Statistica I

Esercitazione 6: analisi della varianza

Tommaso Rigon

Università Milano-Bicocca

Descrizione del problema

- I dati a disposizione sono i 455 **voti registrati** dei **primi tre appelli** dell'esame Statistica I per quattro anni accademici (2020/2021, 2021/2022, 2022/2023).

- Siamo interessati a capire se, in media, vi sono delle **differenze tra i gruppi**.

I dati grezzi

Appello I, $n_1 = 300$

```
  [1] 21 31 25 18 24 26 31 26 29 18 20 24 26 20 29 25 18 21 21 26 24 19
 [23] 24 21 23 31 23 31 22 28 20 24 23 24 26 26 26 25 31 24 26 29 24 23
 [45] 22 26 31 24 30 22 22 25 22 31 26 30 29 24 22 24 22 25 19 27 23 20
 [...]
```

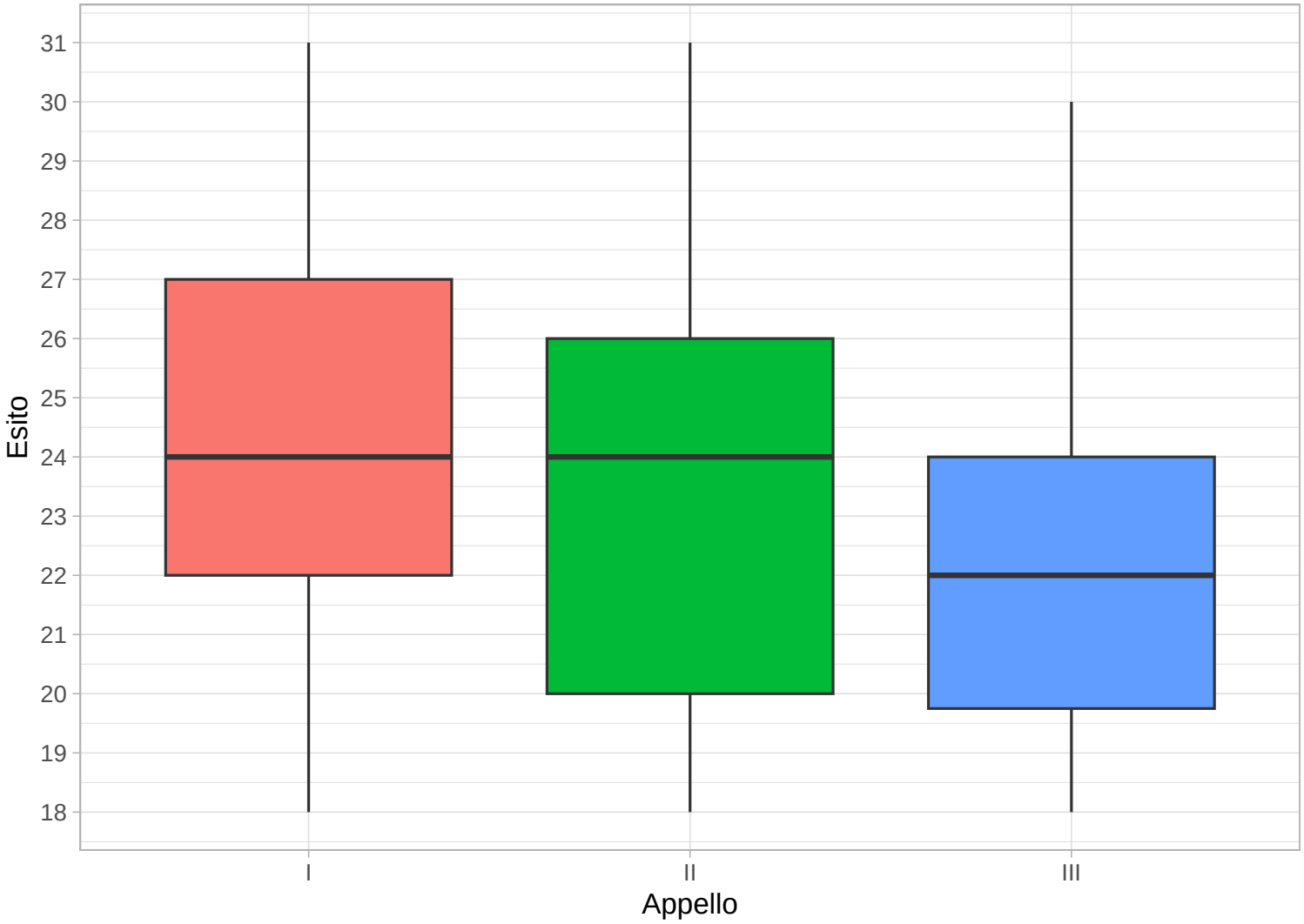
Appello II, $n_2 = 75$

```
 [1] 26 26 24 23 24 21 19 18 21 24 21 19 23 26 26 24 26 26 22 24 30 27
[23] 23 25 28 24 23 25 18 30 23 24 25 27 18 23 20 19 19 26 20 25 25 24
[45] 18 18 18 28 23 18 27 26 18 25 18 24 28 25 24 26 24 21 18 23 31 19
[67] 31 19 24 18 27 23 22 26 18
```

Appello III, $n_3 = 80$

```
 [1] 25 25 23 19 21 23 22 20 25 18 24 21 21 18 20 18 26 21 22 24 22 23
[23] 18 18 26 18 20 25 22 18 20 22 23 25 24 18 23 23 24 18 18 21 20 19
[45] 21 28 21 24 20 24 25 24 21 18 22 30 24 22 20 21 22 24 23 26 22 22
[67] 18 18 27 18 18 20 24 19 29 18 18 20 27 30
```

Boxplot



Alcune statistiche descrittive

- Per semplicità, riportiamo nel seguito alcune statistiche descrittive, ovvero **medie**, **varianze** e numerosità campionaria di ciascun gruppo.

Appello	$\bar{x}_j$	σ_j^2	n_j
I	24.4033	13.114	300
II	23.2267	12.3086	75
III	21.9	9.44	80

Domande

- Si quantifichi, con opportuni indici, il grado di dipendenza in media.
- Si fornisca un'interpretazione dei risultati.

Analisi della varianza I

- La **media globale** è pari a

$$\bar{x} = \frac{300\bar{x}_1 + 75\bar{x}_2 + 80\bar{x}_3}{455} = 23.77.$$

- La **devianza tra i gruppi** è pertanto pari a

$$\mathscr{D}^2_{\text{tr}} = \sum_{j=1}^{2} n_j(\bar{x}_j - \bar{x})^2 =$$

$$= 300(24.4033 - 23.77)^2 + 75(23.2267 - 23.77)^2 + 80(21.9 - 23.77)^2 = 422.2108.$$

- Otteniamo ora le **devianze di ciascun gruppo**, pari a

$$d_1^2 = n_1\sigma_1^2 = 3934.2, \quad d_2^2 = n_2\sigma_2^2 = 923.145, \quad d_3^2 = n_3\sigma_3^2 = 755.2.$$

- Di conseguenza, la **devianza entro i gruppi** è pari a:

$$\mathscr{D}^2_{\text{en}} = d_1^2 + d_2^2 + d_3^2 = 3934.2 + 923.145 + 755.2 = 5612.545.$$

Analisi della varianza II

- Quindi otteniamo che la devianza complessiva è pari a

$$\mathscr{D}^2 = \mathscr{D}^2_{\text{en}} + \mathscr{D}^2_{\text{tr}} = 422.2108 + 5612.545 = 6034.756, \quad (\textbf{6034.769}).$$

- Il rapporto di correlazione è pertanto pari a:

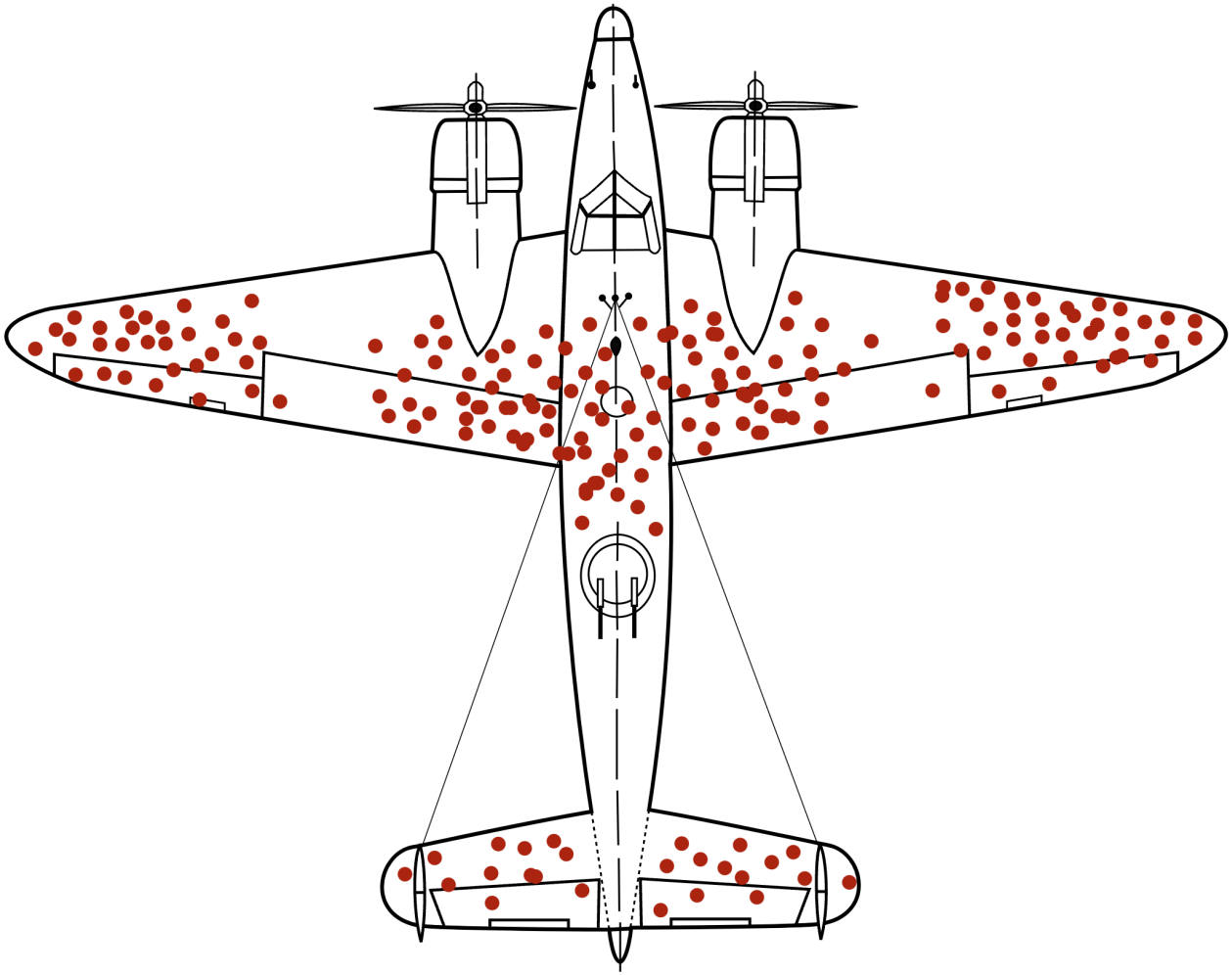
$$\eta^2 = 1 - \frac{\mathscr{D}^2_{\text{en}}}{\mathscr{D}^2} = 1 - \frac{5612.545}{6034.756} = 0.07, \quad (\textbf{0.06997}).$$

- Di conseguenza, la dipendenza in media è **abbastanza debole**, seppur presente.

Commento ai risultati

- Come forse prevedibile, la differenze tra i tre appelli non sono particolarmente marcate. La variabilità individuale prevale di gran lunga rispetto alla tendenza media.
- Detto questo, è possibile notare una lieve ma significativa tendenza: il **voto medio scende** tra un appello ed il successivo. Come mai?
- Smarchiamoci da un pregiudizio: gli esami hanno tutti circa la stessa difficoltà. Del resto, talvolta alcuni esercizi erano proprio gli stessi!
- Ciò che stiamo osservando, invece, è un esempio di ***selection bias*** o **distorsione da selezione**. Nello specifico, si parla di ***survivorship bias***.
- Coloro che hanno superato con ottimi voti l'esame al primo appello, non si presentano al secondo. Chi sostiene la seconda e la terza prova è quindi un "**sopravvissuto**".
- Dato che la popolazione di studenti tra appelli è differente, la media di ciascuna popolazione decresce di volta in volta.

La leggenda degli aeroplani di Wald



- Tra mito e verità: `https://www.ams.org/publicoutreach/feature-column/fc-2016-06`.

Descrizione del problema

- Per valutare tre differenti strategie mediche per trattare l'**ipertensione**, sono state individuate $n = 18$ persone di sesso maschile, leggermente sovrappeso e sedentarie.

I dati grezzi

- Il campione consiste di persone ipertese (pressione sistolica maggiore di 100mmHg) e sono state suddivise in tre gruppi:
  - Il primo gruppo (5 persone) ha seguito una terapia farmacologica.
  - Il secondo gruppo (7 persone) ha seguito una dieta prefissata.
  - Il terzo gruppo (6 persone) ha seguito la dieta del secondo gruppo ma ha anche svolto regolarmente delle attività fisiche.

- La **pressione sistolica** è stata misurata sia all'inizio che dopo 3 mesi dall'ingresso nella studio.

- La seguente tabella mostra, per ognuno dei 18 individui, la **differenza** tra la **pressione iniziale** e quella rilevata **dopo 3 mesi**.

Trattamento	1	2	3	4	5	6	7
Solo farmaco	21	20	7	11	16		
Solo dieta	-9	13	1	2	24	6	9
Dieta ed esercizio fisico	19	18	21	8	8	18	

Domande

- Perchè secondo voi è stata utilizzata la differenza tra la pressione iniziale e quella finale e non direttamente quest'ultima?
- Sulla base dei dati disponibili, quale strategia sembra funzionare meglio? Si risponda con opportuni indici.
- Si quantifichi la correlazione tra strategia utilizzata e la **differenza in pressione sistolica**.

Analisi della varianza I

- È stata utilizzata la differenza perchè in questo modo siamo in grado di misurare il **miglioramento** dato dalla terapia, indipendentemente dal livello di partenza.

- Otteniamo anzitutto alcune analisi descrittive di base, riportate nella tabella seguente.

Trattamento	$\bar{x}_j$	σ_j^2	d_j^2	n_j
Solo formaco	15	28.4	142	5
Solo dieta	6.57	92.24	645.71	7
Dieta & esercizio fisico	15.33	27.89	167.33	6

- Di conseguenza, la strategia che sembra funzionare meglio è quella basata sulla **dieta insieme all'esercizio fisico**, anche se di poco.

- Inoltre, la **media globale** è pari a

$$\bar{x} = \frac{(5\bar{x}_1 + 7\bar{x}_2 + 6\bar{x}_3)}{18} = 11.83.$$

Analisi della varianza II

- La **devianza tra i gruppi** è pertanto pari a

$$\mathscr{D}^2_{\text{tr}} = \sum_{j=1}^{3} n_j(\bar{x}_j - \bar{x})^2 = 5(15-11.83)^2 + 7(6.57-11.83)^2 + 6(15.33-11.83)^2 = 317.42.$$

- Invece, la **devianza entro i gruppi** è pari a:

$$\mathscr{D}^2_{\text{en}} = d_1^2 + d_2^2 + d_3^2 = 142 + 645.71 + 167.33 = 955.04.$$

- Quindi otteniamo che la **devianza complessiva** è pari a

$$\mathscr{D}^2 = \mathscr{D}^2_{\text{en}} + \mathscr{D}^2_{\text{tr}} = 317.42 + 955.04 = 1272.5.$$

- Il **rapporto di correlazione** è pertanto pari a:

$$\eta^2 = 1 - \frac{\mathscr{D}^2_{\text{en}}}{\mathscr{D}^2} = 1 - \frac{955.04}{1272.5} = 0.25.$$

Di conseguenza, la dipendenza in media è moderata.