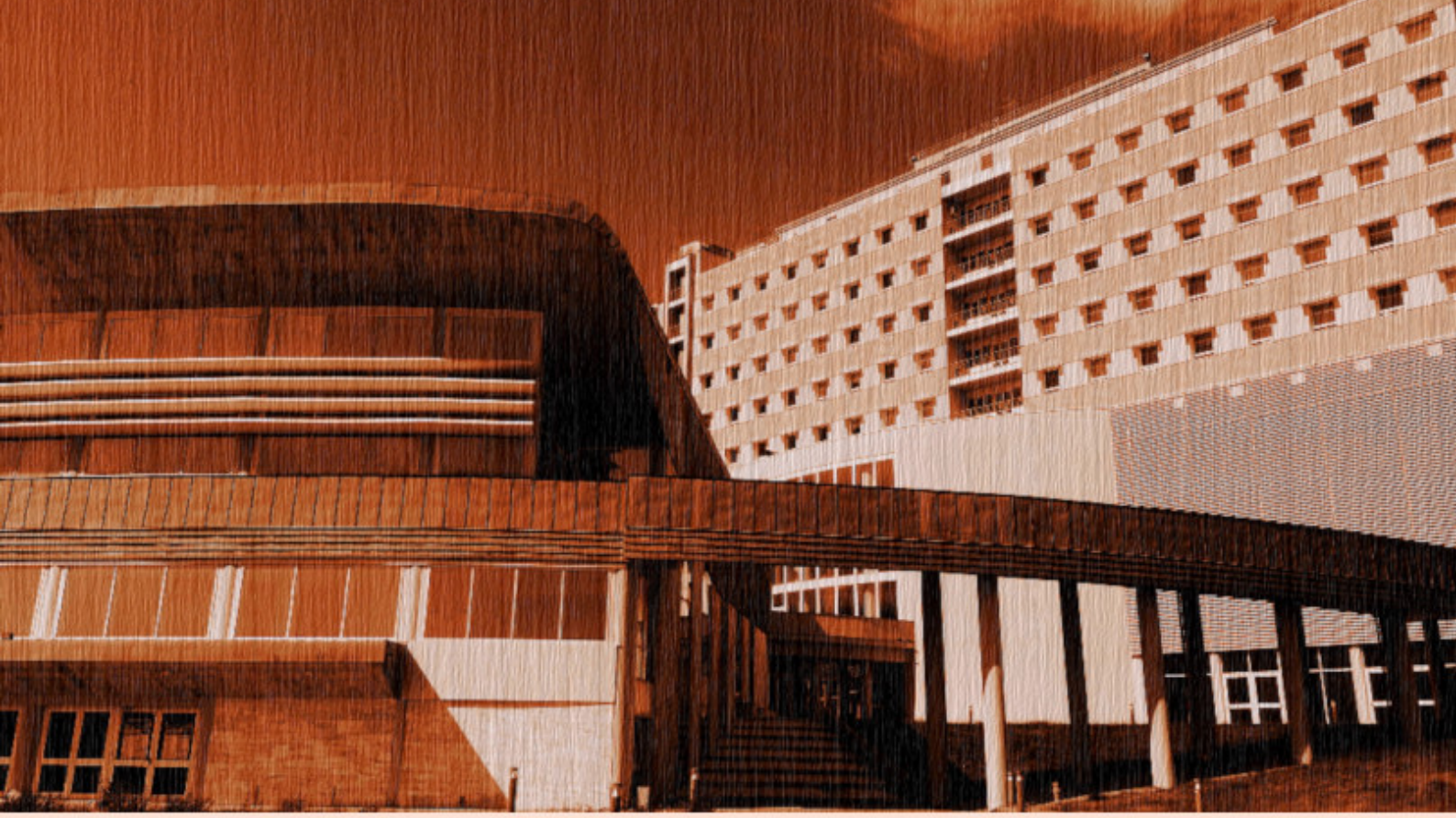


Metodi di Biostatistica, Volume 1

per le Scuole di Dottorato dell'Università Magna Græcia di Catanzaro

Massimo Borelli

http://www.biostatisticaumg.it

First printing, June 2019

Indice

Prima Parte

1. Descrivere i dati

Problema 1.1 *... abbiamo iniziato a scrivere il nostro primo articolo, e ci è stato chiesto di preparare la classica Tabella 1, quella in cui si deve fare la descrittiva del campione. E non sappiamo bene come si debba procedere ...*

1. Roberta Venturella [25], *3 to 5 years later: long-term effects of prophylactic bilateral salpingectomy on ovarian function*: Mean Values $\pm$ SD.
2. Emanuela Chiarella [7], *ZNF521 Represses Osteoblastic Differentiation in Human Adipose-Derived Stem Cells*: Means + SD.
3. Maria Teresa De Angelis, *Short-term retinoic acid treatment sustains pluripotency and suppresses differentiation of human induced pluripotent stem cells* [8]: mean $\pm$ standard error of the mean (SEM).
4. Maria Vittoria Caruso [6] *Influence of IABP-Induced Abdominal Occlusions on Aortic Hemodynamics: A Patient-Specific Computational Evaluation*, citando *Furthermore, they also discovered that the distance between LSA and CT was 241 $\pm$ 23mm*
5. Annalisa Di Cello [9], *A more accurate method to interpret lactate dehydrogenase isoenzymes' results in patients with uterine masses* Mean [Median].
6. Infine, Paolo Zaffino [30], *Radiotherapy of hodgkin and non-hodgkin lymphoma: a nonrigid image-based registration method for automatic localization of prechemotherapy gross tumor volume*: mediana $\pm$ quartili.

- design sperimentale di tipo cross-section ?
- misure ripetute ?

1.1 Misure di tendenza centrale

(dette anche **indici di posizione**)

- **media**
- **mediana**
- **moda**

web La pagina Central Tendency di Wikipedia fornisce un rapido richiamo a numerose altre misure di tendenza centrale: `https://en.wikipedia.org/wiki/Central_tendency`.

varie possibili classificazioni riguardanti il 'tipo' di dato statistico.

- dati **qualitativi** (ad esempio quando un paziente può essere assegnato senza confusione a gruppi disgiunti, tra i quali non è necessario che intercorra una relazione di tipo numerico: per esempio, il genere: femmina o maschio; il gruppo sanguigno A, B, AB oppure 0),
- dati **quantitativi discreti** (tipicamente, elementi che vengono contati ad uno ad uno con numeri interi)
- **quantitativi continui** (evidentemente, tutti i casi rimanenti).

Quesito 1.1.1 Se il donatore di sangue Massimo Borelli riceve nel referto l'indicazione che i trigliceridi ammontano ad 89, e non 89.0 mg/dL, ci conviene interpretare questo dato quantitativo come discreto o continuo?

? scala A.S.A. che valuta il rischio anestesiologico: in base alle condizioni del paziente, e che assume i valori I, II, III, IV e V (modificabili ciascuno con la E di 'emergency'): si tratta di una scala solamente qualitativa, oppure l'ordine conta?

- i dati **categorici**
- i dati **ordinali** (come appunto il rischio anestesiologico)
- i dati **quantitativi discreti**
- i dati **quantitativi continui**

[15] concetto di **scale di misura**

- scala **nominale** (ad esempio nell'epidemiologia, parliamo di individui *S*uscettibili, *I*nfettivi e *R*imossi)
- scala **ordinale**, quella **intervallare** (come ad esempio la scala di temperatura Celsius, che ha uno zero convenzionale ed in cui ha senso solamente parlare di differenze di temperatura)
- scala **a rapporti** (come ad esempio la scala di temperatura Kelvin, che ha uno zero assoluto ed in cui ha senso dire che un corpo a 200 $^\circ K$ ha una temperatura doppia di un corpo a 100$^\circ K$)

Quesito 1.1.2 L'indice di massa corporea BMI (`https://meshb.nlm.nih.gov/record/ui?ui=D015992`) individua negli adulti quattro categorie: below 18.5 (underweight), 18.5-24.9 (normal), 25.0-29.9 (overweight), 30.0 and above (obese). Di che tipo di dato stiamo parlando?

software R
- **fattore** (suddiviso in due o più **livelli**), `factor`, `levels` che possono essere `ordered`
- `numeric`

software Orange
- `categorical`
- `numeric`
- `text` (meta)

1.1.1 Facciamo conoscenza con R

Importare un dataset esistente

? R Studio ? R ?

un nuovo **script**
dataset `iris`, Ronald Fisher [10], Edgar Anderson [2].

```
iris
```

Ora selezioniamo 'schiacciamo' *Run*

`Sepal.Length`, `Sepal.Width`, `Petal.Length`, `Petal.Width` e `Species`.

```
str(iris)
```

Determinare le misure di tendenza centrale

```
attach(iris)
mean(Petal.Length)
median(Petal.Length)
```

tabella di **frequenze assolute**

```
table(Species)
```

Vocabolario 1.1 — dataset bilanciato. Il dataset `iris` si dice **bilanciato** (in inglese **balanced**) in quanto i dati dei tre gruppi presi in esame sono stati osservati rispettando in ciascuno la medesima frequenza assoluta.

```
tapply(Petal.Length, Species, mean)
tapply(Petal.Length, Species, median)
```

Manipolare un dataset

Esercizio 1.1 Cercate di capire da soli cosa succede se eseguite, ad una ad una, le istruzioni qui di seguito elencate. In particolare, cercate di cogliere il senso delle funzioni `head`, `tail` e `names`.

```
iris[1,]
iris[1:6,]
head(iris)
iris[145:150,]
tail(iris)
iris[,1]
iris[,c(3,4,5)]
iris[,3:5]
names(iris)
```

■

Esercizio 1.2 Provate ad eseguire, ad una ad una, le seguenti istruzioni. Scoprirete alla fine due importanti **costanti booleane** di R.

```
levels(Species)
levels(Species)[2]
is.numeric(Species)
is.factor(Species)
is.factor(Petal.Length)
```

■

Esercizio 1.3 Provate ad eseguire, ad una ad una, le seguenti istruzioni e discutetene l'output.

```
iris[order(Sepal.Length), ]
iris[order(Sepal.Width), ]
iris[order(Sepal.Length, Sepal.Width), ]
iris[rev(order(Sepal.Length)), ]
iris[Species == "virginica",]
iris[(Species == "virginica") & (Sepal.Length == 6.3),]
iris[(Species == "virginica") & (Sepal.Length != 6.3),]
```

■

```
length(iris)
length(Petal.Length)
detach(iris)
length(iris)
length(Petal.Length)  ### ???
```

```
with(iris, length(Petal.Length))
```

```
? airquality
attach(airquality)
head(airquality)
```

Vocabolario 1.2 — dataset incompleti. Un dataset si dice **incompleto** quando osserviamo alcuni **missing data** o **valori mancanti**, che in R (ed in Orange) vengono codificati con il simbolo `NA`, acronimo di `N`ot `A`vailable.

Esercizio 1.4 Provate ad eseguire ciascuno di questi comandi e discutetene l'output. Cercate in particolare di capire con precisione il significato della funzione `which`.

```
na.omit(airquality)
complete.cases(airquality)
which(complete.cases(airquality) == TRUE)
```

■

Usare R come una calcolatrice scientifica

```
1/0
0/0
```

Esercizio 1.5 Considerate il numero e ed osservate come funzionano gli arrotondamenti ed i troncamenti in R.

```
exp(1)
round(exp(1), 2)
floor(exp(1))
floor(-exp(1))
trunc(exp(1))
trunc(-exp(1))
```

■

1.2 Misure di dispersione

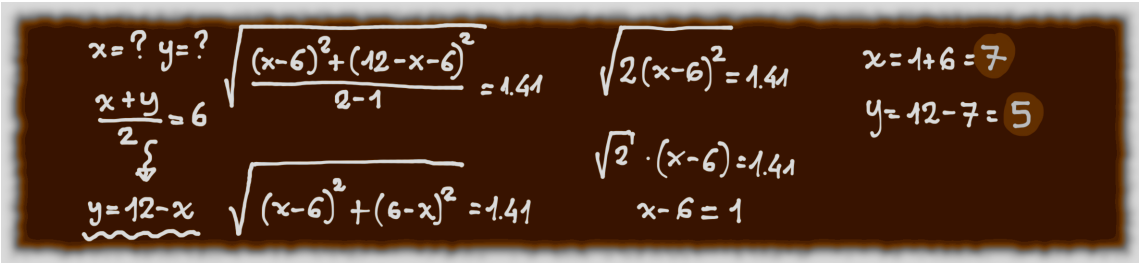
la **variabilità** (o uno dei suoi sinonimi, **eterogeneità**, **dispersione**)

- **deviazione standard**

? **campione**? **popolazione** ?

popolazione: $\sqrt{\frac{\sum_i^n (x_i - m)^2}{n}}$, campione: $\sqrt{\frac{\sum_i^n (x_i - m)^2}{n-1}}$

```
attach(iris)
sqrt(sum((Petal.Length - mean(Petal.Length))^2)/150)
sqrt(sum((Petal.Length - mean(Petal.Length))^2)/149)
sd(Petal.Length)
```



web La pagina Statistical Dispersion di Wikipedia fornisce un rapido richiamo a numerose altre misure di tendenza centrale: `https://en.wikipedia.org/wiki/Statistical_dispersion`.

Quesito 1.2.1 Vi viene in mente un modo per verificare, con R, che la deviazione standard risulta essere definita come la radice quadrata della varianza (che si calcola con la funzione `var`)? No? Allora provate a rileggere a cosa servano i comandi `sd` e `sqrt` qui sopra.

1.2.1 Misure di dispersione con R

La funzione summary

```
summary(Species)
```

```
summary(Petal.Length)
```

Importare un dataset dalla rete

il dataset `cholesterol`

`http://www.biostatisticaumg.it/dataset/cholesterol.csv`

formato .csv

web È essenziale farsi una piccola cultura di base sui formati con cui vengono salvati i dati per poter essere scambiati. Qui vediamo cosa sono i file .csv, `https://en.wikipedia.org/wiki/Comma-separated_values`. Teniamo presente che il carattere 'virgola' si può usare come separatore perché per default nel mondo anglosassone i numeri decimali vengono rappresentati con il punto (ossia $\pi \approx 3.14$) e non come in Italia con la virgola (ossia $\pi \approx 3,14$). Questo implica che se volete condividere il vostro foglio elettronico e all'interno avete delle misure decimali, dovrete assicurarvi di non fare pasticci, agendo sul *Formato* delle *Celle*. Qui potete trovare alcuni tutorial per come esportare il vostro foglio dati in formato `.csv`, in funzione del foglio di calcolo che preferite usare:

- `https://www.youtube.com/results?search_query=csv+export+excel`
- `https://www.youtube.com/results?search_query=csv+export+openoffice`
- `https://www.youtube.com/results?search_query=csv+export+libreoffice`

```
indirizzo = "http://www.biostatisticaumg.it/dataset/cholesterol.csv"
cholesterol = read.csv(indirizzo, header = TRUE)
attach(cholesterol)
tail(cholesterol)
```

```
max(table(idanag))
```

```
which(table(idanag) == max(table(idanag)))
```

I quantili di una distribuzione

```
quantile(TOTchol, 0.95)
```

Da dove esce questo 267?

```
sort(TOTchol)[974]
```

Esercizio 1.6 Verificate che il 513-esimo elemento di `sort(TOTchol)`, ossia il cinquantesimo percentile, è proprio la mediana di `TOTchol`. E verificate anche che il venticinquesimo ed il settantacinquesimo percentile sono proprio il primo ed il terzo quartile che avevamo trovato nel `summary`. ■

1.3 Descrivere i dati nei design cross-section

Siamo quasi pronti per iniziare a dare una (parziale) risposta al problema 1.1

Quesito 1.3.1 Tre amici aprono i loro portafogli, contano il denaro in loro possesso e ci dicono che il valore **medio** è 43.71 euro. Cosa possiamo dedurre?

Quesito 1.3.2 Tre amici aprono i loro portafogli, contano il denaro in loro possesso e ci dicono che il valore **mediano** è 43.71 euro. Cosa possiamo dedurre?

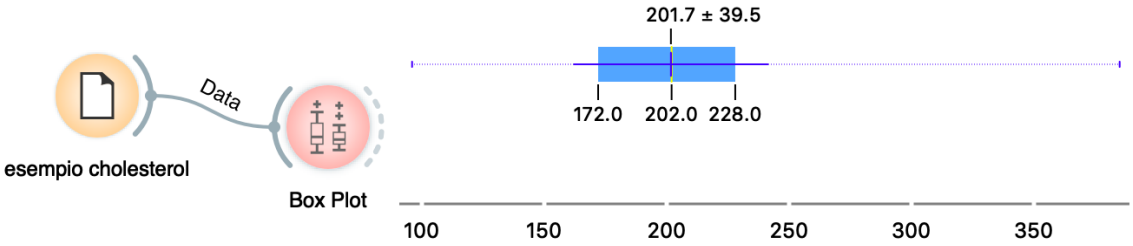
studi di tipo **cross sectional**:

- 'approccio parametrico': la media 'va d'accordo' con la deviazione standard
- 'approccio non parametrico': la mediana 'va d'accordo' con i quartili ed il range

Il consiglio finale: scegliere dopo aver effettuato un'analisi esplorativa

1.3.1 Facciamo conoscenza con Orange

- **widget**
- **canvas**
- **workflow**
- caricare il dataset `iris`
- **diagramma a barre** (ossia, un **barplot**)
- dalla rete il dataset `cholesterol`
- widget `Box Plot`



1.3.2 Il boxplot con R

```
boxplot(Petal.Length ~ Species)

library(ggplot2)
ggplot(iris, aes(x=Species, y=Petal.Length)) + geom_boxplot()
```

- **outlier**

1.3.3 Due presupposti fondamentali

1. operare una scelta consapevole sul grafico opportuno da presentare.

web Su questo primo punto vi invitiamo certamente a leggere con attenzione le raccomandazioni di Claus Wilke nel suo e-book *Fundamentals of Data Visualization* [28], `https://serialmentor.com/dataviz/` per visualizzare quantità, distribuzioni, proporzioni, associazioni, serie temporali e molti molti altri tipi di dato, evitando i possibili tranelli in cui spesso si incorre. Il libro utilizza il pacchetto `ggplot2`, il quale è veramente molto accattivante e molto professionale, ma richiede un po' di tempo per impratichirsene. Date un'occhiata alla sua galleria di immagini `http://www.ggplot2-exts.org/gallery/` per convincervi che vale la pena impararlo ad usare, ad esempio da `https://ggplot2.tidyverse.org/`, oppure, leggendo il libro di Hadley Wickham e Garrett Grolemund [27], `https://r4ds.had.co.nz/`

2. la fase di 'raccolta dei dati': provengono da tabelle di database? siamo stati noi a digitarli? ce li hanno forniti?

1.4 Descrivere i dati nei design a misure ripetute

...

1.4.1 Calcolare i dati di relative gene expression con R

Creare un dataset con R

...

Separare una variabile in base ad un fattore con R

...

1.4.2 L'equazione più pericolosa, parte prima

Id	Well	BiolRep	TechRep	Expression
1	p28	a	1	1.01
2	p28	a	2	0.97
3	p28	b	1	1.05
4	p28	b	2	1.08
5	p28	c	1	0.98
6	p28	c	2	0.92
7	p53	d	1	0.25
8	p53	d	2	0.26
9	p53	e	1	0.31
10	p53	e	2	0.28
11	p53	f	1	0.28
12	p53	f	2	0.30

- Howard Wainer, [26], *The most dangerous equation*
- Tu e Gilthorpe, *The most dangerous hospital or the most dangerous equation?*

concetto di **errore standard della media** (i.e. lo **standard error of the mean**),

$$SEM = \frac{\sigma}{\sqrt{n}}$$

- (**sample size**) n

- **dynamite plot**,

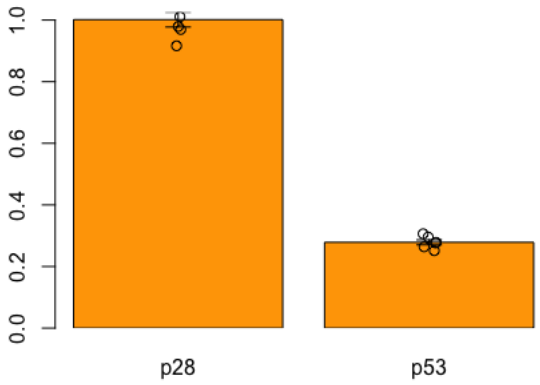
? Meglio di no ?

- il professor Tatsuki Koyama, [18] mostra chiaramente come due dynamite plot apparentemente uguali celino completamente la struttura e la dimensione dei campioni presi in esame
- la comunità degli sviluppatori di R, che non ha mai creato una funzione di base (se non in alcune `library` aggiuntive) per disegnare il dynamite plot
- la comunità degli sviluppatori di Orange, che a quanto pare non ha creato nemmeno library aggiuntive per farlo.

Come uscirne? La mia raccomandazione è:

1. se il nostro design sperimentale non prevede l'utilizzo di misure ripetute, **non** utilizzare mai il dynamite plot
2. se dobbiamo raffigurare esperimenti con misure ripetute, allora sovrapponiamo al dynamite plot anche un **dot plot**, i cui pallini evidenzino i dati grezzi.

1.4.3 Sovrapporre due grafici con R



...

1.5 Esercizi ed attività di approfondimento

■ **Attività 1.1 — misure di tendenza centrale e di dispersione.** ([23, pagina 14]) Considerate il campione dei giorni intercorsi tra il penultimo e l'ultimo periodo mestruale di 500 giovani donne. La colonna delle frequenze riporta il numero di donne che ha riferito rispettivamente quel periodo.

Periodo	Frequenza	Periodo	Frequenza	Periodo	Frequenza
24	5	29	96	34	7
25	10	30	63	35	3
26	28	31	24	36	2
27	64	32	9	37	1
28	185	33	2	38	1

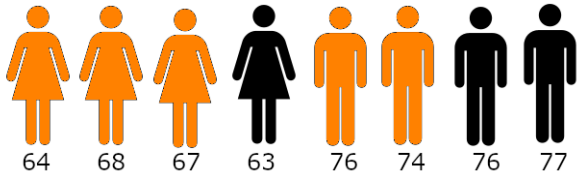
Sapreste determinare 'a mente' la moda? E sempre 'a mente' la mediana? Ed il primo e terzo quartile? E sempre 'a mente', o 'con carta e matita', sapreste dire se ci sono forse delle ragazze outlier? Provate poi ad utilizzare la funzione `rep()` per implementare in R un vettore denominato `periodo`, e controllate con le funzioni `table` e `summary` le due soluzioni da voi proposte. ■

■ **Attività 1.2 — creare ed importare un dataset.** Un mio caro amico di Trieste, il professor Umberto Lucangelo, aveva suggerito ad una laureanda in Odontoiatria uno studio sull'effetto analgesico di tre farmaci. Importate in R il dataset `analgesia.txt` con i comandi seguenti, osservando che si tratta di un dataset in formato `.txt`, il cui carattere di separazione è il segno di tabulazione (e non la virgola) e pertanto utilizziamo il comando `read.table` invece di `read.csv`.

```
www = "http://www.biostatisticaumg.it/dataset/analgesia.txt"
analgesia = read.table(www, header = TRUE)
attach(analgesia)
```

- Quante righe e quante colonne ha il dataset analgesia? Scopritelo con `str`.
- Cosa otteniamo con il comando `table(sex)`: le frequenze assolute o relative dei pazienti?
- Cosa otteniamo con il comando `prop.table(sex)`?
- Cosa otteniamo con il comando `prop.table(table(sex))`?
- Mia moglie, mia mamma e mia suocera affermano che i maschi sopportano meno il dolore delle femmine. Provate a disegnare una coppia di boxplot del `dolore6h` rispetto a `sex`.
- Quanti sono i pazienti maschi che hanno assunto il tramadolo? Scopritelo con `table`.
- Quanti sono i pazienti maschi con `(dolore6h >= 5)` ? Scopritelo con `table`.

■



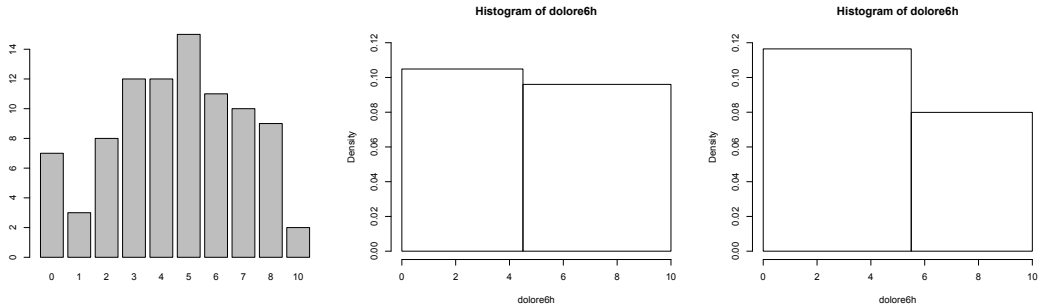
■ **Attività 1.3 — creare ed importare un dataset.** State iniziando uno studio inerente i disturbi dell'alimentazione, ed avete un campione iniziale di otto soggetti, maschi e femmine, alcuni dei quali possiedono una certa mutazione genetica (soggetti arancione della figura qui in alto). I numeri che vedete raffigurati rappresentano il peso di ciascun soggetto. Impostate i dati in un foglio di calcolo elettronico (MS Excel™, Open Office Calc, Google Sheets, Libre Office Calc, ...), salvatelo, ed importatelo in R. Calcolate usando `tapply` la deviazione standard dei pesi dei pazienti con mutazione e di quelli senza mutazione.
Suggerimento: vi tornerà utile la funzione `file.choose()`*; scoprite come funziona cercando qualche esempio in rete.* ■

■ **Attività 1.4 — struttura di un dataset.** Il foglio elettronico è uno strumento molto comodo per raccogliere i dati. Tuttavia, trascrivere dei dati non significa automaticamente aver creato un dataset. Date un'occhiata alle due schermate sottostanti, osservate che entrambe descrivono correttamente l'esperimento dell'attività precedente, ma questi dati non possono venire importati *sic stantibus* in un qualsiasi software oggi esistente di analisi dei dati. ■

	A	B	C
1	peso	MUT	WT
2	F	64, 68, 67	63
3	M	74, 76	76, 77
4			

	A	B
1	F	M
2	64	76
3	68	74
4	67	76
5	63	77
6		

■ **Attività 1.5 — istogramma.** Parliamo un poco della diversità di significato che hanno i grafici a barre dagli istogrammi, riferendoci ancora al dataset `analgesia`.




```
par(mfrow = c(1,3))
barplot(table(dolore6h))
hist(dolore6h, breaks = c(0, 4.5, 10), ylim = c(0, 0.12))
hist(dolore6h, breaks = c(0, 5.5, 10), ylim = c(0, 0.12))
```

A sinistra vediamo un grafico a barre, da cui deduciamo immediatamente che il `dolore6h` modale del nostro campione è 5. Capite che in un diagramma a barre l'informazione essenziale è fornita dalla *altezza* delle colonnine grigie, le quali sono proporzionali alle *frequenze assolute* di `dolore6h`:

```
table(dolore6h)
```

Ora invece osservate cosa succede con i comandi `table(dolore6h < 5)` e `table(dolore6h > 5)`: riuscite a collegare i numeri che avete ottenuto (32, 42, .. , ..) con la *forma* dei rettangoli bianchi degli istogrammi? E se vi chiedessimo di calcolare, 'con il righello' e 'con carta e matita', le aree dei rettangoli bianchi (*base per altezza*), che valori otterreste? Completate dunque da soli la frase: *in un istogramma l'informazione essenziale è fornita dalle . . . . . delle colonne, le quali sono proporzionali alle frequenze . . . . . di* `dolore6h`. In definitiva, in un diagramma a barre non importa la larghezza delle colonne, ma in un istogramma sì. Cercate a tale proposito nella vostra biblioteca universitaria (o in qualche angolo nascosto nella rete) se riuscite a trovare il manuale di Venables e Ripley [24] e guardate la loro figura 5.8 di pagina 127: vedete come la scelta dei **cut-off** (i.e. dei `breaks`) dell'istogramma riesce a cambiare, e di molto, la forma dei rettangoli? ■

2. Simulare i dati sperimentali con R

...

2.1 R è un linguaggio di programmazione

...

Problema 2.1 *.. abbiamo qui la sequenza della proteina Zinc Finger 521, e vorremmo calcolarci lo strand complementare, in maniera semplice.*

2.1.1 Usare il ciclo for e la decisione if

...

La programmazione vector-based

...

2.1.2 Creare le proprie funzioni

...

2.2 Gli eventi casuali con R

$$7,\ 0,\ 5,\ 10,\ 3,\ ?$$

```
orologio = seq(from = 7, by = 5, length.out = 6 )
orologio
orologio %% 12
```

generare sequenze di **numeri pseudocasuali**.

web Per chi fosse interessato, in `https://random.org/` si sfruttano le perturbazioni radio generate dai fenomeni atmosferici, come per esempio le scariche elettriche temporalesche (il cosiddetto rumore atmosferico), per ottenere dei numeri 'veramente casuali' e non generati da alcun algoritmo algebrico.

2.2.1 Randomizzare i pazienti

Problema 2.2 *.. stiamo per iniziare una sperimentazione in doppio cieco e dobbiamo randomizzare 30 pazienti. Come potremmo farlo con il computer?*

la funzione `sample` di R

```
tombola = seq(from = 1, to = 90, by = 1)
sample(tombola, 5)
```

Quesito 2.2.1 Nel gioco della tombola di volta in volta, in maniera imprevedibile, sortiscono dei numeri. Anche nel gioco della roulette nelle case da gioco i numeri si susseguono in maniera imprevedibile. Riuscite a cogliere delle analogie, o delle differenze, tra i due giochi?

```
monetina = sample(c(0,1), size = 5, replace = TRUE)
monetina
```

- **variabile aleatoria di Bernoulli**
- **distribuzione binomiale**
- la funzione `set.seed` per la **riproducibilità**

L'esempio della monetina fornisce allora una soluzione semplice al problema 2.2???
.. stranezza:

```
> set.seed(46521)
> sample(c(0,1), size = 30, replace = TRUE)
 [1] 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1
```

Per non correre rischi come questi: la **randomizzazione adattiva**

...

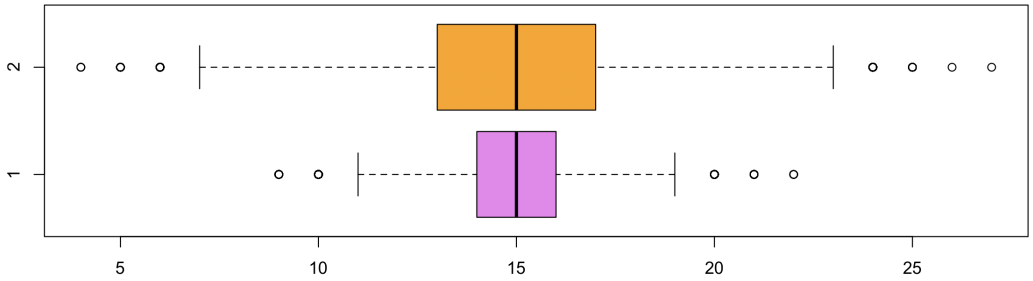


Figura 2.1: Randomizzazione adattiva: il boxplot arancione mostra che simulando per centomila volte il lancio di trenta monetine, non è improbabile che appaiano sequenze che hanno un numero di teste (o croci) anche di molto superiori a venti (o inferiori a dieci). Nel boxplot viola, raffigurante centomila 'randomizzazioni adattive', questo non accade.

2.2.2 Simulare una mutazione genica

Problema 2.3 *.. e come potremmo simulare uno 'snip' nella proteina Zinc Finger 521?*

...

2.3 Le variabili aleatorie con R

2.3.1 Jacob Bernoulli e gli eventi dicotomici

$$\begin{pmatrix} 0 & 1 \\ 1-p & p \end{pmatrix}$$

il dataset `raul`, Annalisa Di Cello [9]

```
www = "http://www.biostatisticaumg.it/dataset/raul.csv"
raul = read.csv(www, header = TRUE)
attach(raul)
names(raul)
table(Esito)
length(Esito)
table(Esito)/length(Esito)
```

$$\begin{pmatrix} 0 & 1 \\ 0.98 & 0.02 \end{pmatrix}$$

Problema 2.4 *Reclutando le prossime dieci pazienti, quale potrebbe essere la probabilità che non vi sia alcun sarcoma tra di esse?*

La distribuzione binomiale

```
caso = sample(c(0,1), size = 100000, prob = c(0.98, 0.02), replace = TRUE)
```

.. idea:

```
tabellona = matrix(caso, nrow = 10000, ncol = 10)
somme = apply(tabellona, MARGIN = 1, FUN = sum)
table(somme)
```

- `dbinom` = **densità** di probabilità della distribuzione **binomiale**

```
teoriche = dbinom(0:10, size = 10, prob =  0.02)
round(teoriche, 3)[1:5]
```

```
[1] 0.817 0.167 0.015 0.001 0.000
```

Spieghiamo i comandi: `size = 10` pazienti, ciascuna ha una `prob = 0.02` maligna. Quale probabilità che nessuna, una, due, tre, eccetera, abbiano il tumore?

- **quantili della distribuzione** = `0:10`

web Se vi interessano le formule della probabilità degli eventi binomiali, le trovate ad esempio su Wikipedia `https://it.wikipedia.org/wiki/Distribuzione_binomiale`

- **probabilità cumulativa** (ovvero la **funzione di distribuzione**) = `pbinom`

Esempio: monetina ripetuto per due volte.

Esercizio 2.1 Provate passo per passo ad eseguire questi comandi e spiegate il loro output:

```
dbinom(x = 0, size = 2, prob = 0.5)
pbinom(q = 0, size = 2, prob = 0.5)

dbinom(x = 1, size = 2, prob = 0.5)
dbinom(x = 0, size = 2, prob = 0.5) + dbinom(x = 1, size = 2, prob = 0.5)
pbinom(q = 1, size = 2, prob = 0.5)

dbinom(x = 2, size = 2, prob = 0.5)
dbinom(x = 0, size = 2, prob = 0.5) + dbinom(x = 1, size = 2, prob = 0.5)
   + dbinom(x = 2, size = 2, prob = 0.5)
pbinom(q = 2, size = 2, prob = 0.5)
```

■

La distribuzione binomiale negativa

Problema 2.5 *Mi sarebbe utile riuscire a stimare quante pazienti dovrò arruolare nello studio prospettico per poter osservare una trentina di sarcomi uterini.*

Studi clinici prospettici: criterio di arresto.

Novità: analisi differenziale di espressione dei geni [22].

web Potete dare un'occhiata su Wikipedia `https://it.wikipedia.org/wiki/Distribuzione_di_Pascal` per farvi qualche idea iniziale.

...

Esercizio 2.2 Per avere un grado di fiducia del 90 per cento di osservare una trentina di sarcomi, quante pazienti dovremmo arruolare nello studio?

```
qnbinom(p = 0.90, size = 30, prob = 0.02)
```

■

2.3.2 L'equazione più pericolosa, parte seconda

Ancora qualche simulazione; `n = 15` lanci di `size = 2` monetine

```
set.seed(234)
rbinom(n = 15, size = 2, prob = 0.5)
```

```
 [1] 1 2 0 2 0 1 2 1 2 1 1 1 1 1 0
```

un milione di lanci, dieci , cento, mille, diecimila monetine

```
set.seed(3456)
monete10    = rbinom(n = 10^6, size = 10,    prob = 0.5)
monete100   = rbinom(n = 10^6, size = 100,   prob = 0.5)
monete1000  = rbinom(n = 10^6, size = 1000,  prob = 0.5)
monete10000 = rbinom(n = 10^6, size = 10000, prob = 0.5)
```

esperimento	n	**media**	**dev. st.** σ	$\sqrt{n}$	**s.e.** $\sigma/\sqrt{n}$
`monete10`	10	5.00	1.60	3.16	0.50
`monete100`	100	50.00	5.00	10	0.50
`monete1000`	1000	500.00	15.80	31.62	0.50
`monete10000`	10000	5000.00	50.00	100	0.50

Colpo di scena: $\sigma/\sqrt{n}$ 'non si muove'

2.3.3 Siméon Poisson e la conta degli eventi

variabile aleatoria di Poisson

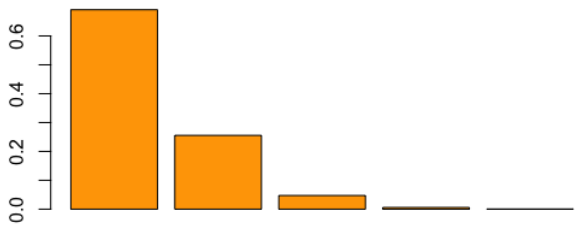
- `dpcR` , reazione a catena della polimerasi digitale
- `edgeR`, analisi empirica dell'espressione genica.

web Se vi interessa approfondire questi due discorsi, iniziate a dare un'occhiata da qui: `https://cran.r-project.org/web//packages/dpcR/dpcR.pdf` e `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/`

esempio di dosimetria / radioterapia oncologica [17, pagina 85]

```
colpiletali = rpois(n = 36, lambda = 0.37)
matrix(colpiletali, nrow = 6)
barplot(dpois(x = 0:4, lambda = 0.37), col = "orange")
```

0	0	2	1	4	0
3	0	0	1	0	0
0	0	1	2	1	1
0	1	0	3	0	0
1	0	1	0	0	2
0	0	0	1	0	1



2.3.4 Carl Gauss, o della normalità

esempio delle monetine

```
monete10000 = rbinom(n = 10^6, size = 10000, prob = 0.5)
hist(monete10000, col = "orange")
```

Esercizio 2.3 Riprendete il dataset `cholesterol` del primo capitolo, ed affiancate all'istogramma viola dei valori del colesterolo HDL dei donatori l'istogramma arancione delle medie di diecimila campioni di cento donatori estratti a caso. Distinguete l'asimmetria dei dati dalla perfetta forma a campana delle medie.

```
indirizzo = "http://www.biostatisticaumg.it/dataset/cholesterol.csv"
cholesterol = read.csv(indirizzo, header = TRUE)
attach(cholesterol)
medieHDLchol = numeric(10000)
for (i in 1:10000) {medieHDLchol[i] = mean(sample(HDLchol, size = 100))}
par(mfrow = c(1,2))
hist(HDLchol, col = "violet")
hist(medieHDLchol, col = "orange")
```

■

web Andate a cercare sul sito di Philipp Plewa l'animazione dedicata al The Central Limit Theorem: `https://bl.ocks.org/pmplewa`. Fantastica, non è vero? La si vede ancor meglio in una delle pagine interne del Chapter 3, Probability Distributions, Discrete and Continuous `https://seeing-theory.brown.edu/`. Inoltre, avremo modo nelle prossime pagine di celebrare anche il nome del cugino di Charles Darwin, sir Francis Galton. Tuttavia qui vi invitiamo a scoprire il suo gioco del *Quincunx* (antica moneta romana): `https://www.mathsisfun.com/data/quincunx.html`. Volendo, potete vedervelo anche nel vostro R:

```
install.packages("visualization")
library(animation)
quincunx()
```

La densità della normale

- **densità della normale standard** = `dnorm`

```
curve(dnorm, from = -3, to = 3)
```

Esercizio 2.4 La variabile aleatoria normale standard è simmetrica rispetto all'origine: il ramo destro della curva è speculare a quello sinistro.

```
dnorm(-1) == dnorm(1)
dnorm(-2.3456) == dnorm(2.3456)
```

■

La probabilità della normale

- `pnorm` = la **funzione di distribuzione cumulativa**

```
pnorm(-1)
1 - pnorm(1)
pnorm(1) - pnorm(-1)
```

Confrontiamo questi valori con il grafico riportato in rete da Wikipedia:

`https://commons.wikimedia.org/wiki/File:Standard_deviation_diagram.svg`

Esercizio 2.5 Verificate che l'area compresa tra -2 e 2, e tra -3 e 3, valgono rispettivamente il 95.4% e il 99.7%. ■

La normale è speciale

Quesito 2.3.1 (Bland [3, pagina 120]) Il picco di flusso espiratorio (PEFR) delle ragazzine di undici anni si distribuisce normalmente con media 300 litri/minuti e deviazione standard 20 litri/minuto.

Quesito 2.3.2 (Pärna et al. [21, pagina 10]) Il consumo di alcol in Estonia nel 1994 si attestava su una media di 128 grammi/settimana con una deviazione standard 147 grammi/settimana.

Pafnutij Čebišëv:

`https://en.wikipedia.org/wiki/Chebyshev%27s_inequality`

Numeri casuali normali

```
set.seed(987)
mu = 3500
sigma = 240
mu + sigma * rnorm(n = 4)
set.seed(987)
rnorm(n = 4, mean = mu, sd = sigma)
```

Quesito 2.3.3 Vi è mai capitato di sentire la frase 'dobbiamo normalizzare i dati'? A vostro giudizio le parole 'normalizzare' o 'standardizzare' hanno lo stesso significato? Discutiamone.

$$Z = \frac{X - \mu}{\sigma}$$

Pertanto, il peso $x = 3800$ grammi di un neonato verrebbe standardizzato ottenendo quello che talvolta si chiama **z score** calcolando

$$z = \frac{3800 - 3500}{240} = 1.25$$

Il grafico quantile-quantile

??? dati normali = istogramma a campana ???

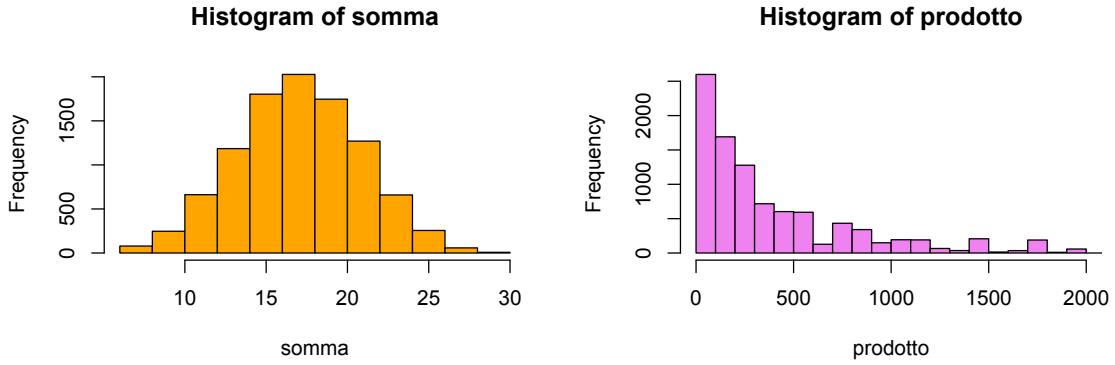
```
set.seed(7155)
normali = rnorm(20)
hist(normali, col = "orange")
```

```
uno = rnorm(70)
due = exp(uno)
par(mfrow = c(1,2))
qqnorm(uno)
qqline(uno)
qqnorm(due)
qqline(due)
```

La distribuzione log-normale

esempio [19]: concentrazione batteri 10^6; una divisione cellulare in più o in meno conduce ad una concentrazione di $2 \cdot 10^6$ o di $5 \cdot 10^5$ = asimmetria

```
d1 = sample(1:6, size = 10000, replace = TRUE)
d2 = sample(1:6, size = 10000, replace = TRUE)
d3 = sample(1:6, size = 10000, replace = TRUE)
d4 = sample(1:6, size = 10000, replace = TRUE)
d5 = sample(1:6, size = 10000, replace = TRUE)
somma = d1 + d2 + d3 + d4 + d5
prodotto = d1 * d2 * d3 * d4 * d5
par(mfrow = c(1,2))
hist(somma, col = "orange")
hist(prodotto, col = "violet", xlim = c(0, 2000), breaks = seq(0,8000,100))
```

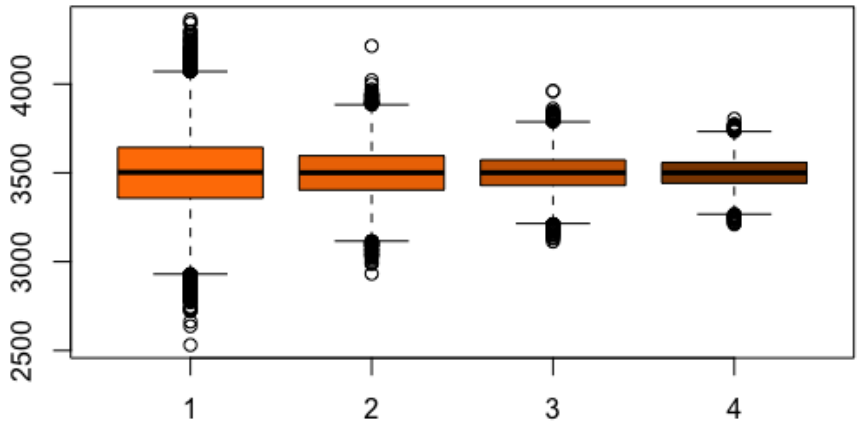


Quesito 2.3.4 Vi ricordate del Problema 1.1, quello di come compilare la Tabella 1 nel paper descrivendo le variabili del vostro studio? In presenza di una distribuzione dei dati come quella colore viola vi verrebbe ancora voglia di utilizzare la convenzione Means $\pm$ SD per riassumere i dati? Ebbene, vi propongo di cercare su PubMed un centinaio di paper che descrivano il Body Mass Index dei loro pazienti, e di contare quanti lo hanno riassunto indicando media e deviazione standard. E poi vi invito a quardare la Figura 1, tipicamente log-normale, di Fonarow G. et al.: `https://www.sciencedirect.com/science/article/pii/S0002870306008271`

2.3.5 L'equazione più pericolosa, parte terza (ed ultima)

?? relazione che intercorre tra i parametri μ e σ che definiscono una variabile aleatoria normale (che nel seguito possiamo considerare come la popolazione di riferimento) e quelli della distribuzione della **media campionaria**, cioè la distribuzione dei dati empirici che si ottengono estraendo dei campioni casuali, di una certa dimensione, dalla popolazione normale??

```
rnorm(n = 176400, mean = 3500, sd = 420)
```



Dimensione n **campione**	**Media** m **della M.C.**	**Deviazione standard** s **della M.C.**
4	3500.4	211.0
9	3500.4	141.5
16	3500.4	106.3
25	3500.4	85.4

ostetriche Andrea, Bruna, Carla e Diana, gruppi di 4, 9, 16 e 25: $\sigma \approx 211.0 \cdot \sqrt{4} \approx 141.5 \cdot \sqrt{9} \approx 106.3 \cdot \sqrt{16} \approx 85.4 \cdot \sqrt{25} \approx 420$.

2.4 Esercizi ed attività di approfondimento

■ **Attività 2.1 — funzioni di R.** Riprendiamo i concetti relativi alle misure di tendenza centrale e di dispersione delle sezioni 1.1 e 1.2. Siccome la deviazione standard (e la varianza) vengono calcolate in relazione alle medie, le quali ovviamente dipendono dalla scala di misura adottata, se si vogliono confrontare due campioni che utilizzano scale di misura diverse (esempio: gradi Kelvin e gradi Farenheit) è opportuno servirsi del **coefficiente di variazione** (talvolta indicato con *RSD*, **relative standard deviation**):

$$CV = \frac{\sigma}{\mu}$$

Realizzate una funzione di R che ne calcoli il valore.
Suggerimento: se non sapete proprio come fare nemmeno dopo aver riletto il paragrafo 2.1.2, guardate questo video.

■

■ **Attività 2.2 — funzioni di R.** Ancora relativamente alle misure di tendenza centrale e di dispersione delle sezioni 1.1 e 1.2, una misura di posizione particolarmente utile ('**robusta**', cioè insensibile agli outlier) è la media del primo e del terzo quartile, che si chiama **midhinge** (il 'perno centrale'):

$$\frac{Q1 + Q3}{2}$$

Realizzate una funzione di R che ne calcoli il valore, utilizzando a vostro piacere le funzioni `quantile`, oppure `summary`.
Suggerimento: osservate come cambia l'output in questi tre esempi, e sfruttate il concetto:

```
summary(iris$Petal.Length)
summary(iris$Petal.Length)[4]
summary(iris$Petal.Length)[[4]]
```

■

■ **Attività 2.3 — variabili aleatorie finite.** ([23, pagina 84]) Nella tabella 4.3 Bernard Rosner riporta le frequenze relative di episodi di otite media riscontrati in una determinata popolazione di neonati durante i loro primi due anni di vita. Pensando alle variabili aleatorie, interpretiamo la prima riga come una rappresentazione dei possibili eventi, la seconda come la densità di probabilità:

$$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 \\ 0.129 & 0.264 & 0.271 & 0.185 & 0.095 & 0.039 & 0.017 \end{pmatrix}$$

Inserite in R questa variabile aleatoria finita con i comandi:

```
evento = 0:6
frequenza = c(0.129, 0.264, 0.271, 0.185, 0.095, 0.039, 0.017)
```

Calcolate il valore atteso (la speranza matematica) e la varianza di questa variabile aleatoria utilizzando le definizioni:

```
speranza = sum(evento * frequenza)
varianza = sum(((evento - speranza)^2) * frequenza)
```

ed osservate che la `varianza` può venir calcolata anche mediante la 'formula' di König (o di Huygens, o di Steiner):

```
sum(evento^2 * frequenza) - speranza^2
```

Generate casualmente 100000 casi di otite media e determinatene la media e la varianza. Cosa notate?

```
casuali = sample(evento, size = 100000, prob = frequenza, replace = TRUE)
mean(casuali)
var(casuali)
```

Da ultimo, confrontate le probabilità teoriche con le frequenze osservate:

```
par(mfrow = c(1,2))
barplot(frequenza)
barplot(table(casuali))
```

■

■ **Attività 2.4 — distribuzione binomiale.** Rileggiamo gli esempi della sezione 2.3.1. Come potremmo usare la funzione `pbinom` per stimare la probabilità che arruolando 10 pazienti sintomatiche, non più di una di loro abbia un sarcoma maligno? E invece, sempre con `pbinom`, la probabilità che almeno due delle dieci pazienti abbia un sarcoma maligno? ■

■ **Attività 2.5 — distribuzione binomiale.** I matematici possono dimostrarvi che, ripetendo per n volte un esperimento bernoulliano di probabilità p, la speranza matematica assume valore $n \cdot p$ e la varianza $n \cdot p \cdot (1-p)$. Sulla falsariga della Attività 2.3, riuscireste a verificare quanto valgono la speranza matematica e la varianza per il sarcoma uterino ($p = 0.02$) in un campione di 450 pazienti sintomatiche? ■

■ **Attività 2.6 — distribuzione di Poisson.** Provate ad ideare una simulazione che vi convinca che in una variabile aleatoria di Poisson di parametro λ la media e la varianza sono coincidenti. ■

■ **Attività 2.7 — distribuzione normale.** [23, pagina 127] Nell'esempio 5.14 Bernard Rosner definisce la misura spirometrica denominata Capacità Vitale Forzata (FVC) come il volume di aria che può essere espirato da una persona con uno sforzo massimale di 6 secondi. Si considera che un bambino abbia una crescita polmonare normale (in senso queteletiano) se la sua FVC standardizzata X stia nel range di $\pm 1.5\sigma$. Supponendo che la FVC standardizzata si comporti come una normale (in senso gaussiano) standard, calcolare la percentuale $Prob(-1.5 \leq X \leq 1.5)$ di bambini che stanno nel range di normalità. ■

■ **Attività 2.8 — distribuzione normale.** [23, pagine 120-121] Negli esempi 5.6 e 5.7 Bernard Rosner afferma che pressione sistolica dei maschi della mia età (giovanissimi, cioè) si distribuisce in maniera gaussiana con media 80 mm Hg e deviazione standard 12 mm Hg. Quanto vale il 90-esimo percentile? E quanti potrebbero essere in percentuale i maschi che hanno una sistolica inferiore a 60 mm Hg? ■

■ **Attività 2.9 — distribuzione normale.** Giovanni Gallo [12] nelle sue tabelle di misurazioni ecografiche biometriche relative all'accrescimento fetale, indica che alla 42-esima settimana di gestazione un feto, 'normalmente', ha un diametro biparietale BPD compreso tra 93 mm (quinto percentile) e 101 mm (novantacinquesimo percentile). Sapreste calcolare la deviazione standard della ipotetica distribuzione gaussiana? ■

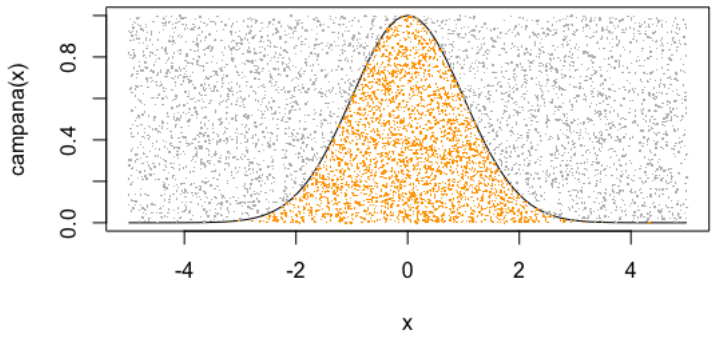
■ **Attività 2.10 — distribuzione uniforme.** La distribuzione aleatoria uniforme è molto semplice da descrivere: essa genera i numeri reali casuali che si dispongono in un intervallo di estremi $[a,b[$ (a compreso, b escluso). Verificate che `valori` vale o zero o $1/4$, ed i `valori` $1/4$ si susseguono sul rettangolo di base da 3 a 7.

```
punti = seq(from = 0, to = 10, by = 0.1)
valori = dunif( punti, min = 3, max = 7)
valori
plot(punti, valori)
punif(5, min = 3, max = 7)
```

Un quesito: `punif(5, min = 3, max = 7)` individua la media, la mediana o la moda della distribuzione? ■

■ **Attività 2.11 — distribuzione uniforme.** Se andate ad esempio su Wikipedia `https://en.wikipedia.org/wiki/Normal_distribution`, oppure se cercate un'immagine della banconota tedesca di dieci marchi dello scorso millennio, `https://www.google.com/search?q=zehn+mark+gauss`, osserverete che la funzione di densità di probabilità della variabile aleatoria normale dipende dalla funzione $\exp(-(x^2/2))$ e presenta accanto a sé uno 'strano' coefficiente, $1/\sqrt{2\pi}$. Questo strano coefficiente ha a che fare con l'area della campana, e viene messo lì in modo tale che l'area valga uno. È soprendente come il numero π salti fuori anche qui, e lo verifichiamo assieme, integrando la funzione $\exp(-(x^2/2))$ con un **metodo Monte Carlo**. Definite la funzione `campana` e generate casualmente una caterva di punti casuali (ad esempio 10^7) che 'bombardino' il rettangolo di base $[-L,L]$ ed altezza $[0,1]$. Individuate quanti ne cadono `aldisotto` della campana con la condizione booleana `ycasuali < campana(xcasuali)` e contateli con `length(aldisotto)`. Adesso dividendo tale quantità per la caterva 10^7 avete una percentuale dell'area del rettangolo di base `2 * L` ed altezza `1` corrispondente proprio all'area arancione della campana di Gauss. A questo punto, indovinate quanto fa $\sqrt{2\pi}$.

```
campana  = function(x){exp(-(x^2)/2)}
L = 5
xcasuali = runif(n = 10^7, min = -L, max = L)
ycasuali = runif(n = 10^7, min = 0, max = 1)
aldisotto = which(ycasuali < campana(xcasuali))
(length(aldisotto)/10^7) * 2 * L * 1
sqrt(2 * 3.1416)
```



Attenzione: non vi venga in mente di dare il comando `plot(xcasuali, ycasuali)`, se non volete impallare il vostro computer per qualche minutino. Se volete replicare il mio disegno, scegliete solo alcune miglialia di puntini, con un codice come questo:

```
curve(campana, -L, L)
points(xcasuali[1:5000], ycasuali[1:5000], pch = ".", col = "grey")
alcuni = aldisotto[1:3000]
points(xcasuali[alcuni], ycasuali[alcuni], pch = ".", col = "orange")
```

■

■ **Attività 2.12 — after hour.** Il professor Morrone è stato un esperto enologo e conoscitore di cocktail. E se anche a voi piace degustare di quando in quando un cocktail (attenzione: bevete con moderazione; l'alcol nuoce alla salute vostra, ed altrui) provate il nostro *Raul*. Ecco gli ingredienti:

- mezzo lime tagliato in tre cunei
- 2.5 cucchiaini di zucchero di canna
- uno spruzzo di Angostura
- 20 ml di Triple Sec
- 20 ml di Curacao
- soda
- una fetta d'arancia per decorazione

In un bicchiere tumbler si pestino con il muddler i tre cunei di lime ricoperti con lo zucchero di canna. Aggiungere uno spruzzo di Angostura e ricoprire di ghiaccio tritato; aggiungere il Triple Sec ed il Curacao e completare con uno splash di soda. Guarnire con fetta d'arancia e servire con cannuccia corta biodegradibile. Mescolare prima di assaggiare. ■

II Seconda Parte

3. C'era una volta il p-value

Problema 3.1 *... bisognerebbe poter dare una qualche significatività ai nostri dati ...*

Problema 3.2 *... vorrei chiedere se il numero di pazienti coinvolti nel mio studio rappresenta una quantità che si possa definire statisticamente significativa ...*

Sfida:

`https://iopscience.iop.org/issue/2041-8205/875/1`

3.1 Il risultato è statisticamente significativo. E dunque?

17 aprile 2014: '*because there is no reliable method for predicting whether a woman with fibroids may have a uterine sarcoma*'.

`http://www.biostatisticaumg.it/dataset/raul.csv`

- CA125 ?
- LDH-1 ed LDH-3 ?

grafici con Orange:

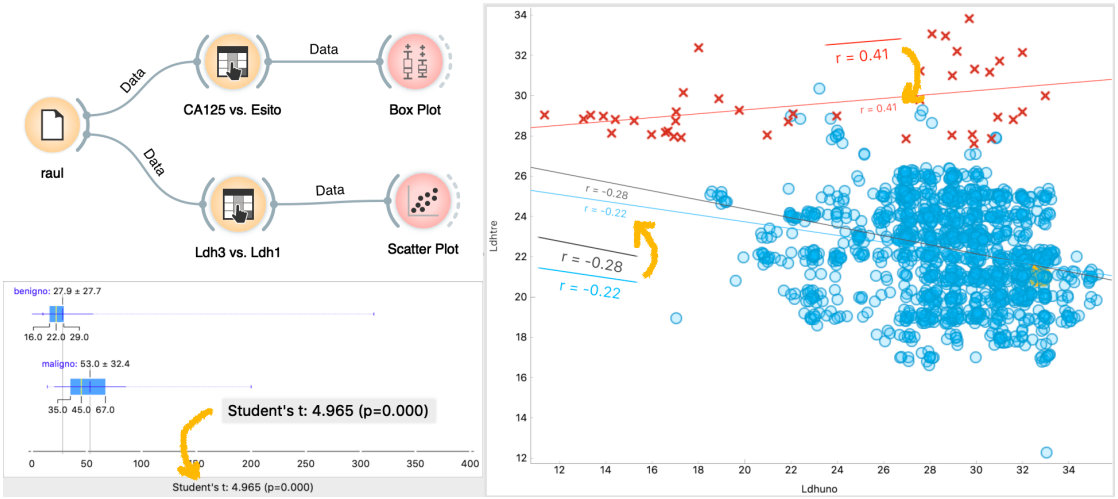


Figura 3.1: Il p-value ci garantisce di aver individuato un biomarcatore clinico efficace?

3.2 Come nacque il t test

Problema 3.3 *... innanzitutto vorrei capire se, in media, i valori del CA125 delle pazienti con il sarcoma siano differenti da quelli delle signore con la patologia benigna.*

web Ci sono mille e mille storie da leggere sulla vita e le opere di William Gosset. Per esempio, questa: `https://www.encyclopediaofmath.org/index.php/Gosset,_William_Sealy`

To test whether it is advantage to kiln-dry barley seed before sowing, seven varieties of barley were sown (both kiln-dried and not kiln-dried) in 1899 and four in 1900; the results are given in the table (...), expressed in Lbs. head corn per acre.

Not Kiln-Dried	Kiln-Dried	Difference
1903	2009	+106
1935	1915	-20
1910	2011	+101
2496	2463	-33
2108	2180	+72
1961	1925	-36
2060	2122	+62
1444	1482	+38
1612	1542	-70
1316	1443	+127
1511	1535	+24

```
differenzaresa = c(106, -20, 101, -33, 72, -36, 62, 38, -70, 127, 24)
mean(differenzaresa)
```

3.2.1 il comando t.test di R

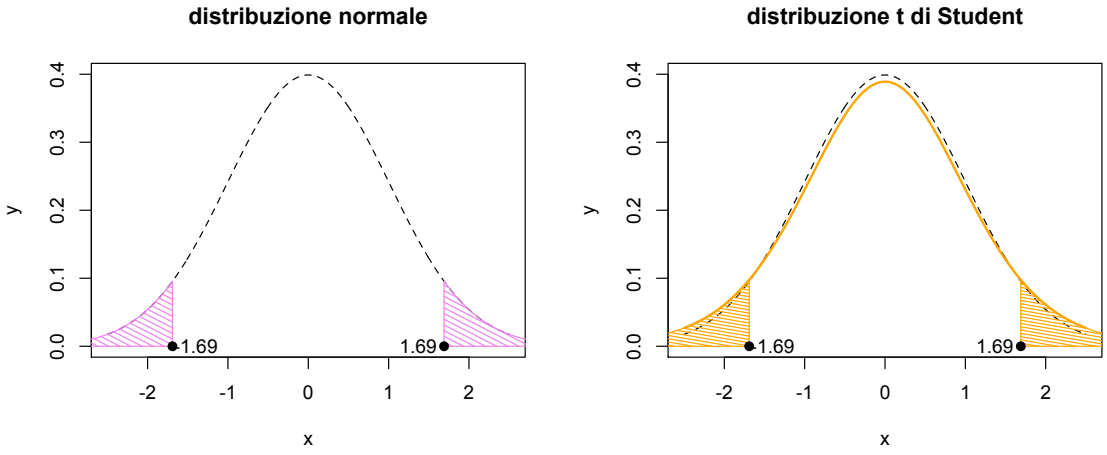
```
t.test(differenzaresa)$estimate
t.test(differenzaresa)$null.value
```

- **rapporto segnale - rumore = test statistic**

$$t = \frac{m - \mu}{s/\sqrt{n}}$$

```
( mean(differenzaresa) - 0 ) / ( sd(differenzaresa) / sqrt(11) )
t.test(differenzaresa)$statistic
```

```
2 * (1 -  pnorm(1.690476))
```




```
length(differenzaresa) - length(mean(differenzaresa))
t.test(differenzaresa)$parameter
```

```
2 * (1 -  pt(q = 1.690476, df = 10))
t.test(differenzaresa)$p.value
```

```
t.test(differenzaresa)
```

3.2.2 il test t di Student tra due campioni

```
www = "http://www.biostatisticaumg.it/dataset/raul.csv"
raul = read.csv(www, header = TRUE)
attach(raul)
names(raul)
table(Esito)
```

modo di procedere scomodo:

```
verde = split(Ca125, Esito)[[1]]
rosso = split(Ca125, Esito)[[2]]
t.test(verde, rosso)
```

Quesito 3.2.1 Il test t di Student ci assicura che possiamo affermare con elevatissimo grado di fiducia che la glicoproteina `Ca125` ha valori mediamente più bassi nelle donne con patologie uterine benigne rispetto a quelle affetta da sarcoma. La domanda importante è: il p-value = 0.000014 ci garantisce che il `Ca125` sia un biomarcatore efficace nella predizione della malattia?

Esercizio 3.1 A vostro giudizio, la media delle età (`Anni`) delle pazienti con il sarcoma è differente da quella delle pazienti con la patologia benigna? Adattate i comandi di `verde` e `rosso` che abbiamo appena visto. ■

3.3 Ascesa e declino del p-value

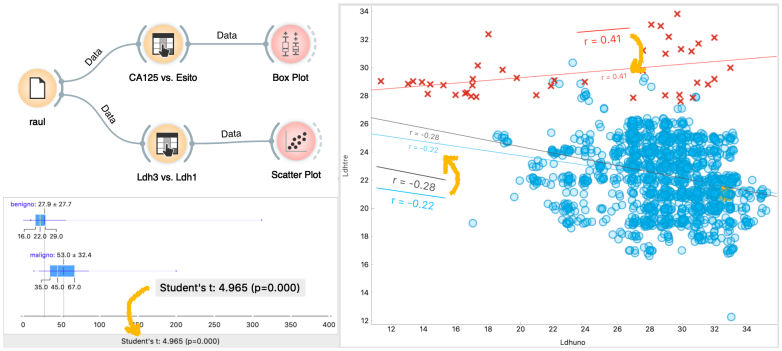
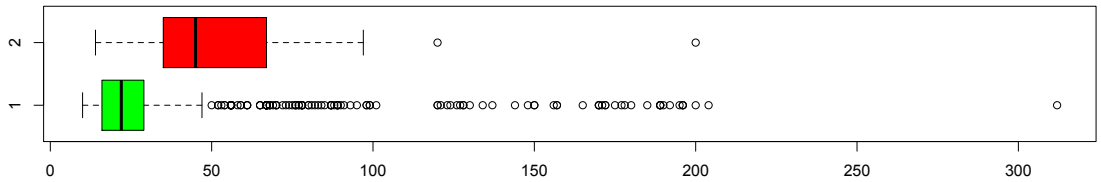


Figura 3.2: Il p-value ci garantisce di aver individuato un biomarcatore clinico efficace?

```
boxplot(verde, rosso, col = c("green", "red"), horizontal = TRUE)
quantile(rosso, .2)
quantile(verde, .9)
```



???? **p-value = 0.000014** ???? Supponiamo CA125 uguale a 32.

1 donna su 5 tra i sarcomi, CA125 < 32, 1 donna su 10 tra le benigne, CA125 > 32

1 sarcoma su 7, CA125 < 28 = media di `verde`

1 benigna su 4, CA125 < 16 1 outlier maligna, CA125 = 14

??? una paziente con CA125 > 100 ??? benigne? maligne (2 volte su 42)

!!!!

ulteriori esempi interessanti: https://it.padlet.com/massimo_borelli/ftmi0gw48r67

storie da ricordare:
1937, sir Ronald Aymler Fisher, *The Design of Experiments*[11]:

> Thus if he wishes to ignore results having probabilities as high as 1 in 20 – the probabilities being of course reckoned from the hypothesis that the phenomenon to be demonstrated is in fact absent –

2008 Stephen Ziliak, Deirdre McCloskey,*The Cult of Statistical Significance*[31] , https://www.deirdremccloskey.com/docs/jsm.pdf

2005, John Ioannidis, *Why Most Published Research Findings Are False*[16] https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.0020124

2015, Richard Horton, *Offline: what is medicine's 5 sigma?*[14]: https://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736%2815%2960696-1.pdf


annuncio - tavola rotonda

Biostatistica: Miseria e Nobiltà

Giovanni Morrone
Fulvio Zullo
Massimo Borelli

Chi fa ricerca nelle scienze della vita e nelle scienze mediche sa che la valutazione dei dati quantitativi riveste un ruolo centrale nel giudicare i risultati di un esperimento o nella diagnosi di una patologia. D'altro canto fare ricerca in maniera continuativa vuol dire porre in atto un confronto aperto e paritetico con la comunità scientifica che, nella veste di osservatore critico e revisore, formula obiezioni ed appunti sui metodi e sui contenuti dell'attività svolta.

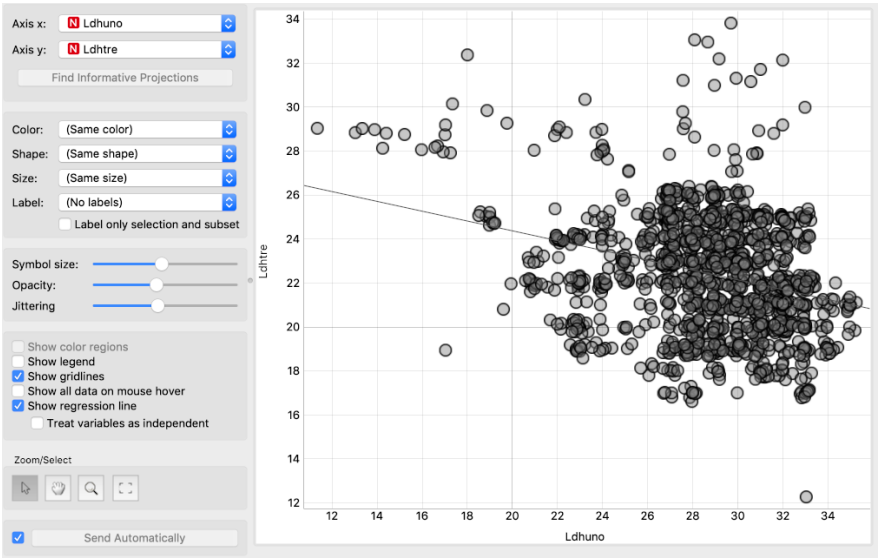
In questa cornice, la Biostatistica risulta essere un prezioso alleato oppure un crudele avversario di chi fa ricerca? Recentemente, Ronald Wasserstein (American Statistical Association) e Richard Horton (the Lancet) hanno posto in evidenza la pesante 'scure' che da anni ormai pende sul collo della Scienza; il ruolo del boia, in questa metafora, purtoppo appartiene proprio alla Biostatistica. L'impiego non appropriato di metodi di analisi e di modelli, spinti in parte dalla frenetica richiesta del *publish or perish*, conduce alla diffusione nella comunità scientifica anche di risultati erronei, o quantomeno non supportati dall'evidenza.

giovedì 7 giugno 2016 ore 11.00

2019, Nature, *Scientists rise up against statistical significance*[1]: https://www.nature.com/articles/d41586-019-00857-9

.. come facciamo a pubblicare i nostri lavori senza essere rigettati dai referee che non sono al corrente di questa profonda mutazione che sta sopraggiungendo?

3.4 La retta di regressione



Problema 3.4 *... noi saremmo interessati a capire come si correlano gli isoenzimi LDH1 ed LDH3 nelle nostre pazienti con la patologia uterina ...*

3.4.1 Covarianza e correlazione

$$cov(x,y) = \frac{1}{n-1}\sum_{1}^{n}(x_i - m_x)(y_i - m_y)$$

$$cor(x,y) = \frac{cov(x,y)}{s_x \cdot s_y}$$

```
cov(Ldhuno, Ldhtre)
cor(Ldhuno, Ldhtre)
```

??? segno meno ???

Esercizio 3.2 Vi ricordate il dataset `airquality` di cento pagine fa? Provate a calcolare il coefficiente di correlazione della temperatura `Temp` rispetto ai livelli di `Ozone`:

```
attach(airquality)
cor(Temp, Ozone)
```

Oops! Cosa succede? Allora provate così:

```
cor(Temp, Ozone, use = "complete.obs")
```

Confrontate questo valore con quel $r = -0.28$ degli isoenzimi Ldh, ed in particolare confrontate quella nuvola di punti con questa:

```
plot(Temp, Ozone)
```

■

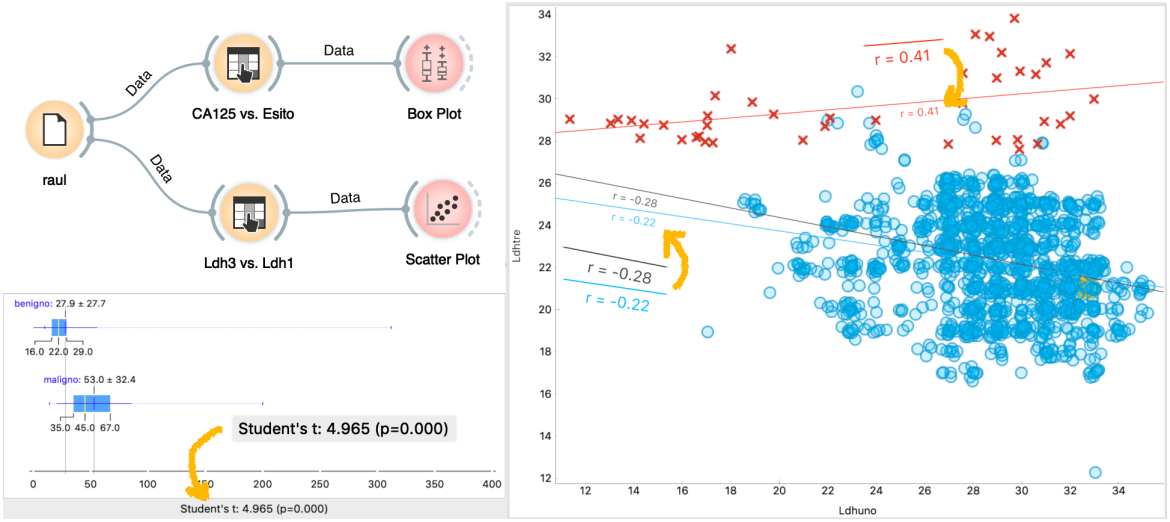
3.4.2 L'idea di Francis Galton

Nel 1886 Francis Galton, *Regression Towards Mediocrity in Hereditary Stature* [13]. **retta di regressione** (? 'retta di progressione'?)

web In rete ci sono due siti mooolto carini che si occupano di queste faccende in maniera visuale:

- `https://setosa.io/`
- `https://seeing-theory.brown.edu`

?? retta rossa + retta azzurra ??? oppure retta nera ???



scoperta (idea furba):

$$\rho = b \cdot \frac{\sigma_x}{\sigma_y}$$

```
-0.223 * sd(Ldhuno) / sd(Ldhtre)
cor(Ldhuno, Ldhtre)
```

Capitolo 4

- il **modello lineare**
  - il **t test**

3.5 Esercizi ed attività di approfondimento

■ **Attività 3.1 — covarianza e correlazione.** Calcolate il coefficiente di correlazione tra la lunghezza dei petali e la lunghezza dei sepali nel dataset più petaloso che c'è, `iris`. ■

■ **Attività 3.2 — correlazione non vuol dire rapporto di causa-effetto.** Si sono sparsi fiumi di inchiostro (e forse se ne dovranno spargere ancora) per spiegare che la correlazione non misura quanto un fenomeno influisca su un altro. Andate a sorridere sul sito `http://www.tylervigen.com/spurious-correlations` nel quale ad esempio si pone in relazione il tasso di divorzi nel Maine con il locale consumo di margarina; o la spesa per la ricerca scientifica e spaziale degli Stati Uniti e i suicidi per impiccagione, soffocamento e strangolamento. ■

4. Che cos'è un modello lineare

Problema 4.1 *... come faccio ad affermare che gli isoenzimi LDH3 ed LDH1 hanno un comportamento diverso nelle pazienti con una patologia benigna rispetto ai casi maligni?*

1. il concetto di **modello lineare**
2. la **selezione del modello**

senza parlare di 'significatività'

> We agree, and call for the entire concept of statistical significance to be abandoned. (...) Rather, and in line with many others over the decades, we are calling for a stop to the use of P values in the conventional, dichotomous way - to decide whether a result refutes or supports a scientific hypothesis.(...) Whatever the statistics show, it is fine to suggest reasons for your results, but discuss a range of potential explanations, not just favoured ones. Inferences should be scientific, and that goes far beyond the merely statistical. Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often more important than statistical measures such as P values or intervals. (Nature [1]:)

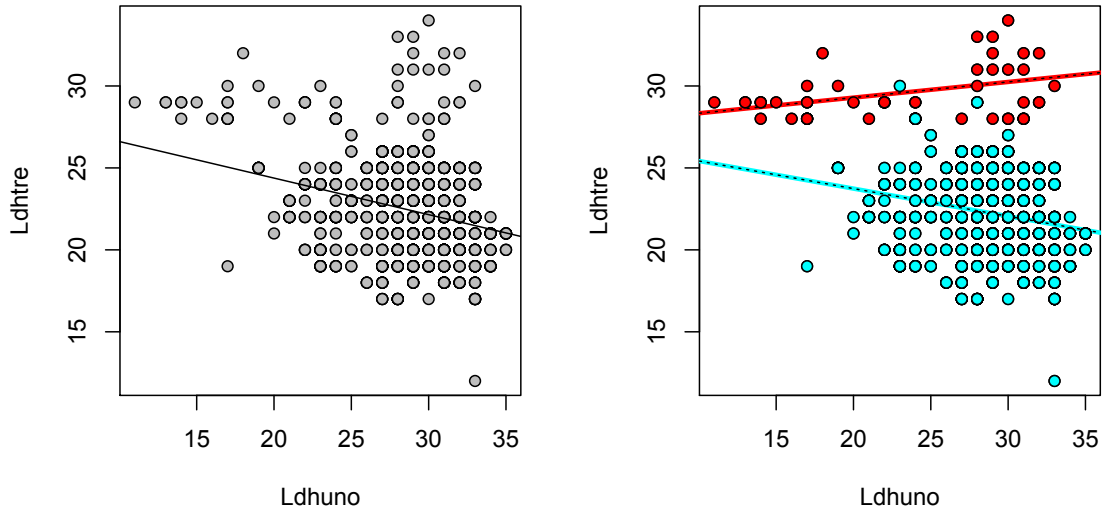


Figura 4.1: Due possibili modelli lineari che 'spiegano' il comportamento di `Ldhtre` in funzione di `Ldhuno`. perché dovremmo preferire quello di destra a quello di sinistra? La risposta ci arriverà alla fine di questo Capitolo, nella Sezione 4.5.

```
formula1 = Ldhtre ~ Ldhuno
```

```
formula2 = Ldhtre ~ Ldhuno * Esito
```

4.1 I dettagli da conoscere

il dataset `fresher`: si tratta di un gruppo di 65 ragazze e ragazzi iscritti (tanti anni fa) al primo anno di medicina dell'università di Trieste:

```
www = "http://www.biostatisticaumg.it/dataset/fresher.csv"
fresher = read.csv( www, header = TRUE )
attach(fresher)
str(fresher)
```

```
> relation1 = weight ~ height
```

- la parola **predittore**

- la retta di regressione del tipo $y = a + b \cdot x$: intercetta a? pendenza b?

- Teorema di Gauss - Markov
- 'Best Linear Unbiased Estimate' ('BLUE')
- metodo dei minimi quadrati (Ordinary Least Square, OLS)

```
lm(relation1)
```

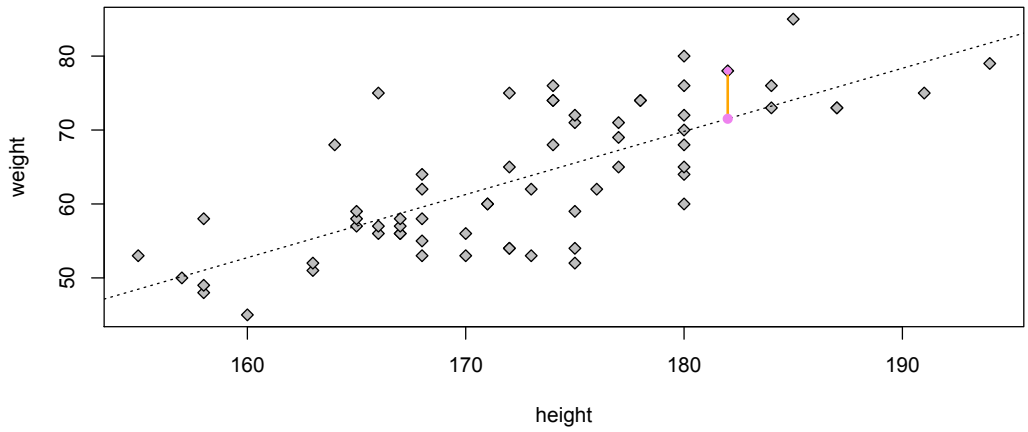
$a = -83.9$ e $b = 0.854$,

$$y = -83.9 + 0.854 \cdot x$$

- due tipi di informazioni: `Coefficients` e `Residuals`

```
model1 = lm(relation1)
summary(model1)
```

4.1.1 I residui di un modello lineare



- **residui**
- residuo arancione:
  - $x = 182$
  - $a + b \cdot 182 = -83.9 + 0.854 \cdot 182 \approx 71.5$
  - $78 - 71.5 = 6.5$

il comando `resid()`

```
summary(resid(model1))
```

Molto interessante: la media è zero;

4.1.2 La devianza di un modello lineare

```
sum(resid(model1)^2)
deviance(model1)
```

4.1.3 La componente aleatoria di un modello lineare

residuo $\varepsilon = 6.5$:

$$78 = -83.9 + 0.854 \cdot 182 + 6.5$$

- **effetti fissi**
- **componente aleatoria**, **componente stocastica**, **effetti casuali**

capire il `Residual standard error`:

Esercizio 4.1 Generate 65 numeri casuali gaussiani con media nulla e deviazione standard 6.46:

```
numericasuali = rnorm(65, mean = 0, sd = 6.46)
```

Mediante questi `numericasuali` generate 65 `pesifittizi`, utilizzando gli effetti fissi a e b:

```
pesifittizi = -83.9 + 0.854 * height + numericasuali
```

Ora disegnate due grafici: a sinistra la nube di punti reale; a destra, la nube di punti fittizi. Non sembra anche a voi che i due grafici si assomiglino?

```
par(mfrow = c(1,2))
plot(height, weight)
plot(height, pesifittizi)
```

■

Confrontate situazione dataset `raul`:

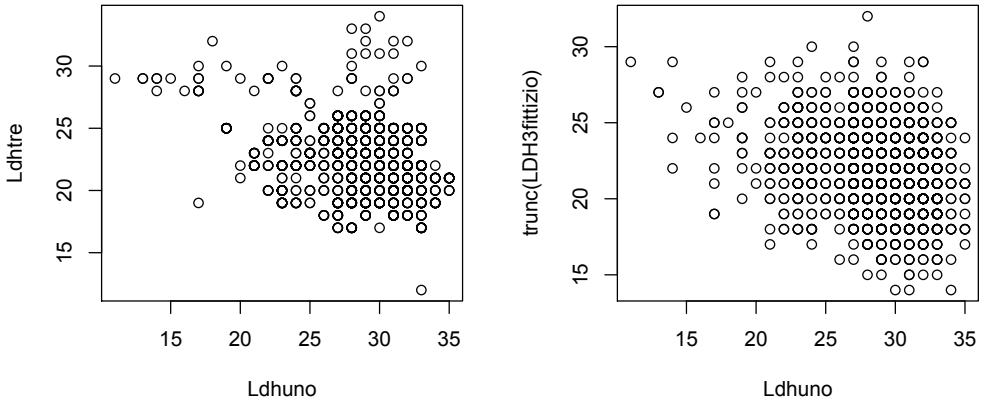
```
formula1 = Ldhtre ~ Ldhuno

modello1 = lm(formula1)
modello1
summary(modello1)$sigma
```

- $y = 28.8 - 0.223 \cdot x$
- residui ε , media nulla e deviazione standard $\sigma = 2.5$

Ripetiamo:

```
numericasuali = rnorm(1610, mean = 0, sd = 2.5)
LDH3fittizio = 28.8 - 0.223 * Ldhuno + numericasuali
par(mfrow = c(1,2))
plot(Ldhuno, Ldhtre)
plot(Ldhuno, trunc(LDH3fittizio))
```



Prima risposta al Problema 4.1: ... *come faccio ad affermare che gli isoenzimi LDH3 ed LDH1 hanno un comportamento diverso nelle pazienti con una patologia benigna rispetto ai casi maligni?*

4.1.4 Il modello nullo è importante

.. una fattucchiera: predire il `weight` conoscendo il `day` del compleanno:

```
formulasciocca = weight ~ day
modellosciocco = lm(formulasciocca)
modellosciocco
mean(weight)
```

$$y = 63.52 + 0.000016 \cdot x$$

Notazione di Wilkinson e Rogers [29]:

```
relation0 = weight ~ 1
modellonullo = lm(relation0)
summary(modellonullo)
mean(weight)
sd(weight)
```

interpretare il `modellonullo` – Capitolo 1 statistica descrittiva, modo '**parametrico**'.

4.1.5 Hirotugu Akaike, un nome da ricordare per sempre

dataset `fresher`

pendenza del `model1` era $b = 0.854$

```
( b = summary(model1)$coefficient[2] )
```

Sezione 3.4.2: coefficiente ρ

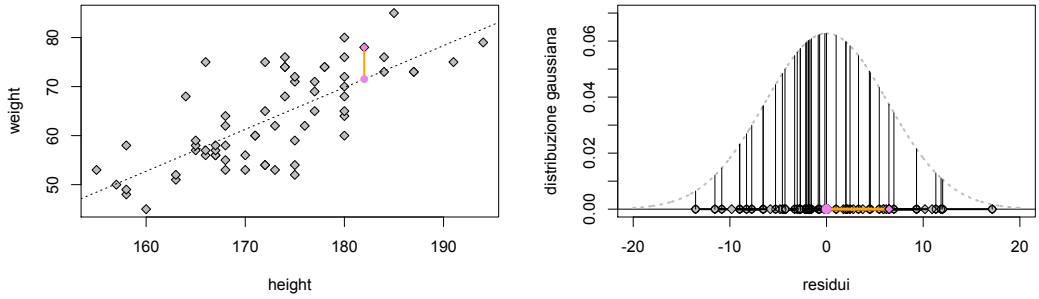
```
b * sd(height) / sd(weight)
cor(height, weight)
```

coefficiente di determinazione R^2, `Multiple R-squared`:

```
cor(height, weight)^2
summary(model1)$r.squared
```

.. 'quanto manca' alla perfezione del 100 per cento??

```
deviance(model1)
model0 = lm(weight ~ 1)
deviance(model0)
1 - deviance(model1)/deviance(model0)
```



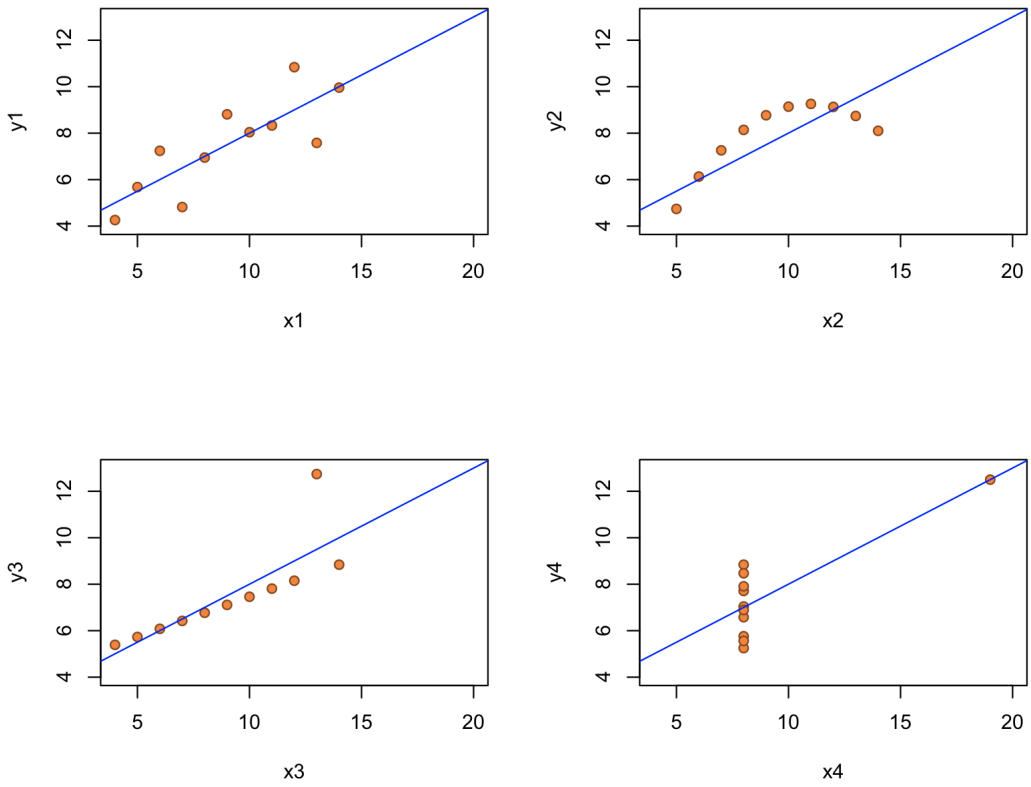

```
logLik(model1)
```

.. ma soprattutto:

```
AIC(model1)
```

4.1.6 La diagnostica del modello lineare

`https://en.wikipedia.org/wiki/Anscombe%27s_quartet`



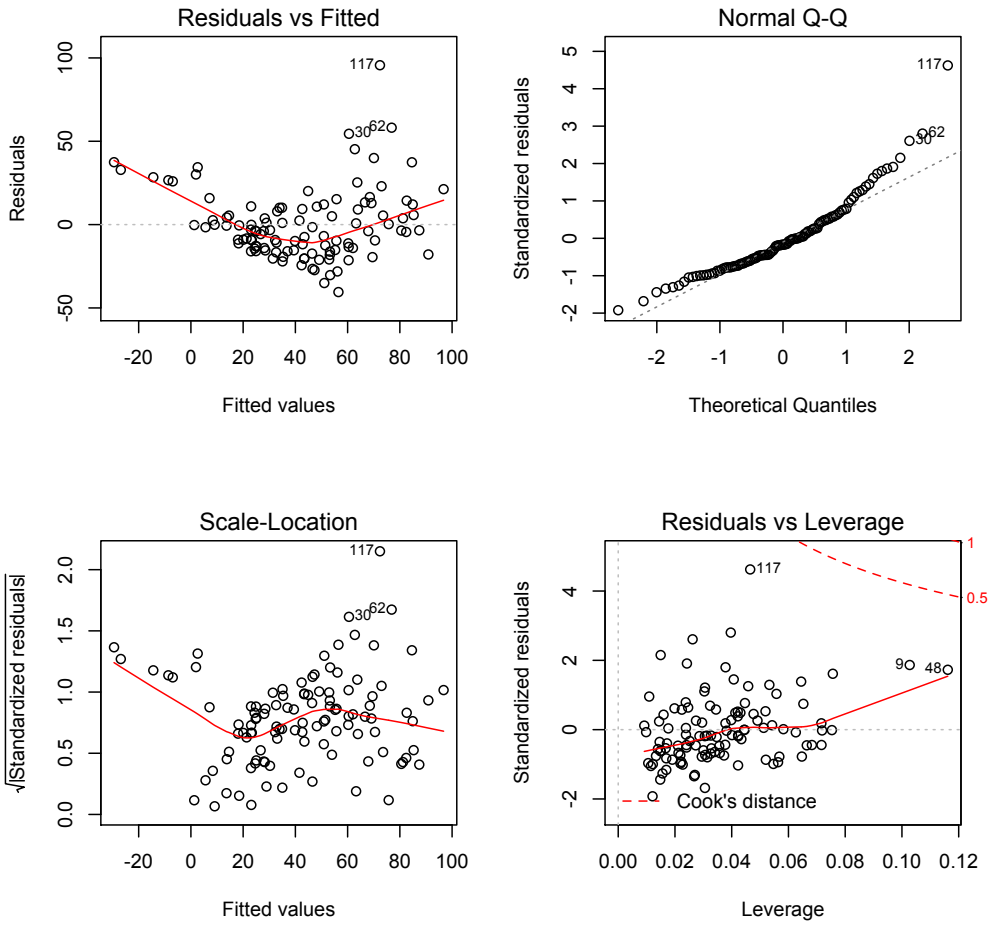
- $y = 3 + 0.5 \cdot x$
- $R^2 = 2/3 = 0.66$

Esercizio 4.2 Provate a riproporre da voi il disegno che trovate su Wikipedia, e che è del tutto analogo a quello raffigurato qui sopra. Avrete bisogno del comando `attach(anscombe)` per importare i dati, e poi con `str(anscombe)` oppure direttamente con `anscombe` vi potrete fare un'idea di quanti siano i dati e di come si chiamino le variabili. Ricordatevi che con lo stranissimo comando `par(mfrow = c(2,2))` si riescono ad organizzare quattro `plot` in maniera affiancata, due per due. ■

```
attach(airquality)
ipotesi1 = Ozone ~ Solar.R + Wind + Temp
modellodiscutibile = lm(ipotesi1)
summary(modellodiscutibile)
```

? ottimo p-value, R^2 supera il 60 per cento: potrebbe andare bene, no? No!

```
par(mfrow = c(2,2))
plot(modellodiscutibile)
```



Esercizio 4.3 Controllate che nel dataset `anscombe`, il terzo esempio esibisce un punto isolato:

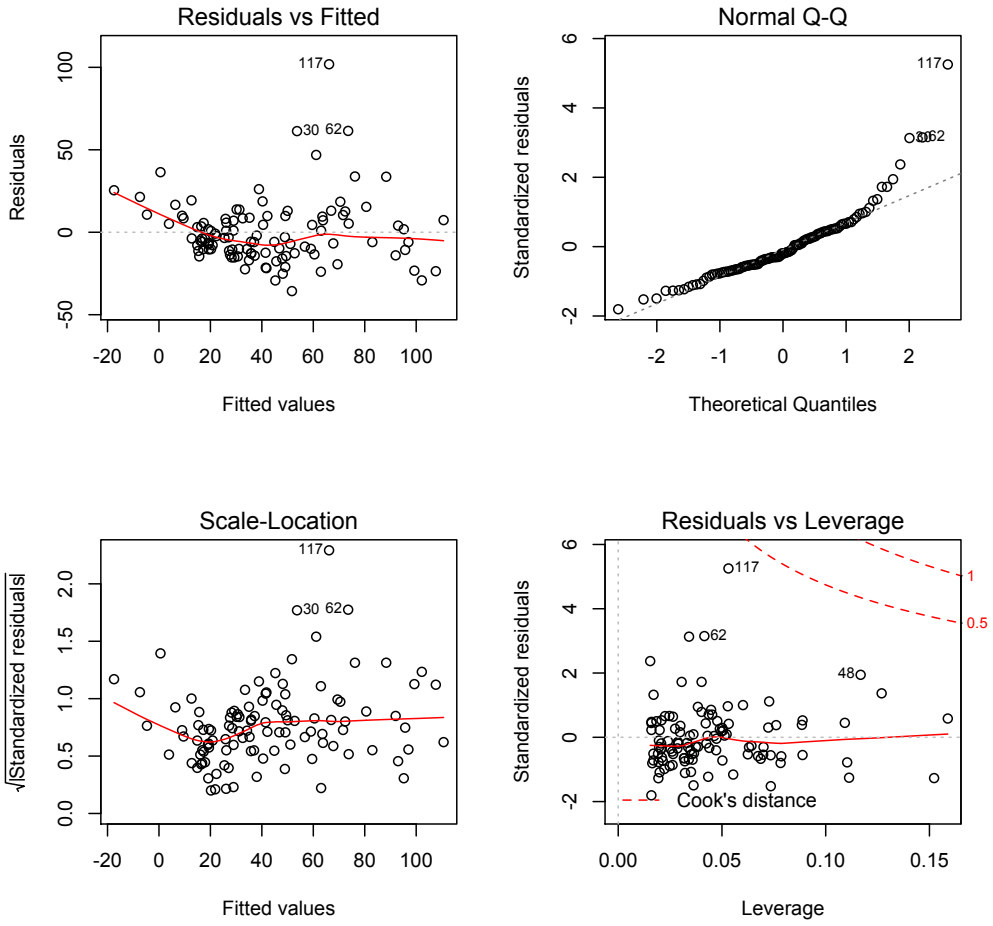
```
> attach(anscombe)
> m3 = lm(y3 ~ x3)
> par(mfrow = c(2,2))
> plot(m3)
```

■

4.1.7 Lineare non è sinonimo di rettilineo

```
ipotesi2 = Ozone ~ Solar.R + Wind + Temp + I(Temp^2)
modellomigliore = lm(ipotesi2)
summary(modellomigliore)

par(mfrow = c(2,2))
plot(modellomigliore)
```



4.1.8 Anche il t test è un modello lineare

ricordo:

```
differenzaresa = c(106, -20, 101, -33, 72, -36, 62, 38, -70, 127, 24)
t.test(differenzaresa)
```

il modello nullo:

```
ipotesinulla = differenzaresa ~ 1
modelloGosset = lm(ipotesinulla)
summary(modelloGosset)
```

calcoliamo:

```
sd(differenzaresa)
sd(differenzaresa)/sqrt(11)
```

$$t = \frac{x_m - \mu}{s/\sqrt{n}} = \frac{33.7 - 0}{19.95} = 1.69$$

Esercizio 4.4 Riprendete la Sezione 3.2.2 nella quale avevamo considerato le 1568 pazienti benigne del dataset `raul` e le 42 maligne 'splittandole' in due gruppi (quello `verde` e quello `rosso`) in modo tale da capire se il biomarcatore `Ca125` si esprimesse in maniera differente tra i due gruppi, in senso statistico. Ecco uno stralcio dell'output:

```
t = -4.9067, df = 42.584, p-value = 1.401e-05
...
mean of x mean of y
 27.94388  52.97619
```

Controllate che il seguente `modellobiomarker` offre delle informazioni apparentemente diverse:

```
ipotesibiomarker = Ca125 ~ Esito
modellobiomarker = lm(ipotesibiomarker)
summary(modellobiomarker)
```

■

Esercizio 4.5 Riprendete l'esercizio precedente, e convincetevi immediatamente sul fatto che il `Ca125` non è affatto un valido predittore dell'`Esito`, esaminando i plot diagnostici:

```
par(mfrow = c(2,2))
plot(modellobiomarker)
```

■

dataset `fresher`. Anche il `gender` è un predittore:

```
relation2 = weight ~ gender
model2 = lm(relation2)
summary(model2)
```

vedremo ora come 'unire' i predittori.

4.2 Ancova: unire i predittori numerici ai fattori

Vi ricordate la domanda:

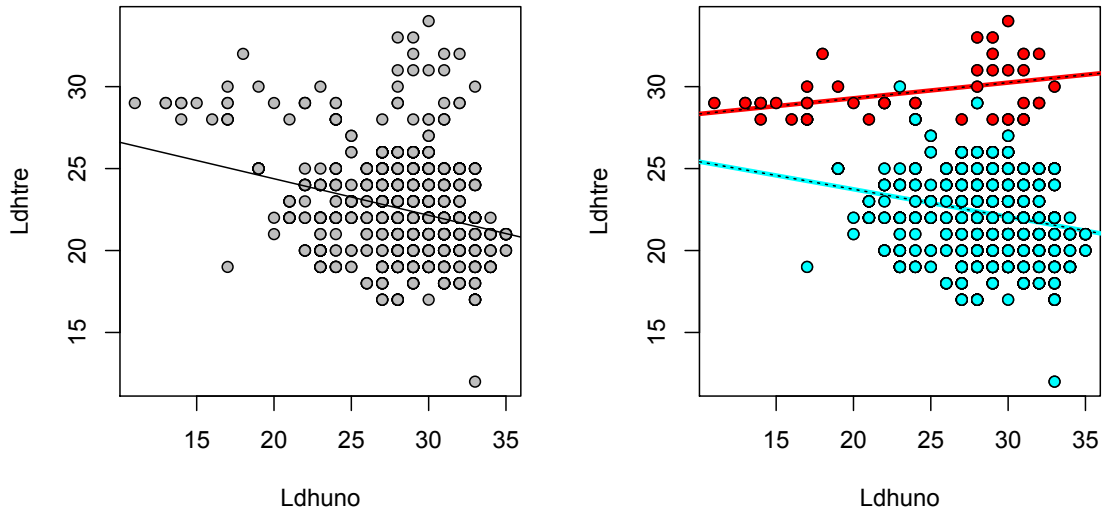


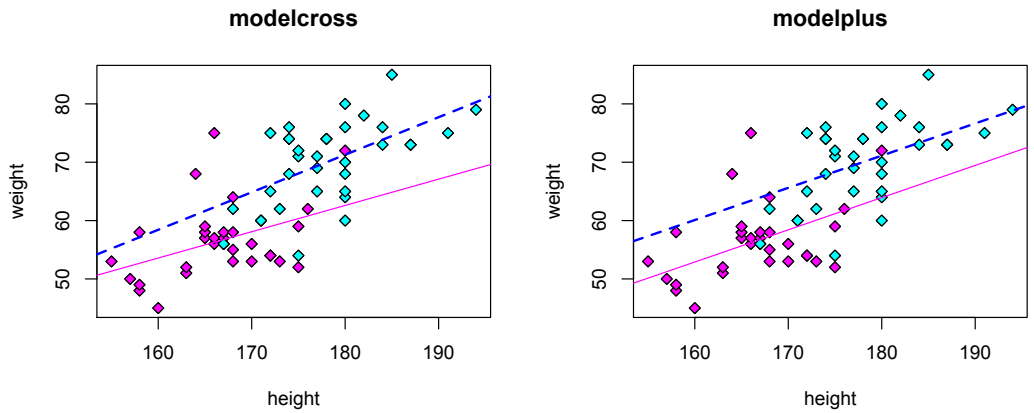
Figura 4.2: Due possibili modelli lineari che 'spiegano' il comportamento di `Ldhtre` in funzione di `Ldhuno`. perché dovremmo preferire quello di destra a quello di sinistra?

```
formula1 = Ldhtre ~ Ldhuno

ipotesibiomarker = Ca125 ~ Esito

formula2 = Ldhtre ~ Ldhuno * Esito
```

come 'mostrare' che `formula2` è 'meglio' di `formula1`?



Spieghiamo :

```
relationcross = weight ~ height * gender
relationplus = weight ~ height + gender
```

```
modelcross = lm(relationcross)
summary(modelcross)
```

la retta rosa $y = -18.50 + 0.45 \cdot x$
la retta blu: $y = (-18.50 - 25.96) + (0.45 + 0.19) \cdot x$, ossia $y = -44.46 + 0.64 \cdot x$.

Wilkinson e Rogers, simbolo :

```
relationcross = weight ~ height * gender
relationcross = weight ~ height + gender + height:gender
```

```
modelplus = lm(relationplus)
summary(modelplus)
```

la retta rosa $y = -35.4 + 0.55 \cdot x$
la retta blu è $y = -28.2 + 0.55 \cdot x$

4.3 Facciamo il punto della situazione

Riassumiamo, dataset `fresher`:

```
relation0 = weight ~ 1
relation1 = weight ~ height
relation2 = weight ~ gender
relationplus = weight ~ height + gender
relationcross = weight ~ height * gender
```

1. `model0` modello lineare nullo, retta orizzontale $y = 63.5$, statistica descrittiva
2. `model1` la retta di regressione $y = -83.9 + 0.85 \cdot x$
3. `model2` t test di Student, due rette orizzontali, la retta 'rosa' $y = 56.6$ e la retta 'azzurra' $y = 70.2$.
4. `modelplus` *ancova without interaction*.
5. `modelcross` *ancova with interaction*.

paradigma: principio di parsimonia di Ockham, *frustra fit per plura quod fieri potest per pauciora*.

Si intuisce immediatamente il fatto che, aumentando il numero di parametri di un modello e calibrandoli sui dati raccolti nel nostro dataset di interessa andiamo incontro al rischio di **overfitting**: `https://en.wikipedia.org/wiki/Overfitting`

```
AIC(model0, model2, model1, modelplus, modelcross)
```

`modelplus` potrebbe essere il **modello minimale adeguato**?

4.4 Anova: la generalizzazione del t test

dataset `fresher`

```
levels(gym)
```

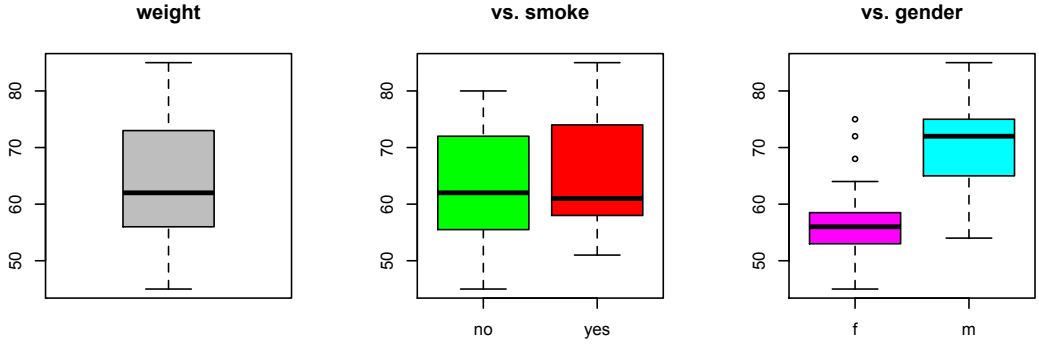
```
relation3 = weight ~ gym
```

Osservate:

```
t.test(relation3)
```

$$t = \frac{x_m - \mu}{s/\sqrt{n}}$$

ecco il perché del nome *An.o.va.*, *An*alysis *o*f *va*riance

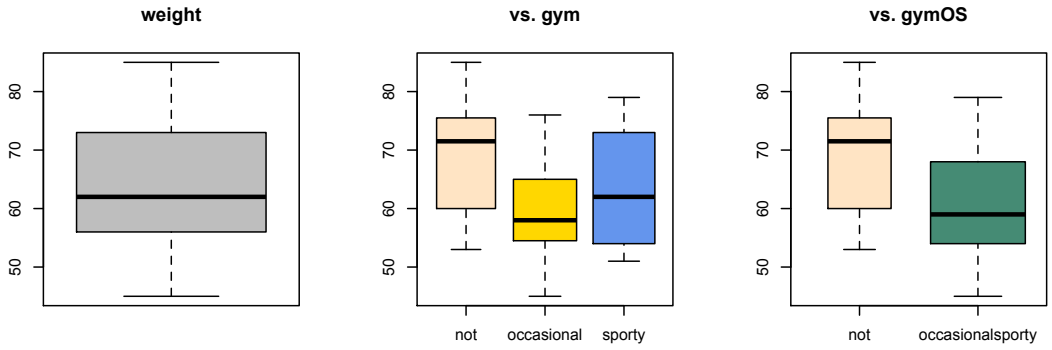


gruppo	media	dev. standard
tutti i 65 `fresher`	63.5	9.6
51 non fumatori	63.1	9.5
14 fumatori	65.1	10.3
32 femmine	56.6	6.4
33 maschi	70.2	7.1

Proseguiamo, due relazioni:

```
relation0 = weight ~ 1
relation3 = weight ~ gym
```

```
boxplot(weight ~ gym, main = "vs. gym", col = c("bisque", "gold",
    "cornflowerblue"), varwidth = TRUE)
```




```
model3 = lm(relation3)
AIC(model0, model3)
```

'Quando c'era il p-value', avremmo detto che 'la Anova è significativa'

osservazione: `occasional` e `sporty` hanno mediane simili. Proviamo a raggrupparli?

```
gymOS = gym
levels(gymOS)
levels(gymOS)[2] = "occasionalsporty"
levels(gymOS)[3] = "occasionalsporty"
levels(gymOS)
```

```
relation3bis = weight ~ gymOS
model3bis = lm(relation3bis)
AIC(model0, model3bis, model3)
```

C'era una volta:

- **multiple comparison**
- **confronti multipli**
- **post-hoc analysis**

C'era una volta: 'correzione di Bonferroni' ... 'residui studentizzati in maniera onesta secondo Tukey' .... del 'false discovery rate secondo Benjamini' ...

Esercizio 4.6 Verificate che adottare il modello `model3ter` è veramente una pessima idea:

```
gymNS = gym
levels(gymNS)[1] = "notsporty"
levels(gymNS)[3] = "notsporty"
relation3ter = weight ~ gymNS
model3ter = lm(relation3ter)
AIC(model0, model3bis, model3, model3ter)
```

■

? **two-way anova** ? **two-way anova with interaction** ?

```
relationtwowayplus = weight ~ gym + sport
relationtwowaycross = weight ~ gym * sport
```

4.5 La meta finale: condurre un'analisi multivariabile

Vi ricordate la domanda:

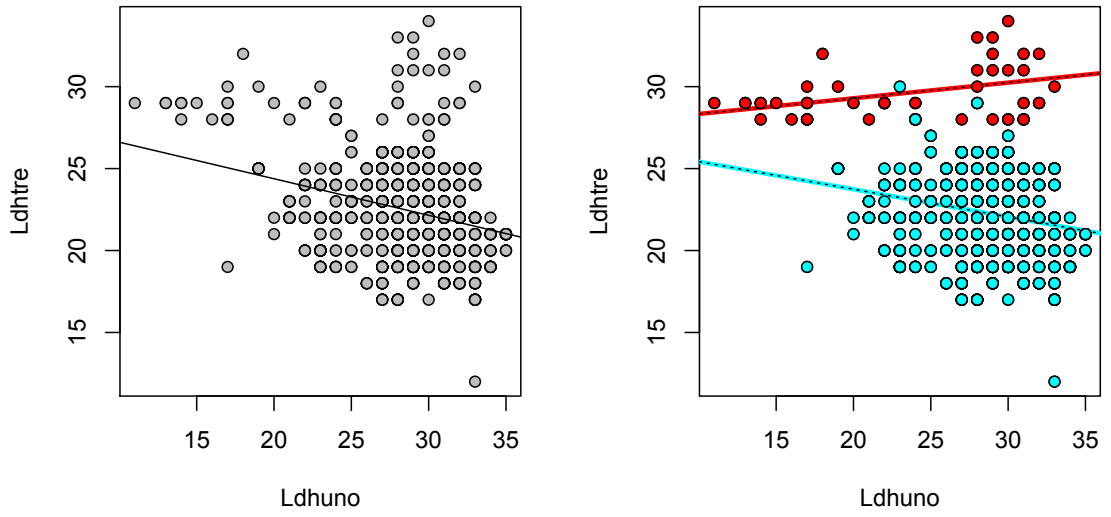


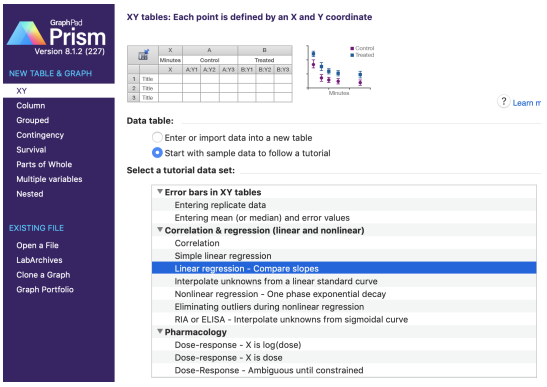
Figura 4.3: Due possibili modelli lineari che 'spiegano' il comportamento di `Ldhtre` in funzione di `Ldhuno`. perché dovremmo preferire quello di destra a quello di sinistra?

Abbiamo quindi il modo di rispondere:

```
uno = lm(Ldhtre ~ Ldhuno)
due = lm(Ldhtre ~ Ldhuno * Esito)
AIC(uno, due)
```

come condurre un'analisi multivariabile

notate la differenza essenziale:



Step 1: ricercare il modello minimale adeguato di tipo additivo.

```
maximalrel = weight ~ gender + height + shoesize + smoke + gym +
     + systolic + diastolic
maximalmodel = lm(maximalrel)

step(maximalmodel)

formulaTemporanea01 = weight ~ height + shoesize + gym
modelloTemporaneo01 = lm(formulaTemporanea01)
```

Step 2: si controllano i fattori a più livelli.

```
formulaTemporanea02 = weight ~ height + shoesize + gymOS + heartrate
modelloTemporaneo02 = lm(formulaTemporanea02)
AIC(modelloTemporaneo01, modelloTemporaneo02)
```

Step 3: ricercare se vi siano termini di interazione.

```
formulaTemporanea03 = weight ~ height * shoesize + gym
formulaTemporanea04 = weight ~ height + shoesize * gym
formulaTemporanea05 = weight ~ height * gym + shoesize
modelloTemporaneo03 = lm(formulaTemporanea03)
modelloTemporaneo04 = lm(formulaTemporanea04)
modelloTemporaneo05 = lm(formulaTemporanea05)
AIC(modelloTemporaneo01, modelloTemporaneo03, modelloTemporaneo04,
 modelloTemporaneo05)
```

Step 4: ricercare se vi siano termini di curvatura.

```
formulaTemporanea06 = weight ~ height + shoesize + gym + I(height^2)
formulaTemporanea07 = weight ~ height + shoesize + gym + I(shoesize^2)
modelloTemporaneo06 = lm(formulaTemporanea06)
modelloTemporaneo07 = lm(formulaTemporanea07)
AIC(modelloTemporaneo01, modelloTemporaneo06, modelloTemporaneo07)
```

Step 5: stabilire se l'intercetta abbia rilevanza nel modello

```
formulaTemporanea08 = weight ~ height + shoesize + gym - 1
modelloTemporaneo08 = lm(formulaTemporanea08)
AIC(modelloTemporaneo01, modelloTemporaneo08)
```

Step 6: eseguire la diagnostica del modello

Esercizio 4.7 Eseguite la diagnostica del modello minimale adeguato:

```
modminadeg = lm(weight ~ height + shoesize + gym)
par(mfrow = c(2,2))
plot(modminadeg)
```

■

4.5.1 Interpretare il modello minimale adeguato

```
modminadeg = lm(weight ~ height + shoesize + gym)
summary(modminadeg)
```

commento: predittori non interagiscono

commento: `gender`? `shoesize` **variabile proxy**

```
table(gender, shoesize > 40)
```

interpretare: studente di statura 182 , scarpe numero 43 e che non è particolarmente sportivo:

$$-71.8 + 0.356 \cdot 182 + 1.91 \cdot 43 = 75.3$$

errore standard residuale 4.5? ricordatevi `rnorm`

4.6 Riassuntone del capitolone

...

4.7 Perle di saggezza: tecniche di linearizzazione

4.7.1 Il modello iperbolico.

...

4.7.2 Il modello esponenziale.

...

4.7.3 Il modello *maxima function*.

...

4.7.4 Il modello potenza.

...

4.7.5 Il modello logistico.

1838 Verhulst,

$$y = \frac{K}{1 + A \cdot \exp(B \cdot x)}$$

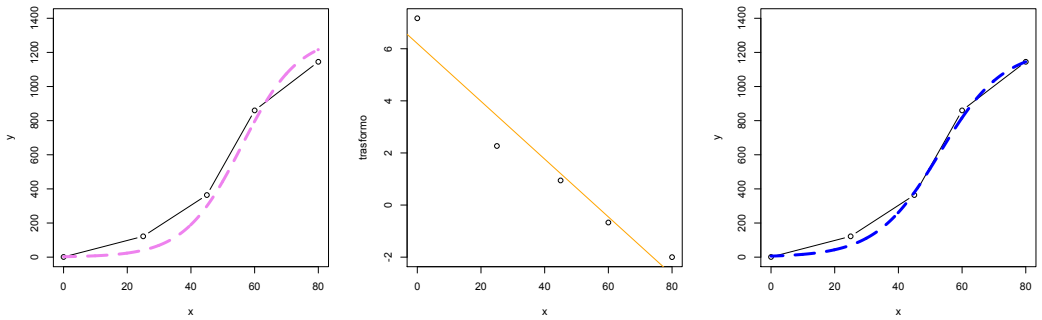
Teresa Calimeri, [4], dataset `mieloma.csv`:

```
www = "http://www.biostatisticaumg.it/dataset/mieloma.csv"
mieloma = read.csv(www, header = TRUE)
attach(mieloma)
```

```
x = time[1:5]
y = paraprotein[1:5]
y[1] = 1
```

```
stimaK = 1300
```

```
trasformo = log(- 1 + stimaK/y)
modello = lm (trasformo ~ x)
modello
```



$\exp(6.202) = 494$ e -0.111

```
( A = exp(modello$coefficients[[1]]) )
( B = modello$coefficients[[2]] )
```

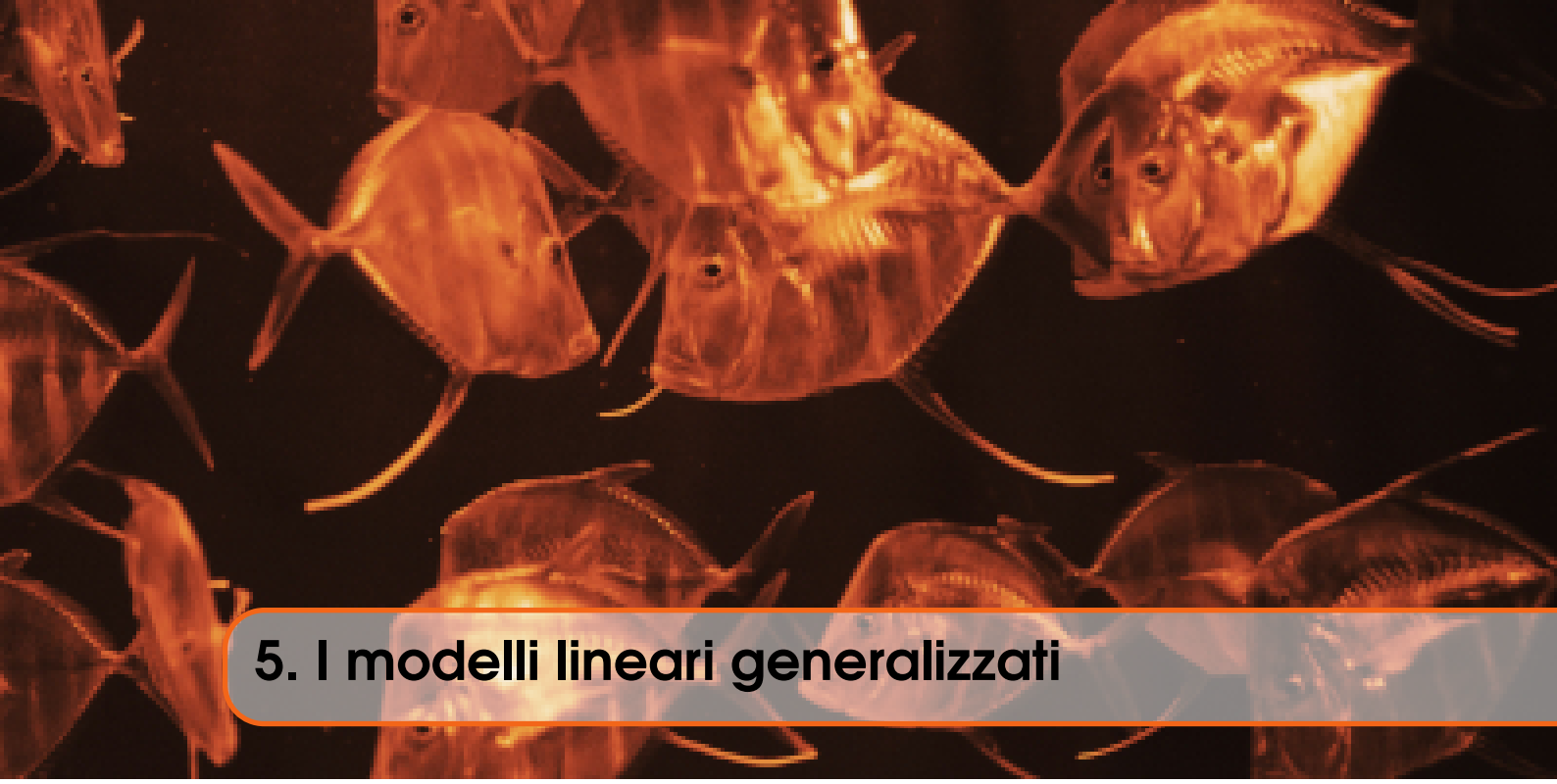
```
curva = nls(y ~ kappa / ( 1 + A * exp(B * x) ),
             start =  list(kappa = 500, A = 2.52, B = -0.43), trace = TRUE)
summary(curva)
```

4.8 Esercizi ed attività di approfondimento

■ **Attività 4.1 — retta di regressione.** Una delle prime cose che si impara a proposito delle retta di regressione è che essa "passa" per il baricentro della nuvola di punti. Per baricentro, o centro di massa, in statistica si intende proprio il punto che ha per coordinate la media dei dati delle ascisse, e la media dei dati delle ordinate. Per esercizio, provate a realizzare con R il grafico di sinistra della Figura 4.1, facendo apparire un bel punto blu a forma di diamante che indichi il baricentro, e verificate che la retta di regressione lo attraversa. ■

■ **Attività 4.2 — Ancova.** Leggete il paper di Roberta Venturella [25] dall'indirizzo `http://www.biostatisticaumg.it/biostatistica/jmig.pdf` e verificate quanto viene affermato usando i dati originali del dataset `salpinge` pubblicati all'indirizzo: `http://www.biostatisticaumg.it/dataset/salpinge.csv` ■

■ **Attività 4.3 — Ancova.** Molto interessante anche il dataset `heather` che potete importare dall'indirizzo `http://www.biostatisticaumg.it/dataset/heather.csv`. Trovate il modello minimale adeguato, rispetto alla risposta `znf`. ■



5. I modelli lineari generalizzati

dataset `raul` (risk assessment in uterine lesions) di Annalisa Di Cello [9]:

```
www = "http://www.biostatisticaumg.it/dataset/raul.csv"
raul = read.csv(www, header = TRUE)
attach(raul)
names(raul)

indicedirischio1 = Esito ~ Anni + Ca125 + Ldhuno + Ldhtre
```

5.1 I dettagli da conoscere

Vocabolario 5.1 — modello lineare generalizzato. Un modello lineare generalizzato è un insieme di tre strumenti matematici:

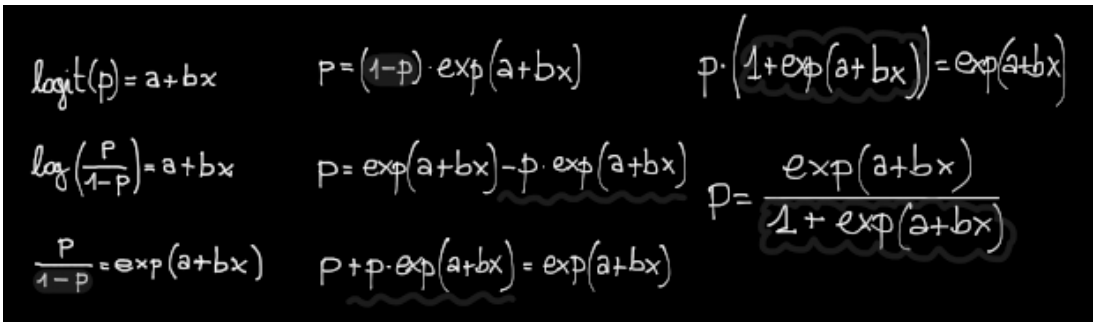
1. una **relazione**, che usualmente viene denominata il **predittore lineare**, la quale connette la risposta del dataset con una o più covariate. Ad esempio, il nostro `indicedirischio1`.
2. una famiglia di **variabili aleatorie** adatte a modellare la risposta (o, per essere giustamente più precisi, i residui del predittore lineare). Ad esempio, la distribuzione binomiale per l'`Esito`.
3. una **funzione di collegamento** che trasforma ('inietta') il valore atteso della variabile aleatoria che modella la risposta nel valore medio del predittore lineare. E qui dobbiamo fare qualche parole di approfondimento, perché nei modelli lineari del Capitolo precedente questo argomento era stato sottaciuto, in quanto 'invisibile'.

5.1.1 La funzione di collegamento

prova:

```
indicedirischio2 = Esito ~  Ca125
```

unità logistica, o **logit**: $logit(p) = \log(p/(1-p))$.



$$p = \frac{\exp(a + b \cdot x)}{1 + \exp(a + b \cdot x)}$$

Esercizio 5.1 Disegnate una sigmoide. Vedete che la ascissa y (e sottolineo ascissa, ossia, l'asse orizzontale) è una variabile numerica continua, mentre `sigmoide` diventa un valore di probabilità, sempre compreso tra 0 ed 1 (sull'asse verticale delle ordinate):

```
y = seq(-6, 6, 0.1)
p = exp(y) / (1 + exp(y))
plot(y, p)
```

■

5.1.2 Ad ogni variabile aleatoria, la sua funzione di link

un riassuntino della situazione:

modello	**variabile aleatoria**	**(inversa della) funzione di link**
lineare	`family = gaussian`	identità, $v = u$
regressione logistica	`family = binomial`	sigmoide, $v = exp(u)/(1+exp(u))$
regressione di Poisson	`family = poisson`	esponenziale, $v = exp(u)$

5.1.3 Interpretare una regressione logistica

il caso di un fattore

un esempio didattico, *fake*:

```
Ca125High = Ca125 > median(Ca125)
table(Esito, Ca125High)
```

odds ratio : $(818 \cdot 38)/(750 \cdot 4) \approx 10.4$.

Altri dettagli utili da conoscere: `https://en.wikipedia.org/wiki/Odds_ratio`

log - odds ratio:

```
log((818 * 38) / ( 750 * 4))
```

Esito vs. Ca125High	FALSE	TRUE
Benigno	818	750
Maligno	4	38

```
indicedirischio3 = Esito ~  Ca125High
logistico3 = glm(indicedirischio3, family = binomial)
summary(logistico3)
```

```
4 / (818 + 4)
```

```
log(0.00486618/(1-0.00486618))
```

```
exp(-5.3206) / (1 + exp(-5.3206))
```

```
exp(-2.9825) / (1 + exp(-2.9825))
38/(750+38)
```

il caso di una variabile numerica

```
indicedirischio2 = Esito ~  Ca125
modello = glm(indicedirischio2, family = binomial)
summary(modello)
```

non possiamo creare una tabella

ipotizziamo una paziente, `Ca125` = 132

```
-4.132 + 0.014 * 132
```

```
exp(-2.284) / (1+exp(-2.284))
```

Esercizio 5.2 Commentate l'andamento di questa sigmoide.

```
plot(Ca125, (as.numeric(Esito)-1))
xx = seq(0,300)
pl = -4.132 + 0.014 * xx
yy = exp(pl)/(1+exp(pl))
lines(xx,yy, col = "orange")
```

■

Esercizio 5.3 Andate a leggere l'articolo di Richard Moore in cui si definisce l'indice R.O.M.A. `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3594101/` ed in particolare cercate la sezione 'Predictive Probability Calculations'. Guardate la formula relativa alla 'Predicted Probability (PP)' e scoprite l'errore di stampa. ■

5.1.4 Problemi con lo standard error

Debora De Bartolo, `tossicologia`

```
www = "http://www.biostatisticaumg.it/dataset/tossicologia.csv"
tossicologia = read.csv(www, header = TRUE)
attach(tossicologia)
table(causa)

...
causa
 incidentedomestico incidenteelicottero
                  1                    1
  incidentestradale            suicidio
                182                    1
```

```
prova = glm(alcolemia ~ genere + eta + causa + tossicologico,
    family = binomial)
summary(prova)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-16.7264	1455.3977	-0.01	0.9908
generem	1.0717	0.5783	1.85	0.0638
eta	0.0016	0.0082	0.19	0.8500
causaincidenteelicottero	0.0187	2058.2429	0.00	1.0000
causaincidentestradale	15.1462	1455.3976	0.01	0.9917
causasuicidio	0.0078	2058.2429	0.00	1.0000
tossicologicopositivo	0.0460	0.4819	0.10	0.9240

collinearità. Lucrezia Ietri , dataset `epilessie`

```
detach(tossicologia)
www = "http://www.biostatisticaumg.it/dataset/epilessie.csv"
epilessie = read.csv( www , header = TRUE )
attach(epilessie)
```

```
prova = glm(prognosi ~ primoEEG + causa + eoneuro + svilpsicom
       + etaallesordio, family = binomial)
summary(prova)
```

	Estimate	Std. Error	z value	Pr(>|z|)
..	..	..	..	..
causasintomatico	-15.5730	1696.2449	-0.01	0.9927
eoneurop	16.1697	1696.2452	0.01	0.9924
..	..	..	..	..

```
table(eoneuro, causa)

...
       causa
eoneuro nonsintomatico sintomatico
      n             35           1
      p              1          14
```

	Estimate	Std. Error	z value	Pr(>|z|)
..	..	..	..	..
causasintomatico	0.0436	0.9656	0.05	0.9640
..	..	..	..	..

5.1.5 La sovradispersione

Eusebio Chiefari [5], dataset gdm:

```
www = "http://www.biostatisticaumg.it/dataset/gdm.csv"
gdm = read.csv(www, header = TRUE)
attach(gdm)
prova = glm( factor(GDM) ~ Ngrav  + BMIpreGrav , family = binomial)
summary(prova)

...
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.11413    0.29777 -13.816  < 2e-16 ***
Ngrav        0.46692    0.06688   6.981 2.93e-12 ***
BMIpreGrav   0.09382    0.01124   8.344  < 2e-16 ***
...
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 2562.1  on 2283  degrees of freedom
Residual deviance: 2430.2  on 2281  degrees of freedom
AIC: 2436.2
...
```

sovradispersione, **parametro di dispersione**,

family = quasibinomial:

```
prova2 = glm( factor(GDM) ~ Ngrav  + BMIpreGrav , family = quasibinomial)
summary(prova2)

...
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.11413    0.29961 -13.732  < 2e-16 ***
Ngrav        0.46692    0.06730   6.938 5.15e-12 ***
BMIpreGrav   0.09382    0.01131   8.292  < 2e-16 ***
...
(Dispersion parameter for quasibinomial family taken to be 1.012403)
    Null deviance: 2562.1  on 2283  degrees of freedom
Residual deviance: 2430.2  on 2281  degrees of freedom
AIC: NA
...
```

5.2 La meta finale: valutare l'accuratezza del modello logistico

Shadi Najaf, dataset `roma`:

```
www = "http://www.biostatisticaumg.it/dataset/roma.csv"
roma = read.csv(www, header = TRUE)
attach(roma)
tail(roma)
```

Moore at al.:

$$P.I. = -12.0 + 2.38 \cdot \log(HE4) + 0.0626 \cdot \log(CA125)$$
$$P.I. = -8.09 + 1.04 \cdot \log(HE4) + 0.732 \cdot \log(CA125)$$

Quindi, Moore et al. [20]:

```
moorerelation = Histology ~ Menopause * logHE4 + Menopause * logCA125
```

Shadi:

```
maximalrelation = Histology ~ logHE4 + logCA125 + logCA19.9 + logCEA
                      + AgePatient + Menopause


maximalmodel = glm(maximalrelation, family = binomial)
step(maximalmodel)


attempt0 = Histology ~ Menopause + logHE4 + logCA125

attempt1 = Histology ~ Menopause * logHE4 + logCA125

attempt2 = Histology ~ logHE4 + Menopause * logCA125

attempt3 = Histology ~ Menopause + logHE4 * logCA125

moorerelation = Histology ~ Menopause * logHE4 + Menopause * logCA125
```

```
modelattempt0 = glm(attempt0, family = binomial)
modelattempt1 = glm(attempt1, family = binomial)
modelattempt2 = glm(attempt2, family = binomial)
modelattempt3 = glm(attempt3, family = binomial)
mooremodel = glm(moorerelation, family = binomial)


AIC(modelattempt0, modelattempt1, modelattempt2, modelattempt3, mooremodel)
```

Per completezza, curvatura:

```
attempt4 = Histology ~ Menopause + logHE4 + logCA125 + I(logCA125^2)

attempt5 = Histology ~ Menopause + logHE4 + I(logHE4^2) + logCA125

modelattempt4 = glm(attempt4, family = binomial)

modelattempt5 = glm(attempt5, family = binomial)

AIC(modelattempt0, modelattempt4, modelattempt5, mooremodel)

summary(modelattempt0)
```

Fisher Scoring

$$P.I. = -14.38 + 2.34 \cdot \log(HE4) + 0.68 \cdot \log(CA125)$$
$$P.I. = -13.44 + 2.34 \cdot \log(HE4) + 0.68 \cdot \log(CA125)$$

5.2.1 La curva ROC

Vi interessa la storia della receiver operating characteristic curve? Eccola: `https://en.wikipedia.org/wiki/Receiver_operating_characteristic#History`

```
Menopause01 = -1 + as.numeric(Menopause)
iSN =  -14.38 + 0.94 * Menopause01 + 2.34 * logHE4 + 0.68 * logCA125
```

```
install.packages("pROC")
library(pROC)
```

```
iSNroc = roc(Histology ~ iSN, auc=TRUE)
iSNroc
```

```
   plot(iSNroc):
```

5.3 Esercizi ed attività di approfondimento

■ **Attività 5.1 — la curva ROC.** Riprendete in esame il dataset `raul` e definite l'indice predittivo come descritto in [9]:

```
umgrisk = Ldhtre + 24 / Ldhuno
```

A questo punto, valutate l'area sotto la curva ROC e disegnatela, e capirete l'emozione di Annalisa Di Cello quando si è accorta di questo risultato:

```
annalisa = roc(Esito ~ umgrisk, auc=TRUE)
annalisa
plot(annalisa)
```

■

Bibliografia

Articoli

[2] Edgar Anderson. “The species problem in Iris”. In: *Annals of the Missouri Botanical Garden* 23.3 (1936), pagine 457–509 (citato a pagina 11).

[4] T Calimeri et al. “A unique three-dimensional SCID-polymeric scaffold (SCID-synth-hu) model for in vivo expansion of human primary multiple myeloma cells”. In: *Leukemia* 25.4 (2011), pagine 707–712 (citato a pagina 78).

[5] Carmelo Capula et al. “A new predective tool for the early risk assessment of gestational diabetes mellitus”. In: *Primary Care Diabetes* 10.5 (2016). PMID: 27268754, pagine 315–323 (citato a pagina 87).

[6] Maria Vittoria Caruso, Attilio Renzulli e Gionata Fragomeni. “Influence of IABP-Induced abdominal occlusions on aortic hemodynamics: a patient-specific computational evaluation”. In: *ASAIO Journal* 63.2 (2017), pagine 161–167 (citato a pagina 9).

[7] Emanuela Chiarella et al. “ZNF521 Represses Osteoblastic Differentiation in Human Adipose-Derived Stem Cells”. In: *International journal of molecular sciences* 19.12 (2018), pagina 4095 (citato a pagina 9).

[8] Maria Teresa De Angelis et al. “Short-term retinoic acid treatment sustains pluripotency and suppresses differentiation of human induced pluripotent stem cells”. In: *Cell death & disease* 9.1 (2018), pagina 6 (citato a pagina 9).

[9] Annalisa Di Cello et al. “A more accurate method to interpret lactate dehydrogenase (LDH) isoenzymes’ results in patients with uterine masses”. In: *European Journal of Obstetrics and Gynecology and Reproductive Biology* in press (2019). DOI: 10.1016/j.ejogrb.2019.03.017. URL: 10.1016/j.ejogrb.2019.03.017 (citato alle pagine 9, 28, 81, 91).

[10] Ronald A Fisher. “The use of multiple measurements in taxonomic problems”. In: *Annals of eugenics* 7.2 (1936), pagine 179–188 (citato a pagina 11).

[13] Francis Galton. “Regression towards mediocrity in hereditary stature.” In: *The Journal of the Anthropological Institute of Great Britain and Ireland* 15 (1886), pagine 246–263 (citato a pagina 53).

[14] Richard Horton. “Offline: what is medicine’s 5 sigma?” In: *The Lancet* 385.9976 (2015), pagina 1380 (citato a pagina 51).

[16] John PA Ioannidis. “Why most published research findings are false”. In: *PLoS medicine* 2.8 (2005), e124 (citato a pagina 51).

[19] Eckhard Limpert, Werner A. Stahel e Markus Abbt. “Log-normal Distributions across the Sciences: Keys and Clues”. In: *BioScience* 51.5 (2001), pagine 341–352 (citato a pagina 36).

[20] Richard G Moore et al. “Comparison of a novel multiple marker assay vs the Risk of Malignancy Index for the prediction of epithelial ovarian cancer in patients with a pelvic mass”. In: *American journal of obstetrics and gynecology* 203.3 (2010), 228–e1 (citato a pagina 88).

[21] Kersti Pärna et al. “Alcohol consumption in Estonia and Finland: Finbalt survey 1994-2006”. In: *BMC Public Health* 10.1 (2010), pagina 261 (citato a pagina 34).

[25] Roberta Venturella et al. “Three to five years later: long–term effects on ovarian function of prophylactic bilateral salpingectomy”. In: *Journal of Minimally Invasive Gynecology* 24.1 (2017). PMID: 27621194, pagine 145–150 (citato alle pagine 9, 80).

[26] Howard Wainer. “The most dangerous equation”. In: *American Scientist* 95.3 (2007), pagina 249 (citato a pagina 20).

[29] GN Wilkinson e CE Rogers. “Symbolic description of factorial models for analysis of variance”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 22.3 (1973), pagine 392–399 (citato a pagina 61).

[30] P Zaffino et al. “Radiotherapy of Hodgkin and non-Hodgkin lymphoma: A nonrigid image-based registration method for automatic localization of prechemotherapy gross tumor volume”. In: *Technology in cancer research & treatment* 15.2 (2016), pagine 355–364 (citato a pagina 9).

[31] Stephen T Ziliak e Deirdre N McCloskey. “The cult of statistical significance”. In: *Ann Arbor: University of Michigan Press* 27 (2008) (citato a pagina 51).

Libri

[3] Martin Bland. *An introduction to medical statistics.* Ed. 3. Oxford University Press, 2000 (citato a pagina 34).

[11] Ronald Aylmer Fisher. *The design of experiments*. Oliver e Boyd; Edinburgh; London, 1937 (citato a pagina 51).

[15] Sergio Invernizzi. *Matematica nelle Scienze Naturali*. Trieste: Edizioni Goliardiche, 1996. ISBN: 8886573170 (citato a pagina 10).

[17] Michael C Joiner e Albert Van der Kogel. *Basic clinical radiobiology*. CRC press, 2016 (citato a pagina 31).

[23] Bernard A Rosner. *Fundamentals of biostatics*. Duxbury Press, 1995 (citato alle pagine 22, 39, 40).

[24] William N. Venables e Brian D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: `http://www.stats.ox.ac.uk/pub/MASS4` (citato a pagina 24).

[27] Hadley Wickham e Garrett Grolemund. *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc., 2016 (citato a pagina 18).

Risorse Web

[1] Valentin Amrhein, Sander Greenland e Blake McShane. *Scientists rise up against statistical significance*. `https://www.nature.com/articles/d41586-019-00857-9`. Accessed: 2019-03-20. 2019 (citato alle pagine 51, 55).

[12] Giovanni Gallo. *Biometria Fetale*. `http://web.tiscali.it/giovannigallo/tabelle/tabelle.htm`. Accessed: 2019-04-08 (citato a pagina 40).

[18] Tatsuki Koyama. *Beware of Dynamite*. `http://biostat.mc.vanderbilt.edu/wiki/pub/Main/TatsukiRcode/Poster3.pdf`. Accessed: 2019-03-20 (citato a pagina 21).

[22] Mark D Robinson, Davis J McCarthy e Gordon K Smyth. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*. `https://www.bioconductor.org/packages/devel/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf`. Accessed: 2019-04-06 (citato a pagina 29).

[28] Claus O. Wilke. *Fundamentals of Data Visualization*. `https://serialmentor.com/dataviz/`. Accessed: 2019-03-27 (citato a pagina 18).

Indice analitico