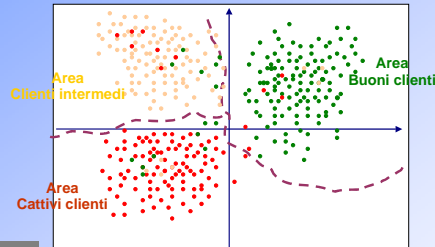


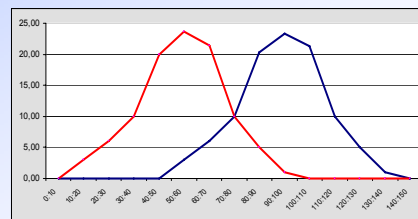
# L'Analisi Multidimensionale dei Dati

## Analisi Discriminante

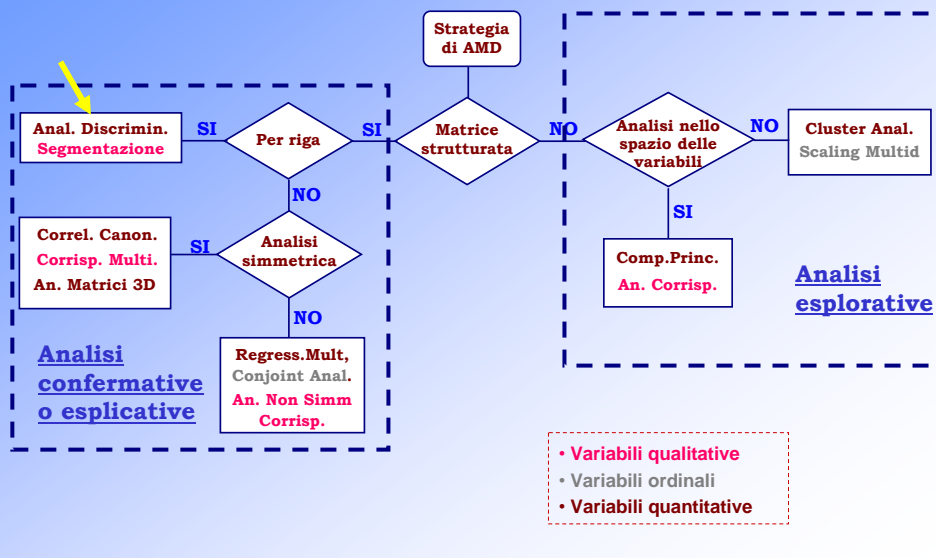
Una Statistica da vedere



		Classificazione a priori			
		Buoni	Intermedi	Cattivi	TOT.
Classificazione a posteriori	Buoni	580	95	38	713
	Intermedi	120	380	64	564
	Cattivi	50	75	298	423
	TOT.	750	550	400	1700



## Matrici e metodi



## I metodi per le decisioni: l'Analisi discriminante

• L'AD rientra fra i metodi utilizzati nel trattamento di matrici che comprendono una variabile **privilegiata qualitativa** che costituisce il **fenomeno da spiegare** e **p** variabili quantitative, che costituiscono l'insieme delle **variabili esplicative**.

• Gli individui sono **ripartiti a priori in k gruppi**, sulla base delle differenti modalità di una variabile qualitativa che costituisce la variabile da spiegare.

### OBIETTIVO

**DESCRITTIVO:** Confrontare la classificazione a priori con quella ottenuta **sulla base dei valori assunti dalle variabili esplicative**.

**DECISIONALE:** Assegnare un nuovo individuo, **di cui non si conosce il gruppo di appartenenza**, ad uno dei gruppi, con la **minima probabilità di commettere un errore di classificazione**.

## I dati nell'Analisi Discriminante

• **variabili esplicative**

1 ..... j .....

1	
⋮	
⋮	$E_1$
⋮	
⋮	$E_2$
⋮	
⋮	
⋮	
⋮	$E_k$
⋮	
⋮	

• **variabile di risposta**

1
1
1
2
2
2
k
k
k

## I dati di base

- Una matrice partizionata di dati quantitativi
- La codifica disgiuntiva della variabile di risposta qualitativa

$$X =$$

	1	.....	p	1	.....	k		
1	$E_1$			1	0	0	0	0
⋮				0	1	0	0	0
⋮	$E_2$							
⋮								
$i$	⋮							
⋮								
⋮	$E_k$			0	0	1	0	0
$n$								

## Notazioni : i gruppi

Gli individui sono ripartiti in  $k$  gruppi definiti a priori sulla base delle differenti modalità di una variabile qualitativa che costituisce la variabile da spiegare.

Questa suddivisione individua una partizione in  $k$  gruppi  $E_1, E_2, \dots, E_k$  tale che:

$$\cap \{E_1, E_2, \dots, E_k\} = 0 ; \quad \cup \{E_1, E_2, \dots, E_k\} = I$$

-

-

## Notazioni: le matrici

•  $\mathbf{X}$  : Matrice dei dati quantitativi ( $n, p$ )

•  $\mathbf{g}_j$ : baricentro del gruppo  $\mathbf{G}_j$

•  $\mathbf{g}$ : baricentro della nube totale

$\mathbf{W}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} \{(\mathbf{x}_i - \mathbf{g}_j)(\mathbf{x}_i - \mathbf{g}_j)'\}$  • Matrice var-cov gruppo  $\mathbf{G}_j$  ( $j=1, \dots, k$ )

$\mathbf{B} = \sum_{j=1}^k \left\{ \frac{n_j}{n} (\mathbf{g}_j - \mathbf{g})(\mathbf{g}_j - \mathbf{g})' \right\}$  • Matrice var-cov tra i gruppi

$\mathbf{V} = \frac{1}{n} \sum_{i=1}^n \{(\mathbf{x}_i - \mathbf{g})(\mathbf{x}_i - \mathbf{g})'\}$  • Matrice var-cov totale

## il caso dei punti pesanti ....

☑  $r = 1, 2, \dots, k$  ;  $\text{card}(E_r) = n_r$   $\sum n_r = n = I$

☑ peso individui:  $p_1, p_2, \dots, p_n$  (spesso  $p_1 = p_2 = \dots = p_n$ )

☑ peso classi :  $q_1, q_2, \dots, q_k$  ( $q_i = n_i / n$ )

☑ baricentro gruppo  $r$  :  $\mathbf{g}_r = \sum_i \mathbf{e}_i \cdot p_i$

☑ baricentro:  $\mathbf{g} = \sum_r q_r \cdot \mathbf{g}_r$

**Notazioni**

☑ varianza gruppo  $r$ :  $\mathbf{W}_r = \frac{1}{n_r} \sum_{\mathbf{e}_i \in E_r} p_i (\mathbf{e}_i - \mathbf{g}_r)(\mathbf{e}_i - \mathbf{g}_r)'$

☑ varianza comune entro le classi:  $\mathbf{W} = \sum_r q_r \mathbf{W}_r$

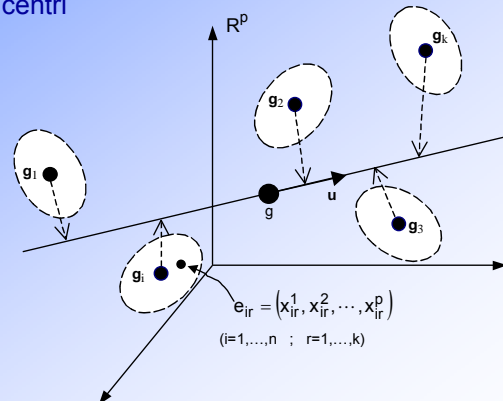
☑ varianza tra le classi:  $\mathbf{B} = \sum_r q_r (\mathbf{g}_r - \mathbf{g})(\mathbf{g}_r - \mathbf{g})'$

☑ varianza totale:  $\mathbf{V} = \sum_i p_i (\mathbf{e}_i - \mathbf{g})(\mathbf{e}_i - \mathbf{g})' = \mathbf{B} + \mathbf{W}$

## Il caso generale: $p$ variabili e $k$ gruppi

### OBIETTIVO

Ricerca le  $k-1$  combinazioni lineari  $y_j$  delle  $p$  variabili esplicative  $x_1, x_2, \dots, x_p$  in grado di determinare il sottospazio tale che le proiezioni dei baricentri  $g_j$  ( $j=1, \dots, k$ ) risultino il più distanziate possibile, avendo al tempo stesso il massimo raggruppamento degli individui attorno ai rispettivi baricentri



## L'Analisi Fattoriale Discriminante (AFD)

**Analisi in  $R^p$ :** L'Analisi Fattoriale Discriminante ricerca un sottospazio di  $R^p$  che minimizzi le distanze tra individui di uno stesso gruppo e massimizzi invece quelle tra individui di gruppi diversi.

**Obiettivo:**  $\left\{ \begin{array}{l} \text{Massimizzare la varianza tra i gruppi} \\ \text{Minimizzare la varianza 'inerzia entro i gruppi} \end{array} \right.$

### **Discriminazione lineare:**

*Ipotizza la distribuzione multinormale delle variabili osservate e l'uguaglianza delle matrici di varianza-covarianza dei gruppi (ipotesi di omoschedasticità).*

*Utilizza la metrica di Mahalanobis definita dalla matrice  $W^{-1}$ .*

### **Discriminazione quadratica:**

(Metodo di Sebestyen)

*Rifiuta l'ipotesi di omoschedasticità. Nel calcolo della distanza di un individuo da un gruppo considera  $k$  metriche definite dalle matrici  $W_r^{-1}$  ( $r=1, \dots, k$ ), definendo così un peso tanto maggiore quanto più il gruppo risulta disperso attorno al proprio baricentro.*

## la ricerca delle funzioni lineari discriminanti

### OBIETTIVO

Ricerca le  $k-1$  combinazioni lineari  $y_j$  delle  $p$  variabili esplicative  $x_1, x_2, \dots, x_p$  che consentono di separare al meglio i  $k$  gruppi.

#### • Prima funzione lineare discriminante: $y_1$

La prima combinazione lineare sarà la funzione:  $y_1 = a_{11}x_1 + a_{12}x_2 + \dots + a_{1p}x_p$

che rende massima la varianza **tra** le classi e, quindi, minima la varianza **entro** le classi.

#### • Seconda funzione lineare discriminante: $y_2$ e successive

La seconda combinazione lineare:  $y_2 = a_{21}x_1 + a_{22}x_2 + \dots + a_{2p}x_p$  sarà quella che, non correlata con la prima, renderà massima la varianza tra le classi, e così via

## Discriminazione lineare: le ipotesi

#### • Ipotesi di omoschedasticità:

uguaglianza delle matrici di var-cov entro i gruppi

#### • Ipotesi di multinormalità:

le variabili esplicative hanno una distribuzione multinormale

## La soluzione dell'ADL

**Obiettivo:**  $Q = \frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{W}\mathbf{u}} = \max$

**Soluzione:**  $\frac{\partial Q}{\partial \mathbf{u}} = \frac{2\mathbf{B}\mathbf{u}\mathbf{u}'\mathbf{W}\mathbf{u} - 2\mathbf{W}\mathbf{u}\mathbf{u}'\mathbf{B}\mathbf{u}}{(\mathbf{u}'\mathbf{W}\mathbf{u}) \cdot (\mathbf{u}'\mathbf{W}\mathbf{u})} = 0$

$$\Rightarrow \mathbf{B}\mathbf{u}\mathbf{u}'\mathbf{W}\mathbf{u} = \mathbf{W}\mathbf{u}\mathbf{u}'\mathbf{B}\mathbf{u}$$

**posto:**  $\lambda = \frac{\mathbf{u}'\mathbf{B}\mathbf{u}}{\mathbf{u}'\mathbf{W}\mathbf{u}} \Rightarrow \mathbf{B}\mathbf{u} = \lambda\mathbf{W}\mathbf{u} \Rightarrow \mathbf{W}^{-1}\mathbf{B}\mathbf{u} = \lambda\mathbf{u}$

**posto:**  $\mathbf{v} = \mathbf{W}\mathbf{u} \Rightarrow \mathbf{B}\mathbf{W}^{-1}\mathbf{v} = \lambda\mathbf{v}$

L'autovalore  $\lambda$  esprime il potere discriminante associato a ciascuna funzione lineare discriminante

## I legami con l'ACP

**ACP con  
metrica M**

$$\mathbf{X}'\mathbf{X} \mathbf{M} \mathbf{u} = \lambda \mathbf{u}$$

**AFD  
lineare**

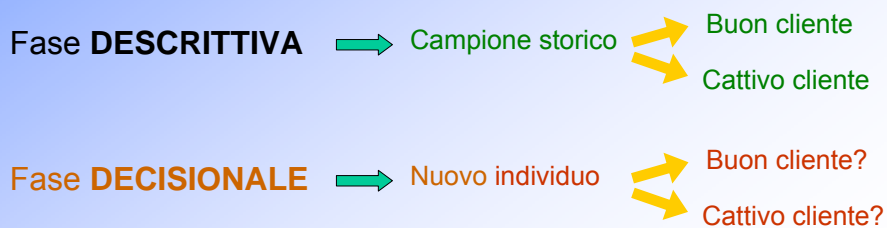
$$\mathbf{B} \mathbf{W}^{-1} \mathbf{v} = \lambda \mathbf{v}$$

Quindi... effettuare un'AFD sulle  $n$  unità utilizzando la metrica euclidea  $\mathbf{M}=\mathbf{I}$  equivale ad effettuare un'ACP sui  $k$  baricentri utilizzando la metrica  $\mathbf{W}^{-1}$ .

## • Esempio: il Credit Scoring

• Il **Credit Scoring** è un metodo statistico che consente di calcolare la probabilità, e quindi il rischio, di insolvenza legata alla concessione di un credito ad un cliente nuovo o già esistente.

• Il principio di base del **Credit Scoring** è che i **buoni clienti** ed i **cattivi clienti** si differenziano sulla base di caratteristiche osservabili sotto forma di variabili quantitative e qualitative.



## • Esempio: il Credit Scoring per i Prestiti Personali

Class.  
a priori

1 = Buon Cliente (max 1 insoluto)    2 = Cliente Medio (max 2-3 insoluti)    3 = Cattivo Cliente (>3 insoluti)

Variabili

- Età
- Reddito dichiarato (Lire × 1000)
- Debt burden (% indebitamento)
- Anni all'ultima occupazione
- Anni cliente (0 se nuovo)
- Importo Prestito (Lire × 1000)
- Durata Prestito (in mesi)

## • La matrice dei dati

		Variabili esplicative							Classificaz. a priori
		$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$x_8$
Buoni Clienti	i1	50	2.100	10	14	12	7.000	24	1
	i2	40	2.300	20	12	20	5.000	12	1
	i3	41	2.600	24	15	18	8.000	36	1
	i4	44	1.900	16	20	15	5.000	12	1
Clienti Intermedi	:	:	:	:	:	:	:	:	:
	i751	55	2.080	31	5	6	12.000	48	2
	i752	61	1.929	45	8	1	6.000	24	2
Cattivi Clienti	:	:	:	:	:	:	:	:	:
	i1301	31	1.200	50	6	0	10.000	48	3
	i1302	25	880	40	2	0	15.000	48	3
	i1700	34	1.400	45	7	0	15.000	48	3

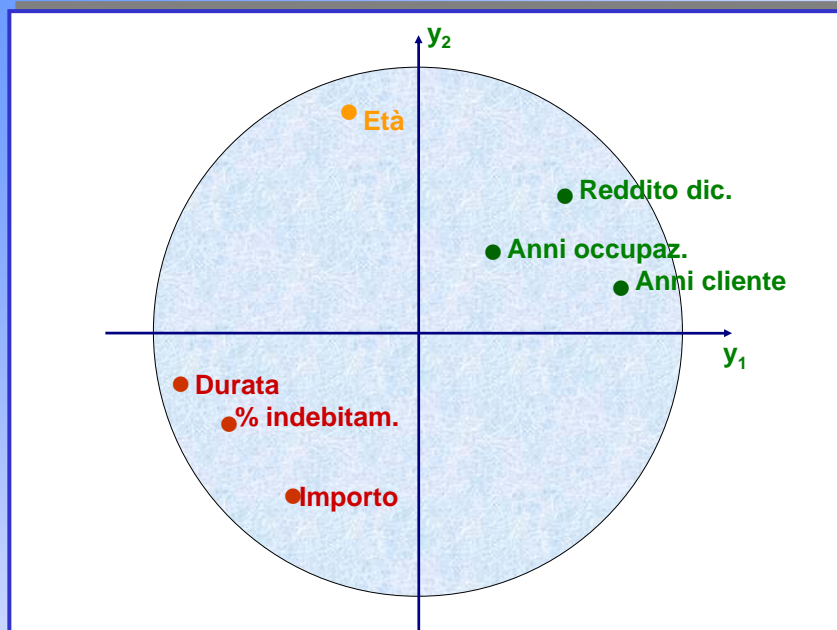
## • Autovalori e variabilità spiegata

Num	Valore	%	% cum.	Istogramma
1	.773	54.3	54.3	.....
2	.652	45.7	100.0	.....

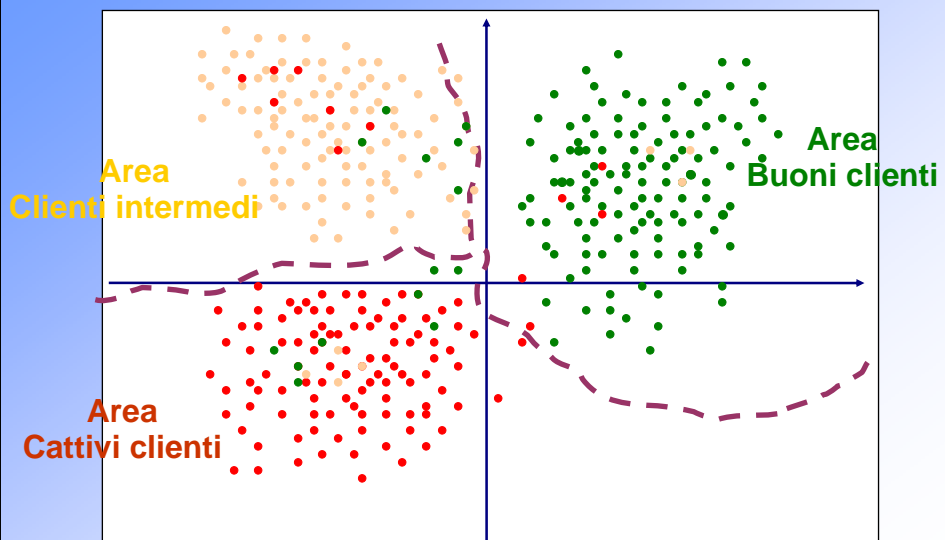
## • Correlazioni variabili-fattori

	(1)	(2)
Età	-0.239	0.923
Reddito dic.	0.556	0.494
% indebitamento	-0.780	-0.416
Anni occupazione	0.439	0.344
Anni cliente	0.862	0.299
Importo	-0.473	-0.622
Durata	-0.835	-0.354

• Cerchio delle correlazioni



• Rappresentazione delle unità



## • La regola di decisione

Ogni individuo verrà classificato nel gruppo il cui baricentro risulta più vicino...

$$e_i \rightarrow G_j \Rightarrow d^2(e_i, g_j) = \min[d^2(e_i, g_j) \quad j=1, \dots, k]$$

... e sarà quindi possibile *confrontare la nuova partizione* così ottenuta *con quella definita a priori* sulla base delle *k* modalità della variabile da spiegare.

## Le regole di classificazione

### 1. Regola grafica

Proiezione di una nuova unità come elemento supplementare sul piano fattoriale discriminante:

$$\phi_\alpha^S = (\mathbf{e}_s - \mathbf{g})' \mathbf{W}^{-1} \mathbf{u}$$

### 2. Regola numerica

Confronto delle *k* distanze dai baricentri dei gruppi:

$$e_s \rightarrow E_r \Leftrightarrow d^2(e_s, \mathbf{g}_r) = \min[d^2(e_s, \mathbf{g}_r) ; r = 1, \dots, k]$$

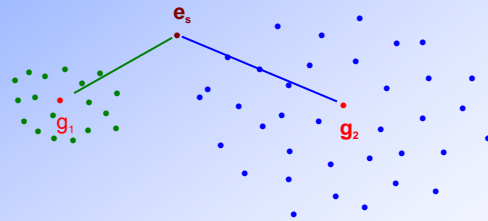
## La scelta della distanza<sup>2</sup>

**1. Distanza di Mahalanobis:**  $(e_s - g_r)' W^{-1} (e_s - g_r)$

**Vantaggi:** di facile calcolo

**Svantaggi:** non tiene conto della possibile diversa forma delle classi

**2. Distanza di Sebestyen:**  $(e_s - g_r)' W_r^{-1} (e_s - g_r)$



Il punto  $e_s$  risulta più vicino al baricentro del gruppo 1 che non a quello del gruppo 2.

Tuttavia, la differente dispersione dei due gruppi mette bene in evidenza come il punto debba essere considerato più logicamente parte del gruppo 2 che non del gruppo 1.

La distanza di Sebestyen, considerando k distanze dai baricentri, ciascuna ponderata con l'inverso della variabilità del gruppo corrispondente, consente di risolvere questo problema.

## • Valutazione dei risultati

		Classificazione a priori			
		Buoni	Intermedi	Cattivi	TOT.
Classificazione a posteriori	Buoni	580	95	38	713
	Intermedi	120	380	64	564
	Cattivi	50	75	298	423
	TOT.	750	550	400	1700

% ben classificati →  $(580+380+298)/1700 = 74,0\%$

## Riepiloghiamo?...

- Consideriamo una matrice con n unità su cui sono osservate p variabili;
- Le unità sono a priori ripartite in k gruppi sulla base delle diverse modalità di una variabile da spiegare;
- Si effettua un'analisi fattoriale discriminante per valutare il potere discriminante delle variabili esplicative;
- Il numero di fattori che è possibile estrarre è uguale a k-1 (*normalmente*  $k \leq 3$ );
- Si proiettano le unità sul piano fattoriale ;
- Se i gruppi risultano ben separati, allora le variabili esplicative hanno un buon potere discriminante e si può passare alla fase decisionale;
- Una nuova unità, di cui ovviamente non si conosce il gruppo di appartenenza, viene proiettata in supplementare sul piano determinato dalle unità attive;
- In base ad uno dei criteri di classificazione scelti, l'unità verrà assegnata al gruppo al quale risulta più vicina.

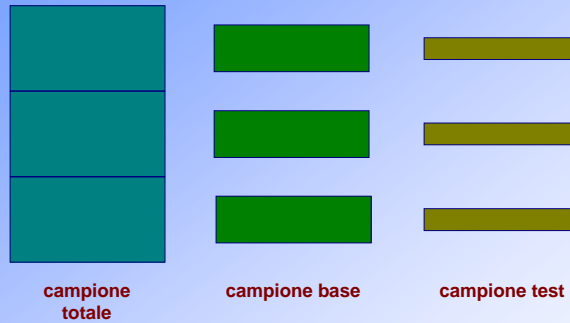
## La validazione dei risultati

		Classificazioni a posteriori			
		Buoni	Intermedi	Cattivi	
Classificazioni a priori	Buoni	580	120	50	750
	Intermedi	95	380	75	550
	Cattivi	38	64	298	400

% ben classificati  $\Rightarrow (580+380+298)/1700 = 74,0\%$

## La stabilità dei risultati: il campione test

**k=3**



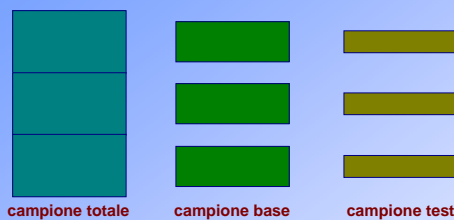
Sul campione base viene costruita la regola di classificazione che viene poi applicata al campione test. Su entrambi i campioni viene calcolata la % di corretta classificazione.

	E <sub>1</sub>	E <sub>2</sub>	E <sub>3</sub>
E <sub>1</sub>			
E <sub>2</sub>			
E <sub>3</sub>			

$$\% \text{ b.c.} = \frac{\text{Traccia}}{n}$$

## La stabilità dei risultati

**k=3**



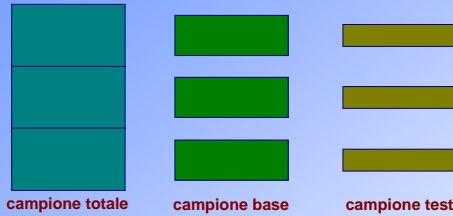
### Cross-validation

1. Si scelgono  $n_1$  elementi che costituiscono il campione di base, con i rimanenti  $n_2$  a formare il campione test, facendo in modo che  $n_2$  risulti molto più piccolo di  $n_1$  e che  $n_1/n_2$  sia un numero intero;
2. Si costruisce la regola di decisione sul campione di base e la si testa sul campione test, calcolando la % di b.c.,  $p_i$ ;
3. Si ripete lo stesso procedimento per  $i = \frac{n_1}{n_2}$  volte;
4. Si stima la % di b.c. complessiva come media delle  $p_i$ :

$$\% \text{ b.c.} = \frac{n_2}{n_1} \sum_{i=1}^{n_1/n_2} p_i$$

## La stabilità dei risultati

**k=3**

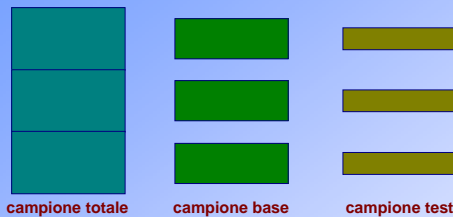


### Jackknife

1. Dal campione totale viene escluso un elemento, che costituisce il campione-test, tutti gli altri a costituire il campione di base;
2. Si costruisce la regola di decisione sul campione di base e la si testa sull'elemento escluso;
3. Si definisce una variabile  $a$  che assume valore 1 se l'elemento è classificato correttamente, 0 altrimenti;
4. Si ripete il procedimento escludendo, uno alla volta, tutti gli elementi;
5. Si stima la % di b.c. come:  $\% \text{ b.c.} = \sum_{i=1}^n \frac{a_i}{n}$

## La stabilità dei risultati

**k=3**



### Bootstrap

1. Dal campione totale si estraggono, con reintroduzione,  $n_1$  elementi che costituiscono il campione di base e  $n_2$  elementi che costituiscono il campione-test, con  $n_1+n_2=n$ ;
2. Si costruisce la regola di decisione sugli  $n_1$  elementi del campione di base e la si testa sugli  $n_2$  elementi del campione test;
3. Si calcola la % di b.c. ;
4. Si ripete il procedimento centinaia o migliaia di volte;
5. Si stima la % di b.c. come media delle % di b.c. precedentemente calcolate.

## I criteri di selezione delle variabili

Riducono i costi di gestione, elaborazione e rilevazione legati al numero di variabili considerate, migliorando inoltre l'affidabilità dei risultati

### Metodi di selezione stepwise

**Seleziona, al passo q, il gruppo di q variabili per le quali risulta:**

1. Criterio della traccia di Lawley-Hotelling

$$tr(\mathbf{W}_q^{-1}\mathbf{B}_q) = \max$$

2. Criterio del determinante di Wilks

$$\frac{\det(\mathbf{W}_q)}{\det(\mathbf{V}_q)} = \max$$

3. Criterio della % di ben classificati

$$\% \text{ b.c.} = \max$$

## Due casi particolari:

### • Caso di variabili qualitative e quantitative

1. ACM sulla tabella dei dati ricodificati

2. Scelta del numero di dimensioni (Fattori)

3. AD sulle nuove variabili

### • Caso di due gruppi

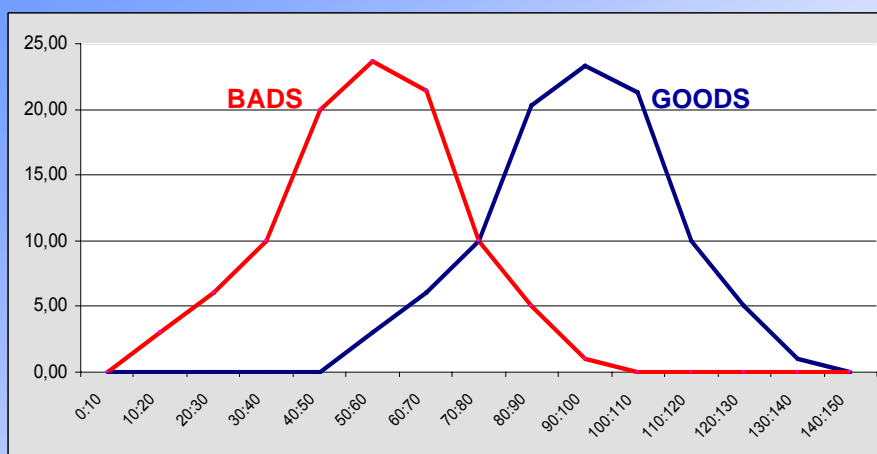
• Una sola funzione discriminante

• Un solo punto di taglio

• Esempio di tabella dei punteggi

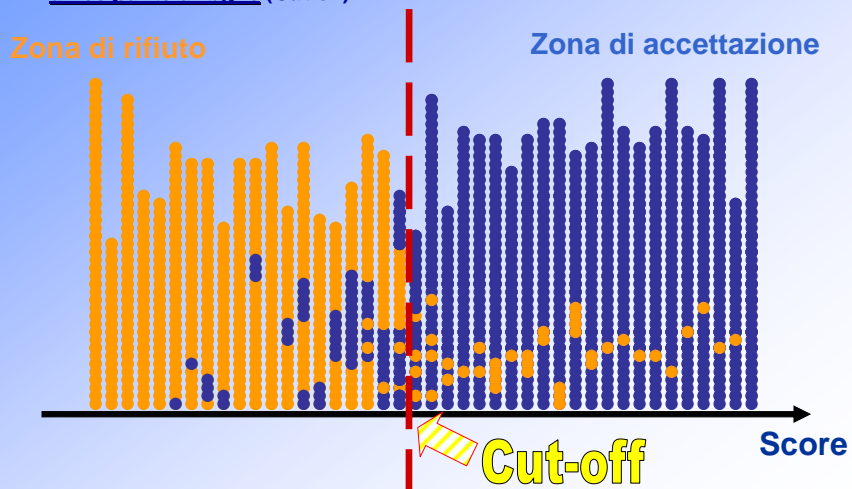
Classi di punteggi	GOODS	% GOODS nelle classi	% cum. GOODS	BADS	% BADS nelle classi	% cum. BADS
0:10	0	0,00	0,00	0	0,00	0,00
10:20	0	0,00	0,00	15	3,00	3,00
20:30	0	0,00	0,00	30	6,00	9,00
30:40	0	0,00	0,00	50	10,00	19,00
40:50	0	0,00	0,00	100	20,00	39,00
50:60	36	3,00	3,00	118	23,60	62,60
60:70	72	6,00	9,00	107	21,40	84,00
70:80	120	10,00	19,00	50	10,00	94,00
80:90	244	20,33	39,33	25	5,00	99,00
90:100	280	23,33	62,67	5	1,00	100,00
100:110	256	21,33	84,00	0	0,00	100,00
110:120	120	10,00	94,00	0	0,00	100,00
120:130	60	5,00	99,00	0	0,00	100,00
130:140	12	1,00	100,00	0	0,00	100,00
140:150	0	0,00	100,00	0	0,00	100,00
TOT	1200	100,00		500	100,00	

• Scelta del punto di taglio



## Il punto di taglio

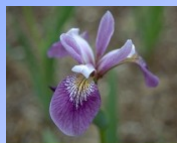
- Nel caso in cui è  $k=2$ , esiste un' unica funzione discriminante (*Score*);
- La regola di decisione per classificare un nuovo individuo sarà quindi individuata da un unico punto di taglio (*Cut-off*).



## Un altro esempio: gli Iris di Fisher



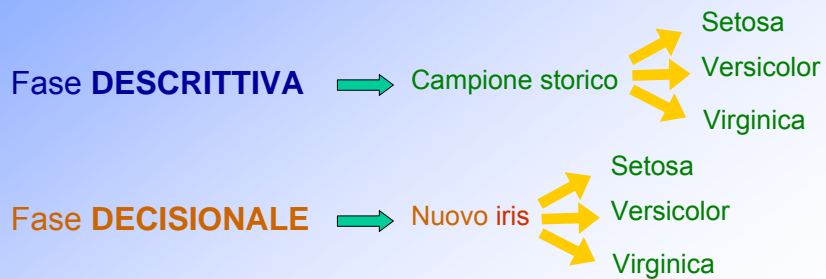
• *Iris setosa*



• *Iris versicolor*



• *Iris virginica*



• Esempio: gli iris di Fisher

Classe. a priori

1 = Setosa      2 = Versicolor      3 = Virginica

Variabili

- sepal length
- sepal width
- petal length
- petal width

• La matrice dei dati

		Variabili esplicative				Classificaz. a priori			
		x <sub>1</sub>	x <sub>2</sub>	x <sub>3</sub>	x <sub>4</sub>	x <sub>5</sub>			
Setosa	i1	50	33	14	14	1			
	i2	46	34	14	12	1			
	i3	46	36	10	15	1			
	i4	51	33	17	20	1			
	:	:	:	:	:	:			
Versicolor	i51	65	28	46	15	2			
	i52	62	22	45	15	2			
	:	:	:	:	:	:			
Virginica	i148	67	31	56	24	3			
	i149	63	28	51	15	3			
	i150	69	31	51	23	3			

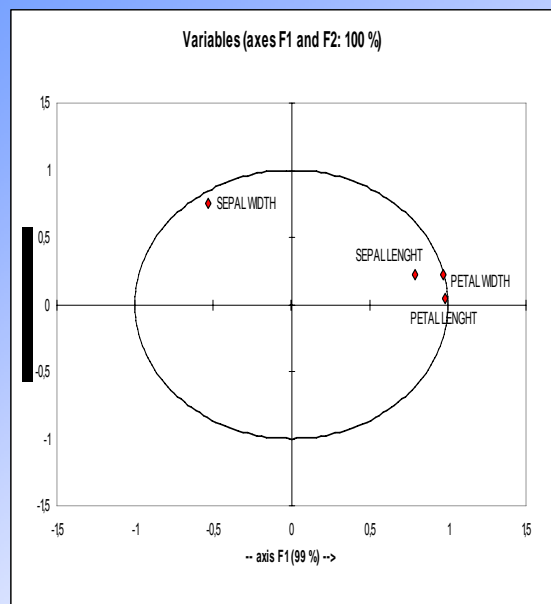
• **Autovalori e variabilità spiegata**

Num	Valore	%	% cum.	Istogramma
1	.32.19	99.12	99.12	.....
2	0.28	0.88	100.0	...

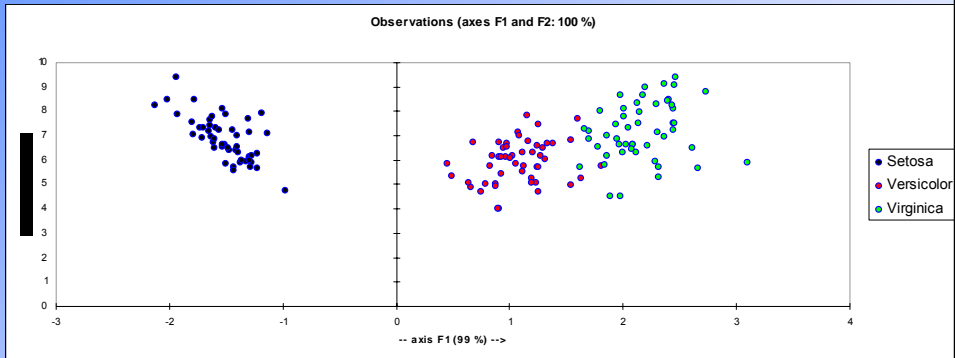
• **Correlazioni variabili-fattori**

	(1)	(2)
Sepal Length	0.792	0.218
Sepal width	-0.531	0.758
Petal length	0.985	0.046
Petal width	0.973	0.223

• **Cerchio delle correlazioni**



## • Rappresentazione delle unità



## • Valutazione dei risultati

		Classificazione a posteriori			
		Setosa	Versicolor	Virginica	TOT.
Classificazione a priori	Setosa	50	0	0	50
	Versicolor	0	48	2	50
	Virginica	0	1	49	50
	TOT.	50	49	51	150

% ben classificati →  $(50+48+49)/150 = 98,0\%$