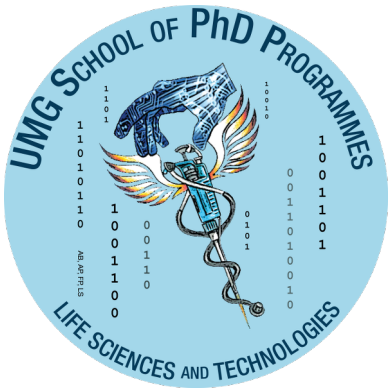
SOCIETA' DEI MATEMATICI
E NATURALISTI DI MODENA
www.socnatmatmo.unimore.it



Massimo Borelli, Ph.D.

il dataset `tossicologia` analizzato con R

9 gennaio 2017

Riproduzione pittorica di foto d'epoca a cura di Caterina Stella.

Giovanni Canestrini, nato a Revò (Tn) nel 1835 e defunto a Padova nel 1900, fu fondatore nel 1865 della Società dei Naturalisti e Matematici di Modena. Nata con lo scopo di promuovere lo studio delle Scienze Naturali e Matematiche, di favorire i legami fra studiosi, ricercatori, collezionisti e diffondere cultura scientifica a tutti i livelli per una equilibrata crescita civile, la Società dei Naturalisti e Matematici di Modena ha accolto nelle sua fila, tra gli altri, Charles Darwin, Louis Pasteur e Thomas Henry Huxley.

caro Massimo,
vorrei effettuare una analisi di tipo epidemiologico relativa al mio campione di persone (per età, sesso, positività all'esame tossicologico) e sapere inoltre se il campione possa essere analizzato in relazione al livello di alcolemia.

1 Struttura del dataset `tossicologia`

Si tratta di un dataset di 185 osservazioni di 8 variabili, in formato *.csv* (il carattere separatore è infatti una virgola): `ID` è un contatore di identificazione delle persone decedute, di cui conosciamo l'`anno` del decesso, il loro `genere` (ossia `f` e `m`, una variabile fattore a due livelli), la loro `eta` al momento del decesso ed una generica descrizione della `causa` di morte. Inoltre si conoscono gli esiti degli esami `tossicologico` ed `alcolemia` (entrambe variabili fattore a due livelli, `negativo` e `positivo`; di quest'ultima, il livello positivo si ha quando la variabile numerica `alcol` è maggiore di zero). Importiamo il dataset in R[2] con i seguenti comandi:

```
www = "http://www.biostatisticaumg.it/dataset/tossicologia.csv"
tossicologia = read.csv(www, header = TRUE)
# tossicologia = read.csv(file.choose(), header = TRUE)
attach(tossicologia)
head(tossicologia)
str(tossicologia)
```

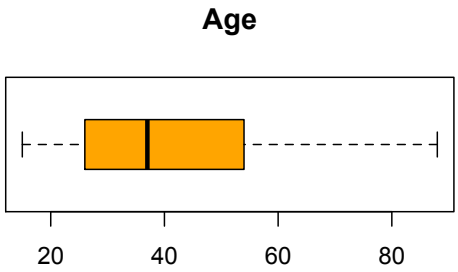
Il comando `head` ci consente di visualizzare le prime righe del dataset:

	ID	anno	genere	eta	causa	tossicologico	alcol	alcolemia
1	1	2006	m	75	incidentestradale	negativo	17	positivo
2	2	2006	m	74	incidentestradale	negativo	3	positivo
3	3	2006	m	67	incidentestradale	negativo	0	negativo
4	4	2006	m	52	incidentestradale	negativo	4	positivo
5	5	2006	m	23	incidentestradale	negativo	0	negativo
6	6	2006	m	36	incidentestradale	negativo	8	positivo

2 Visualizziamo il dataset `tossicologia`

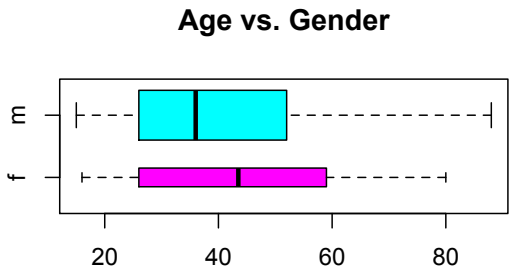
Abbiamo molte possibilità per visualizzare i dati del nostro dataset, dalle più spartane alle più ricercate. Per descrivere l'età dei soggetti possiamo utilizzare un boxplot:

```
boxplot(eta, horizontal = TRUE, col = "orange", main = "Age")
```



Se vogliamo stratificare l'età rispetto al genere, possiamo fare così:

```
boxplot(eta ~ genere, horizontal = TRUE, col = c("magenta", "cyan"),
   varwidth = TRUE, main = "Age vs. Gender")
```



Il box di colore magenta è più sottile del box color ciano, in quanto la dimensione dei box è proporzionale alla (radice quadrata della) numerosità dei maschi e delle femmine:

```
table(genere)
```

	genere
f	22
m	163

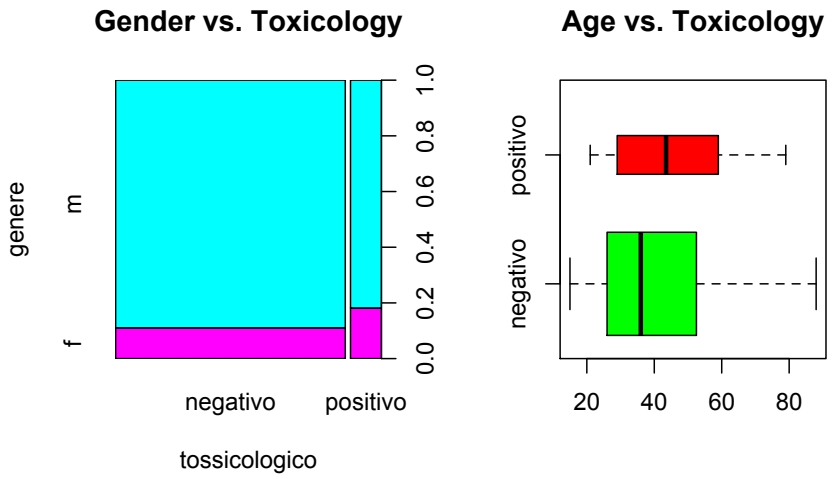
Per quanto riguarda la `causa` del decesso, una tabella riassume molto meglio i dati di quanto ogni grafico possa farlo:

```
table(causa)
```

	causa
incidentedomestico	1
incidenteelicottero	1
incidentestradale	182
suicidio	1

Vediamo ora l'esame `tossicologico`, stratificato sia per genere che per età:

```
par(mfrow = c(1,2))
plot(genere  ~ tossicologico, col = c("magenta", "cyan"),
    main = "Gender vs. Toxicology")
boxplot(eta ~ tossicologico, horizontal = TRUE, col = c("green", "red"),
    varwidth = TRUE, main = "Age vs. Toxicology")
```



Nel pannello di sinistra abbiamo raffigurato un 'grafico a mosaico' (*mosaic plot*), che è un modo efficace di visualizzare i dati delle tabelle a doppia entrata:

	negativo	positivo
f	18	4
m	145	18

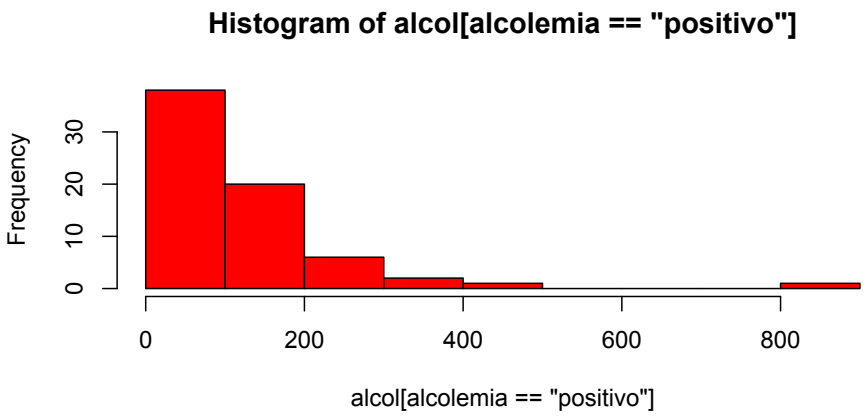
Con la versatile funzione `summary` del linguaggio R [2] otteniamo immediatamente delle statistiche descrittive:

```
summary(eta)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
15.00	26.00	37.00	42.01	53.75	88.00	7

Per mettere infine in luce i livelli di `alcol` per le persone del **livello** `positivo` del **fattore** `alcolemia`, possiamo usare il comando:

```
hist(alcol[alcolemia == "positivo"], col = "red")
```



3 Modelliamo la risposta `alcolemia`

Vogliamo capire se la risposta `alcolemia` possa essere associata a qualcuna delle covariate del dataset. Siccome `alcolemia` è una variabile dicotomica, utilizzeremo un **modello** di regressione **lineare generalizzato**[1] con una famiglia di variabili aleatorie binomiali.

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

	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

Non ci stupisce affatto che la covariata `causa`, molto sbilanciata, non rappresenti un valido predittore del modello. Eliminiamola e ripetiamo l'analisi:

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

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.5528	0.6635	-2.34	0.0193
generem	1.0409	0.5780	1.80	0.0717
eta	0.0008	0.0082	0.09	0.9246
tossicologicopositivo	0.0797	0.4813	0.17	0.8685

Un p-value molto elevato, $p = 0.92$, potrebbe indurci a ritenere che l'età non sia associata, in senso statistico, alla positività alcolemica. Semplifichiamo, pertanto, ulteriormente il modello:

```
modello = glm(alcolemia ~ genere + tossicologico,
    family = binomial)
summary(modello)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.5128	0.5600	-2.70	0.0069
generem	1.0713	0.5767	1.86	0.0632
tossicologicopositivo	0.0475	0.4798	0.10	0.9212

Ecco dunque che potremmo avere delle evidenze che il genere possa avere un ruolo, in senso statistico:

```
modello = glm(alcolemia ~ genere, family = binomial)
summary(modello)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.5041	0.5528	-2.72	0.0065
generem	1.0678	0.5756	1.86	0.0636

Tenendo tuttavia conto del messaggio fornito dal software (*Residual deviance: 239.25 on 183 degrees of freedom*), dobbiamo ipotizzare che vi sia una sovradispersione dei dati; pertanto, cerchiamo di correggere questo fatto modificando il parametro di dispersione:

```
finale = glm(alcolemia ~ genere, family = quasibinomial)
summary(finale)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.5041	0.5558	-2.71	0.0074
generem	1.0678	0.5787	1.85	0.0666

4 Conclusioni

Vediamo come possiamo riassumere la nostra analisi, ipotizzando di dover scrivere in un articolo i 'Materiali e Metodi'.

4.1 Matherials and Methods

Data have been analyzed by means of R software [2]. Continuous covariates have been summarized by means of medians, min-max range and interquartile range, while categoric variables have been tabulated with absolute frequencies. Inferential analysis on alcoholemic response has been led by means of binomial distributed generalized linear models (i.e. the multivariate logistic regression); selection within models has been issued by means of a top-down procedure, correcting for overdispersion[1]. In all inferential instances an $\alpha = .05$ level has been assumed.

Riferimenti bibliografici

[1] Julian J Faraway. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models.* CRC press, 2005.

[2] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2014.