Le misure di sintesi

1

Le misure di sintesi

- Le misure di sintesi descrivono un insieme di dati attraverso un valore (o modalità) rappresentativo del carattere.
- Alcune sono definiti solo per caratteri quantitativi (medie analitiche), altre per caratteri quantitativi e qualitativi (medie di posizione).

2

La media aritmetica

- La media aritmetica è la più diffusa misura di sintesi per caratteri quantitativi.
- Date *n* osservazioni, $x_1, x_2, \dots, x_n$, la media aritmetica è data da

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

- In forma compatta

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

3

- La media aritmetica è calcolata considerando tutti i dati!
- Se il dato più piccolo oppure il dato più grande sono anomali (estremi), tale misura di sintesi potrebbe rivelarsi poco opportuna.
- La media aritmetica è influenzata da valori anomali (misura di sintesi <u>non robusta</u>).
- Dato anomalo: *outlier*

4

- Es. 1: si calcoli la media aritmetica

  12 12 14 15 16 16 18 19 21 29

  n = 10

$$\bar{x} = \frac{12 + 12 + \cdots + 21 + 29}{10} = \frac{172}{10} = 17{,}2$$

5

- Es. 2: si calcoli la media aritmetica

  12 12 14 15 16 16 18 19 21 **290**

  n = 10

$$\bar{x} = \frac{12 + 12 + \cdots + 21 + 290}{10} = \frac{433}{10} = 43{,}3$$

6

La media aritmetica per una distribuzione di frequenze

- Per una distribuzione di frequenze con modalità $x_1, x_2, \dots, x_k$ e frequenze $n_1, n_2, \dots, n_k$, la media aritmetica è calcolata come

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{n}$$

- Perché? $\underbrace{x_1, \dots, x_1}_{n_1\ volte}, \underbrace{x_2, \dots, x_2}_{n_2\ volte}, \dots, \underbrace{x_k, \dots, x_k}_{n_k\ volte}$

7

La media aritmetica per una distribuzione di frequenze

- In forma compatta

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i n_i}{n}$$

8

- Esempio: numero dipendenti (Tabella 1, col. 2).
  Si calcoli la media aritmetica

x_i	n_i	$x_i n_i$
1	2	2
2	6	12
3	9	27
4	3	12
Totale	20	53

$$\bar{x} = \frac{53}{20} = 2{,}65$$

9

La media aritmetica per una distribuzione di frequenze in classi

- Per una distribuzione di frequenze in classi si considerano i valori centrali di ogni classe $c_1, c_2, \ldots, c_k$.
- La media aritmetica (approssimata) è

$$\bar{x} \approx \frac{c_1 n_1 + c_2 n_2 + \cdots + c_k n_k}{n}$$

- In sintesi

$$\bar{x} \approx \frac{\sum_{i=1}^{k} c_i n_i}{n}$$

10

- Esempio: fatturato 2012 (Tabella 1, col. 4)
  Si calcoli la media aritmetica

Classe	n_i	c_i	$c_i n_i$
2 – 3	3	2,5	7,5
3 – 4	5	3,5	17,5
4 – 5	8	4,5	36
5 – 6	4	5,5	22
Totale	*20*		*83*

11

- In formula

$$\bar{x} \approx \frac{83}{20} = 4{,}15$$

- La media aritmetica così calcolata coinciderebbe con la media aritmetica esatta se ogni valore centrale fosse uguale alla media aritmetica dei valori della classe.

12

La media aritmetica ponderata

- *n* osservazioni, $x_1, x_2, \dots, x_n$
- *n* corrispondenti pesi, $p_1, p_2, \dots, p_n$
- La media aritmetica ponderata (o pesata) è

$$\bar{x} = \frac{x_1 p_1 + x_2 p_2 + \cdots + x_n p_n}{p_1 + p_2 + \cdots + p_n}$$

- In forma compatta

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i p_i}{\sum_{i=1}^{n} p_i}$$

13

- Viene utilizzata quando si vuole attribuire una diversa importanza (peso) ai dati.
- La media dei voti universitari è una media ponderata con pesi pari ai CFU.
- Esempio: lo studente De Luca ha i seguenti voti (crediti):

  26 (9) 28 (12) 30 (12)

$$\bar{x} = \frac{x_1 p_1 + x_2 p_2 + x_3 p_3}{p_1 + p_2 + p_3}$$

$$\bar{x} = \frac{26 \cdot 9 + 28 \cdot 12 + 30 \cdot 12}{9 + 12 + 12} = \frac{930}{33} = 28{,}18$$

14

- La media calcolata per distribuzioni di frequenze è un esempio di media aritmetica ponderata con $p_i = n_i$.
- Infatti

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{n}$$

e poiché

$$n = n_1 + n_2 + \cdots + n_k$$

$$\bar{x} = \frac{x_1 n_1 + x_2 n_2 + \cdots + x_k n_k}{n_1 + n_2 + \cdots + n_k}$$

15

Alcune proprietà della media aritmetica

- Si definiscono 4 proprietà:

1. La somma dei valori $x_1, x_2, \ldots, x_n$ è pari a *n* volte la media aritmetica

$$\sum_{i=1}^{n} x_i = n \cdot \bar{x}$$

16

2. La somma delle differenze (scarti) dei valori dalla media è pari a 0

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

17

3. Se un insieme di *n* dati $x_1, x_2, \dots, x_n$ è suddiviso in *L* sottoinsiemi di numerosità $n_1, n_2, \dots, n_L$ tali che

$$n_1 + n_2 + \cdots + n_L = n$$

e con medie $\bar{x}_{(1)}, \bar{x}_{(2)}, \dots, \bar{x}_{(L)}$ allora la media complessiva è

$$\bar{x} = \frac{\bar{x}_{(1)}n_1 + \bar{x}_{(2)}n_2 + \cdots + \bar{x}_{(L)}n_L}{n_1 + n_2 + \cdots + n_L}$$

(la media complessiva è la media ponderata delle medie dei sottoinsiemi con pesi pari alle numerosità degli stessi)

18

- Lo stipendio medio mensile nell'Italia del Nord, calcolato su 22,8 milioni di persone, è pari a 1784, mentre nell'Italia centrale, calcolato su 5,1 milioni di persone, è 1004.
- Calcolare lo stipendio medio dell'Italia Centro-Nord.

$$\bar{x}_{(1)} = 1784, n_1 = 22{,}8$$
$$\bar{x}_{(2)} = 1004, n_2 = 5{,}1$$

$$\bar{x} = \frac{1784 \cdot 22{,}8 + 1004 \cdot 5{,}1}{22{,}8 + 5{,}1} = \frac{45795{,}6}{27{,}9} = 1641{,}42$$

19

4. Data una variabile *X* e considerata una trasformazione lineare

$$Y = a + bX$$

allora la media della variabile *Y* è pari a

$$\bar{Y} = a + b\bar{X}$$

20

- Lo stipendio annuale del sig. Smith è pari a 6000 euro più il 10% del fatturato aziendale X. Se il fatturato medio è pari a 300000, qual è lo stipendio medio del sig. Smith?
- Lo stipendio del sig. Smith è

$$Y = 6000 + 0{,}10 \cdot X$$

- Quindi

$$\bar{Y} = 6000 + 0{,}10 \cdot \bar{X}$$

$$\bar{Y} = 6000 + 0.10 \cdot 300000 = 36000$$

La trimmed mean

- Consiste in una media aritmetica calcolata su una percentuale dei valori centrali.
- La trimmed mean all'α% è calcolata sull'α% dei valori centrali.
- È indicata con $\bar{x}_{TR}(\alpha\%)$
- È una media non influenzata da valori anomali (misura di sintesi robusta).

- Es. 1: si calcoli la trimmed mean all’80%
- ~~12~~ 12 14 15 16 16 18 19 21 ~~29~~
- *n* = 10
- Si considera l’80% dei valori centrali (0,80·10 = 8).
- Si escludono i 2 valori più estremi (il minore e il maggiore).
- Quindi

$$\bar{x}_{TR}(80\%) = \frac{12 + 14 + \cdots + 19 + 21}{8} = \frac{131}{8} = 16{,}375$$

23

- Es. 2: si calcoli la trimmed mean all’80%
- ~~12~~ 12 14 15 16 16 18 19 21 **~~290~~**
- *n* = 10
- Si considera l’80% dei valori centrali (0,80·10 = 8).
- Si escludono i 2 valori più estremi (il minore e il maggiore).
- Quindi

$$\bar{x}_{TR}(80\%) = \frac{12 + 14 + \cdots + 19 + 21}{8} = \frac{131}{8} = 16{,}375$$

24

- Es. 3: si calcoli la trimmed mean al 90%

Numero dipendenti	*Frequenze (assolute)*
1	2
2	6
3	9
4	3
Totale	20

25

- n = 20
- Si considera il 90% dei valori centrali ($0{,}90 \cdot 20 = 18$).
- Si escludono i 2 valori più estremi.
- Il minore è 1, il maggiore è 4.

26

- La distribuzione di frequenze da considerare diviene

Numero dipendenti	*Frequenze (assolute)*
1	~~2~~ **1**
2	6
3	9
4	~~3~~ **2**
Totale	~~20~~ **18**

$$\bar{x}_{TR}(90\%) = \frac{1 \cdot 1 + 2 \cdot 6 + 3 \cdot 9 + 4 \cdot 2}{18} = \frac{48}{18} = 2{,}67$$

27

La media geometrica

- Dati *n* valori positivi, $x_1, x_2, \dots, x_n$, la media geometrica è data da

$$\bar{x}_g = \sqrt[n]{x_1 \cdot x_2 \cdot \cdots \cdot x_n}$$

28

- Il suo principale utilizzo riguarda tassi di variazione percentuali (TVP), osservati nel tempo.
- Obiettivo: calcolare il tasso medio di variazione percentuale (TMVP) attraverso la media geometrica dei coefficienti di incremento (CI).
- Il coefficiente di incremento è dato da

$$CI = 1 + \frac{TVP}{100}$$

29

Anno	*Tassi di variazione (%)*	*Coefficienti di incremento*
1	8%	1,08
2	10%	1,10
3	12%	1,12
4	14%	1,14

30

Si ottiene

$$\bar{x}_g = \sqrt[4]{1{,}08 \cdot 1{,}10 \cdot 1{,}12 \cdot 1{,}14}$$

$$\bar{x}_g = \sqrt[4]{1{,}5168}$$

$$\bar{x}_g = 1{,}1098$$

(media geometrica dei tassi di incremento)

31

E il tasso medio di variazione?

Poiché

$$CI = 1 + \frac{TVP}{100}$$

allora

$$\bar{x}_g = 1 + \frac{TMVP}{100}$$

$$TMVP = 100(\bar{x}_g - 1)$$

32

Nell'esempio

$$TMVP = 100(1{,}1098 - 1)$$

Il tasso medio di variazione (%) è

$$TMVP = 10{,}98$$

33

Proprietà della media geometrica

- Il logaritmo naturale della media geometrica è pari alla media aritmetica semplice dei logaritmi dei valori.
- Infatti, applicando il logaritmo naturale si ottiene

$$\log \bar{x}_g = \log \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$$

$$\log \bar{x}_g = \log(x_1 \cdot x_2 \cdot \dots \cdot x_n)^{1/n}$$

$$\log \bar{x}_g = \frac{1}{n} \log(x_1 \cdot x_2 \cdot \dots \cdot x_n)$$

$$\log \bar{x}_g = \frac{1}{n} (\log(x_1) + \log(x_2) + \dots + \log(x_n))$$

34

La mediana

- La mediana è il valore (o modalità) centrale di un insieme di dati ordinati in senso crescente. Può essere individuata per caratteri quantitativi e caratteri qualitativi ordinabili.
- La mediana, indicata con *Me*, è preceduta dal 50% dei dati più piccoli ed è seguita dal 50% dei dati più grandi.
- Non è influenzata da valori anomali (misura di sintesi robusta).
- Distinzione tra

1. numero di osservazioni *n* dispari
2. numero di osservazioni *n* pari

35

1. *n* dispari: ordinati i dati in senso crescente, la mediana è il valore in posizione *(n+1)/2.*

   Es. 1 (*n*=5)

   23 25 27 42 22

   Ordinati in senso crescente

   22 23 25 27 42

   Il valore centrale è 25, in posizione (5+1)/2 = 3

   Dunque *Me* = 25

36

2. *n* pari: ordinati i dati in senso crescente, è la media aritmetica dei valori in posizione *n/2* e *n/2+1.*

Es. 2 (*n*=6)

23 25 27 42 22 18

Ordinati in senso crescente

18 22 23 25 27 42

I valori centrali sono 23 e 25, in posizione 6/2=3 e 6/2+1=4

Dunque

$$Me = \frac{23 + 25}{2} = 24$$

37

Mediana per distribuzione di frequenze

Numero dipendenti	*Frequenze (assolute)*
1	2
2	6
3	9
4	3
Totale	20

- n = 20 (pari)
- Mediana = media aritmetica dei valori in posizione *n/2* e *n/2+1.*

38

- Valore in posizione 20/2 = 10? E' il valore 3
- Valore in posizione 20/2 +1 =11? E' il valore 3

Numero dipendenti	*Frequenze (assolute)*	*Frequenze cumulate*
1	2	2
2	6	8
3	9	17
4	3	20
Totale	20	

$$Me = \frac{3+3}{2} = 3$$

39

Mediana per distribuzione di frequenze in classi

Classi	*Frequenze (assolute)*
2 – 3	3
3 – 4	5
4 – 5	8
5 – 6	4
Totale	*20*

1. Individuazione della classe mediana
2. Individuazione della mediana all'interno della classe mediana (metodo analitico e grafico).

40

1. Individuazione della classe mediana

- Si considerano le frequenze relative cumulate

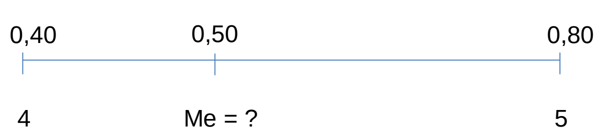
Classi	*Frequenze (assolute)*	*Frequenze relative*	*Frequenze relative cumulate*
2 – 3	3	0,15	0,15
3 – 4	5	0,25	0,40
4 – 5	8	0,40	0,80
5 – 6	4	0,20	1
Totale	*20*		

- La classe mediana è la classe 4-5, poiché 0,40 < **0,50** < 0,80.

41

2. Individuazione della mediana (metodo analitico)

- Il metodo analitico si basa su una proporzione.



$$(5-4):(0{,}80-0{,}40)=(Me-4):(0{,}50-0{,}40)$$

42

2. Individuazione della mediana (metodo analitico)

$$(5-4):(0{,}80-0{,}40)=(Me-4):(0{,}50-0{,}40)$$

$$(Me-4)=\frac{(5-4)(0{,}50-0{,}40)}{(0{,}80-0{,}40)}$$

$$Me=4+\frac{0{,}10}{0{,}40}$$

$$Me=4{,}25$$

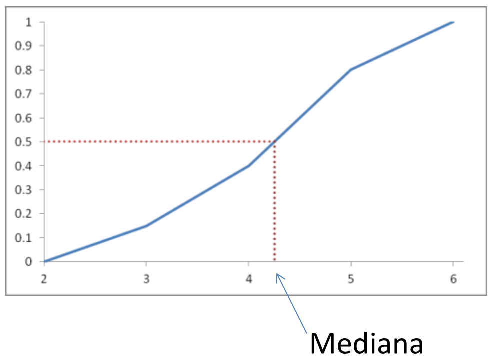
43

2. Individuazione della mediana (metodo grafico)

- Il <u>metodo grafico</u> si basa sul poligono cumulativo (con frequenze relative cumulate).

44

Poligono cumulativo con frequenze relative cumulate



45

Mediana per caratteri qualitativi ordinabili
Giudizi sulla politica sanitaria di 203 individui

Giudizio	n_i	N_i
Scarso	13	13
Sufficiente	48	61
Buono	77	138
Ottimo	65	203
Totale	203	

46

- Poiché n = 203 è dispari, la mediana è la modalità centrale, ovvero la modalità in posizione (n+1)/2 = 102
- Il giudizio mediano è «Buono».

Robustezza della mediana

- Es. 1 (n=10):
- 12 12 14 15 $\boxed{16\ \ 16}$ 18 19 21 29
- La mediana è (16+16)/2 = 16
- *Me* = 16
- Es. 2 (n=10):
- 12 12 14 15 $\boxed{16\ \ 16}$ 18 19 21 **290**
- *Me* = 16
- La mediana non è influenzata da 290 (valore anomalo).

I quartili

- Si definiscono 3 quartili.
- Il primo quartile (Q_1) è quel valore preceduto dal 25% (1/4) dei dati e seguito dal 75% (3/4).
- Il secondo quartile è quel valore preceduto dal 50% dei dati e seguito dal 50% (è proprio la mediana).
- Il terzo quartile (Q_3) è quel valore preceduto dal 75% dei dati e seguito dal 25%.

49

Individuazione di Q_1

- Per l'individuazione di Q_1 si guarda alla frequenza relativa cumulata immediatamente maggiore di 0,25.

50

Individuazione di Q_1

x_i	n_i	N_i	F_i
1	2	0,10	0,10
2	6	0,30	0,40
3	9	0,45	0,85
4	3	0,15	1
Totale	20	1	

- La F_i di riferimento è 0,40
- $Q_1 = 2$

51

Individuazione di Q_1

- Se risulta una frequenza relativa cumulata uguale a 0,25 allora Q_1 è la media aritmetica del valore corrispondente e del successivo valore.

52

Individuazione di Q_1

x_i	n_i	f_i	F_i
10	2	0,10	0,10
20	3	0,15	0,25
30	10	0,50	0,75
40	5	0,25	1
Totale	20	1	

- Q_1 = (20 + 30) / 2 = 25

53

Individuazione di Q_3

- Per l'individuazione di Q_3 si guarda alla frequenza relativa cumulata immediatamente maggiore di 0,75.

54

Individuazione di Q_3

x_i	n_i	f_i	F_i
1	2	0,10	0,10
2	6	0,30	0,40
3	9	0,45	0,85
4	3	0,15	1
Totale	20	1	

- La F_i di riferimento è 0,85
- Q_3 = 3

55

Individuazione di Q_3

- Se risulta una frequenza relativa cumulata uguale a 0,75 allora Q_3 è la media aritmetica del valore corrispondente e del successivo valore.

56

Individuazione di Q_3

x_i	n_i	f_i	F_i
10	2	0,10	0,10
20	3	0,15	0,25
30	10	0,50	0,75
40	5	0,25	1
Totale	20	1	

- $Q_3 = (30 + 40) / 2 = 35$

57

Quartili per distribuzioni di frequenze in classi

1. Individuazione della classe che contiene Q_1 (Q_3).
2. Individuazione di Q_1 (Q_3) nell'ambito della classe di cui al punto 1.

58

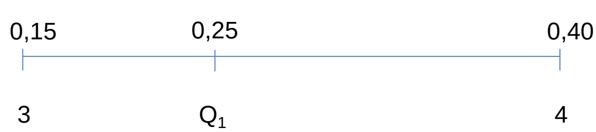
Classi	*Frequenze (assolute)*	*Frequenze relative*	*Frequenze relative cumulate*
2 – 3	3	0,15	0,15
3 – 4	5	0,25	0,40
4 – 5	8	0,40	0,80
5 – 6	4	0,20	1
Totale	*20*		

- La classe che contiene Q_1 è la classe 3-4, poiché 0,40 > **0,25**.

59

Metodo analitico

Proporzione



$$(4-3):(0{,}40-0{,}15)=(Q_1-3):(0{,}25-0{,}15)$$

60

$$(4-3):(0{,}40-0{,}15)=(Q_1-3):(0{,}25-0{,}15)$$

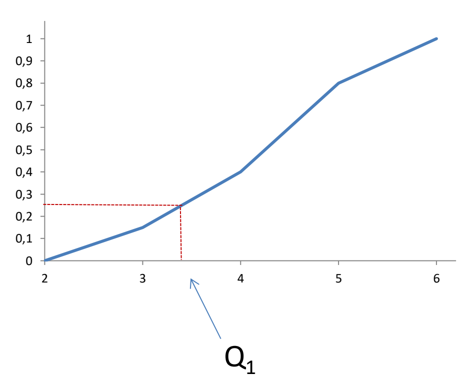
$$(Q_1-3)=\frac{(4-3)(0{,}25-0{,}15)}{(0{,}40-0{,}15)}$$

$$Q_1 \ =3+\frac{0{,}10}{0{,}25}$$

$$Q_1 \ =3{,}4$$

61

Metodo grafico



62

I percentili

- I percentili fanno riferimento ad una suddivisione dei dati in cento parti.
- Es. 1: il 34° percentile (P_{34}) è quel valore che è preceduto dal 34% dei dati ed è seguito dal 66%.
- Es. 2: il 90° percentile (P_{90}) è quel valore che è preceduto dal 90% dei dati ed è seguito dal 10%.
- Da notare che
- $P_{25} = Q_1$
- $P_{50} = Me$
- $P_{75} = Q_3$

63

Calcolo del 90° percentile (P_{90})

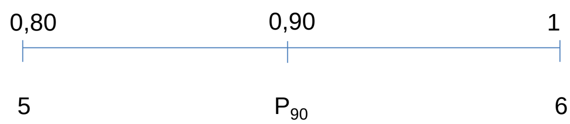
Classi	n_i	f_i	F_i
2 – 3	3	0,15	0,15
3 – 4	5	0,25	0,40
4 – 5	8	0,40	0,80
5 – 6	4	0,20	1
Totale	*20*	*1*	

- La classe che contiene P_{90} è la classe 5-6, poiché 1 < **0,90**.

64

Metodo analitico

Proporzione



$$(6-5):(1-0{,}80) = (P_{90}-5):(0{,}90-0{,}80)$$

65

$$(6-5):(1-0{,}80) = (P_{90}-5):(0{,}90-0{,}80)$$

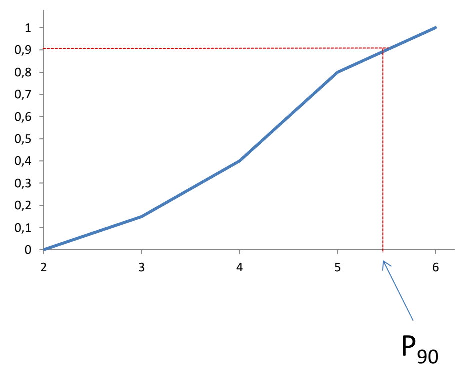
$$(P_{90}-5) = \frac{(6-5)(0{,}90-0{,}80)}{(1-0{,}80)}$$

$$P_{90} = 5 + \frac{0{,}10}{0{,}20}$$

$$P_{90} = 5{,}5$$

66

Metodo grafico



67

La moda

- La moda è la modalità che si presenta con la massima frequenza (il maggior numero di volte).
- Può essere individuata per variabili e mutabili.

68

x_i	n_i
1	2
2	6
3	9
4	3
Totale	20

- La moda è 3, poiché il valore 3 viene osservato 9 volte
- *Mo* = 3

69

Moda per distribuzioni di frequenze in classi

Classi	n_i	h_i
2 – 3	4	4
3 – 5	13	6,5
5 – 10	1	0,2
10 – 20	2	0,2
Totale	*20*	

- La classe modale è la classe 3-5, poiché ad essa è associata la maggiore densità di frequenza.

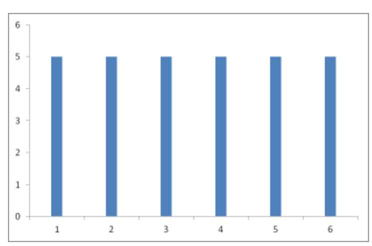
70

- Distribuzione
  - zeromodale: nessuna moda
  - unimodale: una sola moda
  - bimodale: due mode
  - plurimodale: più mode

71

Esempio di distribuzione zeromodale

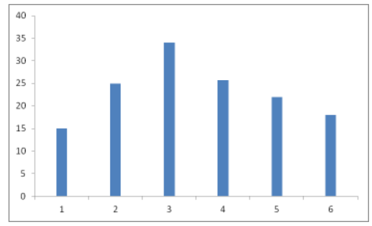
x_i	n_i
1	5
2	5
3	5
4	5
5	5
6	5
Totale	*30*



72

Esempio di distribuzione unimodale

x_i	n_i
1	15
2	25
3	34
4	28
5	26
6	12
Totale	*140*

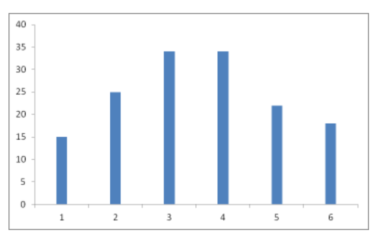


$Mo = 3$

73

Esempio di distribuzione bimodale

x_i	n_i
1	15
2	25
3	34
4	34
5	26
6	12
Totale	*146*



$Mo = 3$
$Mo = 4$

74

- Da notare che per distribuzioni di frequenze con classi di uguale ampiezza la classe con maggiore densità di frequenza è la classe con maggiore frequenza assoluta.

75