Modelli statistici: sessione 2

Francesco Lagona
Università Roma Tre

1 Importare dati in R Studio

Un file di dati può essere importato all'interno di R Studio usando il tasto "Import Dataset" disponibile nella finestra "Environment". Diversi formati sono disponibili. Dopo aver creato un progetto "Session2", importiamo il file excel "mmse" che vi è stato dato a lezione.

In laternativa, il file excel può essere trasfromato in un file di testo, collocato nella directory di lavoro del progetto "session2", e letto con il comando

```
> mmse <- read.table("mmse.txt")
```

Il file "mmse" contiene risultati di un'indagine realizzata su un campione di anziani cinesi. Le variabili hanno il seguente significato:

- age: età (in mesi) al momento dell'indagine
- gender: 1 = maschio
- urbrur: 1 = residente in area non urbana (0 = residente in area urbana)
- act: 1 = stile di vita sedentario
- adl2: 1 = un limite adl (0 = 0 limiti adl)
- adl3: 1 = 2 o più limiti adl (0 = 0 limiti adl)
- n1: numero di risposte corrette nel test MMSE

Visualizzando le statistiche elementari delle variabili, osserviamo che esse hanno un'intepretazione differente a seconda della natura della variabile.

```
> summary(mmse)
```

```
      age           gender           urbrur            act        
 Min.   : 960   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:1033   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1114   Median :1.0000   Median :1.0000   Median :0.0000  
 Mean   :1113   Mean   :0.6004   Mean   :0.6458   Mean   :0.4216  
```

```
 3rd Qu.:1201   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1271   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
      adl2             adl3             n1
 Min.   :0.0000   Min.   :0.000   Min.   : 0.00
 1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:12.00
 Median :0.0000   Median :0.000   Median :19.00
 Mean   :0.1318   Mean   :0.238   Mean   :15.84
 3rd Qu.:0.0000   3rd Qu.:0.000   3rd Qu.:22.00
 Max.   :1.0000   Max.   :1.000   Max.   :23.00
```

2 Tabelle e grafici

Il comando `table` è utile per calcolare distribuzioni di frequenze di variabili non continue

```
> tab.gender <- table(mmse$gender)
> tab.gender

   0    1
3160 4748

> tab.gender.rel <- table(mmse$gender)/length(mmse$gender)
> tab.gender.rel

        0         1
0.3995953 0.6004047
```

e per disegnarne i relativi diagrammi a barra

```
> barplot(tab.gender, names.arg=c("female","male"))
> title(main="gender distribution")
```

Per un carattere continuo, è invece più appropriato un istogramma

```
> hist(mmse$age,freq=F,main="age distribution",xlab="age")
```

oppure un diagramma a scatola (boxplot), dove il segmento centrale indica la mediana, gli estremi della scatola indicano i quartili, e le "ali" si estendono per l'intero campo di variazione dei dati.

```
> boxplot(mmse$age, range=0)
> title(main="age distribution")
```

3 Distribuzioni condizionate e congiunte

E' spesso assai utile ispezionare i dati costruendo grafici alla ricerca di eventuali associazioni e dipendenze. Se abbiamo a che fare con due variabili qualitative, è possibile costruire una tabella e calcolare un chi quadro

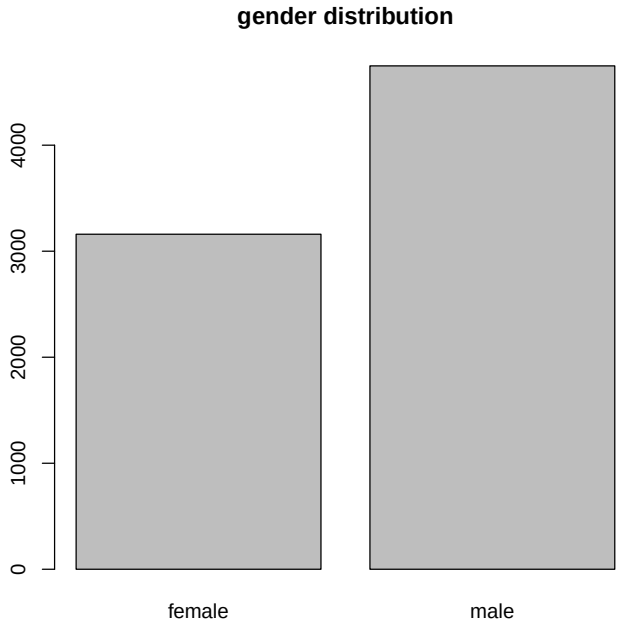

Figura 1: Distribuzione del genere nel dataset MMSE

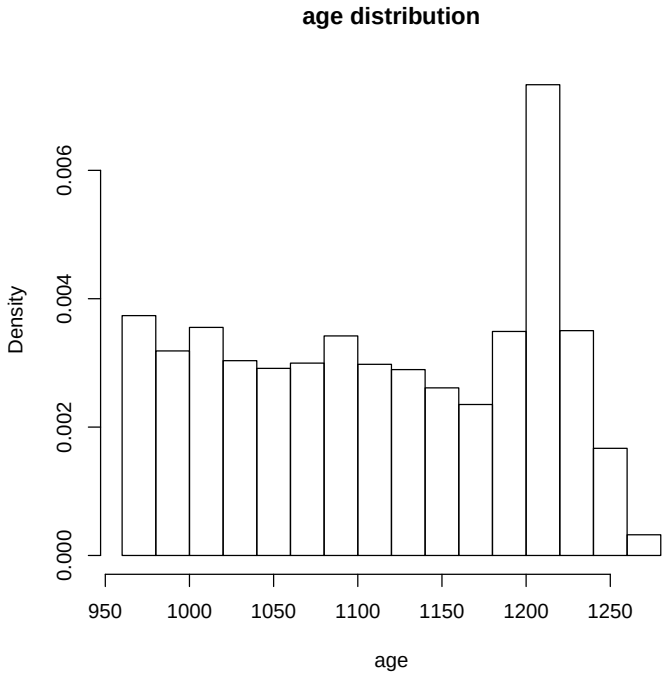

Figura 2: Istogramma dell'età nel dataset MMSE

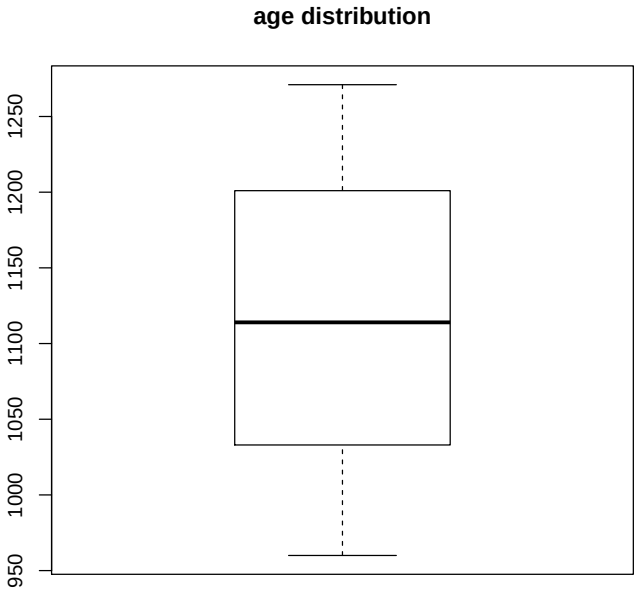

Figura 3: Diagramma a scatola dell'età nel dataset MMSE

```
> # diamo un nome alle modalità di genere e stile di vita
> gender <- ifelse(mmse$gender==0,"female","male")
> act <- ifelse(mmse$act==1,"sedentary","active")
> tab <- table(gender,act)
> tab

        act
gender   active sedentary
  female   1510      1650
  male     3064      1684

> summary(tab)

Number of cases in table: 7908
Number of factors: 2
Test for independence of all factors:
        Chisq = 218.23, df = 1, p-value = 2.203e-49
```

Una tabella a doppia entrata può essere visualizzata da un diagramma a mosaico che descrive la distribuzione condizionata di una variabile data l'altra (figura 4)

```
> mosaicplot(tab, main="gender and lifestyle")
```

Quando invece vogliamo confrontare una variabile qualitativa e una quantitativa, è appropriata una tabella ANOVA

```
> urbrur <- ifelse(mmse$urbrur==0, "urban", "rural")
> n1 <- mmse$n1
> anova(lm(n1 ~ urbrur))

Analysis of Variance Table

Response: n1
            Df Sum Sq Mean Sq F value    Pr(>F)
urbrur       1   5737  5736.7  114.39 < 2.2e-16 ***
Residuals 7906 396498    50.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Si osservi che, sebbene la varianza between sia molto minore della varianza within, il test F è significativo, a causa dell'elevata dimensione del campione. Graficamente, la dipendenza tra le due variabili può essere resa da un boxplot condizionato (figura 5)

```
> boxplot(n1 ~ urbrur, range=0)
> title(main="correct answers by type of residence")
```

Figura 4: Un esempio di diagramma a mosaico

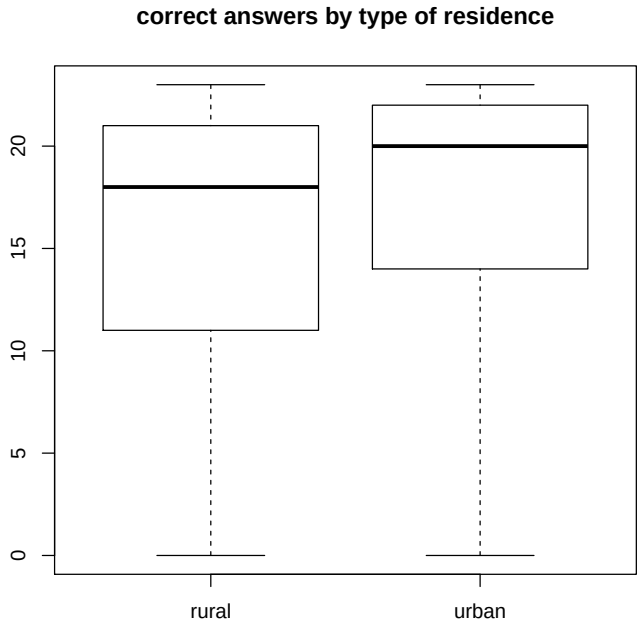

Figura 5: Un esempio di boxplot condizionato

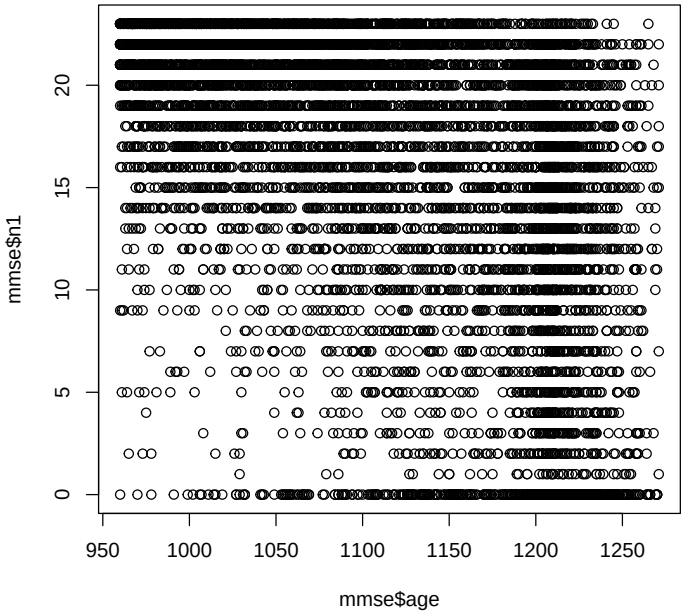

Figura 6: Età e numero di risposte corrette nel dataset mmse.

4 Non sempre è tutto ovvio

Supponiamo di essere interessati all'associazione tra età e numero di risposte corrette nel test MMSE. Il diagramma a dispersione delle due variabili (figura 6) non è molto informativo. Il motivo risiede nei diversi campi di variazione delle due variabili.

```
> plot(mmse$age,mmse$n1)
```

Alternativamente, possiamo suddividere in classi la variabile "age" e costruire il boxplot condizionato della figura 7.

```
> age.cut <- cut(mmse$age,breaks = seq(min(mmse$age),max(mmse$age),length=5))
> boxplot(mmse$n1 ~ age.cut,range=0)
```

La relazione tra età e numero di risposte corrette è confermata dall'analisi della regressione e dalla tabella ANOVA.

```
> reg <- lm(mmse$n1 ~ mmse$age)
> summary(reg)
```

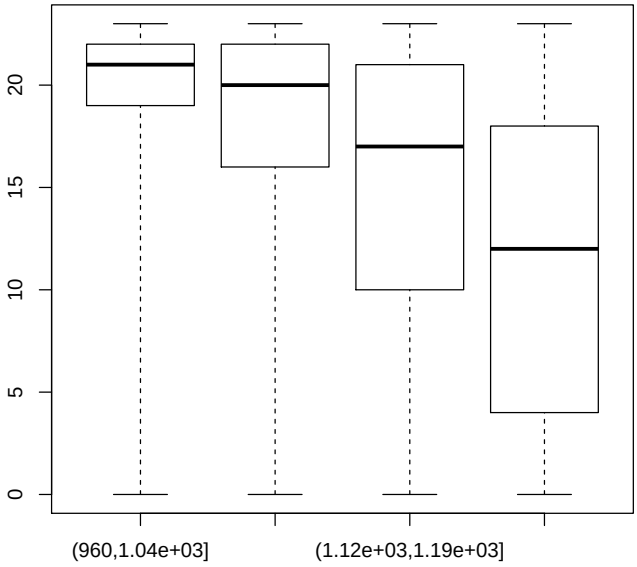

Figura 7: Età e numero di risposte corrette nel dataset mmse.

```
Call:
lm(formula = mmse$n1 ~ mmse$age)

Residuals:
    Min      1Q  Median      3Q     Max
-21.721  -3.278   1.352   4.311  12.986

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.5697824  0.8924620   65.63   <2e-16 ***
mmse$age    -0.0383842  0.0007991  -48.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.276 on 7906 degrees of freedom
Multiple R-squared:  0.2259,        Adjusted R-squared:  0.2258
F-statistic:  2307 on 1 and 7906 DF,  p-value: < 2.2e-16

> anova(reg)

Analysis of Variance Table

Response: mmse$n1
            Df Sum Sq Mean Sq F value    Pr(>F)
mmse$age     1  90865   90865  2307.2 < 2.2e-16 ***
Residuals 7906 311369      39
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```