Corso di Modelli per l’analisi statistica

Prof. G. Scandurra

a.a. 2020-2021

Indicatori sintetici

- Tendenza centrale
  (media, mediana, moda)
- Variabilità
  (varianza, campo di variazione, concentrazione, eterogeneità)
- Forma
  (asimmetria)

Tendenza centrale: la media

- Il modo più intuitivo per sintetizzare un insieme di valori passa attraverso il calcolo della media
- Media (aritmetica) = punto di equilibrio o baricentro dell’insieme di valori
- È una media analitica, funzione di tutti i valori

Calcolo della media dei ricavi

- Conoscendo i ricavi dei 9 punti vendita dell'azienda, posso essere interessato ad ottenere il ricavo medio, un unico valore rappresentativo dell'intero insieme
- Si sommano i ricavi di tutti i punti vendita e il risultato si divide per il numero delle osservazioni (n=9)

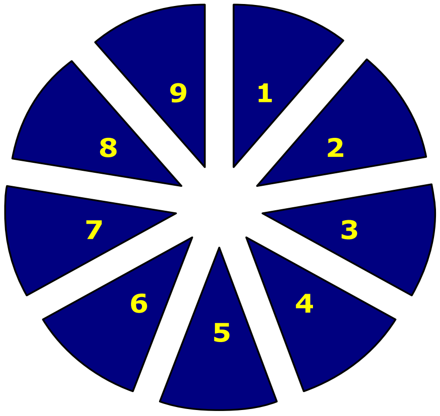
Calcolo della media

Punti vendita	Ricavi
1	350
2	200
3	600
4	500
5	270
6	180
7	205
8	340
9	280

Σ=2925

Somma dei ricavi (Intensità totale del carattere) = 350 + 200 + 600 + 500 + 270 + 180 + 205 + 340 + 280 = 2925

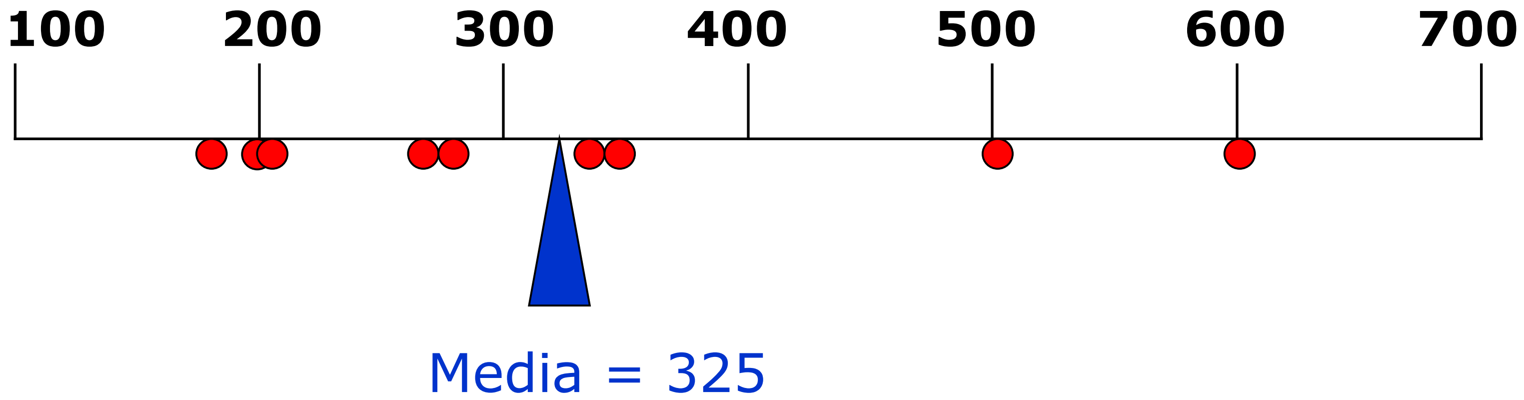
Media dei ricavi = 2925:9=325



L'intera torta rappresenta la somma dei ricavi di tutti i punti vendita

La singola fetta rappresenta la media dei ricavi

Formula della media

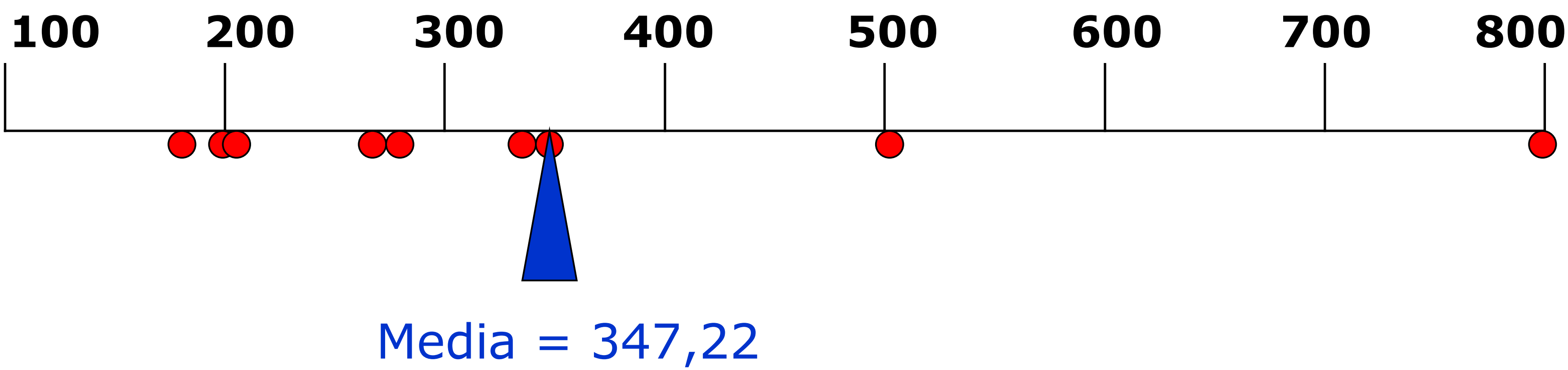


- Dati n valori osservati x_1, x_2,..., x_n di un carattere quantitativo X

$$\bar{x}_a = \frac{1}{n}(x_1 + x_2 + \ldots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Effetto dei valori estremi

Se il valore estremo fosse 800 invece di 600
➡ la media aumenterebbe
(il punto di equilibrio si sposta verso destra)



La media aritmetica risente fortemente dei valori estremi

Media di una distribuzione di frequenza

Addetti (valori x_j)	Numero punti vendita (frequenze n_j)	x_j*n_j
3	2	3*2=6
4	1	4*1=4
6	3	6*3=18
7	1	7*1=7
10	2	10*2=20

6 è il numero complessivo di addetti nei primi 2 punti vendita

18 è il numero complessivo di addetti nei 3 punti vendita in ciascuno dei quali lavorano 6 addetti

$$\sum_{j=1}^{K} n_j = n = 9 \qquad \sum_{j=1}^{K} x_j \cdot n_j = 55$$

55 è il numero complessivo di addetti (l'intensità totale del carattere)

Media $$\overline{x}_a = \frac{\sum_{j=1}^{K} x_j \cdot n_j}{\sum_{j=1}^{K} n_j} = \frac{\sum_{j=1}^{K} x_j \cdot n_j}{n} = \frac{55}{9} = 6{,}11$$

Media di una distribuzione di frequenza con classi di valori

Classi di superficie (in ettari)	Numero aziende (n_j)	Valore centrale classi (c_j)	c_j*n_j
0-1	120	0,5	60
1-2	160	1,5	240
2-3	220	2,5	550
3-5	212	4	848
5-10	205	7,5	1537,5
10-20	110	15	1650
20-40	65	30	1950
40-80	21	60	1260

$$\bar{x}_a \approx \frac{\sum_{j=1}^{K} c_j n_j}{n} = \frac{8095,5}{1113} = 7,27$$

La superficie media di una azienda agricola è di 7,27 ettari

$$n = \sum_{j=1}^{K} n_j = 1113 \qquad \sum_{j=1}^{K} c_j n_j = 8095,5$$

Fonte: Borra-Di Ciaccio, pag. 71

Media ponderata

- Uno studente ha sostenuto i seguenti esami del I anno del corso di laurea di EA.
- Come calcola la media dei voti?

N.	*Esame*	*voto*	*cfu*
1	Ec. Aziend./Rag. Gen I mod	27	9
2	Ist. diritto pubblico	22	6
3	Metodi di matematica applicata	25	9
4	Macroeconomia	20	6
5	Rag. Gen II mod /Bilancio e principi	28	9

Media ponderata: calcolo

N. Esame	*voto (x_i)*	*cfu (p_i)*	*voto*cfu (x_i*p_i)*
1	27	9	243
2	22	6	132
3	25	9	225
4	20	6	120
5	28	9	252

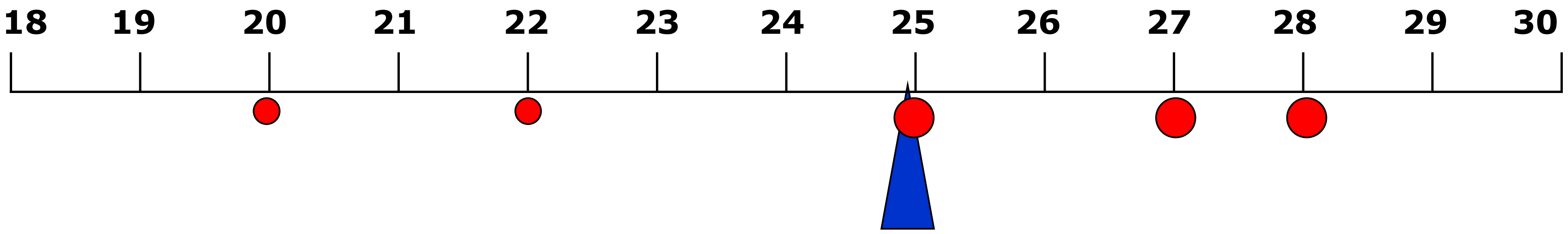
$$\sum_{i=1}^{n} p_i = 39 \qquad \sum_{i=1}^{n} x_i p_i = 972$$

$$\overline{x}_a = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i} = \frac{972}{39} = 24{,}92$$

Il voto medio (su 39 cfu) è pari a 24,92

Media ponderata

i due voti più bassi pesano di meno nel calcolo della media perché sono due esami da 6 cfu



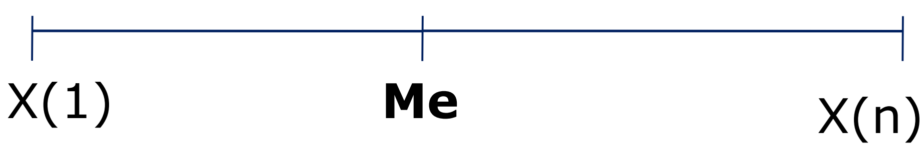
Media ponderata = 24,92

Mediana

- È il valore che occupa la posizione centrale nell'insieme ordinato di tutti i valori $x_{min} = x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} = x_{max}$

Tra x(1) e Me è contenuto il 50% dei valori

Tra Me e x(n) è contenuto il restante 50% dei valori



- È una media di posizione

Come individuare la posizione centrale o rango della mediana

- Insieme di n valori

- n dispari → la posizione centrale è data da $\frac{(n+1)}{2}$

$$Me = x_{\left(\frac{n+1}{2}\right)}$$

- n pari → le posizioni centrali sono due, $\frac{n}{2}$ e $\frac{n}{2}+1$

$$x_{\left(\frac{n}{2}\right)} \leq Me \leq x_{\left(\frac{n}{2}+1\right)}$$

Di solito

$$Me = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$$

La scelta tra media e mediana

FIGURA 30

«Prendiamo la media aritmetica dell'altezza e spaventiamo gli avversari o li vogliamo illudere prendendo il mediano?»

Fonte: Walter Kramer (2009), Le bugie della statistica, Nimesis

Moda

- È la modalità più frequente
- In un insieme di valori: quel termine che si ripete più volte
- In una distribuzione di frequenza: quella modalità che ha la frequenza più alta
- In una distribuzione di frequenza con classi di valori: ogni valore della classe con la più alta densità di frequenza

Moda di un insieme di valori

Punti vendita	Genere respons.
1	maschio
2	maschio
3	femmina
4	femmina
5	maschio
6	maschio
7	maschio
8	femmina
9	femmina

La modalità del carattere "Genere del responsabile" che si ripete più volte (5 volte) è "maschio"

Moda="maschio"

La maggioranza dei punti vendita ha come responsabile un uomo

Moda di una distribuzione di frequenza

Addetti (valori distinti)	Numero punti vendita (frequenze)
3	2
4	1
6	$\boxed{3}$
7	1
10	2

La frequenza maggiore è 3

La modalità del carattere "Numero di addetti" cui è associata la frequenza maggiore è 6

Moda=6

La maggioranza dei punti vendita ha un numero di addetti pari a 6

Moda

- Può non esistere
- Può non essere unica
- Può essere una modalità “poco rappresentativa” del fenomeno

- Per chi vende abbigliamento, la moda rappresenta un parametro utile per decidere in merito a come rifornire il negozio: saranno ordinati più capi delle taglie più diffuse

La scelta tra media, moda e mediana

Consideriamo ora le tre misure della tendenza centrale per calcolare lo stipendio medio di un gruppo di 41 dipendenti di una compagnia.

X = una persona

Numero dei dipendenti	Stipendio	
XX	€ 5000	
XXXXXX	€ 8000	
XXXXXXXX	€ 12 000	— — moda: quello riscontrato più frequentemente
XXXX	€ 20 000	
X	€ 25 000	— — mediana: quello a metà, con venti dipendenti sopra e venti sotto
XXXX	€ 32 000	
XXX	€ 37 000	
XXXXX	€ 40 000	
XX	€ 47 000	
XXXX	€ 55 000	
X	€ 260 000	
X	€ 450 000	

Il valore medio = **€ 42 000**
Il valore modale (con otto dipendenti) = **€ 12 000**
Il valore mediano = **€ 25 000**

Fonte: Magnello e Van Loon (2011), La statistica a fumetti, Raffaello Cortina Editore

Calcolo dei valori medi in base al tipo di carattere

		Caratteri	
	Quantitativi	Qualitativi ordinati	Qualitativi sconnessi
Media	✓		
Mediana	✓	✓	
Moda	✓	✓	✓

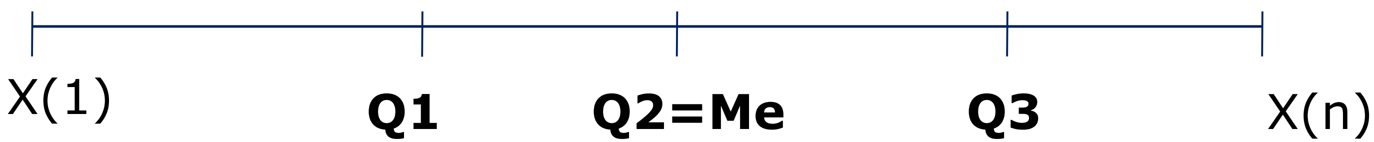
Quartili

- Sono 3 indici di posizione, Q_1 Q_2 e Q_3

$$x_{min} = x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)} = x_{max}$$

Tra $x_{(1)}$ e Q_1 è contenuto il 25% dei valori (più bassi)

Tra Q_3 e $x_{(n)}$ è contenuto il 25% dei valori (i più alti)



Tra Q_1 e Q_2 è contenuto il 25% dei valori

Tra Q_2 e Q_3 è contenuto il 25% dei valori

Primo quartile Q_1

- Q1 Primo quartile: è preceduto dal 25% dei termini (e seguito dal 75%)
- n dispari $Q_1 = x_{\left(\frac{n+1}{4}\right)}$
- n pari $Q_1 = \dfrac{x_{\left(\frac{n}{4}\right)} + x_{\left(\frac{n}{4}+1\right)}}{2}$

Se n (o n+1) non è divisibile per 4, il rango può non essere un numero intero

- In ogni caso, Q_1 è il primo valore x_i in corrispondenza del quale la frequenza cumulata relativa $F_j \geq 0{,}25$

Terzo quartile Q_3

- Q3 Terzo quartile: è preceduto dal 75% dei termini (e seguito dal 25%)
- n dispari $Q_3 = x_{\left(\frac{3}{4}(n+1)\right)}$
- n pari $Q_3 = \dfrac{x_{\left(\frac{3n}{4}\right)} + x_{\left(\frac{3n}{4}+1\right)}}{2}$

Se n (o n+1) non è divisibile per 4, il rango può non essere un numero intero

- In ogni caso, Q_3 è il primo valore x_i in corrispondenza del quale la frequenza cumulata relativa $F_j \geq 0{,}75$

Calcolo dei quartili

Ricavi	Ricavi (valori ordinati)	Freq. cum. rel.
350	$X_{(1)}=180$	1/9=0,11
200	$X_{(2)}=200$	2/9=0,22
600	$X_{(3)}=205$	3/9=0,33
500	$X_{(4)}=270$	4/9=0,44
270	$X_{(5)}=280$	5/9=0,56
180	$X_{(6)}=340$	6/9=0,67
205	$X_{(7)}=350$	7/9=0,78
340	$X_{(8)}=500$	8/9=0,89
280	$X_{(9)}=600$	9/9=1

La prima F_i ad essere maggiore o uguale a 0,25 è la terza

$$Q_1 = x_{(3)} = 205$$

Il 25% dei punti vendita con i ricavi più bassi registrano ricavi che non superano 205 mila euro

La prima F_i ad essere maggiore o uguale a 0,75 è la settima

$$Q_3 = x_{(7)} = 350$$

Per essere nel 25% dei punti vendita con i ricavi più alti si devono superare 350 mila euro di ricavi

Percentili

Sono quei valori che dividono la distribuzione in cento parti di uguale numerosità

Mediana=50-esimo percentile

Q3= 75-esimo percentile

P10 = decimo percentile: lascia alla sua sinistra il 10% dei valori

P90 = novantesimo percentile: lascia alla sua destra il 10% dei valori