Laboratorio di Statistica Applicata - 2

Il database `movies`

Federico Plazzi

15 Dicembre 2015

Analisi del database `movies`

Ottenere il database `movies`

- Far partire R
- Menu `Packages` > `Install package(s)...`
- Selezionare un mirror (tipicamente il più vicino geograficamente)
- Selezionare il pacchetto `ggplot2movies`
- Rispondere di sì ad entrambe le richieste di R
- Il database è un oggetto R del tipo "`data.frame`" e si chiama `movies`

Analisi del database `movies`

Ottenere il database `movies`

- Far partire R
- Menu `Packages` > `Install package(s)...`
- Selezionare un mirror (tipicamente il più vicino geograficamente)
- Selezionare il pacchetto `ggplot2movies`
- Rispondere di sì ad entrambe le richieste di R
- Il database è un oggetto R del tipo "`data.frame`" e si chiama `movies`

Il database `movies`

Il database `movies` è un dataframe (cioè una tabella) che contiene 58.788 film, per ciascuno dei quali sono disponibili alcuni dati forniti dall'Internet Movie DataBase (IMDB; `http://imdb.com`). Ad esempio, per vedere solo le prime venti righe è sufficiente digitare `movies[1:20,]`.

Le variabili del database `movies`

Cosa c'è nelle colonne?

- `title`: il titolo del film
- `year`: anno di uscita nelle sale
- `length`: la lunghezza in minuti
- `budget`: il budget in dollari, se disponibile
- `rating`: il voto medio da parte degli utenti IMDB
- `votes`: il numero di utenti IMDB che hanno assegnato un voto al film
- `r1-r10`: decili della distribuzione dei voti
- `mpaa`: classificazione MPAA
- `action`, `animation`, `comedy`, `drama`, `documentary`, `romance`, `short`: 1 se il film è classificabile in quel genere, 0 in caso contrario

Per trovare tutte queste informazioni, è sufficiente digitare `?movies`.

Il database `film.selezione`

Per prima cosa, è bene lavorare con un sottoinsieme di film per cui tutti i dati, compresi i voti, siano affidabili: estraiamo perciò tutti i film che hanno ricevuto un voto da almeno 1.000 utenti e inseriamoli in un nuovo dataframe, `film.selezione`; eliminiamo poi i film classificati come `Short`. Quali comandi bisogna usare?

Il database `film.selezione`

Per prima cosa, è bene lavorare con un sottoinsieme di film per cui tutti i dati, compresi i voti, siano affidabili: estraiamo perciò tutti i film che hanno ricevuto un voto da almeno 1.000 utenti e inseriamoli in un nuovo dataframe, `film.selezione`; eliminiamo poi i film classificati come `Short`. Quali comandi bisogna usare?

```
film.selezione <- movies[movies$votes >= 1000,]
film.selezione <- film.selezione[film.selezione$Short == 0,]
```

Il database `film.selezione`

Per prima cosa, è bene lavorare con un sottoinsieme di film per cui tutti i dati, compresi i voti, siano affidabili: estraiamo perciò tutti i film che hanno ricevuto un voto da almeno 1.000 utenti e inseriamoli in un nuovo dataframe, `film.selezione`; eliminiamo poi i film classificati come `Short`. Quali comandi bisogna usare?

```
film.selezione <- movies[movies$votes >= 1000,]
film.selezione <- film.selezione[film.selezione$Short == 0,]
```

D'ora in poi, lavoriamo su `film.selezione`: può convenire dare un `attach(film.selezione)`!

Esercizi di base

Gestione dei dati

- Richiamare i film con almeno 8 di voto
- Richiamare tutte le commedie o tutti i film d'animazione
- Richiamare tutti i film degli anni '70
- Richiamare tutti i film per cui la presenza dei genitori è consigliata per i bambini sotto i 10 anni di età (categoria MPAA "PG")

Elaborazioni di base

- Calcolare la media degli utenti che assegnano un voto ad un film su IMDB
- Calcolare media, varianza e deviazione standard della lunghezza dei film
- Calcolare il voto medio dei film che sono classificati in più di un genere

Esercizi avanzati

Per i padawan...

- Verificare se i film d'azione hanno dei voti significativamente maggiori delle commedie
- Verificare se i film con una qualche classificazione MPAA sono significativamente più lunghi degli altri
- Verificare se esiste una correlazione significativa tra lunghezza e voto di un film
- Verificare se le commedie di durata superiore a 110' sono significativamente di più dei film d'animazione di durata superiore a 110'

Per i maestri Jedi della statistica...

Selezionare da `film.selezione` i film classificati in un solo genere e salvarli in un nuovo dataframe, `film.per.genere`. Ci sono delle differenze significative nel budget tra i sei diversi generi? E se oltre al genere consideriamo anche il voto (<6 o ≥ 6)?