Statistica Descrittiva II

Serie statistiche monovariate

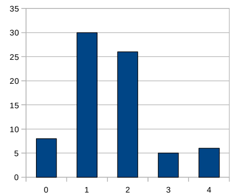
- Indici di posizione
- Indici di variabilità
- Indici di asimmetria
- Indici di normalità
- Outlier
- Box-plot

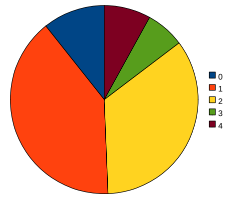
Organizzazione dei dati

Carattere: Cellulari posseduti

Popolazione: 75 studenti di Biotecnologie

i	m_i	n_i
1	0	8
2	1	30
3	2	26
4	3	5
5	4	6
	Totali	75





- Rappresentazioni complete ma “ingombranti”
- Utile un indice più “maneggevole” → indice sintetico

Indici di posizione: Media

Indicano in maniera sintetica il “centro” della statistica

- Media: calcolo il valore medio delle osservazioni
  - Calcolo standard $\bar{m}=\frac{1}{N}\sum_{i=1}^{N} o_i$

$$\bar{m}=\frac{1}{75}(0+0+0+0+0+0+0+0+1+1+1+....)=1{,}61$$

  - Calcolo in tabella (I)
    sfrutto le frequenze assolute

$$\bar{m}=\frac{1}{N}\sum_{i=1}^{N} o_i=\frac{1}{N}\sum_{i=1}^{M} n_i * m_i$$

$$\bar{m}=\frac{121}{75}=1{,}61$$

m_i	n_i	$n_i m_i$
0	8	0
1	30	30
2	26	52
3	5	15
4	6	24
Totali	75	121

Indici di posizione: Media

- Calcolo in Tabella (II)
  - Sfrutto le frequenze relative

$$\bar{m}=\frac{1}{N}\sum_{i=1}^{5} m_i * n_i$$

$$\bar{m}=\frac{m_1 n_1+m_2 n_2+m_3 n_3+m_4 n_4+m_5 n_5}{N}$$

$$\bar{m}=\frac{m_1 n_1}{N}+\frac{m_2 n_2}{N}+\frac{m_3 n_3}{N}+\frac{m_4 n_4}{N}+\frac{m_5 n_5}{N}$$

$$\bar{m}=m_1 f_1+m_2 f_2+m_3 f_3+m_4 f_4+m_5 f_5$$

  - In generale

$$\bar{m}=\sum_{i=1}^{M} m_i * f_i$$

m_i	n_i	f_i	$m_i f_i$
0	8	0,107	0
1	30	0,400	0,4
2	26	0,347	0,69
3	5	0,067	0,2
4	6	0,080	0,32
Totali	75	1	1,61

Media: principali proprietà

Date $O=\{ o_i \}$ osservazioni

- La media è compresa fra le osservazioni minima e massima

$$min(o_i) \le \bar{o} \le max(o_i)$$

- La somma degli scarti dalla media è nulla

$$\sum_{i=1}^{N} (o_i - \bar{o}) = 0$$

- La somma del quadrato degli scarti dalla media è minima

$$\sum_{i=1}^{N} (o_i - \bar{o})^2 \le \sum_{i=1}^{N} (o_i - p)^2 \forall\, p$$

Media: altri tipi di dati

- La media si basa sulla somma delle osservazioni

  Non può essere calcolata per dati qualitativi
- Caratteri continui.
  - Osservazioni

    Nessun problema
  - Dati elaborati (raccolta fatta da terzi)
    - Istogramma
    - Tabella organizzata per classi

      Manca la modalità!

Media: Raccolta dati per classi

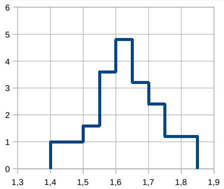
Idea: uso come modalità il valore centrale (medio) della classe

- Tabella ad entrata semplice
  1. Ad ogni c_i associo il valore centrale $\bar{c}_i$
  2. Calcolo usuale con $\bar{c}_i$ al posto di m_i

inf_i	sup_i	$\bar{c}_i$	f_i	$\bar{c}_i\, f_i$
1,40	1,50	1,450	0,10	0,1450
1,50	1,55	1,525	0,08	0,1220
1,55	1,60	1,575	0,18	0,2835
1,60	1,65	1,625	0,24	0,3900
1,65	1,70	1,675	0,16	0,2680
1,70	1,75	1,725	0,12	0,2070
1,75	1,85	1,800	0,12	0,2160
		totale	1	1,6315

- Istogramma:
  1. Da ogni classe ricavo $\bar{c}_i$
  2. Converto densità in frequenza

     $$f_i = (sup_i - inf_i) d_i$$
  3. Calcolo usuale $\bar{o} = \sum_{i=1}^{M} \bar{c}_i\, f_i$



Media: Outlier

- Outlier: valore poco plausibile o errato
  - Errori di misura
  - Dato volutamente poco plausibile
- Come si modifica la media ?
  - *1* outlilier (u) *N-1* osservazioni "valide" (o_i)

    $$\bar{m} = \frac{1}{N} \sum_{i=1}^{N-1} o_i + \frac{u}{N}$$
  - *U* outliers (u_j) *N-U* osservazioni "valide" (o_i) (Ovviamente N >> U)

    $$\bar{m} = \frac{1}{N} \sum_{i=1}^{N-U} o_i + \frac{1}{N} \sum_{j=1}^{U} u_j$$

Media: considerazioni

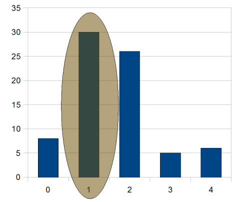
- Può essere calcolata solo per caratteri numerici.
- Possibile calcolarla anche per rilevazioni a classi
- Sensibile agli outliers
  - Peggiora al crescere del # di outlier (*U*)
  - Migliora al crescere di *N*
- La media non è un osservazione
  - (solo in rarissimi casi lo è)

Indici di posizione: Moda

Indicano in maniera sintetica il “centro” della statistica

- Moda: scelgo il valore più presente fra le osservazioni
  - Frequenza (relativa o assoluta) maggiore
  - Barra “più alta” in un diagramma a barre

i	m_i	n_i
1	0	8
2	1	30
3	2	26
4	3	5
5	4	6
	Totali	75



Moda: altri tipi di dati

- La moda si basa sulle frequenze assolute

  Può essere calcolata anche per i valori qualitativi
- Caratteri continui.
  - Osservazioni

  osservazioni uguali sono rare → raggruppo in classi
  - Dati elaborati (raccolta fatta da terzi)
    - Istogramma
    - Tabella organizzata per classi

Classe modale = classe con la frequenza maggiore

Moda - Considerazioni

- Può essere calcolata per tutti i dati
- La moda è un osservazione
- Poco sensibile agli outliers

  Difficilmente gli outliers hanno alte frequenze
- Non è unica (esempio a lato):
  - Due massimi → serie bimodale
  - Tre massimi → serie trimodale
  - ….

i	m_i	n_i
1	A	8
2	B	30
3	C	26
4	D	30
5	E	6
	Totali	100

Indici di posizione: Mediana

Mediana: osservazione che bipartisce le osservazioni

- Procedimento di calcolo (esempio cellulari N = 75)
    1. Ordino le osservazioni
    2. Calcolo la dimensione delle “metà” $\frac{N-1}{2}=37$
    3. Leggo l'osservazione successiva (38).

37 valori

0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 3 3 3 3 3 4 4 4 4 4 4

37 valori

Indici di posizione: Mediana II

Se il numero di osservazioni è pari?

Esempio N = 12

- Procedimento di calcolo
    1. Ordino le osservazioni
    2. Calcolo la dimensione delle “metà” $\frac{12-1}{2}=5{,}5$
    3. Se intera leggo il dato successivo
       altrimenti medio le due osservazioni limite (6 e 7)

1 2 3 4 5 6 7 8 9 10 11 12

$$\frac{6+7}{2}=6{,}5$$

Mediana: altri tipi di dati

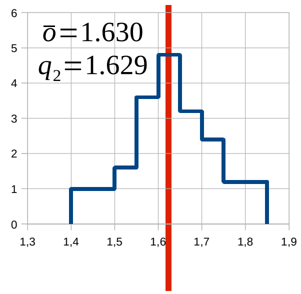
- La mediana si basa sulle osservazioni ordinate

  Non può essere calcolata per dati non ordinabili
- Caratteri continui.
    - Osservazioni

      Nessun problema
    - Dati elaborati
        - Istogramma
        - Tabella organizzata per classi

Mediana: Raccolta dati per classi

- Area Istogramma = *1;*
- Area alla sx q_2 = 0.5
- Metodo di calcolo



1. Trovo la classe $\bar{i}$ in cui cade q_2
   (F_1= 0.10; F_2= 0.18; F_3= 0.36; F_4= 0.60).
2. Calcolo l'area residua.

$$0.5-F_{\bar{i}-1}=0.14$$

3. Calcolo la base *b* del rettangolo “parziale”

$$b\frac{f_{\bar{i}}}{sup_{\bar{i}}-inf_{\bar{i}}}=0.14\rightarrow b\,4.8=0.14\rightarrow b=0{,}029$$

4. Trovo la mediana

$$q_2=inf_{\bar{i}}+b=1.629$$

Mediana: Raccolta dati per classi

- Tabella ad entrata semplice:
- Cerco la modalità $F(q_2) = 0.5$;
- *Esempio di calcolo*
  1. Trovo la classe $\bar{i}$ in cui cade q_2 (4).
  2. Trovo a quanto di f_i mi serve.

$$\frac{0.5-F_{\bar{i}-1}}{f_{\bar{i}}}=\frac{0.5-0.36}{0.24}=\frac{0.14}{0.24}=\frac{7}{12}$$

  3. Applico la stessa proporzione all'intervallo

$$(sup_{\bar{i}}-inf_{\bar{i}})\frac{0.5-F_{\bar{i}-1}}{f_{\bar{i}}}=(1.65-1.60)\frac{7}{12}=0.029$$

  4. Calcolo q2

$$inf_{\bar{i}}+(sup_{\bar{i}}-inf_{\bar{i}})\frac{0.5-F_{\bar{i}-1}}{f_{\bar{i}}}=1.629$$

inf_i	sup_i	n_i	f_i	F_i
1,40	1,50	5	0,10	0,10
1,50	1,55	4	0,08	0,18
1,55	1,60	9	0,18	0,36
1,60	1,65	12	0,24	0,60
1,65	1,70	8	0,16	0,76
1,70	1,75	6	0,12	0,88
1,75	1,85	6	0,12	1
	tot	50	1	

Mediana - Considerazioni

- Molto spesso è un osservazione:
  - Sempre per *N* dispari.
  - Spesso per *N* pari (media di osservazioni uguali).
- Può essere calcolata per tutti i dati tranne i non ordinabili
- Dati in classi:
  - Istogramma: Bipartizione dell'area sottesa.
  - Tabella entrata semplice: $F(q_2) = 0.5$.
- Poco sensibile agli outliers
  - spostano la mediana di *U* posizioni
  - Il valore numerico degli outliers è ininfluente.

Indici di Posizione: Quartili

Invece di bipartire le osservazioni le dividono in 4 parti

- Sono 3
  - q_1 preceduto da (alla "sua sx" ha) ¼ dei dati
  - q_2 mediana
  - q_3 seguito da (alla "sua dx" ha) ¼ dei dati
- Alcuni autori aggiungono:
  - q_0 minimo valore delle osservazioni
  - q_4 massimo valore delle osservazioni
- Metodo di calcolo non è unico (come per la mediana)

Indici di Posizione: Quartili - II

Procedimento di calcolo q_1 (e q_3)

- Ordino le osservazioni
- Calcolo la dimensione dei "quarti" $\frac{N-1}{4}=18{,}5$
- Se intera leggo l'osservazione successiva altrimenti medio le due osservazioni limite

← 19 valori $\frac{1+1}{2}=1$ 56 valori →

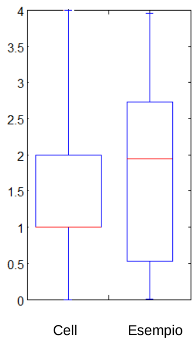
37 valori →

0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2
2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 ← 37 valori
2 2 2 2 3 3 3 3 3 4 4 4 4 4 4

← 56 valori $\frac{2+2}{2}=2$ 19 valori →

Box-plot (prima versione)

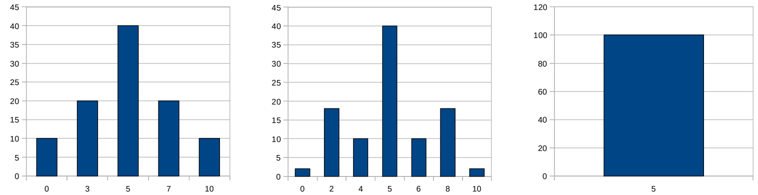
- Rappresentazione grafica legata ai quartili
- Diverse versioni
- Versione base
  - Rettangolo fra q_1 e q_3
  - Mediana (q_2) evidenziata (rosso)
  - Due "baffi" (segmento)
    - da centro del lato dx e q_4
    - da centro del lato sx e q_0



- Alcuni quartili possono essere sovrapposti (Cell)

Indici di variabilità.

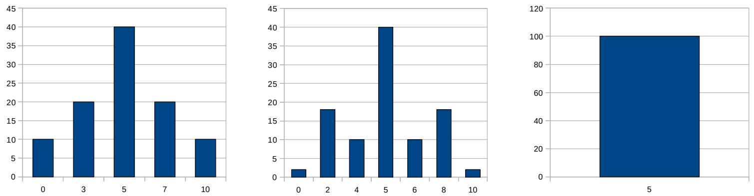
- Esempi di popolazioni in cui moda = media = mediana = 5



- Statistiche molto diverse fra loro.
- Si introduce il concetto di variabilità.
  "Propensione" delle osservazioni ad allontanarsi dal loro centro.
- Osservazione: Serve una "distanza" per poter valutare la variabilità. → No caratteri non ordinabili.

Campo di variazione o range

- range:differenza fra la massima e la minima osservazione



range = 10 - 0 = 10 range = 10 range = 0

- Considerazioni:
  - Facilissimo da calcolare
  - Istogrammi: I valori estremi del grafico
  - Molto sensibile alla presenza di outliers.
    (gli outliers per definizione son valori estremi!)

Varianza della popolazione σ^2

Media degli scarti dalla media al quadrato.

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{N}(o_i-\bar{m})^2$$

- Esempio: osservazioni

O = {2 5 6 7}

  - Media = (2 + 5 + 6 +7)/4 = 5
  - Scarti dalla media = { -3 0 1 2}
  - Scarti dalla media al quadrato = {9 0 1 4 }
  - Media degli scarti dalla media al quadrato
    Varianza = (9 + 0 + 1 +4)/4 = 3,5

Varianza: formule di calcolo

- Definizione

$$\sigma^2 = \frac{\sum_{i=1}^{M} n_i (m_i - \bar{m})^2}{N}$$

$$\sigma^2 = \sum_{i=1}^{M} f_i (m_i - \bar{m})^2$$

m_i	n_i	f_i	$m_i f_i$	$m_i - \bar{m}$	$(m_i - \bar{m})^2$	$f_i(m_i - \bar{m})^2$
0	10	0,1	0	-5	25	2,5
3	20	0,2	0,6	-2	4	0,8
5	40	0,4	2	0	0	0
7	20	0,2	1,4	2	4	0,8
10	10	0,1	1	5	25	2,5
totali	100		$\bar{m}$ =5			6,6

- Formula "breve"

$$\sigma^2 = \sum_{i=1}^{M} f_i (m_i - \bar{m})^2 = \sum_{i=1}^{M} f_i (m_i^2 - 2 m_i \bar{m} + \bar{m}^2)$$

$$\sigma^2 = \sum_{i=1}^{M} f_i m_i^2 - 2 \sum_{i=1}^{M} f_i m_i \bar{m} + \sum_{i=1}^{M} f_i \bar{m}^2$$

$$\sigma^2 = \sum_{i_1}^{M} f_i m_i^2 - 2 \bar{m} \sum_{i=1}^{M} f_i m_i + \bar{m}^2 \sum_{i=1}^{M} f_i$$

$$\sigma^2 = \sum_{i_1}^{M} (f_i m_i^2) - 2 \bar{m} \bar{m} + \bar{m}^2 1 = \sum_{i_1}^{M} (f_i m_i^2) - \bar{m}^2$$

Varianza: formule di calcolo

- Definizione

$$\sigma^2 = \frac{\sum_{i=1}^{M} n_i (m_i - \bar{m})^2}{N}$$

$$\sigma^2 = \sum_{i=1}^{M} f_i (m_i - \bar{m})^2$$

m_i	n_i	f_i	$m_i f_i$	$m_i - \bar{m}$	$(m_i - \bar{m})^2$	$f_i(m_i - \bar{m})^2$
0	10	0,1	0	-5	25	2,5
3	20	0,2	0,6	-2	4	0,8
5	40	0,4	2	0	0	0
7	20	0,2	1,4	2	4	0,8
10	10	0,1	1	5	25	2,5
totali	100		$\bar{m}$ =5			6,6

- Formula "breve"

m_i	n_i	f_i	$m_i f_i$	m_i^2	$f_i (m_i^2)$
0	10	0,1	0	0	0
3	20	0,2	0,6	9	1,8
5	40	0,4	2	25	10
7	20	0,2	1,4	49	9,8
10	10	0,1	1	100	10
totali	**100**		**$\bar{m}$ =5**		**31,6**

$$\sigma^2 = (\sum_{i=1}^{M} f_i m_i^2) - \bar{m}^2$$

$$\sigma^2 = 31.6 - 5^2 = 6.6$$

Varianza e scarto quadratico medio

- Calcolo di σ^2 nel caso di
  - Istogrammi
  - Rilevazione per classi di modalità

Stessa soluzione della media: si utilizza il valore di centro classe come modalità e si procede normalmente

inf_i	sup_i	$\bar{c}_i$	f_i	$\bar{c}_i f_i$	$\bar{c}_i - \bar{m}$	$(\bar{c}_i - \bar{m})^2$	$f_i(\bar{c}_i - \bar{m})^2$
0	4	2	0,1	0,2	-4	16	1,6
4	6	5	0,4	2,0	-1	1	0,4
6	8	7	0,4	2,8	1	1	0,4
8	12	10	0,1	1,0	4	16	1,6
				6,0			**4,0**

- Spesso si indica lo scarto quadratico medio σ

$$\sigma = \sqrt{\sigma^2} = \sqrt{(\sum_{i=1}^{M} f_i m_i^2) - \bar{m}^2}$$

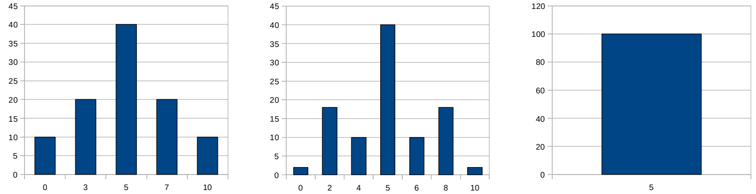
Coefficiente di variazione (cv)

La varianza è sensibile alla media

- Esempio 1: osservazioni O = {2 5 6 7}
  - Media = $\bar{O}$ = (2 + 5 + 6 +7)/4 = 5
  - Varianza = σ^2_O = (9 + 0 + 1 +4)/4 = 3,5
- Esempio 2: cambio unità di misura P = {20 50 60 70}
  - Media = $\bar{P}$ = (20 + 50 + 60 +70)/4 = 50
  - Varianza = σ^2_P = (900 + 0 + 100 +400)/4 = 350
- Introduco il coefficiente di variazione $cv = \frac{\sigma}{|\bar{m}|}$

$$cv(P) = \frac{\sqrt{350}}{|50|} = \frac{\sqrt{3.5 \cdot 100}}{|5 \cdot 10|} = \frac{10\sqrt{3.5}}{10|5|} = cv(O)$$

Varianza: considerazioni



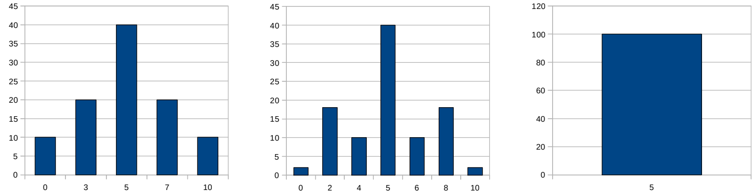
range = 10; σ^2 = 6.6 range = 10; σ^2 = 4.4 range = σ^2 = 0

- Considerazioni:
  - Migliore del range.
  - Istogrammi: uso il valore di centro classe
  - Sensibile alla presenza di outliers (come la media).
  - Sensibile alla media (indice assoluto)

Distanza Interquartile

Rappresenta la distanza fra il terzo ed il primo quartile

$$D = q_3 - q_1$$



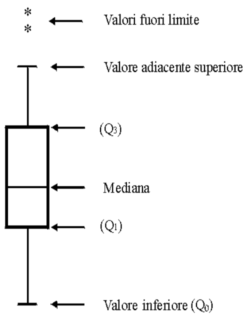
range = 10; σ^2 = 6.6 range = 10; σ^2 = 4.4 range = σ^2 = 0

D= 7-3 = 4 D = 6 - 4 = 2

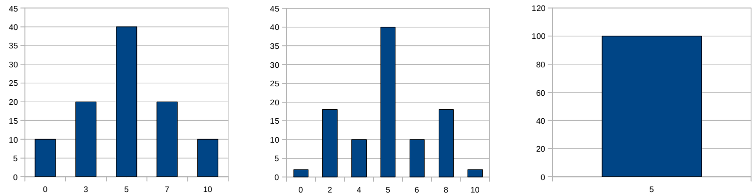
- Risulta meno sensibile al valore degli outlier
  - Il valore numerico del outlier non conta
  - Spesso viene usata per indicare gli outlier (boxplot v2.0)

Box-plot (seconda versione)

- Considerazione: 50% dati è fra q_1 e q_3
- Il 100% dei dati dovrebbe stare fra
  - Valore Adiacente Superiore = q_3+ k*D
  - Valore Adiacente Inferiore = q_1- k*D
- Versione 2 (k =1)
  - Rettangolo fra q_1 e q_3
  - Mediana (q_2) evidenziata
  - Due "baffi" (segmento)
    - da lato a min (q_4 e vas)
    - da lato max (q_0 e vai)
  - uso simboli discreti per i valori restanti

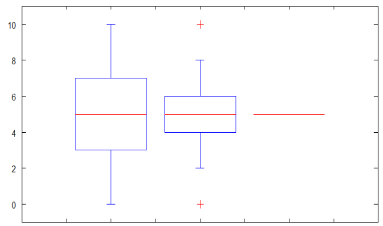


BoxPlot: strumento grafico di confronto



CV = 10; σ^2 = 6.6 CV = 10; σ^2 = 4.4 CV = σ^2 =D = 0

D= 7-3 = 4 D = 6 - 4 = 2



Indici di forma

Indici sintetici principali:

- Posizione: indica il “centro” delle osservazioni
- Variabilità: indica quanto le osservazioni si discostano dal “centro”

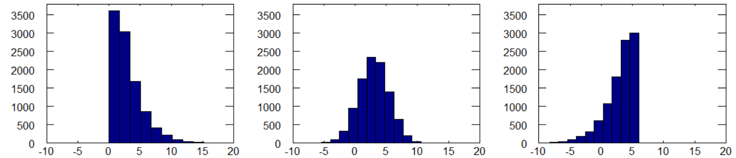
Spesso insufficienti, si usano anche

- Indici Sintetici di Forma: descrivono la “forma” della distribuzione delle osservazioni.
  - Simmetria: indica quanto la distribuzione sia asimmetrica rispetto al valore “centrale”
  - Normalità: quanto distribuzione è simile alla distribuzione di rifermento normale.

Indici di asimmetria

Asimmetria: distribuzione delle osservazioni rispetto al valore centrale (ovvero se sono in maggioranza maggiori o minori).

Esempio di 3 popolazioni (N= 1000) con $\bar{o} \approx 3$ e $\sigma^2 \approx 6$.

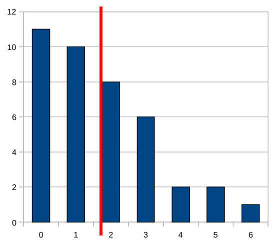


- Asimmetria Positiva:la distribuzione mostra una “coda” più accentuata verso destra (primo istogramma)
- Asimmetria Negativa:la distribuzione mostra una “coda” più accentuata verso sinistra (terzo istogramma)

Momento centrale terzo

Idea: medio gli scarti dalla media al cubo.

- Perché al cubo?
  - serve il segno, quindi potenza dispari.
  - con lo scarto semplice la somma è sempre nulla!
- Esempio



m_i	n_i	f_i	$m_i f_i$	$m_i - \bar{m}$	$(m_i - \bar{m})^3$
0	11	0,28	0	-1,7	-4,91
1	10	0,25	0,25	-0,7	-0,34
2	8	0,2	0,4	0,3	0,03
3	6	0,15	0,45	1,3	2,2
4	2	0,05	0,2	2,3	12,17
5	2	0,05	0,25	3,3	35,94
6	1	0,03	0,15	4,3	79,51
	40		1,7		

Momento centrale terzo standardizzato

- Momento centrale terzo:

$$\mu_3 = \sum_{i=1}^{M} f_i (m_i - \bar{m})^3$$

  - Calcolo in tabella
  - Valore legato a σ

m_i	n_i	f_i	$m_i f_i$	$m_i - \bar{m}$	$(m_i - \bar{m})^3$	$f_i (m_i - \bar{m})^3$
0	11	0,28	0	-1,7	-4,91	-1,35
1	10	0,25	0,25	-0,7	-0,34	-0,09
2	8	0,2	0,4	0,3	0,03	0,01
3	6	0,15	0,45	1,3	2,2	0,33
4	2	0,05	0,2	2,3	12,17	0,61
5	2	0,05	0,25	3,3	35,94	1,8
6	1	0,03	0,15	4,3	79,51	1,99
	40		1,7			3,06

- Momento terzo standardizzato
  - Stesso significato momento terzo
  - Meno sensibile alla variabilità

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{3{,}06}{(1{,}55)^3} = 0.83$$

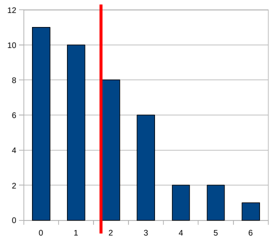
(primo) Indice di skewness di Pearson

Idea: se la moda “dista” molto dalla media le osservazioni non sono simmetriche

$$\frac{\bar{m}-moda}{\sigma}$$

Esempio

$$\frac{1{,}7-0}{1{,}57}=1{,}08$$

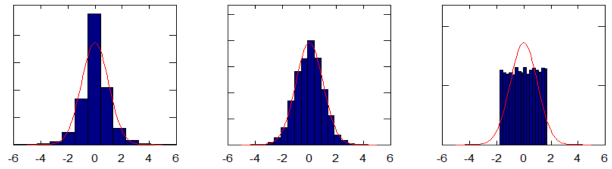


m_i	n_i	f_i	$m_i f_i$
0	11	0,28	0
1	10	0,25	0,25
2	8	0,2	0,4
3	6	0,15	0,45
4	2	0,05	0,2
5	2	0,05	0,25
6	1	0,03	0,15
	40		**1,7**

Indici di Curtosi (Kurtosi)

Curtosi: distribuzione delle osservazioni “vicine” a quelle della distribuzione normale

Esempio: 3 popolazioni (N= 1000) con $\bar{o} \approx 0$ e $\sigma^2 \approx 1$ e $\gamma_1 \approx 0$.



- Iper-normalità: l'istogramma delle osservazioni tendono mostrate un picco vicino al valore centrale (primo istogramma)
- Ipo-normalità: le osservazioni tendono a distribuirsi in maniera piatta (terzo istogramma)

Momento centrale quarto

- Momento centrale quarto

$$\mu_4=\sum_{i=1}^{M} f_i(m_i-\bar{m})^4$$

- Momento centrale quarto normato

$$\beta_2=\frac{\mu_4}{\sigma^4}$$

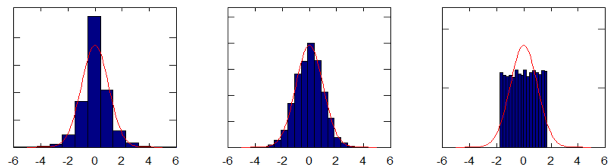
  - β_2 *sempre positivo.*
  - Una gaussiana ha $\beta_2=3$.
  - $\beta_2>3$ => iper-normale
  - $\beta_2<3$ => ipo-normale

Eccesso Curtosi

- Eccesso curtosi

$$\gamma_2=\frac{\mu_4}{\sigma^4}-3$$

  - Una gaussiana ha $\beta_2=3$ quindi $\gamma_2=0$.
  - $\gamma_2>0$ => iper-normale
  - $\gamma_2<0$ => ipo-normale



$\gamma_2=3.5$ $\gamma_2=0.027$ $\gamma_2=-1.955$