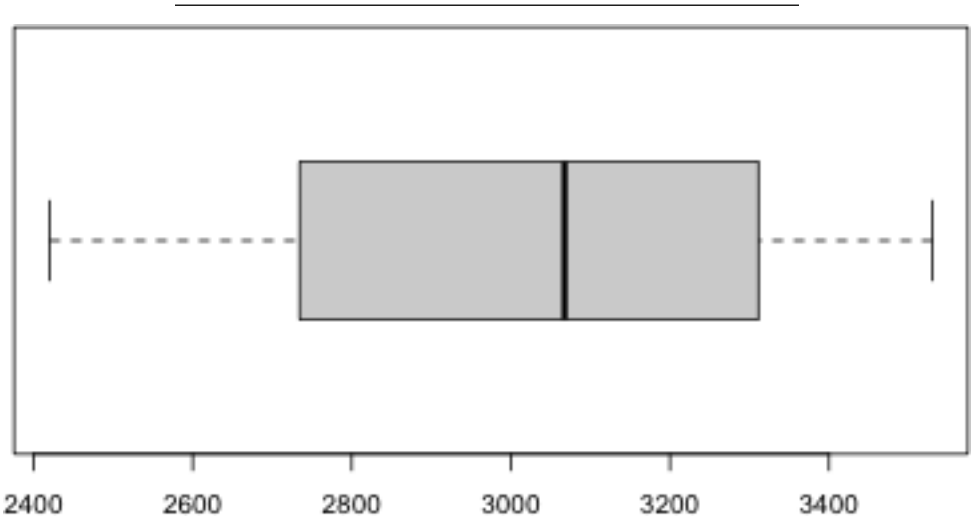
Ancora sulle rappresentazioni grafiche di una variabile quantitativa

Il diagramma a scatola con baffi (boxplot)

Come si è visto il riassunto dei 5 numeri fornisce i 3 quartili che danno idea di diversi aspetti della distribuzione: tendenza centrale (la mediana), dispersione (scarto interquartile), simmetria (indice di Galton). Inoltre il minimo e il massimo danno indicazione di cosa accade sulla coda destra e sinistra.

Un'idea molto semplice ed efficace per riassumere graficamente le caratteristiche essenziali di una variabili quantitativa è quella che conduce al cosiddetto **diagramma a scatola con baffi** (*box and whiskers plot*). Conviene illustrarlo con un esempio utilizzando la funzione di `R` che permette di ottenerlo cioè `boxplot()`.

```
fivenum(neo$Peso)
boxplot(neo$Peso, horizontal=TRUE)
```




```
# Si noti il parametro horizontal che ci consente di mettere la
# scatola verticalamente o orizzontalmente
```

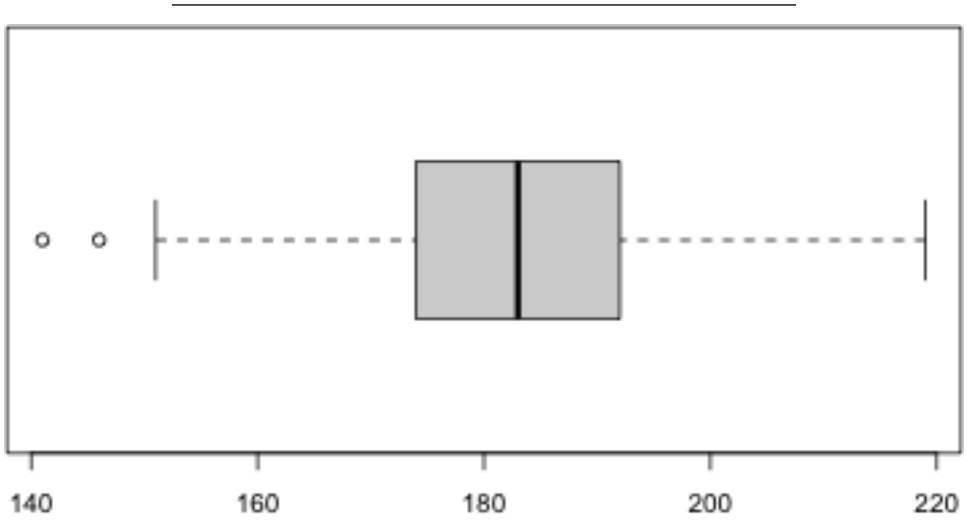
```
## [1] 2420.0 2734.5 3067.5 3311.5 3530.0
```

Come si vede si sono usati i dati forniti dal riassunto dei 5 numeri per disegnare la scatola. La scatola è posizionata con riferimento a un asse lungo cui sono i valori della variabile così che i bordi corrispondano al I e al III quartile (Q1 e Q3) mentre al suo centro la linea è in corrispondenza della mediana (Q2):

1. la larghezza della scatola ci dà quindi un'indicazione della dispersione;
2. la sua posizione ci da una indicazione di dove si trova il 50% dei valori centrali della variabile e assieme alla linea in corrispondenza della mediana ci dà indicazione dei vaolri centrali;
3. la posizione della mediana nella scatola ci informa sulla simmetria;
4. i baffi ci danno indicazione del comportamento sulle code.

Se tuttavia proviamo lo stesso grafico con una diversa variabile

```
fivenum(Cars93$Length)
boxplot(Cars93$Length, horizontal=TRUE)
```




```
## [1] 141 174 183 192 219
```

In questo caso il baffo sinistro non si estende fino al minimo. Inoltre vediamo sul grafico due punti rappresentati separatamente.

Questo perchè in realtà nel diagramma a scatola si evidenziano quei punti che sono ritenuti distanti dal resto dei dati (a destra o a sinistra). Tali valori sono detti *outliers* e vengono rappresentati isolatamente sul grafico.

La definizione di *outlier* nel diagramma a scatola si deriva dalla seguente regola:

- sono *oulier* (valori anomali) quei valori che sono più distanti dai bordi della scatola (cioè dai quartili) più di una volta e mezza la differenza interquartile $SI = Q3 - Q1$. Pertanto tutti i punti che sono superiori a $Q3 + 1.5SI$ o inferiori a $Q1 - 1.5SI$ verranno annotati separatamente sul grafico;
- se essi esistono (a destra o a sinistra) di conseguenza il baffo non va esteso fino al massimo o al minimo valore osservato, va invece esteso:
  - a destra, fino al più grande valore che non sia segnalato come outlier (cioè il massimo valore osservato che risulti inferiore a $Q3 + 1.5SI$);
  - a sinistra, fino al più piccolo valore che non sia segnalato come outlier (cioè il minimo valore osservato che risulti superiore a $Q1 - 1.5SI$).

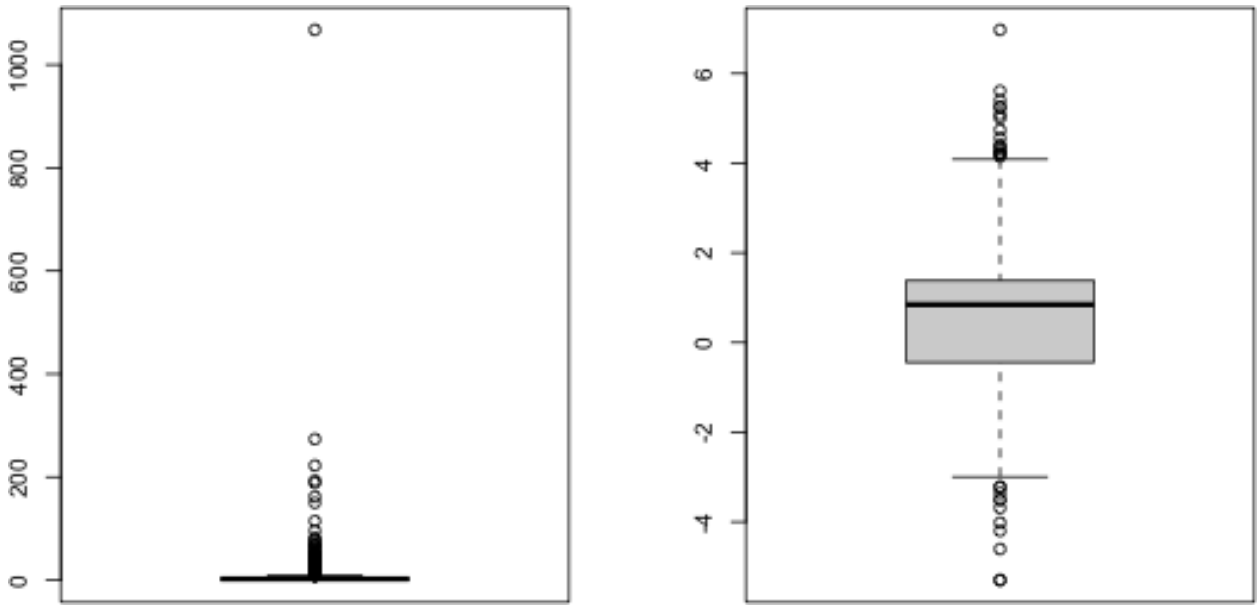
Il diagramma a scatola è una rappresentazione grafica usata molto frequentemente anche perchè è adeguata anche nel caso di un insieme molto limitato di casi (n=20 o 30).

Attenzione però che nel caso i valori siano pochi la regola con cui si determinano i quartili (che ha margini di arbitrarietà come abbiamo già discusso) può cambiare l'apparenza del grafico.

Anche nel caso del diagramma a scatola la presenza di code molto lunghe e/o dati molto estremi esso potrebbe risultare inefficace. Vediamo cosa accade ad esempio con la variabile *AutoBi*LOSS$ che abbiamo visto esser caratterizzata da una coda a destra lunghissima (forte asimmetria positiva) asimmetrica.

```
par(mfrow=c(1,2))
boxplot(AutoBi$LOSS)

# i valori elevatissimi presenti nella variabile vengono rappresentati
# separatamente, ma questo fa sì che la scatola diventi difficile da
# apprezzare sul medesimo grafico.
```

```r
# Proviamo a usare la trasformata logaritmica
boxplot(log(AutoBi$LOSS))
```



La funzione di ripartizione empirica

Una semplice idea per rappresentare graficamente un insieme di dati è quello di ottenere una versione empirica della funzione di ripartizione (o della funzione cumulata) corrispondente quindi alla analoga funzione $F(x)$ definita per una variabile aleatoria X.

Com'è noto per una variabile aleatoria continua X la funzione di ripartizione (anche detta distribuzione cumulata di probabilità) è definita come $F(x) = Pr(X \leq x)$, ovvero fornisce la probabilità che si ottenga un valore aleatorio della X inferiore o uguale a x.

Tale funzione è definita per $x \in \mathbb{R}$

- monotona non decrescente;
- pari a 0 per valori inferiori o uguali al limite inferiore del supporto della variabile X;
- pari a 1 per valori superiori al limite superiore del supporto della variabile X.

Inoltre l'inveresa della funzione di ripartizione fornisce i quantili per cui il quantile $x_p = F^{-1}(p)$. Ovviamente il quantile per ogni $0 \leq p \leq 1$ esiste se la funzione di ripartizione non ha discontinuità.

L'equivalente empirico di tale funzione, denotato con $\hat{F}(x)$ avendo osservato l'insieme di dati $x_1, x_2, \ldots, x_n$ per una variabile quantitativa si ottiene cercando il valore che $\forall x \in \mathbb{R}$ fornisce la proporzione di unità inferiori o pari a x, ovvero

$$\hat{F}(x) = \text{proporzione}(x_i \leq x) = \frac{\text{numero di valori} \leq x}{n}$$

Riprendendo la notazione introdotta per i dati ordinati $x_{(1)} \leq x_{(2)} \leq \ldots, \leq x_{(n-1)} \leq x_{(n)}$ la funzione di ripartizione empirica (cumulata empirica) sarà pari a:

- $\hat{F}(x) = 0$ se $x \leq x_{(1)}$;
- in $x = x_{(1)}$ la funzione fa un salto ed è pari a $\frac{1}{n}$ e rimane costante fino al prossimo valore $x_{(2)}$;
- in $x = x_{(2)}$ la funzione fa un ulteriore salto ancora di altezza $\frac{1}{n}$;
- e così via fino a $x_{(n)}$ partire dal quale la funzione assume il valore 1;

- se vi sono valori ripetuti il salto sarà pari a $\frac{m}{n}$ ove m è il numero di valori ripetuti.

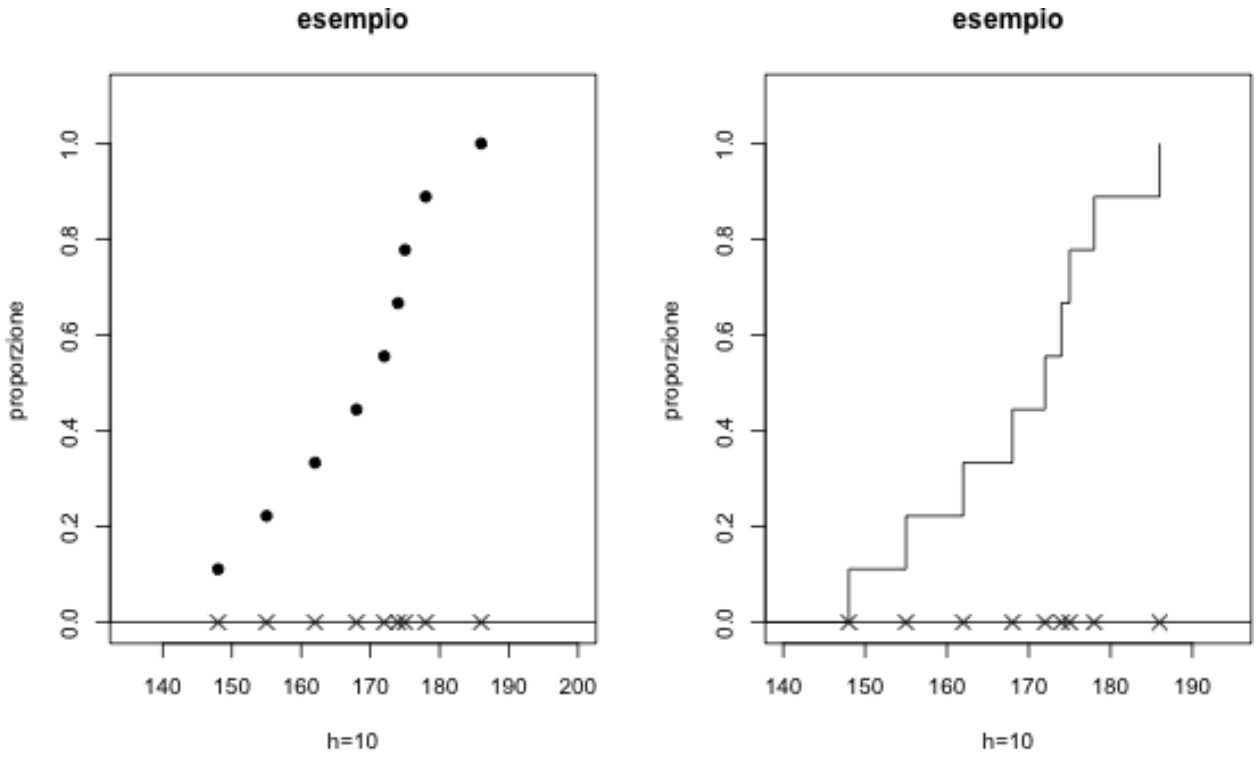
Si tratta quindi di una funzione a scalini definita come:

$$\hat{F}(x) = \begin{cases} 0 & \text{se} \quad x < x_{(1)} \\ \frac{i}{n} & \text{se} \quad x_{(i)} \leq x < x_{(i+1)} \\ 1 & \text{se} \quad x \geq x_{(n)} \end{cases} \qquad \text{for} i = 1, 2, \ldots, n-1$$

.

'E piuttosto agevole tracciare tale grafico in R. Vediamo un semplice esempio.

```
par(mfrow=c(1,2))
# consideriamo l'insieme di 9 dati nel vettore xx
xx<-c(148, 155, 162, 168, 172, 174, 175, 178, 186)

# rappresentiamo i punti su un grafico
# i parametri ylim e xlim permettono di definire
# i limiti per le coordinate x e y, pch permette di scegliere il simbolo
# (4 è il simbolo X), cex le dimensioni del simbolo.
nn<-length(xx)
plot(xx,rep(0,nn), ylim=c(0,1.1), xlim=c(135,200), pch=4, cex=1.5,
     main="esempio", xlab="h=10", ylab="proporzione")
abline(0,0)
# abline(a,b) traccia una retta con intercetta e coefficiente angolare
# con a e b assegnati
# points() permette di aggiungere al grafico esistente nuovi punti
points(c(0,xx),c(0,1:9)/nn, pch=19)
# e il parametro type="s" consente di ottenere una funzione a gradini
plot(xx,rep(0,nn), ylim=c(0,1.1), xlim=c(140,195), pch=4, cex=1.5,
     main="esempio", xlab="h=10", ylab="proporzione")
abline(0,0)
points(c(0,xx),c(0,1:9)/nn, type="s")
```

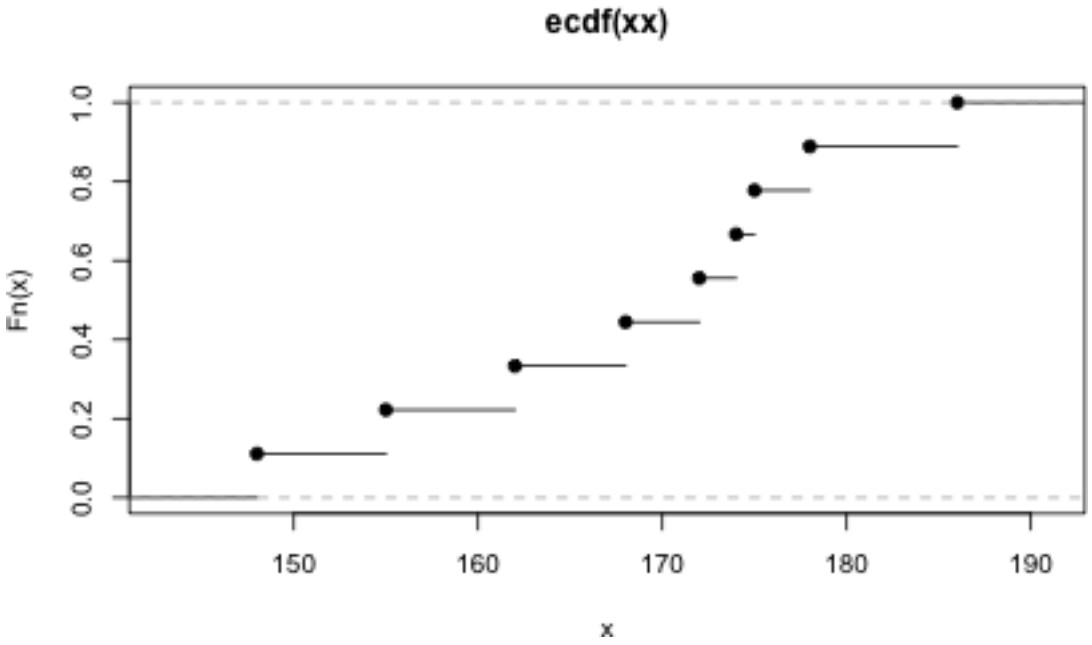



```
par(mfrow=c(1,1))
```

Come si vede la funzione ha incrementi pari a $1/n$ ogni volta che si è in corrispondenza di un nuovo dato osservato. La linea verticale nel gradino è graficamente utile ad apprezzare l'altezza dello stesso ma dal punto di vista formale la funzione dovrebbe presentarsi come una funzione costante a tratti.

Esiste la funzione `ecdf` predefinita in `R` per ottenere la funzione di ripartizione empirica. Essa consente numerose varianti e genera un oggetto che può essere direttamente fornito alla funzione `plot`

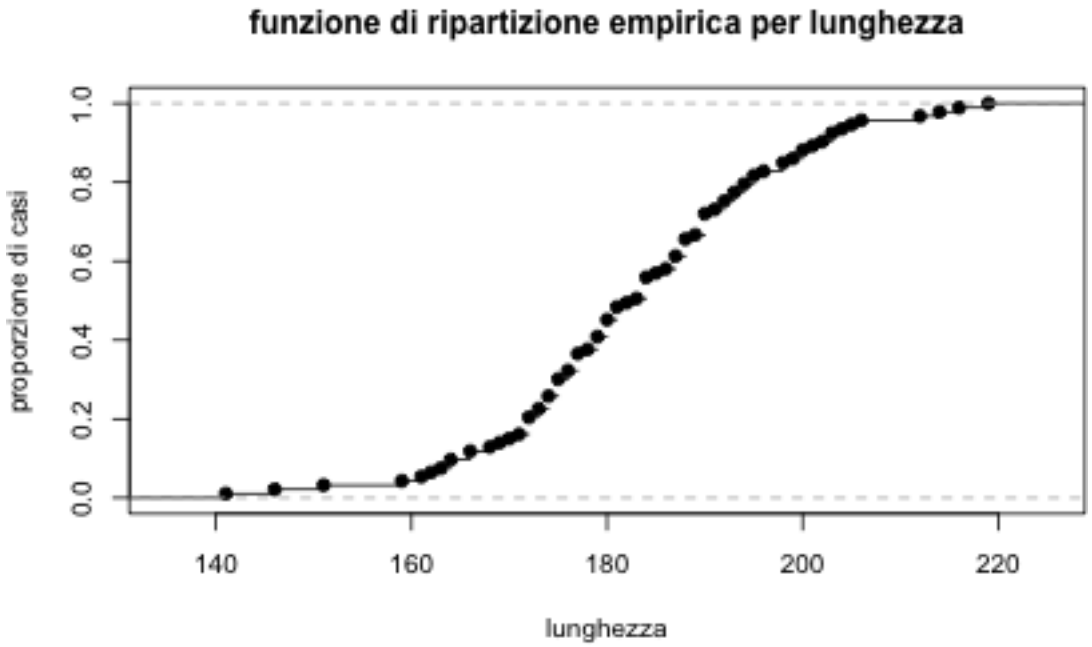
```
mio<-ecdf(xx)
plot(mio)
```



La funzione `ecdf()` è basata sulla funzione `stepfun()` che è appositamente costruita per trattare funzioni a gradini cui si rinvia per dettagli.

Proviamo a ottenere ora la funzione di ripartizione empirica per i dati sulla lunghezza delle auto.

```
plot(ecdf(Cars93$Length), main="funzione di ripartizione empirica per lunghezza",
   xlab="lunghezza", ylab="proporzione di casi")
```



Tale rappresentazione grafica è di interesse per vari motivi:

- conserva l'informazione su ogni singolo valore;
- è del tutto adeguata anche con variabili quantitative discrete;
- permette di visualizzare dove si trovi la mediana o alcuni percentili (la funzione dei quantili è in effetti

l'inversa della funzione di ripartizione). Da questo si intende chiaramente come solo alcuni quantili sono definiti, mentre per altri occorre ricorrere ad approssimazioni (ad esempio considerando interpolazioni lineari fra i due quantili con l'ordine più prossimo);
- la sua interpretazione non è agevole e immediata, ma è evidente che i punti in cui la curva cresce più veleocemente sono quelli in cui si sono osservati molti casi;
- si potrebbe ricorrere a definizioni alternative (ad esempio considerando numero di $x_i < x$). Ottenendo una curva che sarà leggermente spostata rispetto a quella già definita ma con la stessa forma;
- a partire da essa si potrebbe costruire una versione empirica della funzione di sopravvivenza $S(x) = 1 - F(x)$. A questa funzione occorrerebbe dedicare un capitolo a parte vista la sua rilevanza per l'analisi di dati di durata soprattutto in ambito medico (relativi, ad esempio, all'analisi dei tempi impiegati per guarire da una malattia). In effetti esistono in `R` pacchetti dedicati a questo;
- è estremamente vantaggiosa se si vuole confrontare la distribuzione empirica con una distribuzione teorica (aspetto che verrà trattato più avanti).

Resta evidente che una più accurata rappresentazione della distribuzione dei dati si avrebbe considerando piuttosto che la versione empirica della loro funzione di ripartizione la versione empirica delle funzione di densità (che, ricordiamo, in distribuzioni di probabilità per variabili aleatorie continue sarebbe la derivata della funzione di ripartizione).

I modi per rappresentare empiricamente la funzione di densità empirica per un insieme di dati saranno oggetto dei prossimi paragrafi.

L'istogramma

L'istogramma è senz'altro una delle più diffuse e conosciute tecniche di rappresentazione grafica di un carattere quantitativo, i principi che ne ispirano la costruzione e l'interpretazione sono semplici e sono descritti in un qualsiasi testo di statistica di base.

L'istogramma in realtà costituisce uan forma semplice per ottenere una approssimazione empirica della funzione di densità ed è dall'estensione di tale idea che si possono poi considerare tecniche più complesse per la determinazione della curva di densità e per ottenere una versione "liscia" dell'istogramma.

Si suppone di disporre di un insieme di n dati $x_1, x_2, \ldots, x_n$ relativi ad una variabile quantitativa X.

Abbiamo già visto come sia possibile sintetizzare i dati di una variabile quantitativa in una tabella di frequenza: la variabile X è opportunamente categorizzata e trasformata in un fattore mediante un operazione di raggruppamento in classi; si determina cioè una sequenza di valori $z_0 < z_1 < \ldots < z_{I-1} < z_I$ che definiscono una successione di intervalli disgiunti (classi) $(z_{i-1}, z_i]$ (con $i = 1, 2, \ldots, I$) e si determinano le frequenze relative

$$f_i = \frac{\text{numero di valori nell'intervallo}(z_{i-1}, z_i]}{n}.$$

La rappresentazione grafica mediante istogramma della distribuzione di frequenze relative riassunta dalle I coppie $((z_{i-1}, z_i], f_i)$ si effettua costruendo dei rettangoli la cui base coincida con gli intervalli $(z_{i-1}, z_i]$ e la cui altezza sia tale da far corrispondere le frequenze relative f_i con le aree dei rettangoli.

Tale risultato si ottiene introducendo il concetto di **densità di frequenza relativa**. L'altezza di ogni rettangolo è quindi determinata in modo tale da consentire una corrispondenza fra l'area del rettangolo A_i che insiste su $(z_{i-1}, z_i]$ e la frequenza relativa f_i: ovvero $f_i \propto A_i$. Quindi l'altezza del rettangolo che insiste sull'intervallo $(z_{i-1}, z_i]$ deve essere proporzionale a

$$\frac{f_i}{z_i - z_{i-1}}. \tag{1}$$

Di solito si calcola l'altezza così che l'area di ogni rettangolo sia pari esattamente alla frequenza relativa, così che $A_i = f_i$, e in tal caso l'area complessiva all'interno dei rettangoli che compongono il grafico risulterà pari ad 1.

Le altezze dei rettangoli sono quindi poste pari alla frequenza relativa nell'intervallo diviso per l'ampiezza dell'intervallo stesso. Esse **non** rappresentano quindi le frequenze relative per gli intervalli in questione bensì la **densità di frequenze relative** (salvo nel caso in cui le basi dei i rettangoli abbiano tutti ampiezza pari a 1), ovvero la quota di frequenze relative che insiste su un generico intervallo unitario in $(z_{i-1}, z_i]$.

L'istogramma quindi rappresenta un primo, semplice, tentativo di ottenere una versione empirica della funzione di densità sottostante ai dati. Quindi

$$\hat{f}(x) \geq 0$$

$$\text{prop}(a \leq \text{units} \leq b) = \int_a^b f(x)dx$$

In linea di principio possiamo calcolare il valore approssimativo della proporzione di casi fra due numeri reali a e b sommando le aree dei rettangoli rappresentate dall'istogramma e compresi nell'intervallo (considerando porzioni dei rettangoli se a o b non coincidono con gli estremi delle classi).

Come detto nel fare la suddivisione in classi si sono perse informazioni e quindi la vera proporzione di casi (che potrei ottenere solo se tornassi alle informazioni originali rinunciando al riassunto grafico) nei dati sarà verosimilmente diversa.

L'assunzione sottostante alla rappresentazione grafica è la costanza della densità nell'intervallo, ovvero su ogni sottointervallo Δx interno a una classe di valori la frequenza relativa dipende esclusivamente dall'ampiezza Δx (uniforme distribuzione nelle classi).

L'istogramma presenta caratteristiche che lo rendono particolarmente utile in molte situazioni applicative, ma ha anche alcuni inconvenienti.

I principali vantaggi sono:

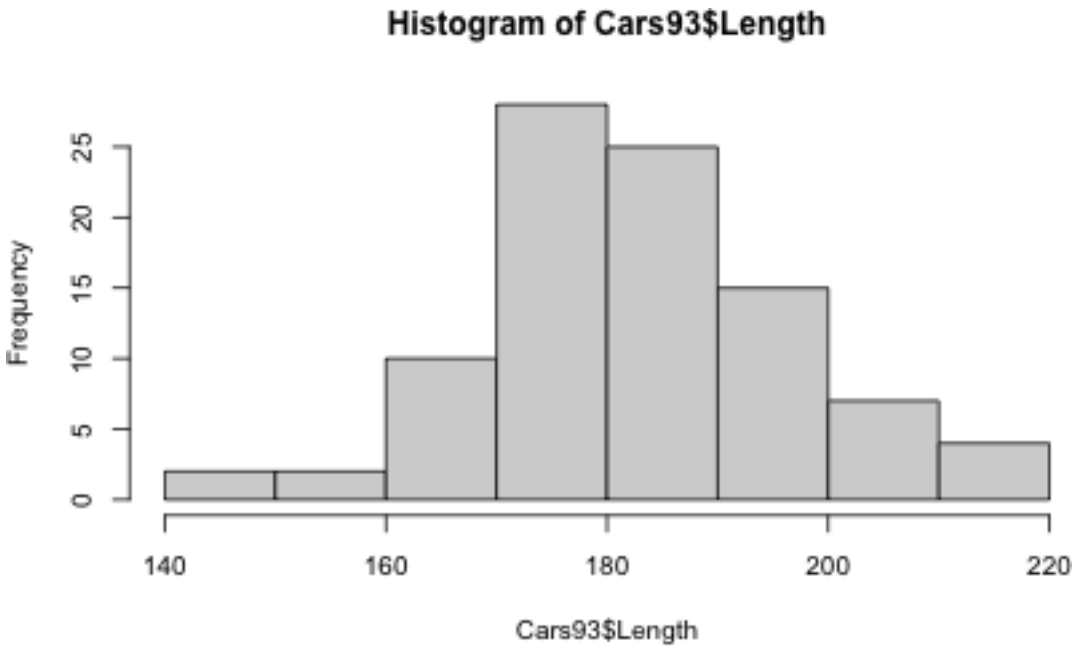
- è semplice da interpretare;
- è immediato da costruire;
- è facile da utilizzare.

I suoi limiti sono invece i seguenti:

- il grafico presenta forti discontinuità. In particolare si potrebbe dedurre che vi è un salto fra la densità in due valori di x molto prossimi ma che appartengono a due diverse classi (è ovvio che tale conclusione potrebbe essere diversa per una scelta alternativa delle classi) e talvolta (con pochi dati) il grafico potrebbe anche essere sensibilmente diverso per scelte alternative delle classi: in genere la scelta di classi più o meno ampie equivale alla scelta di un diverso grado di lisciamento e di regolarità della rappresentazione grafica;
- la densità è assunta costante in ogni intervallo e tale assunto è di solito discutibile.

Ottenere un istogramma in `R` è semplice in quanto la funzione `hist()` automaticamente, determina le classi, calcola le frequenze nelle classi, e produce il grafico. ad esempio:

```
hist(Cars93$Length)
```



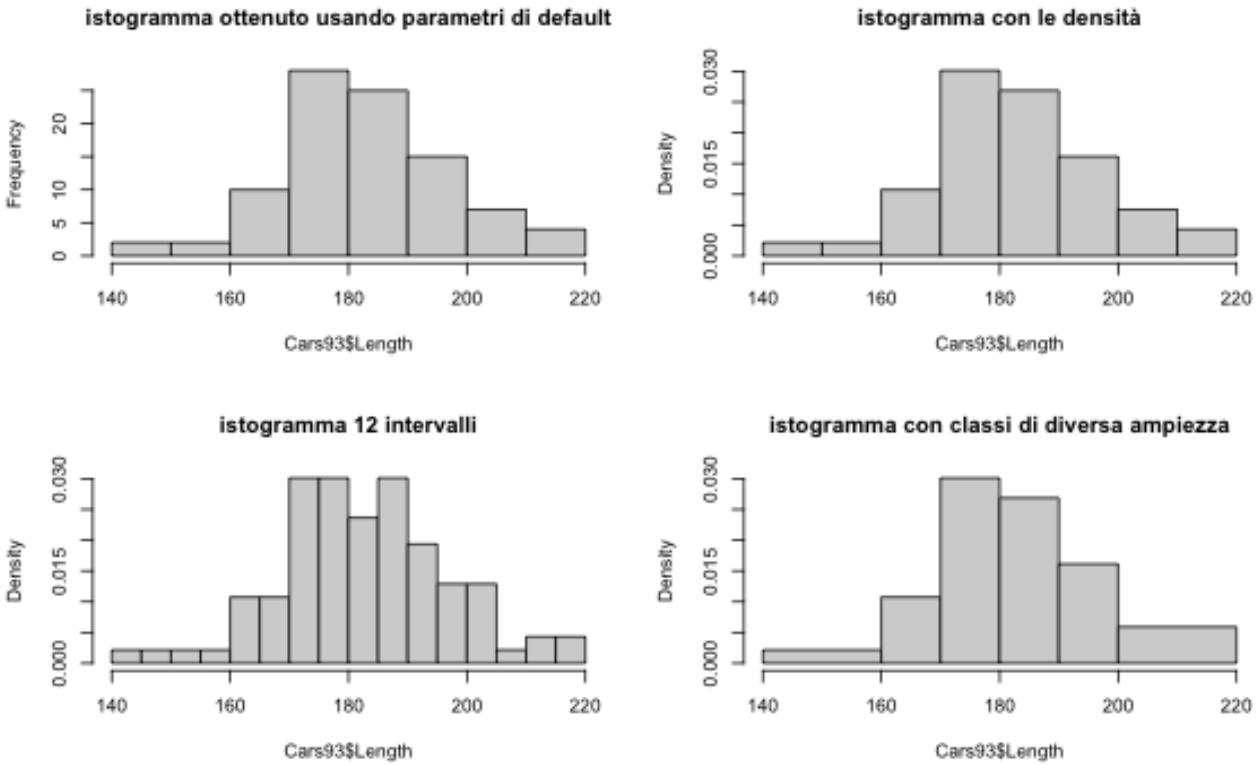
Come si vede il grafico presenta classi tutte della medesima ampiezza e presenta sull'asse i valori relativi alle frequenze assolute. Il grafico è in questo caso non rappresenta le densità, tuttavia la forma resta invariata in quanto le ampiezze delle classi sono uguali e le altezze dei rettangoli sono proporzionali alle densità.

Se si vuole però un grafico di densità, occorre specificare il parametro `prob=TRUE`

Il numero delle classi M che vengono usate in `R` è determinato con una regola detta regola di Sturges per cui $M = 1 + \frac{log(n)}{log(2)}$. Quindi cresce (lentamente) con la dimensione del vettore di dati. E' possibile usare regole diverse (si veda l'help della funzione `hist`).

Ovviamente i parametri della funzione ci consentono di variare tale scelta e anche di usare classi di ampiezza diversa. Il parametro `breaks=` consente di fare entrambe le cose. Gli esempi sotto che illustrano ancora quanto detto sopra: si noti come aumentando il numero delle classi si rischi di avre un grafico meno regolare troppo dipendente dalla specificita dell'insieme di dati.

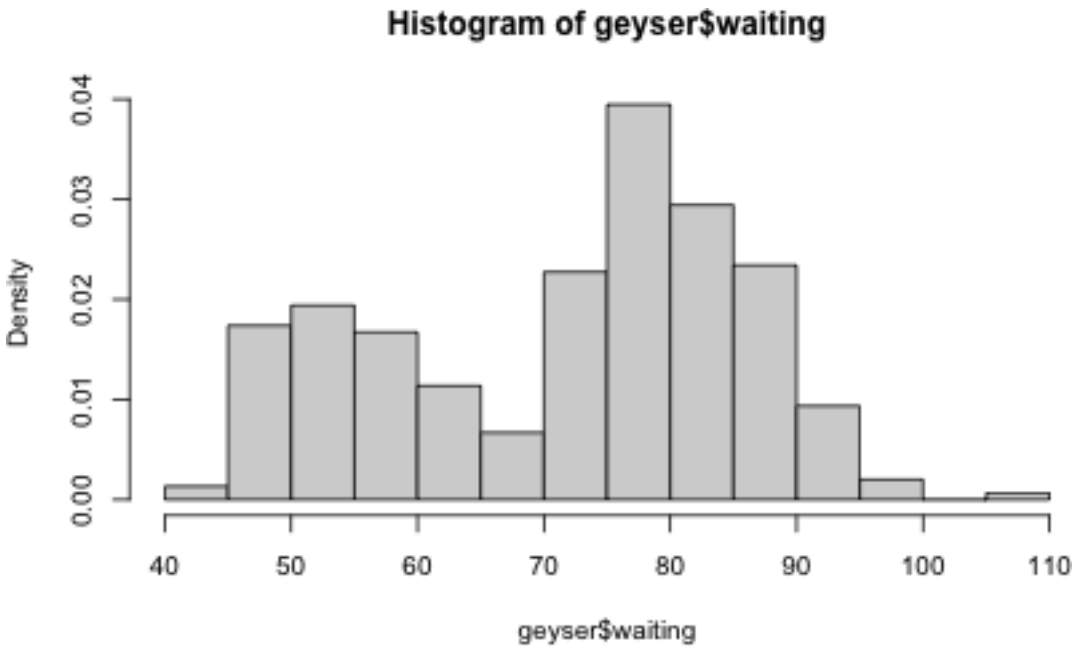
```
par(mfrow=c(2,2))
hist(Cars93$Length, main="istogramma ottenuto usando parametri di default")
hist(Cars93$Length, prob=TRUE, main="istogramma con le densità")
hist(Cars93$Length, prob=TRUE, breaks=12, main="istogramma 12 intervalli")
hist(Cars93$Length, prob=TRUE, breaks=c(140,160,170,180,190,200,220),
     main="istogramma con classi di diversa ampiezza")
```




```
par(mfrow=c(1,1))
```

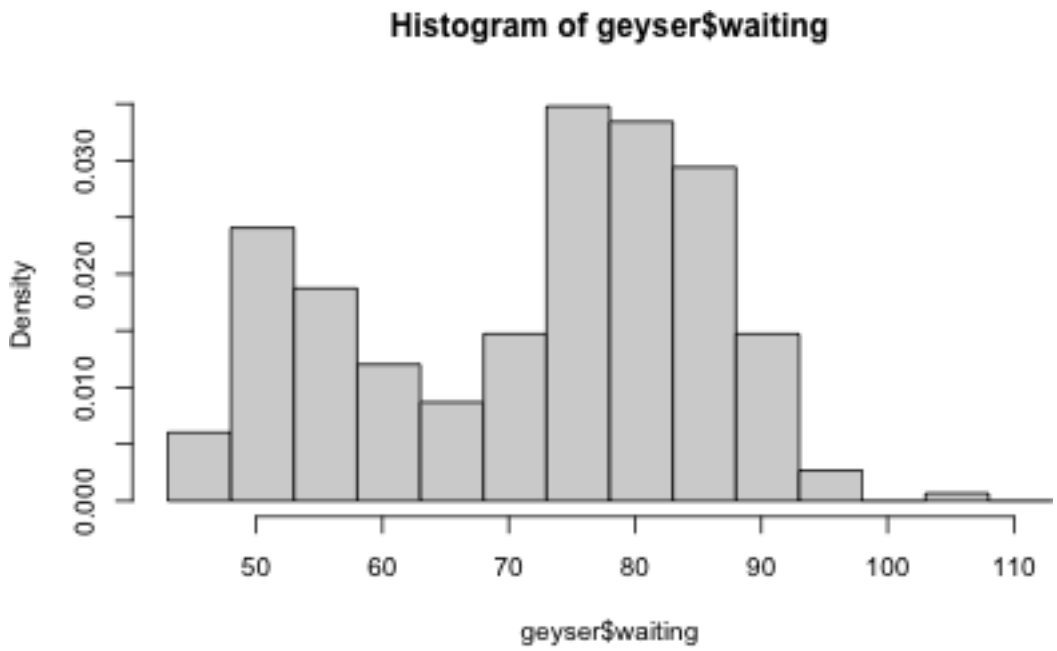
Per illustrare ancora quanto la rappresentazione con istogramma possa fornire grafici diversi variando alcuni criteri di costruzione, si considerino i dati `geyser` presenti nel package `MASS`. Essi si riferiscono ai tempi di attesa fra due eruzioni del geyser *Old faithful* nel parco di Yelloswstone negli USA. Disegniamo l'istogramma

```
library(MASS)
data(geyser)
hist(geyser$waiting, prob=TRUE)
```



Si consideri ora una diversa suddivisione in classi mantenendo la stessa ampiezza per la variabile cambiando semplicemente l'origine della prima classe.

```
library(MASS)
data(geyser)
hist(geyser$waiting, breaks=c(43,(5*(1:14)+43)), prob=TRUE)
```

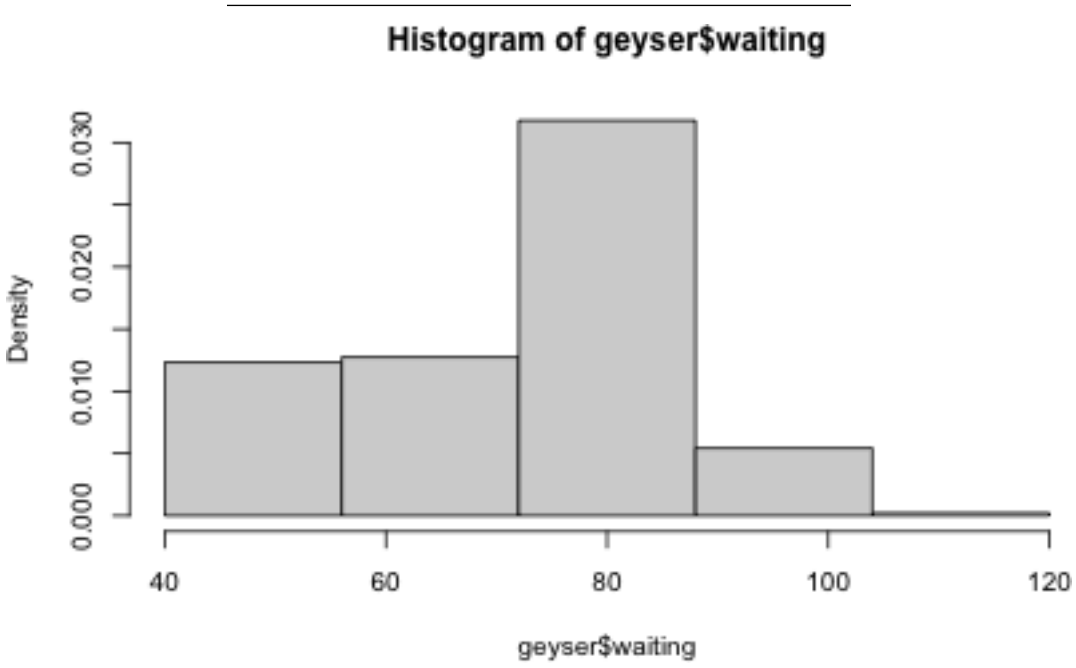


L'ampiezza delle classi è la stessa nei due casi, il punto di origine delle classi è invece spostato. Si noti che pur essendo le caratteristiche del grafico simili (entrambe le rappresentazioni grafiche consentono di cogliere le due gobbe della distribuzione), la collocazione dei due massimi nelle due gobbe e la distanza che li separa appare diversa. Se, inoltre si guarda al primo grafico si deduce che la densità di frequenza nel punto $x = 72$ è la stessa che nel punto $x = 74$, mentre si arriverebbe a ben altra conclusione riguardo la densità locale nei

due punti se si guarda al secondo grafico.

Se inoltre utilizzassi solo 4 classi perderei la caratteristica saliente di questi dati che è data dalla presenza delle due gobbe

```
library(MASS)
data(geyser)
hist(geyser$waiting, breaks=c(40,56,72,88,104,120), prob=TRUE)
```



La moda

Nel caso di un modello probabilistico per una variabile aleatoria continua X con densità $f(x)$ è possibile definire la **moda Mo** come

$$Mo = \arg\max_{x \in R} f(x)$$

ovvero il valore cui corrisponde la densità più alta.

Ovviamente essa risulta definita anche per una variabile aleatoria discreta, limitandosi in questo caso alla ricerca del valore di probabilità più elevato limitatamente ai valori del supporto della variabile aleatoria.

La versione empirica della moda è quindi definita in analogia per cui:

- nel caso di una variabile categoriale (un fattore) essa è definita come la modalità cui corrisponde la frequenza relativa più elevata;
- nel caso di una variabile quantitativa discreta essa è definita come il valore cui corrisponde la frequenza massima;
- nel caso di una variabile quantitativa continua essa si può ottenere dopo la trasformazione della variabile in fattore utilizzando il raggruppamento in classi. In tal caso si determina la ***classe modale*** che è quella classe cui corrisponde la densità di frequenza più elevata (non si guarda quindi in questo caso alla semplice frequenza nella classe).

La moda è un indice di tendenza centrale che viene spesso impiegato e peraltro è *l'unico indice di tendenza centrale* che si può determinare per una variabile categoriale.

Vale la pena di notare che, in particolare nel caso di variabili quantitative continue, essendo la moda in corrispondenza di un punto di massimo assoluto per la funzione di densità, si possono talvolta reperire altri valori però corrispondenti a punti di massimo relativo (mode secondarie). Se una distribuzione oltre alla moda principale (quella per cui si osserva il massimo assoluto) mostra una o più mode secondarie si parla di distribuzione **multimodale** (**bimodale** se i punti di massimo sono due).

Si noti che in distribuzioni unimodali la posizione della moda rispetto alla mediana e alla media fornisce indicazioni sulla asimmetria della distribuzione. In distribuzioni simmetriche i tre indici sono circa nella stessa posizione, mentre se c'è asimmetria positiva (coda destra lunga) si ha $Mo < Me < M$. L'ordinamento si inverte nel caso di asimmetria negativa.

Nel caso dei dati del geyser visti sopra la distribuzione presenta una evidente bimodalità.

Il metodo del nucleo per il "lisciamento" di una curva di densità

A partire dalla definizione di densità di frequenza introdotta per l'istogramma è possibile ottenere la densità di frequenza relativa locale in un qualsiasi punto x come

$$d(x) = \frac{numero\ di\ valori\ in\ (x - \frac{h}{2}; x + \frac{h}{2}]}{hn} \tag{2}$$

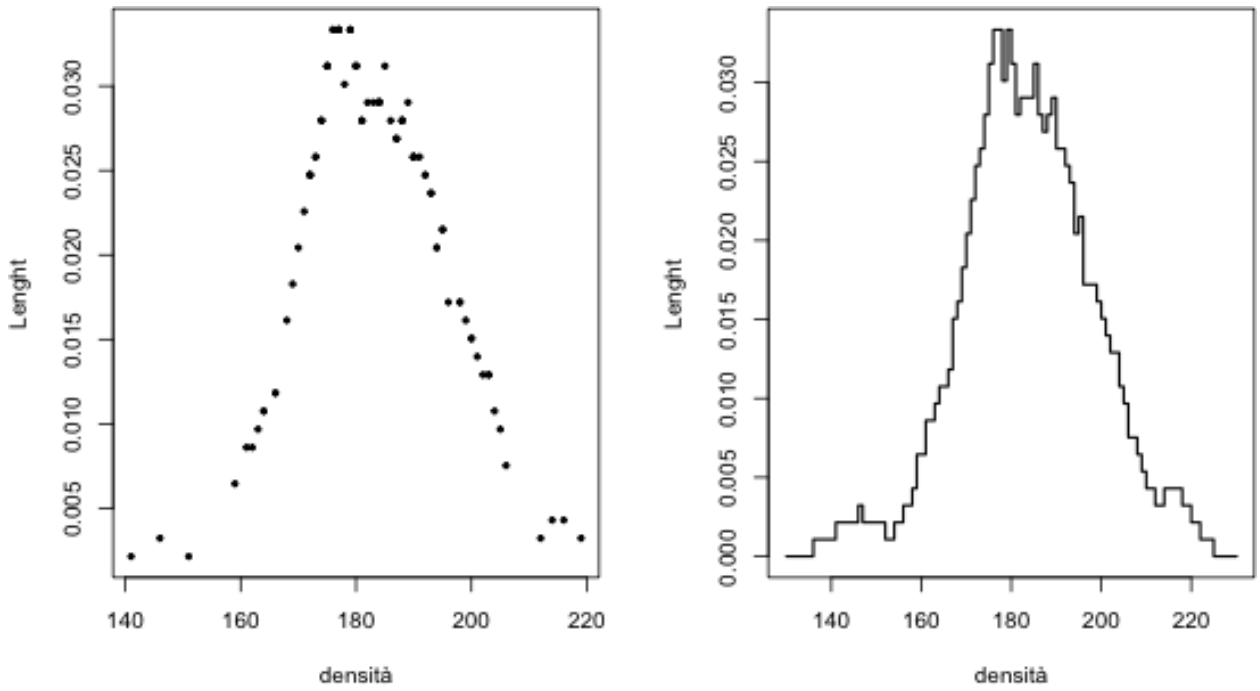
'E come se si prendesse una classe di ampiezza h e la si muovesse ponendola in corrispondenza di ciascun valore osservato

Si potrebbe quindi tracciare il grafico della curva $d(x)$, osservando che essa risulterà avere delle discontinuità nei punti $x_i \pm \frac{h}{2}$ e sarà costante nei tratti intermedi. L'inconveniente legato alle discontinuità presenti nel grafico, peraltro meno rilevanti se n è grande, può essere tuttavia superato come si vedrà nel successivo paragrafo.

La funzione $d(x)$ quindi associa ad ogni valore x la densità di frequenze relative misurata con riferimento ad un intervallo di ampiezza h centrato su x. La curva di densità che descrive tale funzione è uno strumento utile per l'analisi di un insieme di dati quantitativi: come per l'istogramma, esso fornisce infatti una rappresentazione grafica di immediata comprensione. Dove la densità di frequenza è elevata il numero di unità che insiste sull'intervallo di ampiezza h centrato su x è maggiore che in punti ove la densità risulta minore. Si ha in definitiva uno sguardo sull'intera distribuzione di valori senza tuttavia dover imporre come per l'istogramma una particolare scelta di classi di valori.

Per esempio, usando tale criterio (con h=10) per i dati sulla lunghezza delle auto si otterrebbe:

```
par(mfrow=c(1,2))
# versione della densità ottenuta applicando la definizione in (2)

nucl<-function(x,xx,h=1) {
  y=0
  n=length(xx)
  nucl=0
  for (i in 1:n){
    nucl=nucl+(abs(x-xx[i]) <= h/2)*1/(n*h)}
  nucl
}
plot(sort(Cars93$Length),nucl(sort(Cars93$Length),Cars93$Length,10),
     cex=.5, pch=19, ylab="Lenght", xlab="densità")

# versione del grafico in cui il calcolo viene fatto per tutti
# i punti di x e non solo in quelli osservati

plot(seq(130,230),nucl(seq(130,230),Cars93$Length,10), lwd=1.5,
     type="s", ylab="Lenght", xlab="densità")
```




```
par(mfrow=c(1,1))
```

I valori di densità che descrive tale funzione forniscono uno strumento utile per l'analisi di un insieme di dati quantitativi: come per l'istogramma, esso fornisce infatti una rappresentazione grafica di immediata comprensione. Dove la densità di frequenza è elevata il numero di unità che insiste sull'intervallo di ampiezza h centrato su x è maggiore che in punti ove la densità risulta minore. Si ha in definitiva uno sguardo sull'intera distribuzione di valori senza tuttavia dover imporre come per l'istogramma una particolare scelta di classi di valori.

Si noti che tracciare un istogramma equivale alla determinazione della densità secondo la definizione introdotta nella 2 sopra solo per i punti x al centro degli intervalli $x = \frac{z_{z-i}+z_i}{2}$. La particolarità è che tale valore della densità viene estesa a tutti i valori $x \in (z_{i-1}; z_i]$. L'istogramma è quindi una soluzione molto particolare e non del tutto efficiente del problema di determinare la funzione di densità di frequenza $d(x)$.

Determinazione di una funzione di densità empirica con il metodo del nucleo

Al fine di superare alcuni degli inconvenienti già citati nel caso dell'istogramma è quindi possibile approssimare la curva di densità semplicemente applicando la definizione 2.

La 2 può essere opportunamente riscritta come segue. Si definisca

$$W(x) = \begin{cases} 1 & \text{se} \quad |x| < \frac{1}{2} \\ 0 & \text{altrimenti} \end{cases}$$

allora

$$d(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} W\left(\frac{x - x_i}{h}\right) \tag{3}$$

La 3 è equivalente alla 2 ma, come vedremo, rende naturale l'estensione e la generalizzazione di quest'ultima.

Come già notato la curva che si ottiene con la 2 è discontinua nei punti $x_i \pm \frac{h}{2}$, ove x_i è un generico valore corrispondente ad uno dei dati osservati, anche se l'entità delle discontinuità è ridotta rispetto a quello che accade con l'istogramma. Ciò è dovuto alla discontinuità della funzione W.

In realtà la formulazione in 3 equivale ad assumere che l'impatto sulla curva di densità di un qualsiasi valore x_i sia distribuito in ugual misura con peso $\frac{1}{h}$ su tutto l'intervallo $(x - \frac{h}{2}; x + \frac{h}{2}]$, cosicchè comunque il peso complessivo di ogni singola osservazione x_i sia pari a $\frac{1}{n}$.

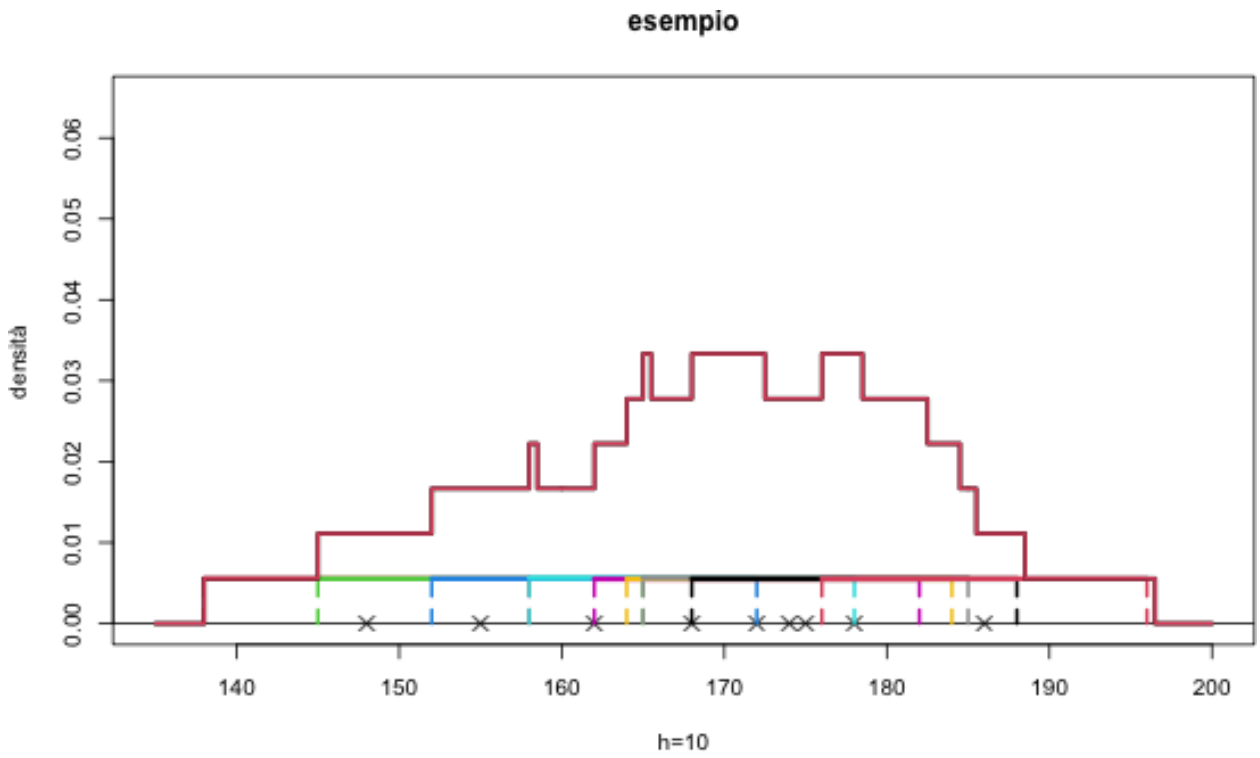
La funzione $W(u)$ è un rettangolo di ampiezza h ed il calcolo nella 3 equivale a posizionare un rettangolo di ampiezza h in corrispondenza di ogni valore osservato x, a valutare la frequenza relativa dei casi che sono all'interno dell'intervallo e a determinare la densità di frequenze relative in x come il rapporto fra la frequenza relativa ottenuta e l'ampiezza dell'intervallo stesso.

Una interpretazione equivalente si ha immaginando di posizionare una scatola di ampiezza h e altezza $\frac{1}{n}$ in corrispondenza di ogni dato osservato x_i, in modo che il centro della scatola coincida con x_i. Il valore di $d(x)$ è pari alla somma delle altezze di tutte le scatole che comprendono il valore x diviso per n.

Illustriamo quanto detto nel seguente esempio

```
par(mfrow=c(1,1))
# consideriamo ancora l'insieme di 9 dati nel vettore xx
xx<-c(148, 155, 162, 168, 172, 174, 175, 178, 186)

plot(xx,rep(0,9), ylim=c(0,0.065), xlim=c(135,200), pch=4, cex=1.5,
     main="esempio", xlab="h=10", ylab="densità")
abline(0,0)

# infine rappresentiamo le funzioni W, ovvero dei rettangoli centrati su ciascun dato
yc=0
for (i in 1:9){
yc[i]<- jitter(1/180) #la funzione jitter() consente di spostare un dato
                     #di una piccola quantità, è utile per rappresentare
                    # dati che apparirebbero altrimenti sovrapposti
segments(xx[i]-10,yc[i],xx[i]+10,yc[i], col=(i+1), lwd=3) # la funzione segments
                                  # aggiunge una linea su un grafico fra
segments(xx[i]-10,0,xx[i]-10,yc[i], col=(i+1),lwd=2, lty=2)
segments(xx[i]+10,0,xx[i]+10,yc[i], col=(i+1),lwd=2, lty=2)}

# Successivamente costruiamo una funzione che calcoli per ogni x
# la somma delle ordinate delle funzioni W(x) (la somma delle altezze
# dei rettangoli di ampiezza h)

rett<-function(x,xx,h=1) {
  n=length(xx)
  rett=0
  for (i in 1:n){
    rett=rett+(abs(x-xx[i]) <= h/2)*1/(n*h)}
  rett
}

# ora sovrapponiamo al grafico: la funzione points() permette di aggiungere
# punti su un grafico esistente
xr<-seq(135,200,.5)
points(xr,rett(xr,xx,20), lwd=2.5, type="s")

# aggiungiamo un altro esempio con rettangoli più ampi
```

```
points(xr,rett(xr,xx,20), lwd=2, type="s", col=2,
  main="esempio con rettangoli più ampi", ylab="h=20", xlab="densità")
```



Una naturale estensione della 3 si ha se si assume che ogni singolo dato osservato x_i abbia un impatto sulla densità in x decrescente con la distanza $x - x_i$. Ciò implica che ad esempio si possa usare una funzione per $W(u)$ diversa dal rettangolo.

In generale possiamo introdurre una funzione $K(u)$, detta **nucleo**, che abbia le seguenti caratteristiche

- $\int_{\infty}^{-\infty} K(u)du = 1$,
- $K(u)$ è una funzione simmetrica rispetto a 0,
- $K(u)$ assume solo valori positivi.

E la funzione $d(x)$ avendo osservato un insieme di dati viene determinata come

$$d(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{4}$$

Si noti che la funzione $W(u)$ definita precedentemente ha le caratteristiche di una funzione nucleo. Altre scelte ragionevoli per $K(u)$ sono, ad esempio le seguenti:

$$K(u) = \begin{cases} \frac{3}{4}(1 - \frac{1}{5}u^2)5^{-\frac{1}{2}} & \text{se|u|<5} \\ 0 & \text{altrimenti} \end{cases} \tag{5}$$

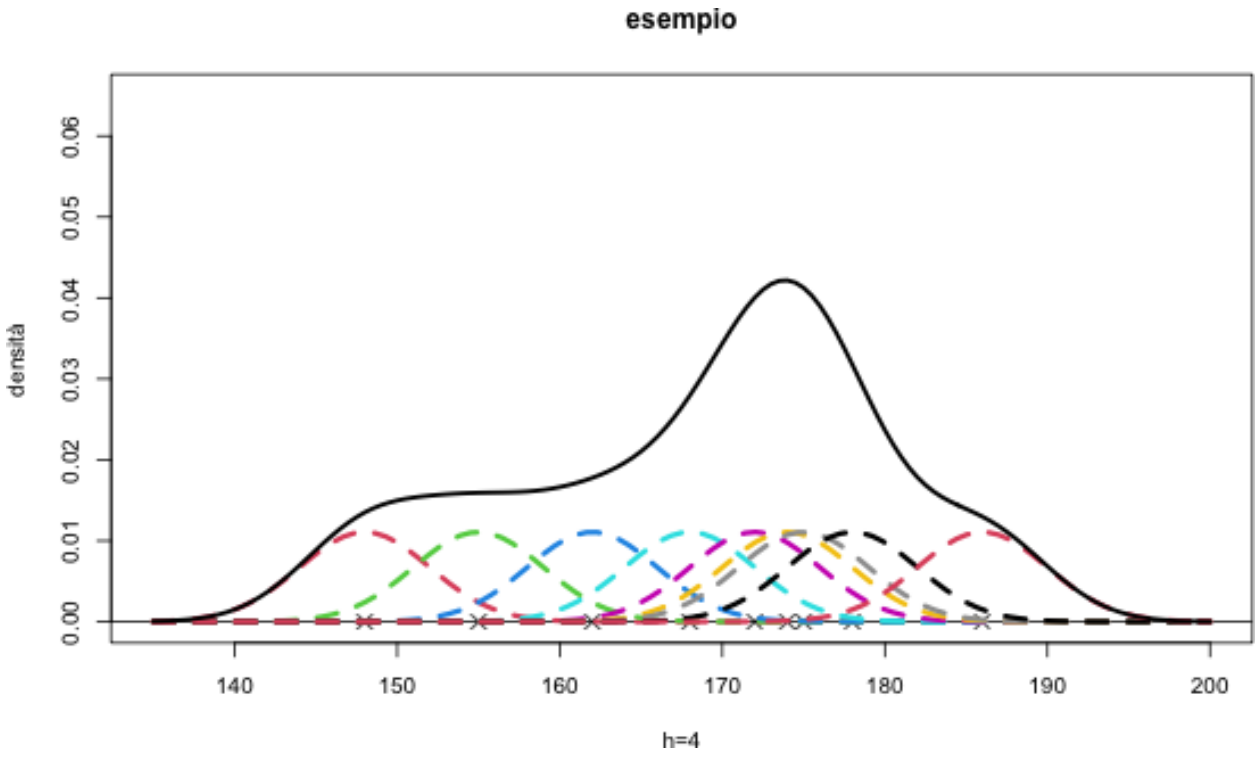
$$K(u) = \begin{cases} 1 + cos2\pi & \text{se|u|< 0.5} \\ 0 & \text{altrimenti} \end{cases} \tag{6}$$

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}. \tag{7}$$

L'ultima funzione è facilmente identificabile come una funzione di densità di una normale standard (nucleo gaussiano). L'interpretazione della 4 data per la funzione $W(u)$ (detta nucleo rettangolare) è valida anche in questo caso, basti pensare che in corrispondenza di ogni osservazione x_i stavolta si posiziona invece che una scatola un cumulo centrato su x_i la cui forma è data da una delle funzioni nucleo descritte precedentemente divise per hn. Il valore di $d(x)$ è pari alla somma del valore dell'ordinata di ciascuna degli n cumuli.

Riconsideriamo i dati dell'esempio precedente e utilizziamo il nucleo gaussiano. Nel grafico rappresentiamo delle piccole gaussiane (che hanno area sottostante pari a $1/n$) centrate su ciascun dato. La curva che si ottiene è la somma delle ordinate di ciascun nucleo.

```
par(mfrow=c(1,1))

xx<-c(148, 155, 162, 168, 172, 174, 175, 178, 186)
plot(xx,rep(0,9), ylim=c(0,0.065), xlim=c(135,200), pch=4, cex=1.5,
     main="esempio", xlab="h=4", ylab="densità")
abline(0,0)

# rappresentiamo le funzioni K, come delle guassiane divise per n
n=length(xx)
for (i in 1:n){
  curve(dnorm(x,xx[i],4)/n, col=(i+1), lwd=3, lty=2, add=TRUE)}

# Successivamente costruiamo la funzione che calcoli per
# ogni x la somma delle ordinate delle funzioni K (la somma
# delle altezze delle gaussiane)

nuclg<-function(x,xx,h=1) {
  nucl=0
  nn<-length(x)
  n=length(xx)
  for (i in 1:nn){
  nucl[i]= sum(dnorm(x[i],xx,h)/n)}
  nucl
}
lines(xr,nuclg(xr,xx,4),lwd=2.5)
```



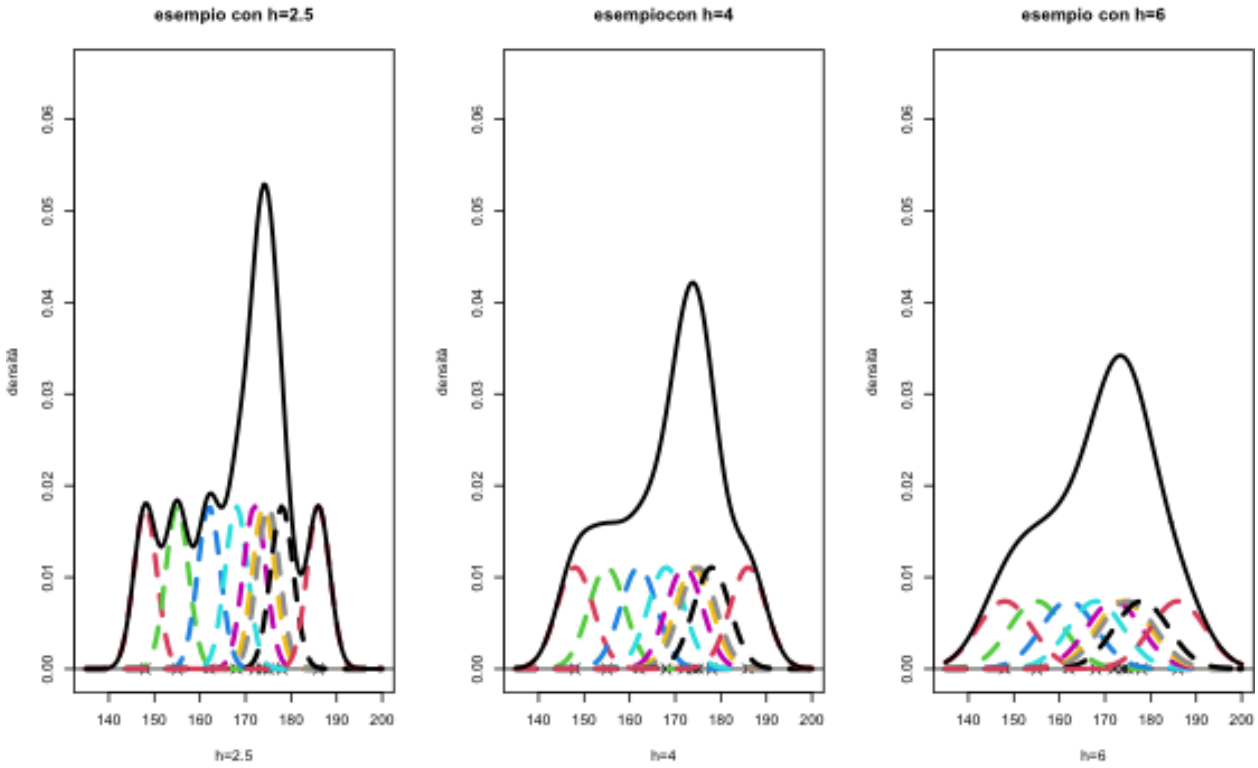
Il parametro di lisciamento

Il comportamento della funzione $d(x)$ ottenuta impiegando la 4 eredita alcune delle proprietà del nucleo utilizzato. In particolare, se $K(u)$ è essa stessa una funzione di densità come nel caso visto del nucleo gaussiano, allora:

- $d(x)$ ha le stesse proprietà di una funzione di densità;
- se la funzione nucleo è derivabile allora lo è anche $d(x)$.

Il grado di lisciamento della curva $d(x)$ dipende tuttavia dal valore di h che è appunto detto parametro di lisciamento. Infatti come si deduce dalla formula 4, h è un fattore di scala che compare fra gli argomenti della funzione K. Valori di h grandi implicano un impatto su $d(x)$ di un generico dato x_i anche per valori molto distanti da esso mentre valori di h piccoli fanno sì che il peso di x_i abbia un ruolo nel determinare il valore di $d(x)$ solo quando x non è molto distante da x_i stesso.

Riprendiamo l'esempio visto e otteniamo le densità ottenute con il metodo del nucleo con tre diversi valori di lisciamento diversi.



La figura a destra con h è più piccolo e con nuclei più appuntiti dà come risultato risultato una curva meno liscia. L'opposto vale per la figura a destra ove i nuclei troppo ampi nascondono le variazioni della densità in alcune aree.

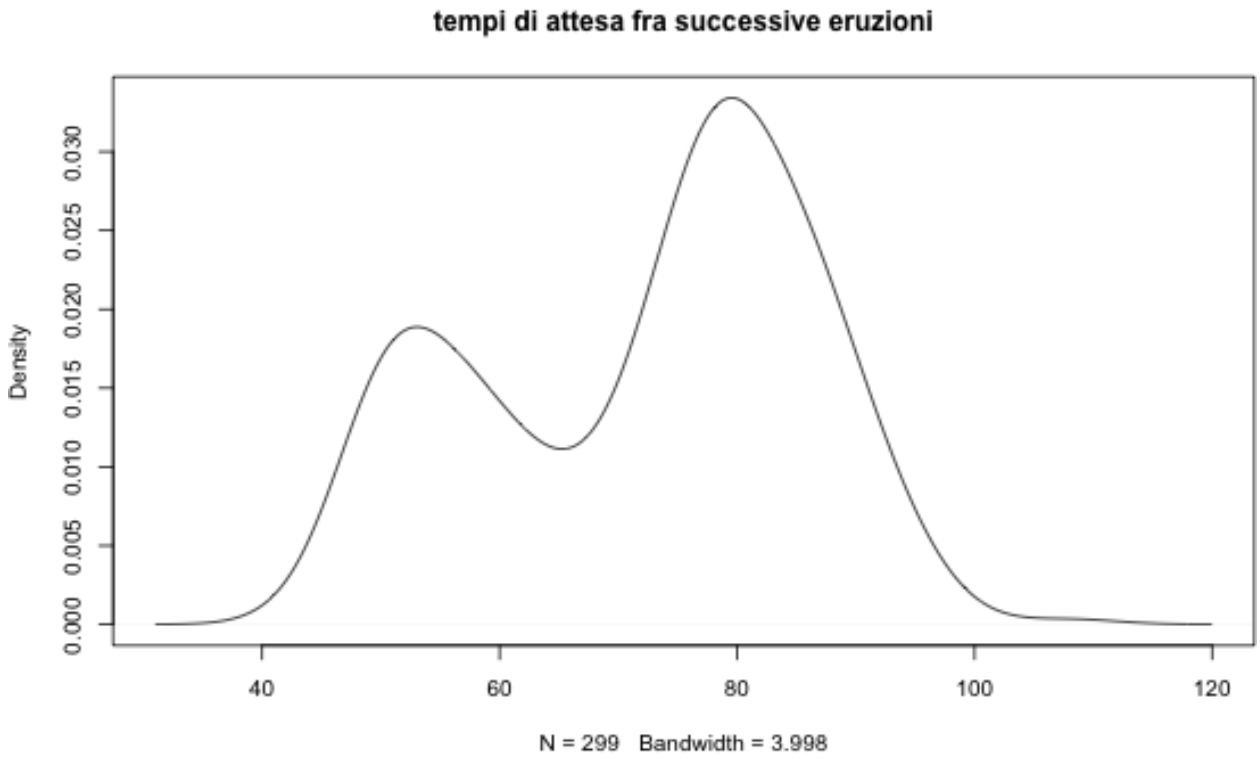
La scelta di h dipende da considerazioni analoghe a quelle fatte per la scelta dell'ampiezza delle classi nel caso dell'istogramma. Un valore del parametro di lisciamento piccolo rischia di introdurre un alto numero di brusche variazioni nella curva di densità in particolar modo nelle regioni dove si osservano pochi dati mentre un valore alto di h permette di determinare una funzione liscia al prezzo di oscurare caratteristiche locali della curva di densità (ad esempio nascondendo effettive caratteristiche bimodali della distribuzione dei dati).

`R` contiene alcune funzioni che consentono il calcolo della curva di densità, e a partire da questo produca un grafico della curva di densità scegliendo un appropriato nucleo e il valore di h opportuno (in `R` il parametro di lisciamento è denotato con `bw` riferito al termine *bandwidth*) .

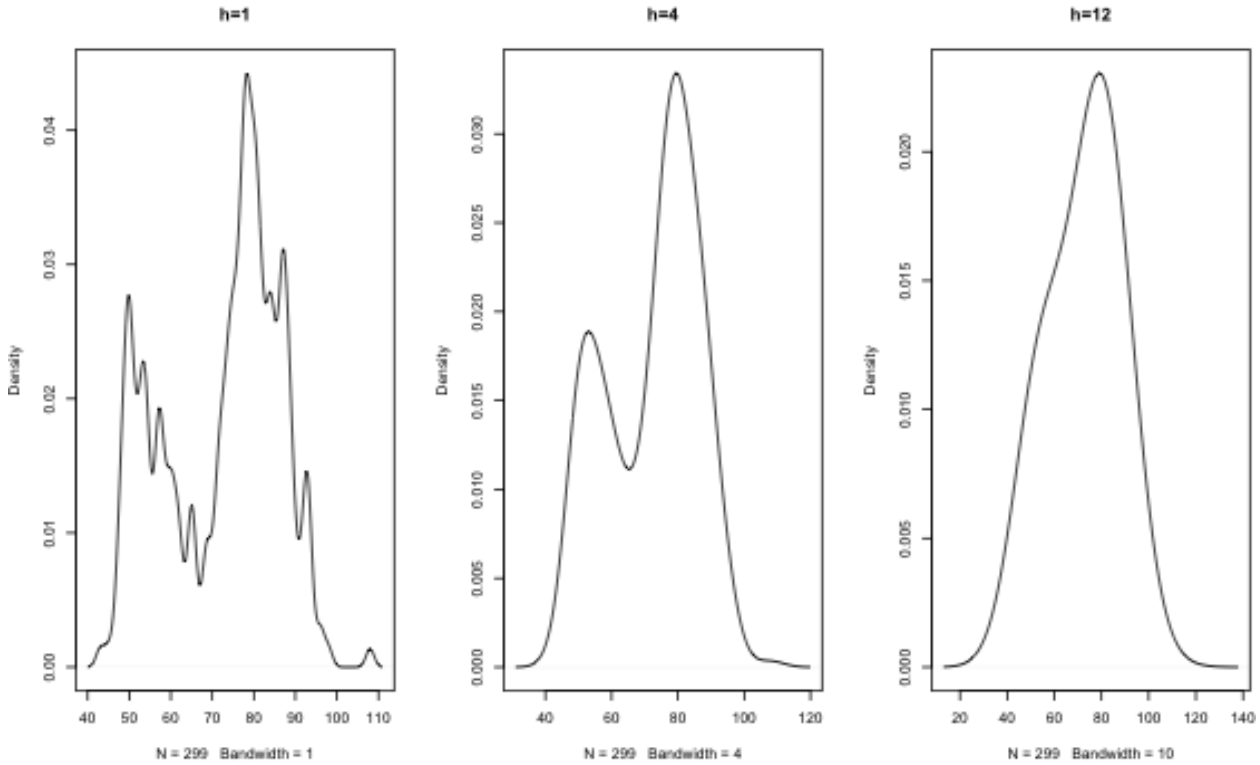
In particolare, se si vuole un grafico della densità con il metodo del nucleo si può utilizzare la funzione `density()`. In essa il grado di lisciamento viene controllato mediante il parametro `bw=` ed è anche possibile scegliere funzioni nucleo alternative (il default è il nucleo gaussiano e va detto che in gnere la scelta del nucleo è meno cruciale rispetto a quella del parametro di lisciamento).

Vediamo cosa accade usando tale funzione con i dati geyser.

```
par(mfrow=c(1,1))

den<-density(geyser$waiting)
plot(den, main="tempi di attesa fra successive eruzioni")
```




```
# Nel grafico è riportato il valore del parametro di lisciamento.
# La funzione density() potrebbe quindi essere chiamata direttamente
# all'interno di altre funzione come la plot() per ottenere il grafico
# della curva di densità.
# Si guardi cosa accade se si varia il parametro di lisciamento

par(mfrow=c(1,3))
plot(density(geyser$waiting, bw=1), main="h=1")
plot(density(geyser$waiting, bw=4), main="h=4")
plot(density(geyser$waiting, bw=10), main="h=12")
```




```
# con un parametro di lisciamento troppo alto si perde la bimodalità
# con uno troppo piccolo appaiono delle  variazioni brusche nella
# densità che riflettono densità locali (è come fare un istogramma
# con troppe classi rispetto alla numerosità dei dati)
par(mfrow=c(1,1))
```

Con insiemi di dati che presentano forte asimmetria il valore di h scelto potrebbe essere adeguato per le regioni in cui la densità è alta e essere invece troppo piccolo per ottenere un buon grado di lisciamento lungo la coda. è possibile in tal caso introdurre criteri flessibili che permettano di utilizzare valori di h diversi in relazione alla diversa densità locale dei dati.

In generale, è consigliabile ottenere differenti curve di densità per diversi valori di h e giudicare a posteriori quale valore di h mostri di cogliere adeguatamente le caratteristiche salienti dell'insieme di dati.

Tuttavia è possibile tenere presenti alcuni semplici criteri pratici. è ad esempio ragionevole attendersi che il valore di h ideale sia inversamente proporzionale ad una funzione di n. In genere si consiglia di scegliere h proporzionale a $n^{-\frac{1}{5}}$. Se si usa una funzione nucleo gaussiana il valore di h di default è $h = 0,9An^{-\frac{1}{5}}$ ove

$$A = \frac{min(\text{scarto quadratico medio}, \text{scarto interquartile})}{1.34}.$$

Per le altre opzioni della funzione `density` si rinvia a quanto contenuto nella documentazione in linea di `R`. Una seconda funzione, anche più generale, per determinare la funzione di densità con il metodo del nucleo è la `bkde()` nel package `KernSmooth`, che funziona in modo analogo alla `density()`.