(Bio)Statistica con R

Parte II

UNIVERSITÀ DEGLI STUDI DI FOGGIA

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

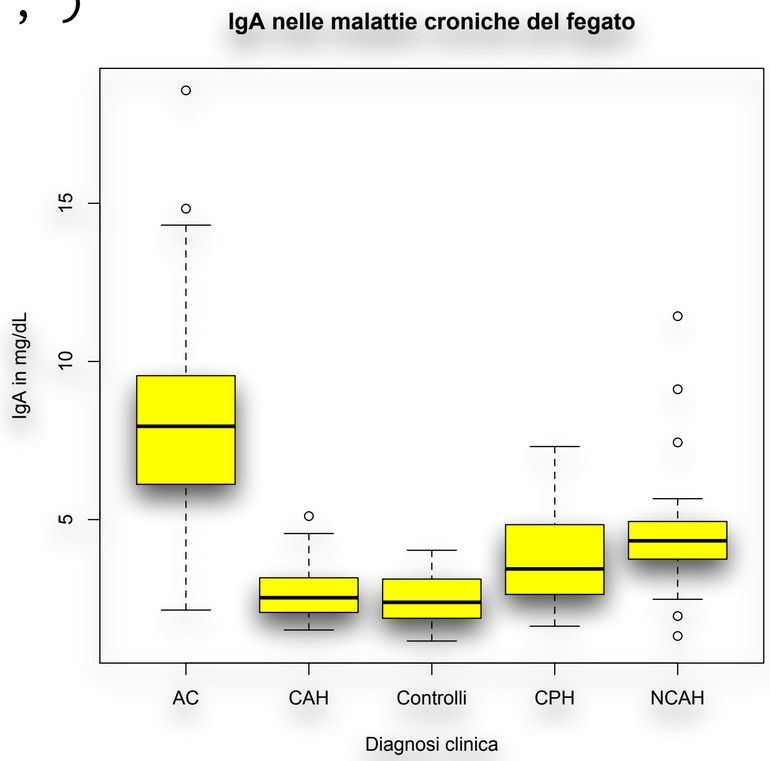
(Bio)Statistica con R — Parte II

- L'impostazione di questa seconda parte dell'esercitazione di (bio)statistica con R è quella di *imparare* attraverso una serie diesempi che includono sia i dati da elaborare sia il codice **R** che li elabora.
- I file contengono i dati da elaborare in formato **csv** (comma separated value), il formato dati raccomandato per **R**. I file **csv** possono essere generati con Excel e LibreOffice Calc semplicemente selezionando il formato al momento di salvare i dati.
- Da notare che negli esempi forniti il separatore nei file **csv** è sempre il punto e virgola (**;**). **R** riconosce il punto e virgola come separatore di campo quando viene specificato il parametro **sep=";"**. Cambiando il valore di tale parametro è possibile importare dati delimitati per esempio con la virgola (**sep=","**), con uno spazio vuoto (**sep=" "**) o con qualsiasi altro separatore.
- Anche i parametri **quote** e **dec** servono a indicare il comportamento dei comandi di importazione **read.table**, **read.csv**, **read.csv2**.

(Bio)Statistica con R — Parte II

- Ecco un primo esempio per iniziare a familiarizzare con dati e script R. D'ora in avanti supporremo che i file siano registrati e letti dalla directory di lavoro impostata all'inizio della sessione **R**.
- Scarichiamo e salviamo nella directory di lavoro il file Boxplot.csv; quindi eseguiamo il seguente codice:

```
# importo i dati
>mydata <- read.table(file="Boxplot.csv", header=TRUE, sep=";")
# traccio i boxplot delle IgA per ciascuna diagnosi
>boxplot(IgA~Diagnosi, data=mydata,
    main="IgA nelle malattie croniche del fegato",
    xlab="Diagnosi clinica", ylab="IgA in mg/dL",
    notch=FALSE, col="yellow")
```



Il grafico illustra la concentrazione delle IgA (in g/L) in un gruppo di soggetti sani (Controlli) e la confronta graficamente con quella rilevata in soggetti con cirrosi alcolica (AC), epatite cronica attiva (CAH), epatite cronica persistente (CPH), epatite alcolica non cirrotica (NCAH).

(Bio)Statistica con R — Parte II

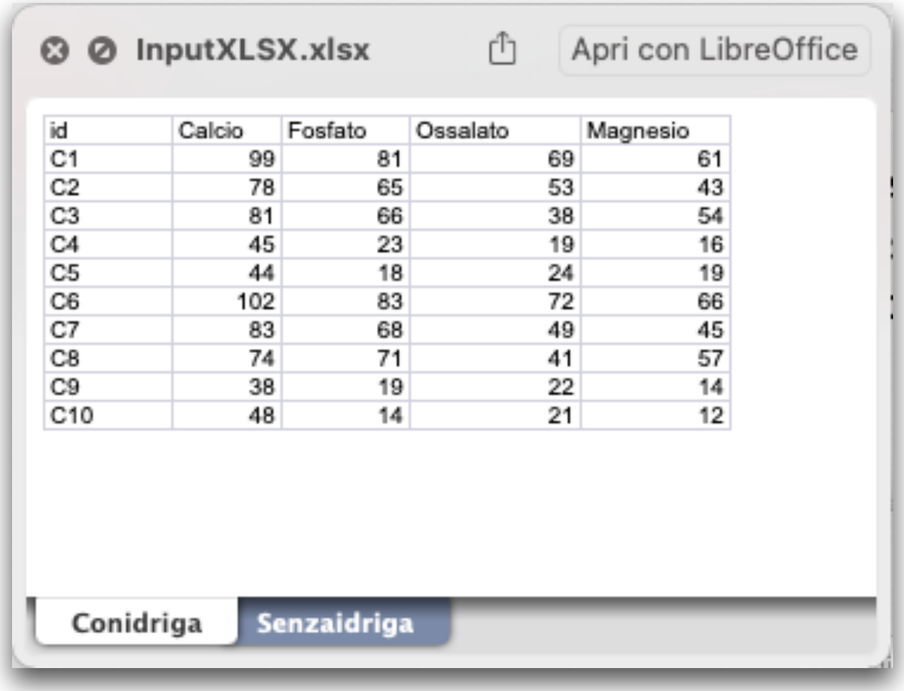
- In quest'altro esempio vediamo come importare direttamente in **R** file di tipo Excel (xlsx). Scarichiamo questo file di esempio: **InputXLSX.xlsx** e registriamolo nella directory di lavoro.
- Quindi installiamo il package *xlsx* ed attiviamolo mediante i seguenti comandi (il package *xlsx* richiede Java sul computer):

```
> install.packages("xlsx")
> library(xlsx)
```

- Possiamo infine importare il file Excel con il seguente comando:

```
> mydata <- read.xlsx("InputXLSX.xlsx",
    sheetName = "Conidriga")
```

- Per le molte altre applicazioni della libreria `xlsx` si rimanda alla sua documentazione che si trova sul CRAN anche digitando semplicemente "package R xlsx" nella casella di ricerca di Google.



id	Calcio	Fosfato	Ossalato	Magnesio
C1	99	81	69	61
C2	78	65	53	43
C3	81	66	38	54
C4	45	23	19	16
C5	44	18	24	19
C6	102	83	72	66
C7	83	68	49	45
C8	74	71	41	57
C9	38	19	22	14
C10	48	14	21	12

(Bio)Statistica con R — Parte II

Gestione dei dati mancanti

- Scarichiamo e salviamo il file InputNA.csv
- Il file contiene la concentrazione delle **IgA** (in g/L) in un gruppo di soggetti sani (**Controlli**) e in soggetti con cirrosi alcolica (**AC**), epatite cronica attiva (**CAH**), epatite cronica persistente (**CPH**), epatite alcolica non cirrotica (**NCAH**) organizzati in cinque colonne.
- Le cinque classi di pazienti contengono ciascuna un numero differente di casi, come possiamo vedere nello stralcio del contenuto qui a lato:
- Carichiamo i dati ed esaminiamo alcune colonne con i seguenti comandi:

```
> mydata <- read.table("InputNA.csv", header=TRUE, sep=";")
> colMeans(mydata[c("Normali")])
> colMeans(mydata[c("CPH")])
> colMeans(mydata[c("CPH")], na.rm=TRUE)
```

Normali	NCAH	CPH	CAH	AC
1.22	7.44	2.45	2.35	3.51
2.81	4.58	1.63	3.21	4.23
4.02	3.71	3.44	3.88	7.66
2.23	4.94	2.47	1.56	9.54
2.35	3.49	1.95	1.78	11.35
1.64	3.88	4.56	2.49	6.43
2.08	4.71	7.31	3.11	5.28
1.96	4.32	5.78	4.56	2.14
1.54	4.9	3.4	5.11	4.76
1.63	11.43	5.12	2.36	7.91
3.25	4.63	6.88	2.98	9.33
2.9	4.11	3.21	2.53	18.57
3.44	5.03	3.64	1.77	8.81
2.55	9.12	2.8	1.51	14.31
1.18	1.32	3.47	2.93	10.83
1.78	4.33			8.48
2.56	5.66			9.56
1.36	4.08			9.01
1.83	2.48			12.44
2.4	1.95			7.61
2.61	3.75			7.03

(Bio)Statistica con R — Parte II

Gestione dei dati mancanti

- Nell'oggetto **mydata** che contiene i dati importati si può vedere che **R** ha sostituito automaticamente i dati mancanti con la sigla **NA** (che sta per Not Available),
- La media nella colonna **Normali**, nella quale non vi sono dati mancanti, viene calcolata e visualizzata senza problemi:

```
> colMeans(mydata[c("Normali")])
2.457
```

- La media della colonna `CPH` non può essere calcolata causa dei dati mancanti, e viene restituito `NA`:

```
> colMeans(mydata[c("CPH")])
NA
```

- Con il parametro **na.rm=TRUE** che rimuove i dati mancanti la media della colonna `CPH` viene invece calcolata correttamente:

```
> colMeans(mydata[c("CPH")])
3.874
```

- Con la funzione **na.omit()** è possibile eliminare definitivamente da una tabella i casi con dati mancanti.

Normali	NCAH	CPH	CAH	AC
1.22	7.44	2.45	2.35	3.51
2.81	4.58	1.63	3.21	4.23
4.02	3.71	3.44	3.88	7.66
2.23	4.94	2.47	1.56	9.54
2.35	3.49	1.95	1.78	11.35
1.64	3.88	4.56	2.49	6.43
2.08	4.71	7.31	3.11	5.28
1.96	4.32	5.78	4.56	2.14
1.54	4.9	3.4	5.11	4.76
1.63	11.43	5.12	2.36	7.91
3.25	4.63	6.88	2.98	9.33
2.9	4.11	3.21	2.53	18.57
3.44	5.03	3.64	1.77	8.81
2.55	9.12	2.8	1.51	14.31
1.18	1.32	3.47	2.93	10.83
1.78	4.33			8.48
2.56	5.66			9.56
1.36	4.08			9.01
1.83	2.48			12.44
2.4	1.95			7.61
2.61	3.75			7.03

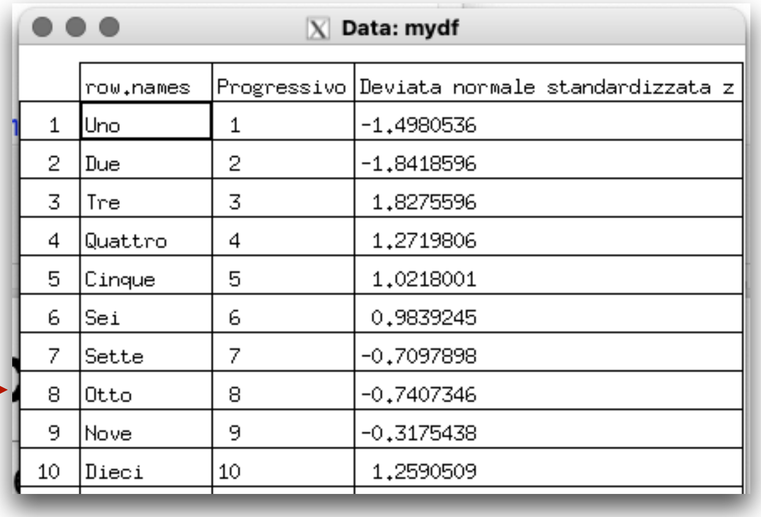
(Bio)Statistica con R — Parte II

Inserimento manuale dei dati

- Se normalmente i dati sono importati dall'esterno, in alcuni casi potrebbe essere utile gestirli direttamente dalla console di **R**. Per questo vediamo un esempio che illustrano la sintassi da utilizzare per inserire direttamente da tastiera array (vettori) numerici e non, e combinarli in dataframe (dataset o tabelle) assegnando i nomi alle variabili.

- L'esempio genera due vettori, li combina in una matrice, assegna i nomi alle variabili (colonne) e assegna un descrittore ai casi (righe):

```
> x <- 1:10 # PRIMO VETTORE – genero gli interi da 1 a 10
> y <- rnorm(10) # SECONDO VETTORE – genero dieci valori di deviata normale standardizzata z
> mydf <- data.frame(x,y) # combino i vettori x e y nel dataframe mydf
> col.names(mydf) <- c("Progressivo",
    "Deviata normale standardizzata z") # assegno i nomi alle variabili
> row.names(mydf) <- c("R1", "R2", "R3", "R4", "R5", "R6",
  "R7", "R8", "R9", "R10") # assegno un descrittore ai casi/righe
> View(mydf) # esamino il dataframe
> fix(mydf) # consente anche di modificare manualmente il contenuto mediante l'editor di dati R
```



Data: mydf

	row.names	Progressivo	Deviata normale standardizzata z
1	Uno	1	-1.4980536
2	Due	2	-1.8418596
3	Tre	3	1.8275596
4	Quattro	4	1.2719806
5	Cinque	5	1.0218001
6	Sei	6	0.9839245
7	Sette	7	-0.7097898
8	Otto	8	-0.7407346
9	Nove	9	-0.3175438
10	Dieci	10	1.2590509

(Bio)Statistica con R — Parte II

Test chi-quadrato (χ^2)

- Nel caso delle scale nominali e delle scale ordinali esiste un solo modo per esprimere le osservazioni in modo quantitativo (numerico): **contare gli eventi**.
- Per verificare se un evento si verifica in due o più gruppi/categorie con la stessa frequenza, o con la frequenza prevista da un modello teorico, si utilizza il test chi-quadrato o una delle sue varianti, il test di Fisher o il test di McNemar.
- Vediamo un esempio di test χ^2:

  La frequenza attesa di nuovi nati di sesso maschile e di sesso femminile è pari a 0.5 per entrambi i sessi. Tra i cariotipi eseguiti su liquido amniotico per diagnosi prenatale nell'arco di due mesi se ne sono osservati 487 di tipo maschile (XY) e 503 di tipo femmine (XX): il numero di casi osservati è in linea con la frequenza attesa?

- Eseguiamo gli opportuni calcoli come di seguito indicato.

(Bio)Statistica con R — Parte II

Test chi-quadrato (χ^2)

```
> casi.osservati <- c(487,503) # inserisco i casi osservati
> freq.attese <- c(0.5,0.5) # inserisco le frequenze attese
# Calcolo il chi-quadrato e lo salvo nell'oggetto Chiquad
> Chiquad <- chisq.test(casi.osservati, p = freq.attese)
> Chiquad$observed # mostra le frequenze osservate
> Chiquad$expected # mostra le frequenze attese
> Chiquad # mostra i risultati del test chi-quadrato
```

- Nella console di **R** sono mostrate le frequenze osservate (**Chiquad$observed**), quindi le frequenze attese (**Chiquad$expected**) e infine viene mostrato il risultato del test chi-quadrato (**Chiquad**):

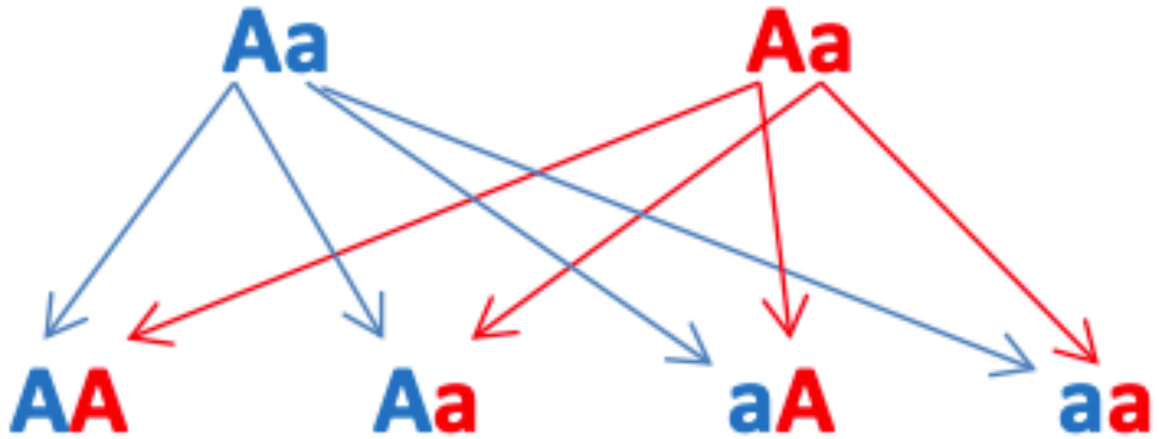
```
Chi-squared test for given probabilities
data: casi.osservati
X-squared = 0.2586, df = 1, p-value = 0.6111
```

- Il valore di *p*, che indica la probabilità di osservare per caso una differenza quale quella effettivamente osservata (487 maschi e 503 femmine contro una frequenza attesa di 495 e 495 rispettivamente), è 0.6111: pertanto dobbiamo ritenere che la differenza osservata sia presumibilmente legata al caso, ovvero statisticamente "non significativa".

(Bio)Statistica con R — Parte II

Test chi-quadrato (χ^2)

- Vediamo un altro esempio. Per l'eredità di un carattere autosomico recessivo presente in entrambi i genitori (padre **Aa** e madre **Aa**) le quattro possibili combinazioni sono:



e hanno frequenza pari a 0.25 (**AA**), 0.50 (**Aa** / **Aa**), 0.25 (**aa**). Da genitori entrambi **Aa** sono nati 85 figli **AA**, 173 **Aa** e 94 **aa**.

- *Il numero di casi osservati è in linea con la frequenza attesa?*

(Bio)Statistica con R — Parte II

Test chi-quadrato (χ^2)

```
> casi.osservati <- c(85,173,94) # inserisco i casi osservati
> freq.attese <- c(0.25,0.50,0.25) # inserisco le frequenze attese
# Calcolo il chi-quadrato e lo salvo nell'oggetto Chiquad
> Chiquad <- chisq.test(casi.osservati, p = freq.attese)
> Chiquad$observed # esamino le frequenze osservate
> Chiquad$expected # esamino le frequenze attese
> Chiquad # esamino i risultati del test chi-quadrato
Chi-squared test for given probabilities
data: casi.osservati
X-squared = 0.5625, df = 2, p-value = 0.7548
```

- Il valore di *p*, che indica la probabilità di osservare per caso una differenza quale quella effettivamente osservata (85 omozigoti **AA** osservati contro 88 attesi, 173 eterozigoti **Aa** osservati contro 176 attesi, 94 omozigoti **aa** osservati contro 88 attesi) in questo caso è 0.7548.
- Anche questa volta la probabilità di osservare per caso una differenza quale quella effettivamente osservata è molto elevata, al punto che dobbiamo ritenere la differenza osservata presumibilmente legata al caso, ovvero statisticamente "non significativa".

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di Fisher e di McNemar

- Il principio che vale per le tabelle nelle quali le osservazioni sono organizzate in 2 righe e 2 colonne è il seguente:

  → si utilizza il test **chi-quadrato** quando le osservazioni sono numerose e indipendenti

  → si utilizza il test di **Fisher** quando sono le osservazioni non sono numerose

  → si utilizza il test di **McNemar** quando le osservazioni non sono indipendenti (dati appaiati)

Osservazioni:	numerose e indipendenti	non numerose	non indipendenti
TEST	**chi-quadrato**	**Fisher**	**McNemar**

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test **chi-quadrato**, di Fisher e di McNemar

- Vediamo un esempio. Applichiamo il test **chi-quadrato** ai dati relativi alla presenza o assenza di emolisi, rispetto ad un valore soglia prefissato, utilizzando due sistemi di prelievo (A e B), in situazioni di estrema difficoltà del prelievo venoso.
- Impiegando il sistema A si sono osservati 68 casi di emolisi su 109 prelievi (pari al 62.4%); con il sistema B si sono osservati 93 casi di emolisi su 125 prelievi (pari al 74.4%).
- *Ci si chiede se il numero di casi di emolisi osservati con l'uno e l'altro sistema di prelievo sia significativamente diverso.* Le osservazioni riguardano in totale 234 differenti prelievi e altrettanti campioni di sangue, sono numerose e sono indipendenti.
- I dati sono stati raccolti in questa tabella:

	Sì emolisi	**No emolisi**
Sistema A	41	68
Sistema B	32	93

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test **chi-quadrato**, di Fisher e di McNemar

```
# importo i dati, dopo averli scaricati da questo link
> mydata <- read.table("Chiquad_2x2.csv", header=TRUE, sep=";", row.names="Prelievo")
# test chi quadrato con la correzione di Yates
> chisq.test(mydata, correct=TRUE)
# test chi quadrato senza la correzione di Yates
> chisq.test(mydata, correct=FALSE)
```

- Il test chi-quadrato con 1 grado di libertà è esatto solo asintoticamente per dimensioni molto grandi dei campioni.

  Pertanto nell'esempio sopra riportato viene applicata la correzione di Yates per la continuità (correct = TRUE) ottenendo $p = 0.06615$ (differenza non significativa):

```
Pearson's Chi-squared test with Yates' continuity correction
data:  mydata
X-squared = 3.3761, df = 1, p-value = 0.06615
```

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di **Fisher** e di McNemar

- Da notare che senza la correzione di Yates per la continuità (**correct = FALSE**) si potrebbe pensare ad una differenza significativa ($p = 0.04783$):

```
Pearson's Chi-squared test
data:  mydata
X-squared = 3.9159, df = 1, p-value = 0.04783
```

- In genere si consiglia di utilizzare il **test di Fisher** quando:

  → il totale delle osservazioni è inferiore a 20 *oppure*

  → in una delle celle abbiamo un valore osservato inferiore a 10 *oppure*

  → in una delle celle abbiamo un valore atteso inferiore a 5

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di **Fisher** e di McNemar

- Vediamo un esempio di applicazione del test di Fisher:

  Nel caso della valutazione di un test diagnostico per una malattia rara è stato possibile reclutare solamente 7 malati. Sono conseguentemente stati reclutati altrettanti soggetti sani di controllo. IL test risultava positivo in 3 malati e negativo in 4 malati. Era invece negativo in 6 e positivo solamente in 1 dei soggetti sani.

- Il numero di osservazioni era quindi molto ridotto e ricorrevano addirittura tutte e tre le condizioni per l'utilizzo del test di Fisher indicate in precedenza.

- I dati sono stati raccolti in questa tabella (che si può scaricare da **qui**):

	Sano	**Malato**
Test positivo	1	3
Test negativo	6	4

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di **Fisher** e di McNemar

- Vediamo un esempio di applicazione del test di Fisher.

```
# importo i dati
> mydata <- read.table("Fisher_2x2.csv", header=TRUE,
    sep=";", row.names="Esito")
# eseguo il test di Fisher
> fisher.test(mydata)
Fisher's Exact Test for Count Data
data:  mydata
p-value = 0.5594
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.003646918 4.442542966
sample estimates:
odds ratio
 0.2480182
```

Il valore di $p = 0.5594$ conferma l'esistenza di una differenza non significativa.

In questo caso viene testata anche l'ipotesi che l'odds ratio sia diverso da 1. Se i limiti di confidenza dell'odds ratio includono il valore 1, la differenza non è significativa.

Qui abbiamo un odds ratio di 0.2480182, i cui limiti di confidenza al 95% vanno da 0.003646918 a 4.442542966 e quindi includono il valore 1.

La conclusione è evidente: se questa è l'incertezza delle nostre conclusioni, incertezza che include il valore 1, dobbiamo dedurne ancora una volta che la differenza tra gli esiti osservata non è significativa.

Va da sé che trarre conclusioni da un numero così ridotto di osservazioni è comunque una pratica possibilmente da evitare.

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di Fisher e di **McNemar**

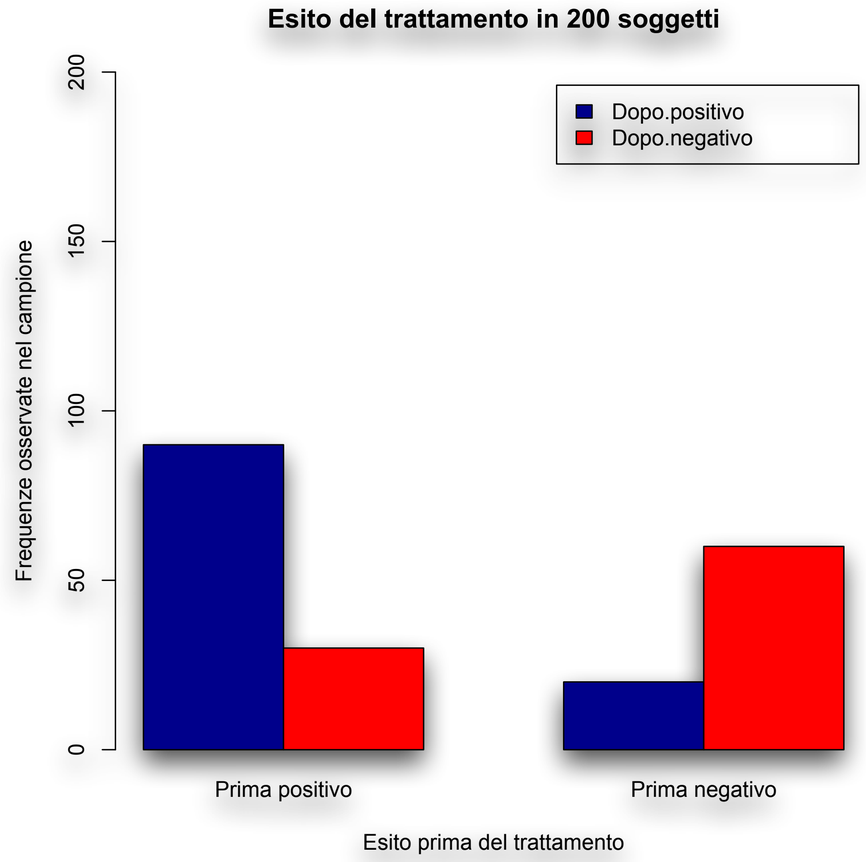
- Il test di **McNemar** viene applicato quando le osservazioni non sono indipendenti (*paired*). Un caso tipico è quello di 200 pazienti quali prima di un trattamento è stato effettuato un test diagnostico, che poteva risultare positivo o negativo. Tutti i 200 pazienti sono stati successivamente sottoposti ad un trattamento terapeutico al termine del quale il test è stato ripetuto. Dei 120 pazienti cui il test è risultato positivo prima del trattamento, dopo il trattamento 90 sono risultati positivi al test e 30 sono risultati negativi al test. Degli 80 pazienti cui il test è risultato negativo prima del trattamento, dopo il trattamento 20 sono risultati positivi al test e 60 sono risultati negativi al test.
- I dati sono stati raccolti in questa tabella (che può essere scaricata da questo **link**):

Esito	**Dopo positivo**	**Dopo negativo**
Prima positivo	90	30
Prima negativo	20	60

(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di Fisher e di **McNemar**

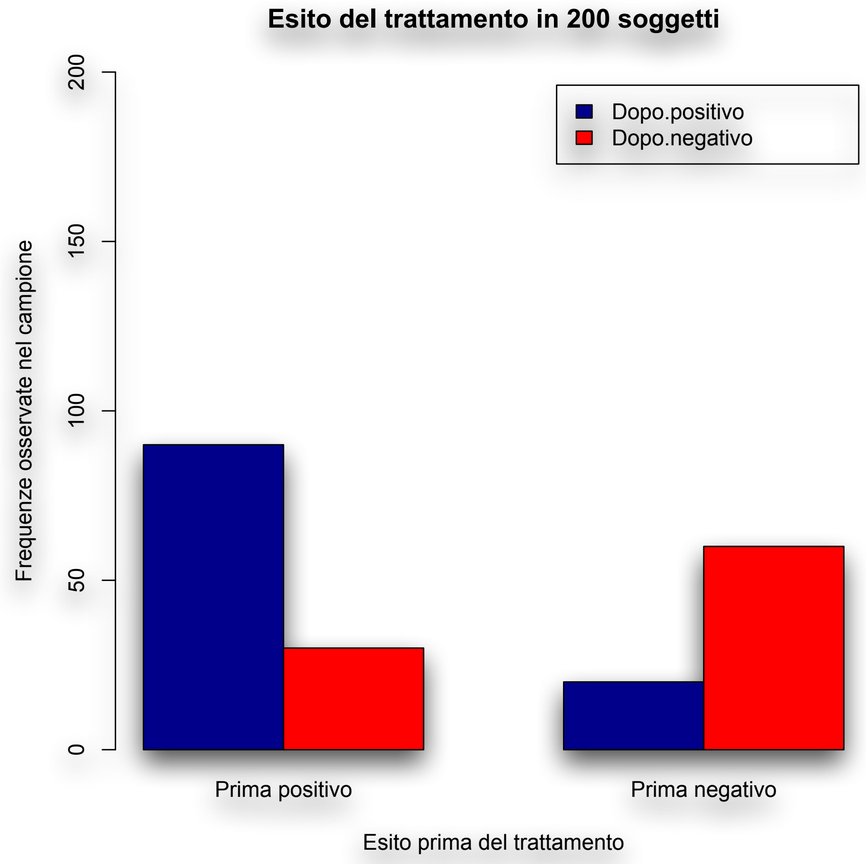
```
# importo i dati
> mydata <- read.table("McNemar_2x2.csv", header=TRUE, sep=";", row.names="Esito")
# genero un oggetto matrice contenente i dati per esigenze delle funzioni successive
> matrix <- data.matrix(mydata)
# eseguo il test di McNemar
> mcnemar.test(matrix, correct=TRUE)
  McNemar's Chi-squared test with continuity correction
data:  matrix
McNemar's chi-squared = 1.62, df = 1, p-value = 0.2031
# eseguo l'analisi dei dati viene mediante un grafico a barre
> barplot(t(matrix), beside=TRUE, legend=TRUE, ylim=c(0,200),
    col=c("darkblue","red"),
    ylab="Frequenze osservate nel campione",
    xlab="Esito prima del trattamento",
    main="Esito del trattamento in 200 soggetti")
```



(Bio)Statistica con R — Parte II

Test per una tabella 2×2: test chi-quadrato, di Fisher e di **McNemar**

- La rappresentazione dei dati sotto forma di un grafico a barre ci aiuta nella sintesi e nella interpretazione dei risultati del test di McNemar (applicato in quanto si trattava di dati appaiati).
- Nel grafico a barre possiamo osservare l'esito del test, positivo o negativo, negli stessi 200 soggetti prima e dopo uno specifico trattamento).
- Il valore di $p = 0.2031$ conferma l'esistenza di una differenza non significativa.



(Bio)Statistica con R — Parte II

Test chi-quadrato per tabelle di contingenza r×c

- Il test chi-quadrato per una tabella di contingenza di **r** righe × **c** colonne è la forma più generalizzata del test. Vediamo un esempio.
- Cinque terreni di coltura selettivi per lo *Streptococcus pyogenes* sono stati provati al fine di valutare la loro capacità di fornire un isolamento selettivo delle colonie dopo la semina di un tampone rinofaringeo.
- L'esito di ciascuna prova è stato registrato, e i risultati sono stati raccolti in questa tabella (che è possibile scaricare da questo **link**):

Esito	Terreno A	Terreno B	Terreno C	Terreno D	Terreno E
No isolamento	39	44	20	41	42
Sì isolamento	177	166	200	183	168

(Bio)Statistica con R — Parte II

Test chi-quadrato per tabelle di contingenza r×c

```
# importo i dati
> mydata <- read.table("Chiquad_rxc.csv", header=TRUE, sep=";", row.names="Esito")
# calcolo il chi-quadrato
> chisq.test(mydata)
# ricalcolo il p-value mediante una simulazione Monte Carlo con un milione di replicati
> chisq.test(mydata, simulate.p.value = TRUE, B = 1000000)
```

- Dopo avere importato i dati, abbiamo calcolato il test con il valore di *p* determinato a partire dalla distribuzione teorica di chi-quadrato:

```
        Pearson's Chi-squared test
data:  mydata
X-squared = 13.6785, df = 4, p-value = 0.008395
```

- Infine, abbiamo nuovamente calcolato il test chi-quadrato valutando il *p*-value con il metodo Monte Carlo (simulate.p.value = TRUE) utilizzando un milione di repliche (B = 1000000).

(Bio)Statistica con R — Parte II

Test chi-quadrato per tabelle di contingenza r×c

- Il valore di *p* corrispondente alla statistica chi-quadrato (χ^2) rappresenta la probabilità di osservare per caso una differenza tra frequenze osservate e frequenze attese della grandezza di quella effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una differenza significativa di incidenza nei diversi gruppi dell'esito della prova (isolamento si / isolamento no).
- In questo caso la probabilità di osservare per caso una differenza tra frequenze osservate e frequenze attese della grandezza di quella effettivamente osservata è dell'ordine dell'8 per mille (0.008293).
- Possiamo concludere che la differenza non sia presumibilmente dovuta al caso, ovvero che sia statisticamente significativa.

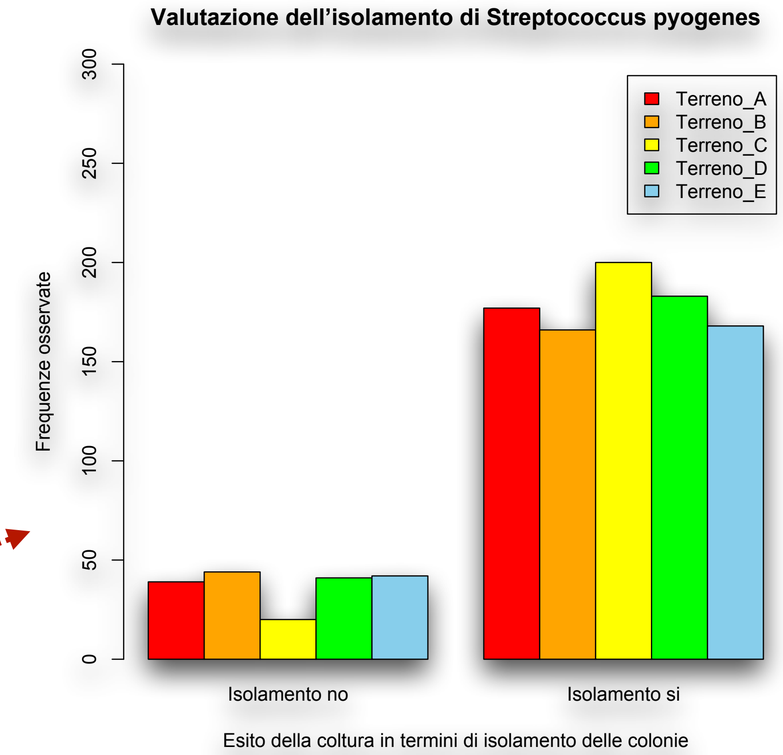
Ma a quale terreno va imputata la differenza osservata?

(Bio)Statistica con R — Parte II

Test chi-quadrato per tabelle di contingenza r×c

- In casi di questo genere può essere utile integrare il risultato numerico con una rappresentazione grafica dei dati utilizzati per calcolare il chi-quadrato.
- Con il codice **R** che segue proviamo a farlo sotto forma di un grafico a barre:

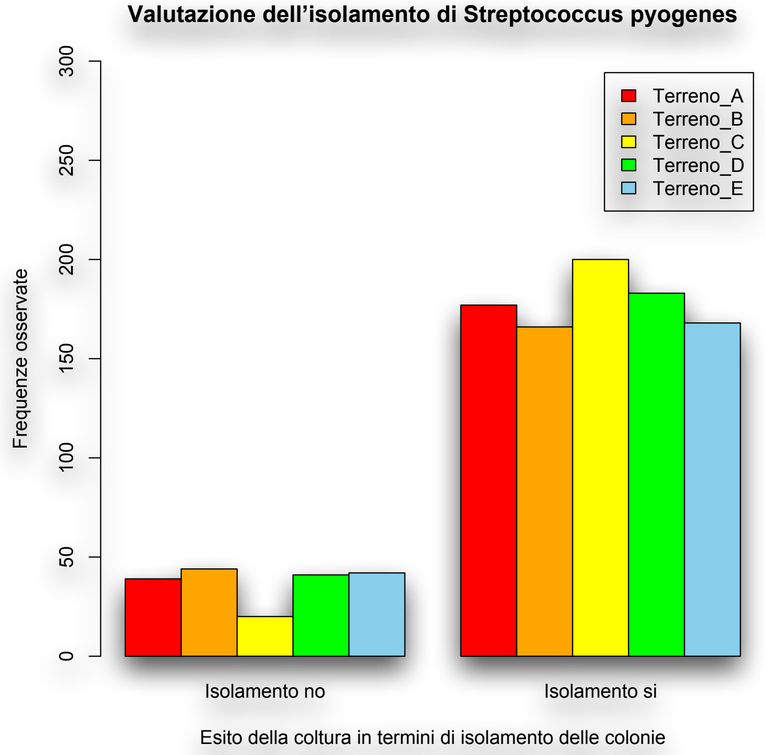
```
# importo i dati
> mydata <- read.table("Chiquad_rxc.csv", header=TRUE, sep=“;",
      row.names="Esito")
# genero un oggetto matrice contenente i dati per il successivo utilizzo
> matrix <- data.matrix(mydata)
# eseguo l’analisi dei dati con un grafico a barre utile per una sintesi dei risultati
> barplot(t(matrix), beside=TRUE, legend=TRUE, ylim=c(0,300),
      col=c("red","orange", "yellow", "green", "skyblue"), ylab="Frequenze osservate",
      xlab="Esito della coltura in termini di isolamento delle colonie",
      main="Valutazione dell’isolamento di Streptococcus pyogenes")
```



(Bio)Statistica con R — Parte II

Test chi-quadrato per tabelle di contingenza r×c

- In effetti il grafico a barre ci aiuta a individuare nel terreno C (in giallo) quello che consente di avere il migliore isolamento delle colonie.
- La conferma numerica del dato è ulteriormente suffragata da una tabella nella quale sono stati calcolati i valori percentuali di successo (isolamento Sì) e di insuccesso (isolamento No) per ciascun terreno:



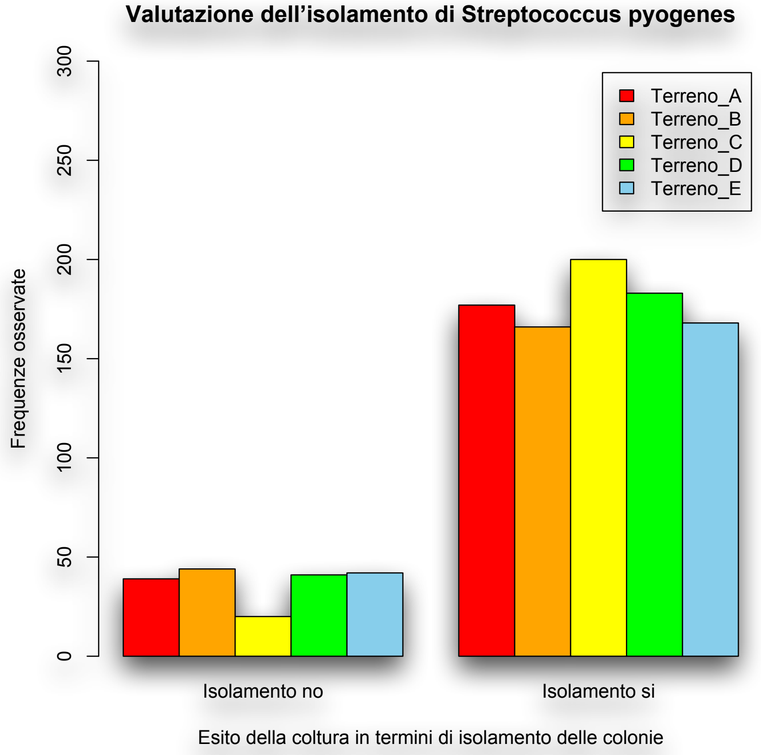
Esito	Terreno A	Terreno B	Terreno C	Terreno D	Terreno E
No isolamento	18.1	21.0	9.1	18.3	20.0
Sì isolamento	81.9	79.0	90.9	81.7	80.0

(Bio)Statistica con R — Parte II

Test chi-quadrato per tabelle di contingenza r×c

- La conclusione è che il terreno di coltura C fornisce una percentuale di isolamento delle colonie del 91% circa, che risulta quindi essere la migliore rispetto a quella di tutti gli altri terreni, e che con lo stesso terreno di coltura C si osserva una percentuale di insuccessi del 9% circa, che risulta inferiore a quella di tutti gli altri terreni.

Esito	Terreno A	Terreno B	Terreno C	Terreno D	Terreno E
No isolamento	18.1	21.0	9.1	18.3	20.0
Sì isolamento	81.9	79.0	90.9	81.7	80.0



(Bio)Statistica con R — Parte II

Verifica della normalità dei dati

- Scarichiamo e salviamo nella directory di lavoro il file **Statelem.csv**. Nella prima riga sono riportati i nomi delle variabili, nelle successive i dati, relativi a quasi settemila casi per quali erano disponibili sesso, età, e i valori di colesterolo, colesterolo HDL, colesterolo LDL e trigliceridi (nel siero, in mg/dL).
- Come si vede alcuni dati possono mancare (nel secondo caso mancano i risultati di colesterolo LDL e trigliceridi, nel terzo caso manca il risultato del colesterolo LDL, e altri valori mancano nei casi successivi):

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

(Bio)Statistica con R — Parte II

Verifica della normalità dei dati

- Nel codice che segue viene analizzata la variabile *Colesterolo*.
- Da notare che viene utilizzata la libreria **nortest** che, se non ancora fatto, va scaricata dal CRAN (in caso contrario si verificherà un errore nell'esecuzione del codice laddove è previsto l'utilizzo della libreria).

```
> mydata <- read.table("Statelem.csv", header=TRUE, sep=";") # importo i dati
> names(mydata) # visualizzo i nomi delle variabili contenute in mydata
> str(mydata) # struttura dell'oggetto mydata
> head(mydata, n=10) # lista dei primi 10 casi di mydata
> tail(mydata, n=5) # lista degli ultimi 5 casi di mydata
> mydata[!complete.cases(mydata),] # mostra i casi con dati NA (Not Available)
> newdata <- na.omit(mydata) # crea un nuovo oggetto denominato newdata omettendo i casi con dati NA
> newdataset <- newdata[c(2,3,4,5,6)] # crea un oggetto denominato newdataset che include solamente le colonne da 2 a 6 con le variabili quantitative
> str(newdataset) # struttura di newdataset
> avector <- newdataset[, "Colesterolo"] # crea un vettore che contiene i dati della variabile "Colesterolo"
> library(nortest) # apre la libreria nortest che contiene vari test di normalità
```

(Bio)Statistica con R — Parte II

Verifica della normalità dei dati

```
> ad.test(avector) # test di Anderson-Darling
> cvm.test(avector) # test di Cramer-von Mises
> lillie.test(avector) # test di Lilliefors (Kolmogorov-Smirnov)
> pearson.test(avector) # test chi-quadrato di Pearson
> sf.test(avector) # test di Shapiro-Francia
> hist(avector, main="Istogramma dei dati osservati", xlab="Colesterolo totale in mg/dL", ylab = "Frequenza") # traccia un istogramma dei dati
> plot(density(avector), main="Distribuzione di densità dei dati osservati", xlab="Colesterolo totale in mg/dL",
   ylab = "Frequenza") # traccia la distribuzione di densità dei dati
> plot(ecdf(avector), main="Distribuzione cumulativa empirica dei dati", xlab="Colesterolo totale in mg/dL",
   ylab = "Frequenza cumulativa") # traccia la distribuzione cumulativa empirica dei dati
# Traccio il grafico che mostra l'adeguatezza dei dati a una distribuzione normale
> zetavector <- (avector-mean(avector))/sd(avector) # calcola la deviata normale standardizzata
> qqnorm((zetavector), main="Quantili campionari vs. quantili teorici", xlab="Quantili teorici",
   ylab = "Quantili campionari") # grafico dei quantili campionari vs. quantili teorici
> abline (0,1) # retta a 45 gradi di riferimento
```

(Bio)Statistica con R — Parte II

Verifica della normalità dei dati

- I test di normalità concordano tutti sul fatto che i valori del *Colesterolo* non sono distribuiti in modo gaussiano.
- La probabilità di osservare per caso uno scostamento dalla distribuzione normale dell'entità di quello osservato per il colesterolo è molto bassa (*p* è variabile con i vari test, ma sempre almeno inferiore a `0.001`), quindi lo scostamento della distribuzione del colesterolo dalla distribuzione gaussiana è da ritenersi statisticamente significativo. Ecco il risultato dei singoli test:

```
> ad.test(avector)
        Anderson-Darling normality test
data:  avector
A = 2.0806, p-value = 2.733e-05
> cvm.test(avector)
        Cramer-von Mises normality test
data:  avector
W = 0.321, p-value = 0.0001917
> lillie.test(avector)
Lilliefors (Kolmogorov-Smirnov) normality test
data:  avector
D = 0.0287, p-value = 9.183e-05
```

```
> pearson.test(avector)
        Pearson chi-square normality test
data:  avector
P = 83.9086, p-value = 0.0001891
> sf.test(avector)
        Shapiro-Francia normality test
data:  avector
W = 0.9931, p-value = 8.518e-09
```

(Bio)Statistica con R — Parte II

Statistiche esplorative

- Nel codice che segue sono impiegate le librerie **Hmisc**, **pastecs** e **psych** che vanno scaricate con il comando **install.packages()**. Il seguente codice esegue alcune statistiche esplorative sul dataset **Statelem.csv**:

```
# importo i dati
> mydata <- read.table("Statelem.csv", header=TRUE, sep=";")
# nomi delle variabili contenute in mydata
> names(mydata)
# struttura dell'oggetto mydata
> str(mydata)
# lista dei primi 10 casi di mydata
> head(mydata, n=10)
# lista degli ultimi 5 casi di mydata
> tail(mydata, n=5)
# mostra i casi con dati NA (Not Available)
> mydata[!complete.cases(mydata),]
```

(Bio)Statistica con R — Parte II

Statistiche esplorative

```
# creo un nuovo oggetto denominato newdata omettendo i casi con dati NA
> newdata <- na.omit(mydata)
# creo un oggetto denominato newdataset che include solamente le colonne da 2 a 6 con le variabili quantitative
> newdataset <- newdata[c(2,3,4,5,6)]
# calcolo la media del colesterolo LDL su mydata
> attach(mydata) # così non è necessario anteporre il nome della variabile mydata
> mean(LDL) # da notare che R restituisce NA anche per la media
> mean(LDL, na.rm=TRUE) # rimuovendo i dati NA questa volta R restituisce il valore della media
```

(Bio)Statistica con R — Parte II

Statistiche esplorative

è ora facile calcolare le altre principali statistiche del colesterolo LDL su mydata

```
> var(LDL, na.rm=TRUE) # calcola la varianza
> sd(LDL, na.rm=TRUE) # calcola la deviazione standard
> min(LDL, na.rm=TRUE) # calcola il valore minimo
> max(LDL, na.rm=TRUE) # calcola il valore massimo
> range(LDL, na.rm=TRUE) # calcola il range
> quantile(LDL, probs = seq (0, 1, 0.25), na.rm=TRUE) # calcola i quartili; sostituendo il valore 0.25 con 0.10 calcola i decili, etc.
```

(Bio)Statistica con R — Parte II

Statistiche esplorative

Mediante la funzione *sapply()* si calcolano le statiche di tutte le variabili (questa volta su *newdataset*, nel quale sono inclusi solamente i dati completi, e dal quale è stata esclusa la variabile qualitativa *Sesso*)

```
> sapply(newdataset, mean) # calcola la media
> sapply(newdataset, sd) # calcola la deviazione standard
> sapply(newdataset, var) # calcola la varianza
> sapply(newdataset, min) # calcola il valore minimo
> sapply(newdataset, max) # calcola il valore massimo
> sapply(newdataset, range) # calcola il range
> sapply(newdataset, median) # calcola la mediana
> sapply(newdataset, quantile) # calcola i quartili
```

(Bio)Statistica con R — Parte II

Statistiche esplorative

```
# Se vogliamo avere rapidamente le statistiche sintetiche di mydata
> summary(mydata)
# Queste sono le statistiche che possiamo ottenere mediante la libreria Hmisc
> install.packages("Hmisc"); library(Hmisc)
> Hmisc::describe(mydata)  # per evitare ambiguità, esplicito la libreria che contiene la funzione "describe" che voglio eseguire
# Queste sono le statistiche che possiamo ottenere mediante la libreria pastecs
> install.packages("pastecs"); library(pastecs)
> pastecs::stat.desc(mydata)
# Queste infine sono le statistiche che possiamo ottenere mediante la libreria psych
> install.packages("psych"); library(psych)
> psych::describe(mydata)
```

(Bio)Statistica con R — Parte II

Statistiche esplorative

- Le librerie **Hmisc**, **pastecs** e **psych** forniscono automaticamente un riepilogo complessivo delle principali statistiche di tutte le variabili. Da notare che nel riepilogo fornito dalla libreria **psych** la colonna *mad* riporta il valore della "median absolute deviation", cioè la mediana delle deviazioni assolute dalla mediana, che è l'equivalente non parametrico della deviazione standard:

	var	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
Sesso*	1	6787	1.43	0.49	1.0	1.41	0.00	1	2	1	0.30	-1.91	0.01
Eta	2	6787	59.01	17.66	62.0	59.89	19.27	3	104	101	-0.41	-0.45	0.21
Colesterolo	3	6787	208.74	42.33	207.0	207.85	41.51	56	435	379	0.27	0.44	0.51
HDL	4	5918	62.15	17.41	60.0	61.19	17.79	5	146	141	0.55	0.38	0.23
LDL	5	2874	125.17	34.04	123.5	124.22	34.84	9	292	283	0.33	0.29	0.64
Trigliceridi	6	5625	117.88	73.88	102.0	106.91	45.96	19	1248	1229	4.11	33.09	0.99

- In una distribuzione perfettamente gaussiana la media coincide con la mediana, e la deviazione standard coincide con la mad. Tanto più i dati si discostano da una distribuzione gaussiana, tanto più la media differisce dalla mediana e/o tanto più la deviazione standard differisce dalla *mad*: questo avviene nel caso dei *Trigliceridi*, che abbiamo visto in precedenza essere distribuiti in modo grossolanamente non gaussiano. Il fatto che si tratti di una distribuzione non gaussiana è confermato anche dal valore del coefficiente di asimmetria (*skew*) e del coefficiente di curtosi (*kurtosis*).

(Bio)Statistica con R — Parte II

CONFRONTO TRA CAMPIONI

- Il confronto tra campioni può essere effettuato sia nel caso di dati indipendenti che nel caso di dati appaiati.
- Accanto alla versione tradizionale parametrica, rappresentata dal test *t* di Student, esistono gli equivalenti non parametrici, che devono essere utilizzati se i dati non sono distribuiti in modo normale.
- Quindi anche nel caso del confronto tra campioni deve essere effettuata una analisi preliminare dei dati per decidere quale sia il test appropriato.
- Per gli esempi seguenti utilizzeremo i dataset **Statind.csv** (confronto tra due campioni indipendenti) e **Statapp.csv** (confronto tra dati appaiati).
- Fare click sul nome per scaricare ciascun file.

(Bio)Statistica con R — Parte II

Confronto tra due campioni indipendenti (test *t* di Student e test non parametrici)

- Nel primo esempio utilizzeremo il file **Statind.csv** contenente dati relativi alla determinazione della concentrazione della riboflavina in due tessuti, il fegato e il muscolo.
- Si tratta ovviamente di campioni indipendenti.
- Nella prima riga sono riportati i nomi delle variabili, nelle successive i dati.
- La variabile *Tessuto* della prima colonna è la variabile classificativa che specifica il tipo di tessuto nel quale è stata determinata la concentrazione di riboflavina, che a sua volta è riportata poi nella variabile numerica *Riboflavina* della seconda colonna.
- Di seguito il codice **R** per effettuare il confronto tra i due campioni.

Tessuto	Riboflavina
Fegato	0.95
Fegato	2.18
Fegato	1.12
Fegato	1.86
Muscolo	0.22
Muscolo	0.18
Muscolo	0.46
Muscolo	0.86
Muscolo	0.64
Muscolo	0.28
Muscolo	0.33
Muscolo	0.35
Muscolo	0.42

(Bio)Statistica con R — Parte II

Confronto tra due campioni indipendenti (test *t* di Student e test non parametrici)

```
# Importo i dati
> mydata <- read.table("Statind.csv", header=TRUE, sep=";")
# Eseguo l'analisi della varianza (test F per il rapporto tra varianze) per verificare se le varianze delle misure effettuate nel tessuto e nel muscolo sono omogenee
> attach(mydata)
> var.test(Riboflavina~Tessuto)  # p-value = 0.02156: le varianze differiscono significativamente!
```

- In effetti l'analisi delle varianze con un p = 0.02156 indica una varianza significativamente differente tra i risultati ottenuti nel tessuto e quelli ottenuti nel muscolo.
- La variabile Riboflavina non è distribuita normalmente (basta verificarlo ad es. con il comando: `shapiro.test(Riboflavina)` che dà un *p*-value = 0.01266 (l'ipotesi nulla che la distribuzione sia normale è rigettata).
- Possiamo utilizzare il test parametrico *t* di Student per campioni indipendenti (Welch), che confronta le medie, solo se specifichiamo che le varianze differiscono significativamente: `t.test(Riboflavina~Tessuto, var.equal=FALSE)`
- Il test produce un p-value = 0.02867, quindi la probabilità di osservare per caso una differenze tra le medie è pari al 2,8% circa. Possiamo quindi concludere che le medie sono significativamente diverse: si osserva nel fegato una concentrazione di riboflavina maggiore di quella osservata nel muscolo.

(Bio)Statistica con R — Parte II

Confronto tra due campioni indipendenti (test *t* di Student e test non parametrici)

- Essendo comunque la distribuzione della Riboflavina non normale, conviene utilizzare test non parametrici, che non sono sensibili alle differenze tra le varianze nei due campioni e non necessitano quindi la correzione che si rende invece necessaria per il test *t* di Student.
- Se utilizziamo il test "Wilcoxon Rank" per campioni indipendenti:

  `> wilcox.test(Riboflavina~Tessuto)`

  otteniamo un *p*-value = 0.002797.
- Se invece applichiamo il test "Kruskal-Wallis":

  `> kruskal.test(Riboflavina~Tessuto)`

  otteniamo un *p*-value = 0.005479.
- In ogni caso sia il test *t* di Student che i due test non parametrici consentono di concludere che esiste una differenza significativa nella concentrazione di riboflavina nei due tessuti.

(Bio)Statistica con R — Parte II

Confronto tra dati appaiati (test *t* di Student e test non parametrici)

- In questo secondo esempio utilizzeremo il file **Statapp.csv** nel quale la variabile *Subito* della prima colonna riporta i valori di aspartato-aminotransaminasi (AST, in U/L) misurati su un campione di siero immediatamento dopo il prelievo, mentre la variabile *Dopo24ore* della seconda colonna riporta i valori determinati sullo stesso campione di siero dopo 24 ore di conservazione dei campioni, tappati per evitare fenomeni di evaporazione, e conservati alla temperatura di + 4 °C.
- Eseguiamo analogamente i test già visti in precedenza:

```
# importo i dati (che non sono distribuiti normalmente)
> mydata <- read.table("Statapp.csv", header=TRUE, sep=";"); attach(mydata)
# test t di Student per dati appaiati
> t.test(Subito, Dopo24ore, paired=TRUE)
# in alternativa si può applicare il test di Wilcoxon (Wilcoxon Signed Rank Test) per dati appaiati (un test non parametrico)
> wilcox.test(Subito, Dopo24ore, paired=TRUE, exact=FALSE)
```

Subito	Dopo24ore
17	16
18	17
19	24
20	21
22	24
24	25
24	27
30	25
37	42
42	40
45	48
52	57
62	60
67	71
95	86
101	106
174	180
327	300
433	440
476	430
495	515
652	631

(Bio)Statistica con R — Parte II

Confronto tra dati appaiati (test *t* di Student e test non parametrici)

- Sia il test *t* di Student per dati appaiati con un valore di $p = 0.4718$ sia il test di Wilcoxon con un valore di $p = 0.6255$ consentono di concludere che la concentrazione dell'AST misurata nel siero conservato per 24 ore alla temperatura di +4 °C non differisce significativamente da quella misurata sul siero immediatamente dopo il prelievo.
- Anche in questo caso la scelta tra test parametrico (test *t* di Student) e non parametrico (test di Wilcoxon) può essere fatta mediante analisi della normalità dei dati. In **R** è facile aggiungere ai dati originari una colonna con una nuova variabile, la *Differenza* (presa con il segno) tra la concentrazione trovata immediatamente dopo il prelievo e quella osservata dopo 24 ore:

```
> mydata = cbind(mydata, mydata$Subito-mydata$Dopo24ore)
> colnames(mydata)[3] = "Differenza"
```

- Se effettuiamo un test di normalità sulla variabile *Differenza* vediamo che essa non è distribuita in modo normale (p=0.0002475 con il test di Shapiro-Wilk).
- Pertanto in questo caso è necessario che le conclusioni siano tratte sulla base del risultato ottenuto con il test di Wilcoxon (test non parametrico).

(Bio)Statistica con R — Parte II

Confronto con una media teorica (test *t* di Student)

- In laboratorio si è preparata mediante pesata e diluizione una soluzione con una concentrazione di 8.0 g/dL di albumina umana. Definiamo questo valore come la "media teorica" della concentrazione dell'albumina in quanto, nelle opportune condizioni, la misura della massa del soluto con una bilancia e la misura del volume della soluzione finale con vetreria tarata di classe A consentono di avere una concentrazione finale nota con una accuratezza che è di alcuni ordini di grandezza superiore a quella di un comune metodo analitico.
- Analizzando il siero con il metodo analitico del quale si intende verificare l'accuratezza, si ottengono i seguenti valori: 8.2, 8.3, 7.9, 8.1 e 8.0 g/dL. Quello che segue è il codice **R** necessario per sapere se i risultati ottenuti si discostano significativamente dal valore assegnato:

```
# inserisco direttamente i valori ottenuti in un vettore
> myvalues <- c(8.2, 8.3, 7.9, 8.1, 8.0)
# confronto i cinque valori ottenuti con il valore assegnato di 8.0 g/dL mediante il test t di Student per una media teorica
> t.test(myvalues, mu = 8)
```

(Bio)Statistica con R — Parte II

Confronto con una media teorica (test *t* di Student)

- I risultati sono:

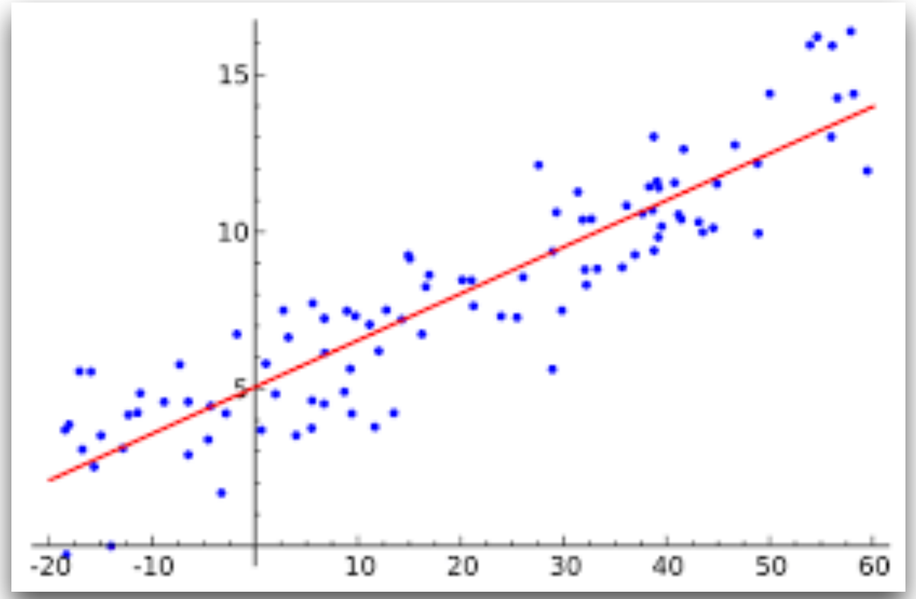
```
        One Sample t-test
data:  myvalues
t = 1.4142, df = 4, p-value = 0.2302
alternative hypothesis: true mean is not equal to 8
95 percent confidence interval:
 7.903676 8.296324
sample estimates:
mean of x
8.1
```

- La media misurata di 8.1 non differisce significativamente dalla media teorica essendo p = 0.2302. La probabilità di osservare per caso una differenza di 0.1 tra il valore teorico e il valore misurato è del 23% circa, troppo elevata per escludere il caso dalle possibili cause della differenza.
- Il dato viene confermato dal fatto che i limiti di confidenza al 95% della media (uguale a 8.1) delle cinque misure effettuate sono rispettivamente 7.903676 (il limite inferiore) e 8.296324 (il limite superiore) e pertanto includono nell'incertezza delle conclusioni il valore 8 della media teorica.

(Bio)Statistica con R — Parte II

REGRESSIONE LINEARE

- La regressione lineare viene tradizionalmente associata al coefficiente di correlazione lineare *r* di Pearson.
- Nonostante quest'ultimo abbia in sé un significato limitato, la possibilità che si ha con **R** di sviluppare una matrice dei coefficienti di correlazione tra più variabili e di generare matrici di scatterplot, che altro non sono che diagrammi cartesiani multipli che illustrano graficamente le relazioni tra dette variabili, rappresenta uno strumento utile per l'**analisi esplorativa dei dati**.



(Bio)Statistica con R — Parte II

Correlazione (coefficiente di correlazione lineare r)

- Scarichiamo e salviamo nella directory di lavoro il file **Statcorr.csv**. I dati contenuti hanno una struttura molto semplice:

643 righe (casi) →

GR	RGO	HB	HCT	HBA2	MCV	HBF	MCH	RDW	FERRO
4.90	97	13.3	40.6	1.8	82.8	0.6	27.1	17.3	106
4.66	81	10.8	34.3	2.6	73.6	1.6	23.2	21.5	148
5.43	57	11.5	36.1	4.8	66.5	2.5	21.1	21.0	104
5.41	63	10.8	39.7	2.5	73.4	1.8	20.0	19.9	74
4.94	60	10.4	32.3	1.4	65.0	0.7	21.1	23.7	17
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

- Le variabili contenute nel file sono gli eritrociti (GR, in 10^12/L), la resistenza globulare osmotica (RGO, in %), l'emoglobina (HB, in g/dL), l'ematocrito (in %), l'emoglobina A2 (in %), il volume globulare medio (MCV, in fL), l'emoglobina F (in %), l'emoglobina corpuscolare media (MCH, in pg), l'ampiezza della distribuzione dei globuli rossi (Red Distribution Width, in %) misurati in 643 soggetti che includevano soggetti sani e soggetti con beta-talassemia, con alfa-talassemia, con anemia sideropenica.
- Utilizzeremo le librerie **Hmisc** e **car** che, se non lo avete ancora fatto, dovete scaricare dal CRAN prima di eseguire l'esempio (in caso contrario si verificherà un errore nell'esecuzione del codice laddove è previsto l'utilizzo delle librerie).

(Bio)Statistica con R — Parte II

Correlazione (coefficiente di correlazione lineare r)

- Eseguiamo nella console di **R** il seguente codice per il calcolo dei coefficienti di correlazione (anche con i relativi livelli di significatività):

```
# importo i dati
> mydata <- read.table("Statcorr.csv", header=TRUE, sep=";")
# calcolo la matrice dei coefficienti di correlazione; il parametro method può essere "pearson" (il classico r), "spearman", "kendall"
> cor(mydata, use="complete.obs", method="pearson")
# calcolo i coefficienti di correlazione con i livelli di significatività
> library(Hmisc)
> x <- as.matrix(mydata) # trasformo mydata in una matrice denominata x
> rcorr(x, type="pearson") # il parametro type può essere "pearson" (il classico r) o "spearman"
```

(Bio)Statistica con R — Parte II

Correlazione (coefficiente di correlazione lineare r)

- Innanzitutto vengono calcolati i coefficienti di correlazione *r* tra tutte le possibili combinazioni di variabili (la matrice di correlazione); la diagonale divide la matrice in due parti simmetriche. I valori del coefficiente di correlazione sulla diagonale sono ovviamente tutti uguali esattamente a 1, in quanto rappresentano la correlazione di ciascuna variabile con sé stessa:

	GR	RGO	HB	HCT	HBA2	MCV	HBF	MCH	RDW	FERRO
GR	1.00	-0.42	0.40	0.48	0.54	-0.41	0.26	-0.42	0.37	0.27
RGO	-0.42	1.00	0.43	0.37	-0.63	0.74	-0.43	0.74	-0.61	-0.02
HB	0.40	0.43	1.00	0.97	-0.03	0.64	-0.13	0.65	-0.54	0.45
HCT	0.48	0.37	0.97	1.00	0.02	0.59	-0.10	0.56	-0.48	0.46
HBA2	0.54	-0.63	-0.03	0.02	1.00	-0.42	0.44	-0.43	0.26	0.30
MCV	-0.41	0.74	0.64	0.59	-0.42	1.00	-0.32	0.97	-0.85	0.26
HBF	0.26	-0.43	-0.13	-0.10	0.44	-0.32	1.00	-0.32	0.30	0.13
MCH	-0.42	0.74	0.65	0.56	-0.43	0.97	-0.32	1.00	-0.85	0.25
RDW	0.37	-0.61	-0.54	-0.48	0.26	-0.85	0.30	-0.85	1.00	-0.31
FERRO	0.27	-0.02	0.45	0.46	0.30	0.26	0.13	0.25	-0.31	1.00

n= 643

(Bio)Statistica con R — Parte II

Correlazione (coefficiente di correlazione lineare r)

- Quindi viene riportato il valore di probabilità *p* di osservare per caso il valore di *r* calcolato:

P

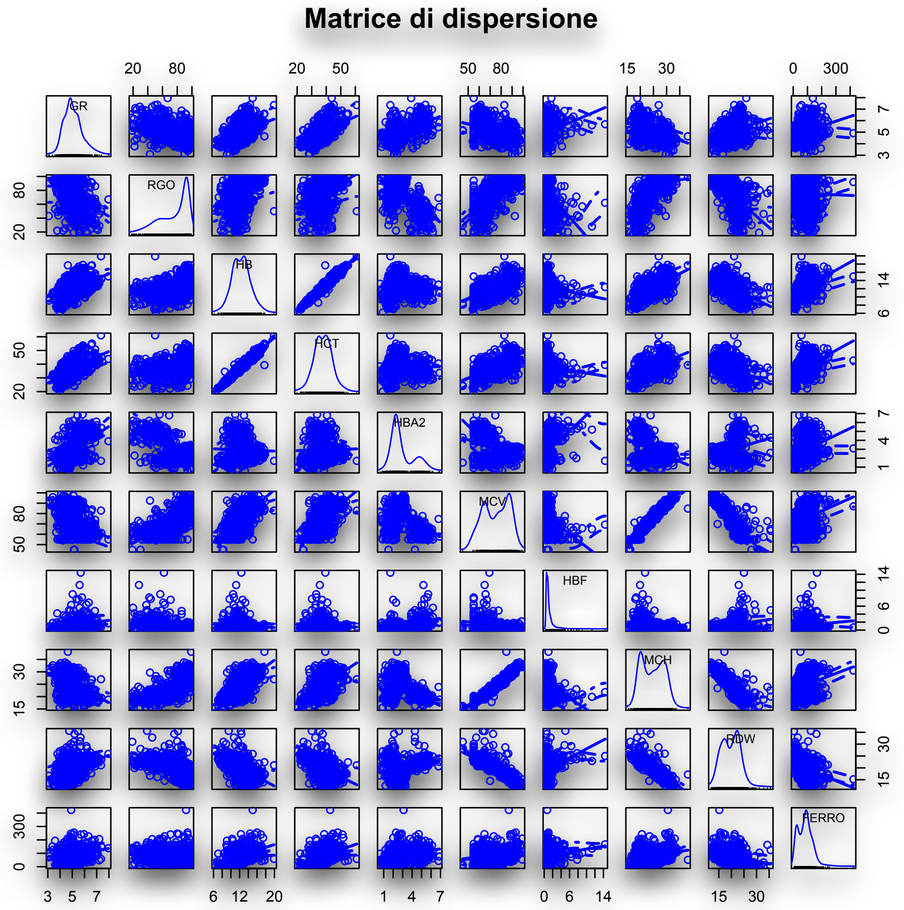
	GR	RGO	HB	HCT	HBA2	MCV	HBF	MCH	RDW	FERRO
GR		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
RGO	0.0000		0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.5911
HB	0.0000	0.0000		0.0000	0.4331	0.0000	0.0015	0.0000	0.0000	0.0000
HCT	0.0000	0.0000	0.0000		0.5489	0.0000	0.0140	0.0000	0.0000	0.0000
HBA2	0.0000	0.0000	0.4331	0.5489		0.0000	0.0000	0.0000	0.0000	0.0000
MCV	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000	0.0000	0.0000
HBF	0.0000	0.0000	0.0015	0.0140	0.0000	0.0000		0.0000	0.0000	0.0006
MCH	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000	0.0000
RDW	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000		0.0000
FERRO	0.0000	0.5911	0.0000	0.0000	0.0000	0.0000	0.0006	0.0000	0.0000	

(Bio)Statistica con R — Parte II

Correlazione (coefficiente di correlazione lineare r)

- Con la libreria **car** mediante una sola riga di codice possiamo rappresentare graficamente le relazioni tra le variabili:

```
> library(car)
> scatterplotMatrix(
    ~GR+RGO+HB+HCT+HBA2+MCV+HBF+MCH+RDW+FERRO,
    reg.line=lm, smooth=TRUE, span=0.5,
    diagonal="density",
    main="Matrice di dispersione",
    data=mydata)
```



- La matrice dei diagrammi di dispersione conferma le forti correlazioni esistenti tra HB e HCT e tra MCH e MCV.
- Questa parte relativa al coefficiente di correlazione può essere utilmente integrata con la parte nella quale r è trattato in forma grafica sotto forma di correlogrammi.

(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Scarichiamo e salviamo il file **Statreglin.csv**; i nomi delle variabili sono nella prima riga, i dati di ciascun caso nelle 2408 righe successive
- Nella prima riga sono riportati i nomi delle variabili, nelle successive i dati, relativi a oltre duemila casi per quali erano disponibili sesso, età, e i valori di colesterolo totale, colesterolo HDL, colesterolo LDL e trigliceridi (concentrazione nel siero, in mg/dL).
- Si tratta degli stessi dati utilizzati in precedenza per le statistiche elementari, dai quali sono stati questa volta esclusi tutti i casi con dati mancanti.
- Da notare che sono utilizzate le librerie **car**, **relaimpo** e **gvlma** che, se non lo avete ancora fatto, dovete scaricare dal CRAN prima di eseguire l'esempio (in caso contrario si verificherà un errore nell'esecuzione del codice laddove è previsto l'utilizzo delle librerie).

Sesso	Eta	Colesterolo	HDL	LDL	Trigliceridi
M	33	56	44	9	19
F	81	101	63	26	62
M	40	127	91	33	35
M	81	100	44	35	100
F	53	163	110	37	97
F	16	107	50	41	54
M	78	107	51	44	82
M	46	97	41	45	86
M	65	107	44	45	107
F	80	123	36	47	226
F	77	109	51	48	52
M	67	110	49	48	104
F	77	105	41	49	81
M	27	113	55	49	44
F	23	159	95	49	151
M	82	113	44	50	151
F	24	128	65	50	93
M	55	140	73	50	88
M	103	108	51	51	61
M	73	109	30	51	152
M	18	120	63	51	37

(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Eseguiamo nella console di **R** il seguente codice al fine di familiarizzare con questo tipo di analisi dei dati:

```
# Importo i dati
> mydata <- read.table("Statreglin.csv", header=TRUE, sep=";")
```

Calcolo la regressione lineare mediante la funzione lm(y~x_1+x_2+…, data=…); y è la variabile dipendente. Se x_1 è l'unica variabile indipendente, viene calcolata la regressione lineare semplice; se x_1+x_2+ …. sono più variabili indipendenti, viene calcolata la regressione lineare multipla.

```
> fit <- lm(LDL ~ Colesterolo + HDL + Trigliceridi, data=mydata)
# calcolo intercetta e coefficienti delle x
> coefficients(fit)
```

- Dopo avere importato di dati, la regressione multipla viene calcolata e salvata in un oggetto denominato **fit**; quindi sono mostrati i coefficienti dell'equazione della regressione lineare multipla $y = a + b \cdot x_1 + c \cdot x_2 + d \cdot x_3$:

```
(Intercept)  Colesterolo         HDL Trigliceridi
 -0.7457405    0.8907005   -0.7904253   -0.1158645
```

a b c d

y = LDL
x_1 = Colesterolo
x_2 = HDL
x_3 = Trigliceridi

- Calcoliamo gli intervalli di confidenza al 95% dell'intercetta e dei coefficienti delle *x*:

```
> confint(fit, level=0.95)
                  2.5 %     97.5 %
(Intercept)  -2.0868561  0.5953751
Colesterolo   0.8844198  0.8969812
HDL          -0.8069285 -0.7739221
Trigliceridi -0.1197046 -0.1120245
```

- Il dato più interessante è che l'intercetta non è significativamente diversa da zero. Pertanto l'equazione riportata sopra può essere semplificata e riscritta come:

$$\textbf{LDL} = 0.8907005 \cdot \textbf{Colesterolo} - 0.7904253 \cdot \textbf{HDL} - 0.1158645 \cdot \textbf{Trigliceridi}$$

(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Esaminiamo un riepilogo dei risultati:

```
> summary(fit) # un riepilogo dei risultati del modello di fitting
> fitted(fit) # ricalcolo i valori mediante l'equazione della retta di regressione
> residuals(fit) # calcolo le differenze residue tra valore osservato e valore calcolato
> anova(fit) # analisi della varianza per le differenze spiegate dalle x
> vcov(fit) # matrice di covarianza dell'intercetta e dei coefficienti delle x
```

- Questi riepiloghi forniscono un'analisi dettagliata e puntuale dei dati, che è necessario avere sempre ben presente, ma che nel nostro caso risulta ridondante; esaminiamo solo la matrice di covarianza tra i coefficienti:

```
                (Intercept)   Colesterolo           HDL  Trigliceridi
(Intercept)   0.4677338712 -1.086941e-03 -2.795948e-03 -4.144704e-04
Colesterolo  -0.0010869406  1.025850e-05 -1.234935e-05 -2.680884e-06
HDL          -0.0027959478 -1.234935e-05  7.082773e-05  8.441593e-06
Trigliceridi -0.0004144704 -2.680884e-06  8.441593e-06  3.834708e-06
```

(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Confrontiamo la regressione a tre variabili (Colesterolo+HDL+Trigliceridi) con quella a due variabili (Colesterolo + HDL) mediante analisi della varianza:

```
> fit1 <- lm(LDL ~ Colesterolo + HDL + Trigliceridi, data = mydata)
> fit2 <- lm(LDL ~ Colesterolo + HDL, data = mydata)
> anova(fit1, fit2)
Analysis of Variance Table
Model 1: LDL ~ Colesterolo + HDL + Trigliceridi
Model 2: LDL ~ Colesterolo + HDL
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2404  79918
2   2405 196298 -1   -116380 3500.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Confrontiamo la regressione a tre variabili (Colesterolo+HDL+Trigliceridi) con quella a due variabili (Colesterolo + HDL) mediante analisi della varianza:

```
> fit1 <- lm(LDL ~ Colesterolo + HDL + Trigliceridi, data = mydata)
> fit2 <- lm(LDL ~ Colesterolo + HDL, data = mydata)
> anova(fit1, fit2)
Analysis of Variance Table
Model 1: LDL ~ Colesterolo + HDL + Trigliceridi
Model 2: LDL ~ Colesterolo + HDL
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   2404  79918
2   2405 196298 -1   -116380 3500.8 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

L’analisi della varianza conferma che la regressione calcolata con tre e quella calcolate con due sole variabili indipendenti differiscono significativamente, pertanto è opportuno utilizzare la prima, più completa, per esprimere i risultati.

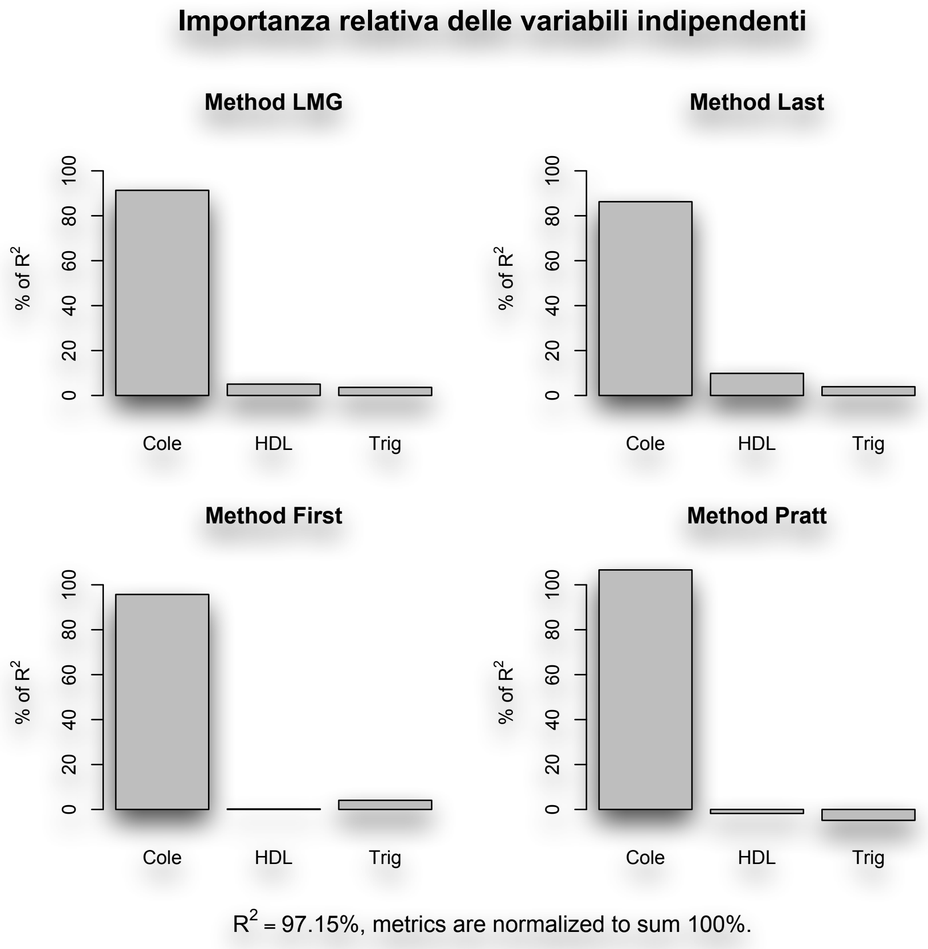
(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Calcoliamo l'importanza relativa di ciascuna variabile indipendente con quattro diversi metodi e mostriamo il grafico:

```
> library(relaimpo)
> myplot <- calc.relimp(fit, type =
    c("lmg","last","first","pratt"), rela = TRUE)
> plot(myplot,
    main = "Importanza relativa delle variabili
    indipendenti")
```

- Il parametro type elenca le metriche da calcolare. Possono essere: "lmg", "pmvd", "last", "first", "betasq", "pratt", "genizi" e "car".
- Il parametro rela=TRUE richiede che le importanze relative sommino 100% (metriche normalizzate).



(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Ripetiamo il calcolo aggiungendo gli intervalli di confidenza dei valori mediante bootstrap:

```
> boot <- boot.relimp(fit, b = 1000, type = c("lmg", "last", "first", "pratt"),
    rank=TRUE, diff=TRUE, rela=TRUE)
> booteval.relimp(boot) # mostra i risultati
```

- Esaminiamo il grafico:

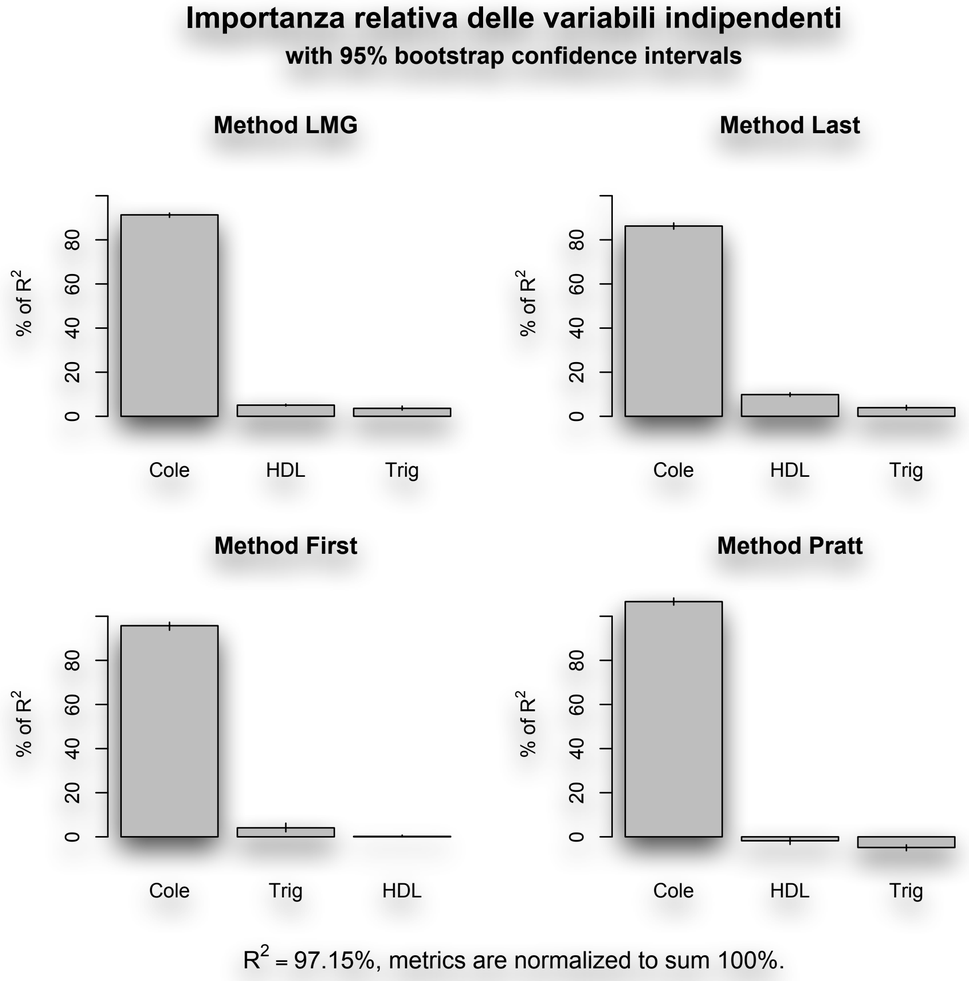
```
> plot(booteval.relimp(boot,sort=TRUE),
  main = "Importanza relativa delle variabili indipendenti")
```

- Nella diapositiva precedente è stato generato il grafico dell'importanza relativa delle variabili indipendenti nel determinare la retta di regressione, mettendo a confronto le conclusioni ottenute con quattro metodi di calcolo, che peraltro forniscono risultati molto simili.
- Con le righe di codice sopra riportate viene generato lo stesso identico grafico, questa volta con i limiti di confidenza calcolati mediante bootstrap (vedi diapositiva seguente).

(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- La figura a lato mostra l'importanza relativa della concentrazione di colesterolo totale (Cole), colesterolo HDL (HDL) e trigliceridi (Trig) – variabili indipendenti – nel determinare la concentrazione del colesterolo LDL (variabile dipendente) utilizzando un modello di regressione lineare multipla.
- I limiti di confidenza sono calcolati mediante bootstrap.



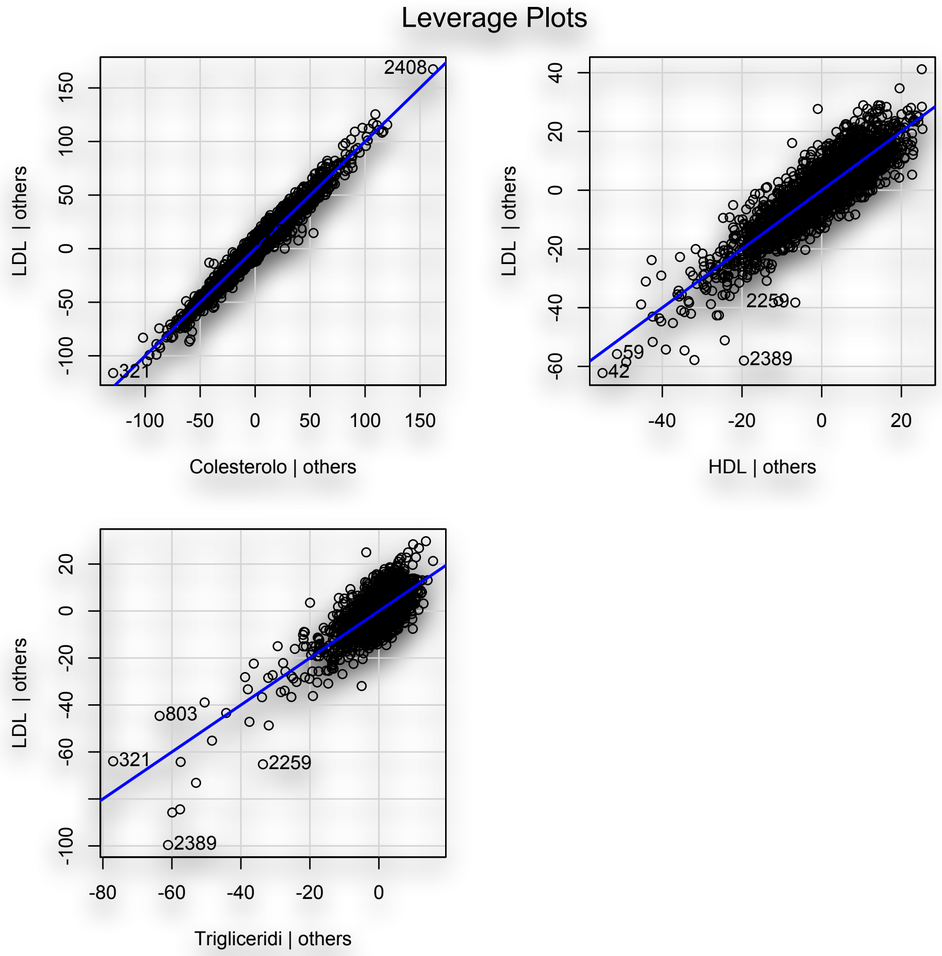
(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- La figura a lato mostra il grafico "leverage plot" dell'influenza dei dati sulle conclusioni, calcolato mediante l'istruzione:

```
> leveragePlots(fit, ask=FALSE)
```

- Il grafico esprime l'influenza delle tre variabili indipendenti (colesterolo, HDL e Trigliceridi) sulle conclusioni (variabile dipendente LDL).
- Per una discussione tecnica sulla costruzione e il significato dei leverage plot visitare il **sito del software JMP**.



(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Per individuare gli outliers, iniziamo caricando la libreria **car**:

```
> library(car)
```

- Identifichiamo gli outliers con l'istruzione **outlierTest**:

```
> outlierTest(fit)
      rstudent unadjusted p-value Bonferroni p
2389 -6.908758         6.2382e-12   1.5022e-08
2259 -5.566038         2.8954e-08   6.9722e-05
855   4.996746         6.2502e-07   1.5050e-03
606  -4.745328         2.2038e-06   5.3068e-03
1551 -4.696544         2.7951e-06   6.7305e-03
753  -4.584611         4.7814e-06   1.1514e-02
```

- Il test di Bonferroni viene impiegato per identificare i dati aberranti, i dati che cioè si discostano in modo "eccessivo" dai rimanenti. Il giudizio finale rimane ovviamente a carico di chi ha raccolto i dati, che dovrà analizzarli per capire le ragioni che hanno determinato la differenza statisticamente "poco plausibile" osservata. I dati numero 2389, 2259, 855, 606, 1551 e 753 si discostano significativamente dagli altri.

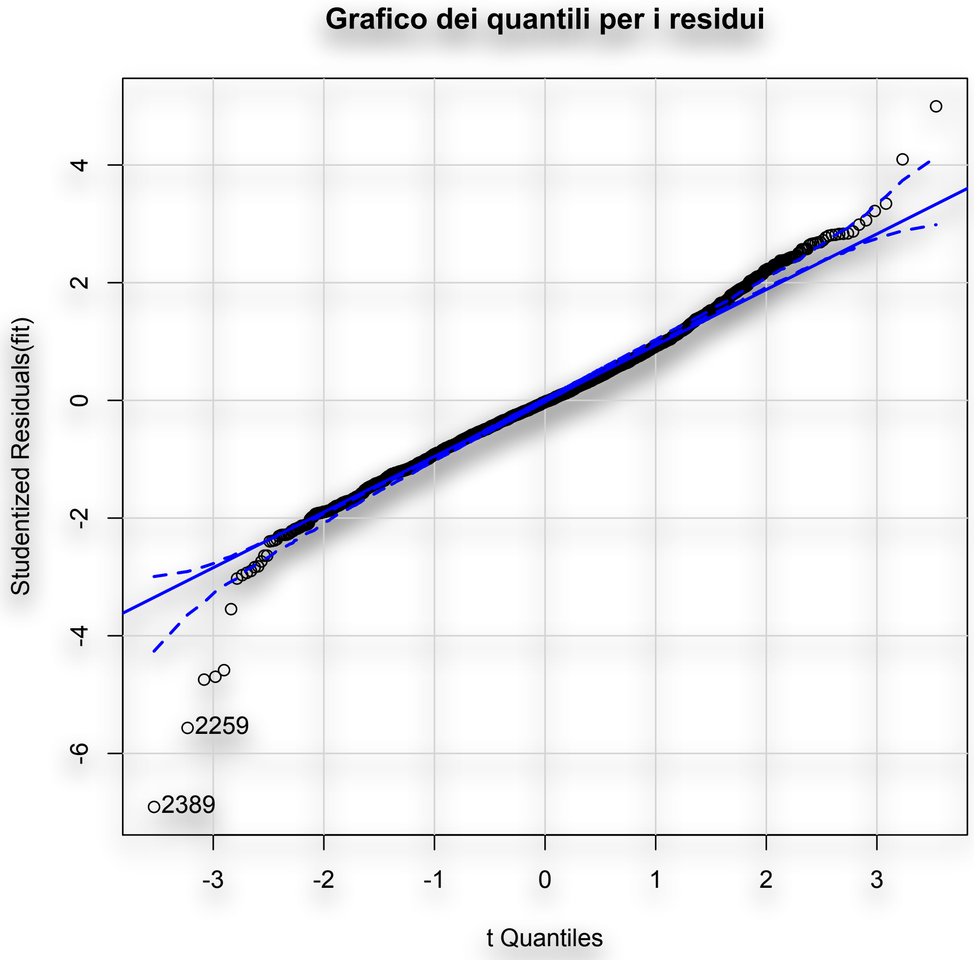
(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Visualizziamo ora il grafico dei quantili per i residui:

```
> qqPlot(fit,
  main = "Grafico dei quantili per i residui")
```

- Il grafico dei quantili per i residui standardizzati (vedi figura) conferma anch'esso la presenza di dati che si discostano molto dalla distribuzione attesa.
- Il confronto tra i quantili campionari e quanto atteso nel caso di una distribuzione gaussiana (linea continua) dimostra inoltre che la distribuzione dei dati campionari non è gaussiana.



(Bio)Statistica con R — Parte II

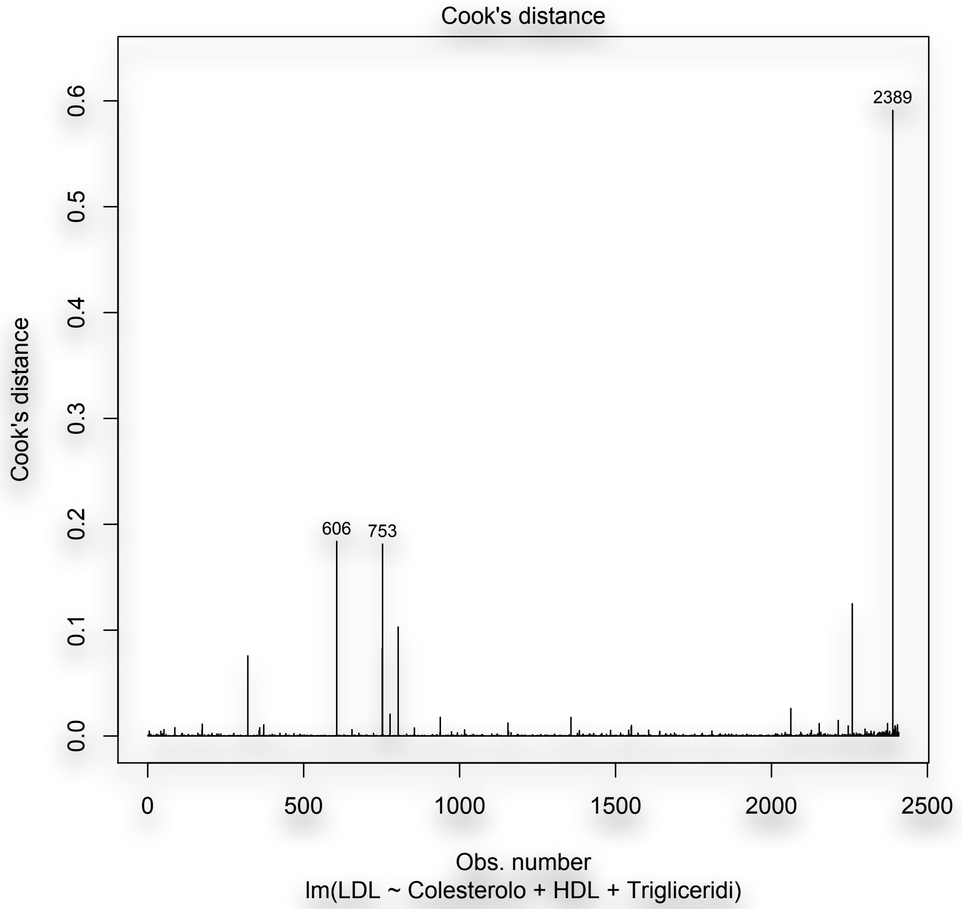
Regressione lineare (semplice e multipla)

- Applichiamo ora la distanza D di Cook per l'individuazione degli outlier:

  # grafico della distanza D di Cook: identifica i valori con D > 4/(n-k-2)

```
> cutoff <- 4/((nrow(mydata) -
    length(fit$coefficients)-2))
> plot(fit, which=4, cook.levels=cutoff)
```

- La distanza D di Cook (vedi figura) misura l'effetto conseguente alla eliminazione di una specifica osservazione. Viene riportato il numero del dato per quelli che determinano l'effetto maggiore, al fine di consentirne la rapida identificazione.
- Sono confermati tra l'altro i dati numero 606, 753 e 2389 che abbiamo già visto identificati sopra con il test di Bonferroni.

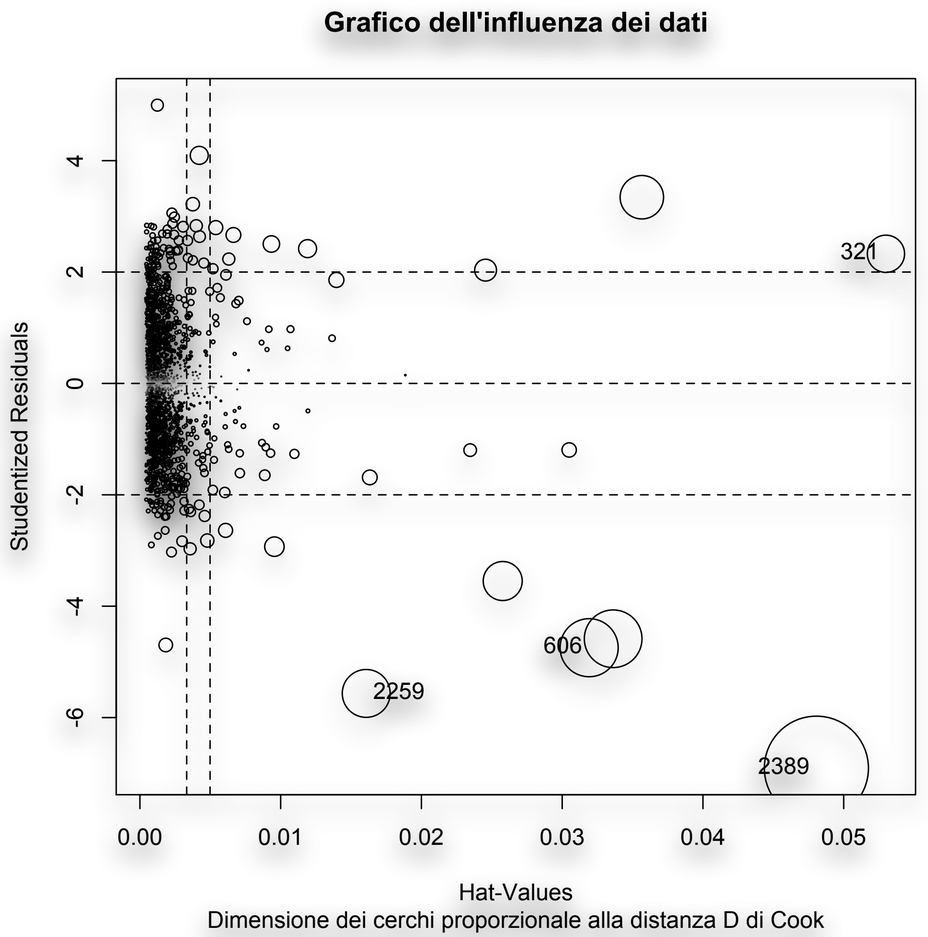


(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Per l'identificazione degli outliers vediamo un altro grafico dell'influenza dei dati sulle conclusioni.
- In questo ulteriore grafico è la dimensione dei cerchi ad essere proporzionale alla distanza di Cook (vedi figura):

```
> cutoff <- 4/((nrow(mydata) - length(fit$coefficients)-2))
> plot(fit, which=4, cook.levels=cutoff)
> influencePlot(fit,
    main = "Grafico dell'influenza dei dati",
    sub = "Dimensione dei cerchi proporzionale alla
    distanza D di Cook")
      StudRes        Hat      CookD
321    2.327997 0.05300108 0.07569068
606   -4.745328 0.03191471 0.18394148
2259  -5.566038 0.01607732 0.12499781
2389  -6.908758 0.04806009 0.59095428
```

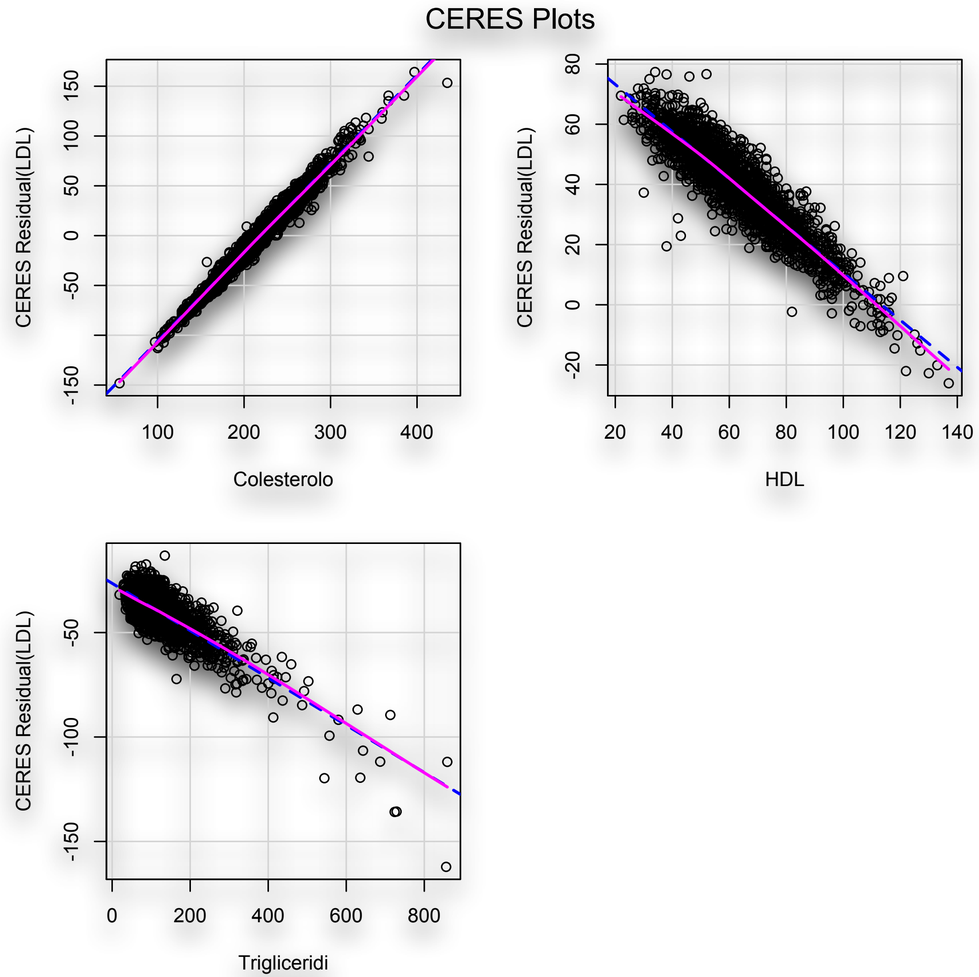


(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Il grafico di Ceres (vedi figura), sempre sviluppato da Cook, conferma l'esistenza di una relazione lineare tra colesterolo totale e colesterolo LDL, mentre il colesterolo HDL e i trigliceridi contribuiscono al colesterolo LDL in modo non lineare:

```
> ceresPlots(fit, ask=FALSE) # test per la linearità
```



(Bio)Statistica con R — Parte II

Regressione lineare (semplice e multipla)

- Ed ecco finalmente l'ultimo blocco di codice che esegue un test globale per l'assunto di linearità (mediante la libreria **gvlma**):

```
> library(gvlma); gvmodel <- gvlma(fit); summary(gvmodel)
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05
Call:
 gvlma(x = fit)
                     Value  p-value                   Decision
Global Stat        557.2861 0.000000 Assumptions NOT satisfied!
Skewness             0.5274 0.467682    Assumptions acceptable.
Kurtosis           412.5266 0.000000 Assumptions NOT satisfied!
Link Function        7.4492 0.006347 Assumptions NOT satisfied!
Heteroscedasticity 136.7829 0.000000 Assumptions NOT satisfied!
```

- Questo test globale e generalista, pur con tutti i limiti derivanti dal comprimere le conclusioni in pochi indici numerici, conferma che l'assunto di linearità non è soddisfatto.

La conclusione è ora abbastanza chiara.

I dati raccolti presentano due tipi di problemi: il primo è che alcuni di essi andrebbero rivalutati per capire il significato del loro eccessivo scostamento dai dati rimanenti; il secondo è che la relazione tra le variabili non è del tutto lineare.

Conseguentemente, la regressione lineare (in questo caso multipla), ancorché sia statisticamente significativa, deve essere intesa come una approssimazione di "grana media" della relazione che intercorre tra le quattro componenti in esame.

(Bio)Statistica con R — Parte II

ANALISI MULTIVARIATA

- **Analisi multivariata** è l'espressione con cui si fa rifermento alle numerose tecniche statistiche che consentono lo studio di sistemi complessi a più variabili.
- Qui vediamo come utilizzare **R** per l'analisi dei gruppi o *cluster analysis*.
- Scarichiamo e salviamo nella directory di lavoro il file **Clusterhclust.csv**.
- Si tratta dei dati relativi alla composizione in calcio, fosfato, ossalato e magnesio di 10 calcoli delle vie urinarie. Il contenuto del file apparirà come a lato, con i nomi delle variabili nella prima riga, i dati di ciascun caso nelle righe successive, e l'identificativo di ciascun caso nella prima colonna (C1 = calcolo 1, etc.):

Id	Calcio	Fosfato	Ossalato	Magnesio
C1	99	81	69	61
C2	78	65	53	43
C3	81	66	38	54
C4	45	23	19	16
C5	44	18	24	19
C6	102	83	72	66
C7	83	68	49	45
C8	74	71	41	57
C9	38	19	22	14
C10	48	14	21	12

(Bio)Statistica con R — Parte II

ANALISI MULTIVARIATA

- Eseguiamo nella Console di **R** il seguente codice:

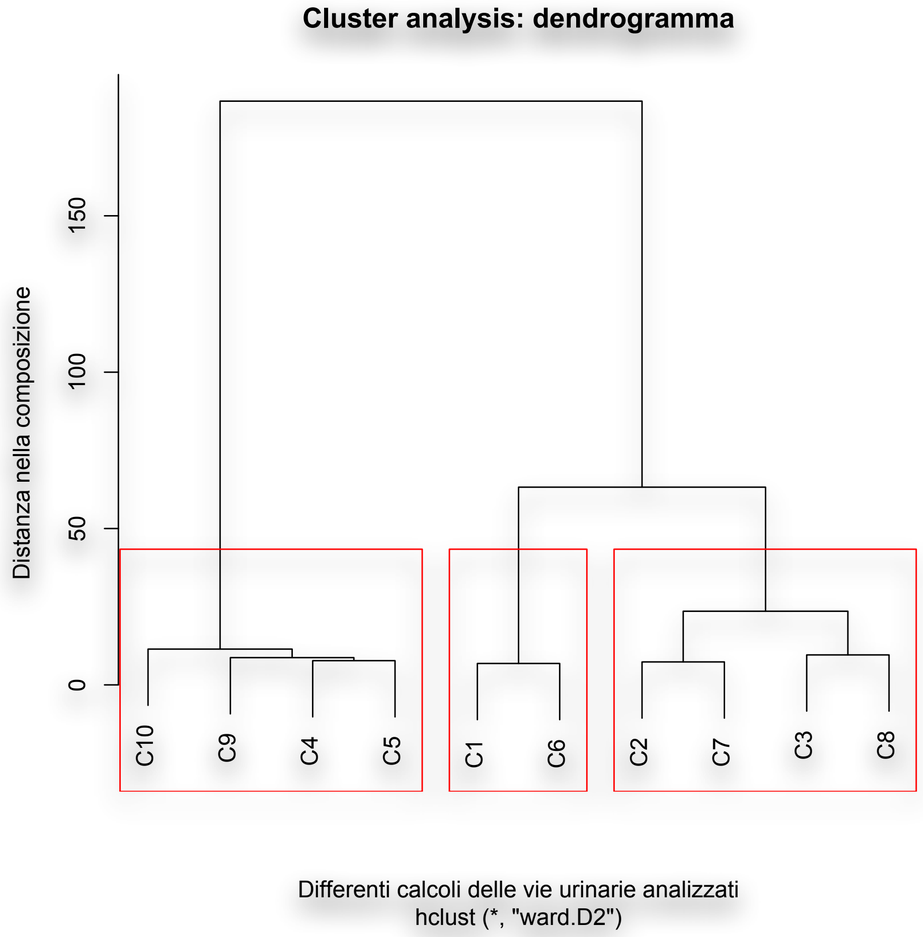
```
# importo i dati
> mydata <- read.table("Clusterhclust.csv", header=TRUE, sep=";", row.names="id")
# effettuo il clustering gerarchico con il metodo di Ward
> d <- dist(mydata, method = "euclidean") # matrice delle distanze euclidee
> fit <- hclust(d, method="ward.D2") # modello di clustering con il metodo Ward
> plot(fit, main="Cluster analysis: dendrogramma",
    xlab="Differenti calcoli delle vie urinarie analizzati",
    ylab="Distanza nella composizione") # traccio il dendrogramma
> groups <- cutree(fit, k=3) # divido in k=3 cluster principali
> rect.hclust(fit, k=3, border="red") # evidenzio i 3 cluster
```

- Al termine comparirà una finestra con il dendrogramma (vedi figura nella diapositiva seguente).

(Bio)Statistica con R — Parte II

ANALISI MULTIVARIATA

- In ascissa sono riportati i singoli calcoli, in ordinata la distanza alla quale questi vanno via via confluendo per "somiglianza" in cluster sempre più estesi.
- Minore la è distanza alla quale avviene la confluenza, maggiore è la somiglianza nella composizione dei calcoli e dei successivi cluster.
- Non esiste un valore soglia della distanza alla quale fermare il processo di raggruppamento per somiglianza dei calcoli, anche se qui sembra ragionevole affermare che:
  - i) si vanno formando tre gruppi/cluster di calcoli, e
  - ii) la composizione chimica dei calcoli C1 e C6 è più simile a quella dei calcoli C2,C7, C3 e C8 che a quella dei calcoli C10, C9, C4 e C5.



(Bio)Statistica con R — Parte II

ANALISI MULTIVARIATA

- Ora scarichiamo e salviamo il file **Clusterpvclust.csv**. Il file contiene i nomi delle variabili nella prima riga, i dati di ciascun caso nelle righe successive, e l'identificativo di ciascun caso nella prima colonna (C1 = calcolo 1, eccetera):

Id	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
Calcio	99	78	81	45	44	102	83	74	38	48
Fosfato	81	65	66	23	18	83	68	71	19	14
Ossalato	69	53	38	19	24	72	49	41	22	21
Magnesio	61	43	54	16	19	66	45	57	14	12

- Si tratta sempre degli stessi dati relativi alla composizione di 10 calcoli delle vie urinarie visti in precedenza, ma attenzione: righe e colonne sono state scambiate tra loro.

(Bio)Statistica con R — Parte II

ANALISI MULTIVARIATA

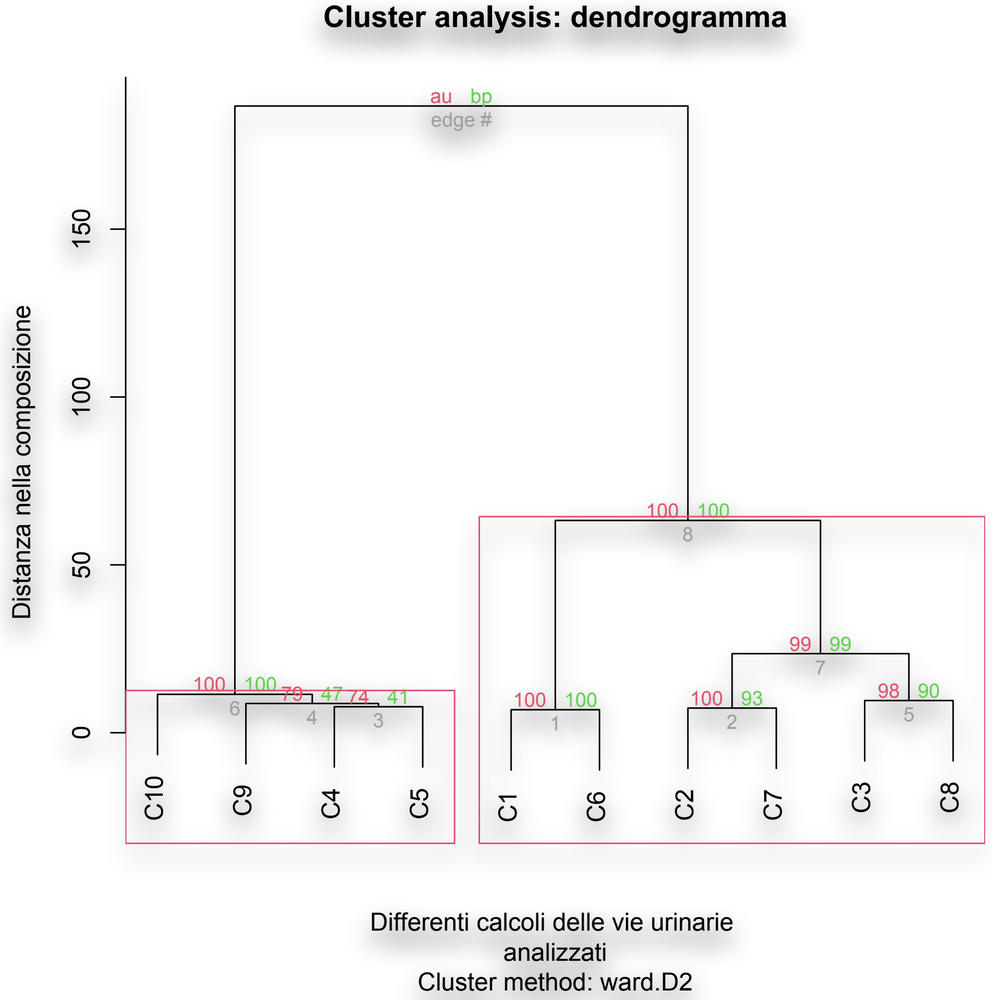
- La trasposizione tra righe e colonne si rende necessaria per utilizzare la libreria **pvclust** (da installare e caricare). Il metodo di clustering gerarchico applicato prevede in più, rispetto al caso precedente, il calcolo mediante bootstrap dei valori di probabilità *p* che caratterizzano i cluster formati in due modi differenti: come "Bootstrap Probability values" (indicati con BP) e come "Approximately Unbiased probability values" (indicati con AU). Eseguiamo questo codice:

```
# importo i dati
> mydata <- read.table("Clusterpvclust.csv", header=TRUE, sep=";", row.names="id")
# installo e carico la libreria necessaria
> install.packages("pvclust"); library(pvclust)
# clustering gerarchico con il metodo di Ward e i valori di p calcolati mediante bootstrap
> fit <- pvclust(mydata, method.hclust="ward.D2", method.dist="euclidean")
# traccio il dendrogramma con i valori di p
> plot(fit, main="Cluster analysis: dendrogramma", xlab="Differenti calcoli delle vie urinarie
    analizzati", ylab="Distanza nella composizione", print.pv=TRUE, print.num=TRUE)
# evidenzio i cluster fortemente supportati dai dati
> pvrect(fit, alpha=0.95, pv="au", type="geq", max.only=TRUE)
# legenda: BP = bootstrap probability, AU = approximately unbiased, p-values = probability value
```

(Bio)Statistica con R — Parte II

ANALISI MULTIVARIATA

- Anche in questo caso al termine comparirà una finestra con il dendrogramma che illustra la confluenza dei calcoli della vie urinarie, in termini di composizione chimica, in due gruppi (cluster) principali.
- Il fatto che i cluster siano tre (dendrogramma precedente) o due (dendrogramma qui a lato) e possano dipendere dal metodo adottato non è particolarmente preoccupante. La **cluster analysis** è un metodo per l'analisi esplorativa dei dati, e le conclusioni non devono essere viste come qualcosa di irreversibile (e quindi nel caso specifico contradditorio).
- Il metodo di clusterizzazione adottato deve essere integrato da una modellizzazione che includa (nel caso specifico) il processo di formazione dei calcoli delle vie urinarie, le patologie che ne sono alla base, i fattori che ne scatenano la formazione, al fine di collegare la somiglianza statistica dei calcoli agli elementi che determinano la loro formazione e la loro confluenza in gruppi significativi.



(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- Vediamo l'applicazione del teorema di Bayes alla diagnostica di laboratorio con un esempio che consente di illustrare come utilizzare **R** per la valutazione di un test diagnostico (nella successiva parte relativa alle curve ROC si trova un indispensabile complemento).
- L'esempio è tratto da *Scott IA, Greenberg PB, Poole PJ. Cautionary tales in the clinical interpretation of studies of diagnostic tests. Internal Medicine Journal 38 (2008) 120–129.*
- Un nuovo test diagnostico è stato provato su 1586 pazienti. Di 744 pazienti che avevano la malattia, 670 sono risultati positivi al test; di 842 pazienti che non avevano la malattia, 640 sono risultati negativi al test.
- Riportiamo i dati nella seguente tabella di contingenza (o anche "matrice di confusione"), dalla quale si deduce che erano 670 i veri positivi (TP=True Positive), 202 i falsi positivi (FP=False Positive), 74 i falsi negativi (FN=False Negative) e infine 640 i veri negativi (TN=True Negative).

	Malattia +	Malattia -	TOTALE
Test +	**670** (TP)	**202** (FP)	872
Test -	**74** (FN)	**640** (TN)	714
TOTALE	744	842	1586

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- Il codice **R** è estremamente conciso se si utilizza la libreria **epiR**, che ovviamente deve essere preventivamente installata e caricata:

```
# installo e carico la libreria necessaria
> library(epiR)
# inserisco i dati direttamente
> data <-
  as.table(matrix(c(670,202,74,640),
  nrow = 2, byrow = TRUE))
# calcolo e mostro tutte le statistiche
> epi.tests(data, conf.level = 0.95)
```

- Il fatto interessante è che le grandezze calcolate sono riportate ciascuna con il rispettivo intervallo di confidenza al 95%

```
        Outcome +    Outcome -      Total
Test +        670          202        872
Test -         74          640        714
Total         744          842       1586

                Point estimates and 95 % CIs:
-------------------------------------------------------------
Apparent prevalence                   0.55 (0.52, 0.57)
True prevalence                       0.47 (0.44, 0.49)
Sensitivity                           0.90 (0.88, 0.92)
Specificity                           0.76 (0.73, 0.79)
Positive predictive value             0.77 (0.74, 0.80)
Negative predictive value             0.90 (0.87, 0.92)
Positive likelihood ratio             3.75 (3.32, 4.24)
Negative likelihood ratio             0.13 (0.11, 0.16)
-------------------------------------------------------------
```

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

Ecco qui le grandezze calcolate con **R**, il loro significato, le formule con cui sono calcolate, e il risultato numerico ottenuto (per semplicità l'intervallo di confidenza viene omesso).

- **Apparent prevalence**
  prevalenza apparente, soggetti con il test positivo: (TP+FP) / (TP+FP+FN+TN) = (670+202) / 1586 = 0.5498108 ≅ 55%
- **True prevalence**
  prevalenza reale, soggetti con la malattia: (TP+FN) / (TP+FP+FN+TN) = (670+74) / 1586 = 0.4691047 ≅ 47%

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- **Sensitivity (Recall)**

  sensibilità, positività nei malati: TP / (TP+FN) = 670 / (670+74) = 0.9005376 ≅ 90%

- **Specificity**

  specifi cità, negatività nei sani: TN / (TN+FP) = 640 / (640+202) = 0.760095 ≅ 76%

- **Positive predicted value (Precision)**

  valore predittivo di un test positivo: TP / (TP+FP) = 670 / (670+202) = 0.7683486 ≅ 77%

- **Negative predicted value**

  valore predittivo di un test negativo: TN / (TN+FN) = 640 / (640+74) = 0.8963585 ≅ 90%

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- **Positive likelihood ratio**

  rapporto di verosimiglianza LR+ per un test positivo: (TP/(TP+FN)) / (FP/(FP+TN)) = (670/(670+74)) / (202/(202+640)) = 3.753726 ≅ 3.75

- **Negative likelihood ratio**

  rapporto di verosimiglianza LR– per un test negativo: (FN/(TP+FN)) / (TN/(FP+TN)) = (74/(670+74)) / (640/(202+640)) = 0.1308551 ≅ 0.13

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

Possiamo inoltre calcolare altri utili indicatori:

- **Accuracy**
  accuratezza diagnostica: (TP+TN) / (TP+FP+FN+TN) = (670+640) / 1586 = 0.8259773 ≅ 83% (il suo complementare 1-accuracy è l'errore di previsione).
- **Odds ratio**
  rapporto LR+ / LR– = 3.753726 / 0.1308551 = 28.6861267 ≅ 28.69
- **NNT** (Number Needed to Treat)
  numero necessario per la diagnosi:
  1 / (sensibilità – (1 – specifi cità)) ≅ 1.51
  È il numero di pazienti da trattare per ottenere un beneficio terapeutico (valore ideale = 1)

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- **False Positive Rate (FPR)**

  Il tasso di falsi positivi viene calcolato come il numero di previsioni positive errate diviso per il numero totale di negativi: FP / (TN+FP) = 202 / (202+640) = 0.2399050 ≅0.24.

  Il miglior tasso di falsi positivi è 0 mentre il peggiore è 1; può anche essere calcolato come 1 – specificità.

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- **F-score**

  È una media armonica ponderata delle metriche *Precision* e *Recall* in modo tale che il punteggio migliore sia 1 e il peggiore sia 0:

  F-score = 2×Recall×Precision / (Recall + Precision) = 2×0.9005376×0.7683486 / (0.9005376+0.7683486) = 0.8292079 ≅ 83%.

  Viene utilizzata per confrontare differenti modelli di classificatore (o di test diagnostici), non la precisione globale del test.

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)

(Bio)Statistica con R — Parte II

STATISTICA BAYESIANA: Valutazione di un test diagnostico

- **Youden index**

  indice di Youden = sensibilità + specificità – 1 = 0.6606326 ≅ 66%

Identifica il *best cut-off*, cioè il valore del *test* che massimizza la differenza tra veri positivi e falsi positivi.

A differenza dell'AUC, l'indice di Youden è una funzione della sensibilità e della specificità massimizzata rispetto al cut-poin ottimale, e fornisce anche una misura immediata del tasso di classificazione corretta globale massimo per un dato marcatore.

	Outcome +	Outcome -	Total
Test +	670	202	872
Test -	74	640	714
Total	744	842	1586

Point estimates and 95 % CIs:

Apparent prevalence	0.55 (0.52, 0.57)
True prevalence	0.47 (0.44, 0.49)
Sensitivity	0.90 (0.88, 0.92)
Specificity	0.76 (0.73, 0.79)
Positive predictive value	0.77 (0.74, 0.80)
Negative predictive value	0.90 (0.87, 0.92)
Positive likelihood ratio	3.75 (3.32, 4.24)
Negative likelihood ratio	0.13 (0.11, 0.16)