Seminario 3

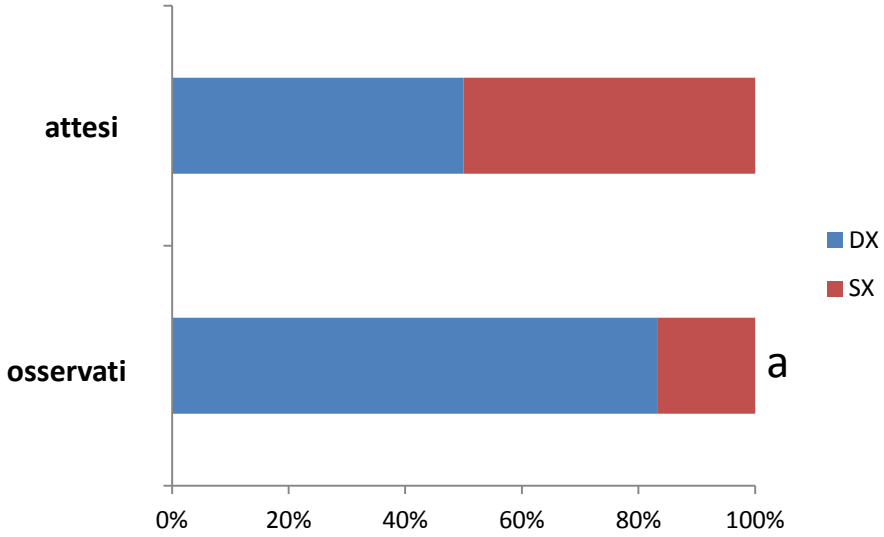
TEST χ^2

Esempio: lancio una moneta 1000 volte ed ottengo 450 volte testa e 550 volte croce, quando mi sarei aspettato una frequenza di 500 e 500
Ammettendo che l'evento testa e l'evento croce siano equiprobabili, cosa posso dire …

In termini più rigorosi devo usare un test di conferma dell'ipotesi, ovvero un test che mi consenta di accettare o scartare l'ipotesi che testa e croce siano effettivamente equiprobabili.

Table Analyzed	Data 1		
Chi-square			
Chi-square, df	5.013, 1		
P value	0,0252		
P value summary	*		
Statistically significant? (alpha<0.05)	Yes		
Data analyzed	testa	croce	Total
teorico	500	500	1000
empirico	450	550	1000
Total	950	1050	2000

Esempio 1 …



[a] = $p<0.001$, $n = 48$, χ^2 test

Table Analyzed	Data 1		
Chi-square			
Chi-square, df	12.00, 1		
P value	0,0005		
P value summary	***		
Statistically significant? (alpha<0.05)	Yes		
Data analyzed	DX	SX	Total
teorico	24	24	48
empirico	40	8	48
Total	64	32	96

Esempio 2 …

	ipotesi A	ipotesi B
giovani	23	26
adulti	24	31
anziani	45	39

Table Analyzed	Data 1
Chi-square	
Chi-square, df	1.419, 2
P value	0,4920
P value summary	ns
Statistically significant? (alpha<0.05)	No
Data analyzed	
Number of rows	2
Number of columns	3

Test esatto di Fisher

Il **test esatto di Fisher** è un test per la verifica d'ipotesi utilizzato nell'ambito della statistica non parametrica in situazioni con due variabili nominali e campioni piccoli. E' usato per verificare se i dati dicotomici di due campioni riassunti in una tabella di contingenza 2x2 siano compatibili con l'ipotesi nulla (H_0) che le popolazioni di origine dei due campioni abbiano la stessa suddivisione dicotomica e che le differenze osservate con i dati campionari siano dovute semplicemente al caso.

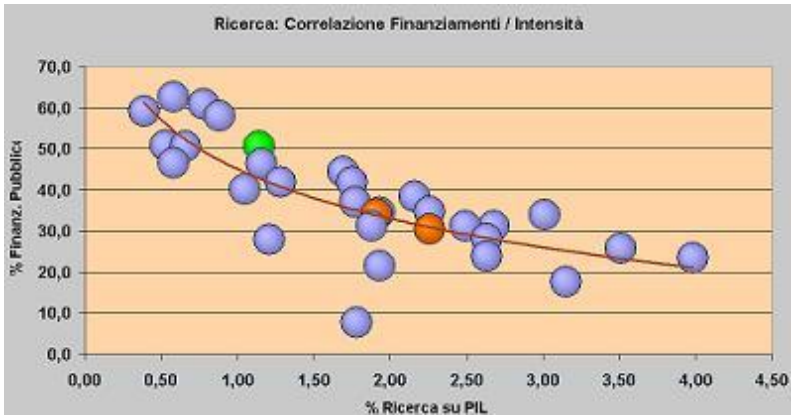
	morti	**vivi**
donne	**1**	**9**
uomini	**3**	**7**

Table Analyzed	Data 1		
Fisher's exact test			
P value	0,5820		
P value summary	ns		
Statistically significant? (alpha<0.05)	No		
Data analyzed	morti	vivi	Total
donne	1	9	10
uomini	3	7	10
Total	4	16	20

CORRELAZIONE

Per **correlazione** si intende una relazione tra due variabili casuali tale che a ciascun valore della prima variabile corrisponda con una certa regolarità un valore della seconda.

Non si tratta necessariamente di un rapporto di causa ed effetto ma semplicemente della tendenza di una variabile a variare in funzione di un'altra.



coefficiente di correlazione = r

$$r = \frac{\Sigma (x-\bar{x})\,(y-\bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \, \Sigma (y-\bar{y})^2}}$$

Esempio 1 …

peso	altezza
56	156
53	159
61	163
67	166
65	175
71	173
73	178
78	180
78	187
83	190
86	192

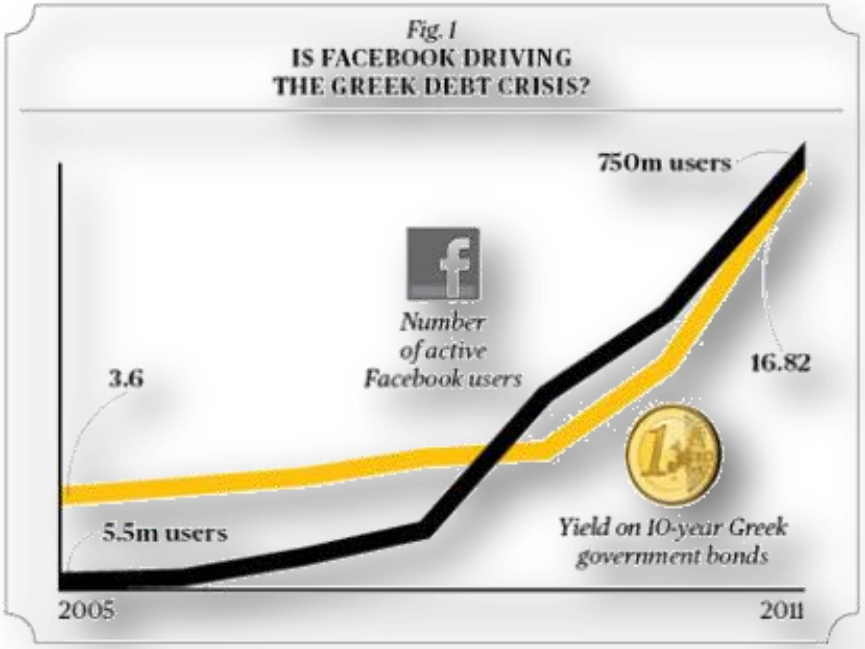
correlazione
0,960777

Esempio 2 …

prezzo moto	età
18.000	0,5
17.000	1
16.500	2
12.000	2,5
11.000	3
9.000	5
6.000	6
6.000	7
3.500	8
1.000	12

-0,95892

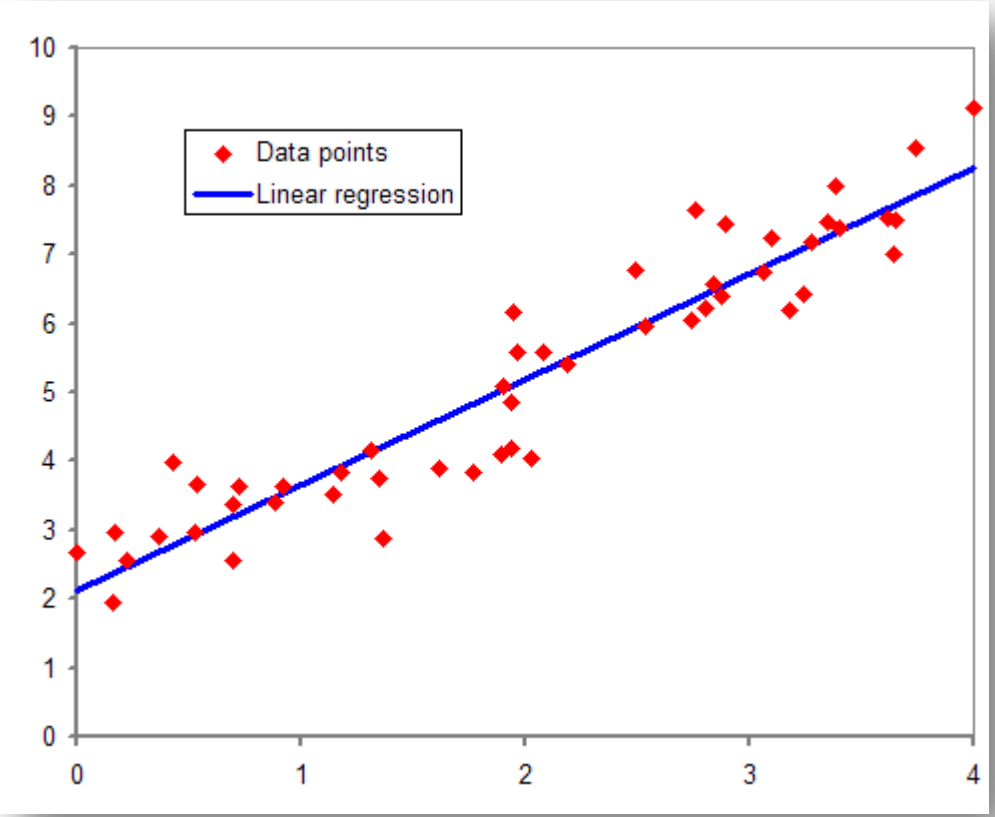
Esempio 3 …



Mangiare peperoncini e' disastroso per la tua salute. Un recente studio ha infatti dimostrato che tutti quelli che hanno mangiato peperoncino dal 1845 al 1850 sono morti. Ora, se i peperoncini sono dannosi, immaginiamoci gli ospedali. Tutto il mondo sa che le tue probabilita' di morire in un ospedale sono molto piu' alte che quelle di morire in qualsiasi altro posto.
- Mauroemme -

REGRESSIONE LINEARE SEMPLICE

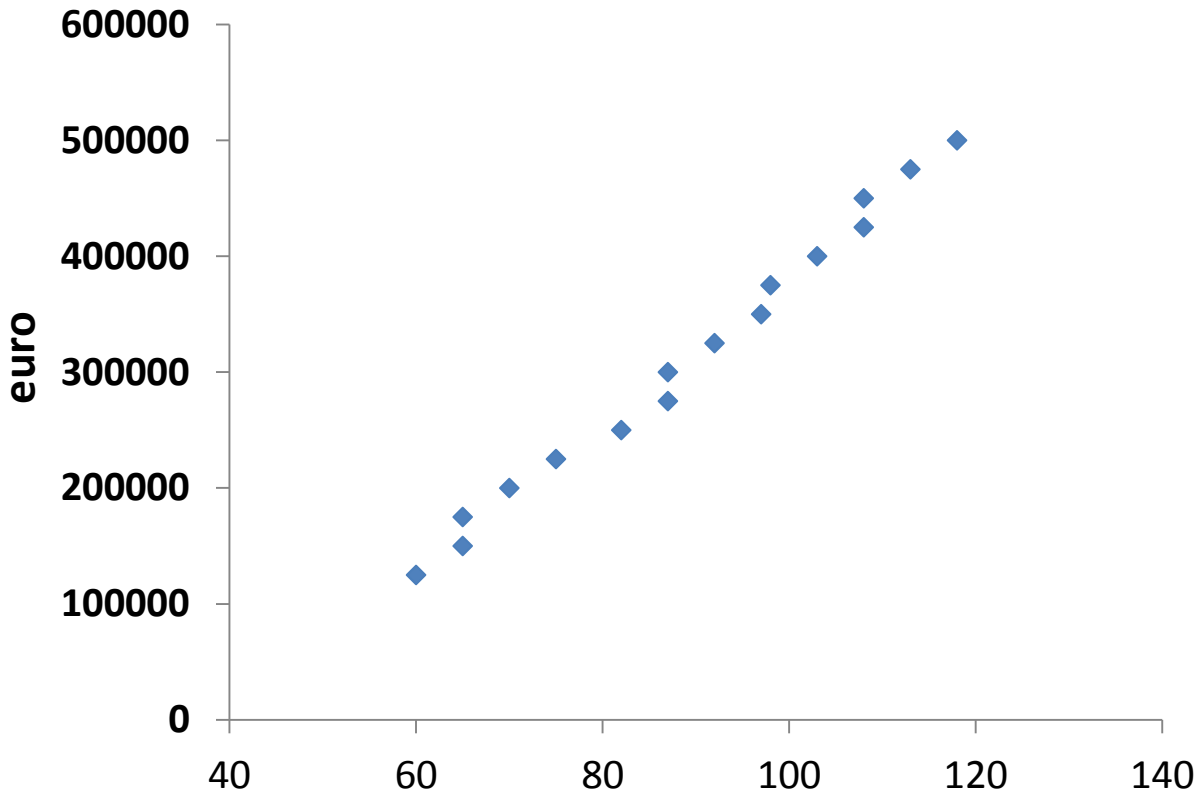
la **regressione lineare** rap-presenta un metodo di stima del valore atteso condizionato di una variabile ***dipendente***, o *endogena*, dati i valori di altre variabili ***indipendenti***, o *esogene*.



Il metodo dei minimi quadrati, deriva una retta che interpola uno scatter di punti minimizzando la somma dei quadrati delle distanze dei punti stessi dalla retta.

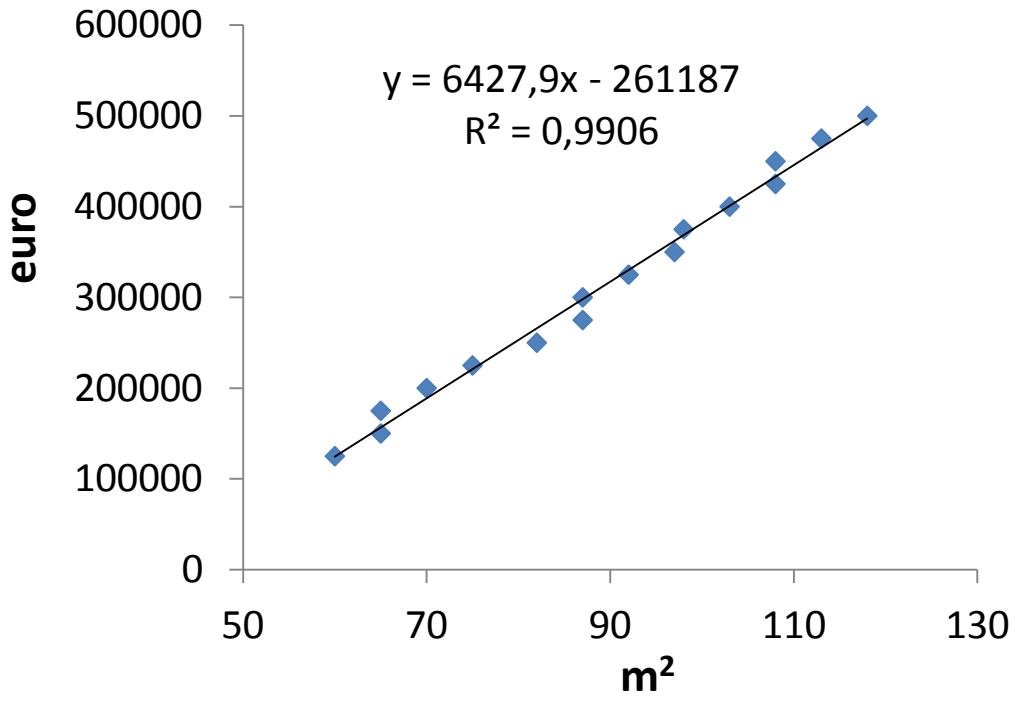
superficie	costo casa
60	125000
65	150000
65	175000
70	200000
75	225000
82	250000
87	275000
87	300000
92	325000
97	350000
98	375000
103	400000
108	425000
108	450000
113	475000
118	500000

r =0,995


600000
500000
400000
300000
200000
100000
0
euro
40
60
80
100
120
140
m²

r =0,995

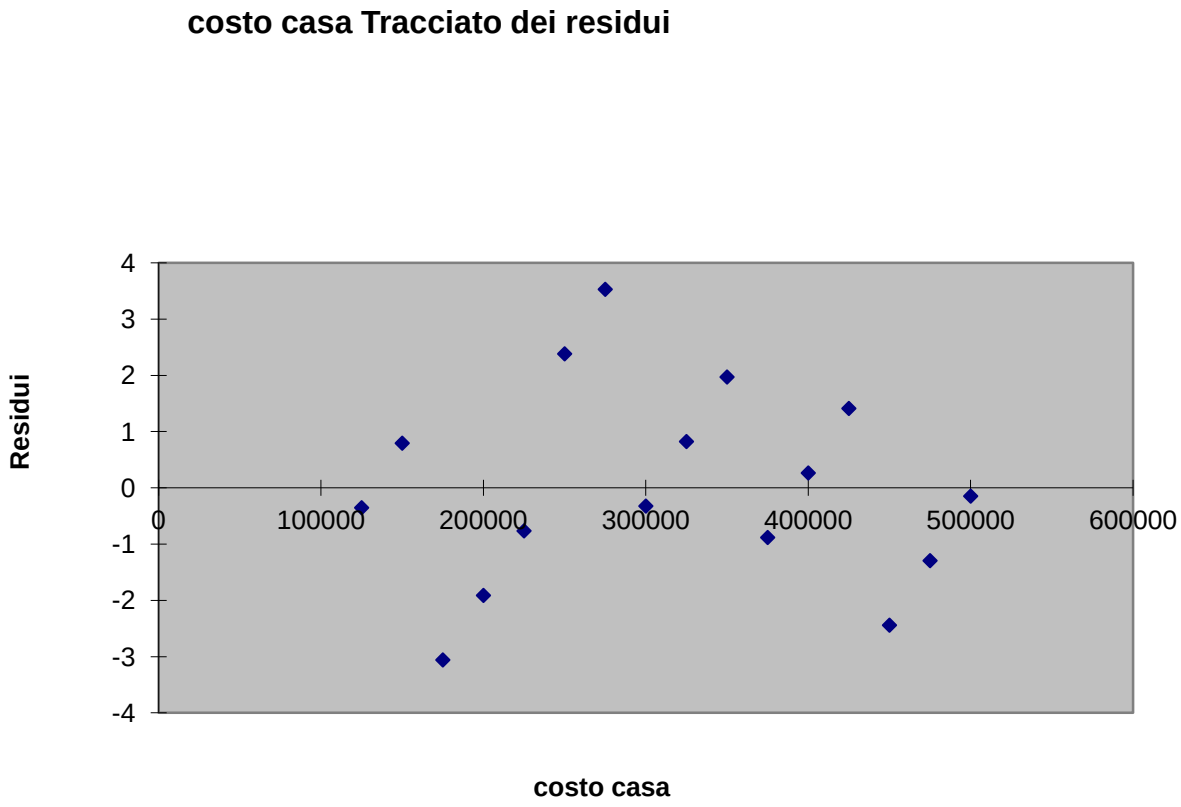
coefficiente di determinazione, (più comunemente R^2), è una proporzione tra la variabilità dei dati e la correttezza del modello statistico utilizzato. Non esiste una definizione concordata di R^2. Nelle regressioni lineari esso è semplicemente il quadrato del coefficiente di correlazione:



Se i tuoi dati non sono precisi, traccia una riga molto grossa.
- Regola di Albinak sui grafici -

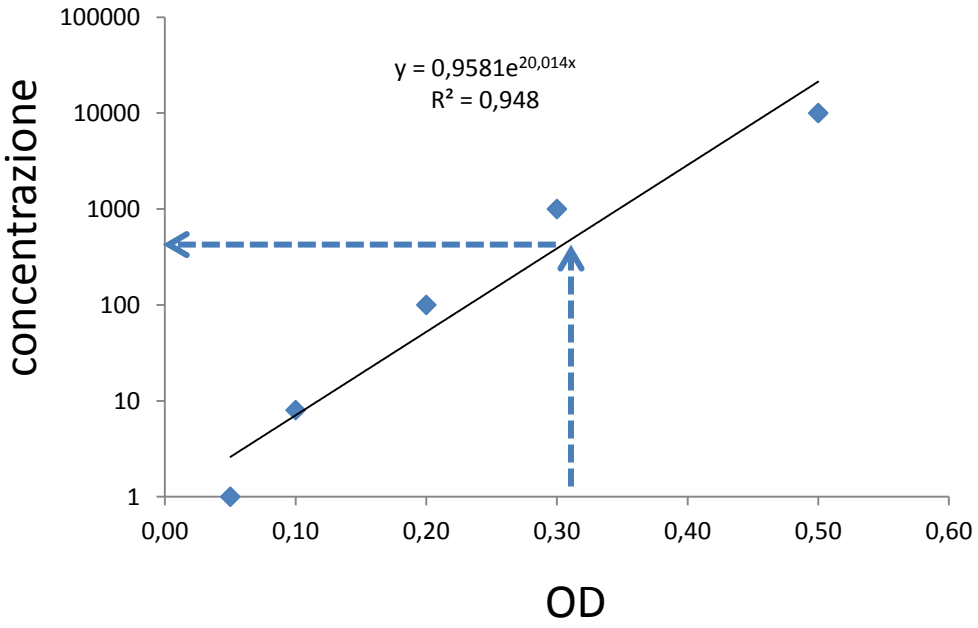
Per la valutare la significativà statistica del modello nel suo insieme viene utilizzato **il test *F basato*** sul rapporto tra varianza spiegata dal modello e varianza residua.
Se il p-value osservato è minore del p-value teorico (solitamente 0.05) il modello utilizzato spiega una quota significativa di varianza del fenomeno.

r = 0,995313
p<0.0001


costo casa Tracciato dei residui
Residui
4
3
2
1
0
-1
-2
-3
-4
0
100000
200000
300000
400000
500000
600000
costo casa

Esempio retta taratura

OD	concentrazione enalita
0,05	1
0,10	8
0,20	100
0,30	1000
0,50	10000



OD campione = 0,32

concentrazione = 579,2416

REGRESSIONE LINEARE MULTIVARIATA

costo casa	n° farmacie vicine	distanza centro	piano	anno costruzione	distanza viabilità	superficie
125000	6	250	6	1954	200	60
150000	3	300	4	1936	250	60
175000	4	250	4	1965	180	65
200000	3	220	3	1968	150	70
225000	1	180	1	1973	190	55
250000	4	169	4	2004	80	82
275000	3	300	5	1998	110	87
300000	1	120	3	2000	120	78
325000	5	89	2	1994	90	83
350000	5	24	1	1993	90	67
375000	3	80	1	2001	100	98
400000	2	100	5	2005	200	90
425000	1	60	3	1958	78	95
450000	5	120	2	1935	80	108
475000	6	25	1	2000	34	89
500000	3	10	1	2010	5	94
	-0,004187254	**-0,855097664**	**-0,57074**	**0,39940657**	**-0,775907084**	**0,827505543**
	NS	**p<0.0001**	**0,0209**	**NS**	**0,0004**	**p<0.0001**

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0,961953367
R al quadrato	0,925354281
R al quadrato corretto	0,875590468
Errore standard	41981,76232
Osservazioni	16

ANALISI
VARIANZA

	gdl	*SQ*	*MQ*	*F*	*Significatività F*
Regressione	6	1,96638E+11	32772964116	18,5949233	0,000133688
Residuo	9	15862215305	1762468367		
Totale	15	2,125E+11			

	Coefficienti
Intercetta	285306,71
n° farmacie vicine	-2867,821061
distanza centro	-632,1291229
piano	-5082,126583
anno costruzione	-84,46606416
distanza viabilità	-44,39216187
superficie	3936,412268

Altro esempio …

anni di vita	colesterolo	glicemia	pressione min	freq card	stile di vita (1 - 5)
56	321	188	188	186	1
67	300	200	200	169	2
68	342	156	190	184	1
71	287	189	156	160	1
73	220	177	180	120	2
75	256	145	138	134	3
77	210	166	100	80	4
81	180	123	77	110	5
84	180	110	80	90	2
87	177	86	92	123	5
96	160	80	60	53	5
	-0,8878	-0,8862	-0,8845	-0,8647	0,7872

r vs. anni di vita

OUTPUT RIEPILOGO

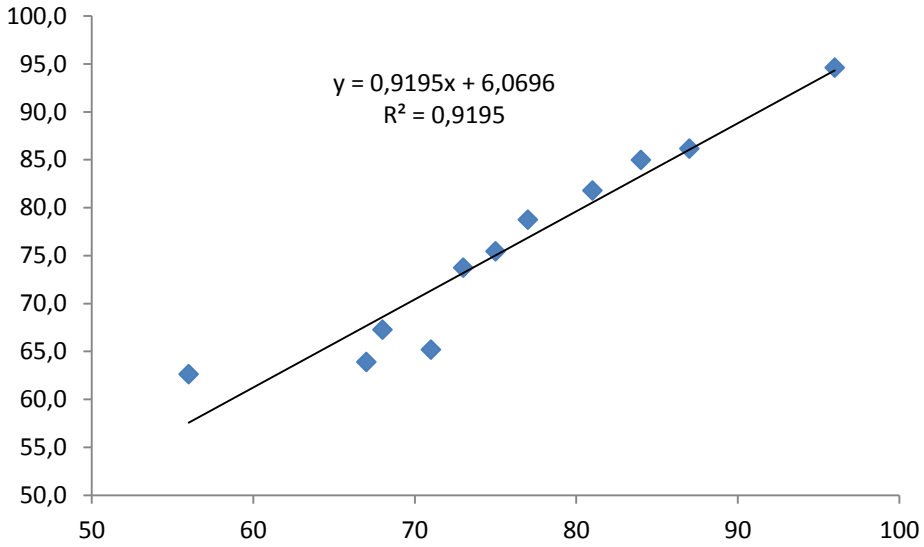
Statistica della regressione	
R multiplo	**0,95892**
R al quadrato	**0,91952**

Verifica dati

	Coefficienti
Intercetta	111,6407
colesterolo	-0,01621
glicemia	-0,14676
pressione min	0,027452
freq card	-0,11672
stile di vita (1 - 5)	0,378917

anni di vita	anni calcolati dal modello
56	62,6
67	63,9
68	67,3
71	65,2
73	73,7
75	75,5
77	78,8
81	81,8
84	85,0
87	86,2
96	94,6

r =0,95892



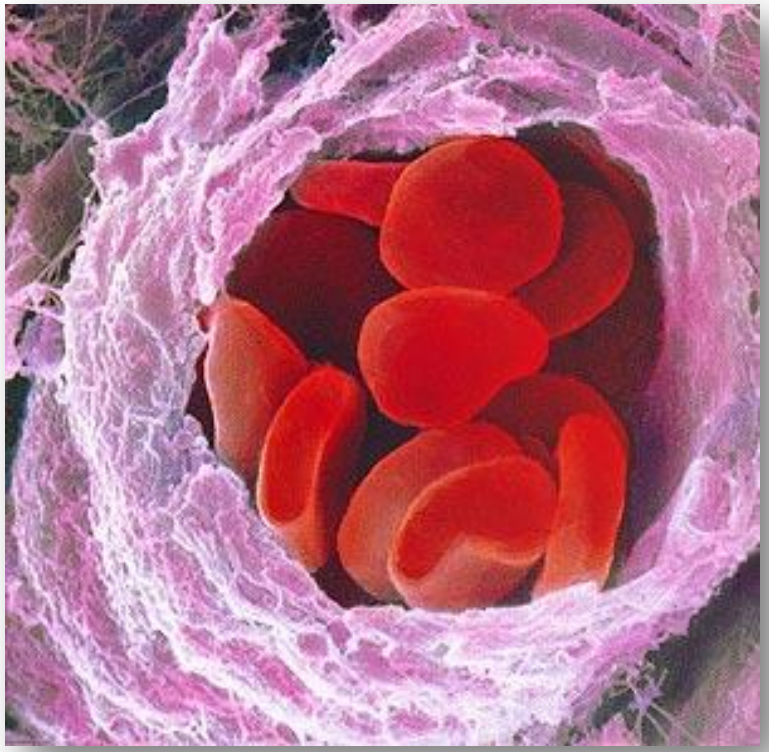


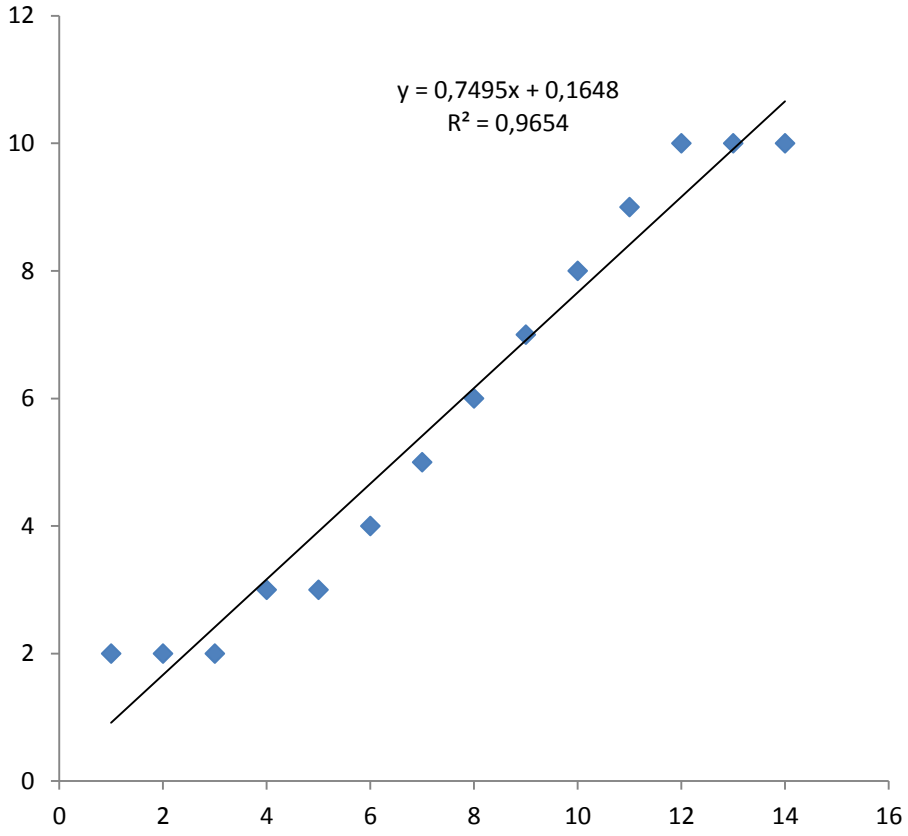
Table 2 *p* value of correlation of MCV with other hematological parameters in CTR and MICRO group of dogs assessed by a linear multivariate regression model

	CTR	MICRO
WBC	NS	$p<0.05$
RBC	$p<0.001$	$p<0.001$
Hgb	$p<0.05$	NS
Hct	NS	$p<0.001$
MCH	$p<0.001$	NS
MCHC	$p<0.001$	NS
RDW	NS	$p<0.001$
PLT	NS	NS
MPV	NS	$p<0.05$

NS, non significant=$p>0.05$

CONFRONTO TRA METODICHE SPERIMENTALI/ANALITICHE

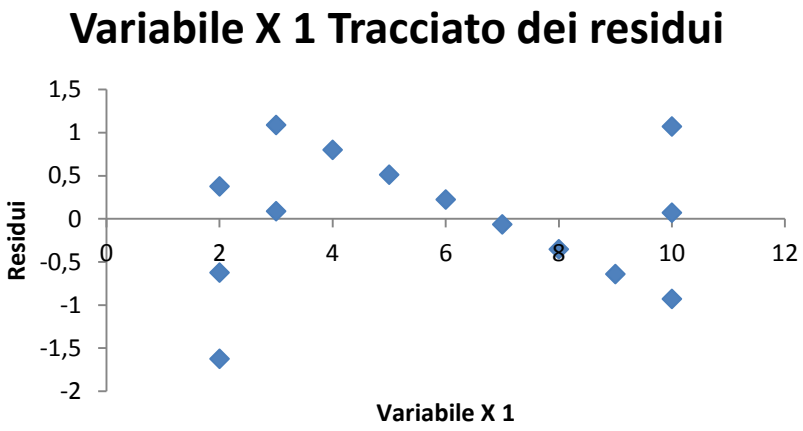
metodo 1	metodo 2
1	2
2	2
3	2
4	3
5	3
6	4
7	5
8	6
9	7
10	8
11	9
12	10
13	10
14	10



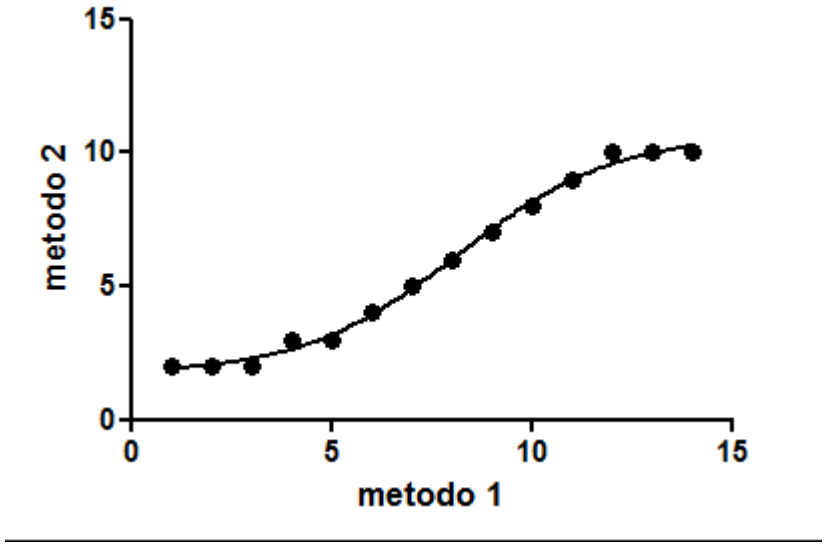
CONFRONTO TRA METODICHE SPERIMENTALI/ANALITICHE

OUTPUT RIEPILOGO

Statistica della regressione	
R multiplo	0,982562
R al quadrato	0,965428
R al quadrato corretto	0,962547
Errore standard	0,809582
Osservazioni	14



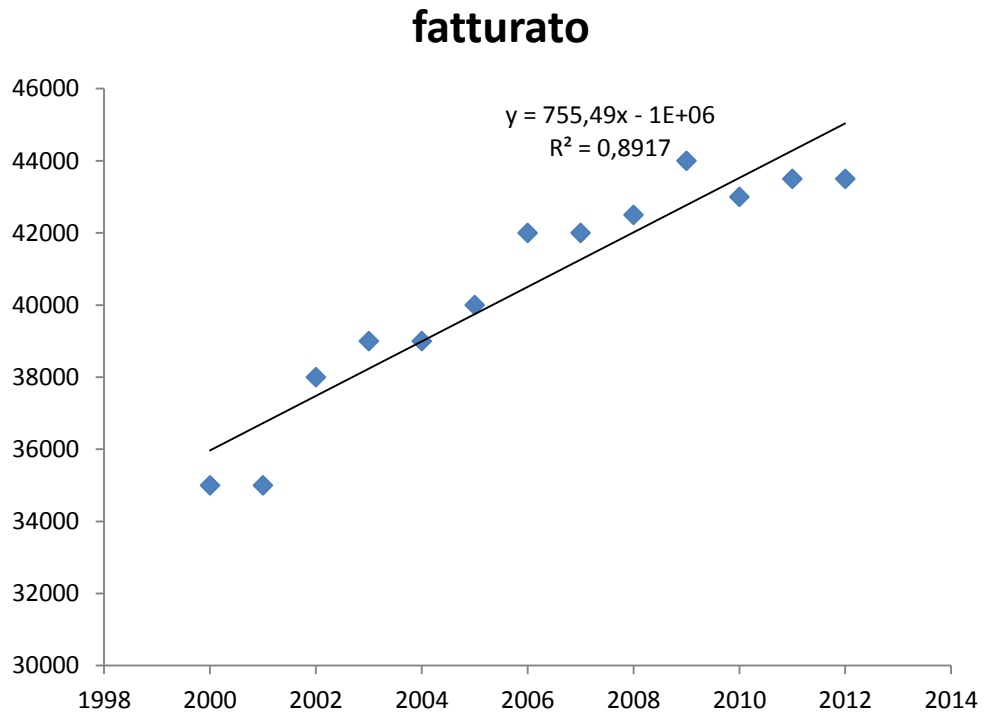
CONFRONTO TRA METODICHE SPERIMENTALI/ANALITICHE



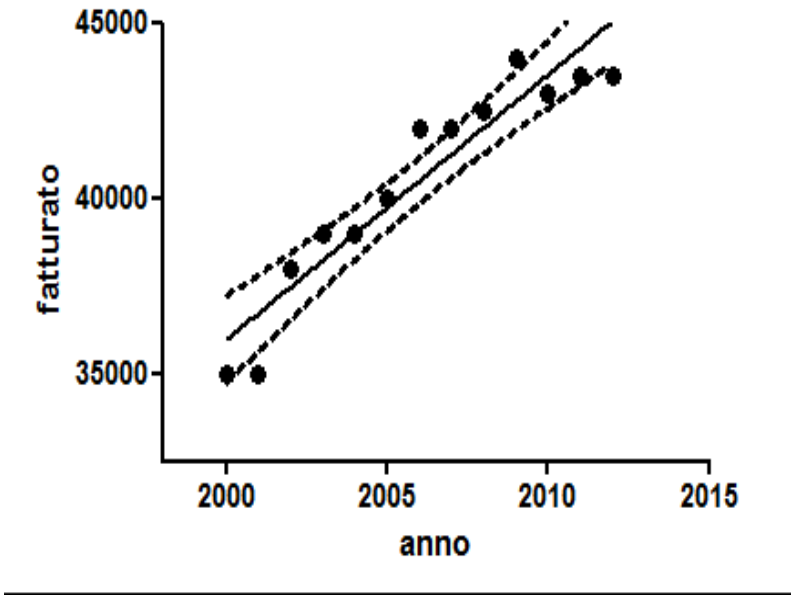
Boltzmann sigmoidal	
Best-fit values	
BOTTOM	1,688
TOP	10,68
V50	8,155
SLOPE	1,952
Std. Error	
BOTTOM	0,2276
TOP	0,3054
V50	0,1873
SLOPE	0,2040
95% Confidence Intervals	
BOTTOM	1.181 to 2.195
TOP	9.999 to 11.36
V50	7.738 to 8.572
SLOPE	1.497 to 2.406
Goodness of Fit	
Degrees of Freedom	10
R^2	0,9958
Absolute Sum of Squares	0,5560
Sy.x	0,2358
Number of points Analyzed	14

LA REGRESSIONE COME METODO PREVISIONALE

anno	fatturato
2000	35000
2001	35000
2002	38000
2003	39000
2004	39000
2005	40000
2006	42000
2007	42000
2008	42500
2009	44000
2010	43000
2011	43500
2012	43500
2012	?
2013	?



LA REGRESSIONE COME METODO PREVISIONALE



Goodness of Fit	
r^2	0,89
Sy.x	1100
Is slope significantly non-zero?	
F	91
DFn, DFd	1.0, 11
P value	< 0.0001
Deviation from zero?	Significant
Data	
Number of X values	13
Maximum number of Y replicates	1
Total number of values	13
Number of missing values	0

2010,08	43582,110	967,3312	967,3312
2010,20	43672,760	982,8868	982,8868
2010,32	43763,420	998,6403	998,6403
2010,44	43854,070	1014,583	1014,583
2010,56	43944,730	1030,705	1030,705
2010,68	44035,380	1046,999	1046,999
2010,80	44126,040	1063,457	1063,457
2010,92	44216,700	1080,071	1080,071
2011,04	44307,350	1096,835	1096,835
2011,16	44398,010	1113,740	1113,740
2011,28	44488,660	1130,782	1130,782
2011,40	44579,320	1147,955	1147,955
2011,52	44669,970	1165,251	1165,251
2011,64	44760,630	1182,666	1182,666
2011,76	44851,290	1200,195	1200,195
2011,88	44941,940	1217,832	1217,832
2012,00	45032,600	1235,574	1235,574

Regressioni non lineari: sigmoide

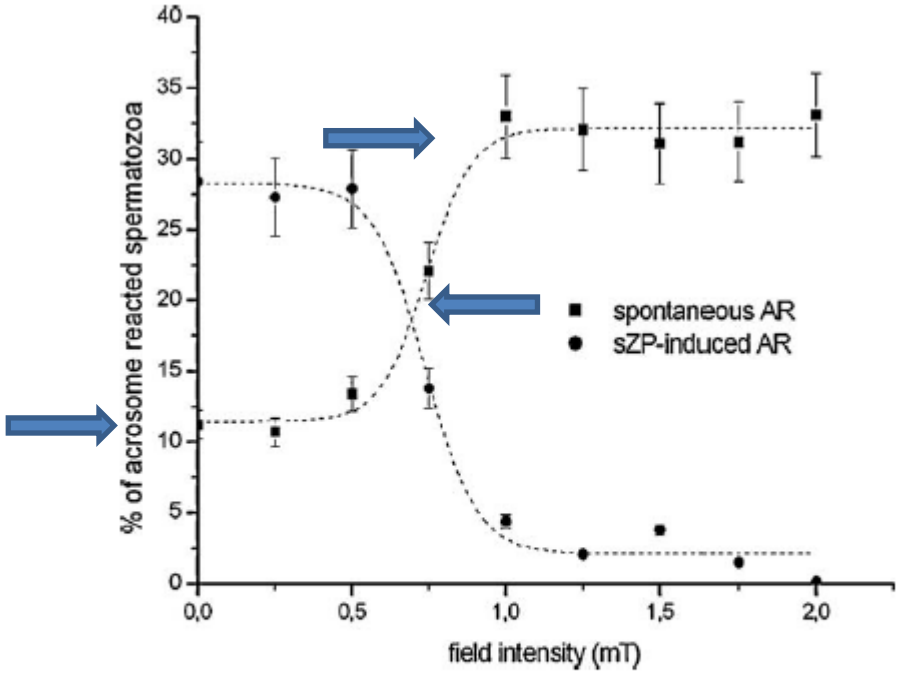
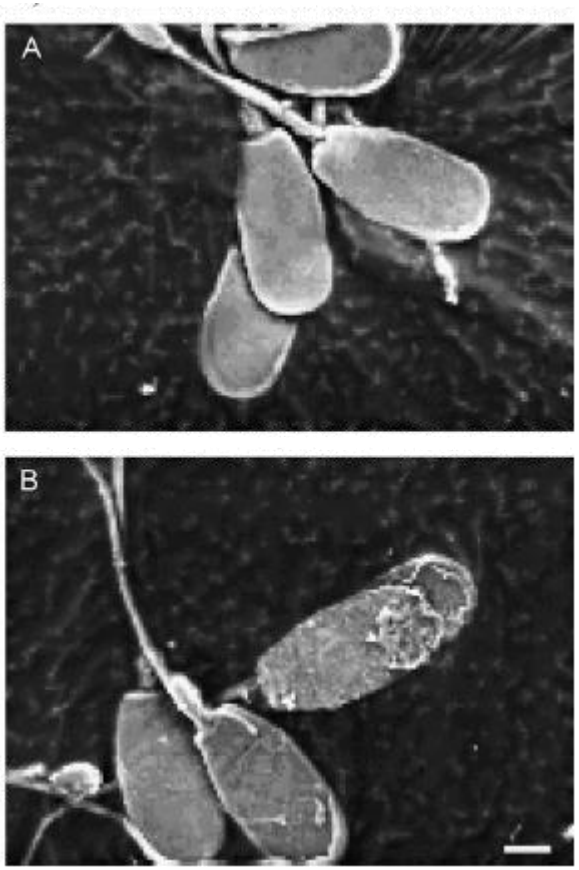
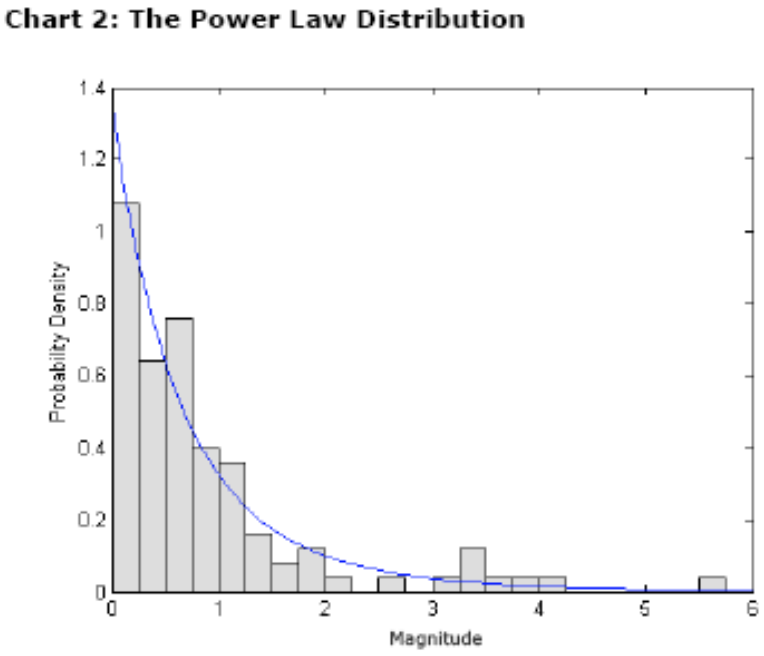


Fig. 3. Relationship between field intensity and the percentage of acrosome loss in alive spermatozoa (black squares) or the percentage of spermatozoa able to respond to sZP coincubation with AR (i.e., capacitated cells) (black circle). All the values are represented as mean ± SD.



Regressioni non lineari: power law

$$y = ax^{-b}$$



Regressione logistica

- La **regressione logistica** è un caso particolare di modello lineare generalizzato avente come funzione link la funzione logit. Si tratta di un modello di regressione applicato nei casi in cui la variabile dipendente *y* sia di tipo dicotomico riconducibile ai valori 0 e 1, come **lo sono tutte le variabili che possono assumere esclusivamente due valori: *vero* o *falso*,*maschio* o *femmina*, *vince* o *perde*, *sano* o *ammalato*, ecc**.

Logit

Da Wikipedia, l'enciclopedia libera.

Il **logit** è una funzione, che si applica a valori compresi nell'intervallo (0,1), tipicamente valori rappresentanti probabilità. Viene definito come

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \ln(p) - \ln(1-p)$$

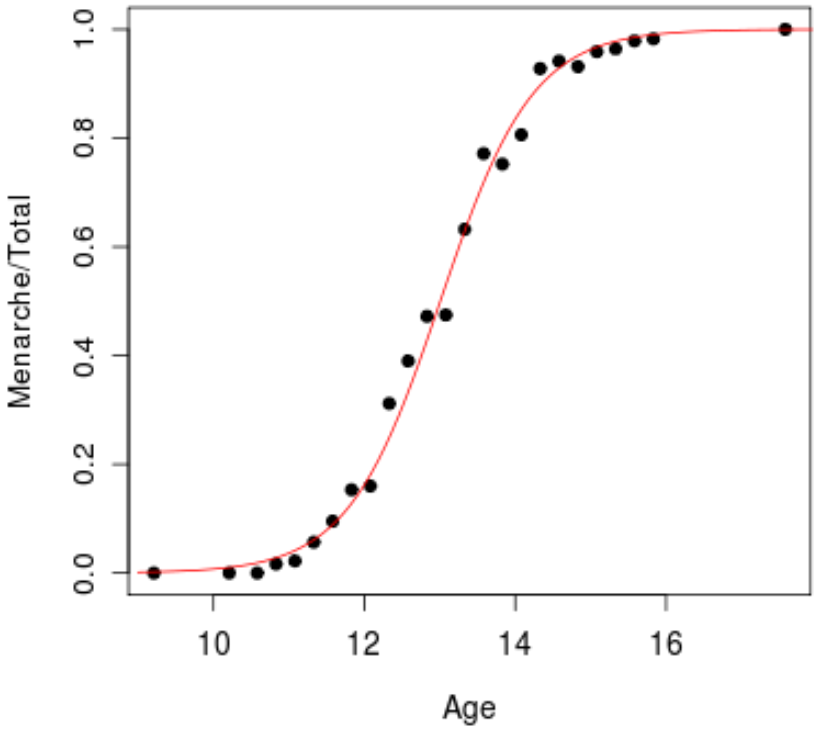
dove *ln* è il logaritmo naturale e $\frac{p}{1-p}$ è detto odds.

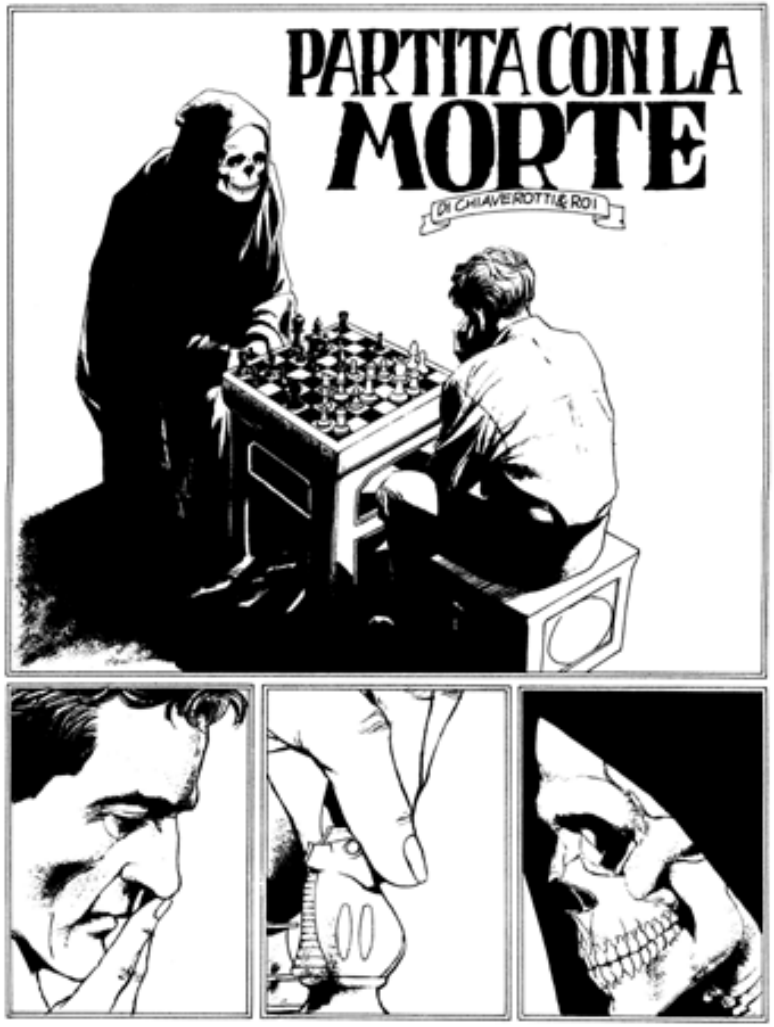
Ha come funzione inversa

$$p = \frac{e^{\operatorname{logit}}}{1 + e^{\operatorname{logit}}}$$

La funzione logit si applica ad esempio nella regressione logistica e nella variabile casuale logistica.

Esempi …


Menarche/Total
0.0
0.2
0.4
0.6
0.8
1.0
10
12
14
16
Age


PARTITA CON LA MORTE
DI CHIAVEROTTI & ROI

CLUSTERING - GROUPING

Windows Phone

Offerta Privati

LG Optimus 7 Ricaricabile	Vantaggi	Mobile Internet	Costo Telefono	
Ricaricabile	SIM Vodafone con 5€ di traffico	3€ settimana 500MB inclusi	399€	Avvisami

LG Optimus 7 Abbonamento	Vantaggi	Mobile Internet	Costo Telefono	Contributo Mensile	Dettagli
Stile Libero New	9 cent al minuto verso tutti	Incluso 2GB al mese	0€	19€	Avvisami
Tutto Facile Small	50€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	44€	Avvisami
Tutto Facile Medium	100€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	69€	Avvisami
Tutto Facile Large	150€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	84€	Avvisami
Tutto Facile Top Club	200€ per chiamare e inviare SMS	Incluso 2GB al mese	0€	100€	Avvisami

sky

Scegli da tre a cinque **GENERI** di Mondo.

INTRATTENIMENTO | BAMBINI | DOCUMENTARI | MUSICA | NEWS

Aggiungi i **PACCHETTI** che ti interessano.

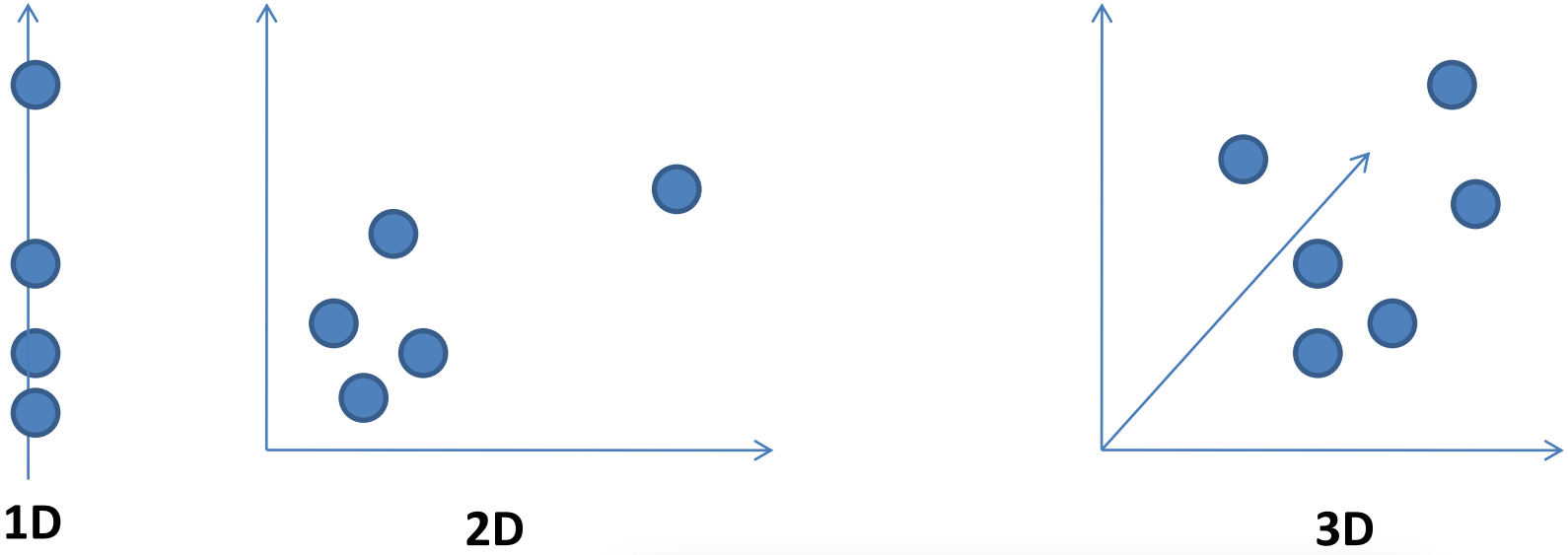
CINEMA | SPORT | CALCIO

Trova la tua combinazione ideale.

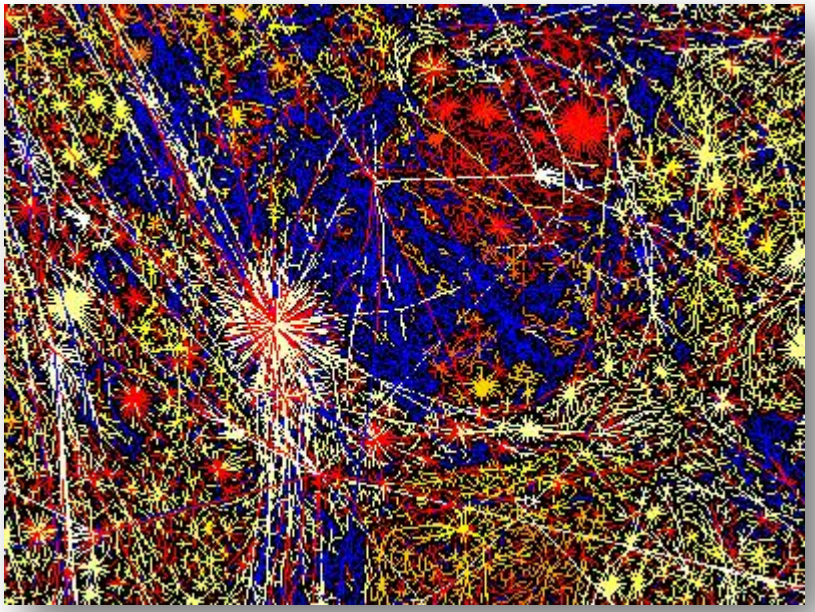
HD SEMPRE INCLUSA	3 GENERI	4 GENERI	5 GENERI
MONDO	19.90€	24.90€	29.90€
MONDO + CINEMA	34€ NOVITÀ	39€ NOVITÀ	43€
MONDO + SPORT CALCIO (1 PACCHETTO a scelta tra Sport e Calcio)		39€ NOVITÀ	43€
MONDO + CINEMA SPORT CALCIO + CINEMA SPORT CALCIO (2 PACCHETTI a scelta tra Cinema Sport e Calcio)		52€ NOVITÀ	56€
MONDO + CINEMA SPORT CALCIO		65€ NOVITÀ	69€

vodafone

CLUSTERING



Le tecniche di *clustering* si basano su misure relative alla somiglianza tra gli elementi. In molti approcci questa similarità, o meglio, dissimilarità, è concepita in termini di distanza in uno spazio multidimensionale.



nD

Due diversi approcci

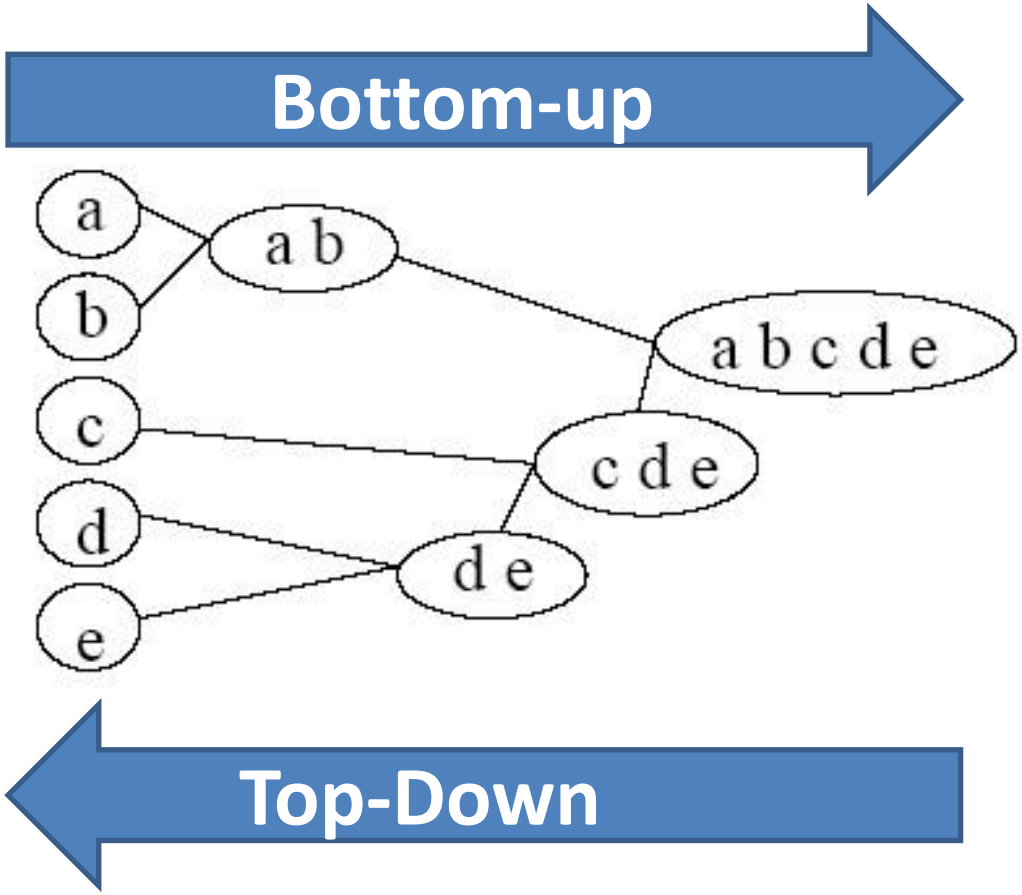
***metodi aggregativi o Bottom-Up*:**

inizialmente tutti gli elementi sono considerati *cluster* a sé, e poi l'algoritmo provvede ad unire i *cluster* più vicini. L'algoritmo continua ad unire elementi al *cluster* fino ad ottenere un numero prefissato di *cluster*, oppure fino a che la distanza minima tra i *cluster* non supera un certo valore, o ancora in relazione ad un determinato criterio statistico prefissato.

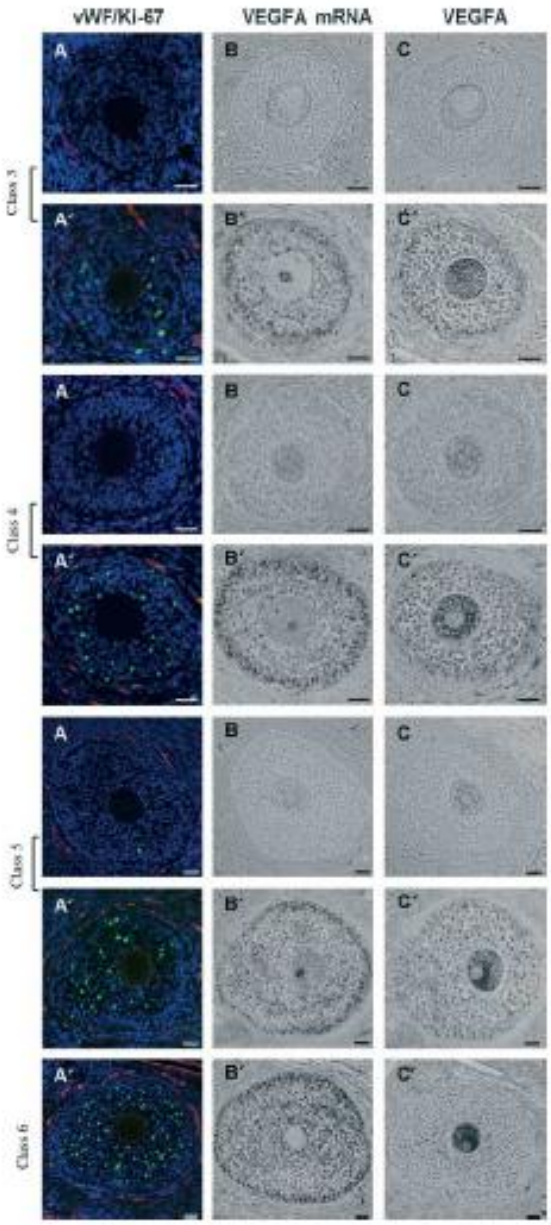
***metodi divisivi o Top-Down*:**

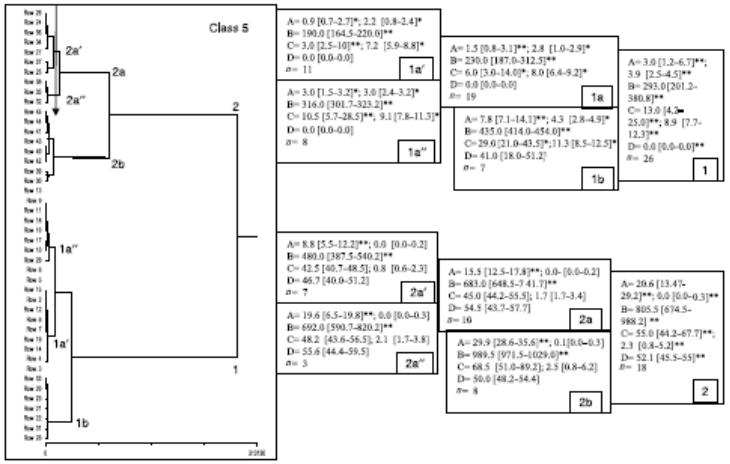
all'inizio tutti gli elementi sono un unico *cluster*, e poi l'algoritmo inizia a dividere il *cluster* in tanti *cluster* di dimensioni inferiori. Il criterio che guida la divisione è naturalmente quello di ottenere gruppi sempre più omogenei. L'algoritmo procede fino a che non viene soddisfatta una regola di arresto generalmente legata al raggiungimento di un numero prefissato di *cluster*.

Clustering gerarchico



dendogramma





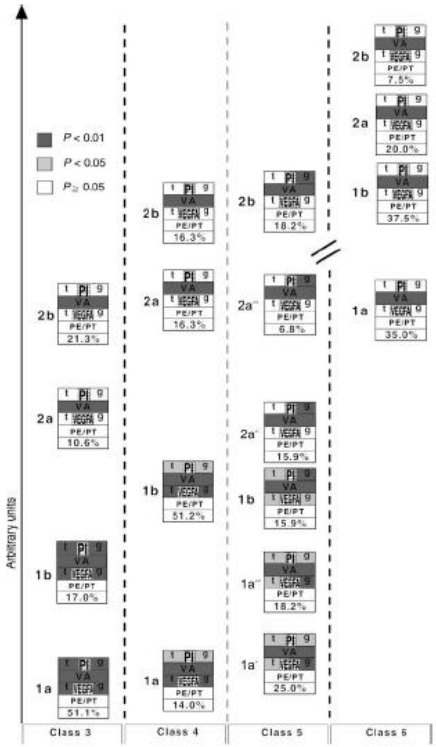


Figure 5 A general model that describes the distribution of the follicular subpopulations identified inside each follicular class by using cluster analysis. The subpopulations were represented within each class and arranged on the *y*-axis by considering the distances between the bifurcation obtained in the dendograms. The grey scale represents significant differences recovered for somatic and vascular parameters. The thickness of the dotted line represents the statistical difference among classes. t PI, theca proliferation index; g PI, granulosa proliferation index; VA, vascular area; t VEGFA, theca *VEGFA* mRNA; g VEGFA, granulosa *VEGFA* mRNA; PE/PT, proportion of proliferating endothelial cells; % value, percentage of preantral follicles belonging to the subpopulation.

CLUSTERING

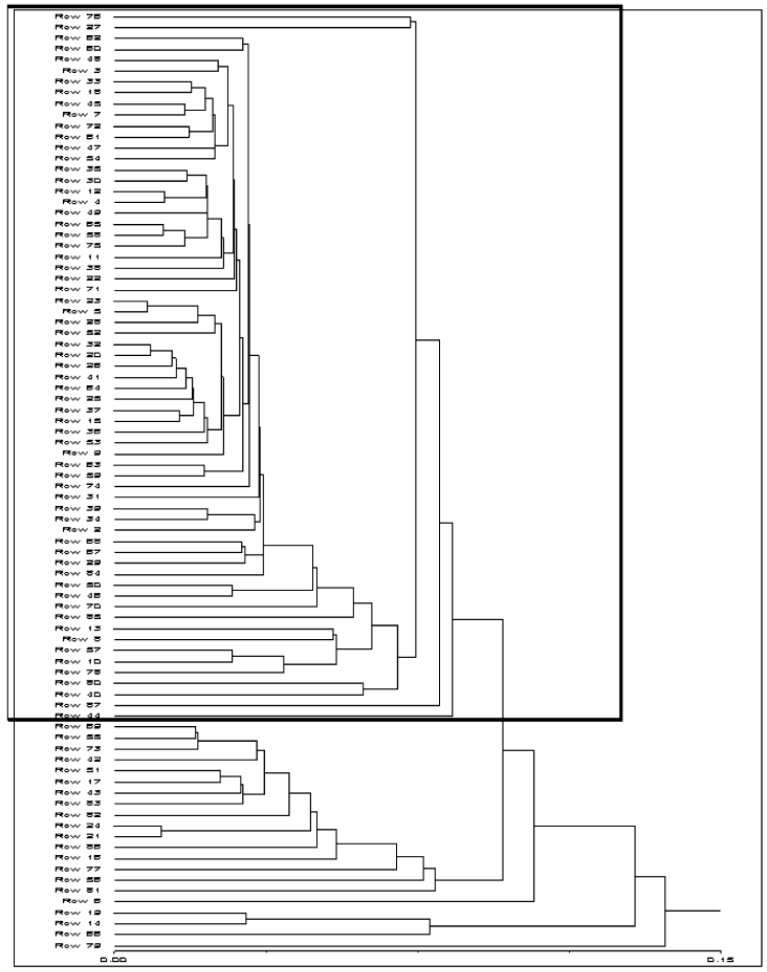


Fig. 2 Multivariate hierarchical cluster analysis of MICRO dogs. The *caption* shows the area of interest where clustering is significant

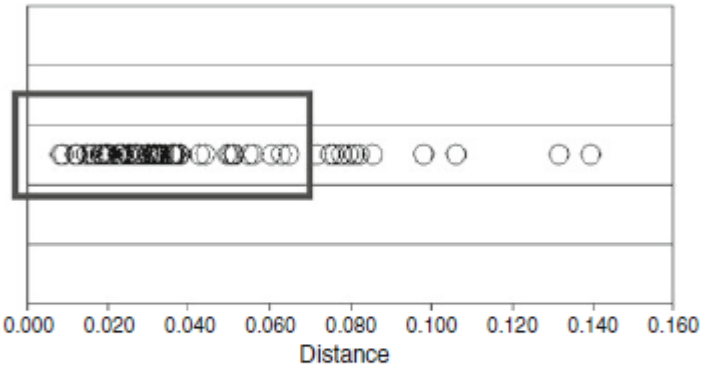
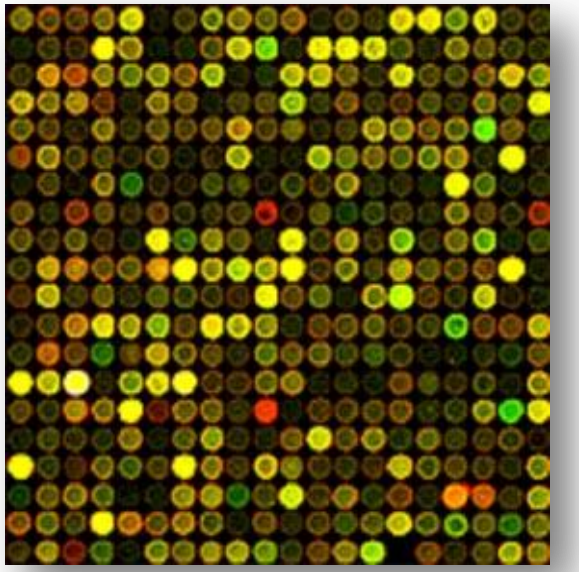
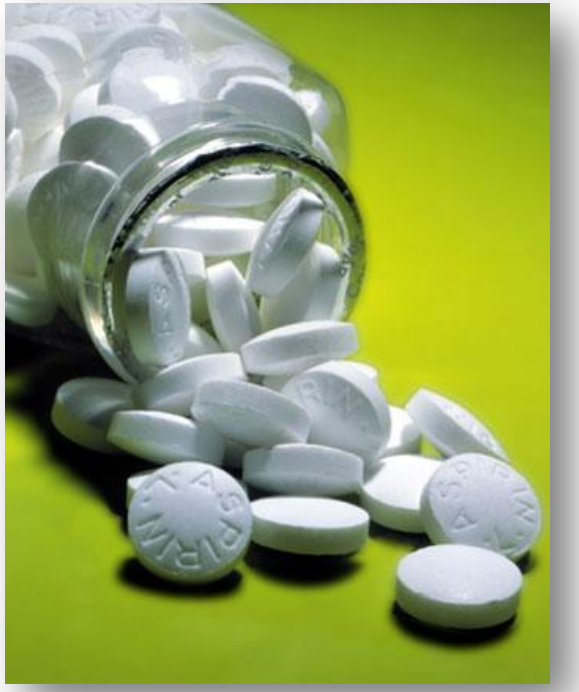
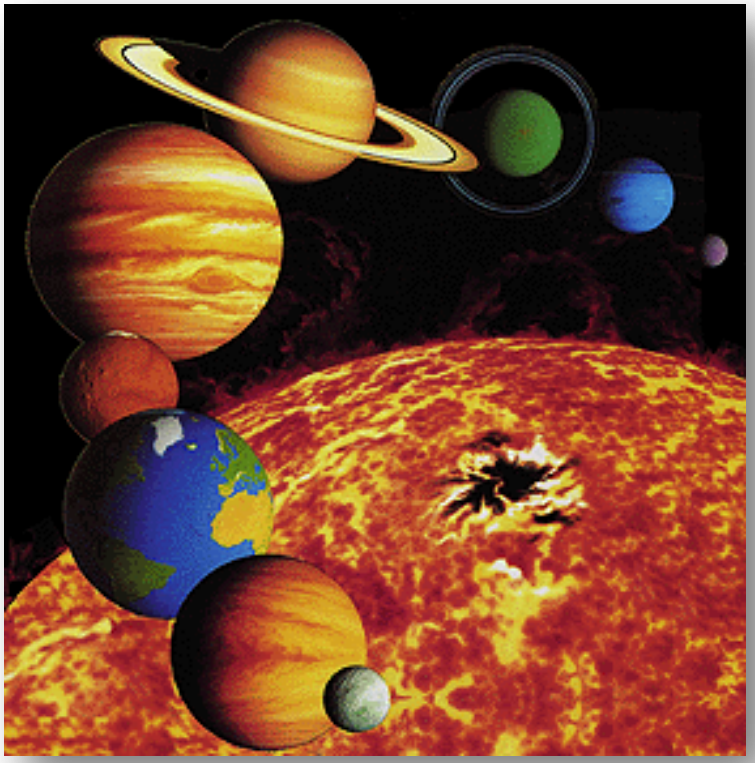


Fig. 3 Mono-dimensional distribution of distances among the dogs as measured by clustering. The *caption* shows a clear subgroup among the MICRO dogs

Individualità
Complessità








SOMEDAY, COMPUTERS WILL TAKE OVER OUR LIVES.
SOMEDAY?
COMPUTER WINS ON "JEOPARDY"
MIKE SMITH

FONTI

- L. Lison: ***Statistica applicata alla biologia sperimentale***. Casa Editrice Ambrosiana. Milano
- A. Camussi , F. Moller , E. Ottaviano, M. Sari Gorla: ***Metodi statistici per la sperimentazione biologica***. Zanichelli. Bologna
- M. R. Middleton: ***Analisi statistica con Excel .*** Apogeo. Milano
- http://www.bayes.it
- http://www.quadernodiepidemiologia.it/epi/HomePage.html
- Statistica 6.0. **User Manual**
- http://it.wikipedia.org/wiki/Portale:Matematica

Non esistono i dati, solo interpretazioni!
- Friedrich Nietzsche -