Corso di Statistica - Esercitazione 2

Dott. Davide Buttarazzi

✉ *d.buttarazzi@unicas.it*

Esercizio 1

La seguente tabella riporta dati relativi al giudizio espresso da alcuni clienti sulla qualità dell'ultimo modello di smartphone prodotto da una nota azienda.

Valutazione	$\mathbf{n}_i$
Inaccettabile	250
Scarsa	500
Accettabile	1500
Buona	2100
Ottima	350
Tot.	**4700**

1. Definire la tipologia della variabile *Valutazione*
2. Calcolare l'indice di asimmetria standardizzato

Soluzioni esercizio 1

1. *Valutazione* è una variabile di tipo qualitativo ordinale
2. L'indice di asimmetria per variabili qualitative ordinali può essere definito come $A^* = \frac{D_{dx} - D_{sx}}{D_{dx} + D_{sx}}$

   Occorre quindi individuare la parte destra e la parte sinistra della distribuzione sulle quali calcolare l'indice di dispersione. Lo mediana rappresenta lo strumento per dividere la distribuzione. Operando con le frequenze assolute cumulate si ha che $\frac{N}{2} = 2350$, pertanto *Buona* risulta essere la modalità mediana. La parte sinistra e destra della distribuzione possono essere così suddivise:

   Parte sinistra

Valutazione	$\mathbf{n}_i$	$\mathbf{N}_i$	$\mathbf{F}_i$
Inaccettabile	250	250	0.106
Scarsa	500	750	0.319
Accettabile	1500	2250	0.957
Buona	2100	2350	1

   Parte destra

Valutazione	$\mathbf{n}_i$	$\mathbf{N}_i$	$\mathbf{F}_i$
Buona	2100	2000	0.85
Ottima	350	2350	1

   $$D_{sx} = 2\sum_{i=1}^{k-1} F_i(1 - F_i) = 2(0.0947 + 0.217 + 0.041) = 0.7054$$

   $$D_{dx} = 2\sum_{i=1}^{k-1} F_i(1 - F_i) = 2[0.85(1 - 0.85)] = 0.255$$

   L'indice di asimmetria normalizzato sarà quindi:

   $A = \frac{D_{dx} - D_{sx}}{D_{dx} + D_{sx}} = \frac{0.255 - 0.7054}{0.255 + 0.7054} = \frac{-0.4504}{0.9604} = -0.47$ (segnale di asimmetria negativa)

Esercizio 2

La seguente tabella riporta il prezzo di mercato di un campione di smartphone considerati da una nota azienda come principali competitor.

Prezzo	200	99	180	450	20	130	100	100	100	360	150	130	200	50	100	195	140	140

1. Definire la tipologia della variabile *Prezzo*
2. Calcolare l'indice di asimmetria di Pearson
3. Calcolare l'indice di asimmetria di Fisher (relativo)
4. Calcolare l'indice di Yule-Bowley
5. Costruire il Boxplot ed individuare eventuali valori anomali

Soluzioni esercizio 2

1. *Prezzo* è una variabile numerica.
2. $A_{Pearson} = \frac{\mu - Me}{\sigma}$

   $\mu = \sum_{i=1}^{N} \frac{x_i}{N} = \sum_{i=1}^{18} \frac{Prezzo_i}{18} = 158$

   $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} = \frac{181374}{18} = \sqrt{10076} = 100.3$

   $Me = x_{(\frac{n+1}{2})} = x_i : F_i \geq 0.5 = 140$

   Pertanto: $A_{Pearson} = \frac{158-140}{100.3} = 0.18$ (segnale di asimmetria a destra)

3. L'indice di asimmetria di Fisher $I_F = \frac{\sum_{i=1}^{n}(x_i-\mu)^3}{n}$ é espresso nell'unità di misura al cubo. Pertanto può essere ricondotto all'unità di misura originale estraendo la sua radice cubica, oppure può essere trasformato in indice relativo dividendolo per un altro indice espresso nella stessa unità di misura, come σ^3.

   Per facilitare i calcoli, è opportuno costruire la seguente tabella:

$\mathbf{x}$	$\mathbf{x} - \mu$	$\mathbf{x} - \mu^3$
200	42	74088
99	-59	-205379
180	22	10648
450	292	24897088
20	-138	-2628072
130	-28	-21952
100	-58	-195112
100	-58	-195112
100	-58	-195112
360	202	8242408
150	-8	-512
130	-28	-21952
200	42	74088
50	-108	-1259712
100	-58	-195112
195	37	50653
140	-18	-5832
140	-18	-5832

Avendo già calcolato $\sigma = 100.3$ nell'esercizio precedente, l'indice di Fisher (relativo) sarà:

$$\frac{I_F}{\sigma^3} = \frac{\left(\frac{\sum_{i=1}^{n}(x_i-\mu)^3}{n}\right)}{\sigma^3} = \frac{\left(\frac{28419282}{18}\right)}{1009027} = 1.564724$$

Lo stesso risultato può essere ottenuto calcolando la media degli scarti standardizzati al cubo, ovvero $I_F^* = \frac{\sum_{i=1}^{n}(\frac{x_i-\mu}{\sigma})^3}{n}$

4. L'indice di Yule-Bowley è definito come $A_{YB} = \frac{(q_3-Me)-(Me-q1)}{(q_3-Me)+(Me-q1)}$. È necessario quindi individuare i quartili della distribuzione.

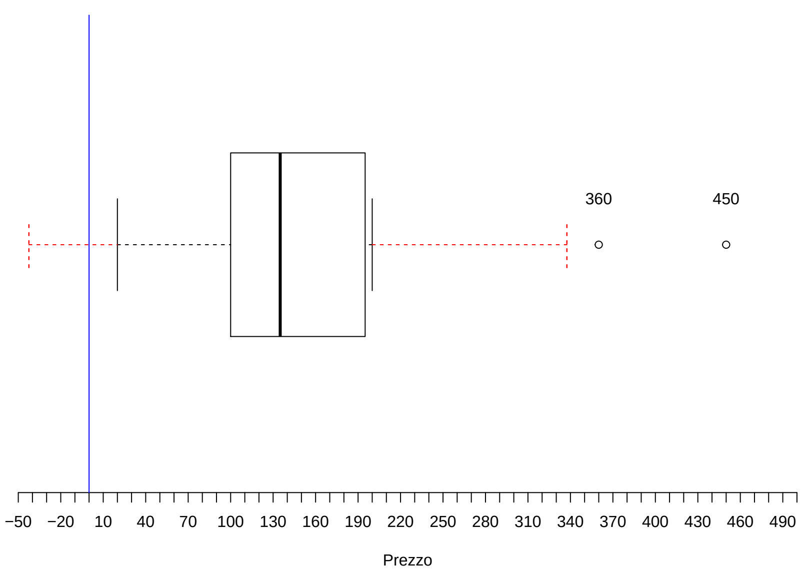
   Per calcolare i quartili occorre riorganizzare la serie grezza in ordine non-decrescente:

   20 50 99 100 100 100 100 130 130 140 140 150 180 195 200 200 360 450

   $Q1 = x_{(\frac{n}{4})} = x_i : F_i \geq 0.25 = 100$
   $Q2 = Me = x_{(\frac{n+1}{2})} = x_i : F_i \geq 0.5 = 140$
   $Q3 = x_{(\frac{3n}{4})} = x_i : F_i \geq 0.75 = 195$

   Quindi: $A_{YB} = \frac{(q_3-Me)-(Me-q1)}{(q_3-Me)+(Me-q1)} = \frac{(195-140)-(140-100)}{(195-140)+(140-100)} = \frac{55-40}{55+40} = \frac{15}{95} = 0.16$ (segnale di asimmetria destra)

5. Il boxplot sulla variabile *Prezzo* pu essere cos rappresentato:



   Eventuali valori anomali possono essere individuati all'esterno dei seguenti valori soglia (in rosso nel grafico):
   $q1 - 1.5 \times (q3 - q1) = -42.5$ (ovvero 0, se si assume che il prezzo degli smartphone sia non-negativo)
   $q3 + 1.5 \times (q3 - q1) = 337.5$
   I valori (prezzi) anomali sono quindi $x_i = 360$ ed $x_i = 450$
   Minimo e massimo relativi (nel grafico i baffi neri non tratteggiati) saranno quindi, rispettivamente, $x_i = 20$ ed $x_i = 200$.

Esercizio 3

Utilizzando i dati dell'esercizio 2:

1. calcolare l'indice di concentrazione di Gini normalizzato per la variabile *Prezzo*
2. rappresentare graficamente la curva di Lorenz

Soluzioni esercizio 3

1. L'indice di concentrazione di Gini normalizzato può essere così calcolato:

$$G^* = \frac{\sum_{i=1}^{n} (p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$$

Dopo aver ordinato in modo non-decrescente la serie grezza, per calcolare l'indice di Gini è opportuno costruire la seguente tabella:

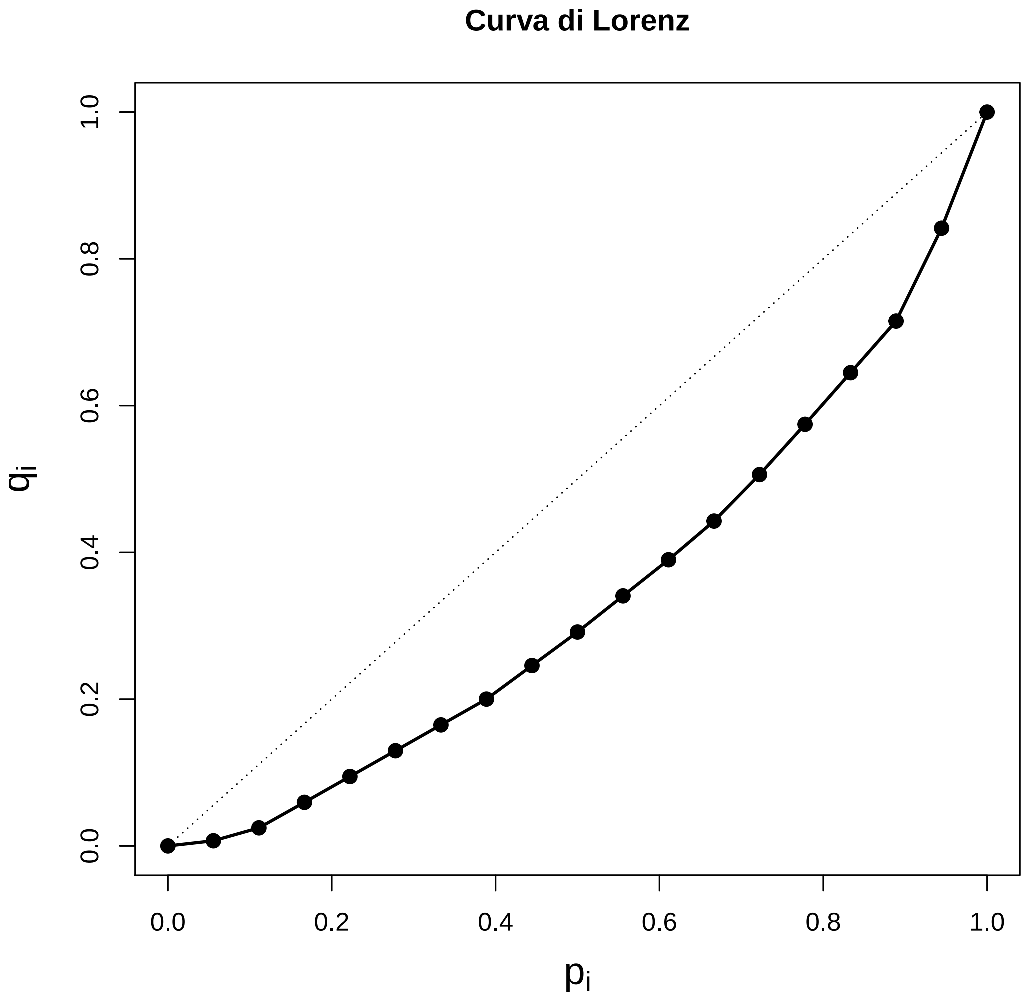
i	$x_i(ord)$	$\sum_{j=1}^{i} x_j$	$p_i = \frac{i}{N}$	$q_i = \frac{\sum_{j=1}^{i} x_j}{N}$	$p_i - q_i$
1	20	20	0.056	0.007	0.049
2	50	70	0.111	0.025	0.086
3	99	169	0.167	0.059	0.107
4	100	269	0.222	0.095	0.128
5	100	369	0.278	0.130	0.148
6	100	469	0.333	0.165	0.168
7	100	569	0.389	0.200	0.189
8	130	699	0.444	0.246	0.199
9	130	829	0.500	0.291	0.209
10	140	969	0.556	0.341	0.215
11	140	1109	0.611	0.390	0.221
12	150	1259	0.667	0.443	0.224
13	180	1439	0.722	0.506	0.216
14	195	1634	0.778	0.575	0.203
15	200	1834	0.833	0.645	0.188
16	200	2034	0.889	0.715	0.174
17	360	2394	0.944	0.842	0.103
18	450	2844	1	1	0
	N=2844				

L'indice di Gini può essere ottenuto quindi sommando tutti gli elementi dell'ultima colonna della tabella costruita $(p_i - q_i)$. Per normalizzarlo occorre dividere per la somma di tutti gli elementi della quarta colonna della tabella ad eccezione dell'ultimo (p_i, per $i = 1, ..., 17$).

Formalmente:

$$G^* = \frac{\sum_{i=1}^{n} (p_i - q_i)}{\sum_{i=1}^{n-1} p_i} = \frac{2.826653}{8.5} = 0.3325$$

2. La curva (o spezzata) di Lorenz si ottiene collegando i punti di coordinate (p_i, q_i), includendo l'origine degli assi tra le coordinate.



La retta tratteggiata rappresenta il caso di equidistribuzione ($p_i = q_i$). L'area compresa tra la retta di equidistribuzione e la spezzata di Lorenz denominata area di concentrazione (ovvero indice di Gini). Tale quantità è calcolabile anche sottraendo ad 1/2 (area al di sotto della retta di equidistribuzione), l'area al di sotto della curva di Lorenz (calcolabile sommando le aree dei 18 trapezi che si ottengono proiettando ortogonalmente i punti di coordinate (p_i, q_i) sull'asse delle ascisse).

Per i procedimenti dettagliati su come ricavare gli indici di posizione e variabilità utilizzati negli esercizi sopra svolti, si faccia riferimento all'esercitazione 1 del 3-11-2016.