Statistica I

Esercitazione 2: indici di posizione

Tommaso Rigon

Università Milano-Bicocca

UNIVERSITA' DEGLI STUDI DI MILANO BICOCCA

Descrizione del problema (continuazione)

- Si faccia riferimento ai dati dell'Esercitazione 1 riguardanti gli **abeti rossi**. Vengono misurati **ulteriori** $n = 30$ **diametri**.

Campione addizionale di abeti rossi (diametro), $n = 30$

```
 [1] 20 20 22 23 28 29 29 31 33 34 34 37 38 38 39 40 42 42 44 44 45 49 54
[24] 54 54 55 58 59 60 68
```

Domande

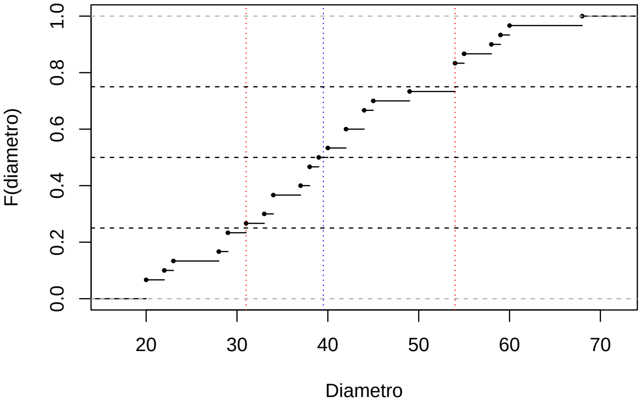
- Si costruisca tale **tabella** indicando frequenze assolute, cumulate assolute e cumulate relative per ciascun diametro.
- Si calcolino la **media aritmetica** e la **mediana**.
- Si calcoli il **primo quartile** ($\mathcal{Q}_{0.25}$), il **terzo quartile** ($\mathcal{Q}_{0.75}$) ed il **quarto decile** ($\mathcal{Q}_{0.4}$), seguendo la definizione data nell'unità C.
- Si rappresentino primo, secondo (mediana) e terzo quartile nel grafico della **funzione di ripartizione**.

Tabella di frequenze

D (diametro)	n_j	N_j	f_j	F_j
20	2	2	0.067	0.067
22	1	3	0.033	0.100
23	1	4	0.033	0.133
28	1	5	0.033	0.167
29	2	7	0.067	0.233
31	1	8	0.033	0.267
33	1	9	0.033	0.300
34	2	11	0.067	0.367
37	1	12	0.033	0.400
38	2	14	0.067	0.467
39	1	15	0.033	0.500
40	1	16	0.033	0.533
42	2	18	0.067	0.600
44	2	20	0.067	0.667
45	1	21	0.033	0.700
49	1	22	0.033	0.733
54	3	25	0.100	0.833
55	1	26	0.033	0.867
58	1	27	0.033	0.900
59	1	28	0.033	0.933
60	1	29	0.033	0.967
68	1	30	0.033	1

Quantili e funzione di ripartizione

- La **media** è circa 40.77, la **mediana** è 39.5. Il **primo quartile** è pari a 31, il **terzo quartile** è pari a 54 mentre il **quarto decile** è pari a 37.



Commenti ai risultati

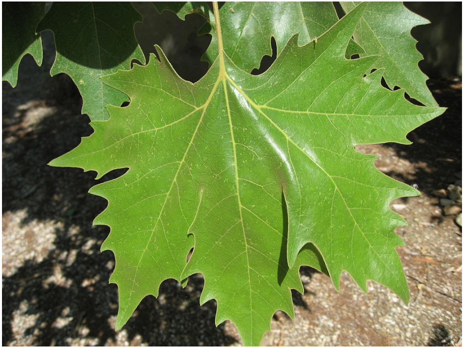
- Questo nuovo insieme di dati è del tutto **compatibile** con i dati degli abeti rossi analizzati nell'Esercitazione 1; qui la numerosità campionaria è $n = 30$.

- Da un punto di vista "tecnico", l'esercizio serve a insegnare come calcolare media, mediana e quantili in presenza di **osservazioni ripetute**.

- Da un punto di vista dell'analisi, i dati confermano quanto già visto nell'Esercitazione 1: il **centro della distribuzione** è circa 40cm; media e mediana sono quasi coincidenti.

- Inoltre, come in parte già notato in precedenza, circa la **metà dei dati** ha diametro compreso tra 31 e 54 cm.

Descrizione del problema



- Siamo interessati a quantificare la **lunghezza** di $n = 100$ foglie di platano, dopo 10 giorni di siccità.

- È noto che, dopo un giorno di pioggia, la **lunghezza** delle foglie **aumenta** di una quota percentuale del 10% più una quota fissa pari a 0.5 mm.

Dati raggruppati e domande

- È riportata di seguito una tabella che riassume le **lunghezze** in mm. Si noti che le classi **non** sono **equispaziate**.

Classe	Frequenza assoluta n_j	Frequenza cumulata N_j
$(120, 135]$	10	10
$(135, 145]$	20	30
$(145, 150]$	60	90
$(150, 165]$	10	100

Domande

- Si calcoli un'approssimazione della **media** e della **mediana**.
- Si calcoli un'approssimazione della media e della mediana **dopo un giorno di pioggia** continuata.

Schema della soluzione I

- I dati sono **raggruppati** per cui non è possibile calcolare la media e la mediana dei dati originari. È però possibile ottenere un'approssimazione.

- La **mediana** è il valore medio delle unità statistiche in posizione 50 e 51. Entrambe queste unità appartengono alla classe $(145, 150]$ e perciò anche la mediana.

- Una possibile **approssimazione** per la mediana si ottiene tramite la formula:

$$\mathrm{Me}_x \approx 145 + (150 - 145)\frac{0.5 - 0.3}{0.9 - 0.3} = 145 + (150 - 145)\frac{50 - 30}{90 - 30} = 146.7.$$

- Siano $m_1, \dots, m_4$ i valori centrali degli intervalli. Allora, un valore approssimato per la **media** è

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{4} m_j n_j = \frac{1}{100}(127.5 \times 10 + \cdots + 157.5 \times 10) = 145.$$

Schema della soluzione II

- Il secondo quesito si risolve rapidamente poichè è equivalente al calcolo della media e della mediana dei **dati trasformati**

$$y_i = 0.5 + (1 + 0.1)x_i, \qquad i = 1, \dots, n.$$

- I dati $y_1, \dots, y_n$ rappresentano le lunghezze i mm dopo un giorno di pioggia continuata.

- Trattandosi di una trasformazione **lineare** monotona crescente, sfruttando le proprietà della mediana si ottiene che

$$\mathrm{Me}_y = 0.5 + (1 + 0.1)\mathrm{Me}_x \approx 161.87,$$

mentre per la media vale che

$$\bar{y} = 0.5 + (1 + 0.1)\bar{x} \approx 160.$$

Descrizione del problema

- In una strada rettilinea sono collocati 5 condomini, chiamati A, B, C, D ed E. Il comune desidera determinare la **posizione ottimale** per un nuovo supermercato.

- I condomini sono occupati dal seguente **numero di inquilini**:

	A	B	C	D	E
Numero di inquilini	6	6	20	12	8

- Inoltre, la **posizione** dei condomini, ovvero i metri di **distanza** dall'inizio della via, sono:

	A	B	C	D	E
Distanza dall'inizio della via (metri)	1000	2000	3000	3100	3150

Domande

- Si indichi la posizione ideale del supermercato, ovvero la **posizione che minimizza il disagio** degli inquilini in termini di **distanza percorsa**, in due casi:
  - Supponendo che disagio cresca **linearmente** con la distanza.
  - Supponendo che disagio cresca con il **quadrato** della distanza.

Schema della soluzione I

- Siano $x_1, \dots, x_{52}$ le **posizioni** dei vari inquilini nella strada rettilinea.
- Sebbene i coinquilini siano 52, le modalità sono solamente 5, dal momento che essi vivono in 5 condomini.
- Nella **tabella** seguente, riscriviamo il testo del problema con una notazione più familiare.

Modalità c_j	Frequenza assoluta n_j	Frequenza cumulata N_j
1000	6	6
2000	6	12
3000	20	32
3100	12	44
3150	8	52

Schema della soluzione II

- Il primo quesito chiede di individuare il valore α, ovvero la posizione ottimale del supermercato, che minimizza la seguente quantità

$$\sum_{i=1}^{52} |x_i - \alpha| = \sum_{j=1}^{5} n_j |c_j - \alpha| = 6|1000 - \alpha| + \cdots + 8|3150 - \alpha|.$$

- Un possibile valore (ce ne potrebbero essere tanti!), che minimizza tale somma è la **mediana**.

- In questo caso, la numerosità campionaria è pari a $n = 52$ e pertanto la mediana è pari alla media dei due valori centrali, ovvero i valori degli individui in posizione 26 e 27.

- Poichè entrambi sono pari a 3000, la mediana è a sua volta pari a 3000.

- Inoltre, poichè $x_{(26)} = x_{(27)} = 3000$, il valore $\alpha = 3000$ è l'**unico** che minimizza la somma degli scarti in valore assoluto.

Schema della soluzione III

- La seconda domanda chiede il valore α che minimizza il **quadrato delle distanza**, ovvero il valore che rende minima la somma

$$\sum_{i=1}^{52}(x_i - \alpha)^2 = \sum_{j=1}^{5} n_j(c_j - \alpha)^2 = 6(1000 - \alpha)^2 + \cdots + 8(3150 - \alpha)^2.$$

- Come visto nell'unità C, tale valore è la **media aritmetica**.

- Si ottiene quindi che il valore cercato è semplicemente

$$\alpha = \bar{x} = \frac{1}{52}\sum_{i=1}^{52} x_i = \frac{1}{52}\sum_{j=1}^{5} n_j c_j = \frac{1}{52}(1000 \times 6 + \cdots + 3150 \times 8) = 2700.$$