Ulteriori Conoscenze di Informatica e Statistica

Carlo Meneghini

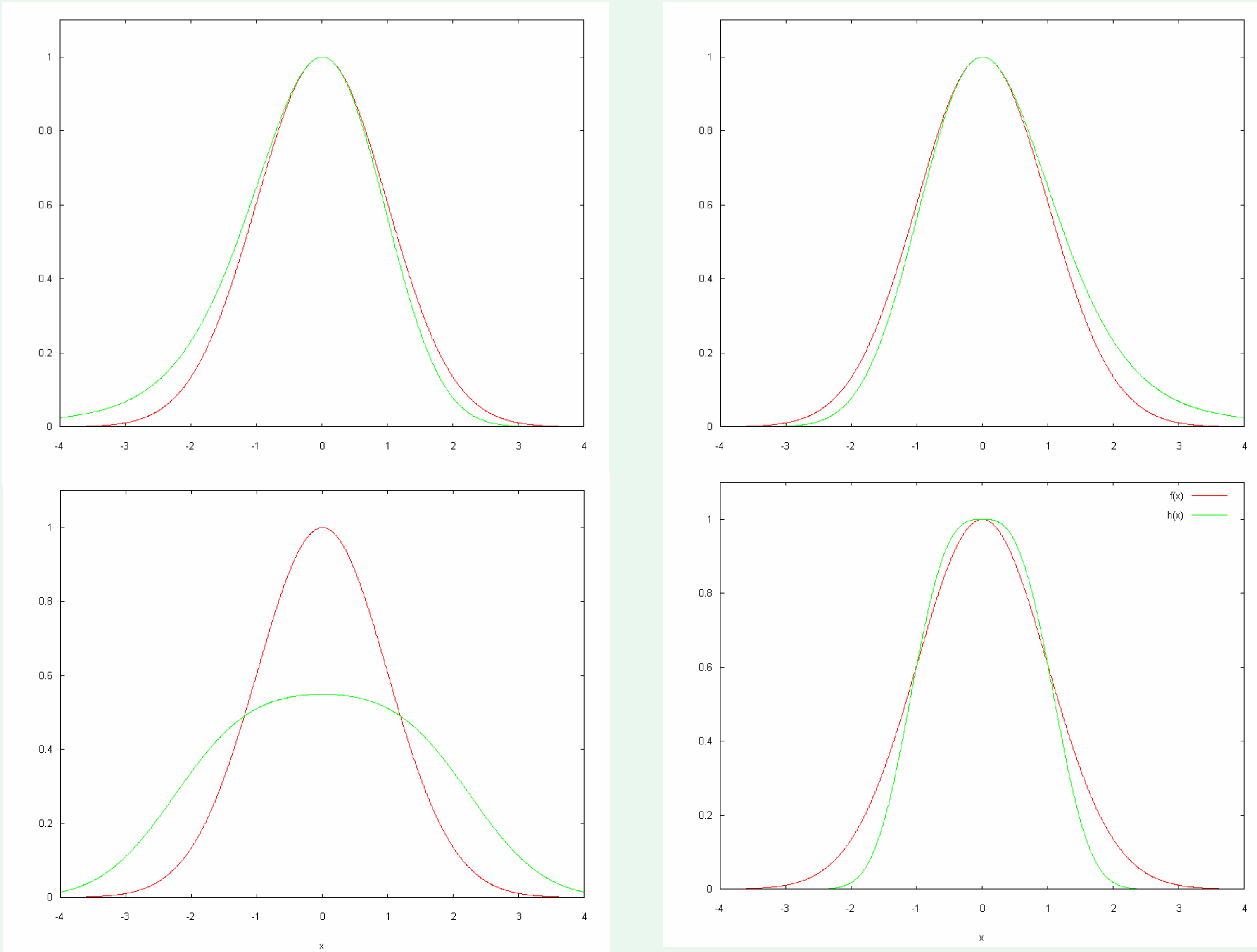
Dip. di fisica – via della Vasca Navale 84,

st. 83 (I piano)   tel.: 06 55 17 72 17

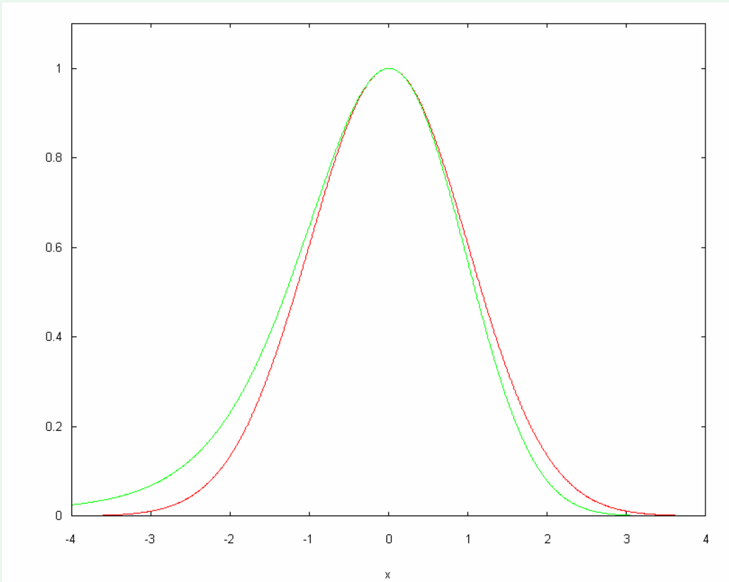
meneghini@fis.uniroma3.it

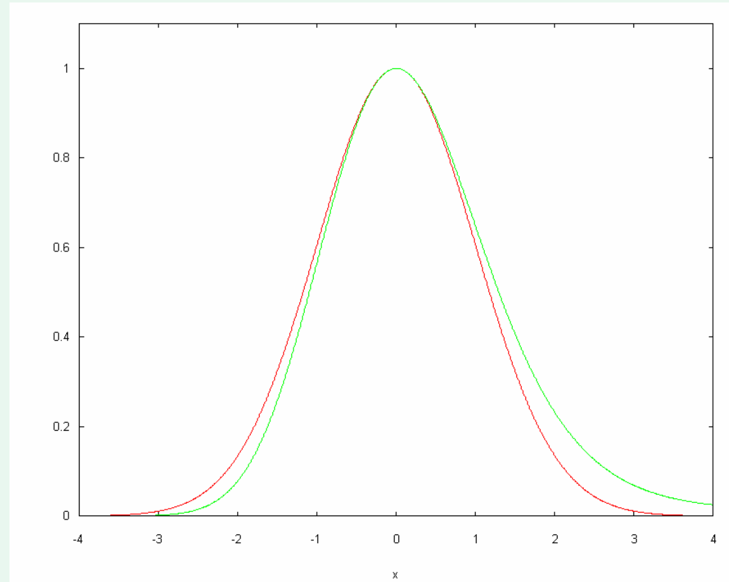
Indici di forma

Descrivono le asimmetrie della distribuzione



Una distribuzione non simmetrica si dice obliqua





Distribuzione obliqua sx:

la media è minore della mediana

Distribuzione obliqua dx:

la media è maggiore della mediana

Momenti della distribuzione

$$m_z = \frac{1}{N}\sum_{i=1}^{N}(x_i)^z = \sum_{i=1}^{N} f_i(x_i)^z$$

$$f_i = \frac{N_i}{N}$$

$$\sum_{i=1}^{N} N_i = N$$

$$M_z = \sum_{i=1}^{N} f_i(x_i - m_o)^z$$

Momenti centrali
(momenti rispetto alla media)

f_i = frequenza relativa

N_i frequenza assoluta

Media = m_o

varianza = σ^2 = M_2

$$\overline{x} = m_o = \sum_{i=1}^{N} f_i x_i$$

$$\sigma^2 = \sum_{i=1}^{N} f_i (x_i - m_o)^2 = \sum_{i=1}^{N} f_i x_i^2 - \overline{x}^2 = \overline{x^2} - \overline{x}^2$$

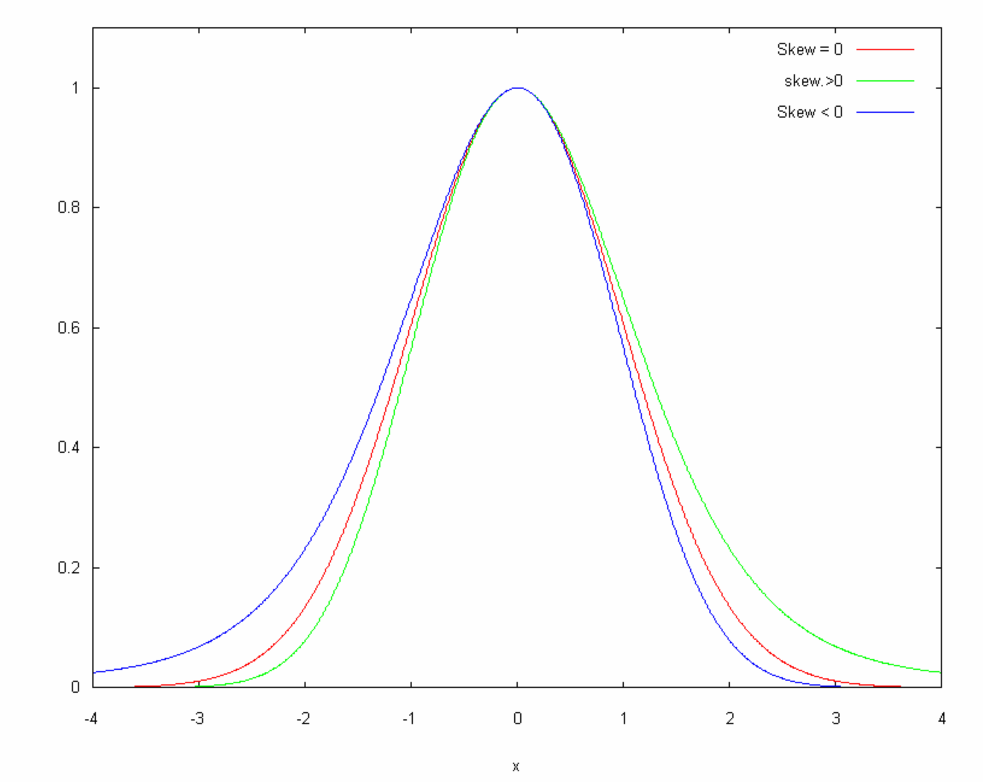
Coefficiente di asimmetria (a_{simm})

$$m_3 = \sum_{i=1}^{N} f_i (x_i - m_o)^3$$

$$a_{simm} = \frac{m_3}{\sigma^3}$$

Skewness

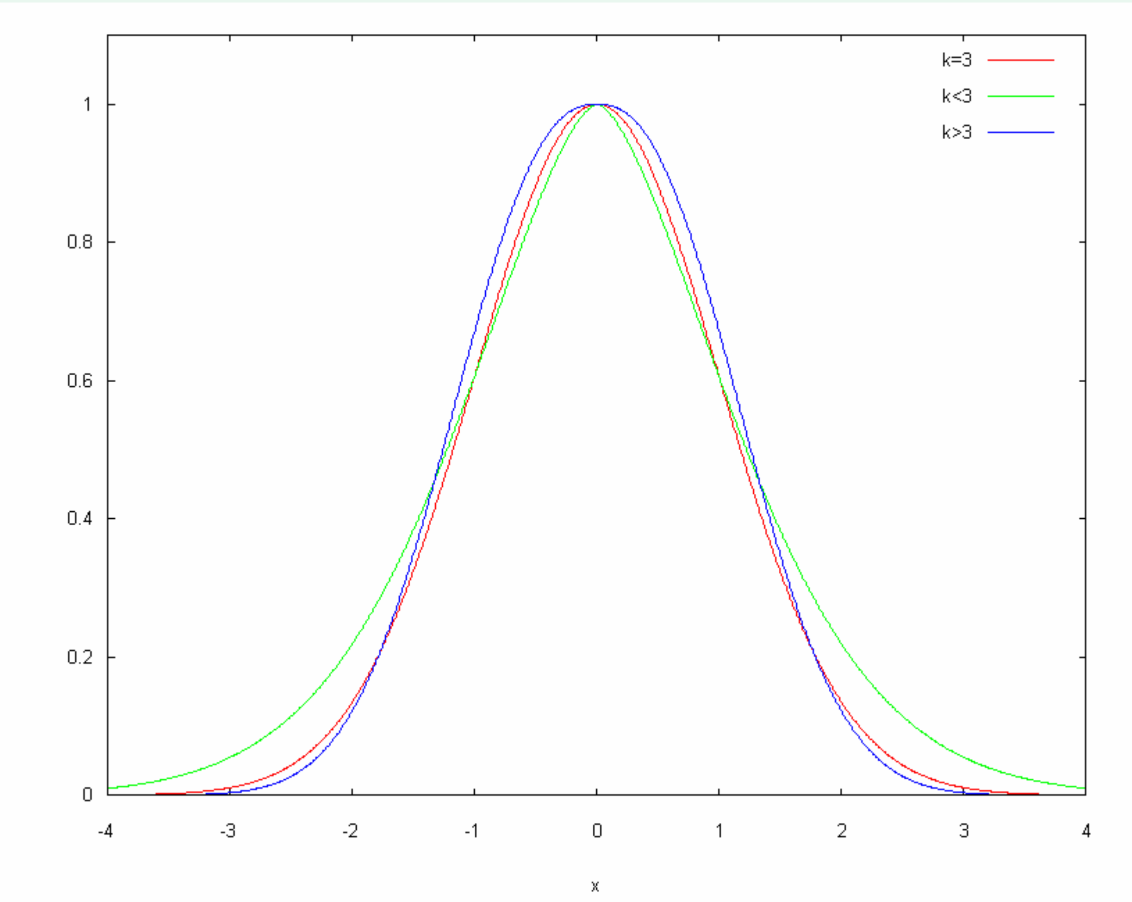
per la distribuzione Normale (Gauss) $a_{simm} = 0$



Curtosi (Kurtosis) (a_4)

$$m_4 = \sum_{i=1}^{N} f_i (x_i - m_o)^4$$

$$a_4 = \frac{m_4}{\sigma^4}$$



per la distribuzione Normale (Gauss)

$a_4 = 3$

$a_4 > 3$ indica una distribuzione più piatta di una Gaussiana

$a_4 < 3$ indica una distribuzione più piccata di una Gaussiana

Centratura e standardizzazione

Se x è una variabile aleatoria caratterizzata da valor medio μ e varianza σ^2 la variabile

$$Z = (x - \mu)/\sigma$$

È una variabile aleatoria con media nulla e varianza unitaria

Tabelle di contingenza

Ad ogni valore i-esimo della variabile (qualitativo o quantitativa) è associato il numero delle osservazioni

Modalità

Guppi

anno	I	II	III	N. stud.
2001	45	35	12	92
2002	42	38	21	98
2003	58	28	14	112
Totali	145	101	47	293

osservazioni

Distribuzioni marginali

anno	I	II	III	N. stud.
2001	45	35	12	92
2002	42	38	21	98
2003	58	28	14	112
Totali	145	101	47	293

$f_{ij}=x_{ij} / N_i$

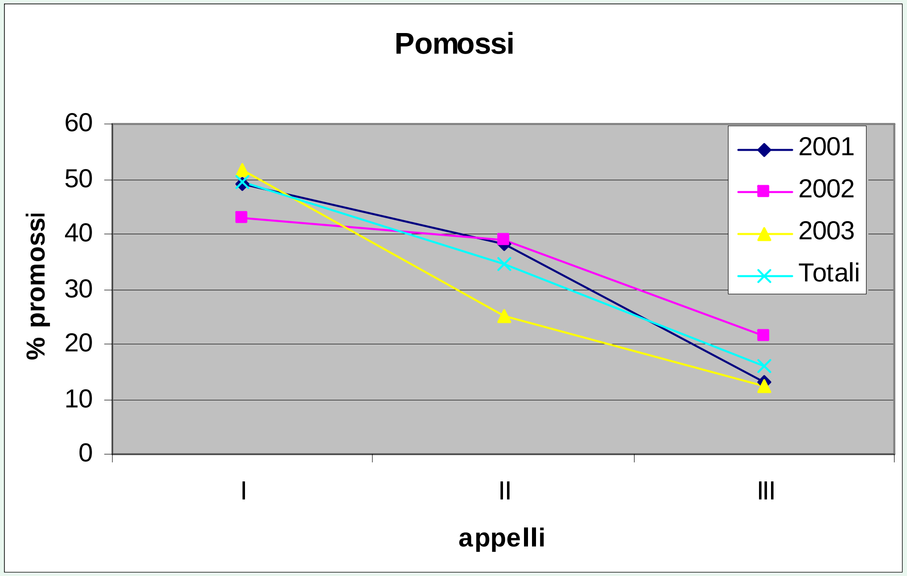
Profili di riga

anno	I	II	III	%. stud.
2001	48.91	38.04	13.04	100
2002	42.86	38.78	21.43	100
2003	51.79	25.00	12.50	100
Totali %	49.49	34.47	16.04	100

x_i

y_j

La percentuale di promossi nei diversi anni (x) dipende dall'appello (y)?



anno	I	II	III	N. stud.
2001	45	35	12	92
2002	42	38	21	98
2003	58	28	14	112
Totali	145	101	47	293

$f_{ij}=x_{ij} / N_j$

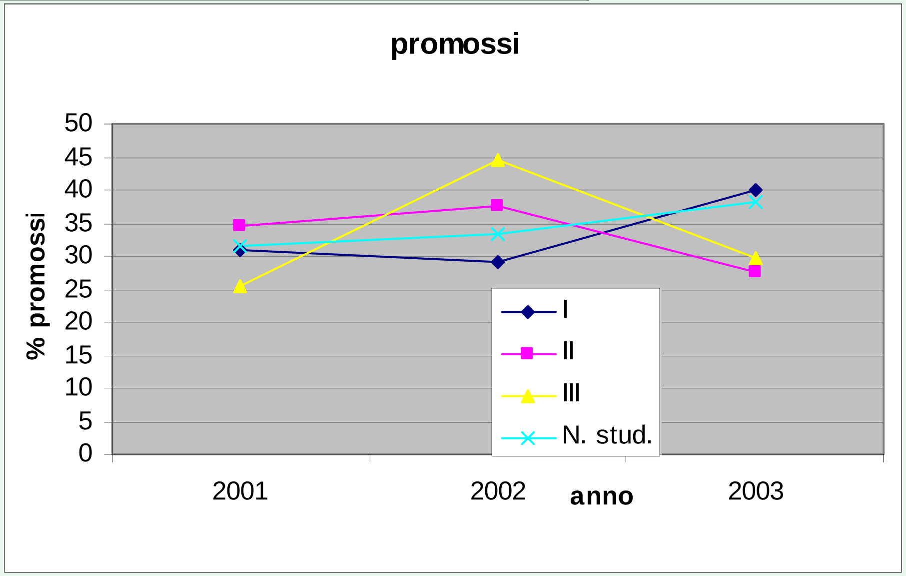
Profili di colonna

anno	I	II	III	%. stud.
2001	31.03	34.65	25.53	31.40
2002	28.97	37.62	44.68	33.45
2003	40.00	27.72	29.79	38.23
Totali %	100	100	100	100

x_i

y_j

La percentuale di promossi nei diversi appelli (y) dipende dall'anno (x)?



anno	I	II	III	N. stud.
2001	45	35	12	92
2002	42	38	21	98
2003	58	28	14	112
Totali	145	101	47	293

$F_{ij}=x_{ij}/N=f_{ij}fi$

$f_i=N_i/N$

anno	I	II	III	N. stud.
2001	15.36	11.95	4.10	31.40
2002	14.33	12.97	7.17	33.45
2003	19.80	9.56	4.78	38.23
Totali	49.49	34.47	16.04	100

N

$f_j=N_j/N$

Confronto tabelle di contingenza

anno	I	II	III	N. stud.
2001	15.36	11.95	4.10	31.40
2002	14.33	12.97	7.17	33.45
2003	19.80	9.56	4.78	38.23
Totali	49.49	34.47	16.04	100

$F_{ij}=x_{ij}/N = f_{ij} fi$

Frequenze osservate

Frequenze attese nel caso random

$F_{ij}=x_{ij}/N = f_{ij} fi$

anno	I	II	III	N. stud.
2001	15.54	10.82	5.04	31.40
2002	16.55	11.53	5.37	33.45
2003	18.92	13.18	6.13	38.23
Totali	49.49	34.47	16.04	100

Nel caso di indipendenza stocastica si la probabilità congiunta è pari al prodotto delle probabilità:

$P(A \cap B) = P(B)\, P(A)$

anno	I	II	III	N. stud.
2001	45	35	12	92
2002	42	38	21	98
2003	58	28	14	112
Totali	145	101	47	293

	Riga 1	Riga 2	Riga 3
Riga 1	1		
Riga 2	0.99	1	
Riga 3	0.91	0.85	1

	col 1	col 2	col 3
col 1	1		
col 2	-0.99	1	
col 3	-0.47	0.57	1

anno	I	II	III	N. stud.
2001	48.91	38.04	13.04	100
2002	42.86	38.78	21.43	100
2003	51.79	25.00	12.50	100
Totali	49.49	34.47	16.04	100

	Riga 1	Riga 2	Riga 3
Riga 1	1		
Riga 2	0.99	1	
Riga 3	0.91	0.85	1

	col 1	col 2	col 3
col 1	1		
col 2	-0.78	1	
col 3	-0.96	0.59	1

anno	I	II	III	N. stud.
2001	31.03	34.65	25.53	31.40
2002	28.97	37.62	44.68	33.45
2003	40.00	27.72	29.79	38.23
Totali	100	100	100	100

	Riga 1	Riga 2	Riga 3
Riga 1	1		
Riga 2	-0.55	1	
Riga 3	-0.04	-0.81	1

	col 1	col 2	col 3
col 1	1		
col 2	-0.99	1	
col 3	-0.47	0.57	1

anno	I	II	III	N. stud.
2001	15.36	11.95	4.10	31.40
2002	14.33	12.97	7.17	33.45
2003	19.80	9.56	4.78	38.23
Totali	49.49	34.47	16.04	100

	Riga 1	Riga 2	Riga 3
Riga 1	1		
Riga 2	0.99	1	
Riga 3	0.91	0.85	1

	col 1	col 2	col 3
col 1	1		
col 2	-0.99	1	
col 3	-0.47	0.57	1

anno	I	II	III	N. stud.
2001	15.54	10.82	5.04	31.40
2002	16.55	11.53	5.37	33.45
2003	18.92	13.18	6.13	38.23
Totali	49.49	34.47	16.04	100

	Riga 1	Riga 2	Riga 3
Riga 1	1		
Riga 2	1	1	
Riga 3	1	1	1

	col 1	col 2	col 3
col 1	1		
col 2	1	1	
col 3	1	1	1

Tabella teorica

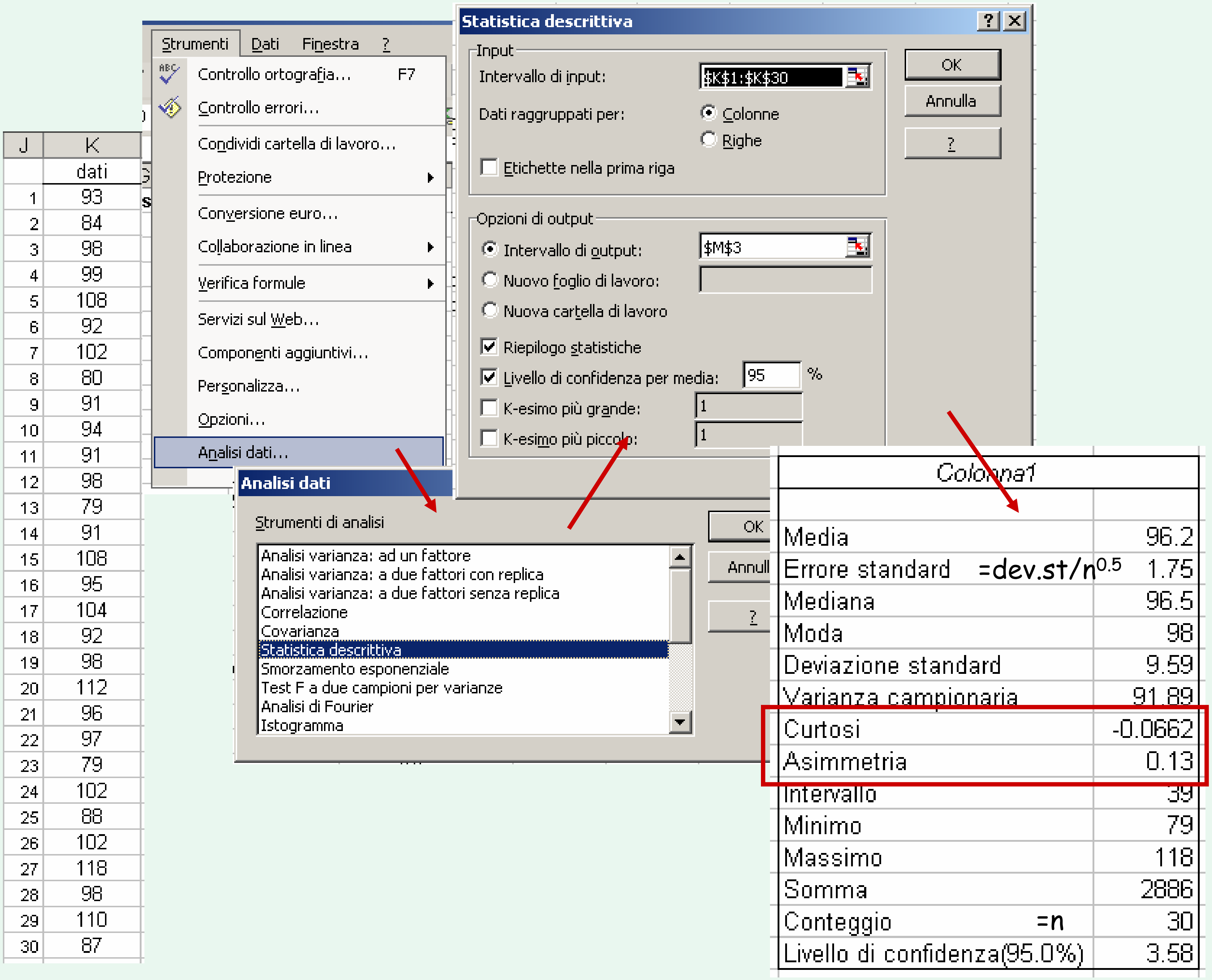
Frequenza e Probabilità

L'insieme dei valori assunti da una variabile aleatoria come risultato di esperimenti e osservazioni costituisce una "**distribuzione**"

Per ogni valore la "frequenza" è il numero di volte che questo valore compare.

La frequenza relativa è il numero di volte che questo compare, normalizzato al numero di osservazioni.

Le frequenze sono **grandezze** "**sperimentali**". La **probabilità** associata ad un dato risultato è il **risultato** di un procedimento **matematico**.


J
K
dati
1 93
2 84
3 98
4 99
5 108
6 92
7 102
8 80
9 91
10 94
11 91
12 98
13 79
14 91
15 108
16 95
17 104
18 92
19 98
20 112
21 96
22 97
23 79
24 102
25 88
26 102
27 118
28 98
29 110
30 87
Strumenti Dati Finestra ?
Controllo ortografia... F7
Controllo errori...
Condividi cartella di lavoro...
Protezione
Conversione euro...
Collaborazione in linea
Verifica formule
Servizi sul Web...
Componenti aggiuntivi...
Personalizza...
Opzioni...
Analisi dati...
Analisi dati
Strumenti di analisi
Analisi varianza: ad un fattore
Analisi varianza: a due fattori con replica
Analisi varianza: a due fattori senza replica
Correlazione
Covarianza
Statistica descrittiva
Smorzamento esponenziale
Test F a due campioni per varianze
Analisi di Fourier
Istogramma
OK
Annull
?
Statistica descrittiva
Input
Intervallo di input: $K$1:$K$30
Dati raggruppati per: Colonne Righe
Etichette nella prima riga
OK
Annulla
?
Opzioni di output
Intervallo di output: $M$3
Nuovo foglio di lavoro:
Nuova cartella di lavoro
Riepilogo statistiche
Livello di confidenza per media: 95 %
K-esimo più grande: 1
K-esimo più piccolo: 1
Colonna1
Media 96.2
Errore standard =dev.st/n^0.5 1.75
Mediana 96.5
Moda 98
Deviazione standard 9.59
Varianza campionaria 91.89
Curtosi -0.0662
Asimmetria 0.13
Intervallo 39
Minimo 79
Massimo 118
Somma 2886
Conteggio =n 30
Livello di confidenza(95.0%) 3.58

Frequenza e Probabilità

Come determinare la probabilità di un evento?

Insieme infinito di "prove"

$$p_i = \lim_{N \to \infty} \frac{n_i}{N} = \lim_{N \to \infty} f_i$$

$$\sum_{i=1}^{M} p_i = 1$$

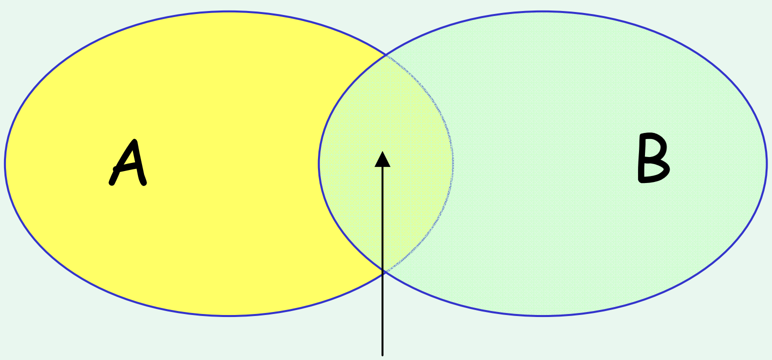
$$p_i \geq 0$$

Calcolo delle probabilità

P(A): probabilità dell'evento A

P(B): probabilità dell'evento B

$$P(A+B) = P(A \cup B) = P(A) + P(B) - P(A \cap B)$$



AB, $A \cap B$,
intersezione

se $P(A \cap B) = 0$ gli eventi si dicono indipendenti

Probabilità composta

$$P(A \cap B) = P(B)\, P(A|B) = P(A)\, P(B|A)$$

per cui la probabilità che due eventi A e B si verifichino contemporaneamente è pari alla probabilità di uno dei due eventi moltiplicato con la probabilità dell'altro evento condizionato al verificarsi del primo.

Nel caso di indipendenza stocastica [$P(B|A) = P(B)$] la probabilità congiunta è pari al prodotto delle probabilità:

$$P(A \cap B) = P(B)\, P(A)$$

Lacio di 2 dadi

Risultati possibili

m_{ij}

2 - 1,1	P_2 = 1/36
3 - 1,2 2,1	P_3 = 2/36
4 - 1,3 2,2 3,1	P_4 = 3/36
5 - 1,4 2,3 3,2 4,1	P_5 = 4/36
6 - 1,5 2,4 3,3 4,2 5,1	P_6 = 5/36
7 - 1,6 2,5 3,4 4,3 5,2 6,1	P_7 = 6/36
8 - 2,6 3,5 4,4 5,3 6,2	P_8 = 5/36
9 - 3,6 4,5 5,4 6,3	P_9 = 4/36
10 - 4,6 5,5 6,4	P_{10} = 3/36
11 - 5,6 6,5	P_{11} = 2/36
12 - 6,6	P_{12} = 1/36

$$\boxed{p_i = \frac{1}{6}}$$

$$\boxed{p_{ij} = m_{ij} p_i p_j = \frac{m_{ij}}{36}}$$

Distribuzione di Bernulli

Descrive una variabile casuale che può assumere solo valori 0,1: X=[0,1]

$P(1) = P(X=1) = p$ $\qquad$ $P(0) = P(X=0) = 1-p$

$$\mu = p$$

$$\sigma^2 = p(1-p)$$

Distribuzione Binomiale

Descrive una variabile casuale X che rappresenta il numero di successi su *n* prove, ognuna con probabilità di successo *p* (X è una sommadi variabili casuali di tipo "bernulli")

$$p(k) = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

k=numero di successi,
n= numero di prove

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

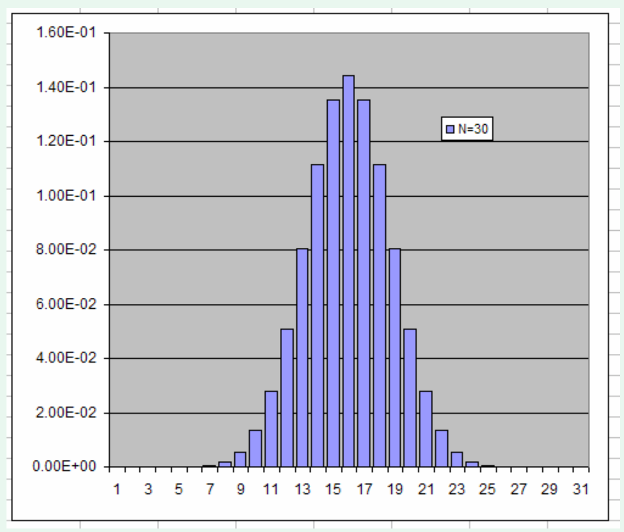
$$\mu = np \qquad \sigma = \sqrt{npq}$$

Distribuzione Binomiale (Bernulli)

Pb.: Lancio 5 (N=5) volte una moneta, quale è la probabilità di avere 3 (k=3) teste

1	●	●	●	○	○
2	○	●	●	●	○
3	○	○	●	●	●
4	●	●	○	●	○
5	●	●	○	○	●
6	○	●	●	○	●
7	●	○	●	●	○
8	○	●	○	●	●
9	●	○	○	●	●
10	●	○	●	○	●

$$P(N,k) = \begin{pmatrix} N \\ k \end{pmatrix} p^k \; q^{N-K} = \frac{N!}{k!(N-k)!} p^k \; (1-p)^{N-k}$$



Valore atteso

$$\mu = np$$

Dev. St.

$$\sigma = \sqrt{npq}$$

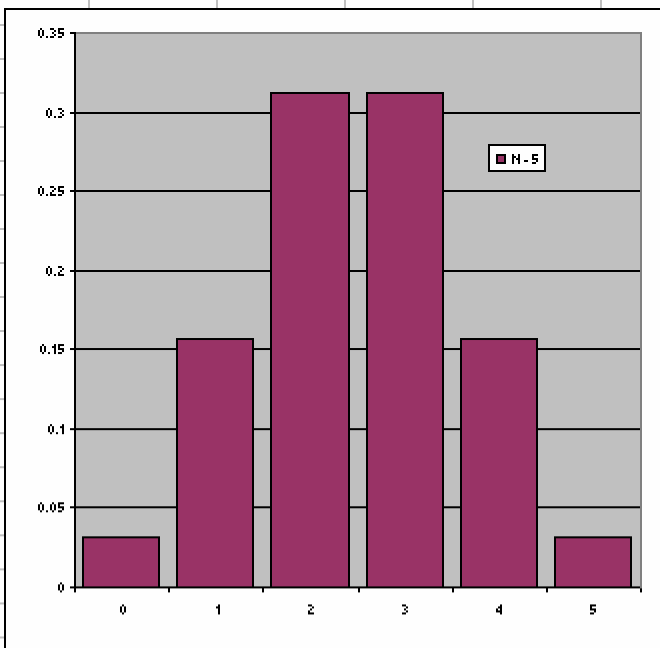
Distribuzione Binomiale (Bernulli)

p = q = 0.5

$$P(N,k) = \begin{pmatrix} N \\ k \end{pmatrix} p^k \; q^{N-K} =$$

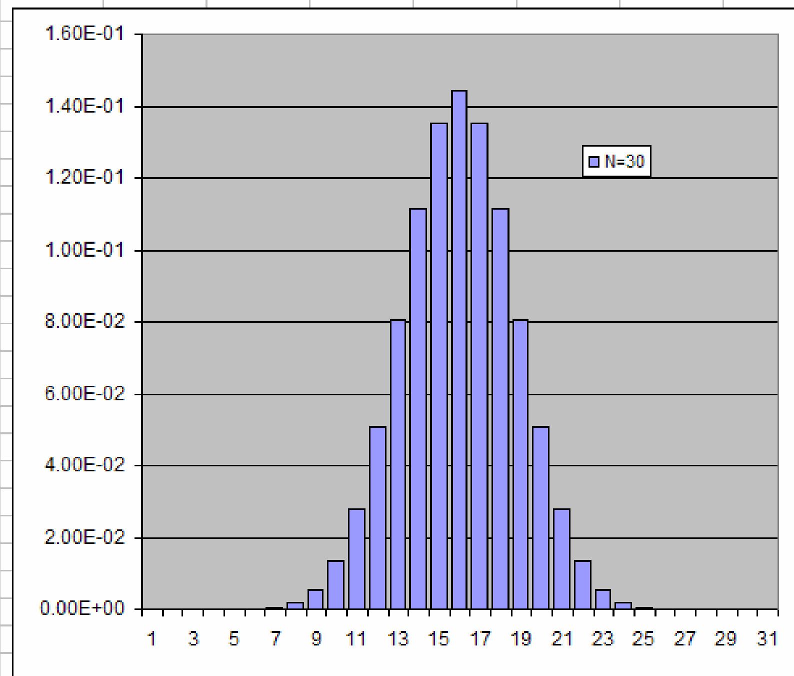
N=5

k	p
0	0.03125
1	0.15625
2	0.3125
3	0.3125
4	0.15625
5	0.03125
N	5



N=30

k	p
0	9.31E-10
1	2.79E-08
2	4.05E-07
3	3.78E-06
4	2.55E-05
5	1.33E-04
6	5.53E-04
7	1.90E-03
8	5.45E-03
9	1.33E-02
10	2.80E-02
11	5.09E-02
12	8.06E-02
13	1.12E-01
14	1.35E-01
15	1.44E-01
16	1.35E-01
17	1.12E-01
18	8.06E-02
19	5.09E-02
20	2.80E-02
21	1.33E-02
22	5.45E-03
23	1.90E-03
24	5.53E-04
25	1.33E-04
26	2.55E-05
27	3.78E-06
28	4.05E-07
29	2.79E-08
30	9.31E-10
N	30



Distribuzione di Poisson

La variabile stocastica X può assumere valori discreti X={0,1,2,3,4...}

Distribuzione Binomiale con:	$N \gg 1$ $p \ll 1$

$$\boxed{p(k) = \frac{\lambda^k}{k!} e^{-k}}$$

Valore atteso: $\mu = \lambda$

varianza: $\sigma^2 = \lambda$

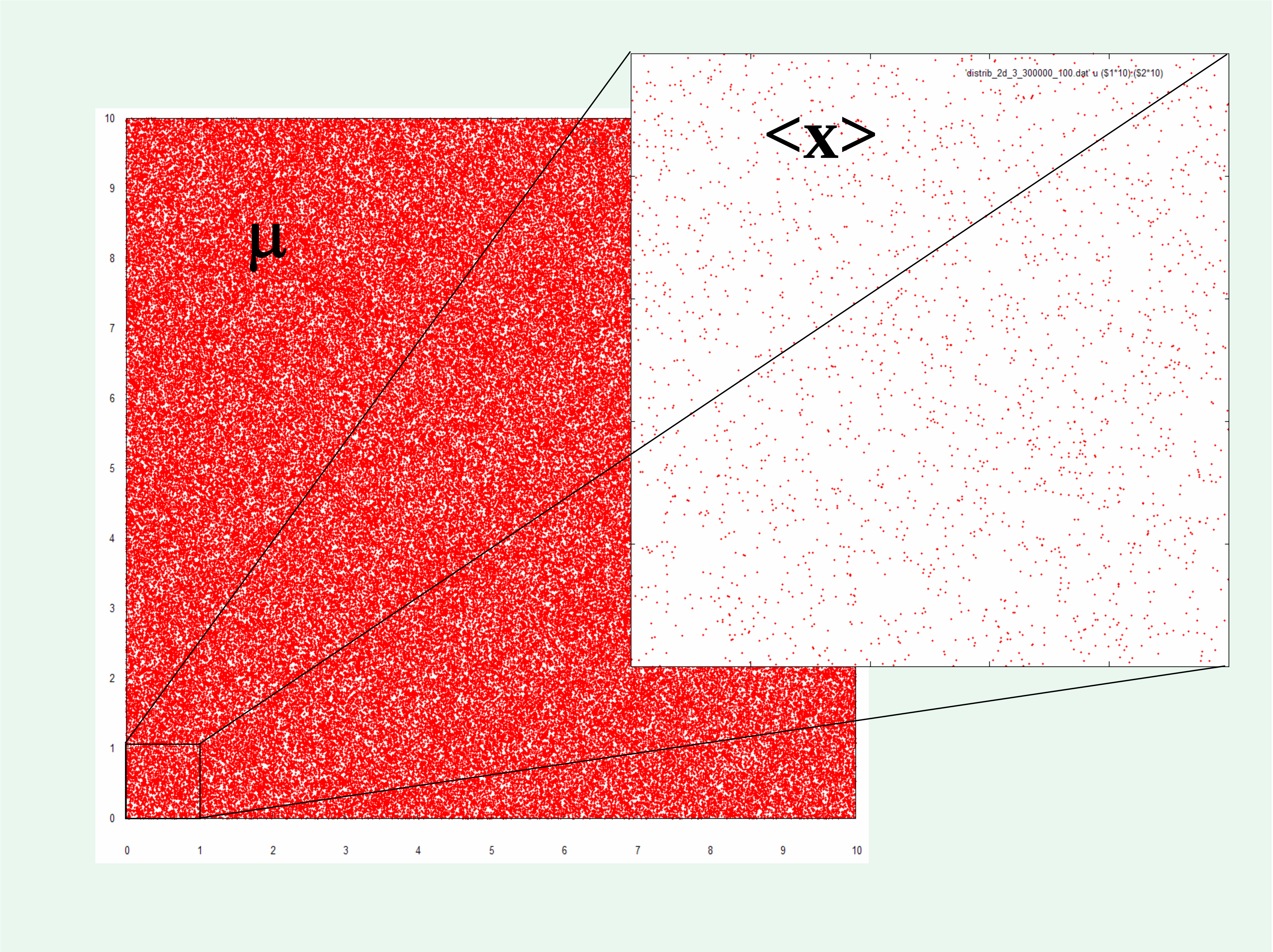
Puó essere usata per descrivere il numero di cellule in una data area, il numero di errori di battitura per pagina, etc...

Descrive distribuzioni di oggetti caratterizzati da:

- densità costante (numero di oggetti proporzionale alla dimensione della regione di campionamento (superficie, volume, lunghezza etc..)

- i conteggi in regioni disgiunte sono indipendenti

- i numero di conteggi tende a zero se le dimensioni della regione tendono a zero.



μ
<x>
'distrib_2d_3_300000_100.dat' u ($1*10):($2*10)
0
1
2
3
4
5
6
7
8
9
10

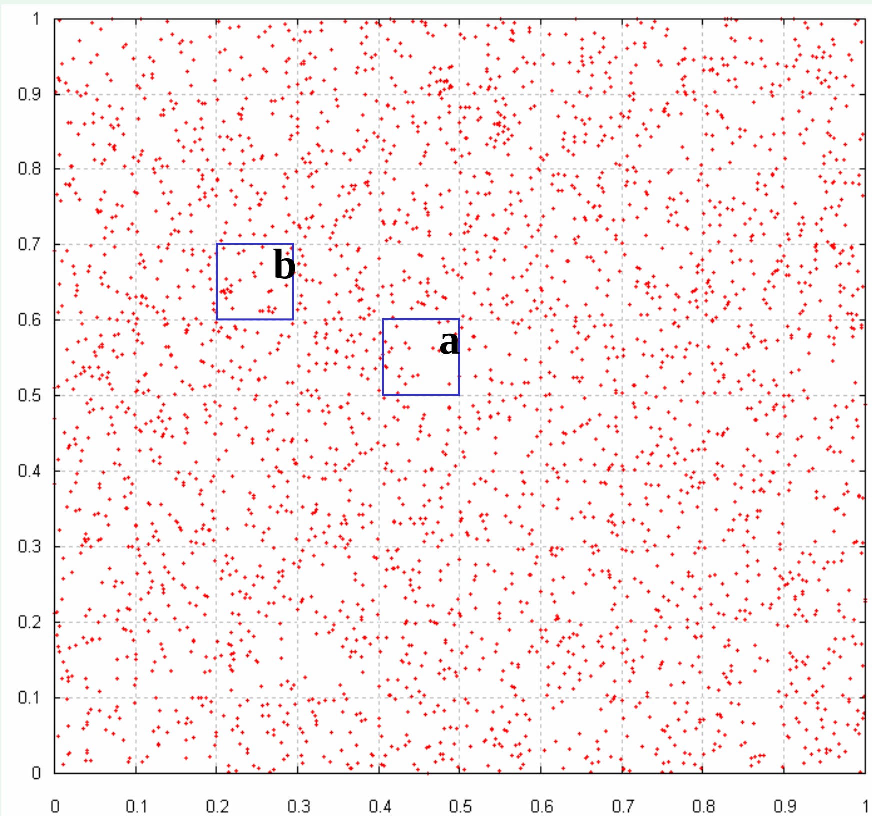
Distribuzioni di probabilità notevoli

Distribuzione Poisson

Distribuzione Binomiale con: $N \gg 1$, $p \ll 1$

$$p(k) = \frac{\lambda^k}{k!} e^{-k}$$

Valore atteso: $\mu = \lambda$



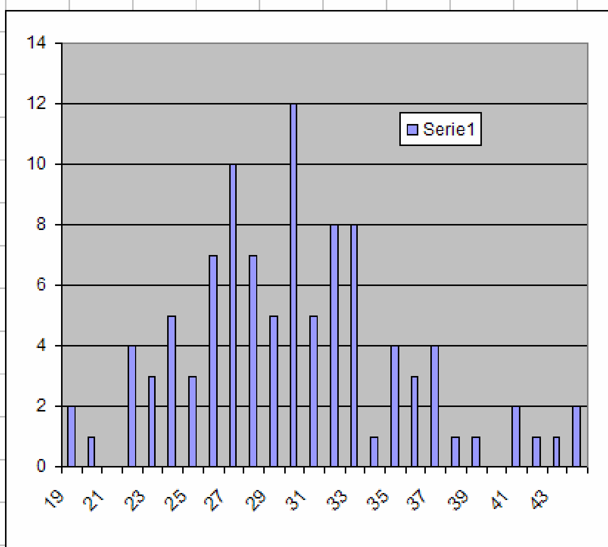
1	27	29	22	31	30	26	28	30	28	26
0.9	33	24	32	28	27	23	32	27	37	33
0.8	29	28	32	22	30	30	35	25	44	29
0.7	30	33	33	27	27	26	39	33	27	31
0.6	41	23	29	36	27	30	30	28	25	31
0.5	22	41	36	24	25	33	30	37	37	42
0.4	32	27	31	20	32	35	35	30	19	44
0.3	30	22	29	38	26	28	27	30	32	34
0.2	37	33	26	26	32	24	35	23	31	28
0.1	33	32	30	26	24	19	27	24	36	43
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

1	27	29	22	31	30	26	28	30	28	26
0.9	33	24	32	28	27	23	32	27	37	33
0.8	29	28	32	22	30	30	35	25	44	29
0.7	30	33	33	27	27	26	39	33	27	31
0.6	41	23	29	36	27	30	30	28	25	31
0.5	22	41	36	24	25	33	30	37	37	42
0.4	32	27	31	20	32	35	35	30	19	44
0.3	30	22	29	38	26	28	27	30	32	34
0.2	37	33	26	26	32	24	35	23	31	28
0.1	33	32	30	26	24	19	27	24	36	43
0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1

STATISTICHE	
MASSIMO	44
MINIMO	19
MEDIA	30
VARIANZA	28.85
DEV.ST.	5.371

CAT.	n_i	f_i
19	2	0.02
20	1	0.01
21	0	0.00
22	4	0.04
23	3	0.03
24	5	0.05
25	3	0.03
26	7	0.07
27	10	0.10
28	7	0.07
29	5	0.05
30	12	0.12
31	5	0.05
32	8	0.08
33	8	0.08
34	1	0.01
35	4	0.04
36	3	0.03
37	4	0.04
38	1	0.01
39	1	0.01
40	0	0.00
41	2	0.02
42	1	0.01
43	1	0.01
44	2	0.02
Totale	100	1

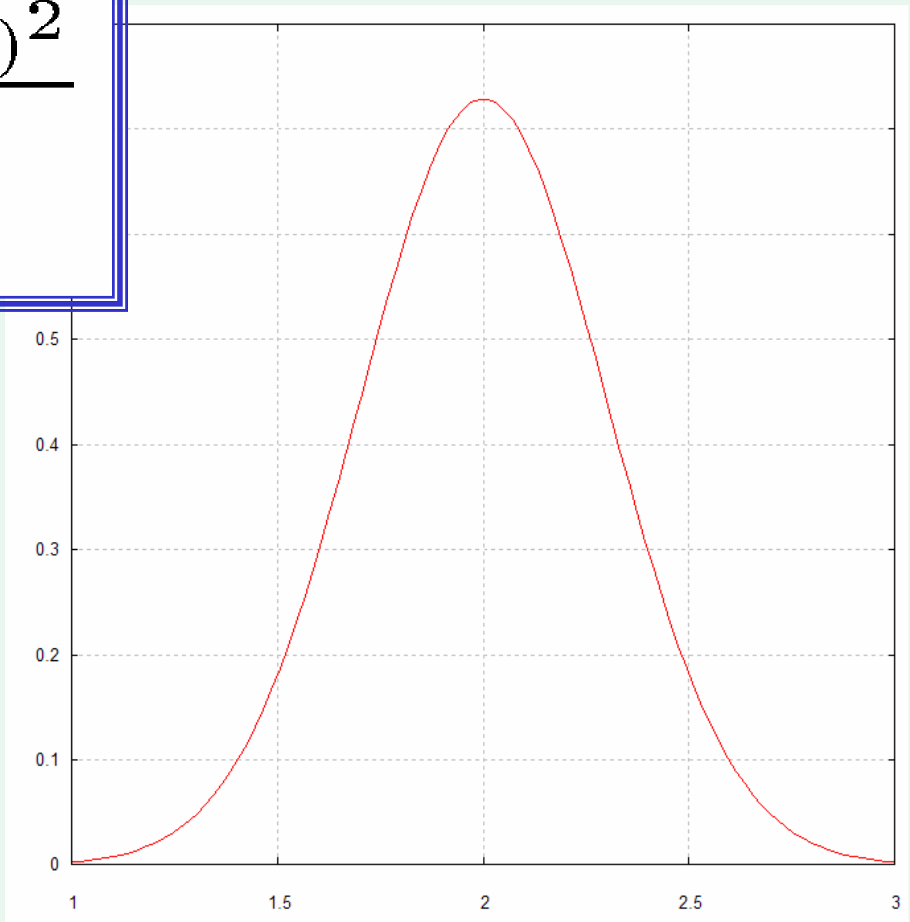
STATISTICHE		STATISTICHE	
MASSIMO	41	MASSIMO	44
MINIMO	22	MINIMO	23
MEDIA	29.2	MEDIA	30.12
VARIANZA	18.58	VARIANZA	23.03
DEV.ST.	4.311	DEV.ST.	4.799
STATISTICHE		**STATISTICHE**	
MASSIMO	41	MASSIMO	38
MINIMO	20	MINIMO	19
MEDIA	29.36	MEDIA	28.72
VARIANZA	28.91	VARIANZA	24.71
DEV.ST.	5.376	DEV.ST.	4.971



Globuli_rossi.xls

Distribuzione di Gauss o Normale

$$\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$



$\bar{x}$ Valor medio

σ Dev. standard

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

Distribuzioni continue

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x)dx$$

probabilità

Densità di probabilità

$$\overline{x} = \int_{x_1}^{x_2} x f(x)dx$$

$$\sigma^2 = \int_{x_1}^{x_2} (x - \overline{x})^2 f(x)dx$$

$$M_z = \int_{x_1}^{x_2} (x - \overline{x})^z f(x)dx$$

Distribuzione uniforme

$$f(x) = \frac{1}{B-A} \qquad A \leq x \leq B$$

$$\overline{x} = \frac{B+A}{2}$$

$$\sigma^2 = \frac{|B-A|}{\sqrt{12}}$$

I risultati di un esperimento sono variabili aleatorie.

Un esperimento non consente di esaminare ogni elemento di una popolazione o di effettuare tutte le misure possibili.

campione

Popolazione

Stime

parametri

$\langle x \rangle$, $\bar{x}$

s

$\mu \quad \sigma$

Teorema del limite centrale

La distribuzione delle medie campionari ($\langle x \rangle_i$) segue una distribuzione normale indipendentemente dalla distribuzione della popolazione d'origine

Il valor medio della distribuzione delle media campionarie è uguale alla media della popolazione d'origine

La deviazione standard dell'insieme di tutte le medie campionarie (errore standard della media σ_x) è una funzione della deviazione standard della popolazione originaria e del numero di elementi del campione.

$$\frac{1}{\sqrt{2\pi\sigma_{\bar{x}}^2}} e^{-\frac{(\bar{x}_i - \mu)^2}{2\sigma_{\bar{x}}^2}}$$

nota: dev.st. della popolazione

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}}$$

Dato un campione *n* estratto da una popolazione N è possibile fornire una stima (<x>, s) dei parametri reali della distribuzione (m, s).

I risultati ottenuti su un campione rappresentano una stima dei valori "veri"

I valori stimati sono variabili aleatorie

Quanto sono accurate queste stime?

Popolazione

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{k=1}^{n_c} x_k p(x_k)$$ Valore atteso (media)

Varianza $$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 = \sum_{k=1}^{n_c} p(x_k)(x_k - \mu)^2$$

Campione

$$\bar{x} = \frac{1}{m}\sum_{j=1}^{m} x_j = \sum_{k=1}^{n_c} x_k f(x_k)$$ Media campionaria

Varianza campionaria $$s^2 = \frac{1}{m-1}\sum_{j=1}^{m}(x_j - \bar{x})^2$$

L’errore standard della media

$$\sigma_{\bar{x}}$$

indica il grado di incertezza da associare alla stima della media ottenuta utilizzando un campione dell’intera popolazione

Accuratezza delle stime

Il valor medio ottenuto da **un** solo campione di **m** elementi è una stima del valore aspettato della popolazione.

L'errore standard della media rappresenta una stima dell'errore fatto nella stima del valore atteso.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}} \sim \frac{s}{\sqrt{m}}$$

Risultato di un'osservazione:

$$\bar{x} \pm \sigma_{\bar{x}}$$

Accuratezza delle stime

Per migliorare la stima del valore atteso si puó ripetere l'esperimento utilizzando K campioni indipendenti

In questo caso la migliore stima del valore atteso è la media delle medie campionarie:

$$\bar{x} = \frac{1}{K}\sum_{i=1}^{K}\bar{x}_i$$

Utilizzando K campioni indipendenti l'errore standard della media è:

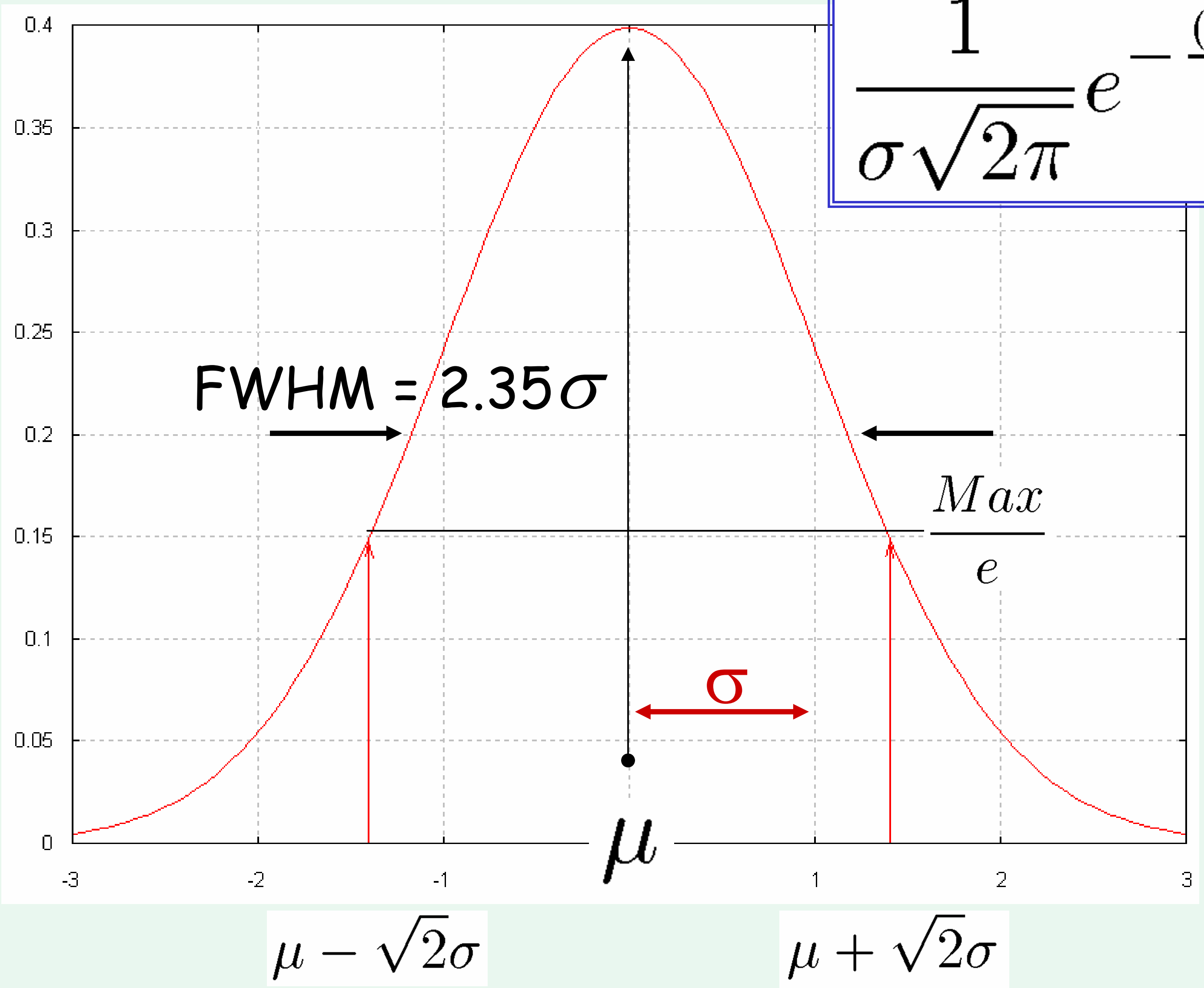
$$\sigma_{\bar{x}}^2 = \frac{1}{K}\sum_{i=1}^{K}(\bar{x}_i - \bar{x})^2$$

Risultato di un'osservazione:

$$\bar{x} \pm \sigma_{\bar{x}}$$

Proprietà della distribuzione di Gauss

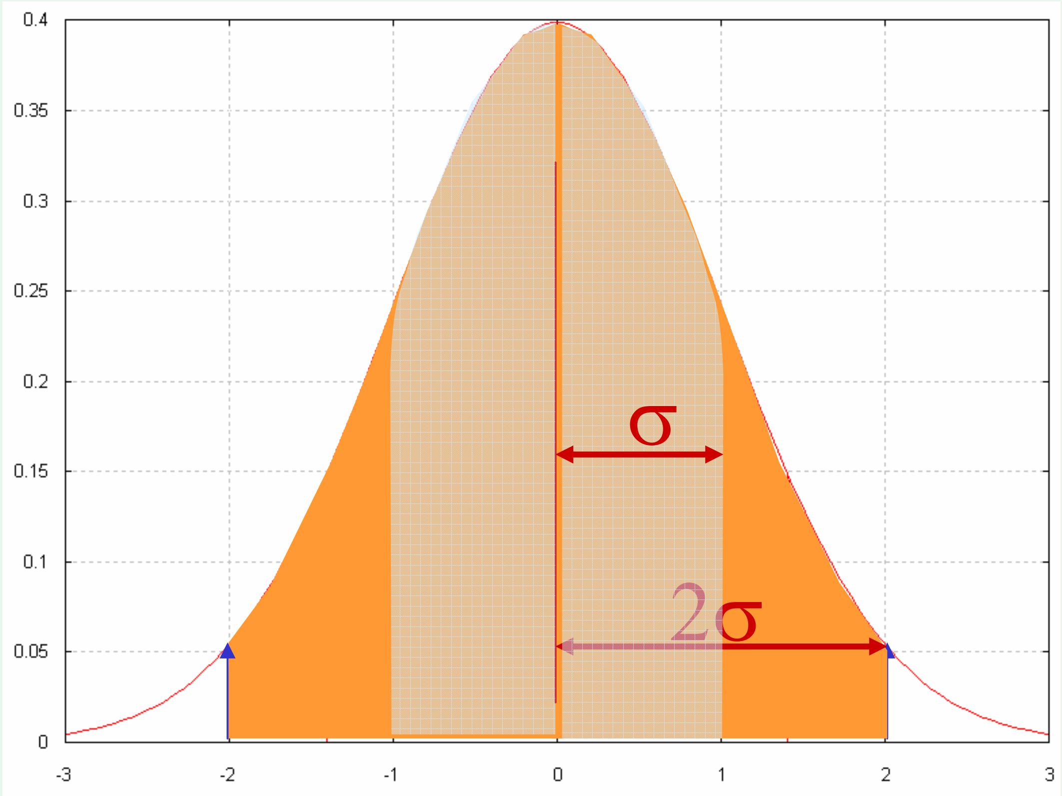
$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



$$\int_{\mu-\sigma}^{\mu+\sigma} g(x)dx = 0.68$$

$$\int_{\mu-2\sigma}^{\mu+2\sigma} g(x)dx = 0.95$$

$$\int_{\mu-3\sigma}^{\mu+3\sigma} g(x)dx = 0.997$$



Date due variabili aleatorie indipendenti X_a, X_b caratterizzate da μ_a σ_a, μ_b σ_b, la variabile $Z= X_a+X_b$ è una variabile aleatoria con:

$$\mu_z = \mu_a+\mu_b$$

$$\sigma_z = \sigma_a+\sigma_b$$

Teorema del limite centrale

La distribuzione delle medie campionarie $\bar{x}$ su campioni di m elementi segue una distribuzione **normale** indipendentemente dalla distribuzione della popolazione d'origine

$$\frac{1}{\sqrt{2\pi\sigma_{\bar{x}}^2}} e^{-\frac{(\bar{x}_i-\mu)^2}{2\sigma_{\bar{x}}^2}}$$

La variabile: $t = \dfrac{\bar{x}-\mu}{\sigma_{\bar{x}}}$

è una variabile aleatoria che, per m molto grande, ha una distribuzione Normale Standard (ha media nulla e varianza unitaria):

$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

Confidenza

Pb.: un'osservazione su un campione di m elementi fornisce come risultato il valor medio $\bar{x}$ di una variabile aleatoria.

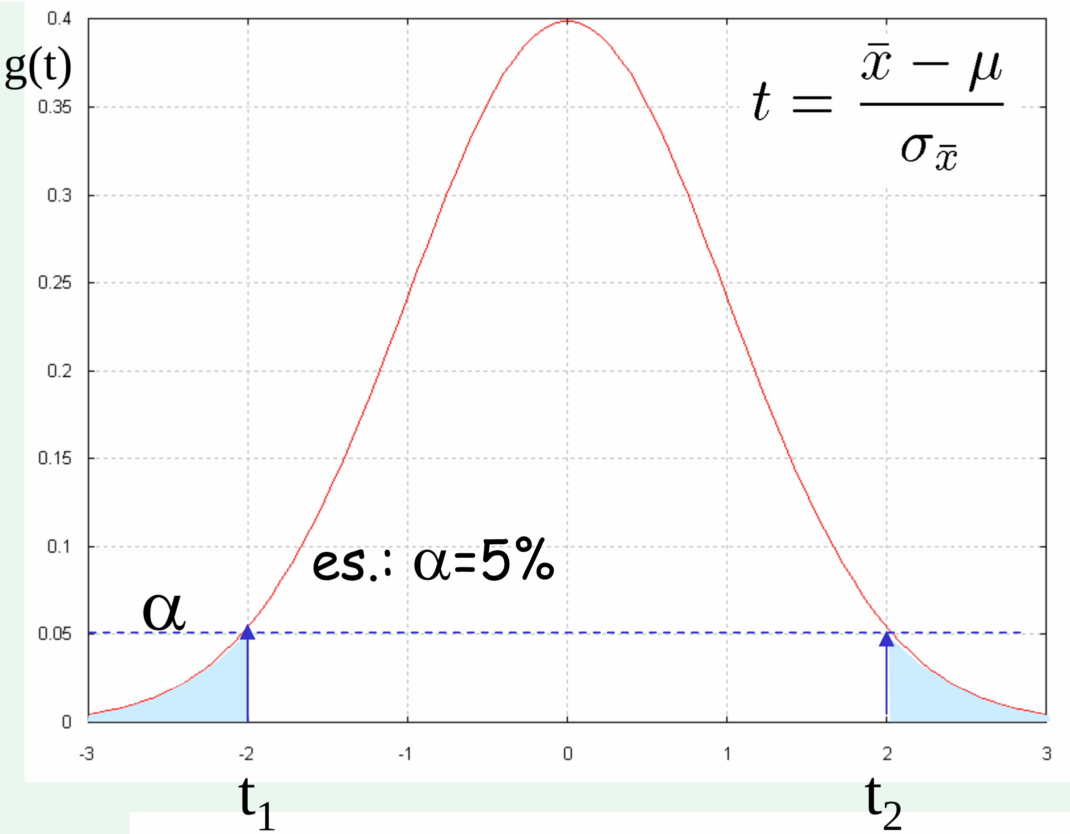
1) costruisco una variabile aleatoria con distribuzione nota, es.:

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \qquad g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

2) sulla base della g(t) determino i valori di t che hanno una bassa probabilità di essere osservati, cioè:

- fisso un livello di confidenza α.
- determino un intervallo di valori $t_{\alpha 1}$–$t_{\alpha 2}$ (intervallo di confidenza) tale che la probabilità di osservare t all'esterno dell'intervallo dato sia minore di α

$$P(t < \alpha) = P(t < t_1) + P(t < t_1) = \int_{-\infty}^{t_1} g(t)dt + \int_{t_2}^{\infty} g(t)dt = \alpha$$



Se $t_1 < t < t_2$ la probabilità di osservare il valore di t, calocolato in base ai dati, è (1-α)

Se $t < t_1$ o $t > t_2$ la probabilità di osservare il valore di t, calocolato in base ai dati, è α

$$P(t > \alpha) = P(t_1 < t < t_1) = \int_{t_1}^{t_2} g(t)dt = 1 - \alpha$$

probabilità che t appartenga all'intervallo t_1 - t_2

$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$P(t_1 < t < t_1) = P(\bar{x} - \sigma_{\bar{x}} t_1 < \mu < \bar{x} + \sigma_{\bar{x}} t_1) = 1 - \alpha$$

dato il valore medio $\bar{x}$, osservato su un campione di **m** elementi, il valore aspettato della popolazione (μ) è contenuto nell'intervallo:

$$[\bar{x} - \sigma_{\bar{x}} t_1 ; \bar{x} + \sigma_{\bar{x}} t_1]$$

con probabilità 1-α.

Se le osservazioni sono distribuite con:

valor medio $\bar{x}$

dev.st. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{m}}$

funzione EXCEL:

CONFIDENZA(α, dev.st, m)

Intervalli di confidenza: varianza nota

Se le osservazioni sono distribuite con:

valor medio $\bar{x}$

dev.st. $\sigma_{\bar{x}} = \dfrac{\sigma}{\sqrt{m}}$

funzione EXCEL: **CONFIDENZA(α, dev.st, m)**

La resistenza elettrica di un cavo viene misurata con uno strumento che ha un'incertezza σ=0.5 Ω. Vengono effettuate 5 misure, ne risulta un valor medio $\bar{R}$=4.52 Ω

CONFIDENZA(0.05, 0.5, 5)=0.438

La resistenza vera del cavo è nell'intervallo :

R = 4.52 $\pm$ 0.44 Ω   oppure: R = [4.08, 4.96]

Nota: α rappresenta il rischio di sbagliare, cioè la probabilità che il valore vero della resistenza sia esterno all'intervallo dato

Intervalli di confidenza: varianza campionaria

Molto piú spesso non conosco la varianza della distribuzione. La migliore stima della varianza in un campione di m elementi è:

$$\sigma_{\bar{x}} = \frac{s}{\sqrt{m}}$$ da cui: $$t = \frac{\bar{x} - \mu}{\sqrt{s^2/m}}$$

$$P(\bar{x} - t_1 \frac{s}{\sqrt{m}} < \mu < \bar{x} - t_1 \frac{s}{\sqrt{m}}) = 1 - \alpha$$

probabilità che il valore vero (μ) sia nell'intervallo:

$$\left[\bar{x} - t_1 \frac{s}{\sqrt{m}}; \bar{x} + t_1 \frac{s}{\sqrt{m}}\right]$$

funzione EXCEL: INV.T(a,m-1) = $t_1 \frac{s}{\sqrt{m}}$.

Nota: la variabile t cosi definita ha una distribuzione nota (t-Student) con ν = m-1 gradi di libertà. La t-Student approssima una distribuzione Gaussiana per ν che tende a infinito

Una misura dell'altezza di un gruppo di 20 studenti fornisce il valore medio: H = 1.68 m con la deviazione standard stimata s = 9 cm.

Determinare l'intervallo di confidenza dell'1%

Dati				
numero di osservazioni: m		20		
valor medio		1.68		
dev. standard: s		0.09		
Confidenza	α	0.01		
	t_a	2.86		
μ	1.68	±	0.06	

inv.T(α; m-1)

$$t_\alpha \frac{s}{\sqrt{m}}$$

valore medio

Test di ipotesi, test statistici

1) ipotesi da verificare

Es.: il valore misurato è compatibile con il valore vero?

2) costruisco una variabile aleatoria con distribuzione nota, es.:

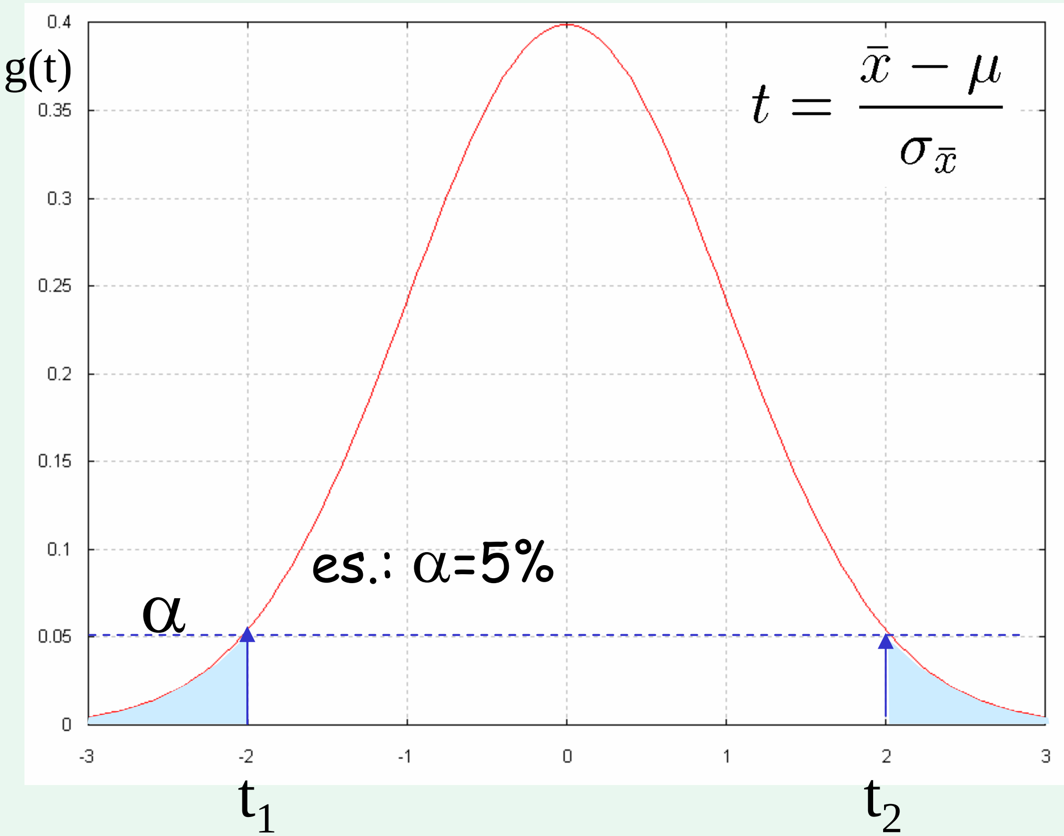
$$t = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}}$$

$$g(t) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

3) sulla base della g(t) determino i valori di t che hanno una bassa probabilità di essere osservati, fissando il livello di confidenza α. Se, in base alla distribuzione scelta, il valore osservato fornisce un valore di t con bassa probabilità di essere osservato, l'ipotesi deve essere rifiutata, altrimenti puó essere accettata.

$$P(t<\alpha) = P(t<t_1) + P(t<t_1) = \int_{-\infty}^{t_1} g(t)dt + \int_{t_2}^{\infty} g(t)dt = \alpha$$

$$P(t>\alpha) = P(t_1<t<t_1) = \int_{t_1}^{t_2} g(t)dt = 1-\alpha$$



~~Se $t_1 < t < t_2$ il risultato (cui è associato il valore t) è compatibile l'ipotesi fatta con una probabilità del P = (1-α)~~

Se $t < t_1$ o $t > t_2$ il risultato (cui è associato il valore t) non è compatibile l'ipotesi fatta con una probabilità del P = (1-α)

<u>Nota</u>: α rappresenta la probabilità di sbagliare e scartare un'ipotesi corretta.