(Bio)Statistica con R

Parte IV

UNIVERSITÀ DEGLI STUDI DI FOGGIA

Dr. Crescenzio Gallo – Università di Foggia – Dipartimenti di Area Medica

(Bio)Statistica con R — Parte IV

Confronto tra due metodi analitici

- La regressione lineare standard non è adatta al confronto tra due variabili legate tra loro da una relazione lineare se anche la variabile indipendente è affetta da errore di misura. Ed è proprio quello che accade quando si confrontano tra di loro i risultati di due metodi analitici per la determinazione della stessa sostanza.
- In questo caso viene suggerito una approccio che prevede l'ispezione dei dati mediante il **diagramma di Bland e Altman**, e l'impiego della **regressione lineare non parametrica** di Passing e Bablok, che assume che entrambe le variabili siano affette da un errore di misura.
- L'approccio globale al confronto tra due metodi con **R** è stato sviluppato con la libreria **MethComp**, eventualmente da installare. Per utilizzare la libreria, la struttura dei dati deve prevedere obbligatoriamente il campo **meth** (il metodo di analisi), il campo **item** (il numero progressivo del campione analizzato), il campo **repl** (il numero del replicato) e il campo **y** (il risultato numerico dell'analisi).

(Bio)Statistica con R — Parte IV

Confronto tra due metodi analitici

- Salviamo nella directory di lavoro il file **MethComp.csv**, che contiene i dati relativi al confronto tra due metodi analitici organizzati esattamente come previsto dalla libreria.
- In questo caso il numero del replicato è sempre uguale a 1 non essendo previste analisi in replicato.
- Eseguiamo il seguente codice:

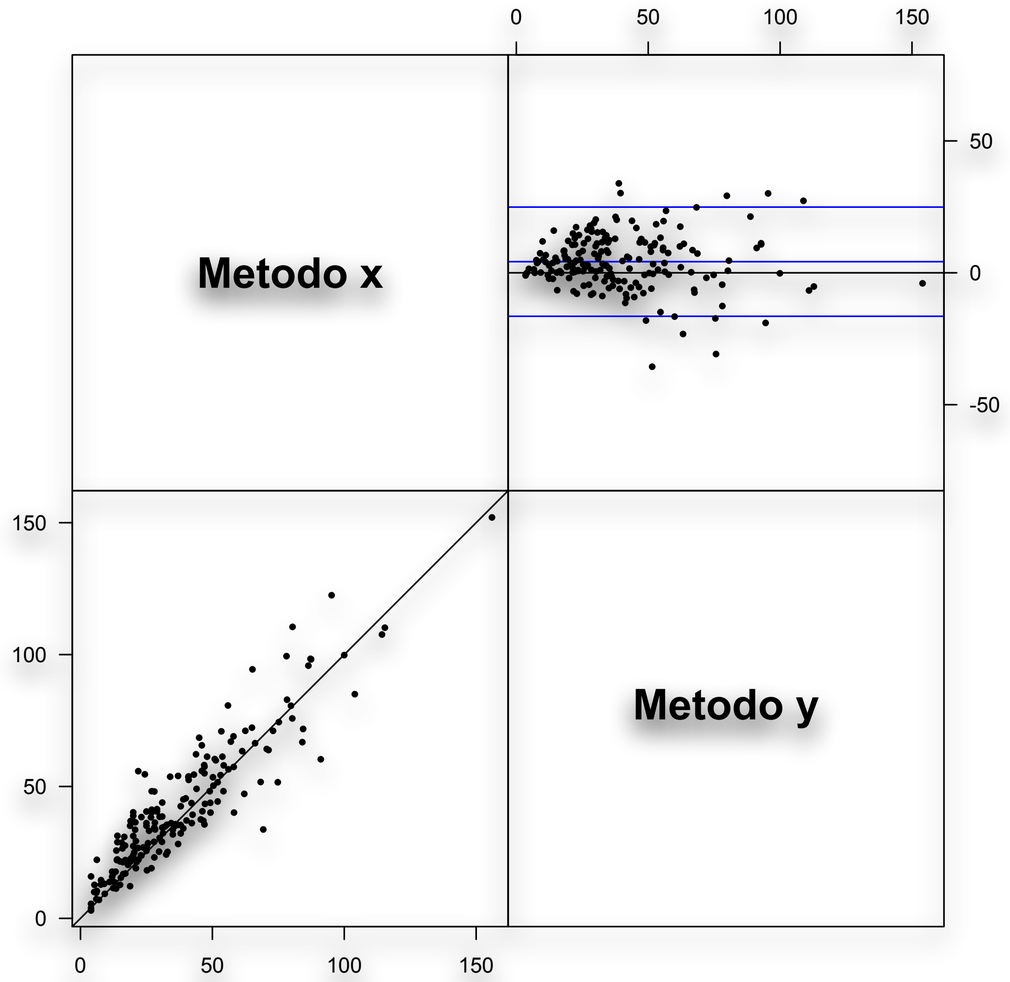
```
> mydata <- read.table("MethComp.csv", header=TRUE, sep=";")
> library(MethComp)
> newdata <- Meth(mydata) # crea un oggetto Meth per la libreria
           #Replicates
Method                1 #Items #Obs: 378 Values:  min  med max
  Metodo x          189    189       189            4 31.0 156
  Metodo y          189    189       189            3 36.3 152
> plot(newdata)
```

meth	Item	repl	y
Metodo x	1	1	4
Metodo y	1	1	3
Metodo x	2	1	4
Metodo y	2	1	3.9
Metodo x	3	1	4
Metodo y	3	1	5.5
Metodo x	4	1	4
Metodo y	4	1	15.9
Metodo x	5	1	5.2
Metodo y	5	1	10
Metodo x	6	1	5.3
Metodo y	6	1	12.7

(Bio)Statistica con R — Parte IV

Confronto tra due metodi analitici

Nella figura a lato vediamo la sintesi grafica dei dati del confronto tra metodi, con il ***diagramma di Bland e Altman*** *in alto a destra e la* ***regressione lineare non parametrica di Passing e Bablok*** *in basso a sinistra.*



(Bio)Statistica con R — Parte IV

Confronto tra due metodi analitici

- Per questo grafico è necessaria l'impostazione dei margini (il relativo problema viene illustrato successivamente):

```
> predef <- par()$mar  # salva i valori predefiniti dei margini
> par(mar = c(5,5,5,4))  # imposta margini più ampi
>BA.plot(newdata, main = "Grafico di Bland e Altman")
```

- Al momento della creazione dell'oggetto **Meth** per la libreria viene fornita una breve sintesi dei dati:

```
> newdata <- Meth(mydata)  # crea un oggetto Meth per la libreria
The following variables from the dataframe
"mydata" are used as the Meth variables:
meth: meth
item: item
repl: repl
   y: y
           #Replicates
Method                1 #Items #Obs: 378 Values:  min  med max
  Metodo x          189    189       189            4 31.0 156
  Metodo y          189    189       189            3 36.3 152
```

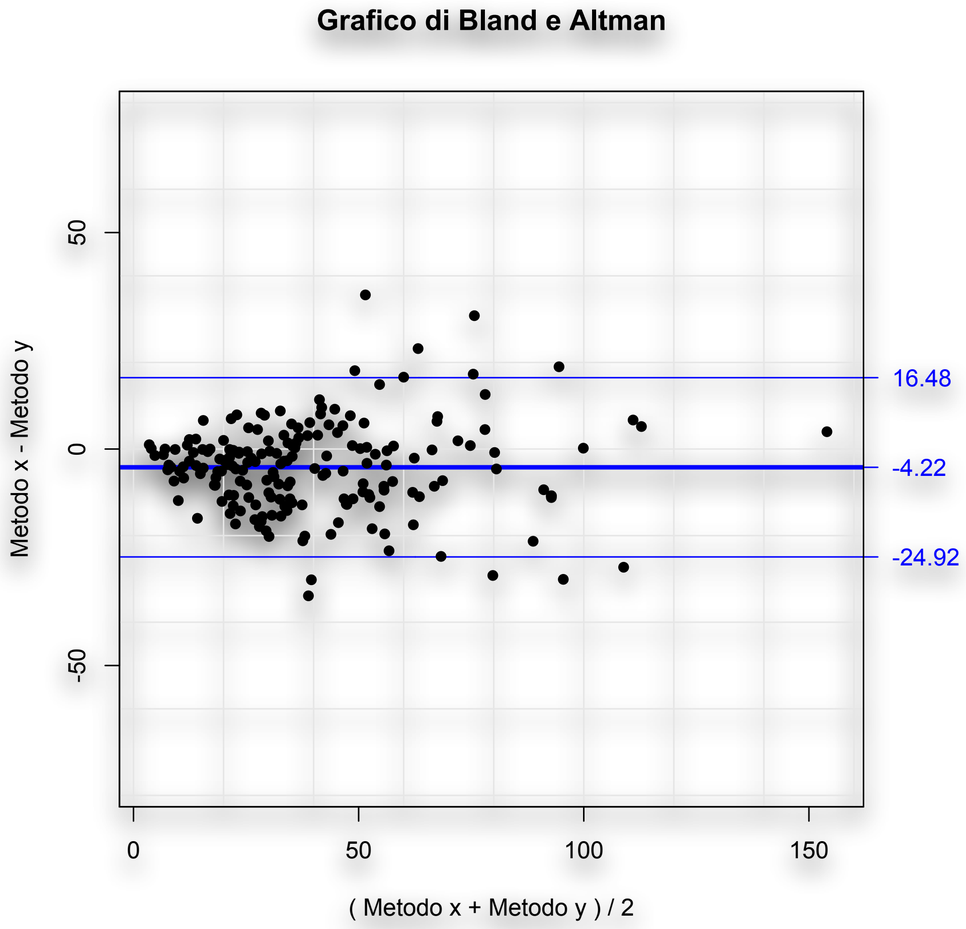


Diagramma di Bland e Altman con la media e i limiti di confidenza al 95% della media.

(Bio)Statistica con R — Parte IV

Confronto tra due metodi analitici

- Infine eseguiamo questo codice:

```
> print(PBreg(newdata)) # statistiche della regressione di Passing e Bablok
$coefficients
            Estimate    2.5%CI 97.5%CI
Intercept 3.9340857 2.1687117 5.37963
Slope     0.9888262 0.9296296 1.06135
> par(mar = predef) # ripristina i valori predefiniti dei margini
```

- Per l'intercetta e per il coefficiente angolare (slope) sono riportati i limiti di confidenza al 95%, che consentono di effettuare immediatamente la valutazione della significatività della differenza dell'intercetta da 0 (zero è il valore atteso dell'intercetta se tra i due metodi non vi è errore sistematico di tipo costante) e del coefficiente angolare da 1 (uno è il valore atteso del coefficiente angolare se tra i due metodi non vi è errore sistematico di tipo proporzionale).

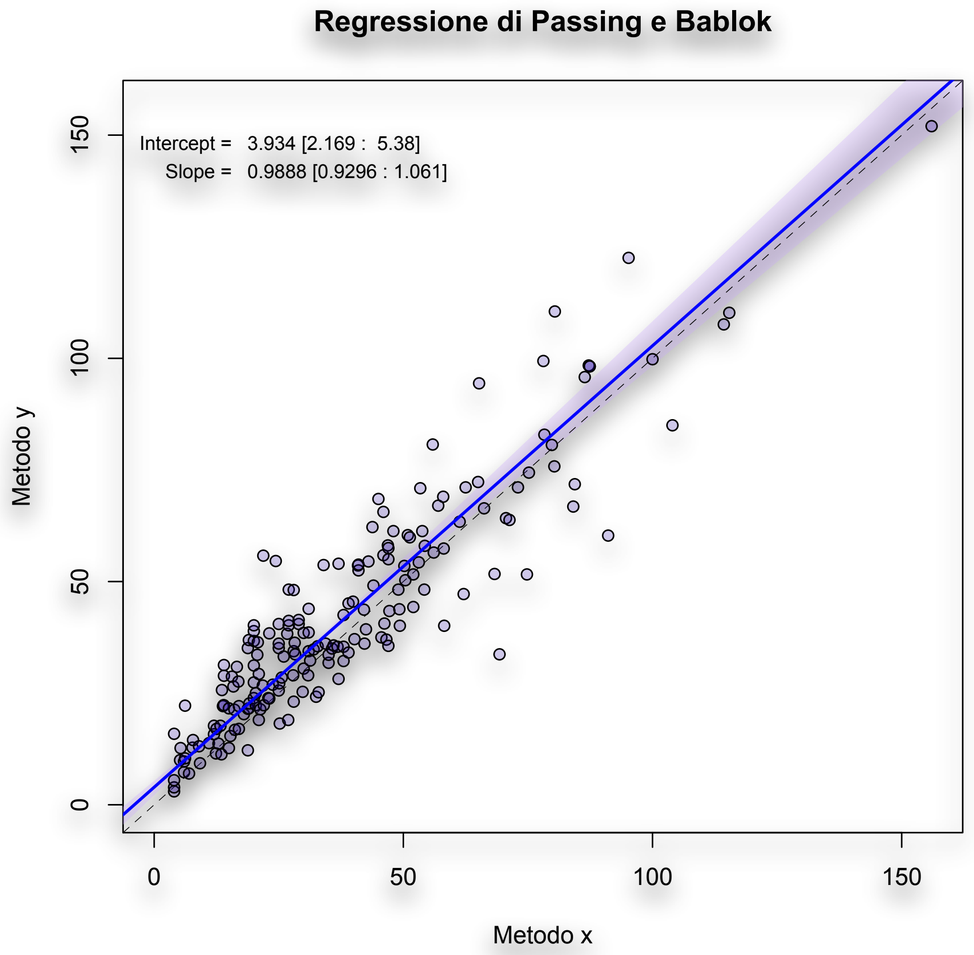
(Bio)Statistica con R — Parte IV

Confronto tra due metodi analitici

- Tracciamo il grafico:

```
> plot(PBreg(newdata),
  main = "Regressione di Passing e Bablok")
```

- Oltre alle statistiche (vise in precedenza) la figura a lato mostra il grafico con la retta teorica di equivalenza "metodo x = metodo y" tratteggiata, e con la retta trovata che conferma graficamente l'esistenza tra i due metodi di una differenza sistematica di tipo costante, che ovviamene necessita di interpretazione dal punto di vista analitico.

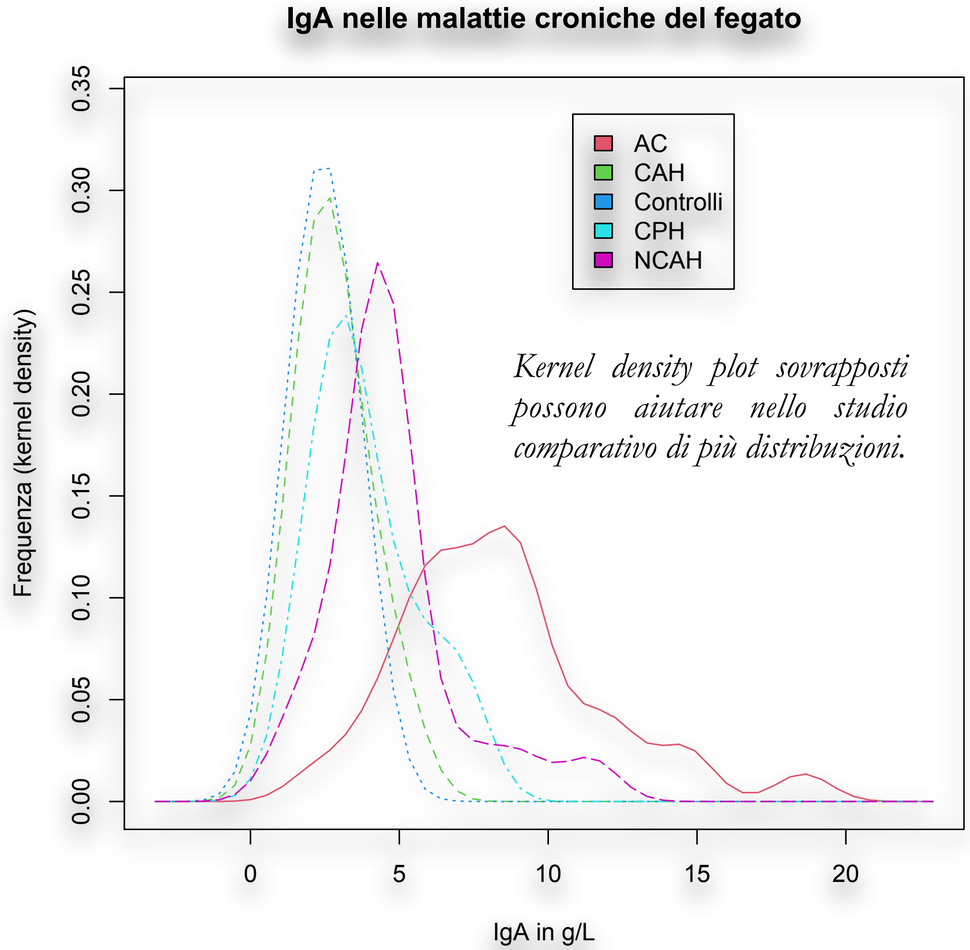


(Bio)Statistica con R — Parte IV

Kernel density plot sovrapposti

- Abbiamo già visto nella Parte III il codice **R** per tracciare due kernel density plot sovrapposti. Qui ne sovrapporremo cinque.
- Utilizziamo gli stessi dati (**Boxplot.csv**) forniti per tracciare i box & whiskers plot; servirà anche la libreria **sm**.

```
> mydata <- read.table("Boxplot.csv", header=TRUE,
  sep=";")
> library(sm); attach(mydata)
> myplot <- factor(Diagnosi,
  levels = c("AC","CAH","Controlli","CPH","NCAH"),
  labels = c("AC","CAH","Controlli","CPH","NCAH"))
> sm.density.compare(IgA, myplot, xlab="IgA in g/L",
  ylab="Frequenza (kernel density)")
> title(main="IgA nelle malattie croniche del fegato")
# aggiungo la legenda
> colfill<-c(2:(2+length(levels(myplot))-1))
> legend(locator(1), levels(myplot), fill=colfill)
```



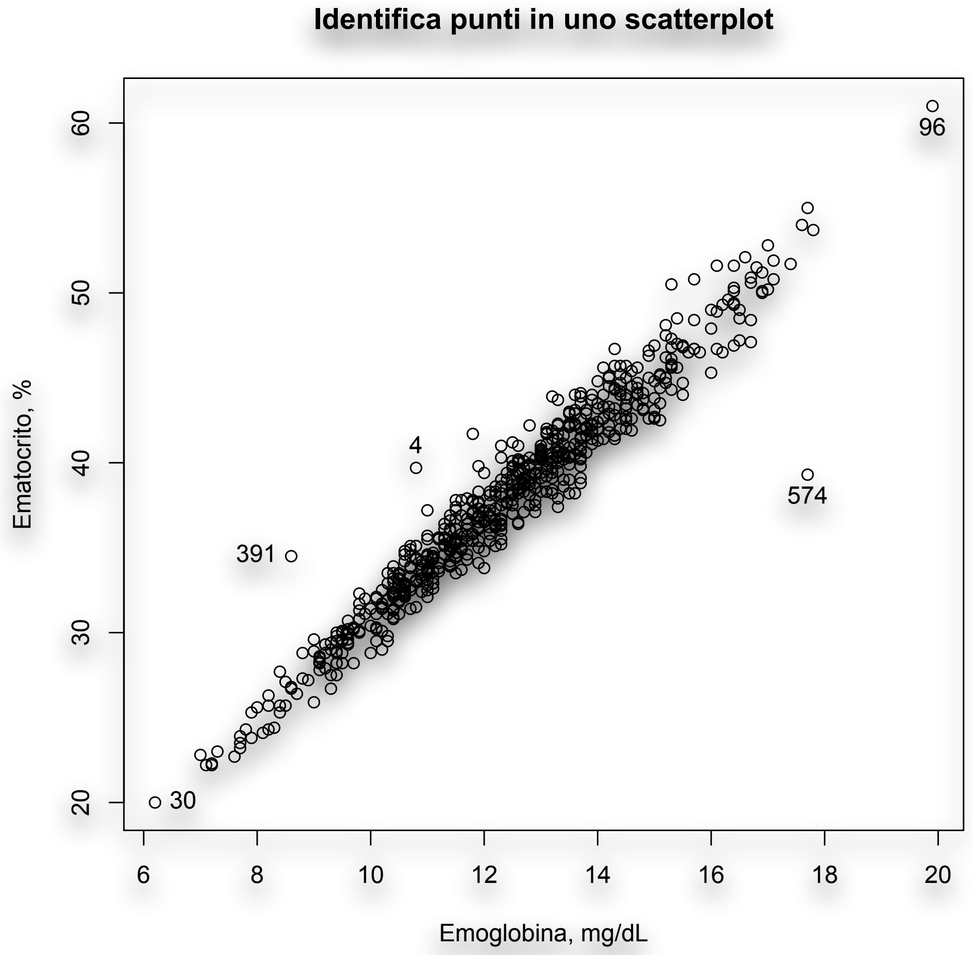
(Bio)Statistica con R — Parte IV

Identificare i punti in uno scatter plot

- Questo codice risponde a un problema banale, ma che si pone sovente: questo punto che si discosta così tanto dagli altri a quale dato corrisponde?
- Utilizziamo il file **Scatterplot.csv**:

```
> mydata <- read.table("Scatterplot.csv", header=TRUE,
   sep=";")
> attach(mydata)
> plot(HB, HCT, main="Identifica punti in uno scatterplot",
   xlab="Emoglobina, mg/dL", ylab="Ematocrito, %", pch=1)
>identify(HB, HCT, plot = TRUE, atpen = FALSE, offset =
0.5, tolerance = 0.25, locatorBell = TRUE)
```

- Posizionarsi nelle vicinanze del punto cui si è interessati e che si vuole identificare e fare click con il tasto sinistro del mouse: nella posizione prescelta comparirà il numero del dato. Per terminare selezionare **Stop** con il tasto destro del mouse. Il risultato è riportato nella figura a lato.



(Bio)Statistica con R — Parte IV

Adattare i margini a un'immagine

- Nel problema del confronto tra metodi avrete certamente notato questa strana riga di codice con il relativo commento:

```
> par(mar = c(5,5,5,4)) # imposta margini più ampi
```

Per avere la spiegazione bisogna scaricate la libreria **MethComp** ed eseguire questo codice:

```
> library(MethComp)
> mydata <- read.table("MethComp.csv", header=TRUE, sep=";")
> newdata <- Meth(mydata) # crea un oggetto Meth per la libreria
> BA.plot(newdata, main = "Grafico di Bland e Altman")
```

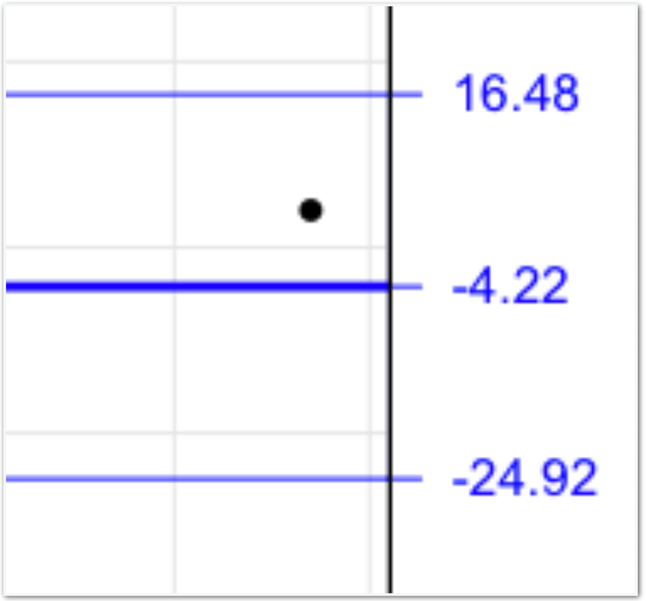
- Come si può vedere i margini sono insufficienti e i valori sulla scala di destra risultano troncati.

(Bio)Statistica con R — Parte IV

Adattare i margini a un'immagine

- Ora eseguiamo questo codice:

```
> mydata <- read.table("MethComp.csv", header=TRUE,
  sep=";")
> library(MethComp)
> newdata <- Meth(mydata) # crea un oggetto Meth per la libreria
> predef <- par()$mar # salva i valori predefiniti dei margini
> par(mar = c(5,5,5,4)) #imposta i nuovi margini
> BA.plot(newdata, main = "Grafico di Bland e Altman")
> par(mar = predef) # ripristina i valori predefiniti dei margini
```



- Come si può vedere i margini sono ora sufficienti a contenere i valori sulla scala di destra.

(Bio)Statistica con R — Parte IV

Inserire più grafici in un'immagine

- Anche questo è un problema banale ma che può creare qualche problema senza trovare la soluzione.
- Scarichiamo e salviamo il file **Verigauss.csv**. Contiene i dati di sesso, età e concentrazione di colesterolo totale, colesterolo HDL, colesterolo LDL e trigliceridi che abbiamo già incontrato in precedenza:

Sesso	Eta	Colesterolo	HDL	LDL	Trigliceridi
M	33	56	44	9	19
M	62	60	5		
F	90	70	30		99
M	75	80	53		
F	32	82	51		23
M	71	84	25		
F	86	89			
F	64	91	35		88
M	90	91	43		
F	95	94	12		260
M	46	97	41	45	86

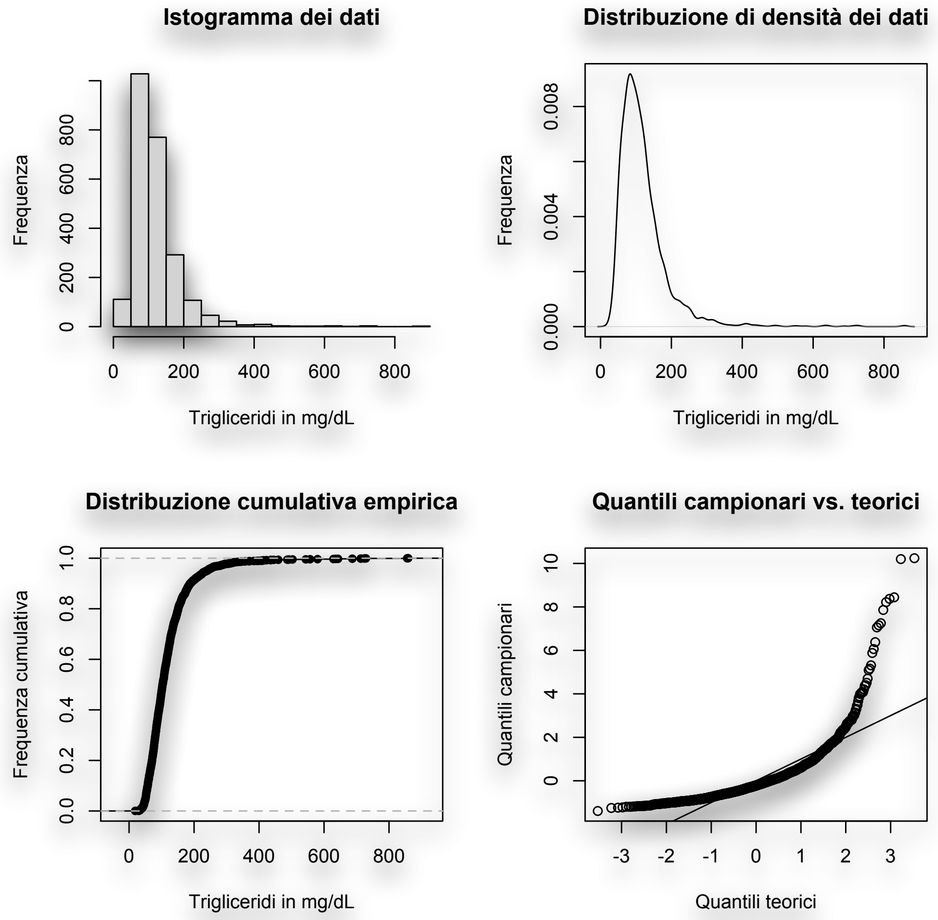
(Bio)Statistica con R — Parte IV

Inserire più grafici in un'immagine

- Eseguiamo questo codice:

```
> mydata <- read.table("Verigauss.csv", header=TRUE, sep=";")
> newdata <- na.omit(mydata) # esclude i casi con dati mancanti
> tri <- newdata$Trigliceridi
> par(mfrow=c(2,2))
> hist(tri, main="Istogramma dei dati",
   xlab="Trigliceridi in mg/dL", ylab = "Frequenza")
> plot(density(tri), main="Distribuzione di densità dei dati",
   xlab="Trigliceridi in mg/dL", ylab = "Frequenza")
> plot(ecdf(tri), main="Distribuzione cumulativa empirica",
   xlab="Trigliceridi in mg/dL",
   ylab = "Frequenza cumulativa")
> qqnorm((tri-mean(tri))/sd(tri),
   main="Quantili campionari vs. teorici",
   xlab="Quantili teorici", ylab = "Quantili campionari")
> abline (0,1) # linea di allineamento teorico di dati gaussiani
```

- Il comando chiave è "par(mfrow=c(2,2))" che predispone la matrice 2 righe x 2 colonne da riempire con i quattro grafici (hist, plot(density, plot(ecdf e qqnorm) per riga, ovvero da sinistra in alto a destra in basso (vedi figura a lato).

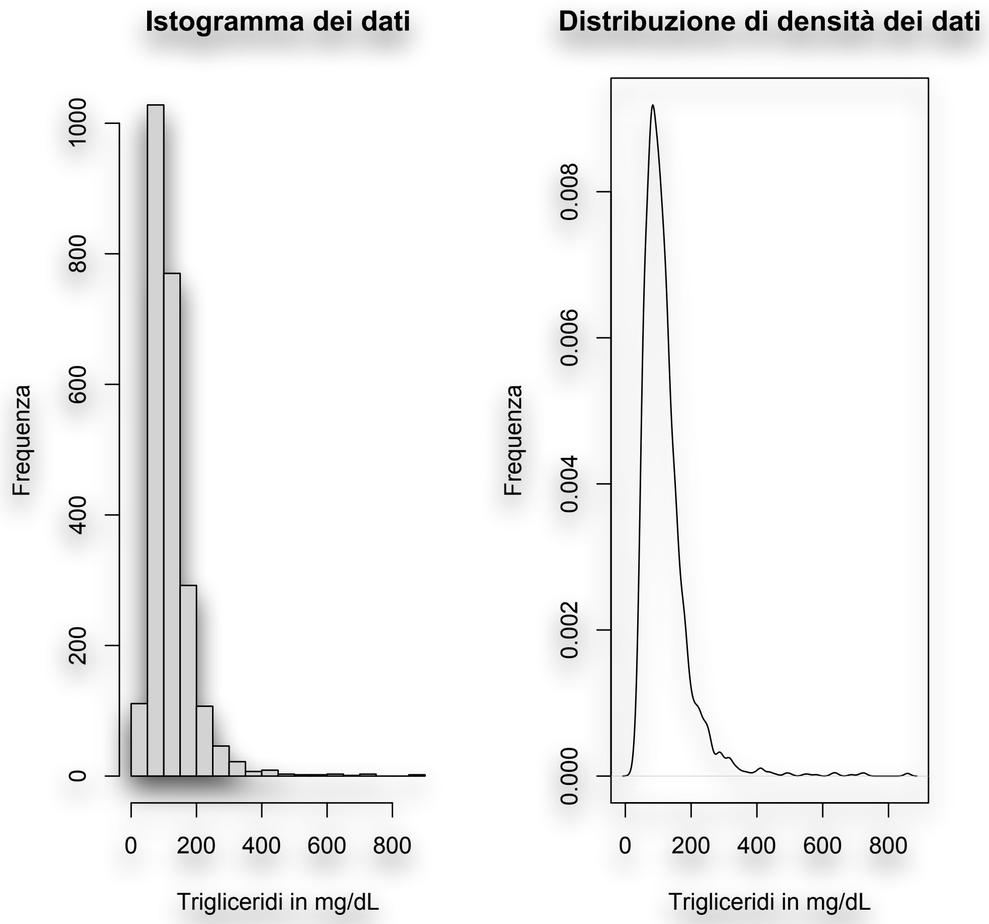


(Bio)Statistica con R — Parte IV

Inserire più grafici in un'immagine

- Le seguenti tre varianti servono come esercizio per familiarizzare con il tema.
- La **prima variante** inserisce due grafici in una riga e due colonne:

```
> par(mfrow=c(1,2))
> hist(tri, main="Istogramma dei dati",
  xlab="Trigliceridi in mg/dL",
  ylab = "Frequenza")
> plot(density(tri),
  main="Distribuzione di densità dei dati",
  xlab="Trigliceridi in mg/dL",
  ylab = “Frequenza")
```

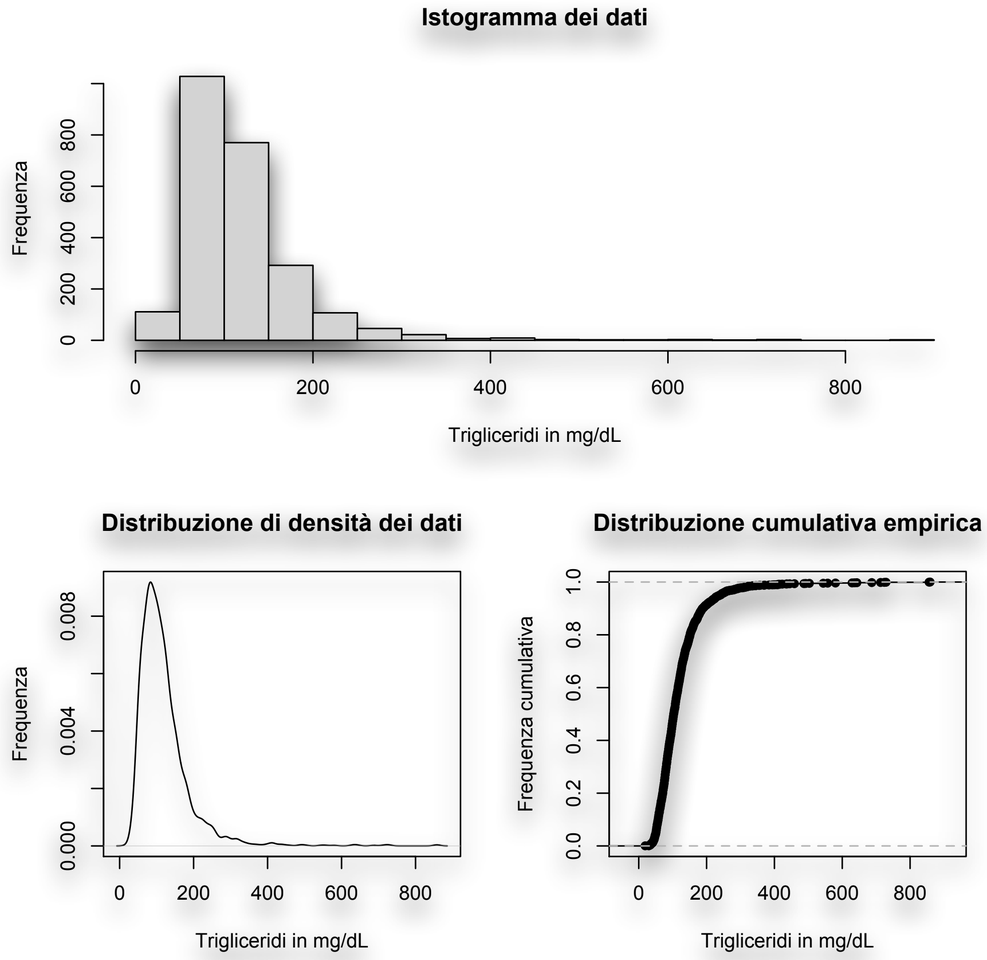


(Bio)Statistica con R — Parte IV

Inserire più grafici in un'immagine

- Questa **seconda variante** inserisce tre grafici, il primo occupa la riga 1, colonne 1 e 2; il secondo la riga 2, colonna 1; il terzo la riga 2, colonna 2:

```
> layout(matrix(c(1,1,2,3), 2, 2, byrow = TRUE))
> hist(tri, main = "Istogramma dei dati",
    xlab = "Trigliceridi in mg/dL",
    ylab = "Frequenza")
> plot(density(tri),
    main = "Distribuzione di densità dei dati",
    xlab = "Trigliceridi in mg/dL",
    ylab = "Frequenza")
> plot(ecdf(tri),
    main = "Distribuzione cumulativa empirica",
    xlab = "Trigliceridi in mg/dL",
    ylab = "Frequenza cumulativa")
```



(Bio)Statistica con R — Parte IV

Inserire più grafici in un'immagine

- Infine questa terza variante inserisce tre grafici, il primo occupa la riga 1, colonna 1; il secondo la riga 2, colonna 1; il terzo la colonna 2, righe 1 e 2:

```
> layout(matrix(c(1,3,2,3), 2, 2,
  byrow = TRUE), widths = c(1,1),
  heights = c(1,1))
> hist(tri, main = "Istogramma dei dati",
  xlab = "Trigliceridi in mg/dL",
  ylab = "Frequenza")
> plot(density(tri),
  main = "Distribuzione di densità dei dati",
  xlab = "Trigliceridi in mg/dL",
  ylab = "Frequenza")
> plot(ecdf(tri),
  main = "Distribuzione cumulativa empirica",
  xlab = "Trigliceridi in mg/dL",
  ylab = "Frequenza cumulativa")
```