6 Indici di forma

6.1 I momenti empirici

Si definisce momento empirico di origine *m* e grado *r* la somma delle potenze *r-me* degli scarti dei singoli valori da *m* divisa per il totale delle osservazioni:

$$\mu_{m,r} = \frac{\sum_{i=1}^{n}(x_i - m)^r}{n}$$

Nel caso si abbia una distribuzione di frequenza, gli scarti vanno ponderati per le rispettive frequenze:

$$\mu_{m,r} = \frac{\sum_{i=1}^{k}(x_i - m)^r n_i}{n} = \sum_{i=1}^{k}(x_i - m)^r f_i .$$

Se l'origine *m* è uguale alla media aritmetica *M*, i momenti vengono definiti "*momenti centrati*" e si indicano semplicemente con μ_r:

$$\mu_r = \frac{\sum_{i=1}^{n}(x_i - M)^r}{n}$$ per una serie di valori

$$\mu_r = \frac{\sum_{i=1}^{k}(x_i - M)^r n_i}{n} = \sum_{i=1}^{k}(x_i - M)^r f_i$$ per una distribuzione di frequenze.

La media aritmetica *M* e la varianza σ^2 sono particolari momenti:

$M = \mu_{0,1}$

$\sigma^2 = \mu_2$

6.2 Asimmetria e curtosi

Una distribuzione di frequenza empirica si definisce *simmetrica* se la prima frequenza assoluta è uguale all'ultima, la seconda alla penultima, la terza alla terzultima e così via:

$n_1 = n_k$

$n_2=n_{k-1}$

$n_3=n_{k-2}$

……

dove *k* è il numero delle modalità o dei valori della variabile osservata.

Una distribuzione si definisce *asimmetrica positivamente* se vi è una maggiore concentrazione delle frequenze in corrispondenza di modalità basse della variabile, viceversa, se le frequenze si addensano maggiormente in corrispondenza di modalità alte della variabile, la distribuzione si dice *asimmetrica negativamente*.

Per una distribuzione simmetrica si verifica che la moda è uguale alla mediana e uguale alla media aritmetica: $M_o=M_e=M$, ma non è vero il contrario.

Se una distribuzione presenta un'asimmetria positiva, allora $M_o<M_e<M$, se invece presenta un'asimmetria negativa, allora $M_o>M_e>M$ (condizione necessaria, ma non sufficiente).

Di seguito vengono riportati alcuni indici, in ordine di importanza crescente, che, tenendo conto di quanto affermato, danno indicazioni sulla simmetria o meno di una distribuzione:

$a_1=(Q_3-Q_2)-(Q_2-Q_1)$

$$A_1=\frac{a_1}{Q_3-Q_1}$$

$$\delta=\frac{M-M_o}{\sigma}\cong\frac{3(M-Q_2)}{\sigma}\qquad\text{(}Pearson\text{)}$$

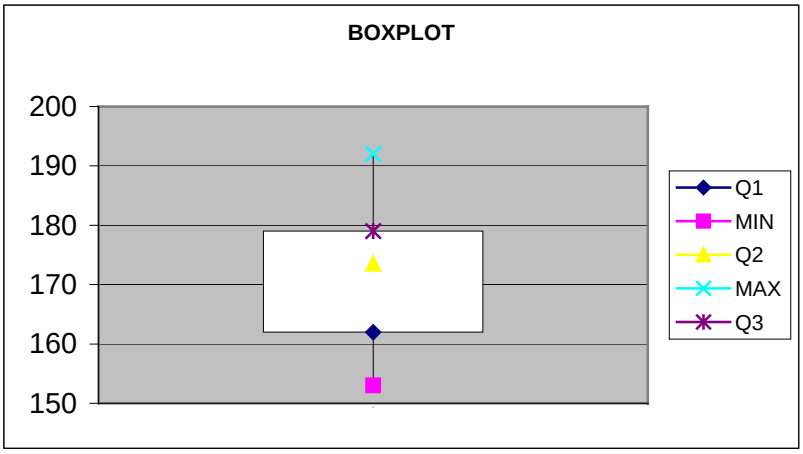
$$\beta_1=\frac{\mu_3}{\mu_2^{3/2}}=\frac{\mu_3}{\sigma^3}\qquad\text{(}Fisher\text{)}$$

I suddetti indici valgono 0 in caso di simmetria, sono positivi in caso di asimmetria positiva e negativi in caso di asimmetria negativa, ma non è detto il contrario.

Gli indici A_1, δ e β_1 sono adimensionali, in quanto hanno numeratore e denominatore espressi nella stessa unità di misura.

6.3 Il boxplot

Il boxplot è un grafico che dà indicazioni sulla simmetria o asimmetria di una distribuzione, in quanto è costituito da una scatola, i cui estremi sono il I ed il III quartile (Q_1, Q_3). La scatola è sezionata dalla mediana (Q_2) ed ha dei baffi in corrispondenza, in genere, dei valori minimo e massimo:



Il boxplot dà indicazioni anche sulla variabilità di una serie; infatti sia *l'intervallo di variazione o range*=MAX-MIN, sia la *differenza interquartile* Q_3-Q_1, possono essere considerati indici di variabilità, seppure grossolani, in quanto il primo non tiene conto delle unità centrali, il secondo dei valori estremi.

6.4 Esempi

Consideriamo la distribuzione di frequenza della variabile "numero di componenti per famiglia", rilevata su un collettivo di 150 famiglie:

x_i	n_i
1	5
2	46
3	35
4	28
5	15
6	10
7	7
8	3
9	1
totale	150

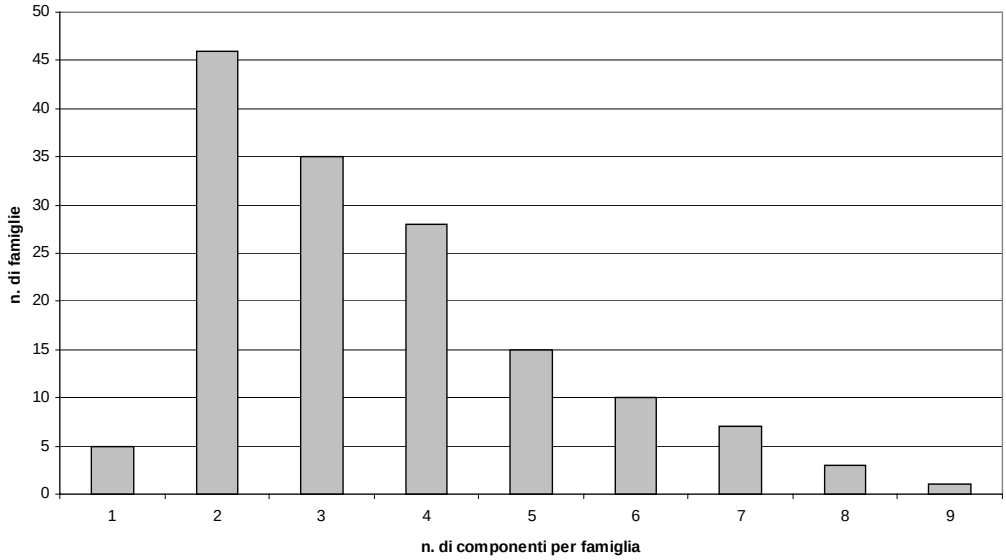
Calcoliamo i tre quartili, quindi gli indici a_1 ed A_1:

x_i	n_i	f_i	F_i
1	5	0,033	0,033
2	46	0,307	0,340
3	35	0,233	0,573
4	28	0,187	0,760
5	15	0,100	0,860
6	10	0,067	0,927
7	7	0,047	0,973
8	3	0,020	0,993
9	1	0,007	1,000
totale	150	1	

$Q_1=2$, $Q_2=M_e=3$, $Q_3=4$,

$a_1=(Q_3-Q_2)-(Q_2-Q_1)=0$, $A_1=\frac{a_1}{Q_3-Q_1}=0$

Gli indici a_1 ed A_1 assumono entrambi valore 0, ma la distribuzione non è simmetrica; piuttosto, sembra esserci un'asimmetria positiva, come si evince anche dalla rappresentazione grafica:



Calcoliamo, adesso, l'indice di Pearson:

x_i	f_i	$x_i f_i$	x_i^2	$x_i^2 f_i$
1	0,033	0,033	1	0,033
2	0,307	0,613	4	1,227
3	0,233	0,700	9	2,100
4	0,187	0,747	16	2,987
5	0,100	0,500	25	2,500
6	0,067	0,400	36	2,400
7	0,047	0,327	49	2,287
8	0,020	0,160	64	1,280
9	0,007	0,060	81	0,540
totale	1,000	3,540		15,353

$M_o=2$

$$M=\sum_{i=1}^{9} x_i f_i = 3{,}54$$

$$\sigma = \sqrt{\sum_{i=1}^{9} x_i^2 f_i - M^2} = \sqrt{15{,}353 - (3{,}54)^2} = 1{,}68$$

$$\delta = \frac{M - M_o}{\sigma} = 0{,}92$$

Il valore di $\delta = 0{,}92$ indica asimmetria positiva, come pure l'indice di Fisher, indice ancora più informativo:

x_i-M	$(x_i$-M$)^3$	$(x_i$-M$)^3 f_i$
-2,540	-16,387	-0,546
-1,540	-3,652	-1,120
-0,540	-0,157	-0,037
0,460	0,097	0,018
1,460	3,112	0,311
2,460	14,887	0,992
3,460	41,422	1,933
4,460	88,717	1,774
5,460	162,771	1,085
		4,411

$$\beta_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3} = \frac{4,411}{(1,68)^3} = 0,93 .$$

Essendo la distribuzione asimmetrica positivamente, si verifica che $M_o<M_e<M$.

Consideriamo adesso la distribuzione delle altezze in *cm* rilevate su un gruppo di 100 studenti:

x_i-x_{i+1}	n_i	$_cx_i$	f_i	$_cx_i f_i$
150-155	2	152,5	0,02	3,05
155-160	4	157,5	0,04	6,30
160-165	8	162,5	0,08	13,00
165-170	14	167,5	0,14	23,45
170-175	24	172,5	0,24	41,40
175-180	23	177,5	0,23	40,83
180-185	15	182,5	0,15	27,38
185-190	7	187,5	0,07	13,13
190-195	2	192,5	0,02	3,85
195-200	1	197,5	0,01	1,98
totale	100		1	174,35

e calcoliamo l'indice di curtosi proposto da Pearson. La curtosi descrive il modo in cui si distribuiscono le frequenze dei valori:

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$$

$_cx_i$-M	$(_cx_i$-M$)^2$	$(_cx_i$-M$)^2 f_i$	$(_cx_i$-M$)^4$	$(_cx_i$-M$)^4 f_i$
-21,85	477,42	9,55	227932,24	4558,64
-16,85	283,92	11,36	80611,99	3224,48
-11,85	140,42	11,23	19718,48	1577,48
-6,85	46,92	6,57	2201,72	308,24
-1,85	3,42	0,82	11,71	2,81
3,15	9,92	2,28	98,46	22,64
8,15	66,42	9,96	4411,95	661,79
13,15	172,92	12,10	29902,19	2093,15
18,15	329,42	6,59	108519,18	2170,38
23,15	535,92	5,36	287212,93	2872,13
		75,83		17491,76

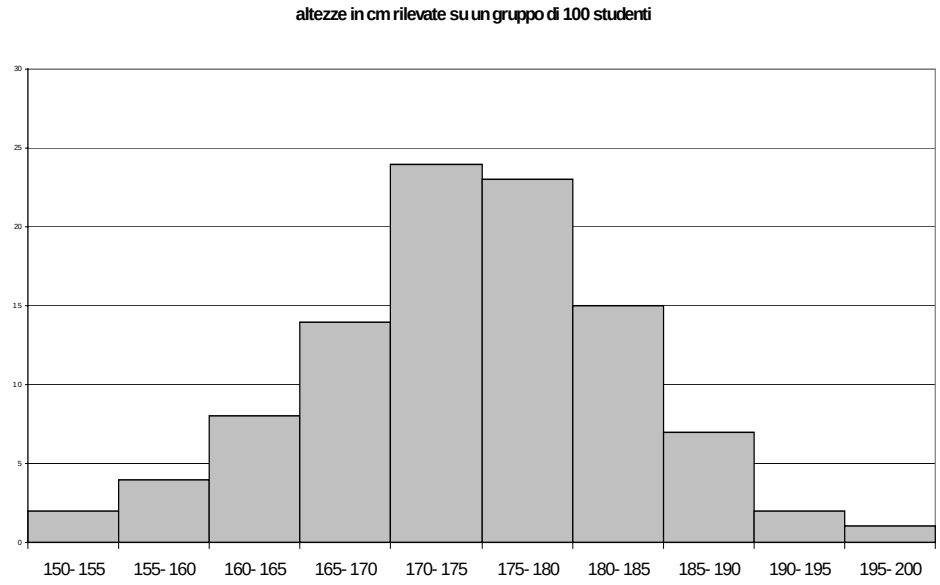
$$M = \sum_{i=1}^{10} {}_c x_i \, f_i = 174{,}35 \ cm$$

$$\mu_2 = \sigma^2 = \sum_{i=1}^{10} \left({}_c x_i - M \right)^2 f_i = 75{,}83 \ cm^2$$

$$\mu_4 = \sum_{i=1}^{10} \left({}_c x_i - M \right)^4 f_i = 17491{,}76 \ cm^4$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{17491{,}76}{(75{,}83)^2} = 3{,}04 \, .$$

Il valore di β_2, molto vicino a 3, indica che la distribuzione è mesocurtica (cfr.par. 5.4). Inoltre, la distribuzione è simmetrica, come si evince anche dalla rappresentazione grafica, per cui può essere ben descritta dal modello di Gauss:


altezze in cm rilevate su un gruppo di 100 studenti
30
25
20
15
10
5
0
150- 155
155- 160
160- 165
165- 170
170- 175
175- 180
180- 185
185- 190
190- 195
195- 200