Misure di *tendenza centrale*, *variabilità* e *forma* di una distribuzione.

Cap. 5 Luccio & Caudek

Nella statistica parliamo di *popolazione* (o universo statistico), quando ci riferiamo alla classe di *tutti* gli *eventi*, mentre parliamo di *campioni* quando ci riferiamo solo ad *alcuni eventi* appartenenti a una certa popolazione.

Nelle ricerche si ricorre a dei sottoinsiemi dell'universo: appunto, i campioni. Dai campioni, poi, si cercherà di risalire alle caratteristiche dell'universo a cui appartengono.

Per descrivere le caratteristiche degli universi come dei campioni, non potrò riferirmi singolarmente a ciascun evento membro, devo invece trovare degli *indici riassuntivi*.

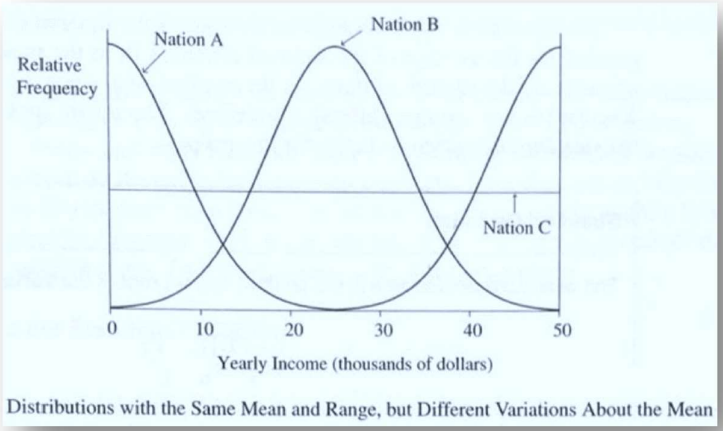
Quando questi indici si riferiscono a popolazioni, si chiamano *parametri* (indicati con lettera in alfabeto greco); quando gli indici si riferiscono invece a campioni, si chiamano *indici statistici* (o statistiche).

Indici di tendenza centrale

Dato un insieme di misure (campione) poniamoci il problema di determinare un valore in grado di "catturare" le caratteristiche della distribuzione nel suo complesso.

Valori *tipici* della distribuzione: Moda, Mediana, e la Media (aritmetica).

Moda Punto centrale della classe di misure più frequente.



Distributions with the Same Mean and Range, but Different Variations About the Mean

Distribuzioni bi(tri)modali: una distribuzione può avere più di una moda se vi sono più massimi, non necessariamente dello stesso valore.

Mediana Si può usare a partire dalla scala ordinale; rappresenta il valore che occupa la posizione centrale quando le osservazioni di un campione sono ordinate in base al loro valore.

Nel 2005 negli USA si sono registrate 30.1 milioni di famiglie con un solo componente, 37.0 milioni con due componenti, 17.8 milioni con tre componenti, 15.3 milioni con quattro componenti, 10.9 milioni con cinque o più componenti.

a) Costruisci la distribuzione delle frequenze assolute e relative e delle frequenze cumulate

b) Rappresenta la distribuzione con metodo grafico

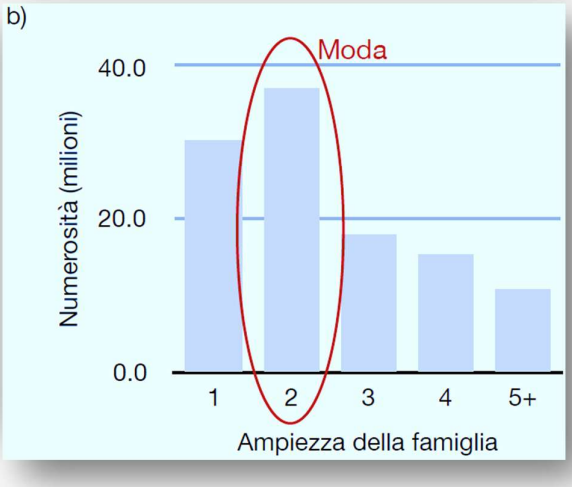
c) Trova moda e mediana (anche con metodo grafico)

Moda = modalità "2 componenti"
Mediana = modalità "2 componenti"

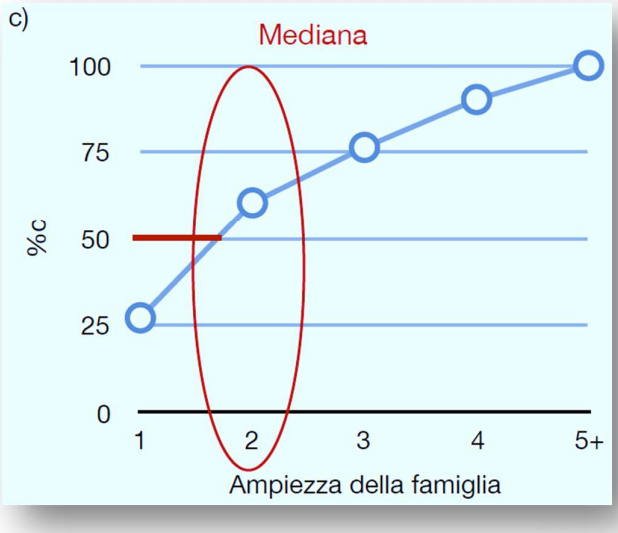
a)

Ampiezza famiglia	f_o	%f	f_c	%c
1	30.1	27.1	30.1	27.1
2	37.0	33.3	67.1	60.4
3	17.8	16.0	84.9	76.4
4	15.3	13.8	100.2	90.2
5+	10.9	9.8	111.1	100
Totale	**111.1**	**100**		

b)



c)



c) Trova moda e mediana (anche con metodo grafico)

Moda = modalità "2 componenti"
Mediana = modalità "2 componenti"

Media Geometrica (γ) La radice n-esima del prodotto degli n dati:

dato il vettore di misure X={1,2,4,6,8,9}, la media geometrica *G* si ottiene come

$$G = \sqrt[6]{1 \cdot 2 \cdot 4 \cdot 6 \cdot 8 \cdot 9}$$

$$= (1 \cdot 2 \cdot 4 \cdot 6 \cdot 8 \cdot 9)^{\frac{1}{6}} = 3.888323;$$

- poco utilizzata nelle scienze sociali;
- utile per rappresentare la tendenza centrale in distribuzioni non simmetriche (linea tratteggiata in Figura 1; *Excel Lezione 10_Foglio 1*).

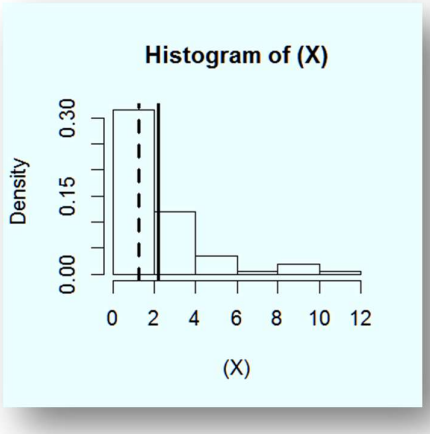


Figura 1.
Linea continua: Media aritmetica.
Linea tratteggiata: G.

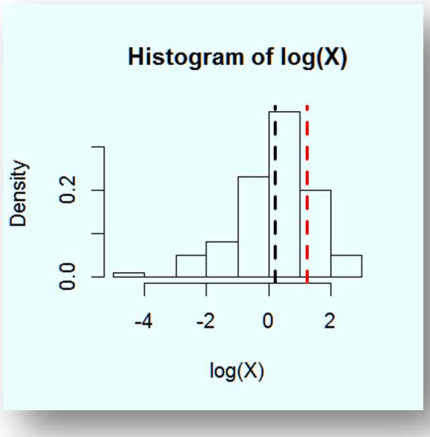


Figura 2
In nero: media di log(X); in rosso: G.

- Relazione con la media di log(x), vedi Figura 2.

  Sfruttando le identità dei logaritmi:

  $exp[log(G)] = G$

  $$log\left(\sqrt[y]{x}\right) = log\left(x^{\frac{1}{y}}\right) = \frac{1}{y} log(x),$$

  possiamo riscrivere *G* come:

  $$G = exp\left[\frac{1}{6}\left(log_e(1) + log_e(2) + (4) + log_e(6) + log_e(8) + log_e(9)\right)\right].$$

G = funzione esponenziale (antilogaritmo) della media dei logaritmi di X (solo valori positivi). Ecco spiegato il funzionamento "migliore" (*si fa trascinare meno verso la coda destra*) in caso di distribuzioni asimmetriche.

Media Armonica (α) Il reciproco della media dei reciproci dei dati: dato il vettore di misure X={1,2,4,6,8,9}, la media armonica *a* si ottiene come

$$a = \frac{1}{\frac{1}{6}\left(\frac{1}{1}+\frac{1}{2}+\frac{1}{4}+\frac{1}{6}+\frac{1}{8}+\frac{1}{9}\right)} = 2.787097.$$

- poco utilizzata nelle scienze sociali;
- in situazioni in cui occorre mediare rapporti o tassi di crescita (velocità)

Esempio: Viaggiate da A a B (200 Km) in 1,5 ore. Avete guidato per il primo tragitto (100 Km) su di una BMW (200 Km/h), e per il secondo tragitto (100 Km) su di una fiat IDEA Multijet 1300 (Diesel, 100 Km/h).

Spazio	Velocità	Tempo
100 Km	200 Km/h	1/2 h
100 Km	100 Km/h	1 h

La velocità (spazio/tempo, media) di crociera è 200 Km / 1.5 h = 133.3333 Km/h.

Media Geometrica delle velocità delle due automobili:

$$G = \sqrt[2]{100 \cdot 200} = 141.4214 \text{ Km/h}.$$

Media Aritmetica delle velocità delle due automobili:

$$\bar{x} = \frac{1}{2}(100 + 200) = 150.000 \text{ Km/h}.$$

Media Armonica

$$a = \frac{1}{\frac{1}{2}\left(\frac{1}{100}+\frac{1}{200}\right)} = 133.3333 \text{ Km/h}.$$

Media aritmetica

Definiamo la media aritmetica come la somma di tutte le osservazioni di una distribuzione (campionaria), divisa per il numero totale delle osservazioni:

dato X={1,2,4,6,8,9}, allora la media sarà

$$\overline{\mathrm{X}} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x_i}}{\mathrm{n}} = \frac{1+2+4+6+8+9}{6} = 5.$$

1. Se un insieme di osservazioni è costituito da due sottoinsiemi disgiunti di grandezza n_1 e n_2, ad esempio $\mathrm{X}_1 = \{1,2,4,6,8,9\}$ e $\mathrm{X}_2 = \{3,4,4,8\}$, e con medie $\bar{X}_1 = 5$ e $\bar{X}_2 = 4.75$, allora la media dell'insieme totale sarà:

$$\overline{\mathrm{X}} = \sum_{\mathrm{i}=1}^{n_1+n_2} \mathrm{X_i} \Big/ (n_1+n_2) = \frac{(1+2+4+6+8+9)+(3+4+4+8)}{10} = 4.9.$$

Più in generale:

$$\overline{\mathrm{X}} = \frac{\boldsymbol{n_1\bar{X}_1 + n_2\bar{X}_2}}{\boldsymbol{n_1+n_2}} = \frac{6\cdot 5 + 4\cdot 4.75}{10} = 4.9.$$

2. Definiamo gli *scarti* come le quantità $d_i = (\mathrm{X_i} - \overline{\mathrm{X}})$, la cui sommatoria è zero:

$$\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\mathbf{X_i} - \overline{\mathbf{X}}) = \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \mathbf{X_i} - \boldsymbol{n}\overline{\mathbf{X}} = \boldsymbol{n}\overline{\mathbf{X}} - \boldsymbol{n}\overline{\mathbf{X}} = \mathbf{0}.$$

3.1 La media risente dei cambiamenti effettuati agli estremi di una distribuzione, mentre la mediana è insensibile a questi cambiamenti.

$\mathrm{X} = \{1,2,4,\mathbf{6},8,9,9\}$; $\overline{\mathrm{X}} = 5.57 < Mediana = 6$

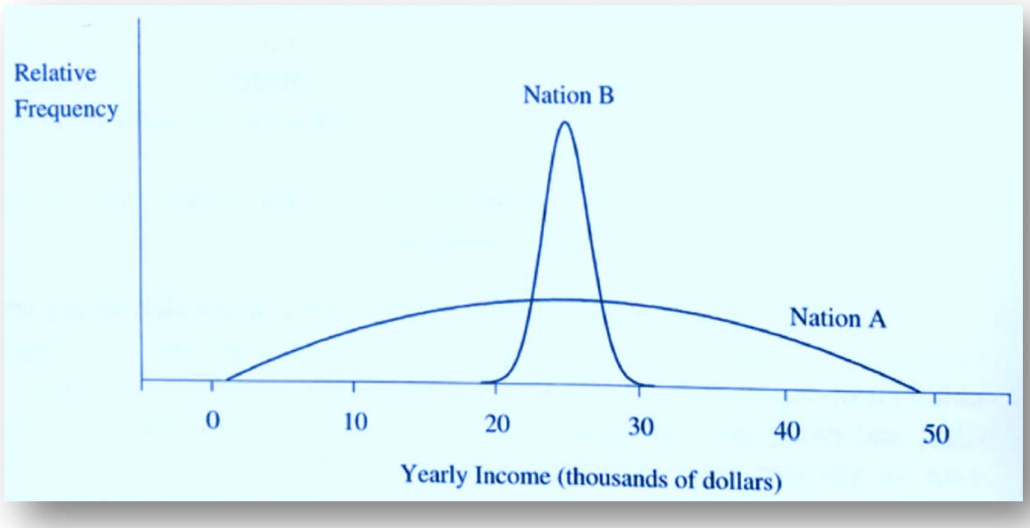
$\mathrm{X} = \{1,2,4,\mathbf{6},8,9,\mathbf{19}\}$; $\overline{\mathrm{X}} = 7 > Mediana = 6$

3.2 La media è più stabile della mediana, ovvero varia di meno al passare da un campione ad un altro.

Indici di variabilità o dispersione

Variabilità dei dati: tendenza delle singole osservazioni di una distribuzione di allontanarsi dalla tendenza centrale.

La dispersione esprime dunque la bontà (o la povertà) della tendenza centrale quale descrittore di una distribuzione.



Gamma: Max – Min. Molto sensibile a valori occasionali ed estremi (outliers), ma adatto per campioni di piccole dimensioni dove non è sempre possibile calcolare altre misure di dispersione (es. Distanza interquartilica).

Distanza interquartilica: la differenza tra il terzo e il primo quartile. Come per la mediana, si dimostra poco sensibile (robusta) ai valori estremi della distribuzione.

La misura di variabilità più importante è la **VARIANZA** (dati almeno su scala intervalli).

Consideriamo le due distribuzioni (funzioni di densità) continue in Figura 3

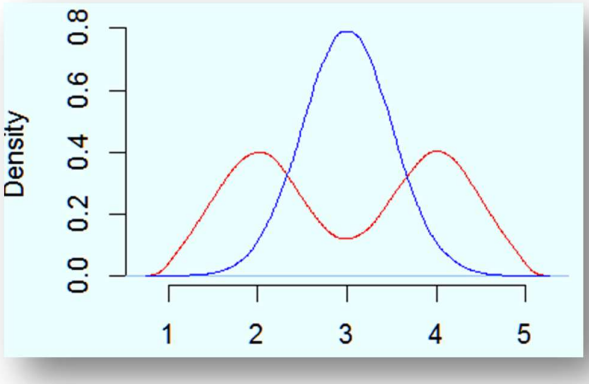


Figura 3 In rosso: bimodale (simmetrica); In blu: unimodale (simmetrica).

Le due distribuzioni hanno stessa media (3), mediana (3), gamma (circa 4 punti) e *distanza interquartilica* in rapporto di 3:1 (1.97 in rosso vs. 0.67 in blu).

La *varianza* rende meglio conto di quanto tipicamente i dati si allontanino dal centro: il rapporto è di 4.7:1 (1.17 in rosso vs. 0.25 in blu), si tratta infatti di una distanza (d_i) non di una posizione come la mediana ed i quartili.

Varianza: la media degli scarti dalla media al quadrato.

$$\boldsymbol{S^2 = \frac{\sum(X_i - \bar{X})^2}{n}}$$

Di cui conosciamo la formula alternativa (5.11 Luccio)

$$S^2 = \frac{\sum(X_i - \bar{X})^2}{n} = \sum \frac{(X_i)^2}{n} - \bar{X}^2$$

Ragionando invece nei termini della varianza di una variabile aleatoria (discreta):

$$\sigma^2 = E(X_i - \mu)^2 = E\left({X_i}^2 - 2X_i\mu + \mu^2\right) = E\left({X_i}^2\right) - 2\mu E(X_i) + \mu^2 = E\left({X_i}^2\right) - \mu^2$$

Deviazione standard: la radice quadrata della varianza.

$$\boldsymbol{S = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n}}}$$

Nota bene. Perché la varianza è calcolata nei termini degli scarti delle singole osservazioni della media, anziché da altre misure di tendenza centrale (moda, mediana, …)?

Si può facilmente dimostrare che *la media aritmetica è il centro di ordine 2* (confronta pg. 92 e 94 Luccio), rendendo minima la distanza (o scarto)

$$d = \sqrt{\sum (X_i - \tau)^2}$$

Infatti, la quantità sotto radice può essere riscritta come segue

$$\sum (X_i - \tau)^2 = \sum [(X_i - \overline{X}) + (\overline{X} - \tau)]^2$$

Sviluppando il quadrato abbiamo che

$$= \sum (X_i - \overline{X})^2 + 2(\overline{X} - \tau) \sum (X_i - \overline{X}) + n(\overline{X} - \tau)^2$$

$$= \sum (X_i - \overline{X})^2 + 2(\overline{X} - \tau)0 + n(\overline{X} - \tau)^2$$

$$= \sum (X_i - \overline{X})^2 + n(\overline{X} - \tau)^2$$

- Quando $\tau = \overline{X}$,

  allora $\sum (X_i - \tau)^2$ avrà il valore minimo possibile in $\sum (X_i - \overline{X})^2$.

- In tutti gli altri casi $\tau \neq \overline{X}$,

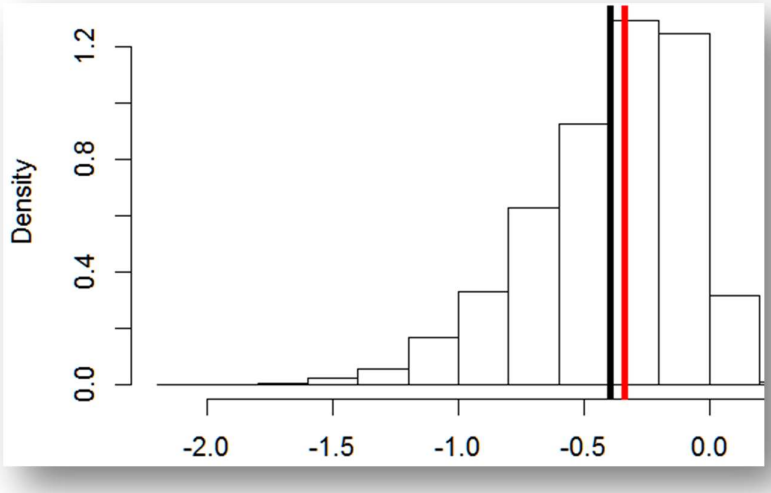
  verrà sempre aggiunta la quantità positiva $n(\overline{X} - \tau)^2$.

- **La varianza assume il minore valore possibile quando viene calcolata a partire dalla media campionaria, rispetto ad un qualsiasi numero arbitrario tau.**
- **La media campionaria è il valore più "vicino" a tutti gli altri punti del campione, rispetto ad un qualsiasi numero arbitrario tau. Rappresenta il miglior indice rispetto al quale misurar la "*tendenza delle singole osservazioni di una distribuzione di allontanarsi dalla tendenza centrale*".**

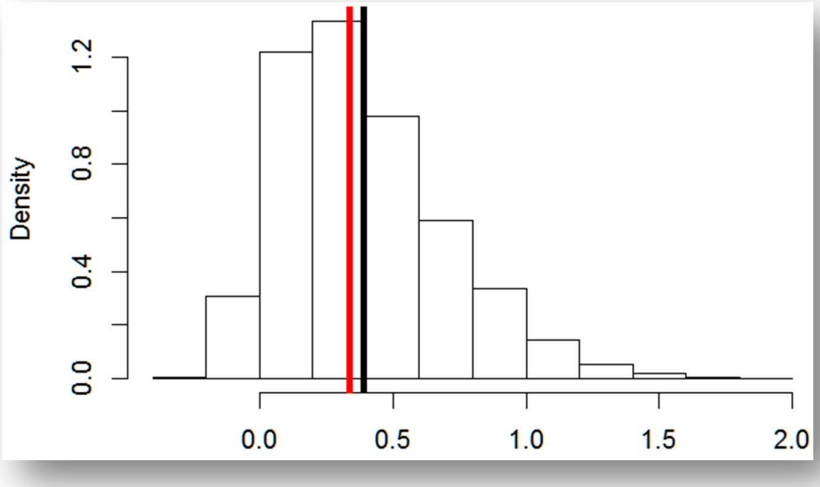
Indici di forma di una distribuzione

- Le distribuzioni di frequenza unimodali possono essere *simmetriche* (media, moda, mediana coincidenti), ma possono essere anche *asimmetriche* (schiacciate - *skewed*) positive o negative.

ASIMMETRIA NEGATIVA La mediana (rosso) è leggermente spostata sulla destra rispetto alla media campionaria (in nero).

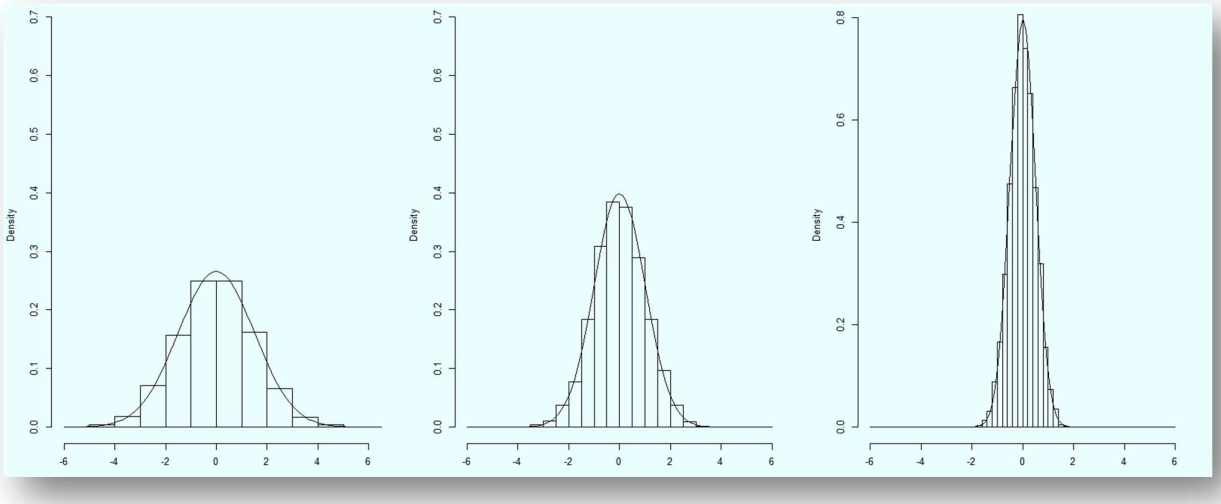


ASIMMETRIA POSITIVA La mediana (rosso) è spostata sulla sinistra, la media (nero) "segue" invece la coda destra.



- Possono avere diversi gradi di curtosi, raggruppandosi più o meno bene attorno alla tendenza centrale.

PLATICURTICA MESOCURTICA LEPTOCURTICA



$N(\mu = 0, \sigma = 1.5)$ $N(\mu = 0, \sigma = 1)$ $N(\mu = 0, \sigma = .5)$

Come si quantifica il grado di asimmetria e di curtosi di una distribuzione?

ASIMMETRIA G1 di Fisher

$$g_1 = \frac{\frac{1}{n}\sum(X_i - \bar{X})^3}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^{\frac{3}{2}}} = \frac{1}{n}\sum(X_i - \bar{X})^3 / s^3$$

E per **piccoli campioni**

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \frac{\frac{1}{n}\sum(X_i - \bar{X})^3}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^{\frac{3}{2}}}$$

Il fattore $\sqrt{n(n-1)}/(n-2)$ approssima il valore di 1 al crescere di n: 1.48 per *n* = 5, 1.19 per *n* = 10, 1.05 per *n* = 30, 1.03 per *n* = 60, …

Sono entrambi valori negativi in caso di asimmetria negativa, e positivi nel caso di asimmetria positiva.

Formula alternativa di Pearson, che chiarisce il rapporto tra media e mediana nel caso di asimmetria:

$$\mathrm{Sk}_2 = 3\frac{(\overline{\mathrm{X}} - Mediana)}{s}$$

CURTOSI g2 di Fisher

$$\mathrm{g}_2 = \frac{\frac{1}{n}\sum(\mathrm{X_i} - \overline{\mathrm{X}})^4}{\left[\frac{1}{n}\sum(\mathrm{X_i} - \overline{\mathrm{X}})^2\right]^2} = \frac{1}{n}\sum(\mathrm{X_i} - \overline{\mathrm{X}})^4/s^4$$

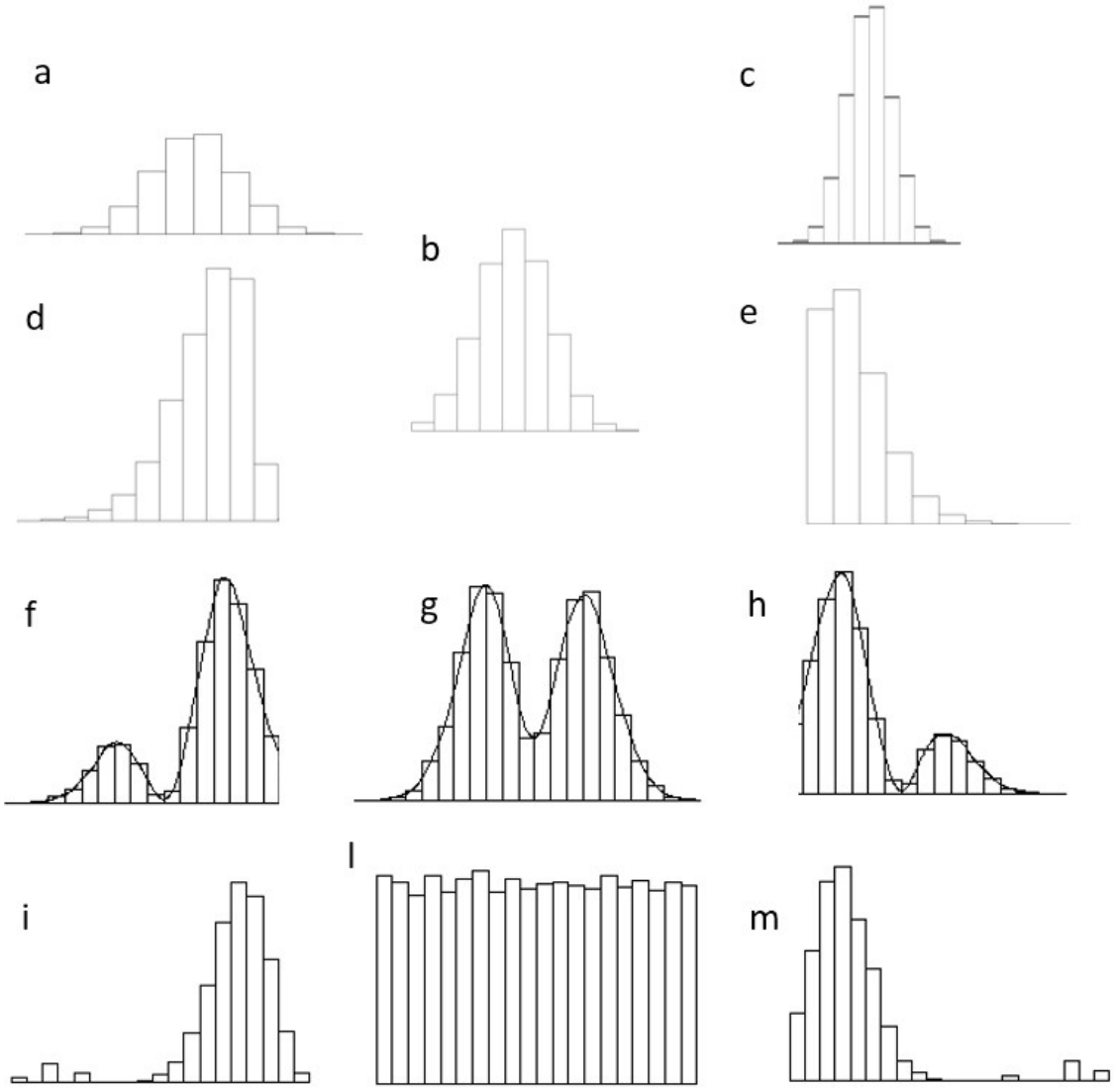
La distribuzione è normale ha valore 3, > 3 se è leptocurtica e < 3 se è platicurtica.

Alcuni manuali, compreso Luccio & Caudek, riportano l'indice come

$$\mathrm{g}_2 - 3.$$

In tal caso la distribuzione normale ha curtosi 0, mentre valori positivi o negativi definiscono rispettivamente distribuzioni leptocurtiche e platicurtiche.

Facciamo un po’ di esercizio.
Descrivere le distribuzioni nel disegno:



Punti z e scale standardizzate vedere il foglio Excel Lezione 10.