Laboratorio di ST1
Lezione 7

Claudia Abundo

Dipartimento di Matematica
Università degli Studi Roma Tre

30 Aprile 2010

Un esercizio riepilogativo

Willerman et al. (1991) collected a sample of 40 psychology students at a large southwestern university. The researchers used Magnetic Resonance Imaging (MRI) to determine the brain size of the subjects. Information about gender and body size (height and weight) are also included. The researchers withheld the weights of two subjects and the height of one subject for reasons of confidentiality.
Dall'esame di 7 variabili su 40 persone vogliamo capire se c'è relazione tra l'intelligenza e la grandezza del cervello.

Preliminarmente elaborariamo misure e grafici che possono esserci utili.

Il Dataset

- Gender: Male or Female
- FSIQ: Full Scale IQ scores based on the four Wechsler (1981) subtests
- VIQ: Verbal IQ scores based on the four Wechsler (1981) subtests
- PIQ: Performance IQ scores based on the four Wechsler (1981) subtests
- Weight: body weight in pounds
- Height: height in inches
- MRI_Count: total pixel Count from the 18 MRI scans

Creiamo una cartella sul desktop e cambiamo directory...

```
brain=read.table("brain.txt", header = TRUE, sep = "", dec = ".")
```

Frequenze della variabile “VIQ”: assolute, relative, cumulate

```
>table(brain$VIQ)
 71  77  83  86  90  91  93  96 100 107 112 114 120 123 126
 129 132 136 145 150
  1   1   2   2   5   1   1   4   1   1   1   1   1   1   2
   5   3   1   3   3

> VIQ_int=cut(brain$VIQ, breaks=c(70, 90, 110, 130, 150))
> table(VIQ_int)
VIQ_int
  (70,90]  (90,110] (110,130] (130,150]
       11         8        11        10
> rel=table(VIQ_int)/length(brain$VIQ)
> rel
VIQ_int
  (70,90]  (90,110] (110,130] (130,150]
    0.275     0.200     0.275     0.250
> cumsum(rel)
  (70,90]  (90,110] (110,130] (130,150]
    0.275     0.475     0.750     1.000
```

Grafici_1: Istogramma

Costruiamo l'istogramma della variabile Weight (virtualmente continua) prima con le frequenze assolute e poi con quelle relative (basta aggiungere prob=T !):

```
hist(brain$Weight, col = 3, main="Istogramma di Weight")
hist(brain$Weight, col = 3, main="Istogramma di Weight", prob=TRUE)
```

Grafici_2: Box Plot

```
boxplot(brain$VIQ, col=3, main="Boxplot di VIQ")
```

Richiamo:
La linea centrale evidenzia la mediana
Sopra e sotto la mediana ci sono il primo e il terzo quartile
la lunghezza della scatola è detta intervallo o scarto interquartile (SIQ)
Le linee al termine dei “baffi” sono poste a Q1 - 1.5SIQ (sotto) e Q3 + 1.5SIQ (sopra)
Le osservazione esterne alle linee che terminano il baffo sono indicate da pallini (potenziali outliers)

```
summary(brain$VIQ)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   71.0    90.0   113.0   112.3   129.8   150.0
```

media

Osserviamo che nel dataset c'è qualche dato "Non Disponibile" (Not Available: NA)

Se proviamo a calcolare la media della variabile "Height"...

```
mean(brain$Height)
[1] NA
```

Possiamo *escludere* i dati NA attraverso il seguente comando: ridenominiamo Height senza NA come H_2, e poi ne potremo calcolare la media

```
H_2=na.exclude(brain$Height)
mean(H_2)
```

Mediana

La Mediana si trova semplicemente con il comando “median”:

```
> median(brain$PIQ)
[1] 115
```

la mediana é la modalitá che divide la popolazione in due parti di uguale numerosità...

Quartili_1

I Quartili dividono la popolazione in quattro parti di uguale numerosità.
il primo ha il 25% della popolazione a sinistra ed il 75% a destra
il secondo coincide con la mediana e ha il 50% della popolazione sia a sinistra che a destra
il terzo ha il 75% della popolazione a sinistra ed il 25% a destra
il quarto ha il 100% della popolazione a sinistra
Possiamo averli tutti con un solo comando:

```
quantile(brain$MRI_Count)
      0%       25%       50%       75%      100%
790619.0  855918.5  905399.0  950078.0 1079549.0
```

Possiamo calcolare gli scarti interquartili, cioè le differenze tra un quartile e l'altro

```
 quantile(brain$MRI_Count, 1)-quantile(brain$MRI_Count, 0.75)
  100%
129471
 quantile(brain$MRI_Count, 0.75)-quantile(brain$MRI_Count, 0.5)
  75%
44679
 quantile(brain$MRI_Count, 0.5)-quantile(brain$MRI_Count, 0.25)
    50%
49480.5
```

Summary

Il comando Summary ci fornisce molte misure, tutte insieme...

```
summary(brain)
```

```
> summary(brain)
    Gender          FSIQ             VIQ             PIQ
 Female:20   Min.   : 77.00   Min.   : 71.0   Min.   : 72.00
 Male  :20   1st Qu.: 89.75   1st Qu.: 90.0   1st Qu.: 88.25
             Median :116.50   Median :113.0   Median :115.00
             Mean   :113.45   Mean   :112.3   Mean   :111.03
             3rd Qu.:135.50   3rd Qu.:129.8   3rd Qu.:128.00
             Max.   :144.00   Max.   :150.0   Max.   :150.00

     Weight          Height        MRI_Count
 Min.   :106.0   Min.   :62.00   Min.   : 790619
 1st Qu.:135.3   1st Qu.:66.00   1st Qu.: 855919
 Median :146.5   Median :68.00   Median : 905399
 Mean   :151.1   Mean   :68.53   Mean   : 908755
 3rd Qu.:172.0   3rd Qu.:70.50   3rd Qu.: 950078
 Max.   :192.0   Max.   :77.00   Max.   :1079549
 NA's   :  2.0   NA's   : 1.00
```

Varianza

La Varianza è una quantificazione della dispersione dei dati relativi ad una variabile. Calcoliamola per Weight ed Height e vediamola graficamente.

```
W_2=na.exclude(brain$Weight)
var(W_2)
[1] 551.2404
hist(brain$Weight)
H_2=na.exclude(brain$Height)
var(H_2)
[1] 15.95722
hist(brain$Height)
```

Controlliamo il coefficiente di variazione...

```
> CV=sd(W_2)/abs(mean(W_2))
> CV
[1] 0.1554326

> CV2=sd(H_2)/abs(mean(H_2))
> CV2
[1] 0.05829422
```

H_2 ha effettivamente varianza più bassa.

Scatterplots

```
plot(brain)
```

Covarianza e Correlazione

Vogliamo capire se c'è una relazione tra la grandezza del cervello (MRI_Count) e l'intelligenza (misurata attraverso i punteggi nei test: FSIQ, VIQ, PIQ), calcoliamo quindi la covarianza e la correlazione!

Ma prima eliminiamo i dati mancanti...

```
  b_2=na.omit(brain)

 cov(b_2$FSIQ, b_2$MRI_Count)
[1] 576686.1
 cor(b_2$FSIQ, b_2$MRI_Count)
[1] 0.3337137

 cov(b_2$VIQ, b_2$MRI_Count)
[1] 499825.9
 cor(b_2$VIQ, b_2$MRI_Count)
[1] 0.3002791

 cov(b_2$PIQ, b_2$MRI_Count)
[1] 619463.9
 cor(b_2$PIQ, b_2$MRI_Count)
[1] 0.3777816
```

Covarianza e Correlazione_2

Però è plausibile che ci sia una relazione tra la grandezza del corpo e quella del cervello, infatti...

```
> cor(b_2$Weight, b_2$MRI_Count)
[1] 0.5133785
> cor(b_2$Height, b_2$MRI_Count)
[1] 0.5883772
```

Vediamo cosa accade “eliminando l’influenza” della grandezza del corpo:

consideriamo soltanto i dati con peso corporeo (Weight) minore della mediana: 146.5

Covarianza e Correlazione_3

```
 SUB=subset(b_2, b_2$Weight<146.5)

 SUB

 cor(SUB$FSIQ, SUB$MRI_Count)
[1] 0.4160623
 cor(SUB$VIQ, SUB$MRI_Count)
[1] 0.3312977
 cor(SUB$PIQ, SUB$MRI_Count)
[1] 0.5104168
```

In alternativa...

```
SUBMAN=subset(SUB, SUB$Gender=="Male")
```

Covarianza e Correlazione_4

Ancora: vediamo come cambia la correlazione se dell'ultimo dataset prendiamo soltanto i maschi...

```
 SUBMAN=SUB[c(2, 5, 12, 18),]
 SUBMAN

 cor(SUBMAN$FSIQ, SUBMAN$MRI_Count)
[1] 0.67509
 cor(SUBMAN$VIQ, SUBMAN$MRI_Count)
[1] 0.281383
 cor(SUBMAN$PIQ, SUBMAN$MRI_Count)
[1] 0.9235757
```

Covarianza e Correlazione_5

O soltanto le femmine...

```
 SUBWOM=SUB[c(-2,-5, -12, -18),]
 SUBWOM

 cor(SUBWOM$FSIQ, SUBWOM$MRI_Count)
[1] 0.1828296
 cor(SUBWOM$VIQ, SUBWOM$MRI_Count)
[1] 0.08573658
 cor(SUBWOM$PIQ, SUBWOM$MRI_Count)
[1] 0.2995408
```

Le correlazioni tra uomini e donne sono notevolmente diverse!

Covarianza e Correlazione_6

Se invece non distinguiamo per peso corporeo, ma prendiamo soltanto i maschi...

```
 SM2=b_2[c(2, 3, 8, 9, 11, 12, 17, 19, 20, 22, 24, 26, 30, 31, 32,
35, 37, 38),]
 SM2

 cor(SM2$FSIQ, SM2$MRI_Count)
[1] 0.4313880

 cor(SM2$VIQ, SM2$MRI_Count)
[1] 0.3265548

 cor(SM2$PIQ, SM2$MRI_Count)
 [1] 0.5284226
```

Paragonando con i risultati relativi alla eliminazione del peso corporeo il primo ed il terzo indice sono più bassi: c'è una relazione tra l'MRI_Count ed il peso, che abbiamo già visto, che influenza gli indici di correlazione!

Intervalli parte 2

```
x <- c(0.39, 1.68, 1.82, 2.35, 0.38, 1.62, 1.70, 1.71,
1.85, 2.14, 2.89, 3.69)
```

In generale, all'aumentare di n l'intervallo di confidenza si restringe.
Si ha:

$$n \geq \left(2z_{1-\alpha/2}\sigma/L\right)^2$$

con L ampiezza dell'intervallo
Esempio
In un esame di psicologia vengono misurati i tempi di reazione di 100 individui, riscontrando un tempo medio di 1 secondo. Da studi pregressi, lo scarto quadratico σ è noto essere pari a 0.05 secondi. Quale deve essere il numero minimo di osservazioni campionarie n per avere un'ampiezza dell'intervallo pari al più a 0.02 secondi ed un intervallo di confidenza pari al 99%.

```
n <- 100
a <- qt(0.995, df = n - 1)
n1 <- (2 * a * (0.05/0.02))^2
cat("(",ceiling(n1),")\n")
```

Intervallo per la varianza

Gli intervalli di confidenza per la varianza forniscono un campo di variazione all'interno del quale ci si aspetta di trovare il parametro incognito σ^2.

$$\sigma^2 > \frac{n-1 \cdot s^2}{\chi^2_{n-1}}$$

```
ic.var <-
function(x, conf.level){
alfa <- 1 - conf.level
n <- length(x)
l.inf <- 0
l.sup <- (n - 1) * var(x)/qchisq(alfa, df = n - 1)
c(l.inf, l.sup)
}
```