

# Calcolo delle probabilità

- Definizione di variabile aleatoria.
- Definizione di funzione di frequenza (densità) di probabilità per variabili aleatorie discrete.
- Definizione di funzione di distribuzione per variabili aleatorie discrete.
- Valore atteso (media) di una variabile aleatoria discreta.
- Valore atteso di una funzione di una variabile aleatoria discreta.
- Distribuzione binomiale.
- Esempi.

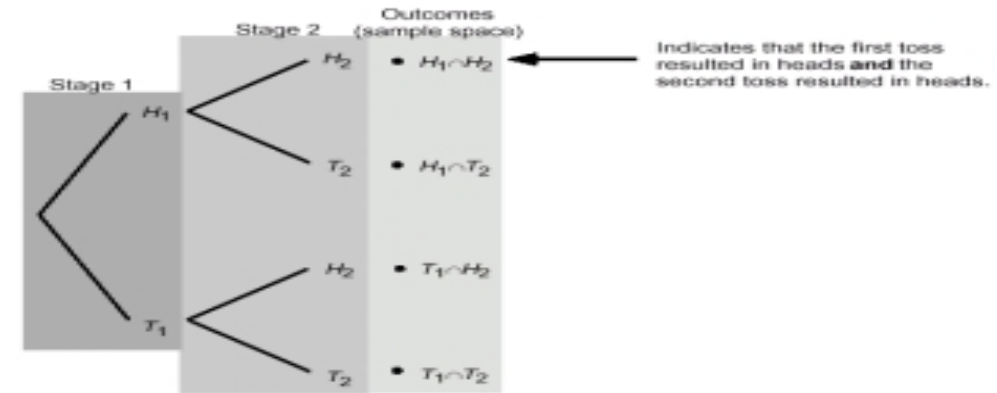


FIGURE 1.1 A tree diagram of the sample space for flipping a coin twice.

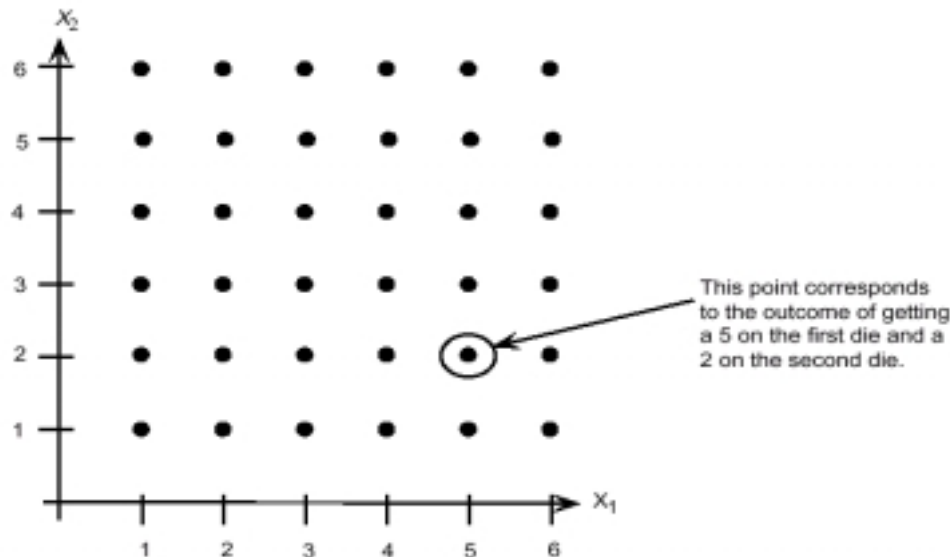


FIGURE 1.2 Coordinate system representation of a sample space for tossing two die.

## Variabili aleatorie (V.a.)

V.a.  $X$  e' una funzione reale degli eventi di uno spazio delle prove (S) probabilizzato

$$E \xrightarrow{X} X(E \subseteq S) = x \in D \subseteq \mathbb{R}$$

S Insieme di definizione di  $X$

$x=X(E)$  Realizzazione della variabile aleatoria  $X$

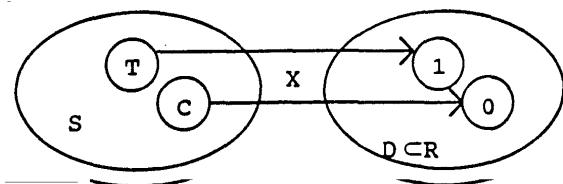
Se  $D$  è un insieme continuo v.a. continua  
(misura velocità, altezza)

Se  $D$  è un insieme discreto v.a. discreta  
(Lancio dado, moneta)

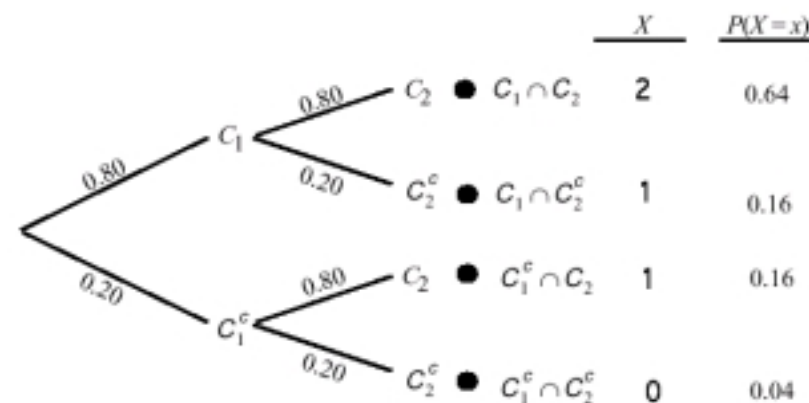
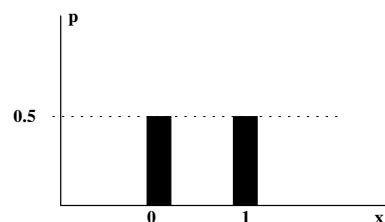
La probabilità di una realizzazione d'una variabile aleatoria è la probabilità che si verifichi l'evento ad essa associato

La probabilità di una realizzazione d'una variabile aleatoria (discreta) è la probabilità che si verifichi l'evento ad essa associato

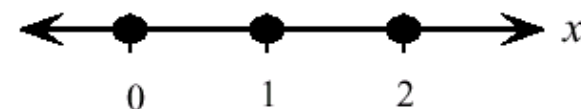
$$P[X(E) = x] = P(E)$$



$X(T)=1, P(T)=0.5$   
 $X(C)=0, P(C)=0.5$   
 $P(1)=0.5$   
 $P(0)=0.5$

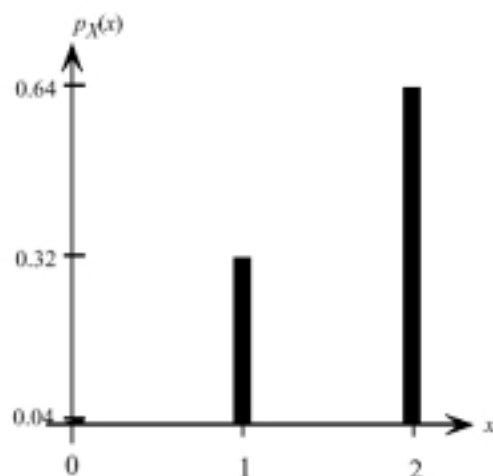


$$P(X = 0) = P(C_1^c \cap C_2^c) = (0.2)(0.2) = 0.04$$



$$p_X(x) = \begin{cases} 0.04 & \text{for } x = 0 \\ 0.32 & \text{for } x = 1 \\ 0.64 & \text{for } x = 2 \\ 0 & \text{elsewhere} \end{cases}$$

$x$	$p_X(x)$
0	0.04
1	0.32
2	0.64



$$\begin{aligned} P(X \geq 1) &= P(X = 1 \text{ or } X = 2) \\ &= P(X = 1) + P(X = 2) \\ &= 0.32 + 0.64 = \mathbf{0.96} \end{aligned}$$

- $0 \leq p_X(x) \leq 1$  for all  $x$  (from the axioms of probability)
- $\sum_x p_X(x) = 1$  (because each point in the sample space is assigned one and only one experimental value by the random variable).

### ESEMPIO DI FUNZIONE DI FREQUENZA ERRATA

$$q(x) = \begin{cases} \frac{1}{9}(x-2)^3 & x = 0, 1, 2, 3, 4, 5 \\ 0 & \text{elsewhere} \end{cases}$$

$$\begin{aligned} \sum_x q(x) &= \frac{1}{9}(0-2)^3 + \frac{1}{9}(1-2)^3 + \frac{1}{9}(2-2)^3 + \frac{1}{9}(3-2)^3 + \frac{1}{9}(4-2)^3 + \frac{1}{9}(5-2)^3 \\ &= -\frac{4}{9} - \frac{1}{9} + 0 + \frac{1}{9} + \frac{4}{9} + \frac{9}{9} = 1 \end{aligned}$$

QUALE DELLE DUE FUNZIONI E' UNA FUNZIONE DI FREQUENZA ?

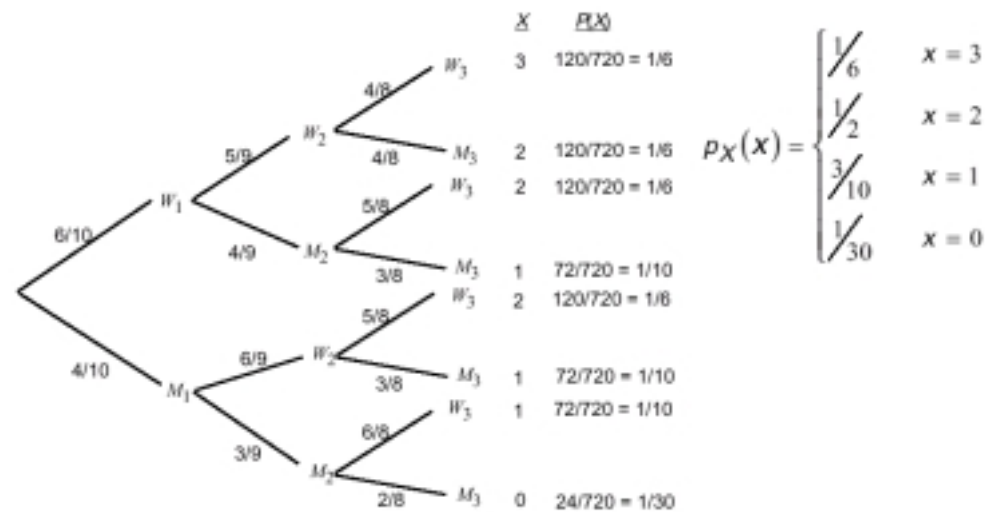
1

$$r(x) = \begin{cases} \frac{1}{9}(x-2)^2 & x = 0, 1, 2, 3, 4, 5 \\ 0 & \text{elsewhere} \end{cases}$$

2

$$p(x) = \begin{cases} \frac{1}{19}(x-2)^2 & x = 0, 1, 2, 3, 4, 5 \\ 0 & \text{elsewhere} \end{cases}$$

In un gruppo composto da 4 studentesse e 6 studenti. Tre sono scelti a caso per una prova alla lavagna. Quale sono le probabilità di avere 0,1,2 o 3 studenti.



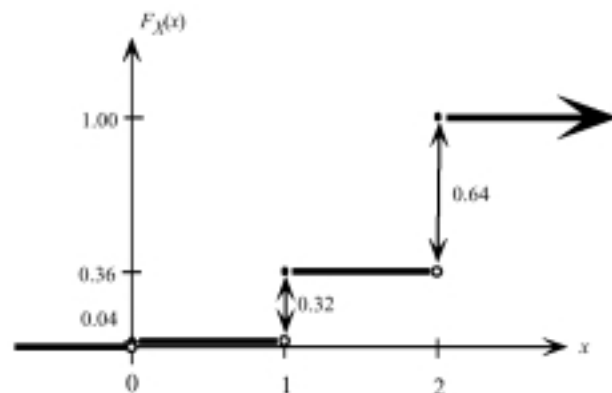
Funzione Distribuzione di Probabilità di V.a discrete

$$F_X(x) = P(X \leq x).$$

$$F_X(x) = \sum_{y=-\infty}^x p_X(y)$$

$$p_X(x) = \begin{cases} 0.04 & \text{for } x = 0 \\ 0.32 & \text{for } x = 1 \\ 0.64 & \text{for } x = 2 \\ 0 & \text{elsewhere} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.04 & 0 \leq x < 1 \\ 0.36 & 1 \leq x < 2 \\ 1.00 & x \geq 2 \end{cases}$$



Event	Formula for Probability of the Event
$\{X = a\}$	Height of jump of graph of $F_X(x)$ at $x = a$
$\{a < X\}$	$1 - F_X(a)$
$\{a \leq X\}$	$1 - F_X(a) + P(X = a)$
$\{X \leq b\}$	$F_X(b)$
$\{X < b\}$	$F_X(b) - P(X = b)$
$\{a < X \leq b\}$	$F_X(b) - F_X(a)$
$\{a < X < b\}$	$F_X(b) - F_X(a) - P(X = b)$
$\{a \leq X \leq b\}$	$F_X(b) - F_X(a) + P(X = a)$
$\{a \leq X < b\}$	$F_X(b) - F_X(a) + P(X = a) - P(X = b)$

**Property 1:** A CDF  $F_X(x)$  is always nondecreasing in  $x$ . That is,  $F_X(a) \leq F_X(b)$  whenever  $a \leq b$ .

**Property 2:** The values of  $F_X(x)$  always lie between 0 and 1; that is,  $0 \leq F_X(x) \leq 1$  for all  $x$ .

**Property 3:**  $F_X(x)$  approaches 0 as  $x$  becomes arbitrarily small, and  $F_X(x)$  approaches 1 when  $X$  becomes arbitrarily large. More formally,

$$\lim_{x \rightarrow -\infty} F_X(x) = 0, \quad \lim_{x \rightarrow \infty} F_X(x) = 1$$

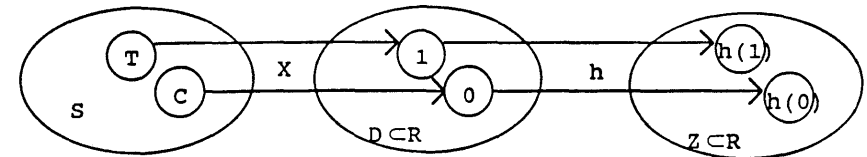
**Property 4:** For any number  $a$ , as  $x$  takes values decreasing to  $a$ , the value of  $F_X(x)$  will approach  $F_X(a)$ . That is,

$$\lim_{x \downarrow a} F_X(x) = F_X(a)$$

**Valore atteso (media) di una variabile aleatoria discreta:**

$$E(X) = \sum_x x p_X(x)$$

Data una variabile aleatoria  $X$ , è sempre possibile definire una funzione della variabile aleatoria  $X$   $g(X)$ , la funzione risulta essere a sua volta una funzione aleatoria.



SI DIMOSTRA:

$$E[g(X)] = \sum_x g(x) p_X(x)$$

### ESEMPIO

Gioco della roulette. Scommettiamo 1 Dollaro a Las Vegas su *dispari*.

Qual è la speranza matematica di vincita.?

$$p_X(1) = P(X=1) = P(\{1, 3, 5, \dots, 35\}) = \frac{18}{38} = \frac{9}{19}$$

$$p_X(-1) = P(X=-1) = 1 - \frac{9}{19} = \frac{10}{19}$$

Risposta

Vincita attesa =  $1 (9/19) + (-1) (10/19) = -0.053 \$$  (circa -100 lire)

### Distribuzione binomiale (distribuzione di Bernoulli)

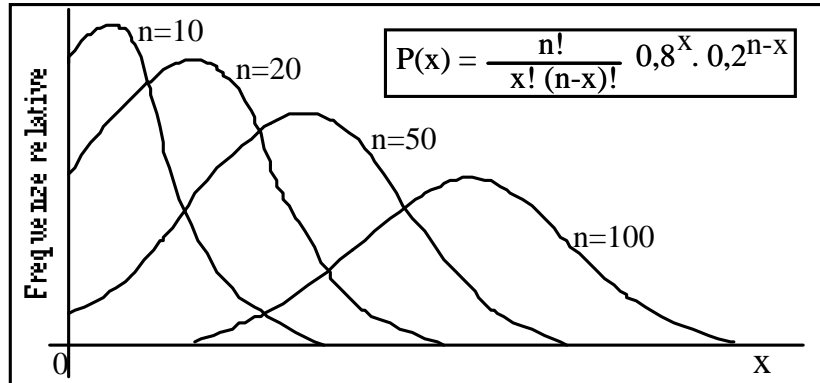
Se  $p$  denota la probabilità che si presenti un evento favorevole in una singola prova e  $q = 1 - p$  la probabilità dell'evento contrario.

La probabilità che l'evento favorevole si presenti esattamente  $x$  volte in una serie di  $n$  prove indipendenti è data dalla espressione seguente:

$$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x}.$$

La distribuzione binomiale o bernoulliana fornisce le risposte al problema delle prove ripetute, stima le probabilità che un evento, con probabilità a priori o frequentista  $p$ , si presenti rispettivamente 0, 1, 2, ...,  $i$ , ...,  $n$  volte, nel corso di  $n$  prove identiche ed indipendenti.

L'equazione della funzione binomiale mostra che la forma della curva binomiale dipende dai valori  $n$ ,  $p$  e  $q$ . Per valori di  $n$  molto grandi, anche quando  $p$  e  $q$  sono molto diversi tra loro, la curva binomiale assume una forma approssimativamente simmetrica, ma quando  $n$  è piccolo e  $p$  e  $q$  sono molto diversi tra loro, la curva appare asimmetrica, con asimmetria tanto maggiore quanto più diversi sono i valori di  $p$  e di  $q$  e quanto più piccolo è il numero delle osservazioni  $n$ .



## Media aritmetica della distribuzione binomiale

$$M = \frac{1}{N} \sum_{i=0}^n x_i \cdot f_i = \sum_{i=0}^n x_i \cdot P(x_i)$$

e poiché

$$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x}$$

si avrà che la media aritmetica è uguale a:

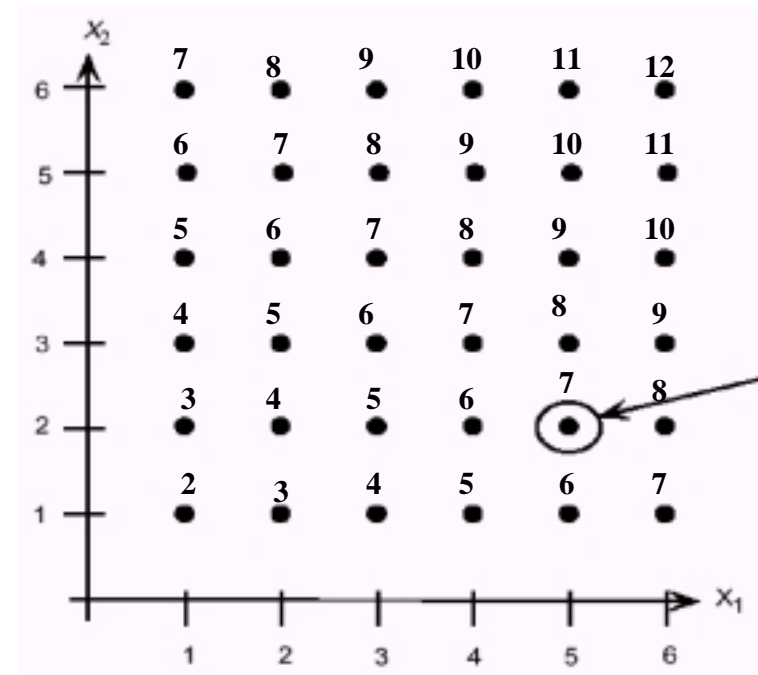
$$M = \sum_{x=0}^n \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x} \cdot x = np \cdot \sum_{x=0}^{n-1} \frac{(n-1)!}{(x-1)!(n-x)!} \cdot p^{x-1} \cdot q^{n-x} =$$

$$= np \cdot (p+q)^{n-1} = n \cdot p \quad (p+q)=1$$

La media aritmetica della distribuzione binomiale è quindi data dalla relazione  $M = n \cdot p$ .

## Calcolo delle probabilità

- Distribuzione binomiale.
- Momenti di una variabile aleatoria discreta
- Varianza e scarto quadratico medio di una variabile aleatoria discreta.
- Proprietà del valore atteso e della varianza
- Varianza della distribuzione binomiale
- Variabili aleatorie continue
- Esempi.



### 7 . Prove ripetute

Se è nota la probabilità che si verifichi un evento in una prova, la probabilità che lo stesso evento si presenti una volta, due volte, eccetera, in  $n$  prove è data dai termini successivi dello sviluppo binomiale.

Infatti, indichiamo con  $p$  la probabilità che si verifichi l'evento e con  $q$  la probabilità dell'evento contrario (cioè che l'evento atteso non si verifichi). Se selezioniamo un particolare insieme di  $r$  prove sul numero complessivo delle  $n$  prove, la probabilità che l'evento si verifichi in tutte queste  $r$  prove e non si presenti nelle altre  $(n-r)$  è:

$$p^r \cdot q^{n-r}$$

Ma un insieme di  $r$  prove può essere selezionato tra le  $n$  in  $C(n,r)$  modi diversi, tutti ugualmente possibili. Quindi la probabilità che l'evento si presenti in  $r$  prove è uguale a:

$$C(n,r) \cdot p^r \cdot q^{n-r} = \frac{n!}{r!(n-r)!} \cdot p^r \cdot q^{n-r}$$

La probabilità che l'evento si verifichi almeno  $r$  volte è uguale alla somma dei primi  $(n-r+1)$  termini.



La deviazione standard non è altro che la radice quadrata del momento secondo rispetto alla media ossia della varianza.

$$\sigma = \sqrt{\sigma^2}$$

Proprietà del valore atteso e della varianza

$$E(X+a) = E(X) + a$$

$$\sigma_{X+a}^2 = \sigma_X^2$$

Verification

$$\begin{aligned} E(X+a) &= \sum_x (x+a)p_X(x) = \sum_x xp_X(x) + \sum_x ap_X(x) \\ &= E(X) + a \sum_x p_X(x) = E(X) + a \end{aligned}$$

$$\begin{aligned} \sigma_{X+a}^2 &= E\{[(X+a) - E(X+a)]^2\} = E\{[X+a - E(X) - a]^2\} \\ &= E\{[X - E(X)]^2\} = \sigma_X^2 \end{aligned}$$

$$E(aX) = aE(X)$$

$$\sigma_{aX}^2 = a^2 \sigma_X^2$$

Verification

$$E(aX) = \sum_x a x p_X(x) = a \sum_x x p_X(x) = aE(X)$$

$$\begin{aligned} \sigma_{aX}^2 &= E\{[aX - E(aX)]^2\} = E\{[aX - aE(X)]^2\} = E\{a^2[X - E(X)]^2\} \\ &= a^2 E\{[X - E(X)]^2\} = a^2 \sigma_X^2 \end{aligned}$$

$$E(X+Y) = E(X) + E(Y)$$

$$\begin{aligned} \sigma_X^2 &= E\{[X - E(X)]^2\} = E\{X^2 - 2XE(X) + [E(X)]^2\} \\ &= E(X^2) - E(2XE(X)) + E\{[E(X)]^2\} \\ &= E(X^2) - 2E(X)E(X) + [E(X)]^2 \\ &= E(X^2) - [E(X)]^2 \end{aligned}$$



### Varianza della distribuzione binomiale

Per la determinazione del momento secondo rispetto all'origine valgono le seguenti relazioni:

$$M_2 = \sum_{i=0}^n x_i^2 \cdot P(x_i) = \sum_{x=0}^n x^2 \cdot \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x} =$$
$$= np + n^2 p^2 - np^2$$

Ma, poiché la varianza o momento secondo rispetto alla media, se espressa in funzione dei momenti rispetto all'origine, è data dalla relazione:

$$\sigma^2 = M_2 - M_1^2 =$$
$$= np + n^2 p^2 - np^2 - n^2 p^2 =$$
$$= np - np^2 = np(1-p) = npq$$

come risultato si ottiene per la deviazione standard la relazione:

$$\sigma = \sqrt{\sigma^2} = \sqrt{npq}.$$

Nella distribuzione binomiale la varianza è inferiore alla media; infatti essa è uguale alla media  $n \cdot p$  moltiplicata per un numero che è inferiore all'unità [  $q=(1-p)$  ] .

### V.A. CONTINUE

Per le v.a continue:  $\text{Prob}[x = x] = 0$

*"Non è possibile estendere la nozione di funzione di probabilità introdotta nel caso delle variabili discrete"*

Si definisce preliminarmente la nozione di Funzione di Distribuzione di Probabilità, analogamente al caso di v.a. discrete:

$$F_x(\bar{x}) = \text{Prob}[X \leq \bar{x}]$$

(la Funzione di Distribuzione di Probabilità è sempre monotona crescente)

Si definisce indirettamente la Funzione di Densità di Probabilità come:

$$f_X(x) = \frac{d}{dx} F_X(x)$$

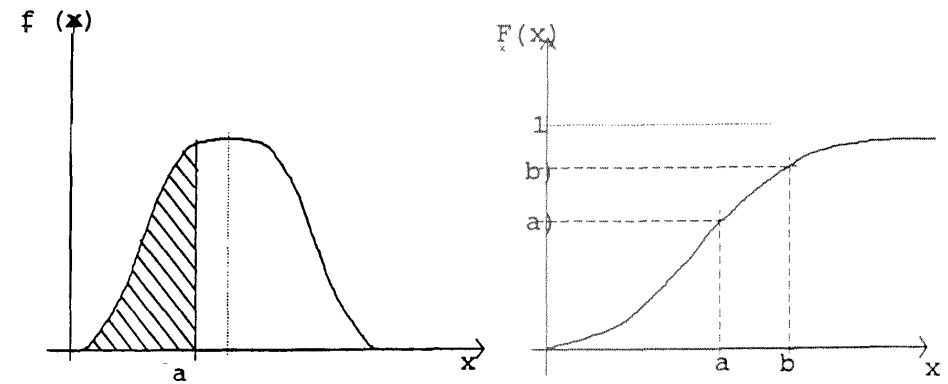
Si noti che  $f_X(x) \geq 0 \quad \forall x$

Ovviamente, da un punto di vista operativo, se è nota la  $f_X(x)$  risulta:

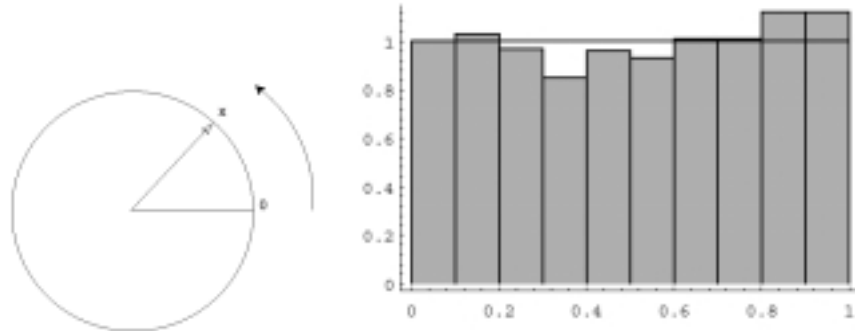
$$F_X(a) = \int_{-\infty}^a f_X(x) dx \quad \text{Per def. di prob.} \Rightarrow F_X(\infty) = \int_{-\infty}^{\infty} f_X(x) dx = 1$$

Si dimostra che:

$$P(a < X \leq b) = F_X(b) - F_X(a)$$



## Variabile uniforme



$$P(0 \leq X \leq 1) \quad P\left(0 \leq X < \frac{1}{2}\right) = P\left(\frac{1}{2} \leq X < 1\right) = \frac{1}{2}$$

$$P(c \leq X < d) = d - c$$

$$P(E) = \int_E f(x) dx \quad P(a \leq X \leq b) = \int_a^b f(x) dx$$

## Funzione densità di probabilità

$$f(x) = \begin{cases} 1, & \text{if } 0 \leq x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

$$P\left(0 \leq X < \frac{1}{2}\right) = P\left(\frac{1}{2} \leq X < 1\right) = \frac{1}{2}$$

$$P(E) = \int_0^{1/2} 1 dx = \frac{1}{2}$$

$$P(E) = \int_a^b 1 dx = b - a$$

## Funzione di distribuzione di probabilità

$$F_X(x) = P(X \leq x) \quad F(x) = \int_{-\infty}^x f(t) dt$$

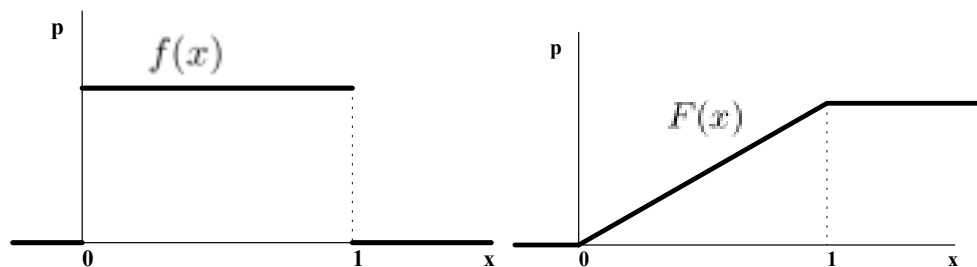
## Funzioni densità e distribuzione di probabilità

$$F_X(x) = P(X \leq x) \quad F(x) = \int_{-\infty}^x f(t) dt$$

$$\frac{d}{dx} F(x) = f(x)$$

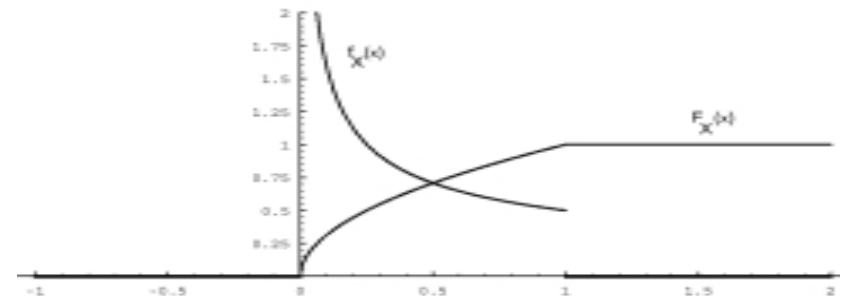
$$F(x)=x \quad f(x)=1 \quad 0 \leq x \leq 1$$

## Variabile uniforme



## Variabile $X = U^2$ (U=var. unif.)

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= P(U^2 \leq x) \\ &= P(U \leq \sqrt{x}) \\ &= \sqrt{x}. \end{aligned} \quad F_X(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ \sqrt{x}, & \text{if } 0 \leq x \leq 1, \\ 1, & \text{if } x \geq 1. \end{cases}$$



$$f_X(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ 1/(2\sqrt{x}), & \text{if } 0 \leq x \leq 1, \\ 0, & \text{if } x > 1. \end{cases}$$

**Valore atteso variabili discrete**

$$E(X) = \sum_x x p_X(x)$$

**Valore atteso lancio del dado**

$$\mu = 1\left(\frac{1}{6}\right) - 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) - 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) - 6\left(\frac{1}{6}\right)$$

**Varianza lancio del dado**

$x$	$m(x)$	$(x - 7/2)^2$
1	1/6	25/4
2	1/6	9/4
3	1/6	1/4
4	1/6	1/4
5	1/6	9/4
6	1/6	25/4

$$E((X - \mu)^2)$$

$$V(X) = \frac{1}{6} \left( \frac{25}{4} + \frac{9}{4} + \frac{1}{4} + \frac{1}{4} + \frac{9}{4} + \frac{25}{4} \right)$$

$$= \frac{35}{12},$$

$$D(X) = \sqrt{35/12} \approx 1.707$$

**Varianza lancio del dado come differenza momenti**

$$V(X) = E(X^2) - \mu^2$$

$$E(X^2) = 1\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 9\left(\frac{1}{6}\right) + 16\left(\frac{1}{6}\right) + 25\left(\frac{1}{6}\right) + 36\left(\frac{1}{6}\right)$$

$$= \frac{91}{6},$$

$$V(X) = E(X^2) - \mu^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

**Valore atteso variabili continue**

$$\mu = E(X) = \int_{-\infty}^{+\infty} x f(x) dx$$

**Valore atteso variabile uniforme**

$$E(X) = \int_0^1 x dx = 1/2$$

**Varianza variabile uniforme**

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

$$V(X) = \int_0^1 \left(x - \frac{1}{2}\right)^2 dx = \frac{1}{12}$$

<sup>1</sup>

## STATISTICA DESCRITTIVA

### Temi considerati

- 1) Aspetti generali
- 2) Distribuzioni statistiche
- 3) Rappresentazioni grafiche
- 4) Misure di tendenza centrale
- 5) Medie ferme o basali
- 6) Medie lasche o di posizione
- 7 ) Dispersione o variabilità
- 8 ) Forma della distribuzione

### Elementi forniti

- 1) Definizione generale
- 2) Esempi

<sup>2</sup>

## Aspetti generali

*Statistica*                      Tecnica speciale per lo studio quantitativo dei fenomeni di massa o collettivi

*Fenomeno di massa*    Fenomeno la cui misura richiede una massa o collezione di osservazioni di altri fenomeni più semplici detti fenomeni singoli o individuali

*Unità statistica*            Risultato di una osservazione sopra uno dei fenomeni individuali da cui risulta il fenomeno collettivo

*Dato statistico*            Risultato di un'operazione compiuta sopra le unità statistiche relative ai fenomeni individuali che rientrano in un fenomeno collettivo

<sup>3</sup>

## Aspetti generali

*Modalità* Qualunque modo di manifestarsi di un fenomeno, espresso in termini o quantitativi o qualitativi

*Variabile statistica*    Qualunque fenomeno che si manifesta o è capace di manifestarsi in almeno due modi diversi, espressi in termini quantitativi o qualitativi. A volte, quando il fenomeno è un carattere qualitativo si usa indicarlo come Mutabile statistica

*Intensità* Dato numerico che esprime una modalità quantitativa di un fenomeno

*Frequenza* Numero di unità statistiche portatrici di una stessa modalità del carattere o fenomeno allo studio

<sup>4</sup>

## Aspetti generali

*Popolazione o universo*

Insieme delle unità statistiche portatrici di un dato carattere o fenomeno che si manifesta secondo due o più modalità differenti e che si vuole studiare o in relazione allo stesso carattere o in relazione ad un altro fenomeno

*Campione*

Porzione della popolazione o universo che interessa, estratta seguendo criteri prefissati e che viene effettivamente sottoposta a rilevazione e ad analisi

5

### **Contenuto della statistica moderna**

**Raccolta, presentazione ed elaborazione numerica delle informazioni, per agevolare l'analisi dei dati ed i processi decisionali**

#### **Statistica descrittiva**

**Insieme dei metodi che riguardano la raccolta, la presentazione e la sintesi di un insieme di dati per descriverne le caratteristiche essenziali**

#### **Statistica inferenziale**

**Insieme dei metodi con cui si possono elaborare i dati dei campioni allo scopo di dedurre omogeneità o differenze nelle caratteristiche analizzate, al fine di estendere le conclusioni alla popolazione**

7

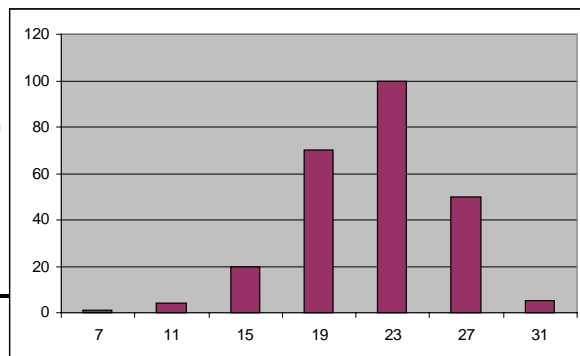
**Consideriamo un esempio di distribuzione di frequenza. Supponiamo di avere un gruppo di persone che classifichiamo secondo la loro età, con il seguente risultato:**

Limiti della classe (in anni)	Modalità o indice della classe : $x_i$	Frequenza o numero dei casi della classe: $f_i$
5 - 9	7	1
9 - 13	11	4
13 - 17	15	20
17 - 21	19	70
21 - 25	23	100
25 - 29	27	50
29 - 33	31	5
		<u>250</u>

**Intervallo di classe: 4 anni**

**Numero delle classi: 7**

**Valore centrale di classe: semisomma estremi**



6

### **Aspetti generali**

#### **Classe o intervallo**

**Raggruppamento di modalità ordinate di un dato carattere comprese tra un valore inferiore ed uno superiore chiamati *limiti di classe***

#### **Distribuzione di frequenza**

**Raggruppamento di una serie di dati in classi, contando quanti valori o unità statistiche appartengono ad ogni gruppo o classe di modalità**

8

### **Distribuzioni statistiche**

#### **Caratteristiche**

**1 - Tendenza centrale**

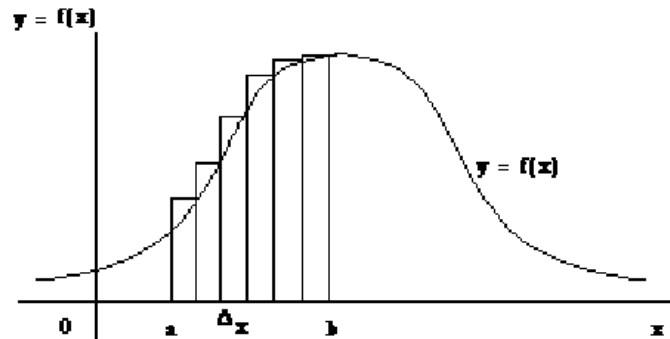
**2 - Dispersione o variabilità**

**3 - Forma della distribuzione**

**Distribuzioni discontinue o discrete** Sono quelle relative a variabili per loro natura discontinue, perché possono assumere solo valori discontinui (isolati), cioè numeri interi e non frazioni di numero. Il suo grafico può essere rappresentato da un istogramma

**Distribuzioni continue** Sono relative a variabili continue caratterizzate da valori compresi in certi intervalli di variazione al loro interno continui, cioè costituiti da infiniti valori. Il suo grafico può essere rappresentato da una curva continua

## 2 Distribuzioni statistiche continue e discrete



$$\int_a^b f(x) \cdot dx = P(a < x < b).$$

$$\int_{-\infty}^{+\infty} f(x) \cdot dx = 1.$$

$$\lim_{\Delta x \rightarrow 0} \sum_{i=1}^{i=n} f_i \cdot \Delta x = \int_a^b f(x) \cdot dx$$

10

## Rappresentazioni grafiche

**Scopo** Evidenziare in modo semplice, *a colpo d'occhio*, le caratteristiche fondamentali di una distribuzione di frequenza, cioè: tendenza centrale, variabilità e forma

### Inconvenienti

Mancano di precisione, sono soggettive e permettono letture diverse degli stessi dati

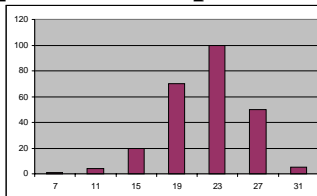
11

## Rappresentazioni grafiche

### Tipi più frequenti per fenomeni discreti

#### - Istogramma

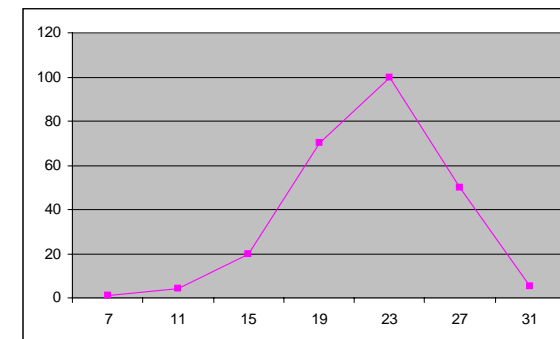
E' un grafico a barre verticali i cui rettangoli vengono costruiti in corrispondenza degli estremi di ciascuna classe. Le misure della variabile o fenomeno sono riportate lungo l'asse orizzontale, la frequenza o numero dei casi in cui si presenta ciascuna classe di misure lungo l'asse verticale. Le superfici dei vari rettangoli sono proporzionali alle frequenze corrispondenti



12

#### - Poligono di frequenza

Figura simile all'istogramma che è di solito utilizzata per rappresentare valori relativi o percentuali. Può essere ottenuto dall'istogramma corrispondente, unendo con una spezzata le frequenze relative ai punti centrali di ogni classe



13

## Rappresentazioni grafiche

### Tipi più usuali per fenomeni qualitativi

#### - *Diagramma a rettangoli distanziati*

E' un grafico a colonne, formato da rettangoli con basi uguali ed altezze proporzionali alle frequenze dei vari gruppi

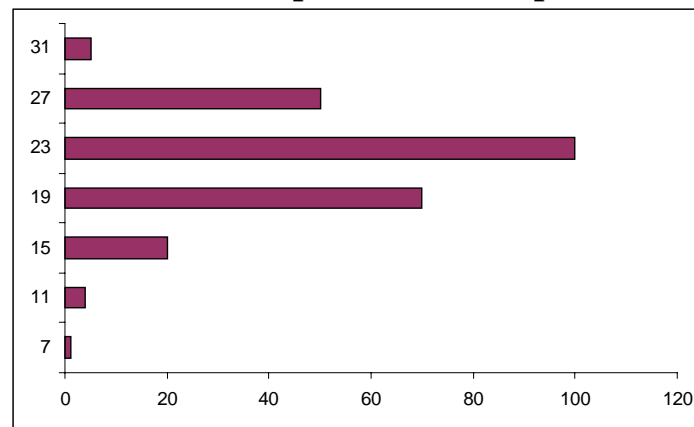
#### - *Diagramma a barre*

Al posto dei rettangoli utilizza linee continue più o meno spesse

14

#### - *Ortogramma o grafico a nastri*

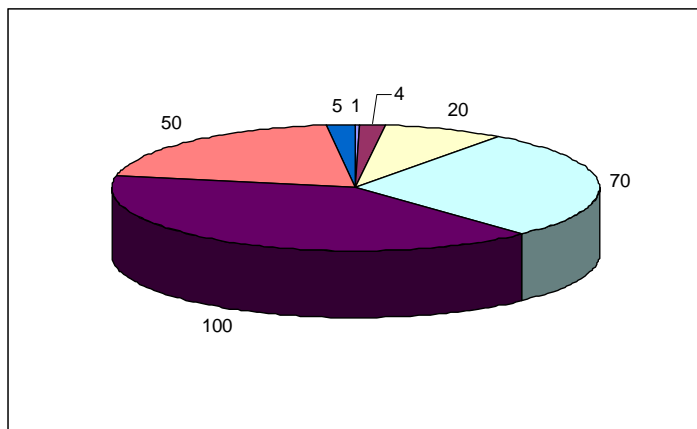
E' uguale ai rettangoli distanziati, ma gli assi sono scambiati per una lettura più facile



15

#### - *Diagramma circolare o a torta*

Si divide un cerchio in parti proporzionali alle frequenze o alla intensità di ciascuna modalità



16

#### - *Diagramma a figure*

La frequenza di ogni modalità viene rappresentata da una figura o da simboli che ricordano facilmente l'oggetto, la cui altezza è proporzionale alla frequenza o alla intensità

#### - *Cartogramma*

Evidenzia distribuzioni territoriali mediante carte geografiche in cui in alcune località sono riportati cerchi proporzionali alle frequenze o alle intensità del fenomeno



## Valore centrale o media

E' una misura attraverso la quale si cerca di *sintetizzare* la variabile descritta dall'intera distribuzione di frequenza, attorno alla quale tendono a distribuirsi tutti gli altri valori della variabile compresi nell'intervallo o campo di definizione del fenomeno osservato e ne rappresenta l'ordine di grandezza

## Condizioni generali a cui deve rispondere un valore centrale.

Esso deve:

- essere definito in maniera obiettiva
- dipendere da tutte le osservazioni della serie
- avere un significato concreto e facile da concepire
- essere semplice a calcolarsi
- essere poco sensibile alle fluttuazioni del campionamento
- prestarsi con facilità ai calcoli algebrici ulteriori

(Queste condizioni non sono sempre rispettate dalle varie medie)

## Misure di tendenza centrale

Il valore medio può essere uno dei valori realmente osservati della serie oppure qualunque altro valore intermedio che non figura effettivamente tra le osservazioni.

I valori medi possono essere raggruppati in due insiemi principali: quello delle medie ferme o basali e quello delle medie lasche o di posizione.

### *Medie ferme o basali*

Sono quelle suscettibili di essere determinate in funzione di tutti i termini della serie a cui si riferiscono

### *Media lasche o di posizione*

Sono quelle il cui valore non dipende da tutti i termini della serie a cui si riferiscono, ma dalla posizione che esse occupano nella serie

## Medie ferme o basali

### *Media aritmetica*

La *media aritmetica* di una variabile statistica X è definita come la somma dei valori  $x_i$  assunti dalla variabile in tutti gli  $f_i$  casi rilevati, divisa per il numero totale degli stessi casi. In formule:

per la semplice: 
$$M_1 = \frac{\sum_{i=1}^{i=N} X_i}{N},$$
 con  $i = 1, 2, \dots, N.$

per la ponderata: 
$$M_1 = \frac{\sum_{i=1}^{i=n} x_i f_i}{\sum_{i=1}^{i=n} f_i},$$
 con  $\sum_{i=1}^{i=n} f_i = N.$

## Misure di tendenza centrale

### Proprietà della media aritmetica

- 1) La *somma algebrica* degli scostamenti dalla media aritmetica di una distribuzione è *uguale a zero*
- 2) La *somma dei quadrati* degli scostamenti dalla media aritmetica di una distribuzione è *un minimo*
- 3) La media aritmetica di una successione  $a_1, a_2, \dots, a_n$ , con pesi  $f_1, f_2, \dots, f_n$  è l'*ascissa del baricentro* di un sistema di forze di gravità di intensità uguale alle  $f_i$ , applicate nei punti di ascissa  $a_i$  di un asse orizzontale.
- 4) Si applica di preferenza a *grandezze additive*
- 5) E' particolarmente utile per il fatto che essa effettua la correzione degli errori accidentali di osservazione, per cui essa è la *stima più precisa di misure ripetute*.
- 6) E' la *più semplice* tra le medie algebriche.

Esempio III.2 Si abbia la distribuzione di una data popolazione di individui secondo la statura, descritta nella tavola e si faccia l'ipotesi che le intensità di ogni classe di statura siano concentrate nel valore centrale; per le classi estreme aperte, si ipotizza che il valore centrale sia rispettivamente di 145 e 185 cm. I calcoli per determinare la media aritmetica ponderata della distribuzione sono i seguenti.

Classi di statura (cm)	Valori centrali delle classi ( $x_i$ )	Frequenze ( $n_i$ )	Prodotti ( $n_i x_i$ )
meno di 150	145	1.500	217.500
150   --160	155	41.200	6.386.000
160   --170	165	18.830	3.106.950
170   --180	175	125.900	22.032.500
180 e oltre	185	14.500	2.682.500
		<u>371.400</u>	<u>62.388.000</u>

Pertanto la media aritmetica ponderata sarà:

$$\bar{x} = \frac{62.388.000}{371.400} = 167,98.$$

### *Media geometrica*

La *media geometrica* di una variabile statistica X che assume il valore  $x_i$  con frequenza  $f_i$  (con  $i=1,2,...,n$ ) tale che la somma delle  $f_i$  sia pari a N osservazioni complessive è definita come la radice di ordine N del prodotto degli n termini  $x_i$  ciascuno elevato alla potenza  $f_i$ . In formule si avrà:

per la semplice  $M_0 = \sqrt[N]{\prod_{i=1}^N x_i}$ , dove  $i = 1, 2, \dots, N$ .

per la ponderata  $M_0 = \sqrt[N]{\prod_{i=1}^n x_i^{f_i}}$ , dove  $\sum_{i=1}^n f_i = N$ .

### *Proprietà della media geometrica*

- 1) Il reciproco della media geometrica è uguale alla media geometrica dei reciproci dei termini.
- 2) La potenza emmesima della media geometrica è uguale alla media geometrica delle potenze emmesime dei termini.
- 3) Il logaritmo della media geometrica è uguale alla media aritmetica dei logaritmi dei termini.
- 4) La media geometrica è utilizzata quando le variabili non sono rappresentate da valori ottenuti come prodotto o rapporto tra valori lineari. Serve per il confronto di superfici o volumi, di tassi di variazione, cioè valori che sono espressi da rapporti.
- 5) Per il calcolo della media geometrica è condizione necessaria che le quantità siano tutte positive.

**Esempio** Supponiamo di impiegare 1 lira ad interesse composto ai seguenti tassi :  $i_1=0,05$  nel primo anno;  $i_2=0,06$  nel secondo anno;  $i_3=0,055$  nel terzo anno;  $i_4=0,07$  nel quarto anno;  $i_5=0,065$  nel quinto anno. Il montante alla fine del primo anno sarà dato dalla relazione  $C_1=1+0,05$ ; alla fine del secondo anno sarà  $C_2=(1+0,05)(1+0,06)$ ; alla fine del terzo anno sarà  $C_3=(1+0,05)(1+0,06)(1+0,055)$  e così via.

Ci si chiede qual'è il tasso medio  $i$  a cui capitalizzare la nostra lira per ottenere alla fine del quinquennio il montante  $C_5$  che rappresenta, evidentemente, l'invariante del problema.

Da quanto detto, si deduce che deve essere:

$$(1+i)^5 = 1,05 \times 1,06 \times 1,055 \times 1,07 \times 1,065$$

da cui si vede che  $(1+i)$  è la media geometrica dei prodotti indicati nel secondo membro e non dei singoli tassi annui, per cui, mediante i logaritmi si calcola:

$$\log(1+i) = \frac{1}{5} (\log 1,05 + \log 1,06 + \log 1,055 +$$

$$+ \log 1,07 + \log 1,065 = 0,025296$$

Risalendo al numero, si ha che  $i = 0,0599$ , pari a 5,99%

**Media armonica** La media armonica è definita come il reciproco della media aritmetica dei reciproci dei termini. E' quindi anche vero che il reciproco della media armonica è la media aritmetica dei reciproci dei termini. In formule:

per la semplice 
$$M_{-1} = \frac{N}{\sum_{i=1}^N \frac{1}{X_i}}$$
, dove i = 1, 2, ....., N.

per la ponderata 
$$M_{-1} = \frac{\sum_{i=1}^n f_i}{\sum_{i=1}^n \frac{f_i}{X_i}}$$
, dove  $\sum_{i=1}^n f_i = N$ .

La media armonica è la stima più corretta della tendenza centrale per distribuzioni di dati in cui devono essere usati gli *inversi o reciproci*, come nel caso di misure dei tempi di reazione.

**Esempio** Si voglia conoscere il consumo medio annuo di rasoi usa-e-getta in Italia, mediante una ricerca diretta sui consumatori.

Non sarà opportuno chiedere: " Quanti rasoi consuma in media all'anno?" perché la domanda così formulata richiede una stima relativa ad un ampio intervallo di tempo; si potrà invece chiedere: " Quanti giorni le dura in media un rasoio?". Immaginiamo di esaminare le risposte di cinque persone:

1a persona	10 giorni in media
2a persona	6 giorni in media
3a persona	30 giorni in media
4a persona	5 giorni in media
5a persona	14 giorni in media
Totale	65

La media aritmetica delle durate è 65:5=13 giorni. Ma da questo dato non è corretto ricavare il consumo medio annuo pari a 365:13=28 ,1 rasoi in media per persona, equivalente per i 5 consumatori considerati a 28 ,1x5=140,5 rasoi di consumo annuo. Infatti con i dati di partenza possiamo ricavare direttamente il consumo globale

Persone	Consumo annuo di rasoi
1a	365:10=36,5
2a	365: 6=60,8
3a	365:30=12,2
4a	365: 5=73,0
5a	<u>365:14=26,1</u>
In complesso	208 ,6 rasoi

mentre in precedenza si era ottenuto il risultato di 140,5 rasoi. Con l'ultimo risultato il consumo pro-capite è 208 ,6:5=41,7 rasoi e la durata media 365:41,7 =8 ,8 giorni. Questo valore si ottiene immediatamente come media armonica dei dati iniziali:

$$M_{-1} = \frac{5}{\frac{1}{10} + \frac{1}{6} + \frac{1}{30} + \frac{1}{5} + \frac{1}{14}} = 8,8.$$

Per comprendere il motivo per il quale si deve adoperare la media armonica e non quella aritmetica delle durate, occorre osservare che il problema riguarda il *consumo*, per cui si deve tenere conto che la prima persona consuma in un giorno 1/10 di rasoio, la seconda consuma 1/6 di rasoio e così via, per cui, nel

complesso, le cinque persone consumano in un giorno la somma delle quantità suddette.

Il valore unico  $1/\bar{x}$  da sostituire a questi consumi diversi, lasciando invariato il consumo complessivo delle 5 persone è dato, perciò, dalla equazione:

$$5 \frac{1}{\bar{x}} = \frac{1}{10} + \frac{1}{6} + \frac{1}{30} + \frac{1}{5} + \frac{1}{14}$$

da cui si ricava la durata media facendo l'inverso della media dei consumi, cioè proprio la media armonica. Avendo quindi rilevato la durata dei rasoi invece del consumo, bisogna tener conto che tra queste due quantità esiste una relazione inversa.

Formula generale per le medie ottenibili a calcolo

*Media di potenze*

La media di potenze di indice  $r$  di una variabile statistica che si presenta con  $n$  modalità differenti, ciascuna avente frequenza  $f_i$ , è quel valore che si ottiene considerando la radice di ordine  $r$  della media aritmetica delle potenze  $r$ -esime delle singole determinazioni. In simboli:

$$M_r = \sqrt[r]{\frac{\sum_{i=1}^n x_i^r f_i}{\sum_{i=1}^n f_i}}$$

La media di potenze si definirà semplice o ponderata, a seconda che le frequenze siano tutte uguali all'unità oppure tra loro diverse.

*Media di potenze di ordine  $r$*

$M_r$  è una funzione continua e crescente con  $r$  e in statistica definisce la Media di potenza di ordine  $r$  della variabile considerata. Essa è pari alla radice  $r$ -esima del Momento di ordine  $r$  rispetto all'origine.

Per  $r = -1$ , la media di potenze è uguale alla media armonica

~~Per  $r = 0$ , la media di potenze è uguale alla media geometrica~~

Per  $r = 1$ , la media di potenze è uguale alla media aritmetica

Per  $r = 2$ , la media di potenze è uguale alla media quadratica

*media armonica < media geometrica*

*media geometrica < media aritmetica*

Medie lasche o di posizione

*Mediana* Se le unità statistiche sono in ordine crescente dei valori della variabile, la mediana è quel valore al di sotto ed al di sopra del quale si situa la metà del numero totale dei casi, cioè divide l'insieme delle unità in due parti di uguale frequenza.

Se il numero dei termini è dispari, la mediana è il valore relativo al termine che occupa il posto di mezzo; se è pari essa è uguale alla media aritmetica dei valori relativi ai due termini che occupano i posti centrali.

Se i dati raggruppati in  $n$  classi, la mediana si calcola con la formula:

$$M_e = Cl_i + \frac{\left[ \left( \frac{N}{2} - \sum_{k=1}^{i-1} f_k \right) \cdot A_i \right]}{f_i}, \text{ con } \sum_{i=1}^n f_i = N,$$

dove  $Cl_i$  è il confine inferiore della classe  $i$ ,  $A_i$  è la sua ampiezza e  $f_i$  è la frequenza della classe.

### *Proprietà della mediana*

- 1) Il numero degli scostamenti positivi è uguale al numero degli scostamenti negativi. Quindi, in una distribuzione o serie di dati ogni valore estratto a caso ha la stessa probabilità di essere inferiore o superiore alla mediana.
- 2) La somma dei valori assoluti degli scostamenti è un minimo in confronto alla somma dei valori assoluti degli scostamenti cui darebbe luogo un altro valore medio qualsiasi diverso dal valore mediano.
- 3) E' una misura robusta: non è influenzata dalla presenza di dati anomali e in particolare dai valori estremi, ma soltanto dal numero delle osservazioni
- 4) La mediana è la misura di posizione o di tendenza centrale utilizzata in quasi tutti i tests non parametrici.

**Esempio** Nella seguente successione di numeri 5, 9, 6, 14, 11, il valore mediano è il 9. Nella seguente 5, 6, 9, 11, 14, 18 formata da un numero pari di termini la mediana è  $(9+11)/2=10$ .

Nel caso di variabili statistiche divise in intervalli, il metodo migliore è quello di costruire la distribuzione cumulativa delle frequenze. Ad esempio, consideriamo la distribuzione:

Classi	$n_i$	$N_i = \sum n_k$
50  --100	110	110
100  --200	400	510
200  --300	90	600
		$N = 600$

La mediana corrisponde alla modalità del l'unità che occupa il posto  $600/2=300$ , quindi si tratta di un valore interno alla classe 100| --200.

Per individuarlo, si fa l'ipotesi di uniforme distribuzione delle unità all'interno della classe e si considera la proporzione:

$$(M_{.1} - 100) : (200 - 100) = (300 - 110) : 400$$

dalla quale si ricava la mediana

$$M_{.1} = 190 \times 100 / 400 + 100 = 147,5$$

che risulta leggermente inferiore al valore centrale della classe 100| --200 alla quale apparteneva il valore che lasciava da una parte e dall'altra lo stesso numero di termini.

### *Quartili*

Il *primo quartile* di una successione di termini non decrescente è quella quantità al di sotto della quale sta 1/4 ed al di sopra della quale stanno i 3/4 dei valori dati.

Il *secondo quartile* coincide con la mediana della distribuzione. Mentre il *terzo quartile* è quella quantità al di sotto della quale stanno i 3/4 ed al di sopra della quale sta 1/4 dei valori dati.

*Il primo quartile è la mediana dei valori inferiori alla mediana e il terzo quartile è il valore mediano dei valori superiori alla mediana della distribuzione.*

### Moda o norma

Si chiama *moda* o *norma* o *valore modale* quella modalità della variabile che si presenta con la frequenza più elevata e la classe in cui essa risulta compresa si chiama *classe modale*. Nel caso di dati raggruppati per classi, se  $i$  è la classe modale, si può calcolare la moda in base alla formula

$$M_d = Cl_i + \frac{[(f_i - f_{i-1}) \cdot A_i]}{(f_i - f_{i-1}) - (f_i - f_{i+1})},$$

dove  $Cl_i$  è il confine inferiore della classe  $i$ ,  $A_i$  è la sua ampiezza e  $f_i$  è la frequenza della classe.

### Proprietà della moda

- 1) La moda rende *massimo il numero degli scostamenti nulli*.
- 2) *Non è influenzata dalla presenza di valori estremi*, tuttavia viene utilizzata solamente per scopi descrittivi, perché è *meno stabile ed oggettiva* di altre misure di tendenza centrale. Essa differisce sia da campione a campione, sia quando con gli stessi dati si formano classi di distribuzione con ampiezza differente
- 3) Se si fanno variare tutti i termini di una serie in base ad una certa legge, la moda della serie data corrisponde alla moda della nuova serie.

**Esempio** Si abbia la seguente distribuzione del numero di frantoi in funzione della capacità annua di produzione di olio:

Capacità produttiva (q.li)	Numero di frantoi
150  --200	60
200  --300	115
300  --500	140
500  --7 50	7 5
7 50  --1000	<u>15</u>
Totale	405

A prima vista si potrebbe ritenere che la moda sia compresa nella classe 300| --500, ma ciò è falso, in quanto le classi hanno ampiezza differente.

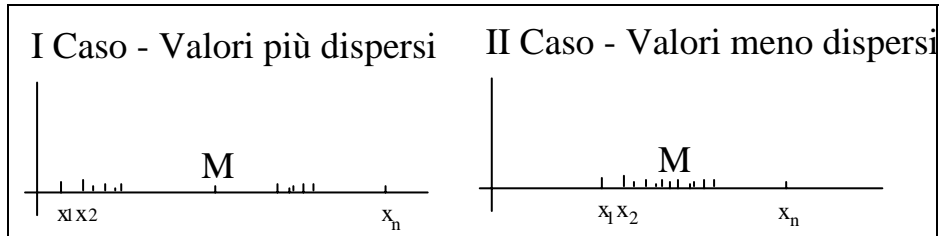
Per trovare la classe modale, quindi occorre anzitutto ridurre le varie classi ad una ampiezza uguale: ad esempio:

Capacità Produttiva	Numero di frantoi		
150  --200	60	7 00  --7 50	15
200  --250	57 ,5	7 50  --8 00	3
250  --300	57 ,5	8 00  --8 50	3
300  --350	35	8 50  --900	3
350  --400	35	900  --950	3
400  --450	35	950  --1000	<u>3</u>
450  --500	35		
500  --550	15		
550  --600	15		
600  --650	15		
650  --7 00	15		
		Totale	405

Così facendo si scopre che la classe modale è la prima.

### Dispersione o variabilità

**Definizione** Si definisce *dispersione* oppure *variabilità* di una distribuzione l'attitudine dei dati a disporsi intorno a un valore medio



**Misure di dispersione o variabilità**

**Campo di variazione**

Intervallo di valori compreso tra il più piccolo ed il più grande dei valori assunti dalla variabile.

**Deviazione semplice media o scostamento semplice medio**

$$S. s. m. \text{ rispetto a } K = \frac{1}{N} \sum_{i=1}^{i=n} |x_i - K| \cdot f_i$$

ove  $N = \sum_{i=1}^{i=n} f_i$ .

E' una misura di dispersione che *dipende da tutti i valori della variabile*, ma non è molto utilizzata, specie negli sviluppi teorici, per la *non derivabilità* della funzione dovuta al valore assoluto degli scarti. Ad essa viene generalmente preferita un'altra misura della variabilità basata sul quadrato degli scarti dalla media.

**Scostamento quadratico o scarto quadratico medio**

$$Scostamento quadratico da K = \sqrt{\frac{\sum_{i=1}^{i=n} (x_i - K)^2 \cdot f_i}{N}}$$

K può essere la media aritmetica, la mediana o qualsiasi altro valore medio preferito. Se K è la media aritmetica, lo scostamento quadratico medio si chiama anche *deviazione standard* e, in genere, viene sempre indicato con la lettera greca  $\sigma$  (sigma) quando ci si riferisce ad una intera popolazione di valori oppure con la lettera latina s (esse), quando ci si riferisce ad un campione di valori tratto dalla popolazione studiata.

E' la misura di dispersione più utilizzata.

**Varianza** E' pari al *quadrato della deviazione standard*:

$$\sigma^2 = \frac{\sum_{i=1}^{i=n} (x_i - M)^2 \cdot f_i}{N}$$

**Devianza** E' pari al *numeratore della varianza*. Ha una grande importanza in statistica, perché può essere scomposta in porzioni che sono di grande utilità per la teoria dell'analisi della varianza.



### *Coefficiente di variazione*

E' pari al *rapporto tra scostamento quadratico medio e media* della distribuzione moltiplicato per cento.

Essendo una misura relativa della variabilità, consente di comparare tra loro due o più distribuzioni le cui unità di misura sono molto diverse. In formule:

$$V = \frac{\sigma}{M} \cdot 100$$

Altre misure di variabilità si possono avere anche con riferimento agli scarti delle singole modalità da altri tipi di medie.

*Nella statistica non parametrica è molto usato lo scostamento quadratico medio dalla mediana della distribuzione.*

### *Momento centrato di ordine k per distribuzioni discrete*

Il k-esimo momento rispetto ad una origine arbitraria A è definito dalla espressione:

$$v_k = \frac{1}{N} \sum_{i=1}^{i=n} (x_i - A)^k \cdot f_i$$

Se A = M, il k-esimo momento rispetto alla media aritmetica (M) è pari a:

$$\mu_k = \frac{1}{N} \sum_{i=1}^{i=n} (x_i - M)^k \cdot f_i$$

Se A = 0, si definisce il k-esimo momento rispetto all'origine degli assi (zero), cioè:

$$m_k = \frac{1}{N} \sum_{i=1}^{i=n} x_i^k \cdot f_i$$

### *Momento di ordine k per distribuzioni continue*

Nel caso di distribuzioni continue il posto del segno di sommatorio è preso dal segno di integrale. Quindi le tre espressioni date per i momenti di ordine k rispetto all'origine arbitraria, rispetto alla media e rispetto a zero sono:

$$v_k = \int_a^b (x - A)^k \cdot f(x) \cdot dx,$$

$$\mu_k = \int_a^b (x - M)^k \cdot f(x) \cdot dx,$$

$$m_k = \int_a^b x^k \cdot f(x) \cdot dx,$$

in cui, al solito l'intervallo (a, b) indica il campo di definizione della funzione o distribuzione.

Da queste definizioni segue immediatamente che:

- *il primo momento rispetto all'origine non è altro che la media aritmetica*

- *il secondo momento centrato rispetto alla media è la varianza.*

Il terzo ed il quarto momento rispetto alla media, divisi rispettivamente per il cubo e per la quarta potenza dello scostamento quadratico medio, sono utilizzati in statistica anche per misurare l'asimmetria e l'appiattimento delle distribuzioni statistiche di tipo campanulare.

### Correzioni di Sheppard

Supporre che una classe possa essere rappresentata dal suo valore centrale comporta errori nel calcolo dei momenti che possono essere corretti con le seguenti formule di Sheppard:

- per il momento secondo:

$$\mu_2 \text{ corretto} = \mu_2 - \frac{h^2}{12}$$

- per il momento quarto:

$$\mu_4 \text{ corretto} = \mu_4 - \frac{h^2}{2} \cdot \mu_2 + \frac{7}{240} h^4$$

dove:  $h$  = ampiezza delle classi

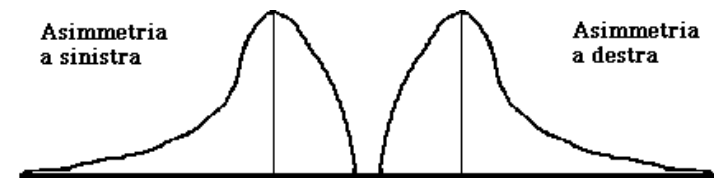
mentre i momenti  $\mu_1$  e  $\mu_3$  non hanno bisogno di essere corretti.

### Forma della distribuzione

**Caratteristiche descrittive della forma:** Asimmetria-  
Appiattimento

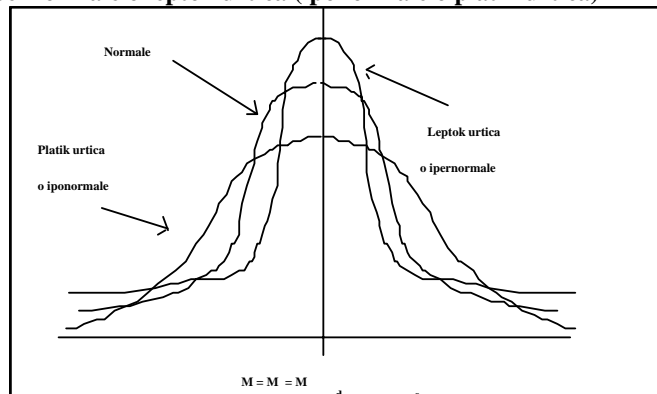
#### ASIMMETRIA (Skewness)

Una distribuzione si dice *simmetrica* se, con riferimento al valore mediano, le frequenze dei valori inferiori a quest'ultimo si distribuiscono così come le frequenze dei valori superiori ad esso ed in maniera *speculare* rispetto al valore centrale. Se ciò non si verifica la distribuzione è asimmetrica: positiva, se la coda si allunga verso i valori positivi, negativa in caso contrario. Una distribuzione simmetrica non è detto che sia unimodale.



#### APPIATTIMENTO (Kurtosis)

L'appiattimento o disnormalità è quella caratteristica che individua le distribuzioni i cui valori centrali e quelli estremi hanno frequenza più elevata (oppure meno elevata) di quella tipica di una distribuzione gaussiana, mentre i valori intermedi tra quelli estremi e quelli centrali hanno una frequenza meno elevata (oppure più elevata). In tal caso si parla di distribuzione ipernormale o leptokurtica (iponormale o platikurtica)



#### Misure di asimmetria

Misura basata sulla relazione tra media, moda e scostamento quadratico medio

$$S_k = \frac{M_1 - M_d}{\sigma}$$

Misura basata sulla relazione tra media, mediana e scostamento quadratico medio

$$S_k = \frac{M_1 - M_e}{\sigma}$$

**Misura basata sul terzo momento rispetto alla media (dipende dall'unità di misura)**

$$S_k = \frac{1}{N} \sum_{i=1}^n (x_i - M_1)^3 \cdot f_i$$

**Misura basata sul terzo momento rispetto alla media (non dipende dall'unità di misura)**

$$\alpha_3 = \frac{\frac{1}{N} \sum_{i=1}^n (x_i - M_1)^3 \cdot f_i}{\sigma^3} = \frac{\mu_3}{\sqrt{\mu_2^3}}$$

**Coefficiente beta uno del Pearson**

$$\beta_1 = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3}$$

### *Misure di appiattimento*

**1) Misura basata sul quarto momento rispetto alla media (dipende dall'unità di misura)**

$$K_u = \frac{1}{N} \sum_{i=1}^n (x_i - M_1)^4 \cdot f_i$$

**2) Misura basata sul quarto momento rispetto alla media (non dipende dall'unità di misura) o coefficiente beta due del Pearson**

$$\alpha_4 = \frac{\frac{1}{N} \sum_{i=1}^n (x_i - M_1)^4 \cdot f_i}{\sigma^4} = \frac{\mu_4}{\mu_2^2}$$

*Nelle distribuzioni gaussiane l'appiattimento misurato dal coefficiente del Pearson è  $\beta_2 = 3$*

## Distribuzioni teoriche

### A. DISTRIBUZIONI DISCRETE

- Distribuzione binomiale
- Distribuzione multinomiale
- Distribuzione di Poisson
- Distribuzione ipergeometrica
- Distribuzione uniforme

### B. DISTRIBUZIONI CONTINUE

- Distribuzione normale o di Gauss
- Distribuzione rettangolare
- Distribuzione esponenziale negativa
- Le curve di Karl Pearson
- Distribuzione Gamma
- Distribuzione Beta
- Distribuzione chi-quadrato
- Distribuzioni F di Fisher e t di Student

## Distribuzione binomiale

### Momento terzo rispetto all'origine

$$M_3 = \sum_{i=0}^n x_i^3 \cdot P(x_i) = \sum_{x=0}^n x^3 \cdot \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x}$$

e poiché  $x^3 = x[x + x(x-1)]$ ,

$$M_3 = np + 3n(n-1)p^2 + n(n-1)(n-2)p^3$$

### Momento terzo rispetto alla media

$$\mu_3 = M_3 - 3 \cdot M_2 \cdot M_1 + 2 \cdot M_1^3 = npq(q-p).$$

### Asimmetria

$$\alpha_3 = \frac{\mu_3}{\sigma^3} = \frac{npq(q-p)}{npq\sqrt{npq}} = \frac{q-p}{\sqrt{npq}}.$$

### Coefficiente di asimmetria del Pearson

$$\beta_1 = \alpha_3^2 = \frac{\mu_3^2}{\mu_2^3} = \frac{(q-p)^2}{npq}.$$

### 10. Appiattimento o Kurtosis o disnormalità

$$\alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = \frac{1}{(npq)^2} [3n^2 p^2 q^2 + npq(1-6pq)] = 3 + \frac{(1-6pq)}{npq}$$

### Coefficiente di appiattimento del Pearson.

$$\beta_2 = \alpha_4 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4} = 3 + \frac{(1-6pq)}{npq}.$$

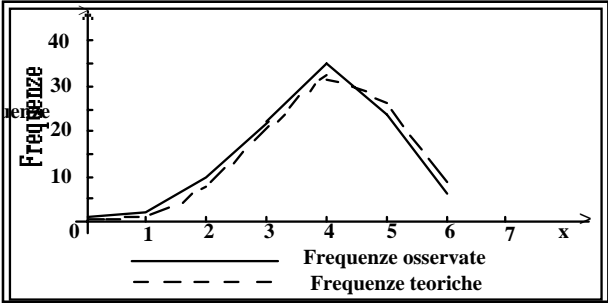
*La curva binomiale è più aguzza della normale  
ossia è leptokurtica o ipernormale  
(Per la curva normale è  $\beta_2 = 3$ )*

**Esempio** In un villaggio si esegue un'inchiesta sul numero dei campi gestiti direttamente dal proprietario. Da precedenti esperienze si sa che la proporzione dei campi gestiti direttamente dal proprietario è pari a  $2/3$ . Estratti 100 campioni di 6 campi ciascuno i risultati sono:

Numero di campi gestiti dal proprietario sui 6 campi esaminati	Proporzione dei campi gestiti dal proprietario	Numero di campioni che danno tale risultato
0	0	1
1	1/6	2
2	2/6	10
3	3/6	22
4	4/6	35
5	5/6	24
6	6/6	6
		100

Adattiamo a tali dati una distribuzione binomiale e confrontiamo le frequenze teoriche calcolate con i risultati della rilevazione campionaria. Inoltre, troviamo media e deviazione standard di questa distribuzione e tracciamo un grafico di confronto tra frequenze osservate e teoriche.

In questo esempio,  $p = 2/3$ ,  $q = 1/3$ , mentre  $n = 6$  ed  $N = 100$ . La tavola che segue mostra le frequenze teoriche trovate calcolando i termini nell'espansione del binomio:



Avremo, quindi la tavola di confronto seguente:

x	Frequenze osservate	Frequenze teoriche
0	1	0,137
1	2	1,646
2	10	8 ,230
3	22	21,948
4	35	32,922
5	24	26,338
6	6	8 ,7 7 9
Totale	100	100,000

x	f	xf	x <sup>2</sup> f
0	1	0	0
1	2	2	2
2	10	20	40
3	22	66	198
4	35	140	560
5	24	120	600
6	6	36	216
100		38 4	1616

per cui sarà:  $M = 38\ 4/100 = 3,8\ 4$ ;  $M^2 = 1616/100 = 16,16$

$s^2 = M_2 - M^2 = 16,16 - (3,8\ 4)^2 = 16,16 - 14,7\ 5 = 1,41$

$s = \sqrt{1,41} = 1,18\ 7$ .

La media e la deviazione standard dei valori teorici, invece, saranno:

$$M = np = 6 \cdot \frac{2}{3} = 4,$$

$$s = \sqrt{(6 \cdot \frac{2}{3} \cdot \frac{1}{3})} = \sqrt{\frac{4}{3}} = 1,153.$$

### Distribuzione multinomiale

Estensione della binomiale per k eventi indipendenti di probabilità  $p_1, p_2, \dots, p_i, \dots, p_n$ , la cui somma è uguale ad 1, che possono comparire nel corso di N prove indipendenti, successive o simultanee.

La probabilità della combinazione di eventi  $n_1, n_2, \dots, n_k$  in N prove è determinata dal multinomio:

$$P_{(n_1, n_2, \dots, n_k)} = \frac{N!}{n_1! n_2! \dots n_k!} p_1^{n_1} p_2^{n_2} \dots p_k^{n_k}$$

Per k qualunque,  $0 \leq \sum_{i=1}^k p_i \leq 1$

### Distribuzione di Poisson

Per n tendente all'infinito (cioè il numero dei dati è molto grande) e p prossimo a zero (cioè la probabilità che l'evento ha di verificarsi è molto piccola), ma tali che  $n \cdot p$  rimanga costante, una buona approssimazione alla binomiale è data dalla *funzione di distribuzione di Poisson*:

$$P(x) = \frac{e^{-M} \cdot M^x}{x!} \quad \text{se} \quad \begin{cases} n \rightarrow \infty \\ p \rightarrow 0 \\ n \cdot p = \text{cost.} \end{cases}$$

La funzione di distribuzione di Poisson fornisce adattamenti molto buoni in problemi riferiti ad un *evento raro*

(Es. manifestazioni di patologie o eventi catastrofici rari)

### La binomiale come approssimazione della poissoniana

$$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x \cdot q^{n-x} \quad (1)$$

Se  $np = a$ , è  $p = a/n$  e sostituendo p con  $a/n$  e q con  $(1 - a/n)$ :

$$P(x) = \frac{n!}{x!(n-x)!} \cdot \left(\frac{a}{n}\right)^x \cdot \left(1 - \frac{a}{n}\right)^{n-x} = \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} \cdot \frac{a^x}{x!} \cdot \left(1 - \frac{a}{n}\right)^{n-x}$$

Poiché  $\lim_{n \rightarrow +\infty} \frac{n(n-1)(n-2)\dots(n-x+1)}{n^x} = 1$  e  $\lim_{n \rightarrow +\infty} \left(1 - \frac{a}{n}\right)^{n-x} = e^{-a}$

Sarà  $P(x) = \frac{a^x}{x!} \cdot e^{-a}$

### Caratteristiche della distribuzione di Poisson

La distribuzione di Poisson è una *distribuzione teorica discreta* che dipende o è totalmente *definita dal solo parametro a*, cioè la *media aritmetica* della distribuzione di Poisson

L'approssimazione di Poisson o *legge dei piccoli numeri* si considera buona quando è  $p < 0,03$  o quando  $np < 5$ .

In questi casi sono assai frequenti le classi con zero o con pochi eventi rispetto a quelle con numerosi eventi.

**Momenti della distribuzione di Poisson**  
**Media o momento primo rispetto all'origine**

$$M = \sum_{x_i=0}^{x_i=+\infty} x_i \cdot P(x_i) = \sum_{x=0}^{x=+\infty} x \frac{a^x}{x!} e^{-a} =$$

$$= e^{-a} \left( 0 + a + 2 \frac{a^2}{2!} + 3 \frac{a^3}{3!} + 4 \frac{a^4}{4!} + \dots \right) =$$

$$= a \cdot e^{-a} \left( 1 + a + \frac{a^2}{2!} + \frac{a^3}{3!} + \frac{a^4}{4!} + \dots \right) = a \cdot e^{-a} \cdot e^a = a$$

*La funzione di distribuzione di Poisson dipende da un solo parametro che coincide con la media aritmetica della distribuzione.*

**Momento secondo rispetto alla media**

$$\mu_2 = \sigma^2 = M_2 - M_1^2 = a + a^2 - a^2 = a = M$$

*La varianza della distribuzione di Poisson è esattamente uguale alla media della distribuzione stessa*

*La deviazione standard della distribuzione di Poisson è uguale alla radice quadrata della media aritmetica della distribuzione.*

**Momento terzo e misura dell'asimmetria**

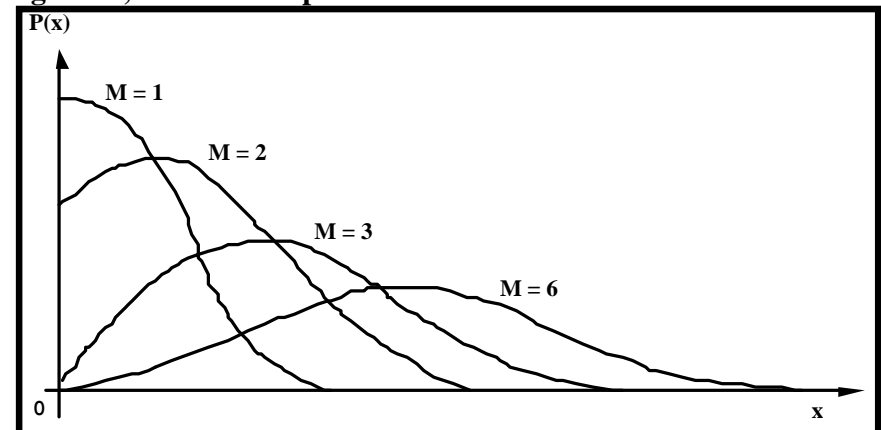
Stesso procedimento già mostrato per il calcolo del momento secondo. Risulta  $\mu_3 = M$ . Per  $p$  tendente a zero,  $np = M$ ,  $q$  tendente all'unità ed  $n$  tendente all'infinito si hanno le seguenti misure di asimmetria:

$$\alpha_3 = \lim_{n \rightarrow +\infty} \frac{q - p}{\sqrt{npq}} = \frac{1}{\sqrt{M}} \quad e \quad \beta_1 = \alpha_3^2 = \frac{1}{M}$$

*La forma della poissoniana è molto asimmetrica per valori piccoli della media aritmetica (inferiori a 3)*

**Forma generale del poligono di Poisson**

Cambia al variare di  $M$  e di  $x$ . Per bassi valori di  $M$  il poligono assume una forma molto asimmetrica, ma man mano che  $M$  diventa più grande, essa diviene più simmetrica.



**Esempio IV.2** Si riporta la distribuzione di frequenza dei decessi di mucche per una malattia rara in 50 provincie durante un periodo di 10 anni:

Numero di mucche decedute	Numero di provincie in 10 anni
0	240
1	150
2	60
3	25
4	17
5	8
<b>Totale</b>	<b>500</b>

Su questi dati calcoliamo le frequenze della distribuzione di Poisson che ha la stessa media e compariamo i risultati con le frequenze osservate.

Si ha:

x	f	xf
0	240	0
1	150	150
2	60	120
3	25	75
4	17	68
5	8	40
<b>500</b>	<b>453</b>	

$$np = M = 453/500 = 0,906;$$

$$E_{-0,906} = 0,4044$$

$$P_0 = 0,4044$$

$$P_1 = 0,4044 * 0,906 = 0,3664$$

$$P_2 = 0,3664 * 0,906/2 = 0,1660$$

$$P_3 = 0,1660 * 0,906/3 = 0,0501$$

$$P_4 = 0,0501 * 0,906/4 = 0,0113$$

$$P_5 = 0,0113 * 0,906/5 = 0,0020$$

X	F	XF	P(X)	N.P(X)
0	240	0	0,4044	202,20
1	150	150	0,3664	183,20
2	60	120	0,1660	83,00
3	25	75	0,0501	25,05
4	17	68	0,0113	5,65
5	8	40	0,0020	1,00
<b>500</b>	<b>453</b>			

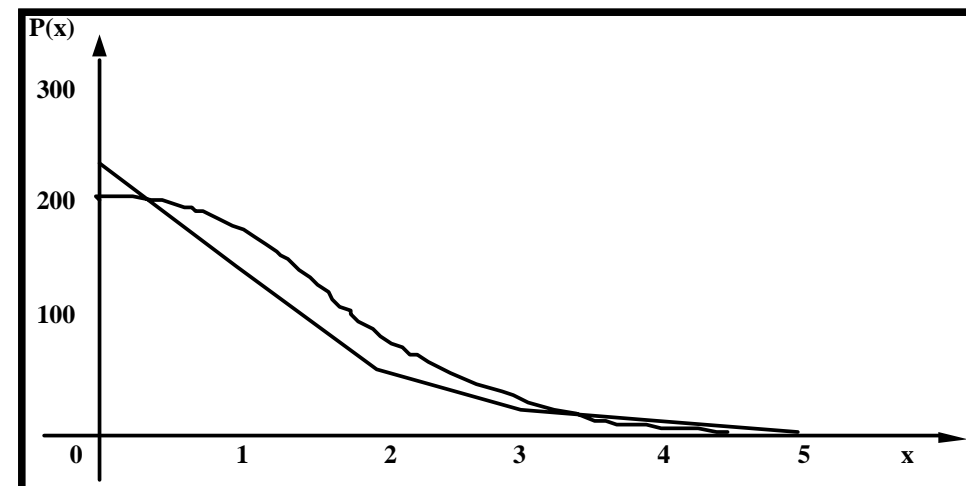
Se si vuole misurare l'adattamento della curva poissoniana ai dati osservati e, nello stesso tempo, si vuole valutarne l'attendibilità, si può calcolare una media quadratica degli scarti tra i valori osservati ed i valori teorici ottenuti, operando nel modo indicato nella tabella che segue:

x	f	N.P(x)	$\Delta = [N.P(x) - f]$	$\Delta^2$	$\Delta^2/N.P(x)$
0	240	202,20	- 37,8	1428,84	7,0665
1	150	183,20	+ 33,2	1102,24	6,0166
2	60	83,00	+ 23,0	529,00	6,3735
3	25	25,05	0	0	0
4	17	5,65	- 11,4	129,96	23,2071
5	8	1,00	- 7,0	49,00	49,0000
<b>500</b>	<b>500,00</b>	<b>0</b>	<b>3239,04</b>	<b>91,6637</b>	



Come si vede, se si considera il test del  $\chi^2$  per misurare la bontà dell'adattamento, si ha un valore di  $\chi^2 = 91,6637$  cioè un valore che può essere superato per effetto del caso solo con una probabilità molto piccola. Ciò vuol dire che l'adattamento ottenuto con la funzione di Poisson non è molto buono e che si dovrebbe procedere alla selezione di una funzione di distribuzione di altro tipo per rappresentare il fenomeno studiato.

Rappresentando in un grafico i risultati ottenuti per i valori teorici e, insieme ad essi, anche i dati osservati si ottiene la distribuzione di Poisson adattata al diagramma di frequenza dato, che si riporta nel grafico seguente:



### *Distribuzione ipergeometrica*

Nella distribuzione binomiale la *probabilità* di un evento si mantiene sempre costante. Quando essa *varia* in funzione degli eventi precedenti, come succede nell'estrazione *senza ripetizione* di alcuni oggetti da un campione di piccole dimensioni, si ha la *distribuzione ipergeometrica*.

Ad esempio, qual'è la probabilità che sia un re la seconda carta estratta da un mazzo di 40 carte? Se il gioco avviene con reimmissione della carta già estratta, la probabilità di estrarre un re è costantemente pari a 4/40. Ma se il gioco si svolge senza reintroduzione nel mazzo della carta estratta per prima, la probabilità che la seconda carta sia un re varia in rapporto all'estrazione della prima: se la prima era un re, la probabilità per la seconda estrazione è pari a 3/39; se la prima era una carta diversa, la probabilità che la seconda sia un re è 4/39.

*La distribuzione ipergeometrica permette di calcolare la probabilità di una data combinazione di eventi quando le probabilità dei vari eventi sono variabili da prova a prova*

$$P(N, n_1; n, r) = \frac{C(n_1, r) \cdot C(N - n_1, n - r)}{C(N, n)} = \frac{\binom{n_1}{r} \binom{N - n_1}{n - r}}{\binom{N}{n}}$$

**N** = unità della popolazione;

**n** = unità del campione, al massimo pari a N;

**n1** = unità della popolazione che hanno la caratteristica in esame, al massimo pari a N;

**r** = unità del campione estratte, che hanno la caratteristica in esame, al massimo pari a n.

La distribuzione ipergeometrica è definita da *tre parametri* ( $N$ ,  $n_1$ ,  $n$ : totale unità della popolazione, unità di popolazione che hanno il carattere in esame, numero di unità estratte) *in funzione di  $r$*  (numero di unità estratte che hanno il carattere in esame).

*Momenti della distribuzione ipergeometrica*

Ponendo  $p=n_1/N$  e  $q=(N-n_1)/N$  si ha:

$$P(N, Np; n, r) = \frac{\binom{Np}{r} \binom{Nq}{n-r}}{\binom{N}{n}}$$

Media aritmetica e varianza:

$$M = np \quad \text{e} \quad \sigma^2 = \frac{npq(N-n)}{N-1}.$$

La media della distribuzione ipergeometrica è uguale a quella della distribuzione binomiale corrispondente, mentre la varianza è inferiore.

La distribuzione ipergeometrica, per  $N$  tendente ad infinito, ossia per  $N$  grande rispetto ad  $n$ , tende alla distribuzione binomiale.

**Esempio IV.3** Un'urna contiene  $N$  biglie, delle quali  $n_1$  bianche e  $N-n_1$  nere. Si estraggono dall'urna  $n$  biglie (con  $n \leq N$ ) senza reimmissione; si vuole determinare la probabilità che delle  $n$  biglie estratte  $r$  siano bianche (con  $r \leq n$ ). Si ha:

$$P(N, n_1; n, r) = \frac{\binom{n_1}{r} \binom{N-n_1}{n-r}}{\binom{N}{n}} =$$

$$= \frac{n_1! (N-n_1)! n! N!}{r! (n-r)! (N-n_1-n+r)! N!}$$

### *Distribuzione normale o di Gauss*

E' la funzione continua più usata nelle applicazioni teoriche e pratiche. Essa è rappresentata da una curva simmetrica di forma campanulare, talvolta chiamata "*curva degli errori accidentali*". Mostra graficamente il numero degli scarti tra osservazioni reali e loro valore teorico, quando tali scarti sono casuali, cioè non dipendono dall'azione di fattori sistematici.

Le stime campionarie di un parametro della popolazione da cui proviene il campione sono descritte dalla funzione esponenziale:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{z^2}{2\sigma^2}}$$

Con forma esplicita degli scarti rispetto alla media, la funzione che descrive la distribuzione normale è:

$$f(x) = \frac{N}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-M)^2}{2\sigma^2}}$$

Introducendo una nuova variabile  $t$  chiamata "unità standard" o "scarto ridotto" oppure "scarto standardizzato" e definita dalla

relazione:

$$t = \frac{x - M}{\sigma}$$

l'equazione della "*curva normale standardizzata*" diventa:

$$f(t) = \frac{N}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}}$$

Molte distribuzioni empiriche seguono la curva normale (caratteri antropometrici e biometrici).

*La curva normale si usa anche per distribuzioni non normali, ma che possono approssimarsi a tale forma con opportune trasformazioni di variabile:* Ad esempio, una distribuzione asimmetrica rispetto alla variabile  $x$  può diventare quasi normale quando invece della variabile  $x$  si considera una sua trasformata del tipo  $\text{radq}(x)$  oppure  $x^2$  oppure  $\log x$  o  $1/x$ , eccetera.

*Ha grande importanza nella teoria statistica per le sue numerose proprietà matematiche:* In teoria dei campioni, si dimostra che, anche quando la variabile di base non ha una distribuzione normale, la media campionaria, attraverso la quale si cerca di stimare la media vera della variabile di base, segue approssimativamente una distribuzione normale.

La distribuzione della variabile normale standardizzata  $t$  si ottiene dalla distribuzione della variabile di base  $x$  con una trasformazione di quest'ultima che trasferisce l'*origine dei valori nel punto medio* della distribuzione di base ed assume come nuova *unità di misura la deviazione standard* della variabile di base.

*La trasformata  $t$  è una variabile caratterizzata dal fatto di avere media uguale a zero e deviazione standard uguale all'unità.*

*La normale come approssimazione della binomiale*

La funzione normale può essere trovata anche come forma limite della funzione binomiale per valori di  $n$  abbastanza grandi. In

$$P(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x}$$

### Alcune proprietà della curva normale

La somma di tutti i valori di  $P(x)$  compresi tra  $x = -\infty$  ed  $x = +\infty$  è pari all'unità e vale la relazione:

$$\int_{-\infty}^{+\infty} P(x).dx = \int_{-\infty}^{+\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-M)^2}{2\sigma^2}} \cdot dx =$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} e^{-\frac{(x-M)^2}{2\sigma^2}} \cdot dx = 1$$

oppure, nella forma standard in cui  $t = (x-M)/\sigma$  e  $dx = \sigma dt$ :

$$\int_{-\infty}^{+\infty} P(t).dt = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{t^2}{2}} \cdot dt = \frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} \cdot dt = 1$$

- Ha un solo massimo per  $x = M$ ;
- Ha due punti di flesso per  $x = M \pm \sigma$ ;
- $f(x) \rightarrow 0$  per  $x \rightarrow \pm \infty$ . cioè l'asse  $x$  è un asintoto;
- Media, mediana e moda coincidono.

Apposite tavole numeriche danno le ordinate della curva corrispondenti ad un dato valore dell'ascissa.

Esse danno anche la misura delle aree comprese sotto la curva a sinistra ed a destra di certe ordinate.

In generale le tavole sono preparate utilizzando i valori della variabile espressi in unità standard.

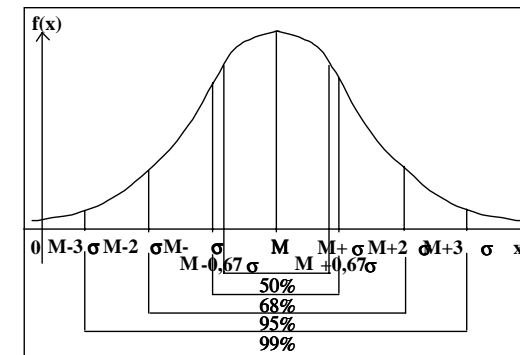
La curva normale è simmetrica rispetto alla verticale passante per la media della distribuzione per cui

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-M)^2}{2\sigma^2}} \cdot dx = \frac{2}{\sigma\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{(x-M)^2}{2\sigma^2}} \cdot dx = \frac{2}{\sigma\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{(x-M)^2}{2\sigma^2}} \cdot dx = 1/2$$

e per la variabile ridotta standardizzata

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{2}} \cdot dt = \frac{2}{\sqrt{2\pi}} \int_{-\infty}^0 e^{-\frac{t^2}{2}} \cdot dt = \frac{2}{\sqrt{2\pi}} \int_0^{+\infty} e^{-\frac{t^2}{2}} \cdot dt = 1/2$$

La distribuzione normale è completamente determinata dalla sua media  $M$  e dalla sua deviazione standard  $\sigma$



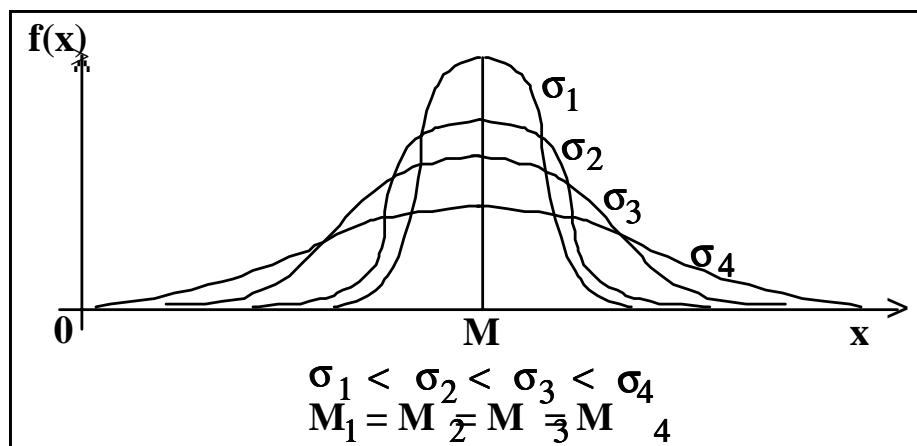
La metà dell'intera area è compresa nell'intervallo  $M \pm 0,67 \sigma$ , il quale è anche chiamato "errore probabile".

Circa il 68 % dell'area è nell'intervallo  $M \pm \sigma$

Circa il 95% insiste sull'intervallo  $M \pm 2\sigma$

Circa il 99% circa ricade nell'intervallo  $M \pm 3\sigma$ .

A parità di valore medio, la forma campanulare è tanto più appiattita quanto più grande è la variabilità



Date  $n$  variabili indipendenti  $x_1, x_2, \dots, x_n$ , se la loro distribuzione è di tipo normale, con media  $m_1, m_2, \dots, m_n$  e varianza  $\sigma_{x_1}, \sigma_{x_2}, \dots, \sigma_{x_n}$  allora la variabile  $X = x_1 + x_2 + \dots + x_n$ , somma di queste variabili è anch'essa distribuita normalmente, con media  $M = m_1 + m_2 + \dots + m_n$  e varianza  $\Sigma = \sigma_{x_1} + \sigma_{x_2} + \dots + \sigma_{x_n}$ , uguali alla somma delle medie ed alla somma delle varianze delle variabili originarie.

#### Momenti della distribuzione normale

La media e la deviazione standard della distribuzione normale sono uguali ad  $M$  ed a  $\sigma$  che compaiono nell'equazione normale.

Il momento di ordine  $k$  rispetto alla media è definito da:

$$\mu_k = \int_{-\infty}^{+\infty} \frac{(x - M)^k \cdot f(x)}{N} \cdot dx =$$

Data la simmetria della distribuzione tutti i momenti di ordine dispari rispetto alla media sono nulli

Per i momenti di ordine pari vale la relazione:

$$\mu_k = (k-1) \cdot \sigma^2 \cdot \mu_{k-2}.$$

per cui si avrà:  $\mu_0 = 1, \mu_1 = 0, \mu_2 = \sigma^2, \mu_4 = 3 \cdot \sigma^4, \mu_6 = 15 \cdot \sigma^6$  e così via.

Il parametro  $\sigma$  della legge normale è dunque lo scarto tipo o deviazione standard di  $x$

La variabile  $t$ , trasformata della  $x$  che ha origine degli assi nel punto medio e unità di misura pari alla deviazione standard della variabile originaria, è detta anche variabile normalizzata o scarto ridotto e la sua legge di distribuzione viene detta forma ridotta della legge normale.

Non dipende da alcun parametro, ha media uguale a zero e deviazione standard uguale all'unità. I momenti dei successivi ordini fino al quarto hanno i valori  $\alpha_3 = 0$  ed  $\alpha_4 = 3$  e lo stesso dicasi per  $\beta_1$  e  $\beta_2$ .

Questo risultato è molto importante, in quanto è a questi valori che vengono paragonati quelli corrispondenti calcolati per le altre distribuzioni di tipo campanulare, al fine di valutarne il grado di asimmetria e di appiattimento.

### Adattamento della distribuzione normale a dati empirici

La densità di frequenza della distribuzione normale in  $x$  è :

$$f(x) = \frac{N}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{(x-M)^2}{2\sigma^2}}$$

che in termini di logaritmi diviene:

$$\log f(x) = -\frac{(x-M)^2}{2\sigma^2} + \log \frac{N}{\sigma\sqrt{2\pi}}$$

e ponendo  $Y = \log f(x)$  ed  $X = (x - M)^2$ , esprimerà la retta:

$$Y = -\frac{1}{2\sigma^2} \cdot X + \log \frac{N}{\sigma\sqrt{2\pi}}$$

che può essere adattata ai dati effettivi con uno dei metodi di interpolazione disponibili.

**Esempio IV.4** Si consideri la seguente tabella delle frequenze dell'altezza in centimetri di 78 piante di una determinata specie. Ci si propone di:

- verificare se la curva normale può essere scelta come una buona curva interpolatrice per tale distribuzione di frequenze;
- calcolare le frequenze teoriche della curva normale che ha la stessa media e la stessa varianza della seriazione data.

Altezza (cm)	Frequenza
10 - 15	4
15 - 20	20
20 - 25	28
25 - 30	12
30 - 35	8
35 - 40	6
	<b>78</b>

Calcoliamo anzitutto i momenti della distribuzione data.

Classi di altezza (cm)	Media di classe (v <sub>i</sub> )	x <sub>i</sub> =(v <sub>i</sub> -22,5)	f <sub>i</sub>	x <sub>i</sub> f <sub>i</sub>	x <sub>i</sub> <sup>2</sup> f <sub>i</sub>	x <sub>i</sub> <sup>3</sup> f <sub>i</sub>	x <sub>i</sub> <sup>4</sup> f <sub>i</sub>
10 - 15	12,5	-2	4	-8	16	-32	64
15 - 20	17,5	-1	20	-20	20	-20	20
20 - 25	22,5	0	28	0	0	0	0
25 - 30	27,5	1	12	12	12	12	12
30 - 35	32,5	2	8	16	32	64	128
35 - 40	37,5	3	6	18	54	162	486
<b>Totali</b>			<b>78</b>	<b>3</b>	<b>18</b>	<b>134</b>	<b>186</b>

$$M = 18 / 78 = 0,23; \quad M_2 = 134 / 78 = 1,72;$$

$$M_3 = 186 / 78 = 2,38; \quad M_4 = 10 / 78 = 0,128;$$

$$\mu_2 = M_2 - M^2 = 1,72 - 0,053 = 1,665.$$

Con le correzioni di Sheppard sul momento secondo sarà:

$$\mu'_2 = \mu_2 - 1/12 = 1,66 - 0,08 = 1,58$$

e quindi  $\sigma = \sqrt{1,58} = 1,26$ .

Per il momento terzo avremo:

$$\mu'_3 = M_3 - 3 \cdot M_2 \cdot M + 2 \cdot M^3 =$$

$$= 2,38 - 1,19 + 0,024 = 1,21$$

$$\alpha_3 = 1,21 / (1,26)^3 = 1,21 / 1,99 = 0,6$$

e per il momento quarto  $\mu_4$  avremo:

$$\mu_4 = M_4 - 4 \cdot M_3 \cdot M + 6 \cdot M_2 \cdot M^2 - 3 \cdot M^4 =$$

$$= 0,128 - 2,19 + 0,55 - 0,01 = 0,478$$

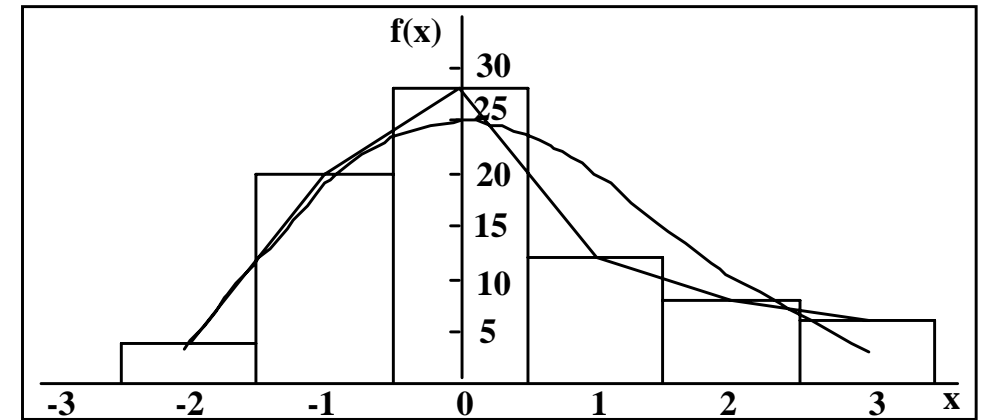
$$\begin{aligned}\mu'_4 &= \mu_4 - 1/2\mu_2 + 7/240 = \\ &= 7,44 - 0,83 + 0,03 = 6,64\end{aligned}$$

$$\alpha_4 = 6,64/(1,26)^4 = 6,64/2,50 = 2,6.$$

Come si vede, i valori di  $\alpha_3$  e  $\alpha_4$  sono quasi prossimi a zero ed a 3; si può quindi concludere che la curva normale può essere adottata come curva interpolatrice, malgrado essa non dia un perfetto adattamento ai dati.

La curva della distribuzione data ha *asimmetria positiva* e quindi essa sarà più allungata verso destra e più inclinata verso sinistra. Inoltre, poichè  $\alpha_4 = 2,6$  è minore di 3, la curva sarà *più appiattita* della curva normale.

La rappresentazione grafica della distribuzione osservata e della curva normale ad essa adattata risulta essere la seguente:



### *Distribuzione rettangolare o uniforme*

E' la più semplice tra le distribuzioni continue. Nell'intervallo tra  $x_1=\alpha$  e  $x_2=\beta$  ha densità di frequenza relativa pari a:

$$f(x) = \frac{1}{\beta - \alpha} \text{ con } (\alpha < x < \beta)$$

quindi ha *densità costante* in tutto l'intervallo compreso tra  $\alpha$  e  $\beta$ .

La rappresentazione grafica di questa distribuzione ha la forma di un rettangolo, che giustifica il suo nome.

E' l'equivalente della distribuzione uniforme discreta considerata nel continuo e la sua *media* e *varianza* sono:

$$M = \frac{\alpha + \beta}{2} \text{ e } \sigma^2 = \frac{(\beta - \alpha)^2}{12}$$

La mediana coincide con la media, mentre la moda o non esiste o ve ne sono tante quanti i valori compresi tra  $\alpha$  e  $\beta$ .

### **Distribuzioni continue**

#### *Distribuzione esponenziale negativa*

Distribuzione continua descritta dalla relazione

$$f(x) = \alpha \cdot e^{-\alpha x} \text{ con } \alpha > 0 \text{ e } x > 0$$

E' una funzione *positiva o nulla continuamente decrescente*, che tende a 0 per  $x$  *tendente all'infinito*. Nel discreto ha l'equivalente nella distribuzione geometrica decrescente. La *media* e la *varianza* sono, rispettivamente:

$$M = \frac{1}{\alpha} \text{ e } \sigma^2 = \frac{1}{\alpha^2} = M^2.$$