Analisi statistica sul dataset "iris"

Carla Perrone

Giugno 2020

Indice

1 Introduzione

In questo elaborato analizzerò i dati contenuti nel dataset "Iris" della library datasets reperibile al seguente link: https://archive.ics.uci.edu/ml/datasets/Iris

1.1 Prima analisi dei dati in possesso

Il dataset è costituito da 150 unità statistiche e i caratteri presi in esame sono 5:

- **Lunghezza del sepalo**(Sepal length)
- **Larghezza del sepalo**(Sepal width): variabile quantitativa continua, la scala di misurazione è una scala di rapporti
- **Lunghezza del petalo**(Petal length): variabile quantitativa continua, la scala di misurazione è una scala di rapporti
- **Larghezza del petalo**(Petal width): variabile quantitativa continua, la scala di misurazione è una scala di rapporti
- **Specie**(Species): variabile qualitativa, la scala di misurazione è una scala nominale

Dopo aver digitato "library(xtable)", con il comando "xtable(iris[1:5,])" si ottiene:

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.10	3.50	1.40	0.20	setosa
2	4.90	3.00	1.40	0.20	setosa
3	4.70	3.20	1.30	0.20	setosa
4	4.60	3.10	1.50	0.20	setosa
5	5.00	3.60	1.40	0.20	setosa

Di seguito riporto la struttura e il summary dei dati:

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

Nel secondo capitolo dell'elaborato vengono analizzati i dati singolarmente (statistica univariata), nel terzo capitolo viene fatta un'analisi di più caratteri simultaneamente (statistica bivariata), in particolare verranno analizzate le seguenti coppie di caratteri:
Infine nell'ultimo capitolo viene fatta un'analisi multivariata.

2 Analisi dei singoli caratteri

Iniziamo con un'analisi preliminare dei singoli caratteri presi singolarmente; per ciascuno di essi, inizialmente, riporteremo i grafici delle frequenze e faremo opportune rappresentazioni grafiche, per poi ipotizzare un modello matematico di ditribuzione.

2.1 Sepal length

La lunghezza del sepalo è un carattere quantitativo continuo, la scala di misurazione è una scala di rapporti. Iniziamo a vedere i dati statistici fondamentali del carattere:

```
> summary(Sepal.Length)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  4.300   5.100   5.800   5.843   6.400   7.900
```

Per il calcolo delle frequenze assolute supponiamo che i valori nel carattere abbiano come modalità l'intervallo

2.2 Sepal width

2.3 Petal length

2.4 Petal width

2.5 Species

Riferimenti bibliografici

[1] Montgomery, D., Runger, G., *Statistics and probability for engineers*, Wiley, 2018.

Appendice

Si riportano, in ordine, i codici relativi alla prima e alla seconda parte dell'esercizio e, a seguire, i rispettivi dati su cui sono stati testati.