Statistica I

Unità N: analisi della varianza

Tommaso Rigon

Università Milano-Bicocca

Anno Accademico 2020-2021

Unità N

Argomenti affrontati

- Rapporto tra medie e varianze condizionate e media e varianza marginali
- Una misura della dipendenza in media
- Analisi della varianza

Descrizione del problema

- Per capire quanto il **tipo di carne** con cui vengono preparati gli hot-dog influenza il loro **contenuto calorico**, sono state misurate le calorie di ciascun hotdog in $n = 54$ confezioni di diverse marche.

- È inoltre noto se l'hot-dog era stato preparato con: carne bovina; carne mista (in larga parte maiale); pollame (pollo o tacchino).

- Siamo interessati a quantificare la **connessione** tra la variabile `carne` e la variabile `calorie`.

- Le prossime pagine mostrano: i dati grezzi; le funzioni di ripartizione empirica; i boxplot; le principali statistiche descrittive dei tre gruppi.

I dati grezzi

carne	calorie	carne	calorie	carne	calorie
Bovina	186	Bovina	181	Bovina	176
Bovina	149	Bovina	184	Bovina	190
Bovina	158	Bovina	139	Bovina	175
Bovina	148	Bovina	152	Bovina	111
Bovina	141	Bovina	153	Bovina	190
Bovina	157	Bovina	131	Bovina	149
Bovina	135	Bovina	132	Mista	173
Mista	191	Mista	182	Mista	190
Mista	172	Mista	147	Mista	146
Mista	139	Mista	175	Mista	136
Mista	179	Mista	153	Mista	107
Mista	195	Mista	135	Mista	140
Mista	138	Pollame	129	Pollame	132
Pollame	102	Pollame	106	Pollame	94
Pollame	102	Pollame	87	Pollame	99
Pollame	107	Pollame	113	Pollame	135
Pollame	142	Pollame	86	Pollame	143
Pollame	152	Pollame	146	Pollame	144

Distribuzione bivariata

- Le osservazioni a nostra disposizione possono essere viste come un insieme di **dati bivariati**.
- Le unità statistiche sono i singoli hot-dog, le due variabili sono `carne` (**qualitativa**) e `calorie` (**numerica**).

hot-dog	carne	calorie
1	Bovina	186
2	Bovina	149
⋮	⋮	⋮
54	Pollame	144

Statistiche descrittive

- È evidente che gli hot-dog preparati con **pollame** sono tendenzialmente più **poveri di calorie**.

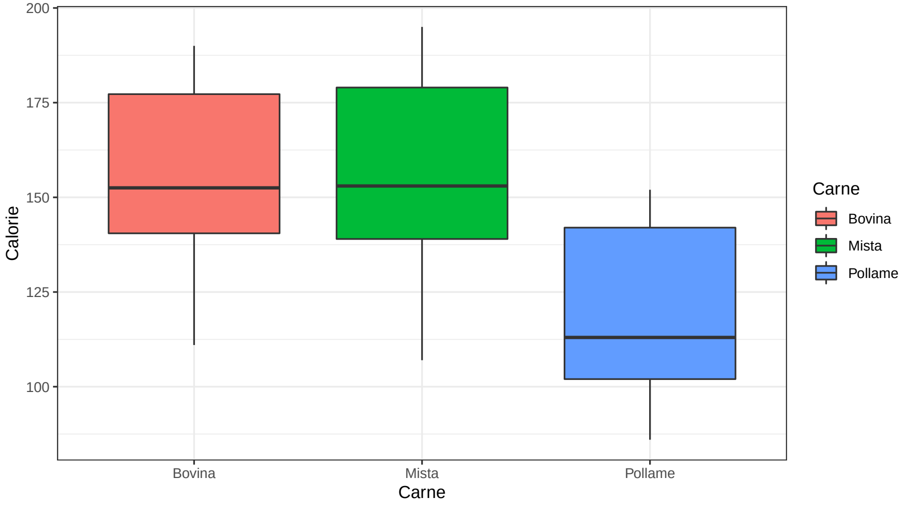
- Questo è confermato dalla media e dalla mediana, riportati nella tabella seguente.

Tipo di carne	Numerosità	Media	Mediana	Deviazione standard
Bovina	20	156.85	152.5	22.07
Mista	17	158.71	153	24.48
Pollame	17	118.76	113	21.88

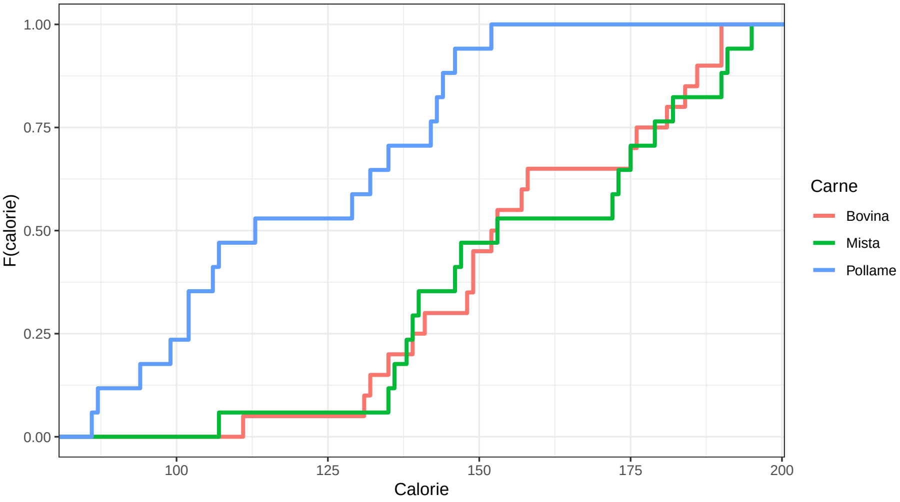
- È inoltre noto che la media complessiva dei dati è $\bar{y} = 145.44$.

- Le variabile `carne` e la variabile `calorie` sono intuitivamente **connesse**. Infatti, le medie di ciascun gruppo sono diverse tra loro.

I boxplot


200
175
150
125
100
Calorie
Bovina
Mista
Pollame
Carne
Carne
Bovina
Mista
Pollame

Le funzioni di ripartizione



Connessione e dipendenza in media

- Siamo interessati a quantificare, con opportuni indici, la **connessione** tra due variabili.

- La connessione è l'equivalente della correlazione quando una delle due variabili è qualitativa.

- Sebbene esistano vari modi per definire la connessione tra variabili, noi ci focalizzeremo principalmente sulle **differenze tra le medie** dei gruppi.

- Quindi, quando la connessione è forte, diremo che c'è **dipendenza in media**, nel senso che le medie "dipendono" dalla variabile qualitativa.

- Viceversa, se le medie dei gruppi sono uguali tra loro, la connessione è debole e parleremo invece di **indipendenza in media**.

Le medie dei gruppi

- In generale, indicheremo con k il **numero di gruppi**.
- Inoltre, le frequenze $n_1, \dots, n_k$ indicano il **numero di osservazioni** per ciascun gruppo, e quindi

$$(\text{numerosità campionaria}) = n = \sum_{j=1}^{k} n_j.$$

- L'insieme di tutte le osservazioni può essere quindi indicato come

$$y_{ij} = (\text{osservazione } i\text{-esima del gruppo } j\text{-esimo}), \qquad i = 1, \dots, n_j, \quad j = 1, \dots, k.$$

- È quindi possibile calcolare le **medie dei gruppi**, che indicheremo come

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \qquad j = 1, \dots, k.$$

- Nel nostro caso avremo $k = 3$ gruppi con frequenze $n_1 = 20$ (carne bovina), $n_2 = 17$ (carne mista) e $n_3 = 17$ (pollame). Inoltre: $\bar{y}_1 = 156.85$, $\bar{y}_2 = 158.71$ e $\bar{y}_3 = 118.76$.

La distribuzione delle medie dei gruppi

Modalità	$\bar{y}_1$	$\bar{y}_2$	$\cdots$	$\bar{y}_k$
Frequenze	n_1	n_2	$\cdots$	n_k

- Consideriamo una distribuzione le cui modalità sono le medie dei k gruppi e le cui frequenze sono le numerosità delle osservazioni nei gruppi.

- **Proprietà**. La media di questa distribuzione è pari a

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{y}_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} y_{ij},$$

ovvero è pari alla media complessiva dei dati.

- **Esercizio**. Si dimostri questa proprietà.

La devianza tra i gruppi

- Lo scopo dell'analisi è identificare un indice di **connessione**. Se le medie dei gruppi sono molto diverse tra loro significa che la connessione è forte.

- Di conseguenza, un possibile indice di connessione potrebbe essere la **varianza delle medie dei gruppi**. Per praticità, in questo contesto si preferisce usare la devianza.

- **Devianza tra i gruppi**. La devianza tra i gruppi è pari a

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

- La devianza è quindi una varianza che non viene divisa per n.

- La devianza tra i gruppi tuttavia dipende dalla scala di y ed è di difficile interpretazione.

Devianza entro i gruppi e devianza totale

- Prima di procedere, consideriamo due ulteriori quantità: le **varianze delle osservazioni in ciascun gruppo** e la **varianza complessiva** σ^2 (o meglio, le rispettive devianze).

- **Devianza entro i gruppi**. La devianza delle osservazioni nel j-esimo gruppo è

$$d_j^2 = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2, \qquad j = 1, \dots, k.$$

Quindi, la devianza entro i gruppi è pari a

$$\mathscr{D}_{\text{en}}^2 = \sum_{j=1}^{k} d_j^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2.$$

- **Devianza totale**. La devianza complessiva delle osservazioni è

$$\mathscr{D}^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y})^2,$$

Decomposizione della devianza

- La **devianza tra i gruppi** misura la dispersione delle medie dei gruppi dal loro centro.
- La **devianza entro i gruppi** misura la dispersione delle osservazioni dal centro del rispettivo gruppo.
- La **devianza totale** misura la dispersione delle osservazioni dalla media dei dati.
- **Teorema (decomposizione della devianza)**. Vale la seguente decomposizione

$$\text{(devianza totale)} = \text{(devianza tra i gruppi)} + \text{(devianza entro i gruppi)}.$$

Più precisamente, avremo che

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

- Di conseguenza si avrà che $0 \leq \mathscr{D}^2_{\text{tr}} \leq \mathscr{D}^2$, suggerendo quindi una normalizzazione per la devianza entro i gruppi.

Dimostrazione

- La dimostrazione è molto semplice:

$$
\begin{aligned}
\mathscr{D}^2 &= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2 = \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(y_{ij} - \bar{y}_j) + (\bar{y}_j - \bar{y})]^2 = \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}[(y_{ij} - \bar{y}_j)^2 + (\bar{y}_j - \bar{y})^2 + 2(y_{ij} - \bar{y}_j)(\bar{y}_j - \bar{y})] = \\
&= \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2 + \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 + 2\sum_{j=1}^{k}(\bar{y}_j - \bar{y})\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j) = \\
&= \mathscr{D}^2_{\text{en}} + \mathscr{D}^2_{\text{tr}}.
\end{aligned}
$$

L'indice di connessione η^2

- Il teorema di decomposizione della devianza consente di definire un indicatore **normalizzato** di connessione.

- **Coefficiente di connessione η^2**. Il coefficiente di connessione η^2 è pari a

$$\eta^2 = \frac{\text{(devianza tra i gruppi)}}{\text{(devianza totale)}} = 1 - \frac{\text{(devianza entro i gruppi)}}{\text{(devianza totale)}},$$

ovvero

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

- L'indice è normalizzato, poiché $0 \leq \eta^2 \leq 1$.

- L'indice η^2 misura la forza della **dipendenza in media**.

Interpretazione di η^2

- L'**interpretazione** dell'indice η^2 è agevole.

- Se le osservazioni non variano entro i gruppi (sono tutte pari alla media del gruppo), allora la devianza entro i gruppi è nulla $\mathscr{D}^2_{\text{en}} = 0$ e la **connessione è massima** e $\eta^2 = 1$.

- La connessione massima si ottiene anche quando la varianza tra i gruppi è molto grande rispetto alla varianza entro i gruppi.

- Se la devianza tra i gruppi è nulla $\mathscr{D}^2_{\text{tr}} = 0$, allora le medie di tutti i gruppi sono uguali tra loro. Di conseguenza la **connessione è minima** e $\eta_2 = 0$.

- Si noti che η^2 non è definito quando $\mathscr{D}^2 = 0$. Questo non costituisce un problema, in pratica, poiché $\mathscr{D}^2 = 0$ significa che tutte le osservazioni sono uguali tra loro.

- Nell'ultimo caso descritto, non c'è nessuna "varianza" da analizzare.

Hot-dog e decomposizione della devianza

- Nel caso degli hot-dog, il coefficiente η^2 è facilmente calcolabile.

- A partire dalla tabella presentata nella slide 5, si ottiene

$$\begin{aligned}
(\text{devianza tra i gruppi}) &= \mathscr{D}^2_{\text{tr}} \approx 17692.2, \\
(\text{devianza entro i gruppi}) &= \mathscr{D}^2_{\text{en}} \approx 28067.78, \\
(\text{devianza totale}) &= \mathscr{D}^2 \approx 45759.33.
\end{aligned}$$

- Pertanto, si ottiene $\eta^2 = 0.39$. Il valore indica la presenza di una discreta ma non eccezionale connessione tra `carne` e `calorie`.

- Questo è probabilmente dovuto al fatto che vi sono poche differenze tra carne bovina e carne mista, in termini di calorie.

- **Esercizio**. Si ottengano le devianze $\mathscr{D}^2_{\text{tr}}, \mathscr{D}^2_{\text{en}}$ e $\mathscr{D}^2$ a partire dalla slide 5.