FACOLTA' DI INGEGNERIA

CORSO DI LAUREA IN INGEGNERIA PER L'AMBIENTE ED IL TERRITORIO

CORSO DI STATISTICA E CALCOLO DELLE PROBABILITA'

PROF. PASQUALE VERSACE

SCHEDA DIDATTICA N°4

ARGOMENTO:

ANALISI DI BASE DEI DATI CAMPIONARI

A.A. 2008-09

ANALISI DEI DATI

Il primo passo necessario in qualsiasi studio d'ingegneria è l'analisi dei dati disponibili, per valutarne la natura ed il grado di incertezza. Esistono diversi metodi di organizzazione, presentazione e sintesi dei dati osservati che ne facilitano l'interpretazione e la valutazione, alcuni dei quali sono descritti in queste pagine.

VISUALIZZAZIONI GRAFICHE

Istogrammi. Un primo passo utile nell'analisi dei dati osservati (o campionari) è quello di rappresentarli con un tipo di grafico a barre. Si considerino, per esempio, i dati presentati nella tabella 1: questi numeri rappresentano le piogge annue registrate nella stazione di Chiaravalle Centrale, i cui valori variano da 535.7 a 2661.9 mm. Si divida questo intervallo di valori in intervalli di 270 mm ciascuno: 511- 780, 781 - 1050, etc. e si contino il numero di occorrenze in ciascun intervallo.

Anno	Piogge annue (mm)		Anno	Piogge annue (mm)		Anno	Piogge annue (mm)
1916	1312.6		1946	2204.8		1976	1201.6
1917	1425.9		1947	1617.4		1977	832.6
1918	879.5		1948	1242.7		1978	1589.9
1919	980.1		1949	1343.1		1979	1605.3
1920	1084.2		1950	1331.3		1980	1054.5
1921	1058.8		1951	2661.9		1981	739.1
1922	535.7		1952	1314.8		1982	1334.9
1923	883.3		1953	1881.3		1983	1491.6
1924	964.5		1954	2044.7		1984	1448.4
1925	1925.9		1955	1515.6		1985	1452.8
1926	1275.6		1956	1287.9		1986	1308.0
1927	2076.4		1957	2131.2		1987	1061.4
1928	1924.2		1958	1670.5		1988	803.4
1929	1641.8		1959	1789.5		1989	1230.4
1930	2126.0		1960	1451.6		1990	944.4
1931	1792.3		1961	972.4		1991	977.0
1932	1370.1		1962	1480.3		1992	1438.6
1933	2127.9		1963	1428.4		1993	1741.6
1934	1433.5		1964	1676.2		1994	895.0
1935	2078.3		1965	1127.9		1995	1602.6
1936	1844.6		1966	1493.2		1996	2416.4
1937	1468.9		1967	1103.2		1997	825.0
1938	1155.0		1968	1177.9		1998	878.0
1939	2007.4		1969	1599.9		1999	1061.6
1940	1783.2		1970	1325.0			
1941	1245.3		1971	1405.2			
1942	1558.2		1972	1908.1			
1943	1835.1		1973	978.6			
1944	1301.5		1974	1235.6			
1945	1251.2		1975	1072.2			

Tab. 1- Serie storica delle piogge annue della stazione di Chiaravalle Centrale.

Rappresentando la frequenza delle occorrenze in ogni intervallo come un rettangolo, si ottiene un *istogramma*, come quello mostrato nella figura 1, in cui le altezze di ogni colonna sono proporzionali al numero di occorrenze in quell'intervallo. Il grafico dà un'immediata visione del campo di variazione dei dati, dei valori più frequenti e del grado dispersione rispetto al valore centrale.

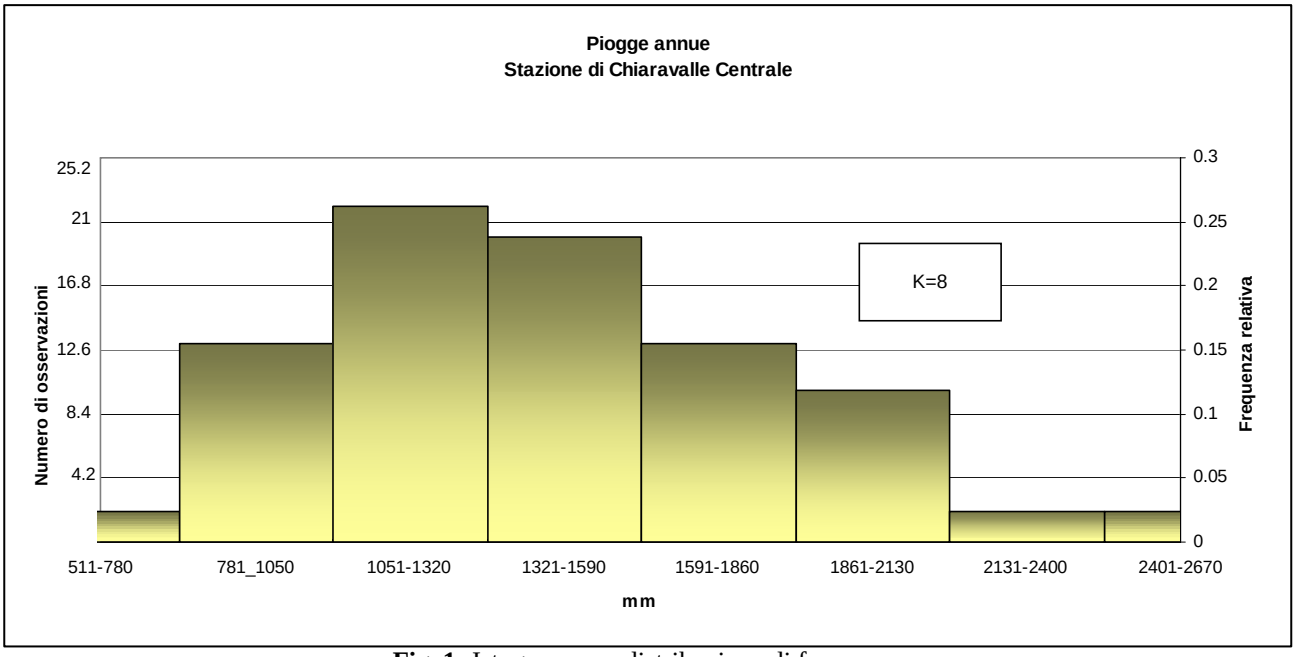


Fig. 1- Istogramma e distribuzione di frequenza

Se la scala delle ordinate dell'istogramma viene divisa per il numero totale di dati disponibili (frequenza relativa), si ottiene la *distribuzione di frequenza*. Nella figura 1 i numeri riportati nella scala a destra sono ottenuti dividendo quelli riportati nella scala a sinistra per il numero totale di osservazioni, cioè 84. Dividendo ancora questa scala per l'ampiezza dell'intervallo (270 mm) si ottiene la *distribuzione di densità di frequenza*, con unità delle ordinate pari a "frequenza per mm". L'area sottesa da questo istogramma è pari all'unità. Questa forma di visualizzazione è preferita quando devono essere comparati differenti set di dati, con ampiezze di intervallo diverse.

Particolare attenzione deve essere posta nella scelta dell'ampiezza di ciascun intervallo di questi diagrammi: il numero di classi, infatti, può condizionare l'impressione che si riceve circa l'andamento dei dati.

Un criterio pratico è stata suggerito da Sturges [1926]: se il numero di dati è *n*, il numero di intervalli *k* tra il valore minimo e massimo osservati dovrebbe essere circa:

$k = 1 + 3.3 \log n$

in cui con log si indica il logaritmo in base 10. Nel caso di Chiaravalle Centrale

$k = 1 + 3.3 \log 84 = 7.35$

La figura 2 riporta due istogrammi, ottenuti sempre con i dati della tabella 1, in cui k=4 e k=30, e mostra l'influenza dell'ampiezza dell'intervallo adottata sulla forma dell'istogramma.
Queste variazioni possono in qualche modo disorientare, ma, in realtà, sono indicative della insufficienza di un singolo campione di dati a rappresentare bene alcuni aspetti relativi al comportamento del fenomeno osservato.

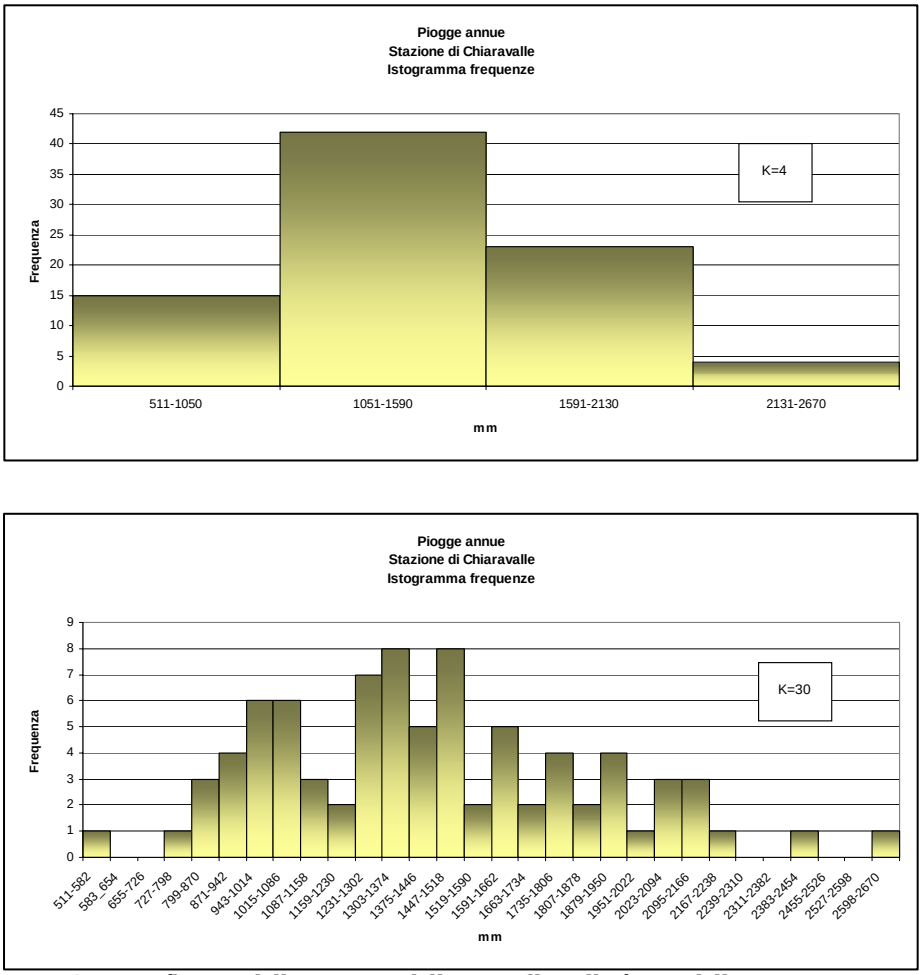


Fig.2 – Influenza dell'ampiezza dell'intervallo sulla forma dell'istogramma

La *distribuzione di frequenza cumulata*, un'altra utile rappresentazione grafica dei dati, si ottiene dalla distribuzione di frequenza calcolando le successive somme parziali delle frequenze associate ad ogni punto di divisione di ciascun intervallo. Questi punti sono diagrammati e connessi da linee rette così da formare una funzione non decrescente (monotona) che varia da zero all'unità. In realtà si evita l'arbitrarietà della scelta degli intervalli riportando un punto per ogni osservazione, cioè riportando i punti di coordinate ($x^{(i)}$, i/n) , dove $x^{(i)}$ è l'i-esimo valore della lista dei dati ordinati in maniera crescente, mentre i/n, definita **Plotting Position (PP(i))**, è la frequenza dei dati che assumono valori minori o al più uguali a $x^{(i)}$ (Figura 3). La definizione di Plotting Position pari a i/n implica che al valore massimo del campione si assegni una frequenza cumulata uguale a uno;

poiché la frequenza costituisce un'approssimazione della probabilità, questo implica l'attribuzione al valore massimo del carattere di limite superiore della distribuzione, ovvero si imporrebbe, erroneamente, che l'evento $X \leq x^{(n)}$ sia un evento certo! Sembra, dunque, opportuno introdurre altre formule di Plotting Position; a riguardo ci si può ricondurre all'espressione generale:

$$PP(i) = \frac{i-a}{n+1-2a} \qquad i = 1,2,...,n$$

nella quale *a* è una costante, il cui valore distingue tra loro le diverse espressioni. Se si pone *a=0* si ottiene la formula di ***Weibull***:

$$PP(i) = \frac{i}{n+1} \qquad i = 1,2,...,n$$

Se invece si impone *a=0.5* si ottiene la formula di ***Hazen***:

$$PP(i) = \frac{i-0.5}{n} \qquad i = 1,2,...,n$$

Intermedia tra le due è la formula di ***Gringorten***, che si ottiene assumendo *a=0.44*:

$$PP(i) = \frac{i-0.44}{n+0.12} \qquad i = 1,2,...,n$$

In Figura 3 si riportano, per i dati campionari di partenza, le frequenze cumulate utilizzando tutte le formule di Plotting Position esposte; dal grafico si evince come non vi siano sostanziali differenze nell'utilizzo di una espressione o di un'altra. L'utilizzo dell'espressione *i/n* è comunque da evitare per il motivo sopra esposto.

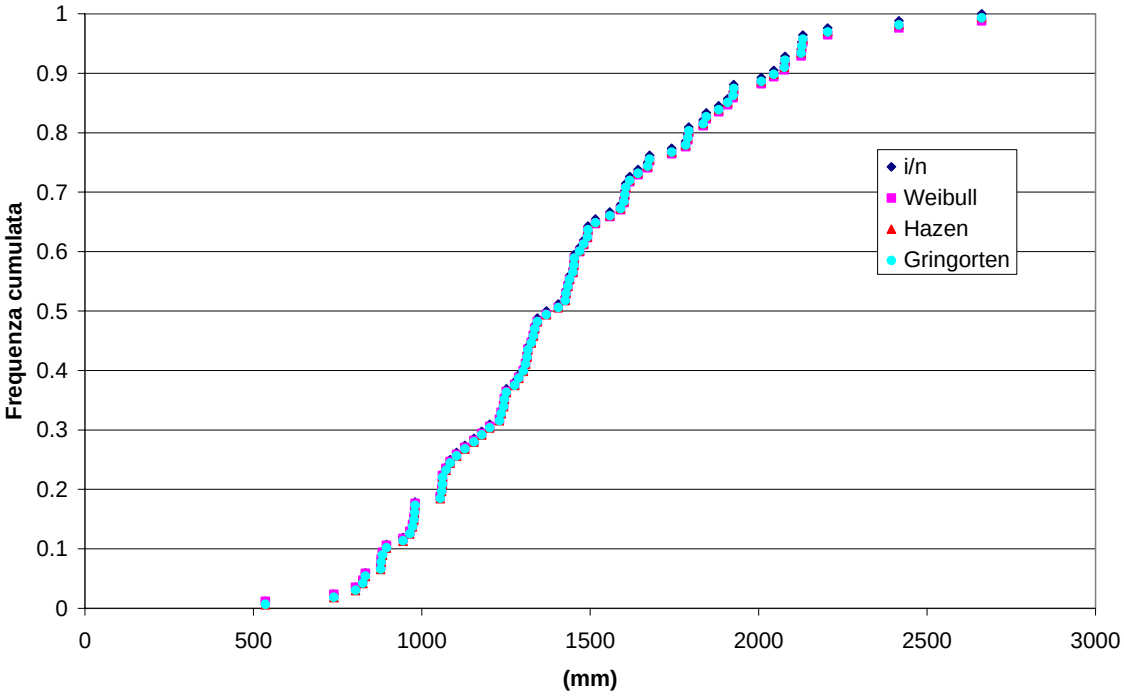


Fig. 3- Distribuzione di frequenza cumulata.

VALORI NUMERICI SINTETICI

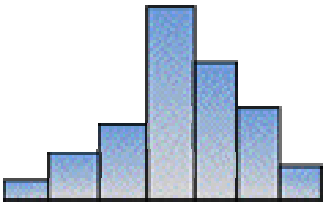
Misure del valore centrale. Il singolo valore più rappresentativo di un campione di dati è il suo valore medio, o media aritmetica. Se la sequenza di valori osservati è indicata con $x_1, x_2, \ldots, x_n$ la *media campionaria* $\bar{x}$ è semplicemente:

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

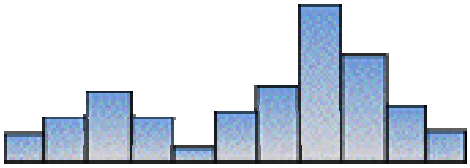
Il valore medio delle piogge annue di Chiaravalle Centrale risulta ad esempio:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{119753.5}{84} = 1425.6 \text{ mm}$$

La media campionaria è usualmente interpretata come valore tipico o valore centrale dei dati.
Un'altra misura della tendenza centrale della distribuzione è la *moda*, il valore che si verifica con la maggiore frequenza nel campione di dati: se tale valore è unico, si ha una distribuzione unimodale o monomorfa; se fra i dati considerati, più d'uno si presenta con la medesima frequenza, l'insieme ha più mode ed i risultati si accumulano a formare due picchi non adiacenti, la distribuzione si definisce bimodale o bimorfa. Se, invece, la distribuzione si presenta con più mode, si chiama plurimodale o plurimorfa. Può accadere che un insieme di dati sia privo di moda: ciò accade quando tutti i dati hanno la stessa frequenza.



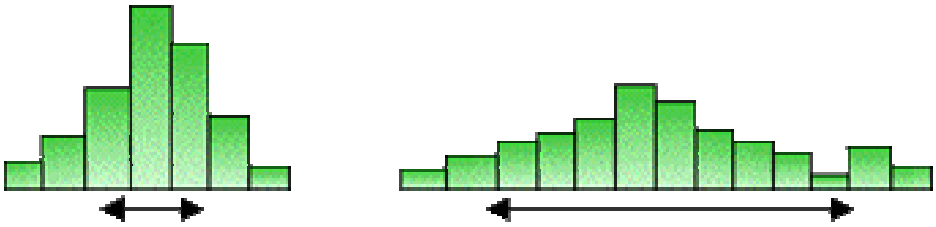
unimodale



bimodale

La *mediana*, invece, detta anche valore centrale, è il valore che in una serie ordinata di osservazioni occupa il posto centrale e suddivide esattamente la serie dei dati in due parti di numerosità uguale, ovvero è il valore per cui la frequenza cumulata è pari a 0.5. Essa è il valore di mezzo se la numerosità *n* del campione è dispari, o la media dei due valori di mezzo se *n* è pari.
La mediana dei dati relativi alla rottura è 1387.65 mm. La moda è pari circa a 1061 (si ripete nel 1987 e nel 1999) ed è unica poiché tutti gli altri valori sono diversi tra loro, e nessuno ha una frequenza maggiore.

Misure della dispersione. Dato un campione di valori osservati è utile sintetizzare con un singolo valore l'informazione riguardante la variabilità delle osservazioni. In passato la misura più spesso considerata era l'intervallo dei valori (in inglese: range). Questo numero, che è semplicemente la differenza tra il massimo ed il minimo valore osservato, ha il vantaggio di essere semplice da calcolare. L'intervallo dei valori osservati, però, dà molto peso ai valori degli estremi, che spesso sono affetti da errori sperimentali, e trascura l'informazione fornita da tutti gli altri dati. Il valore assunto dal range, inoltre, dipende dalla dimensione del campione osservato.



Piccola variabilità Grande variabilità

Una misura più significativa della dispersione è la *varianza campionaria*. E' l'analogo del momento di inerzia, poiché misura i quadrati della distanze da un punto centrale che in questo caso è costituito dalla media campionaria. La varianza campionaria s^2 è definita come:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

dove i quadrati delle distanze sono divise per n per avere una media dei quadrati degli scostamenti ed eliminare così la dipendenza di s^2 dalla dimensione del campione. Lo sviluppo dell'equazione porta ad un'espressione che può risultare più utile per il calcolo di s^2:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2\right)$$

ma poiché:

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

si ha:

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right)$$

La radice quadrata positiva s della varianza campionaria è definita come *scarto quadratico medio* o *deviazione standard campionaria*.

E' un parametro che riguarda la "forma della distribuzione" dei dati piuttosto che il loro valore: praticamente, più piccola è la deviazione standard (o la varianza), più i dati sono raggruppati intorno alla media del campione. L'aggiunta di una costante a tutti i valori osservati altererebbe la media campionaria, ma lascerebbe inalterata la deviazione standard del campione.
Per i dati di Chiaravalle Centrale la varianza campionaria e la deviazione standard sono:

$$s^2 = \frac{1}{n}\left(\sum_{i=1}^{n} x_i{}^2 - n\bar{x}^{-2}\right) = 1/84(185286997 - 170725008) = 173357 mm^2$$

$$s = \sqrt{173357} = 416.36 \text{ mm}$$

Quando si vogliono confrontare le dispersioni relative a più di un tipo di dati, è opportuno disporre di un indice adimensionale, come il *coefficiente di variazione campionario.* Questa quantità *v* è definita come il rapporto della deviazione standard e della media campionarie.

$$v = \frac{s}{\bar{x}}$$

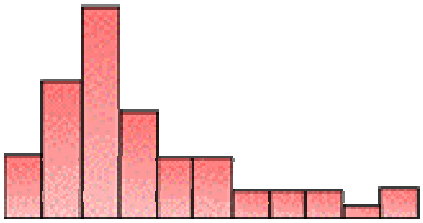
Il coefficiente di variazione campionario dei dati di Chiaravalle Centrale è:

$$v = \frac{s}{\bar{x}} = \frac{416.36}{1425.6} = 0.293$$

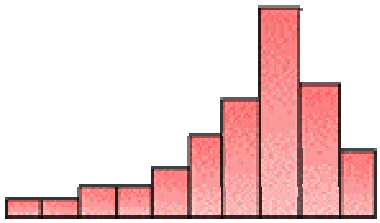
Misura dell'asimmetria. E’ un altro valore sintetico dei dati osservati può essere considerato come una estensione logica del ragionamento che ha condotto alla formula della varianza campionaria. Se la varianza è il momento del secondo ordine calcolato rispetto alla media, così il *coefficiente di asimmetria campionario* è relativo al momento del terzo ordine calcolato sempre rispetto alla media. Per rendere il coefficiente adimensionale il momento è diviso per il cubo della deviazione standard campionaria.
Il coefficiente di asimmetria g_1 fornisce una misura del grado di asimmetria rispetto alla media dei dati ed è pari al momento del terzo ordine della variabile standardizzata:

$$g_1 = \frac{(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$$

Il coefficiente è positivo per istogrammi asimmetrici a destra (cioè con la coda di destra più lontana dalla media) e negativo per quelli asimmetrici a sinistra.



Asimmetria positiva
(Istogramma asimmetrico a destra:
la media è maggiore della mediana)



Asimmetria negativa
(Istogramma asimmetrico a sinistra:
la media è minore della mediana)

Per i valori delle piogge annue di Chiaravalle Centrale si ottiene:

$g_1 = 0.474 > 0$

I tre parametri, $\bar{x}$, s e g_1 costituiscono delle elaborazioni di base dei dati campionari, sufficienti a dare indicazioni generali sulla forma dell'istogramma.

Un quarto parametro sintetico è il *coefficiente di curtosi*, ma è raramente applicato quando si dispone di pochi dati campionari ed ha senso calcolarlo solo per distribuzioni unimodali.

Il coefficiente di curtosi campionario g_2 è legato alla mancanza di picchi e misura in qualche modo il grado di "schiacciamento" dell'istogramma delle frequenze. Esso ha la seguente espressione:

$$g_2 = \frac{(1/n)\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$$

ed è, quindi, pari al momento del quarto ordine della variabile standardizzata cui viene sottratto il valore 3, momento che si ottiene in corrispondenza della cosiddetta curva a campana o *curva normale*. Un valore del coefficiente di curtosi campionario positivo si ottiene in caso di distribuzione eccessivamente alta e con code lunghe (leptocurtosi); viceversa un valore negativo indica una distribuzione eccessivamente appiattita, con code corte (platicurtosi).

Per i dati in esame si ha:

$g_2 = 2.92 - 3 = -0.08 < 0$

e pertanto la curva ha un andamento platicurtico.