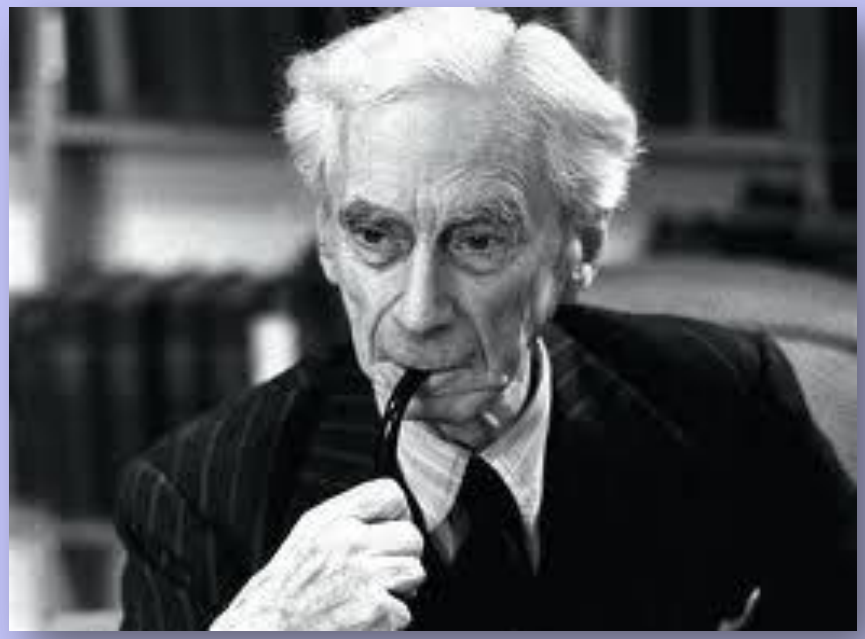
Seminario 1

- Definizione di probabilità
- Popolazione campione e tecniche di campionamento
- Intervallo di confidenza
- Statistica descrittiva

“Il concetto di probabilità è il più importante della scienza moderna, soprattutto perché nessuno ha la più pallida idea del suo significato.”

- Bertrand Russel -



PROBABILITA’

Definizione classica

la probabilità di un evento è il rapporto tra il numero dei casi favorevoli all'evento e il numero dei casi possibili, purché questi ultimi siano tutti equiprobabili.

1. **la probabilità di un evento è un numero compreso tra 0 e 1;**
2. **la probabilità dell'evento certo è pari a 1;**
3. **la probabilità del verificarsi di uno di due eventi incompatibili, ovvero di due eventi che non possono verificarsi simultaneamente, è pari alla somma delle probabilità dei due eventi.**

PROLEMI:

- è una definizione circolare: richiede che i casi possiedano tutti la medesima probabilità, che è però ciò che si vuole definire;

- non definisce la probabilità in caso di eventi non equiprobabili;

- presuppone un numero finito di risultati possibili e di conseguenza non è utilizzabile nel continuo.

PROBABILITA’

Definizione frequentista

La probabilità di un evento è il limite cui tende la frequenza relativa dell'evento al crescere del numero degli esperimenti

- La definizione frequentista si applica ad esperimenti casuali i cui eventi elementari non siano ritenuti ugualmente possibili, ed assume che l'esperimento sia ripetibile più volte, idealmente infinite, sotto le stesse condizioni.

ma…

non tutti gli esperimenti sono ripetibili!!!

PROBABILITA’

Definizione soggettiva

***la probabilità di un evento è il prezzo che un individuo ritiene equo pagare per ricevere 1 se l'evento si verifica, 0 se l'evento non si verifica*.**

Le probabilità degli eventi devono essere attribuite in modo tale che non sia possibile ottenere una vincita o una perdita certa.

La definizione soggettiva consente di calcolare la probabilità di eventi anche quando gli eventi elementari non sono equiprobabili e quando l'esperimento non può essere ripetuto.

Però singoli individui, verosimilmente, avranno diverse propensioni al rischio.

1 numero = 6/90	0,066667	prob 1
2 numero = 5/89	0,05618	prob 2
3 numero = 4/88	0,045455	prob 3
4 numero = 3/87	0,034483	prob 4
5 numero = 2/88	0,023256	prob 5
6 numero = 1/87	0,011765	prob 6

prob TOT (prob 1*2*3*4*5*6) = 1,61E-09 = 1/622.614.630

se due possibilità a schedina: prob = 1/311.307.315

montepremi/probabilità > costo schedina

montepremi = 1E+08

valore schedina = 0,321226

PROBABILITA’

Definizione assiomatica

Dato un qualsiasi esperimento casuale, i suoi possibili risultati costituiscono gli elementi di un insieme non vuoto Ω, detto spazio campionario, e ciascun evento è un sottoinsieme di Ω. La probabilità viene vista, in prima approssimazione, come una misura, cioè come una funzione che associa a ciascun sottoinsieme di Ω un numero reale non negativo tale che la somma delle probabilità di tutti gli eventi sia pari a 1.

- Gli eventi sono sottoinsiemi di uno spazio Ω.
- Ad ogni evento è assegnato un numero reale non negativo $P(A)$, detto *probabilità di A*.
- $P(\Omega)=1$, ovvero la probabilità dell'evento certo è pari ad 1.
- Se l'intersezione tra due eventi A e B è vuota, allora $P(A \cup B)=P(A)+P(B)$.

Va notato che la definizione assiomatica non è una *definizione operativa* e non fornisce indicazioni su *come* calcolare la probabilità.

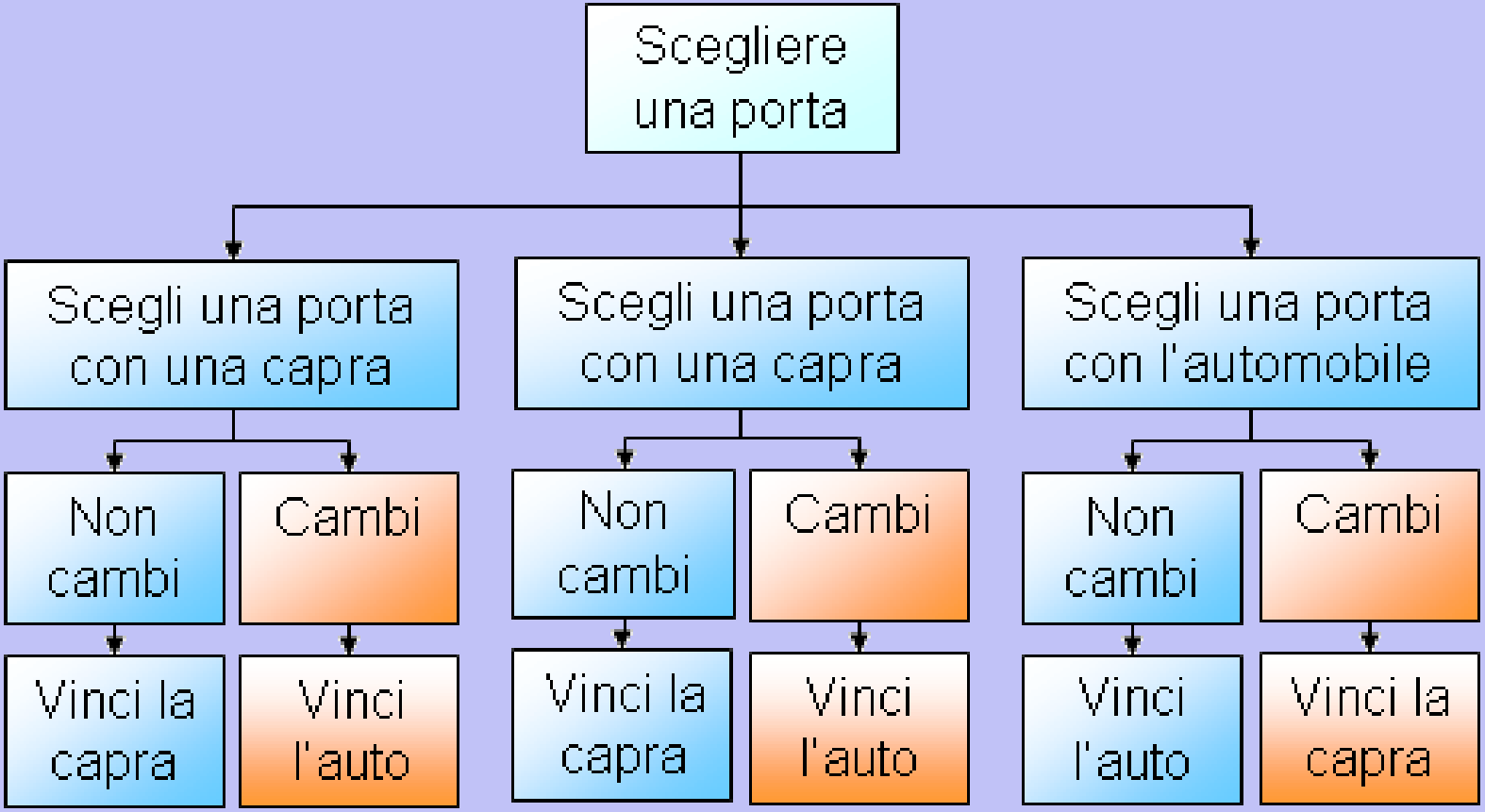
In conclusione …

1) mancanza di una definizione/criterio di misura univoco;

2) talvolta controintuitiva.

Problema di Monty Hall

SOLUZIONE


Scegliere una porta
Scegli una porta con una capra
Scegli una porta con una capra
Scegli una porta con l'automobile
Non cambi
Cambi
Non cambi
Cambi
Non cambi
Cambi
Vinci la capra
Vinci l'auto
Vinci la capra
Vinci l'auto
Vinci l'auto
Vinci la capra

Teorema di Bayes

Qualche definizione:

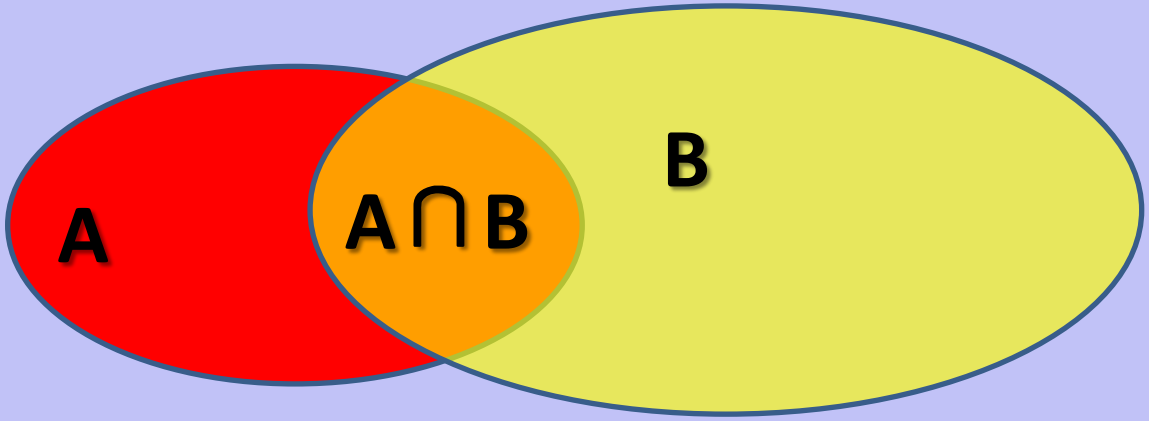
Mettiamo di avere due eventi, A e B

Probabilità A = P(A)

Probabilità B = P(nonA) = 1 – P(A)

Probabilità condizionata = probabilità che si verifichi A a condizione che si sia verificato B = P(A|B)

Probabilità congiunta = probabilità che si verifichino sia A che B = P(A ∩ B)



Teorema di Bayes

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

ovvero

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

Teorema delle probabilità composte

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

Teorema di Bayes

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Rivediamo il problema di Monty Hall…

Mettiamo che la porta 3 sia stata aperta dal conduttore mostrando una capra e che il concorrente abbia selezionato la porta 1.
La probabilità che l'automobile si trovi dietro la porta 2 (ovvero la probabilità di trovare l'auto dopo aver cambiato la scelta iniziale) è :

$$1\text{-}P(A1|C3)$$

dove A1 = l'auto si trova dietro alla porta 1;
C3 = il conduttore seleziona una capra dietro la porta 3.

La probabilità a priori che l'automobile si trovi dietro la porta 1, P(A1), = 1/3
La probabilità che il conduttore trovi una capra dietro la porta 3, P(C3) , è = 1, (il conduttore sa in anticipo dove è l'automobile)
La probabilità che il conduttore selezioni una porta con dietro la capra posto ("a posteriori") che l'automobile sia dietro la porta 1, P(C3|A1) = è 1.
Pertanto, sfruttando il teorema di Bayes la probabilità di trovare l'auto cambiando la scelta iniziale, dopo che il conduttore (onnisciente) ha mostrato una porta con dietro la capra è:

$$1\text{-}P(A1\text{-}C3) = 1 - \frac{P(C3|A1)P(A1)}{P(C3)} = \frac{1\text{-}1*1/3}{1} = \frac{2}{3}$$

http://montyhallproblem.com/
http://www.youtube.com/watch?v=mhlc7peGlGg

DIAGRAMMA DI FLUSSO

- Formulazione quesito 
- Definizione popolazione campionaria
- Esperimento e raccolta dei dati
- Statistica descrittiva
- Tests di verifica dell'ipotesi
- Risposta al quesito

"Lo statistico e' un uomo che fa un calcolo giusto partendo da premesse dubbie per arrivare ad un risultato sbagliato."

- Jean Delacour -

POPOLAZIONE CAMPIONARIA

Il campionamento può avvenire da una popolazione **infinita** o da una **finita.**

Campionamento da popolazioni **finite**:

- **INDAGINE CAMPIONARIA**: indagine su una parte della popolazione, che sarà definito **CAMPIONE**:

- Il numero di unità che compongono il campione è detto **DIMENSIONE CAMPIONARIA**.

- Il rapporto tra la dimensione campionaria n e quella della popolazione N si chiama **FRAZIONE DI CAMPIONAMENTO.**

Elemento cruciale nella definizione del campione è dato dalla regola di selezione, ossia dalla procedura con la quale le unità campionarie sono estratte dalla popolazione.

Campionamento casuale semplice.

I campioni di uguale dimensione hanno la stessa probabilità di essere estratti.

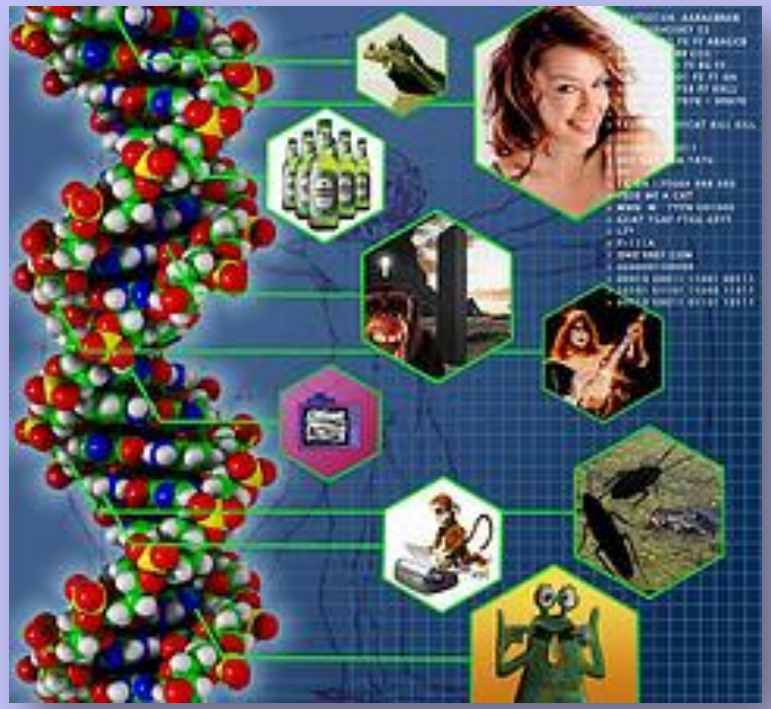
Solitamente viene condotto senza ripetizione. Se N è la dimensione del campione e n quella del campione, la probabilità di inclusione è pari alla frazione di campionamento n/N.

Campionamento stratificato.

Discende dal c.c.s. e si adotta quando si possiedono informazioni aggiuntive sulla popolazione. In base a tali informazioni, **la popolazione viene suddivisa in sottopopolazioni o strati, ognuno dei quali contiene unità tra loro omogenee secondo qualche criterio**. Da ogni strato vengono estratte, tramite c.c.s. le unità da inserire nel campione

Campionamento casuale a grappoli e a stadi.

Per adottare questa tecnica deve esistere un modo di suddividere la popolazione in sottoinsiemi di unità (i grappoli). **Si selezionano quindi, con un'estrazione casuale senza ripetizione, un certo numero di grappoli e si prendono come unità campionarie tutte le unità appartenenti ai grappoli estratti**. Una variante è il campionamento a due stadi, che si differenzia da quello a grappoli poiché nella fase finale di rilevazione viene applicata un'estrazione casuale, creando così un secondo stadio di campionamento.



INTERVALLO DI CONFIDENZA

L’intervallo di valori plausibili in cui cui ricade la stima di un parametro si definisce intervallo di confidenza (o intervallo di fiducia).

I valori estremi dell'intervallo di confidenza si chiamano *limiti di confidenza*.

INTERVALLO DI CONFIDENZA

In genere si considera attendibile un intervallo di confidenza 95%

Possiamo essere in uno dei due seguenti casi:

-Conosciamo la varianza (σ^2) della popolazione (ad esempio da indagini precedenti)

- Non conosciamo la varianza della popolazione (caso più frequente)

Esempio

Caso 1:

con excel: =confidenza(α; σ; dimensioni campione)

alfa =	**0,05**	
varianza=	**34**	
deviazione standard=	**5,830952**	
numerosità campione=	**24**	
intervallo di confidenza=	**2,332824**	
media=	**12**	
media+/- intervallo di confidenza=	**9,667176**	**14,33282**

Alfa è il livello di significatività utilizzato per calcolare il livello di confidenza. Il livello di confidenza è uguale a 100*(1 - alfa)%. In altre parole, un valore alfa di 0,05 indica un livello di confidenza del 95%.

Caso 2

Esamino 240 topini e trovo che il 35 hanno una mutazione xyz di mio interesse. Quale sarà la stima della percentuale di topi affetti, con un IC del 95%?

$$\text{Frequenza} +/- 1.96 * \sqrt{\frac{\text{affetti} * (1 - \text{affetti})}{N}}$$

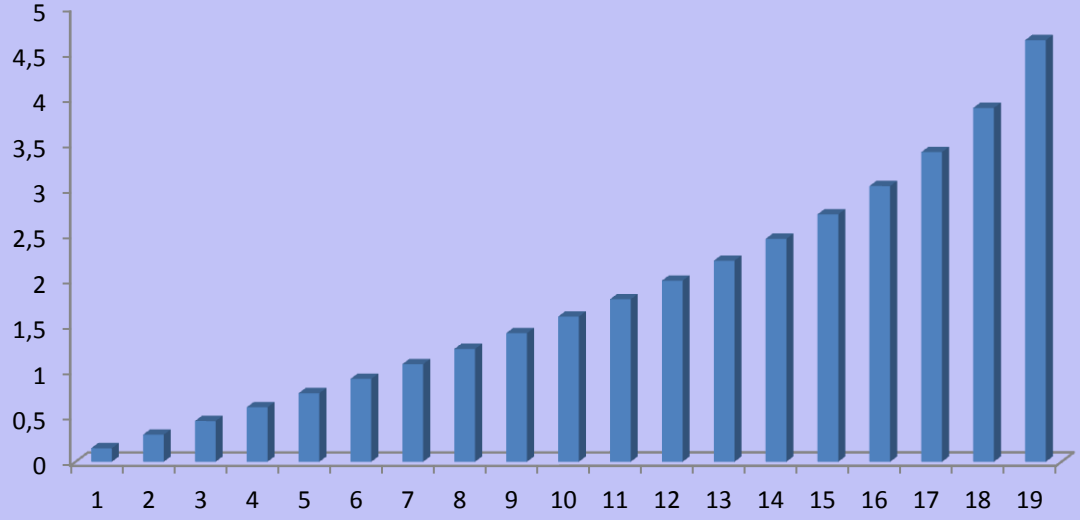
% affetti = 14,58
% non effetti = 85,41
N = 240

% compresa tra **30,53** – **39,46**

Alcune considerazioni…

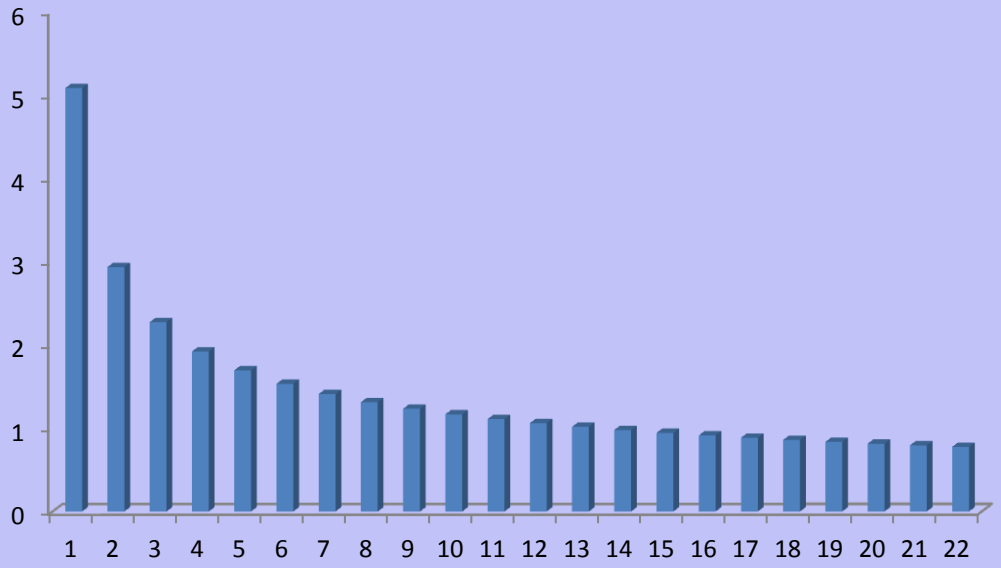
Come varia IC al variare di alfa?

alfa	dev st	numerosità	intervallo
1	5,8	24	
0,95	5,8	24	0,14848
0,9	5,8	24	0,297546
0,85	5,8	24	0,447802
0,8	5,8	24	0,599885
0,75	5,8	24	0,754487
0,7	5,8	24	0,912377
0,65	5,8	24	1,074436
0,6	5,8	24	1,241697
0,55	5,8	24	1,4154
0,5	5,8	24	1,597084
0,45	5,8	24	1,788702
0,4	5,8	24	1,992824
0,35	5,8	24	2,212958
0,3	5,8	24	2,454109
0,25	5,8	24	2,723843
0,2	5,8	24	3,034509
0,15	5,8	24	3,40858
0,1	5,8	24	3,89475
0,05	5,8	24	4,640881
-3,2E-16	5,8	24	



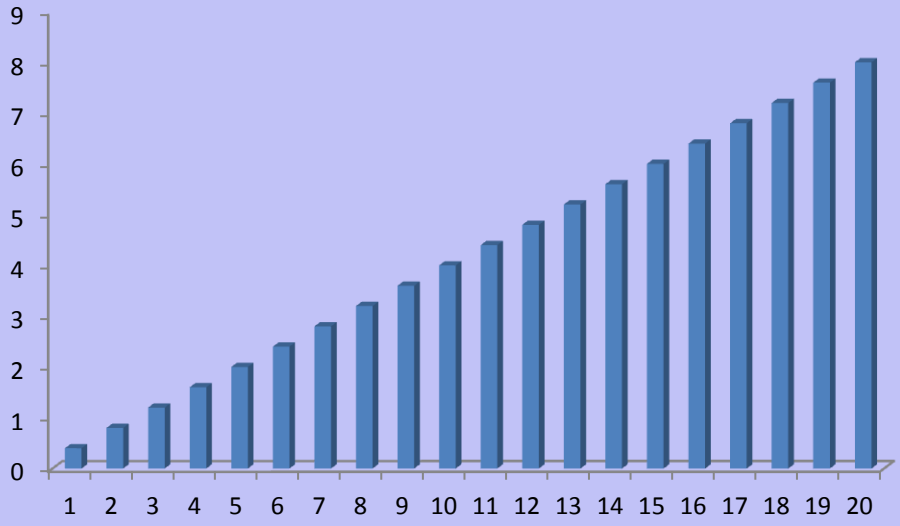
Come varia IC al variare della numerosità del campione?

alfa	dev st	numerosità	intervallo confidenza
0,05	5,8	5	5,083831
0,05	5,8	15	2,935151
0,05	5,8	25	2,273558
0,05	5,8	35	1,921507
0,05	5,8	45	1,69461
0,05	5,8	55	1,532833
0,05	5,8	65	1,410001
0,05	5,8	75	1,312639
0,05	5,8	85	1,23301
0,05	5,8	95	1,166311
0,05	5,8	105	1,109383
0,05	5,8	115	1,060052
0,05	5,8	125	1,016766
0,05	5,8	135	0,978384
0,05	5,8	145	0,944044
0,05	5,8	155	0,913083
0,05	5,8	165	0,884981
0,05	5,8	175	0,859324
0,05	5,8	185	0,835777
0,05	5,8	195	0,814064
0,05	5,8	205	0,793961
0,05	5,8	215	0,775277



Come varia IC al variare della varianza della popolazione?

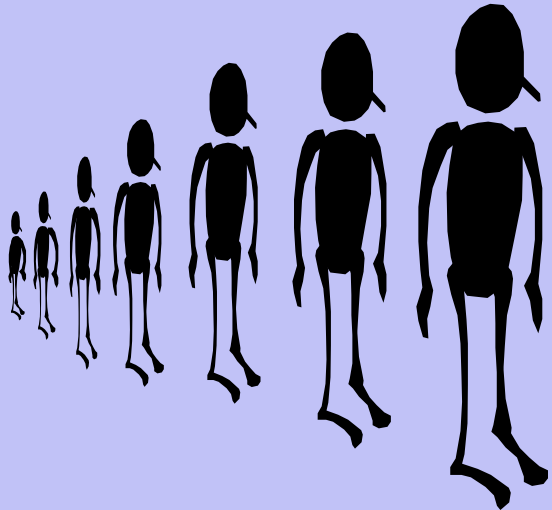
alfa	dev st	numerosità	intervallo confidenza
0,05	1	24	0,400076
0,05	2	24	0,800152
0,05	3	24	1,200228
0,05	4	24	1,600304
0,05	5	24	2,00038
0,05	6	24	2,400456
0,05	7	24	2,800532
0,05	8	24	3,200608
0,05	9	24	3,600684
0,05	10	24	4,00076
0,05	11	24	4,400836
0,05	12	24	4,800912
0,05	13	24	5,200988
0,05	14	24	5,601064
0,05	15	24	6,00114
0,05	16	24	6,401216
0,05	17	24	6,801292
0,05	18	24	7,201368
0,05	19	24	7,601443
0,05	20	24	8,001519



DIMENSIONE CAMPIONARIA

1) La bontà dei risultati ottenibili da un campione non dipende tanto dal numero degli individui che compongono il campione quanto dal ***modo*** con cui essi sono stati selezionati.

2) La dimensione del campione **non è necessariamente proporzionata** alla dimensione della popolazione in studio.



CALCOLO DIMENSIONE CAMPIONARIA

Esempio

Indagine sul consumo proteico giornaliero (grammi) dei cani:

quanti selezionarne?

Definire:

- ampiezza dell'intervallo di confidenza : L
- il livello di fiducia α
- la varianza della popolazione σ

Il veterinario vuole stimare il valore medio entro ± 5 grammi, cioè entro un intervallo di 10 grammi, vuole una fiducia di 1-α = 95%, e da precedenti indagini indica una σ di 20 grammi:

$$n = \frac{Z^2\ \alpha\ 2\ \sigma^2}{L^2} = \frac{1.96^2\ 20^2}{5^2} = 61.47 \qquad Z^2\ \alpha = 1.96$$

E' opportuno usare un campione di 62 cani

IN EPIDEMIOLOGIA

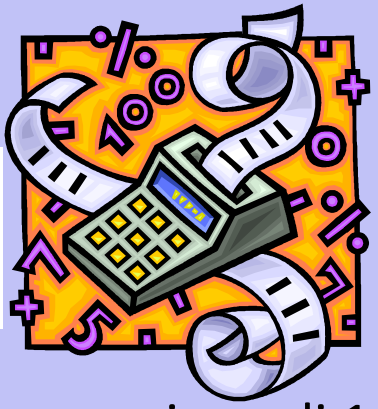
Si usa il concetto di prevalenza:

$$n = \frac{1.96^2 \, P_{att} \, (1 - P_{att})}{D^2}$$

ESEMPIO

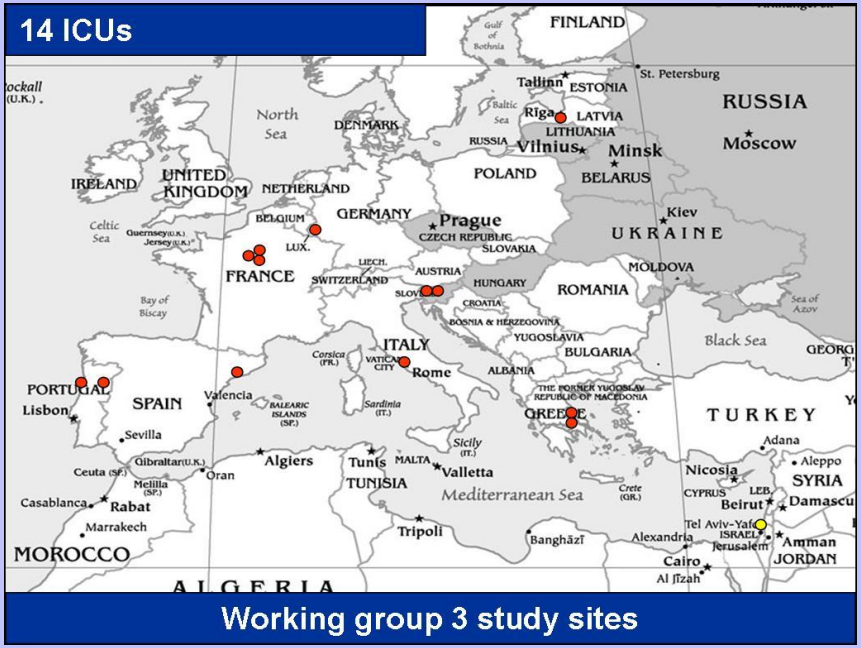
- Si sospetta che la prevalenza di una malattia in una popolazione sia pari a 0.3. Si vuole studiare un campione per stimare la prevalenza della malattia nella popolazione con precisione 0.07 (ossia 7%), ovvero ci si aspetta che i limiti dell'intervallo di confidenza della stima siano compresi fra 0.23 e 0.37. Si vuole calcolare la dimensione del campione necessaria:

$$n = \frac{1.96^2 \; 0.3(1-0.3)}{0.07^2} = 165$$



Per ottenere il tuo scopo, dovrà essere esaminato un campione di 165 animali.

In concreto …



Analisi statistica: manipolazioni misteriose talora bizzarre dei dati di un esperimento per nascondere il fatto che i risultati non hanno significati generalizzabili per l'umanita'. In genere si usano i computer: cio' conferisce un'aurea addizionale di mistero.

http://wpage.unina.it/dambra/ridere/Statistici.htm

DIAGRAMMA DI FLUSSO

- Formulazione quesito 
- Definizione popolazione campionaria 
- Esperimento e raccolta dei dati 
- Statistica descrittiva
- Statistica inferenziale
- Risposta al quesito

STATISTICA DESCRITTIVA

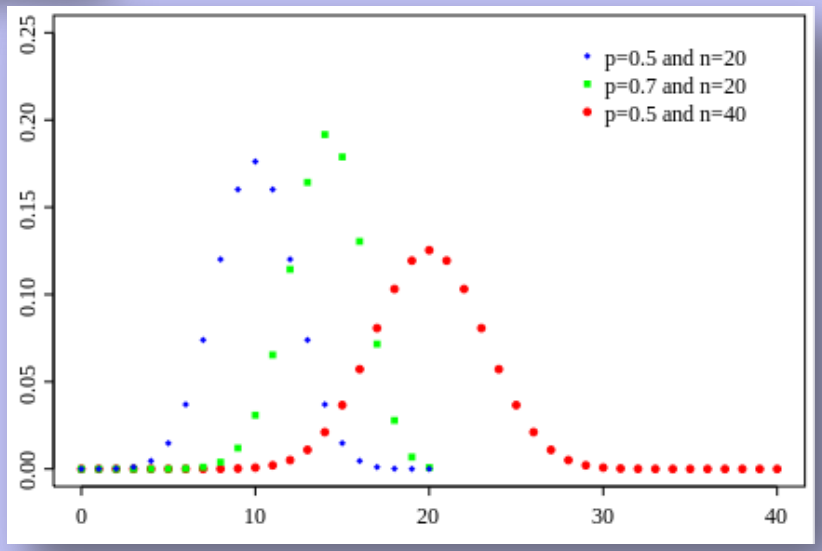
- Tipo di distribuzione
- Misura della tendenza centrale
- Misura della variabilità

intervalli di riferimento

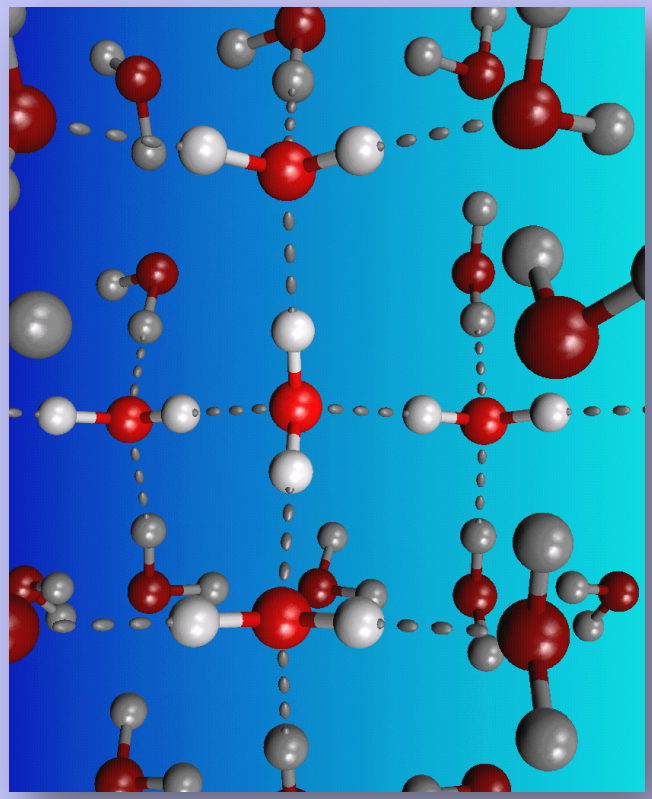
DISTRIBUZIONE BINOMIALE

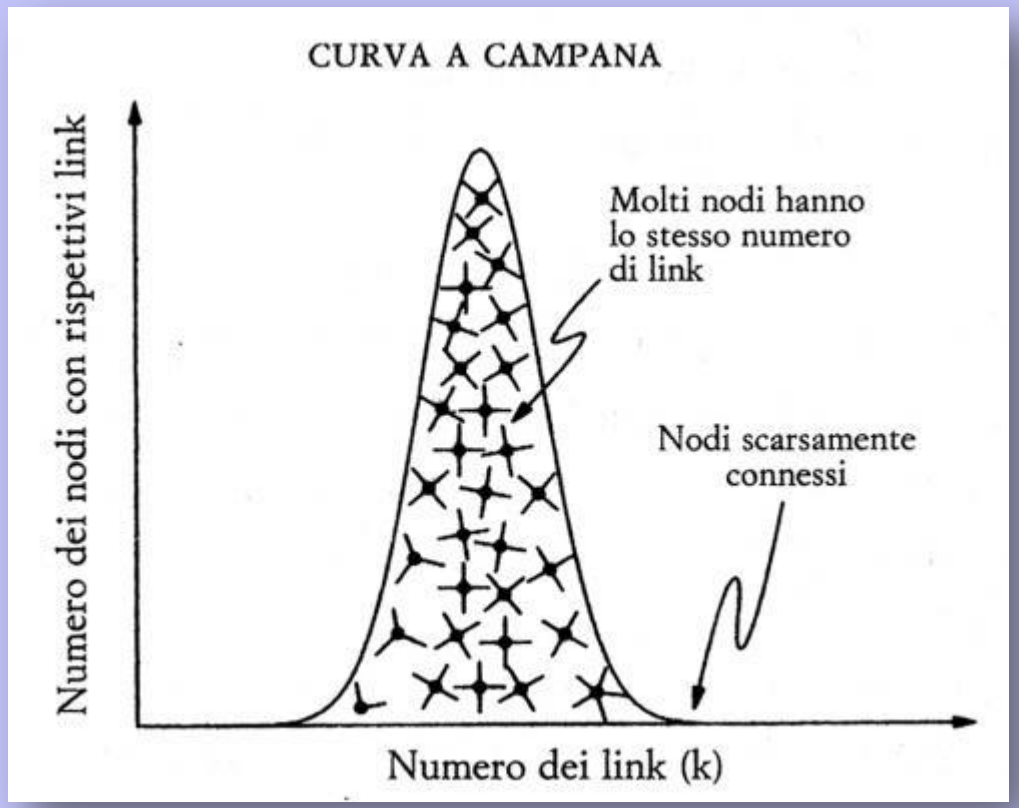
Si dice esperimento di Bernoulli una sequenza di ***n*** prove con le seguenti caratteristiche:

1) il risultato di ogni prova può essere solo "successo" o "fallimento";
2) il risultato di ciascuna prova è indipendente dai risultati delle prove precedenti;
3) la probabilità p di "successo", e quindi la probabilità q = 1 – p di "fallimento", sono costanti in ciascuna prova.

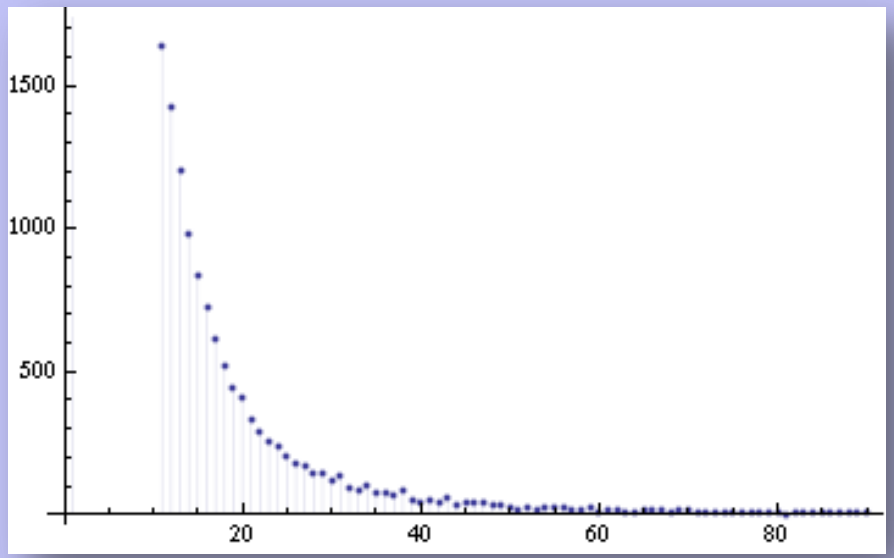


DISTRIBUZIONE DI POISSON





DISTRIBUZIONE DI PARETO

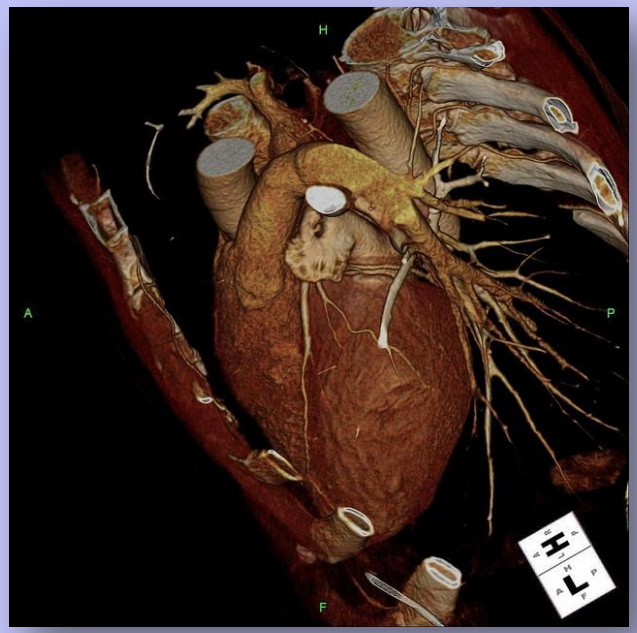


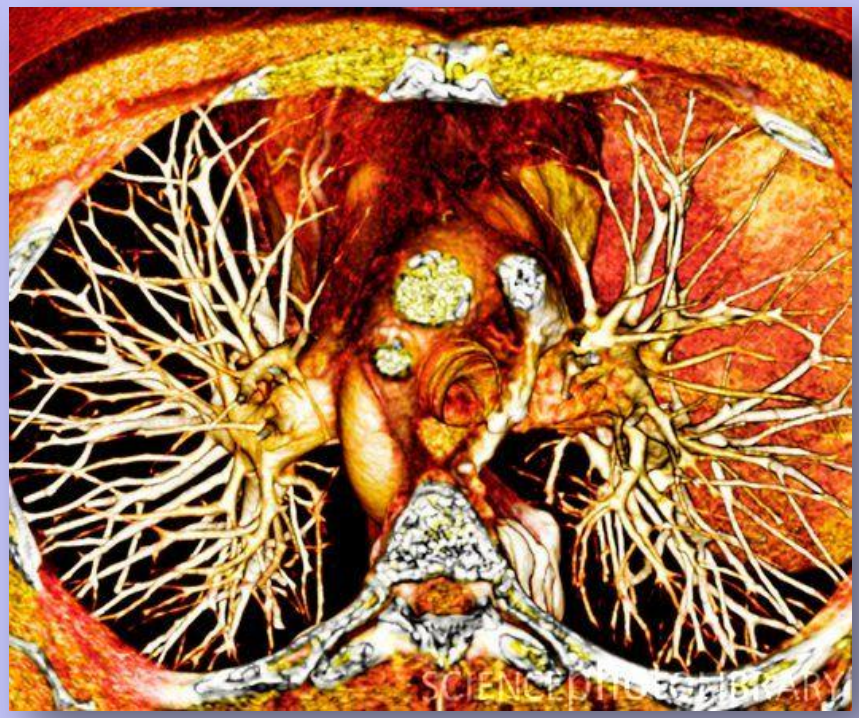
Una **distribuzione a legge di potenza**, o **distribuzione power-law**, o **distribuzione di Pareto**, nella sua forma più generale ha la forma:

$$y = a\, x^{-b}$$

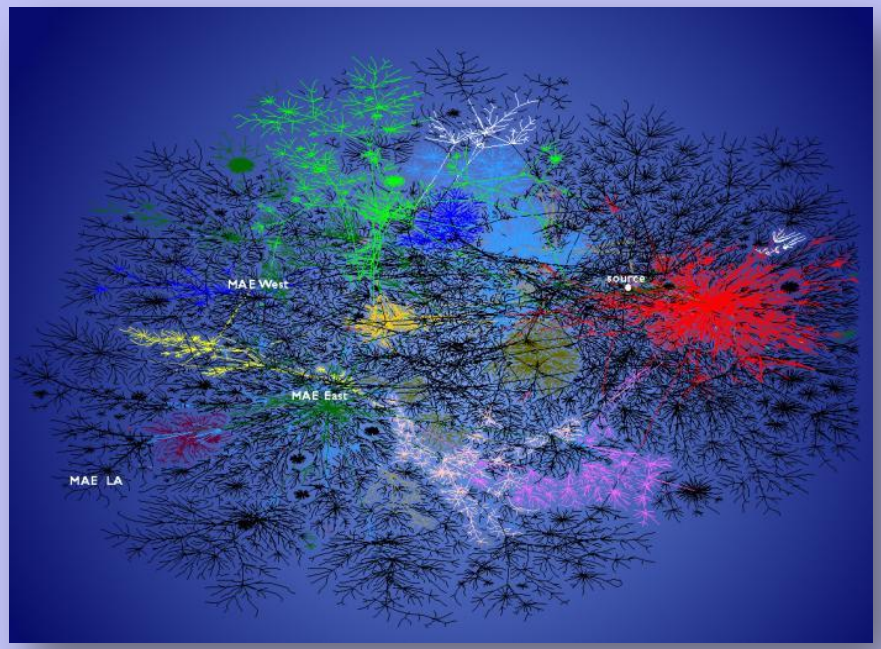
$b > 1$

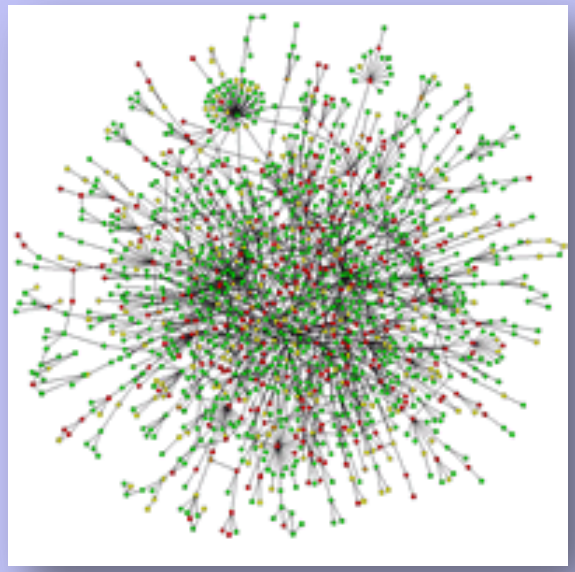
DISTRIBUZIONE DI PARETO

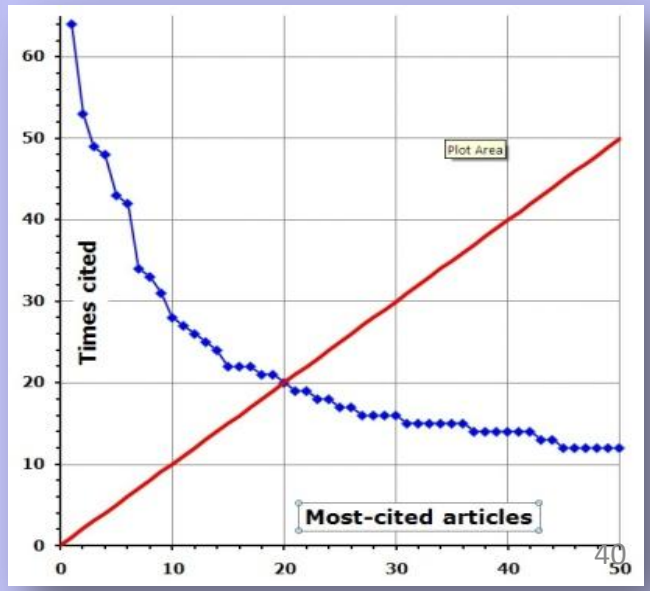




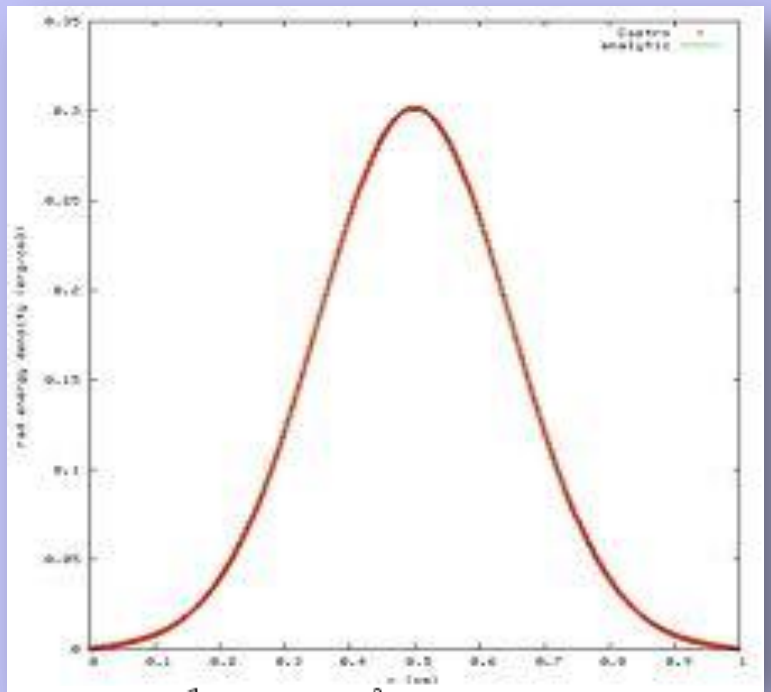
INVARIANZA DI SCALA





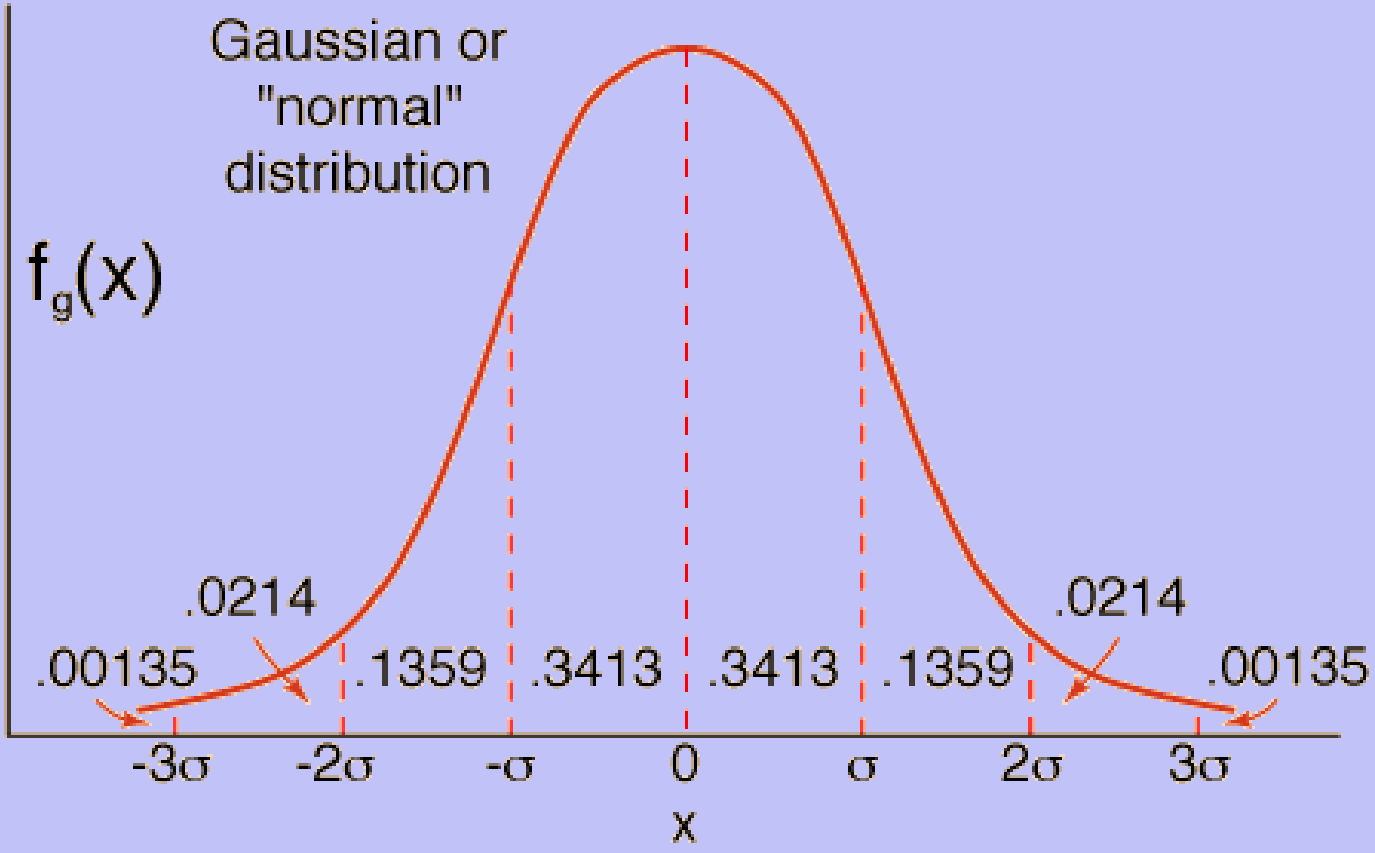


DISTRIBUZIONE GAUSSIANA



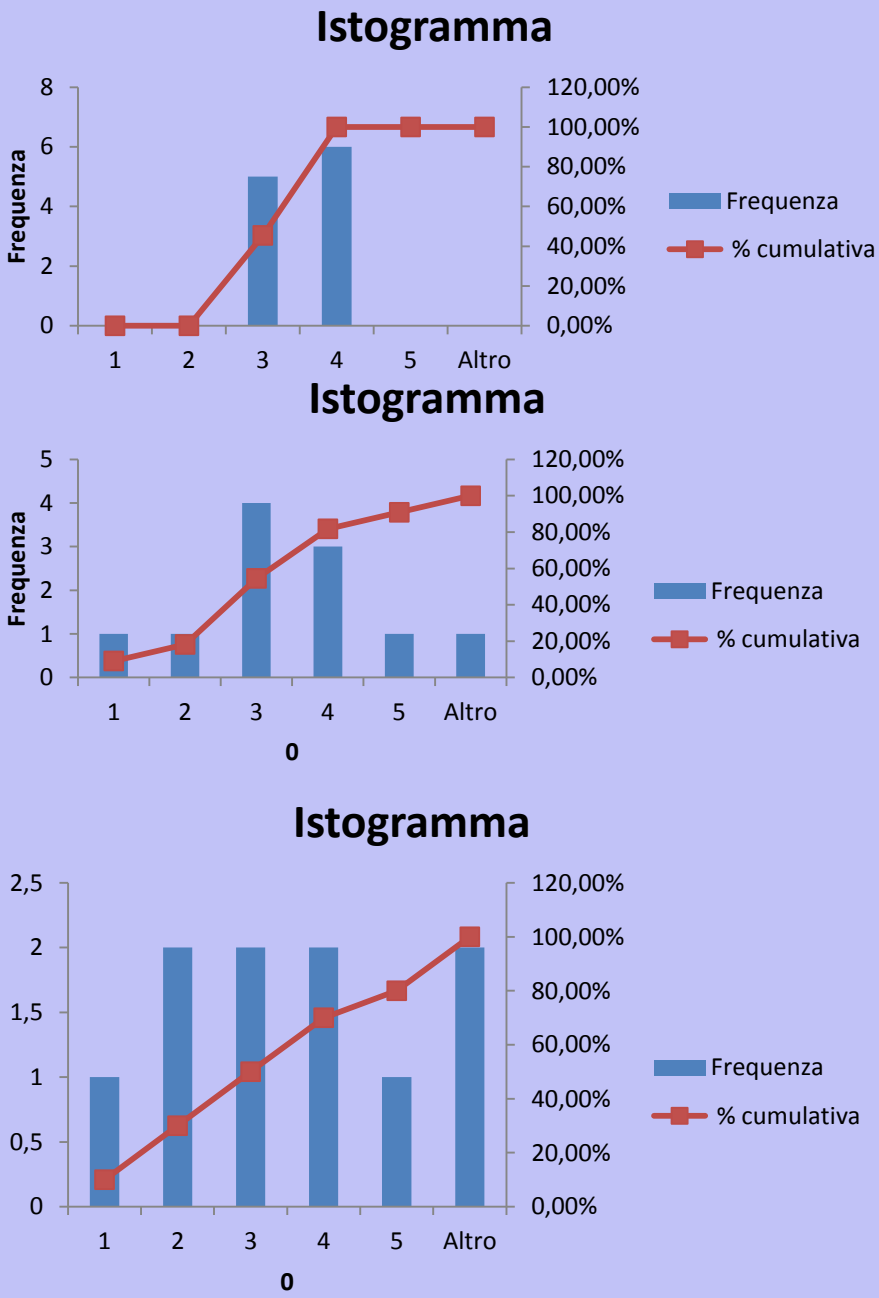
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{con} \quad -\infty < x < \infty$$

FORMA DELLA DENSITA` DI PROBABILITA` GAUSSIANA



	PESO GATTI	
A	B	C
3	2	0,5
4	3	1
3,3	4	2
4	3	3
3	3	4
3	4	5
3	5	6
3	6	7
3,5	4	2
3,3	3	3
3,4	2	4
	1	

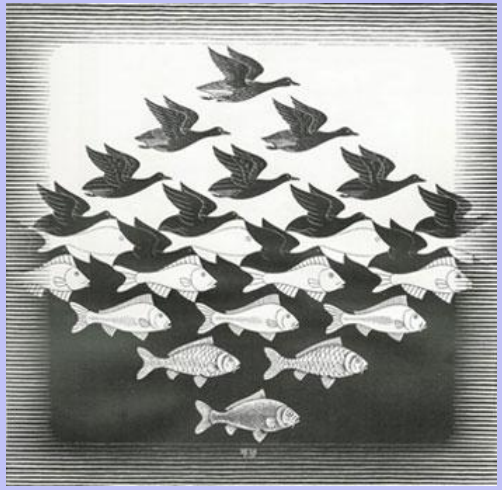
media	3,318182	3,333333	3,409091
varianza	0,147636	1,878788	4,140909
dev. standard	0,384235	1,370689	2,034922
err. standard	0,03493	0,114224	0,184993



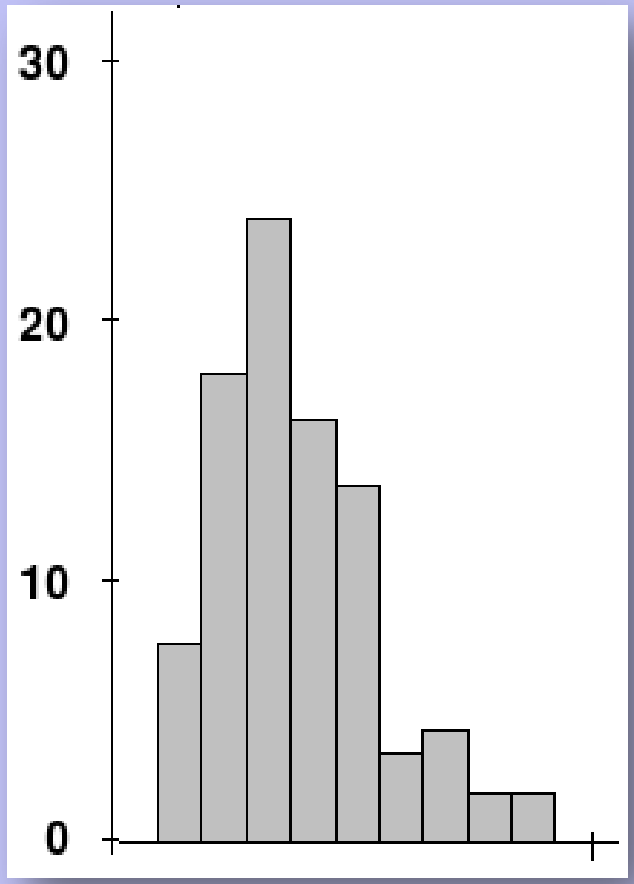
ASIMMETRIA

una distribuzione di probabilità è **simmetrica** quando la sua funzione di probabilità P (nel caso discreto) o la sua funzione di densità di probabilità (nel caso continuo) siano simmetriche rispetto ad un valore fissato x_0:

$$P(x_0 + x) = P(x_0 - x) \text{ oppure } f(x_0 + x) = f(x_0 - x)$$



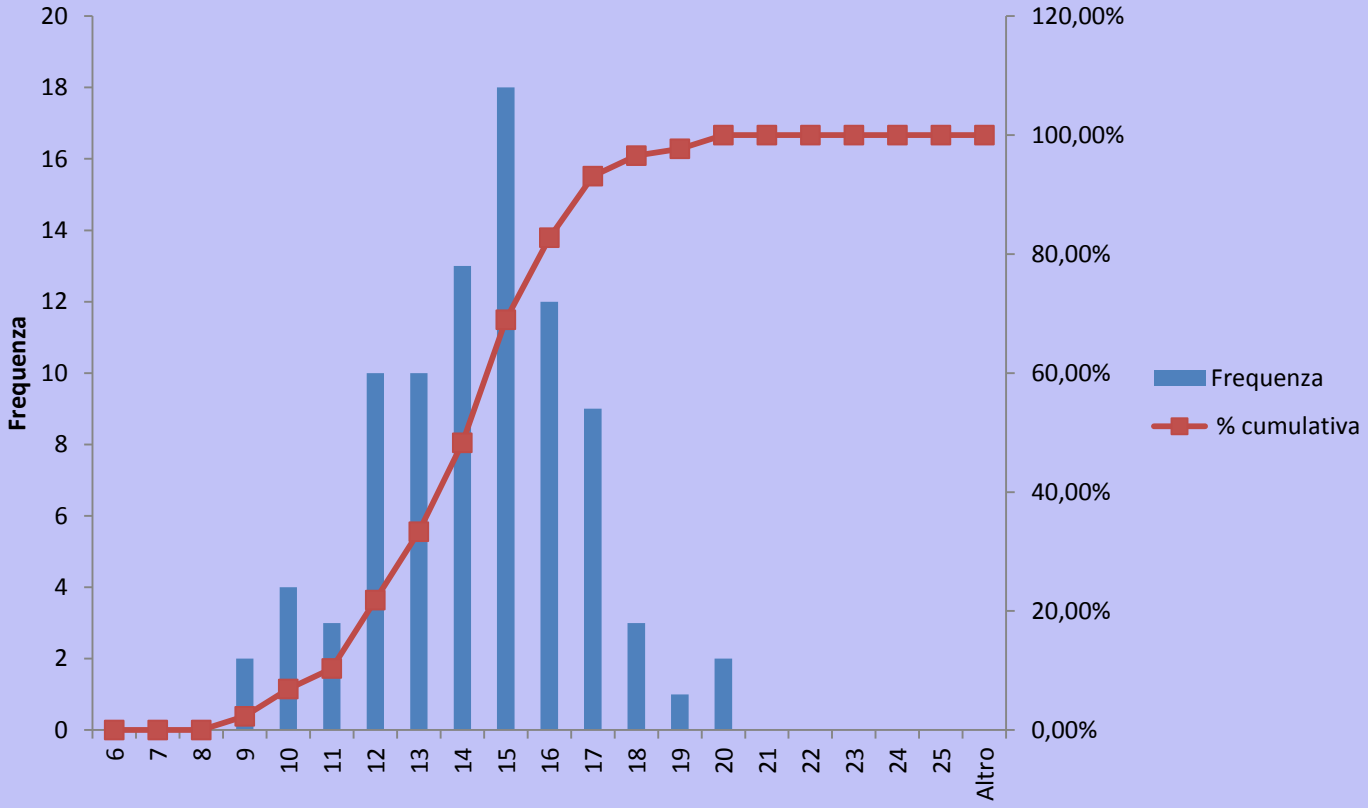
ASIMMETRIA



ASIMMETRIA

- Esistono diversi indici di asimmetria. Per ognuno di essi il valore 0 fornisce una condizione necessaria, ma **non** sufficiente, affinché una distribuzione sia simmetrica.
- Gli indici di asimmetria comunemente utilizzati si basano su alcune proprietà delle distribuzioni simmetriche o, in particolare, della distribuzione normale.
- il valore atteso, la mediana e la moda coincidono.

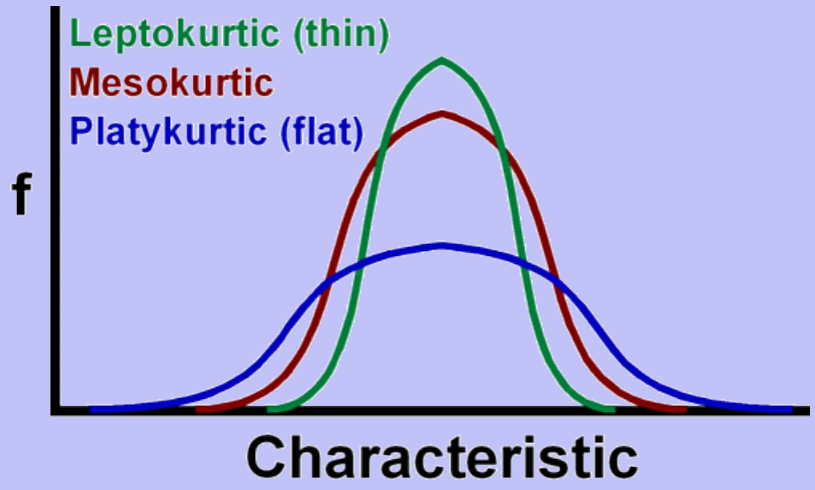
ASIMMETRIA



asimmetria = -0,08511

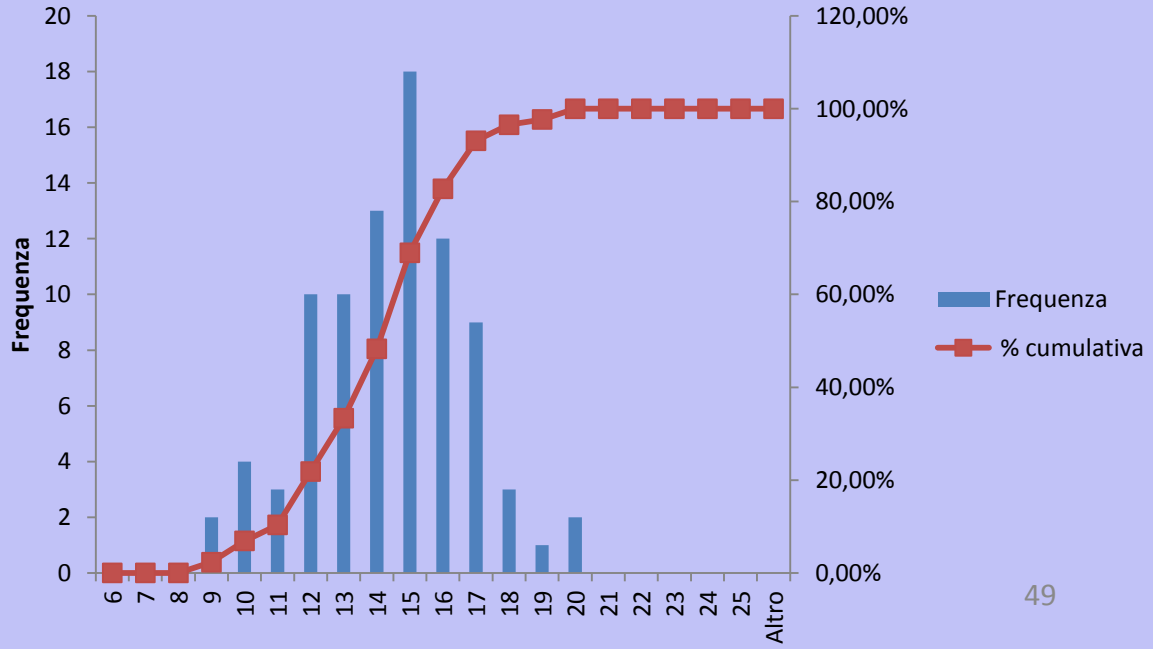
CURTOSI

- La **curtosi** (o **kurtosi**) è un **allontanamento dalla normalità distributiva**, rispetto alla quale si verifica un maggiore *appiattimento* distribuzione platicurtica) o un maggiore *allungamento* (distribuzione leptocurtica).

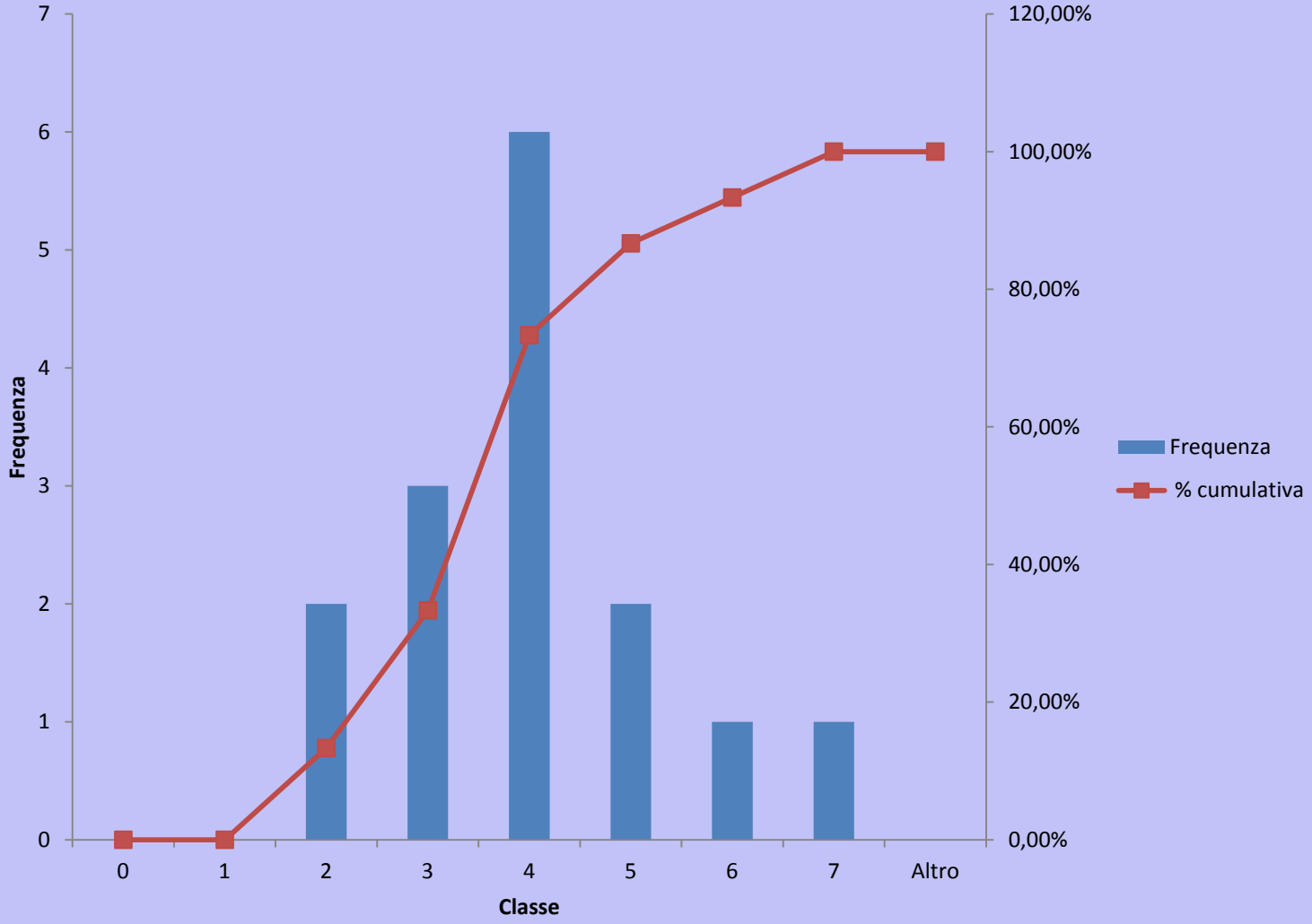


CURTOSI

- Un valore minore di 0 indica una distribuzione platicurtica, mentre un valore maggiore di 0 indica una distribuzione leptocurtica



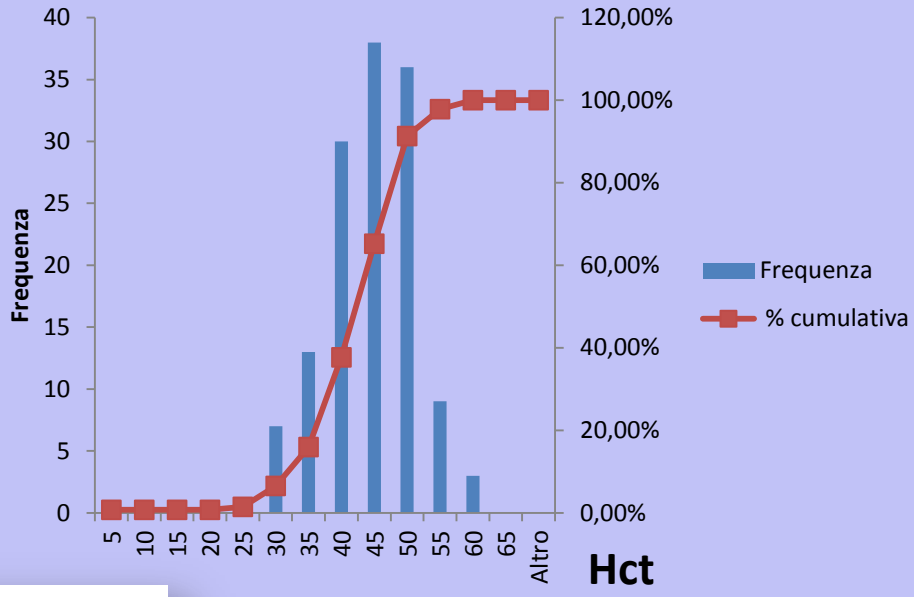
Curtosi = -0,26316

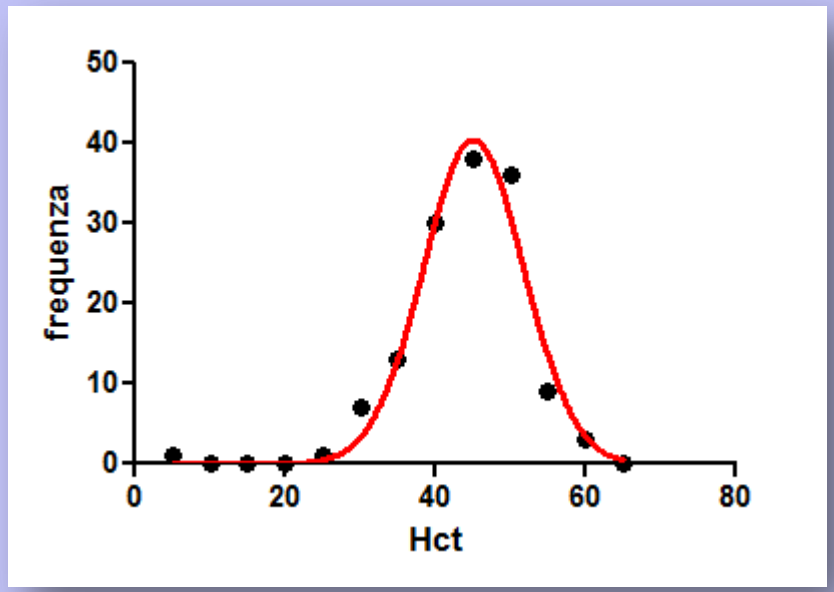


asimmetria	**0,800298**
curtosi	**0,753757**

Valori Hct di 139 cani

	Hct
Media	41,37116
Errore standard	0,626669
Mediana	42,15
Moda	45,4
Deviazione standard	7,361699
Varianza campionaria	54,19461
Curtosi	3,452818
Asimmetria	-0,96729
Intervallo	52,58
Minimo	4,82
Massimo	57,4
Somma	5709,22
Conteggio	138
Più grande(1)	57,4
Più piccolo(1)	4,82
Livello di confidenza(95,0%)	1,239196





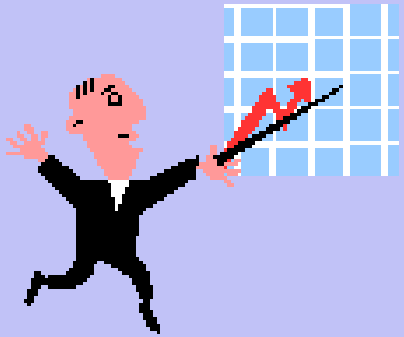
Goodness of Fit:

R^2 = 0,9724

MISURA DELLA TENDENZA CENTRALE

MEDIA ARITMETICA ottenuta sommando un insieme di numeri e quindi dividendo per il conteggio di questi numeri.

MEDIANA ovvero il numero che occupa la posizione centrale di un insieme di numeri. Una metà dei numeri ha un valore superiore rispetto alla mediana, mentre l'altra metà ha un valore inferiore



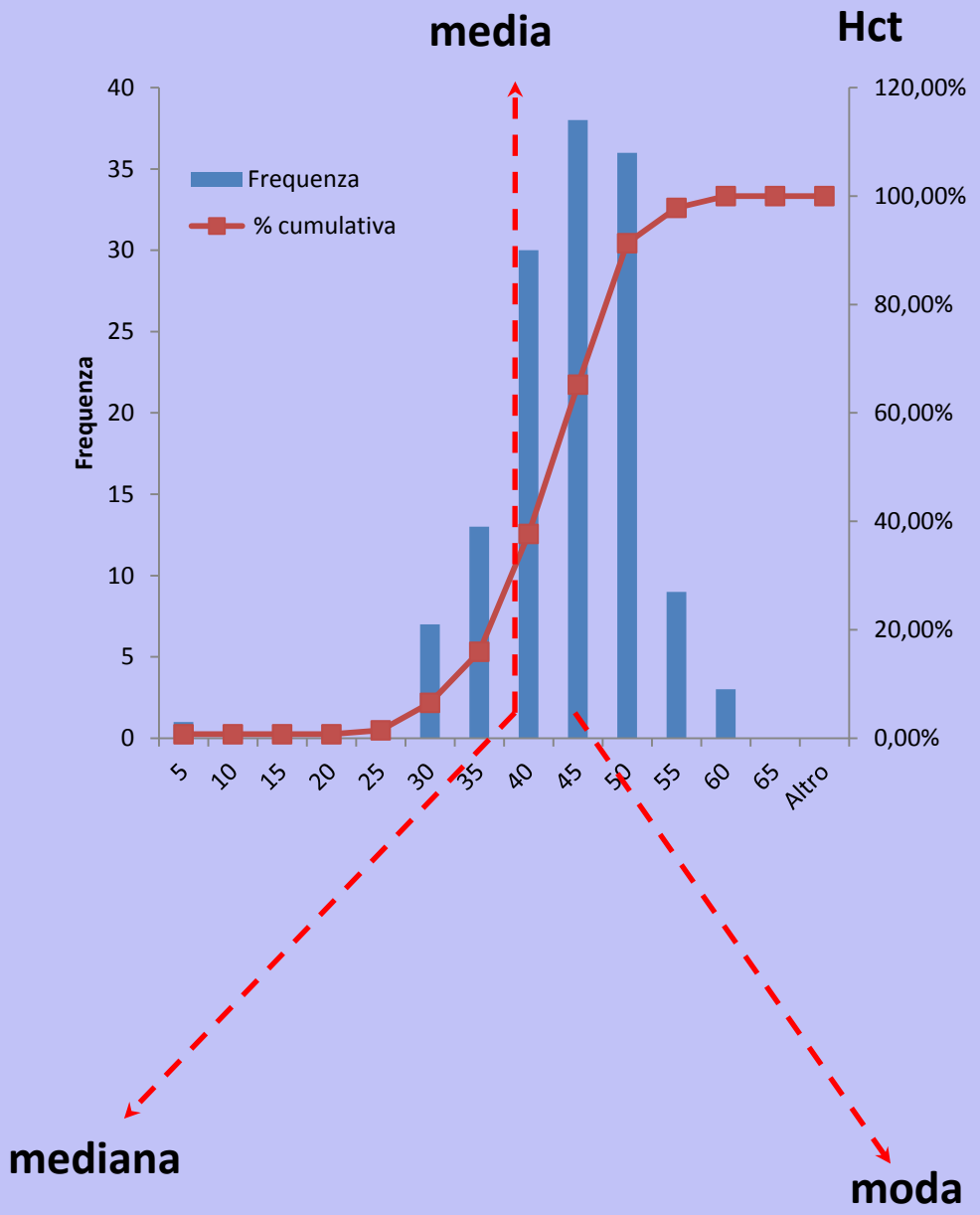
MODA ovvero il numero più ricorrente in un insieme di numeri.

"Non mi fido molto delle statistiche, perchè un uomo con la testa nel forno acceso e i piedi nel congelatore statisticamente ha una temperatura media."

- Charles Bukowski -

ESEMPIO I

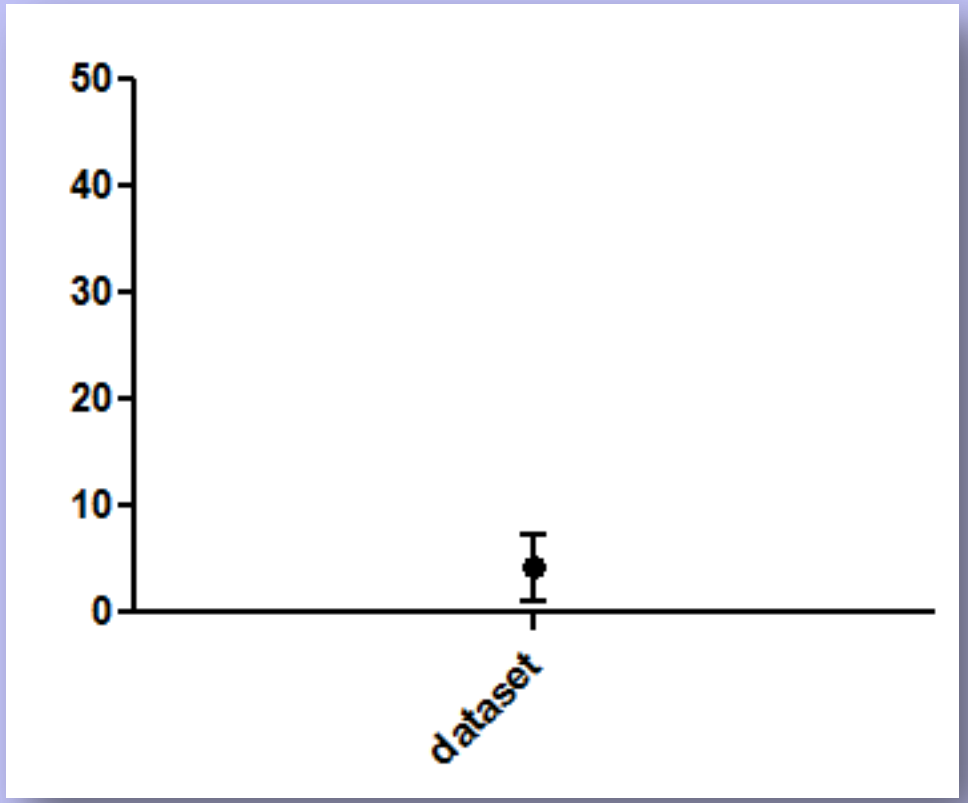
Hct	
Media	41,37116
Errore standard	0,626669
Mediana	42,15
Moda	45,4
Deviazione standard	7,361699
Varianza campionaria	54,19461
Curtosi	3,452818
Asimmetria	-0,96729
Intervallo	52,58
Minimo	4,82
Massimo	57,4
Somma	5709,22
Conteggio	138
Più grande(1)	57,4
Più piccolo(1)	4,82
Livello di confidenza(95,0%)	1,239196



ESEMPIO II

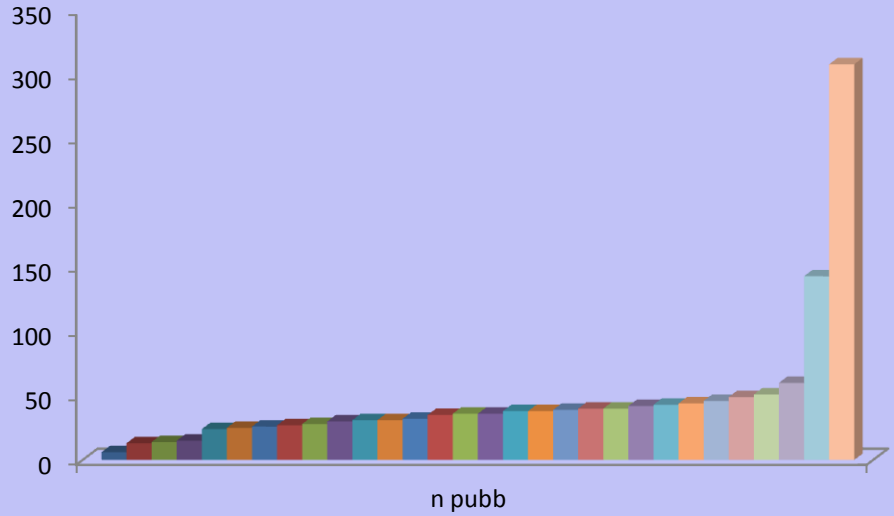
dataset
1
1
1
1
1
1
1
1
1
1
1
2
1
45

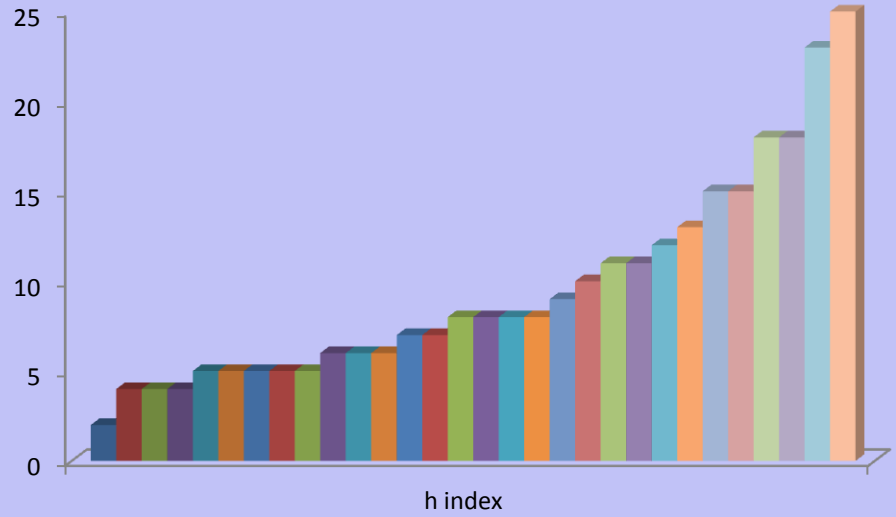
4,214286	media
11,74196	ds
1	mediana
1	10°percentile
1,7	90°percentile

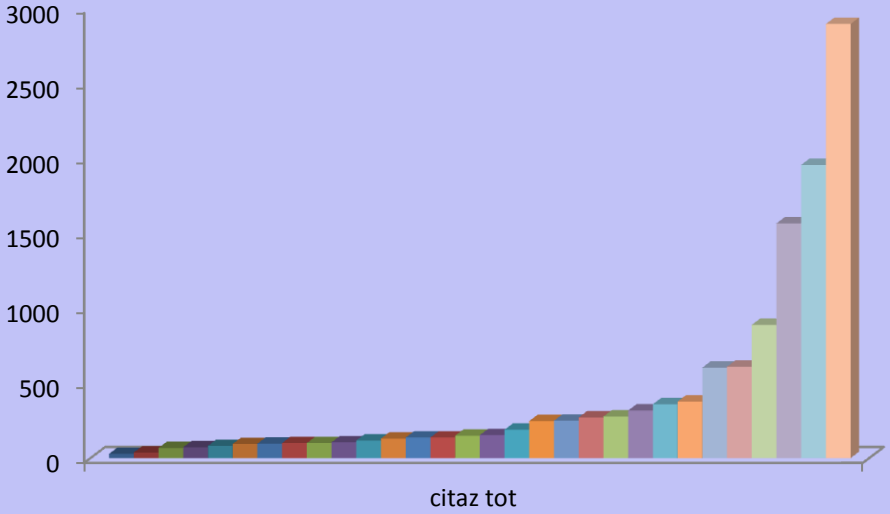


Un esempio reale…

Autore	n pubb	h index	citaz tot
1	36	5	117
2	36	6	96
3	49	12	317
4	25	5	130
5	27	10	247
6	51	18	1565
7	60	18	888
8	143	23	1955
9	42	11	271
10	30	8	153
11	31	7	149
12	46	8	189
13	308	25	2895
14	39	5	100
15	6	2	67
16	24	4	28
17	14	6	72
18	13	4	37
19	40	8	249
20	15	4	277
21	32	11	377
22	43	15	609
23	38	8	137
24	28	5	81
25	44	13	603
26	31	9	105
27	26	6	94
28	40	15	359
29	38	7	137
30	35	5	100







	n pubb	h index	citaz tot
mediana	36,0	8,0	151,0
media	46,3	9,4	413,5

MISURA DELLA VARIABILITA'

VARIANZA $\dfrac{\Sigma (x - \bar{x})^2}{(n - 1)} = \sigma^2$

DEVIAZIONE STANDARD $\sqrt{\dfrac{\Sigma (x - \bar{x})^2}{(n - 1)}} = \sigma$

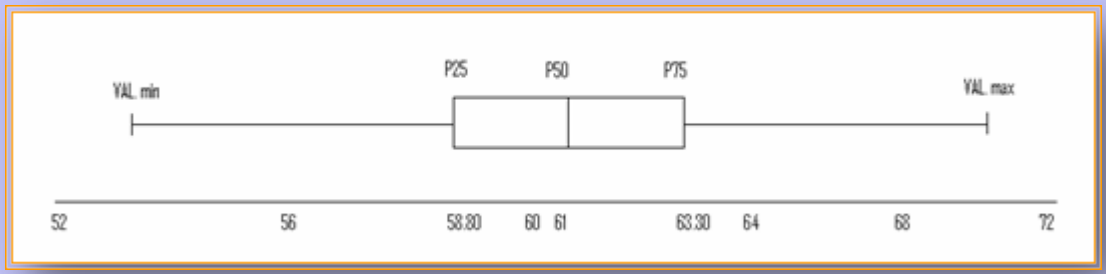
ERRORE STANDARD $\sigma / \sqrt{n}$

PERCENTILE

I **percentili** dividono la distribuzione in cento parti uguali. Dato un campione il percentile ennesimo è il valore che separa il numero percentuale dei dati dal resto. Ad esempio il 50° percentile è la mediana.

La definizione di **percentile** permette di stabilire la percentuale di valori al di sotto di una certa soglia, e anche la percentuale tra due soglie;

Il grafico che permette di visualizzare le suddivisioni percentili (compresa la mediana) e valori estremi delle serie di dati è detto “**box plot**” o "Box and wiskers plot“.

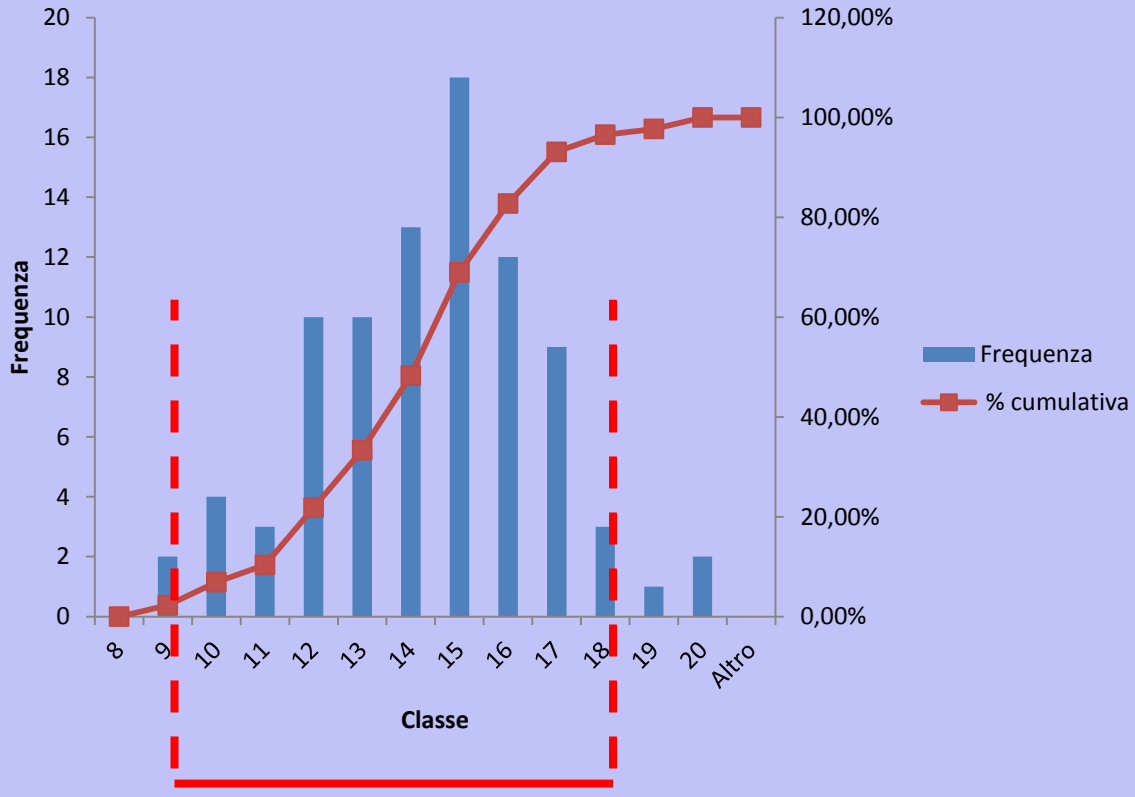


INTERVALLI DI RIFERIMENTO

Gli intervalli di riferimento rappresentano l'ambito di valori che include il 95% dei risultati osservati in un gruppo di controllo di soggetti sani.

INTERVALLI DI RIFERIMENTO I

Distribuzione normale (Hgb)



media	**14,01724**
ds	**2,298443**
min	**9,420356**
max	**18,61413**

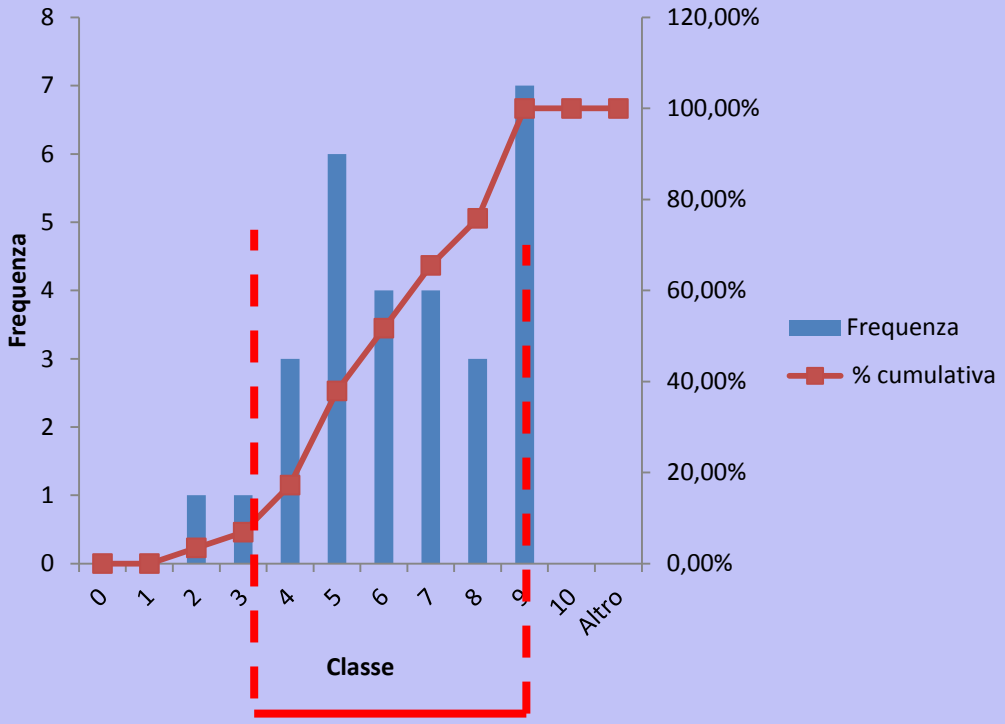
Intervallo di riferimento

INTERVALLI DI RIFERIMENTO II

Distribuzione
non normale

curtosi	-0,88313
asimmetria	-0,23046

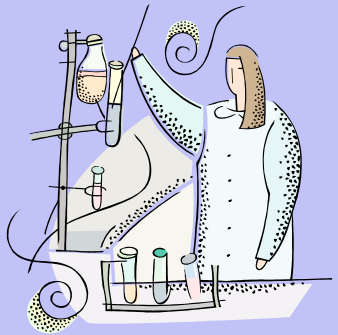
2,5° percentile	3,4
97,5° percentile	9



Intervallo di riferimento

INTERVALLI DI RIFERIMENTO

Un soggetto sano che esegue una analisi di laboratorio ha il 95% di probabilità che il proprio valore risulti all'interno degli intervalli di riferimento;



Un soggetto sano che esegue 20 analisi di laboratorio ha una probabilità di $(0,95)^{20}$ = 0,358486 che tutti i risultati rientrino all'interno degli intervalli di riferimento. Ovvero, su 100 soggetti sani che eseguono ciascuno 20 analisi, solamente il 36% circa avrà tutti i valori all'interno degli intervalli di riferimento.