CAPITOLO 8

L'analisi quantitativa dei fenomeni di insegnamento/apprendimento. Alcuni strumenti della statistica non parametrica: L'analisi implicativa di Gras e l'analisi fattoriale.

8.1 Introduzione

L'obiettivo di questo capitolo è quello di presentare alcuni strumenti statistici utilizzati frequentemente nella ricerca in didattica.

Verranno presentati alcuni esempi riguardanti l'utilizzo del χ^2, l'analisi implicativa di Regis Gras e brevi cenni sull'analisi fattoriale (delle corrispondenze e a componenti principali).

Vengono forniti alcuni riferimenti bibliografici comunemente citati nei lavori di ricerca in Didattica:

- Luigi Vajani, *Statistica Descrittiva*, Milano, ETAS libri, 1978.
- CNRS, *Pratique de l'Analyse des Donnèes*, Paris, Dunod, 1984 (I vol: *Analyse des correspondances exposé élémentaire*; II Vol: *Abrégé théorique études de Cas Modèle*).
- G. Brousseau, *Stratégies de l'analyse statistique*, Université Bordeaux I, LADIST, 1993.
- G. Brousseau, *Fiches de Statistiques non parametriques pour la didactique*, Université Bordeaux I, 1993, LADIST.
- G. B.Flores d'Arcais, *Metodi statistici per la ricerca psicologica*, Firenze, Giunti Barbera, 1968.
- S. Siegel, *Statistica non parametrica per le scienze del comportamento*, Firenze, OS, 1978.
- R. Gras, *L'implication statistique (Nouvelle méthode exploratoire de données, Recherches en Didactique des Mathématiques)*, La Pensée sauvage, Grenoble,1996.
- B. Escofier-Jerôme Pagès, *Analyse factorielles simples et multiples (objectifs, méthodes et interpretation*, Paris, Dunod, 1990.
- Web Site : Groupe International d'Analyse Statistique Implicative, International Group of Implicative Statistic Analysis, Gruppo Internazionale di Analisi Statistica Implicativa: http://math.unipa.it/~grim/asi/asi_index.htm . Nel sito si trovano tutti i proceedings dei convegni del gruppo internazionale dal 2000 ad oggi.
- Statistical Implicative Analysis : theory and applications, Series: Studies in Computational Intelligence , Vol. 127, Editors: R. Gras, E. Suzuki, F. Guillet, F. Spagnolo, 2008, XVI, 514 p. 147 illus., Hardcover, Springer, ISBN: 978-3-540-78982-6.
- "L'analisi a-priori e l'indice di implicazione statistica di Gras", Filippo Spagnolo, (pp. 110, 125) (Italian Version), Quaderni di Ricerca in Didattica (Matematica), n.7, 1997.

- "The statistic implicative index of Gras: a distance between the contingency and the a-priori", Filippo Spagnolo (pp. 110-125) (English Version), Quaderni di Ricerca in Didattica (Matematica), n.7, 1997.
- "Metodologia di analisi di indagini", Regis Gras, (pp. 99 - 109)., Quaderni di Ricerca in Didattica (Matematica), n.7, 1997.

La modellizzazione attraverso argomentazioni statistiche fornisce alla ricerca in didattica delle matematiche una maggiore possibilità di trasferibilità dell'esperienza.

Risulta evidente, come è stato ampiamente dibattuto nei precedenti capitoli, che senza una riflessione teorica dal punto di vista della didattica e quindi della epistemologia dei contenuti matematici, l'argomentazione statistica non avrebbe alcun peso. Soltanto uno studio in parallelo di tutti i possibili percorsi argomentativi della ricerca può portare a risultati considerati attendibili.

8.1.1 I dati.

Ogni ricerca didattica ci porta inevitabilmente a raccogliere dei dati che possiamo considerare formati da una collezione di informazioni elementari. Ogni informazione elementare riporta in generale un comportamento di un allievo in una situazione. Una statistica sarà quindi un insieme composto da: allievo, situazione, comportamento.

L'allievo appartiene ad un campione E osservato, supposto estratto da una popolazione più vasta, o a caso, o seguendo un sistema di situazioni di controllo (ad esempio: livello scolare, sesso, conoscenza personale anteriore...).

La situazione è scelta in un insieme S (di questioni, esercizi...) generata e strutturata da condizioni e parametri di varia natura (sapere in gioco, condizioni materiali, condizioni didattiche...).

I comportamenti (tipici delle conoscenze o di saperi mirati) sono presi in un insieme C di risposte possibili dell'allievo nelle condizioni nelle quali è posto.

Una classe può essere definita come un insieme di allievi E, un corso di Matematica come un insieme di esercizi S, i risultati degli allievi come una certa applicazione di E nell'insieme SxC dove C è l'insieme dei comportamenti di riuscita o di errore, una nota come una applicazione di SxC in R.

La conoscenza di un certo comportamento potrà essere rappresentato da una certa applicazione di un insieme di questioni in un insieme di comportamenti.

8.1.2 Utilizzazione della statistica da parte degli insegnanti e da parte dei ricercatori.

L'insegnante deve prendere rapidamente numerose decisioni e può correggerle molto velocemente se si dovessero rilevare inadatte. Non può aspettare il risultato del trattamento statistico di tutte le sue questioni. L'insegnante deve cercare di utilizzare quei trattamenti statistici che gli consentono rapidamente di trarre certe conclusioni.

Il ricercatore deve seguire un processo opposto:

- Quali ipotesi corrispondono alle questioni che ci interessano?
- Quali dati raccogliere?
- Quali trattamenti utilizzare?
- Quali conclusioni?

Più che la rapidità e l'utilità immediata, è la consistenza, la stabilità, la pertinenza e la sicurezza delle risposte che interessano il ricercatore.

La ricerca con opportuni metodi statistici consentirà:

- di comunicare tra insegnanti le informazioni di cui hanno bisogno e che raccolgono sui risultati degli allievi, il valore dei metodi impiegati...;
- di utilizzare anche con discernimento i risultati delle ricerche in didattica;
- di conoscere le possibilità e i limiti dei metodi statistici e quindi la legittimità delle conoscenze che essi utilizzano nella loro professione;
- di discutere questa legittimità;
- di formulare delle congetture suscettibili di essere sottomesse alla prova della contingenza sperimentale;
- di immaginare la plausibilità di queste congetture;
- di sapere come convertire la loro esperienza in conoscenza;
- di partecipare a delle ricerche.

8.1.3 Le osservazioni.

Una osservazione consiste in una attribuzione di un valore a una variabile a proposito di un individuo: l'oggetto osservato.

La statistica permette principalmente di trattare il caso dove:
- parecchie osservazioni sono raccolte.
- e dove queste osservazioni nel loro insieme:

i) riguardino individui differenti per una stessa proprietà;
ii) riferiscano proprietà differenti per uno stesso individuo;
iii) riguardino sia la i) che la ii).

Se "24" (valore osservato) è attribuito all'allievo X (oggetto d'osservazione), come "risultato dell'esame di matematica" (variabile osservata), l'insieme dei valori o dei casi possibili nel nostro caso è un intero compreso tra 0 e 30.

Le variabili possono essere numeriche, d'intervallo, ordinali, nominali.

- *Variabile Numerica*: quando i valori sono espressi in numeri (appartenenti agli insiemi N, Z, Q, R) e le operazioni che si possono fare con essi hanno un senso per la variabile.
- *Variabile d'Intervallo*: quando solo le differenze tra valori hanno senso mentre la somma non ne ha. Per esempio il punteggio ottenuto in una disciplina sportiva può costituire una variabile d'intervallo.
- *Variabile Ordinale*: quando i valori esprimono soltanto un ordine tra le osservazioni. In una variabile ordinale la somma di due valori non è un valore.
- *Variabile Nominale*: quando i valori sono dei caratteri o attributi. Questa variabile può essere a due valori: un suo attributo e la sua negazione. Anche se essa è espressa da numeri come 0 e 1, una variabile nominale non è numerica: la somma tra due caratteri non è definita, né il loro ordine in generale. Le sole operazioni sono quelle logiche (insiemistiche).

Una variabile numerica è sempre possibile trasformarla in variabile d'intervallo, ordinale o nominale (perdendo delle informazioni); una Variabile d'intervallo può essere trasformata in variabile ordinale o nominale; una variabile ordinale può essere trasformata in variabile nominale. L'inversa non è vera.

8.2 Scheda di lavoro sul Chi Quadro.

Il Prof. Nicalsi constata che 17 dei suoi allievi sono riusciti in un esercizio, mentre 10 hanno sbagliato. Uno dei suoi colleghi il Prof. Carozzi, dichiara che in generale la metà dei suoi allievi sbaglia su questo esercizio.

Il Prof. Nicalsi può pensare che i suoi allievi sono migliori di quelli del Prof. Carozzi?

Il campione è di 27 allievi del Prof.Nicalsi - *una variabile*: il risultato dell'esercizio, questa variabile può prendere due valori (Vero-Falso 1-0).

Esemplificazione: Distribuzione delle osservazioni

0 (falso)	**1 (vero)**
a	c
b	d
e	

Dove a, b, c, d, e, rappresentano gli allievi che hanno risposto correttamente o no alla questione.
[Le osservazioni sono gli elementi del campione]

0	**1**	**Σ**
n_1	n_2	n

Valori effettivi di V (Variabile)
n_1 e n_2 , $n_1+n_2=n$

Nel caso dell'esempio proposto:

0	**1**	**Σ**
10	17	27

8.2.1 Comparazione con una popolazione simile: Ipotesi Nulla.

Questa prova permette di comparare due distribuzioni, o di comparare la distribuzione osservata con una distribuzione simile o teorica dedotta da una ipotesi: per esempio la distribuzione uniforme di tutti i valori reali. L'ipotesi nulla corrispondente sarà:

"Ogni valore della variabile presenta la stessa possibilità di prendere tutti i valori".

Nel caso del Prof. Carozzi sono parecchie le ipotesi che si possono dedurre prenderemo in considerazione la più evidente:

"Il numero degli allievi che riescono sono eguali a quelli che sbagliano, questa legge è eguale per tutte le classi (fatto o modello) e la vostra non dovrà sfuggire a questa regola (Ipotesi)".

Infatti si tratta di sapere in questa ipotesi se questa classe, dove la riuscita è stata del 63% circa (17/63) segua o no (ipotesi contraria) la legge del prof.Carozzi. *L'ipotesi detta nulla associata a un modello è quella che afferma che la contingenza non si discosta molto dal modello*. Nel nostro caso: "La classe del prof. Nicalsi (A) segue la legge della classe del prof. Carozzi (B)". Chiameremo questa ipotesi H_0 . L'ipotesi contraria: "La contingenza si discosta significativamente dal modello".

Se l'ipotesi nulla è contraddetta, cioè a dire se essa ha delle conseguenze contrarie all'osservazione (dunque la contingenza non segue il modello), potremo pensare sia che la classe A non è una classe "nella norma", sia che essa è una classe nella norma ma che la legge del prof. Carozzi è falsa.

Modello: distribuzione teorica.

Nel nostro caso il modello teorico è il seguente:

0	**1**	**Σ**
13,5	13,5	27

Introduciamo adesso la distanza χ^2 , questa distanza è data dalla seguente formula:

$$\chi^2_{\text{osservato}} = \sum_i \frac{(E_T - E_O)^2}{E_T} \qquad [1]$$

Dove E_T= Effettivo teorico delle osservazioni di un valore della variabile (nel nostro esempio 13,5).

Dove E_O= Effettivo corrispondente osservato (nel nostro caso 10 e 17).

Nel nostro caso $\chi^2_{\text{o}} = \frac{(13{,}5-10)^2}{13{,}5} + \frac{(13{,}5-17)^2}{13{,}5} = \frac{2 \times 3{,}5^2}{13{,}5} = 1{,}81$

Se la distanza tra il modello e la contingenza è troppo grande rinunceremo all'ipotesi nulla come mezzo per spiegare la contingenza, si dirà che la si rigetta.

In caso contrario non potremo rigettata, ma non potremo dire che accettiamo l'ipotesi nulla.

Per stabilire che la distanza trovata è grande o piccola ci si riferirà ad una percentuale del 5%, cioè per sapere se un χ^2 è grande bisogna conoscere una distribuzione di numerosi χ^2 calcolati nelle stesse condizioni e sapere che il valore chiamato $\chi^2_{\text{soglia o,o5}}$ lascerà da una parte 95% di questi χ^2.

Nel caso del nostro esempio 1,81 è un valore grande per il χ^2 ?

Per saperlo bisogna considerare una popolazione di χ^2 che si otterrebbe nelle condizioni dell'ipotesi nulla:

"Si suppone che si dispone di un gran sacco contenente gettoni: metà rossi e metà neri. Si effettua un numero elevato di estrazioni con reimbussolamento (per esempio 100), di 27 estrazioni che rappresentano delle classi, come quella del Prof. N., ma che saranno presi in una popolazione rispondente all'ipotesi del Prof. C."

Ad ogni estrazione si ottiene un n_1 e un n_2 che non sono esattamente eguali a n/2. Si effettua il calcolo di χ^2 . La tabella di riferimento relativa al χ^2 si trova in appendice al capitolo e prendendo in considerazione la prima riga osserviamo che, nel nostro esempio, 70% (leggi 0.70) dei χ^2 saranno più grandi di 0,15, e dunque che il 30% sono più piccoli, o ancora che 50% sono più grandi di 0,46. E, nello stesso tempo, solamente il 20% di χ^2 sono più grandi di 1,64.

Se $\chi^2_{\text{soglia 0.05}} < \chi^2_{\text{osservato}}$, allora si rigetterà l'ipotesi nulla, si accetterà quindi che la distribuzione osservata si sarebbe potuta ottenere da un campionamento a caso che seguiva questa legge. Si dirà allora che il χ^2 è significativo, in caso contrario si rigetterà l'ipotesi e il χ^2 non è significativo.

Nel nostro esempio non si può rigettare l'ipotesi secondo la quale gli alunni del prof. Nicalsi appartengono a una popolazione che riesce una volta su due. Il prof. Nicalsi non può affermare che i suoi allievi sono migliori di quelli del prof. Carozzi.

Questo esempio consente di poter estendere l'analisi del χ^2 ad n variabili[1]:

Variabili	v_1	v_2	v_3	...	v_n	Σ
Valori effettivi	n_a	n_b	n_c	...	n_p	n
Distribuzione Teorica Ipotesi nulla	n/p	n/p	n/p	...	n/p	n

Dove si utilizza la formula [**1**].

Si possono studiare i seguenti casi:

- Confronto tra due campioni sperimentali con una variabile;
- Confronto tra più campioni sperimentali con n variabili. In questo caso la formula è

$$\chi^2_{\text{ossrvato}} = \sum_i \sum_j \frac{(E_T - E_O)^2}{E_T} \qquad [2]$$

8.3 Analisi implicativa tra variabili

Il problema che ha cercato di affrontare R.Gras[2] è stato quello di poter rispondere alla seguente questione: "Date delle variabili binarie **a** e **b**, in quale misura posso assicurare che in una popolazione, da ogni osservazione di **a** segue necessariamente quella di **b**?". O anche in maniera più lapidaria :"E' vero che se a allora b?".

In generale la risposta non è possibile ed il ricercatore si deve accontentare di una implicazione "quasi" vera. Con l'analisi implicativa di R. Gras si cerca di misurare il grado di validità di una proposizione implicativa tra variabili binarie e non. Questo strumento statistico viene messo a punto su ricerche riguardanti la Didattica delle Matematiche.

---

[1] Nell'esempio precedente dove la variabile era solatnto una, uno era il grado di libertà. Se le variabili sono due, ad esempio, in una tabella di p righe ed m colonne i gradi di libertà sono: (m-1)(p-1).

[2] Régis Gras, *L'implication statistique..., op. cit.*.

Viene presentata la modellizzazione del caso binario.

Siano date una popolazione E e un insieme di variabili V, si vuole dare significato statistico alla implicazione larga **a ⇒ b**.

Siano A e B gli insiemi delle sotto popolazioni rispettive dove la variabile a e b prendono il valore 1 (vero). L'intensità della implicazione viene espressa formalmente:

$$\varphi(a,\bar{b}) = 1 - \Pr ob\left[card(X \cap \bar{Y}) \le card(A \cap \bar{B})\right] \qquad \textbf{[3]}$$

X e Y sono due sotto insiemi di E, parti aleatorie di E e che hanno la stessa cardinalità rispettivamente di A e B. $\bar{Y}$ è il complementare di Y rispetto ad E. $\bar{B}$ è il complementare di B rispetto ad E. $\bar{b}$ rappresenta il fatto di non possedere il carattere b.

E si dirà:

$$\left[a \Rightarrow b \text{ accettabile alla soglia } \varphi(a,\bar{b}) = 1 - \alpha\right] \Leftrightarrow \Pr ob\left[\left[card(X \cap \bar{Y}\right] \le card(A \cap \bar{B})\right] \le \alpha$$

Dove:

$$n_a = cardA, n_b = cardB, n_{a \wedge \bar{b}} = card(A \cap B).$$

$$q(a,\bar{b}) = \frac{n_{a \wedge \bar{b}} - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} \qquad \textbf{[4]}$$

$q(a,\bar{b})$ rappresenta l'indice d'implicazione: indicatore della non implicazione di a su b.

L'intensità d'implicazione viene data dalla seguente formula:

$$\varphi(a,\bar{b}) = 1 - \Pr ob\left[Q(a,\bar{b}) \le q(a,\bar{b})\right] = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt \qquad \textbf{[5]}$$

$Q(a,\bar{b})$ è la variabile centrata e ridotta dedotta dalla variabile aleatoria $card(X,\bar{Y})$ che segue la legge di Poisson di parametro $n\pi = np(a)p(\bar{b}) = \frac{n_a n_{\bar{b}}}{n}$.

Supponiamo di avere le due variabili con i rispettivi valori ricavati da un questionario di tipo vero/falso:

	a	b
1.	1	1
2.	1	1
3.	0	1
4.	0	0
5.	1	1
6.	1	1
7.	0	1
8.	1	0
9.	0	0
10.	1	1
	6/ 10	7/ 10

Esempio di calcolo dell'indice di intensità di implicazione $a \Rightarrow b$:

$$n_a = 6$$

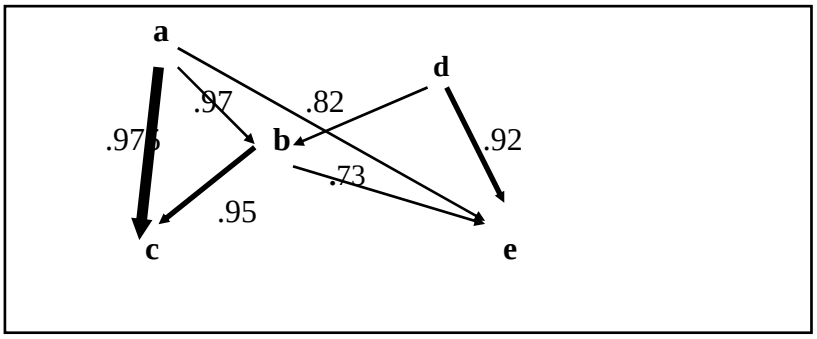
$$n_{\bar{b}} = 3$$

$$n_{a \wedge \bar{b}} = 1$$

$$q(a,\bar{b}) = \frac{1 - \frac{6 \bullet 3}{10}}{\sqrt{\frac{6 \bullet 3}{10}}} = -0{,}67$$

Se l'intensità d'implicazione[3] è abbastanza piccola le due variabili non saranno legate. Se φ(a,b)=0,75 è un livello di confidenza del 75% per l'implicazione in quanto la probabilità che card(X∩Y) sia più piccola di 1 è 0,25 (In generale se il livello di confidenza è inferiore al 95% non lo si ritiene accettabile). Il livello di confidenza va calcolato con la formula[4]. Ma naturalmente se i dati a disposizione sono numerosi sarà necessario l'uso del computer.

La rappresentazione di un grafo di relazione d'ordine parziale indotto dall'intensità di implicazione da la possibilità di visualizzare una situazione didattica dove intervengono più variabili. Lo spessore del braccio indica l'intensità che generalmente è indicata numericamente accanto come nella seguente figura:



[3] Per il calcolo dell'intesità di implicazione si può utilizzare la "*tavola della distribuzione normale*" acclusa a questo capitolo.

[4] Il valore di q deve risultare negativo affinché l'intensità di implicazione φ sia accettabile.
In particolare $q(a,\bar{b}) \le -1.64 \leftrightarrow \varphi(a,\bar{b}) \ge 0.95$.

Viene definita coesione implicativa quando per esempio date tre variabili **a**, **b**, **c** si osserva che $\varphi(a,\bar{b}) = 0.97$, $\varphi(b,\bar{c}) = 0.95$, $\varphi(a,\bar{c}) = 0.97$, allora la classe orientata da **a** verso **c** ammette una buona coesione. Non sarebbe stato lo stesso se avessimo avuto rispettivamente i valori 0.82, 0.38, 0.48.

Nella ipotesi che i valori di $a \Rightarrow b$ e $b \Rightarrow a$ risultano uguali le frecce del grafo non avranno la direzione. La nozione di implicazione statistica è stata estesa anche a variabili modali e variabili numeriche.

E' stato messo a punto dal gruppo I.R.MA.R.[5] un programma su PC che consente di fare l'analisi implicativa abbastanza celermente. Il programma si chiama *CHIC* (Classification Hiérachique Implicative et Cohésive) che permette differenti statistiche:

- statistiche elementari tipo media, varianza, correlazioni tra variabili;
- l'analisi delle similarità di Lerman[6];
- L'analisi implicativa secondo R. Gras, con le seguenti informazioni:

1. Grafo Implicativo;

[5]Istituto di ricerca Matematica di Rennes, Università di Rennes I, prof. Regis Gras. Il programma CHIC è disponibile al seguente indirizzo: Prof.R.Gras, 14 avenue de la Chaise, F-35170 Bruz, Francia. Informazioni si trovano sul sito dell'ARDM (Association Recherches Didactiques Mathematiques) http://ardm.eu/contenu/logiciel-danalyse-de-donn%C3%A9es-chic Il programma comprende le lingue: Francese, inglese, italiano, portoghese, slovacco.

[6]Questa operazione segue l'indice di similarità di Lermann e classifica le variabili secondo livelli gerarchici. L'indice di similarità di Lerman segue la legge di Poisson e viene così definito:

$$s(a,b) = \frac{n_{a\wedge b} - \frac{n_a n_b}{n}}{\sqrt{\frac{n_a n_b}{n}}}$$

ed é legato all' indice di implicazione dalla relazione :

$$\frac{q(a,\bar{b})}{s(a,b)} = -\sqrt{\frac{nb}{n\bar{b}}}$$

L'indice di similarità di Lermann e l'implicazione di Gras pur fornendoci informazioni sulle variabili nella stessa direzione a volte differiscono nel senso che si può avere similarità senza implicazione e viceversa.

Un esempio chiarisce bene il significato di indice di similarità di Lermann:

Esempio 1		**Esempio 2**		**Esempio 3**		**Esempio 4**	
0	0	1	1	0	0	0	1
1	0	1	1	0	0	1	0
1	1	1	1	0	0	0	1
1	1	1	1	0	0	1	0
1	1	1	1	0	0	0	1
-0,44721		0		0		-2,23607	

La similarità (esempi 2 e 3) é 0 quando le due distribuzioni sono identiche, negli altri casi sono negative.

2. Gerarchia implicativa e i Nodi Significativi dove si formano le classi della gerarchia[7];
3. Contribuzione degli individui nei cammini significativi del grafo e alle classi significative della gerarchia;
4. Comparazione tra il grafo implicativo ed il grafo inclusivo[8].

8.3.1 Analisi implicativa e analisi a-priori

Nell'analisi implicativa un problema importante che ci viene trasmesso dalla ricerca in didattica è quello relativo al rapporto tra l'analisi a-priori di una determinata situazione didattica e il rapporto con i dati sperimentali. Nell'analisi fattoriale questo rapporto è controllato dall'utilizzo delle variabili supplementari e Individui supplementari. Le variabili supplementari e gli Individui supplementari hanno lo scopo di costruire un modello statistico dell'analisi a-priori.

Nell'analisi implicativa non ci si era posti il problema sino a questo momento. Attraverso colloqui con Regis Gras ho pensato di introdurre una distanza tra una implicazione a-priori così fatta:

---

[7]La gerarchia implicativa delle classi ci fornisce delle informazioni sulla implicazione tra classi di variabili. Per poter costruire un grafo per la gerarchia implicativa delle calssi è necessario introdurre il concetto di coesione implicativa:

$\varphi(a,\bar{b}) = \Pr ob(Q(a,\bar{b}) > q(a,\bar{b})$

L'entropia dell'evento $(Q(a,\bar{b}) > q(a,\bar{b})$ sara E:

$E = -pLog_2(p) - (1-p)Log_2(1-p)$

$p = \varphi(a,\bar{b})$

E é massima quando p = 0.5

E é minima quando p = 1 o p = 0 $(0Log_2(0) = 0)$

Coesione tra a e b, $c(a,b) = \sqrt{1-E^2}$ se $p \geq 0.5$

in caso contrario $c(a,b) = 0$.

L'implicazione si fa per aggregazioni successive di classi d'implicazione. Il principio è quello di riunire ad ogni passo di aggregazione la coppia di variabili o la coppia di classi di variabili che presentano la massima coesione nella tappa considerata.

L'informazione che se ne ricava fornisce un utile strumento per stabilire quali classi di variabili implicano altre classi di varibili ed a quale livello.

[8] I dati sono registrati in files con estensione **.csv**. Il programma CHIC leggerà sia i dati registrati con estensione .csv che quelli registrati dal foglio lettronico EXCEL. Qualunque estensione excel va sempre riconvertita in files **.csv**.

$$Q(a,\bar{b}) = \frac{0 - \frac{n_a n_{\bar{b}}}{n}}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = -\sqrt{\frac{n_a n_{\bar{b}}}{n}} \qquad \textbf{[6]}$$

dove nella formula **4** si è posto $n_{a\wedge\bar{b}} = 0$, cioè si è considerato il caso a-priori che a→b (nella rappresentazione insiemistica è il caso della inclusione completa). A questo indice Q corrisponde una intensità di implicazione calcolata con la formula **5** ed indicata con Φ.

Siamo adesso nelle condizioni di poter definire una distanza tra questa implicazione teorica a-priori e l'implicazione sperimentale, dove $n_{a\wedge\bar{b}}$ non é necessariamente uguale a 0.

Tale distanza verrà così introdotta:

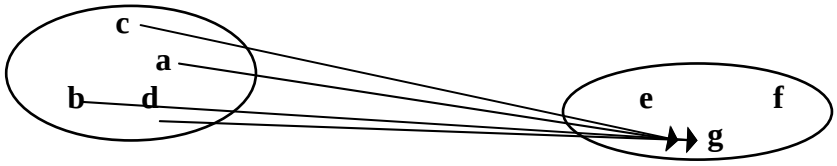
$$\Delta = \sum_{i,j} \frac{1}{\Phi(a_i,\bar{b}_j)} \left[\Phi(a_i,\bar{b}_j) - \varphi(a_i,\bar{b}_j)\right]^2 \qquad \textbf{[7]}$$

dove il termine $\frac{1}{\Phi(a_i,\bar{b}_j)}$ normalizza la distanza.

Quando cioè il valore di questa distanza si avvicina a 0 allora la relazione tra la nostra analisi a-priori e la contingenza non si discostano. Si può sostenere in base ad analisi sperimentali che un $0 \le \Delta \le 0.25$ possa essere considerato accettabile. Il caso limite negativo è 1.

Ma in una analisi a-priori non abbiamo solo il caso di semplici implicazioni, bisognerà quindi estendere la formula **8** al caso di classi di variabili che implicano classi di variabili.

Il primo passo sarà quello di estendere la formula al caso in cui **n** variabili implicano una sola variabile. Ed in questo caso basterà calcolare le singole implicazioni delle variabili cioè i Δ_i e poi fare la media aritmetica delle distanze Δ_i. Possiamo visualizzare questo procedimento:



Ma se dobbiamo prendere in considerazione l'implicazione tra le variabili **a**, **b**, **c**, **d** ed **e**, **f**, **g** dobbiamo allora considerare le distanze $\Delta_{(a,g)}$, $\Delta_{(b,g)}$, $\Delta_{(c,g)}$, $\Delta_{(d,g)}$ e fare una media tra queste 4 distanze, fare la stessa cosa con le variabili e ed f . La formula seguente sintetizza il procedimento:

$$\Delta_{(a,b,c,d),(e,f,g)} = \frac{1}{3}\left[\frac{\Delta_{(a,g)}+\Delta_{(b,g)}+\Delta_{(c,g)}+\Delta_{(d,g)}}{4}+\frac{\Delta_{(a,e)}+\Delta_{(b,e)}+\Delta_{(c,e)}+\Delta_{(d,e)}}{4}+\frac{\Delta_{(a,f)}+\Delta_{(b,f)}+\Delta_{(c,f)}+\Delta_{(d,f)}}{4}\right]$$

[8]

In questo paragrafo si cercherà di portare un esempio applicativo del rapporto tra analisi a-priori e contingenza. Riportiamo i dati relativi ad un precedente lavoro di ricerca didattica riguardante gli Ostacoli Epistemologici[9].

Il modello teorico-sperimentale per l'individuazione degli ostacoli epistemologici è stato analizzato sia attraverso l'analisi fattoriale che l'analisi implicativa. Mentre per l'analisi fattoriale è possibile mettere a punto una analisi a-priori attraverso l'introduzione di variabili supplementari e/o di individui supplementari che mettano in evidenza le caratteristiche fondamentali dell'analisi a-priori, nell'analisi implicativa questo non è stato possibile fare nel lavoro già citato.

In particolare, per quanto riguarda l'analisi fattoriale, la tecnica utilizzata è stata quella di inserire un individuo supplementare che avesse il "profilo" dell'allievo che ha come "ostacolo epistemologico" il Postulato di Eudosso-Archimede (P.E.-A.). In questa sede non ci addentreremo sulle questioni riguardanti il Postulato[10] ma affronteremo le questioni relative al modello statistico per l'interpretazione di una eventuale analisi a-priori.

Per una migliore comprensione del testo introduciamo una tabella riguardante una sintesi dell'analisi a priori relativa ai questionari dell'esperienza finale (Spagnolo, 1995).

[9]F.Spagnolo (1995), Obstacles épistémologiques: *Le Postulat d'Eudoxe-Archimede*, Tesi di Dottorato, Quaderni di Ricerca in Didattica, Supplemento al n.5, Palermo.

[10]Per questo rimandiamo al già citato lavoro.

Q1	Conoscenza del Postulato P.E.-A. in termini operazionali. Bisogna determinare un n tale che il multiplo del segmento **na>b**. (**a<b**). Formulazione diretta del Postulato. La rappresentazione con dei trattini da un legame con la misura. Risposte attese: n>4.
Q2	Questione simile alla precedente, ma il segmento **b** è molto più grande e il segmento **a** è stato disegnato più piccolo. Risposte attese: n>19. Formulazione diretta del Postulato.
Q3	Risponde affermativamente all'esistenza di n / n**a**>**b**. Formulazione diretta del Postulato.
Q4	Dare una giustificazione alla risposta data nella questione precedente. Formulazione diretta del Postulato.
Q5	Conoscenza del P.E.-A. in termini operazionali. Bisogna determinare un n tale che (1/n)**a**<**b**). Risposte attese: n>3. Formulazione inversa del Postulato.
Q6	Risponde affermativamente all'esistenza di un n / (1/n)**a**<**b**. Formulazione inversa del Postulato
Q7	Formulazione linguistica differente della questione precedente: "...è sempre possibile...".**.**
Q8	Cambiamento del punto di vista: Modello del "Veronese", non Archimedeo, in geometria. Risposta attesa: Affermare la non validità del Postulato.
Q9	Cambiamento del punto di vista, l'allievo deve seguire la costruzione evocata nella proposizione X,I[11] di Euclide e concludere con la sua validità (angoli rettilinei).
Q10	Cambiamento del punto di vista, validità della proposizione X,I (angoli curvilinei). La questione 17 indica come confrontare degli angoli curvilinei tra di loro. L'allievo deve effettuare una costruzione e concludere sulla validità del P.E.-A..
Q11	Conteso generalizzato (comparazione di angoli curvilinei e rettilinei). L'allievo deve, in questo caso, rigettare la validità della proposizione X,I.
Q12	Resistere alle contraddizioni con la contingenza: l'allievo da una giustificazione della X,I (un contesto non Archimedeo.
Q13	Per riuscire, l'allievo da un argomentazione per rigettare il procedimento della X,I in un contesto non Archimedeo.
Q14	Conferma: l'ostacolo persiste: Affermare la validità della proposizione X,I per un allievo, è mostrare che esce a provare il suo modello interpretativo in un contesto più generale.
Q15	Conferma della posizione del quesito Q13: Affermare la non validità della proposizione X,1.
Q16	Gli allievi devono cercare una relazione d'ordine tra tre triangoli.
Q17	Relazione d'ordine tra tre triangoli (altri contesti).
Q18	Relazione d'ordine (R.O.) tra tre triangoli (altri contesti).
Q19	R.O. tra angoli rettilinei.
Q20	R.O. tra angoli curvilinei (parabole).
Q21	R.O. tra angoli curvilinei (contingenza o contatto).
Q22	R.O. inclusione tra angoli rettilinei.
Q23	R.O. inclusione tra angoli di contingenza.
Q24	R.O. inclusione tra angoli curvilinei e di contingenza.

In una analisi a-priori riguardante l'ostacolo epistemologico possiamo individuare le seguenti implicazioni tra gruppi di variabili.

[11] Proposizione X. I: *Date due grandezze disuguali, se si sottrae dalla maggiore una grandezza maggiore della metà, dalla parte restante un'altra grandezza maggiore della metà, e così si procede successivamente rimarrà una grandezza che sarà minore della grandezza minore inizialmente assunta.*

Gli elementi ritenuti significativi nel Modello riguardante l'ostacolo epistemologico sono sostanzialmente:

- la resistenza;
- la persistenza;
- il cambiamento del punto di vista;
- la generalizzazione.

In questa sede esamineremo soltanto le seguenti implicazioni:

1. Se l'ostacolo resiste e persiste allora si avrà un cambiamento del punto di vista [Se (Q_{12},Q_{13}, Q_{14}, Q_{15}) allora (Q_8,Q_9,Q_{10})].
2. Se l'ostacolo resiste e persiste allora si avrà una generalizzazione [Se (Q_{12} , Q_{13} , Q_{14} , Q_{15}) allora (Q_{11})].

Gli allievi che avranno un modello di ostacolo che Resiste e Persiste allora saranno anche in grado di cambiare il punto di vista. O anche tutti gli allievi che hanno risposto alle questioni riguardanti la resistenza e persistenza dell'ostacolo hanno risposto al cambiamento del punto di vista.

La formula **[9]** la possiamo anche scrivere:

$$Q(a,\bar{b}) = -\sqrt{\frac{n_a n_{\bar{b}}}{n}} \qquad \textbf{[9]}$$

Il campione osservato constava di 107 allievi del primo anno di corso di laurea in matematica (anno accademico 1994/95).

I questionari sono stati somministrati in un'unica soluzione in un tempo massimo di 2 ore. L'unico supporto a disposizione, oltre ai fogli del questionario, un foglio di carta lucida ottenibile a richiesta.

La tabella a-priori delle implicazioni è la seguente:

$\Phi_{12,8}$=0.99[12]	$\Phi_{12,9}$=0.99	$\Phi_{12,10}$=1
$\Phi_{13,8}$=0.75	$\Phi_{13,9}$=0.79	$\Phi_{13,10}$=0.83
Φ_{1248}=0.99	$\Phi_{14,9}$=0.91	$\Phi_{14,10}$=0.99
$\Phi_{15,8}$=1	$\Phi_{15,9}$=1	$\Phi_{15,10}$=1

La tabella relativa alla contingenza è la seguente:

[12]In questo caso particolare $n_{12} = 14,\ n_{\bar{8}} = 53,\ q_{12,8} = -0.70$.

$\varphi_{12,8}=0.98$	$\varphi_{12,9}=0.88$	$\varphi_{12,10}=0.65$
$\varphi_{13,8}=0.75$	$\varphi_{13,9}=0.79$	$\varphi_{13,10}=0$
$\varphi_{1248}=0.66$	$\varphi_{14,9}=0.99$	$\varphi_{14,10}=0.65$
$\varphi_{15,8}=0.79$	$\varphi_{15,9}=0.82$	$\varphi_{15,10}=0.67$

La relativa distanza è così calcolata:

$$\Delta_{(12,13,14,15),(8,9,10)} = \frac{1}{3}\left\{\left[\frac{\frac{1}{0.99}(0.99-0.98)^2+\frac{1}{0.75}(0.75-0.75)^2+\frac{1}{0.99}(0.99-0.66)^2+(1-0.79)^2}{4}\right]+\left[\frac{\frac{1}{0.99}(0.99-0.88)^2+\frac{1}{0.79}(0.79-0.79)^2+\frac{1}{0.99}(0.99-0.99)^2+(1-0.82)^2}{4}\right]+\left[\frac{(1-0.65)^2+\frac{1}{0.83}(0.83-0)^2+\frac{1}{0.99}(0.99-0.65)^2+(1-0.67)^2}{4}\right]\right\} = 0.116$$

Il valore di questa distanza 0.116 è compreso tra 0 e 0.25 che risulta essere accettabile, quindi l'analisi a-priori riguardante l'implicazione tra la resistenza e la persistenza dell'ostacolo viene confermata dalla contingenza.

La distanza riguardante l'implicazione tra la resistenza e persistenza di un ostacolo e la generalizzazione del contesto ha il seguente valore[13]:

$$\Delta_{(12,13,14,15),(11)=0.319}$$

che risulta di poco maggiore del valore indicato da considerazioni teoriche e sperimentali

Si potrebbero quindi fare delle considerazioni sul modello riguardante l'ostacolo epistemologico:

- Il contesto riguardante la generalizzazione è giocato tutto su una sola questione e forse andrebbero aggiunte delle nuove questioni per puntualizzare meglio questo elemento importante del modello;
- Il contesto riguardante la generalizzazione mette in discussione il modello?
- Questa distanza introdotta è significativa?

[13]Non vengono riportati i calcoli riguardanti questa distanza.

8.4 Alcune osservazioni sull'analisi Fattoriale.

Consideriamo il prodotto cartesiano E (in generale allievi, n, $n \in N$) e V (in generale variabile $m \in N$). Questa è una tipica situazione di rilevazione di dati in didattica. Si pone il problema di poter rappresentare geometricamente in uno spazio ad nxm dimensioni la distribuzione dei due insiemi. L'analisi fattoriale interpreta le rappresentazioni geometriche. Sorta nell'ambito delle Scienze Umane ha avuto parecchie applicazioni nel campo della Psicologia, ma consentendo una analisi su piccoli campioni nel campo della Statistica non parametrica contribuisce ad interpretare significativamente i fenomeni didattici. In questi ultimi dieci anni sono state condotte una serie di ricerche per cercare di affinare lo strumento nel campo delle ricerche didattiche, ma soprattutto di creare dei modelli ad hoc.

In questo paragrafo introdurremo il metodo senza entrare molto nel particolare in quanto richiederebbe una trattazione a parte.

Il punto di partenza è una tabella a doppia entrata I x J del tipo:

J / **I**	**j**	Colonna di margine
i	**k(i,j)**	**k(i)** [totale della linea i]
Linea di margine	**k(j)** [totale della colonna j]	**k** [totale]

- $\mathbf{f_i = k(i)/k}$ **massa della linea i;**
- $\mathbf{f_j = k(j)/k}$ **massa della colonna j.**

La massa di un elemento i o j misura l'importanza relativa di questo elemento.

- $\mathbf{f_j^i} = \{ \mathbf{f_j^i} \ / \ \mathbf{j \in J} \}$**,** insieme degli f_j^i per j che percorre J. **Profilo di una linea.** $\mathbf{f_j^i = k(i,j)/k(i)}$ è la parte relativa, la porzione di j nella i^{esima} linea, $\mathbf{f_j^i}$ insieme di tutti i termini corrispondenti ai diversi elementi di J, è la composizione di questa linea.
- $\mathbf{f_i^j} = \{ \mathbf{f_i^j} \ / \ \mathbf{i \in I} \}$**, Profilo di una colonna.**

Spazio dei profili su J: Un punto di questo spazio è un profilo su J, cioè un insieme di numeri positivi o nulli indicizzati da J e di totale 1.

$\pi_j = \{\pi_j \ / \ j \in J\}, \pi_j \in R^+_0 , \ \sum\{\pi_j \ / \ j \in J\} = 1.$

Ad ogni elemento j di J corrisponde un termine π_j e uno solo del profilo π_j ; ad ogni termine π_j di π_j corrisponde un elemento e uno solo di J: il suo indice j; un profilo π_j è composto da parametri che variano su J con cardinalità **Card J**, i parametri sono legati dalla relazione d'avere come somma 1. Lo spazio dei profili su J è uno spazio a (**Card J** -1) dimensioni.

Spazio dei profili su I:

$\pi_i = \{\pi_i \ / \ i \in I\}, \pi_i \in R^+_0 , \ \sum\{\pi_i \ / \ i \in I\} = 1.$

Lo spazio dei profili su I è uno spazio a (**Card I** -1).

8.4.1 Le rappresentazioni grafiche: le nuvole.

La nuvola N(I):

Nello spazio dei profili su J, ogni linea i della tabella è rappresentata dal suo profilo:

$f_j^i = \{ k(i,j) / k(i) / \ j \in J \},$

al quale si associa la massa di $i : f_i = k(i)/k$; l'insieme dei profili delle diverse linee i, ciascuna munita della massa della linea che rappresenta, costituisce la nuvola N(I):

$N(I) = \{ f_j^i, f_i / i \in I \};$ un elemento della nuvola N(I) è una coppia formata da un profilo di linea e della massa di questa linea.

La nuvola N(J):

Nello spazio dei profili su I, ogni linea i della tabella è rappresentata dal suo profilo:

$f_i^j = \{ k(j,i) / k(j) / \ i \in I \},$

al quale si associa la massa di $j : f_j = k(j)/k$; l'insieme dei profili delle diverse linee i, ciascuna munita della massa della linea che rappresenta, costituisce la nuvola N(J):

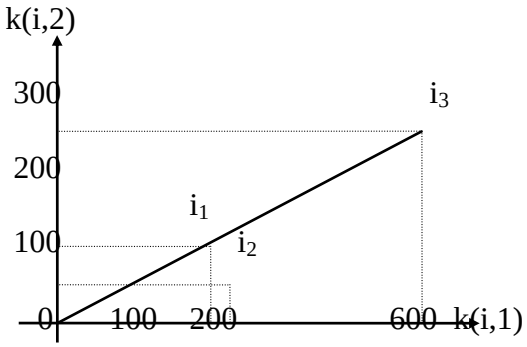
$N(J) = \{ f_i^j, f_j / j \in J \};$ un elemento della nuvola N(J) è una coppia formata da un profilo di linea e della massa di questa linea.

La rappresentazione grafica viene fatta nello spazio dei profili. Per meglio comprendere questa situazione consideriamo il seguente esempio concreto:

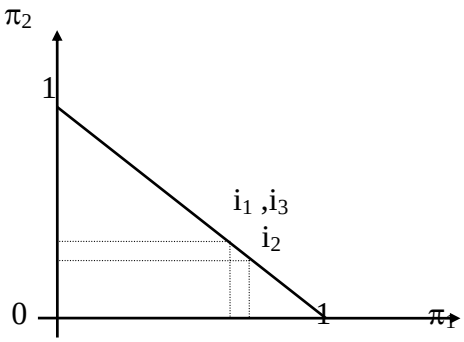
	1	2	Margine
i_1	200	110	310
i_2	210	80	290
i_3	600	330	930

$f_j^{i1} = \{200/310, 110/310\} = \{0,65; 0,35\}$
$f_j^{i2} = \{210/290, 80/290\} = \{0,72; 0,28\}$
$f_j^{i3} = \{600/930, 330/930\} = \{0,65; 0,35\}$

Si possono fare due rappresentazioni differenti, una riguardante il grafico delle linee grezze (a sinistra) e dei profili (a destra):



Spazio delle linee grezze
ascissa di i : k(i,1); ordinata di i : K(i,2)



Spazio dei profili
ascissa di i= f_1^i; ordinata di i=f_2^i

Nello spazio delle linee grezze, i_1 e i_2 sono vicini, mentre i_3 è distante. Ma nello spazio dei profili, i_1 e i_3 coincidono mentre i_2 si distingue. In generale i_1 e i_3 hanno linee proporzionali nella tabella k(i,j) e quindi l'allineamento nel primo grafico, ma hanno lo stesso profilo e quindi sarà situato (nel secondo grafico) nella stessa retta ed avrà il suo profilo confuso con i due precedenti.

8.4.2 Media e centro di gravità; Dispersione e inerzia.

Il centro di gravità di un sistema di punti muniti di massa (numeri positivi o nulli) è una generalizzazione spaziale della nozione di media: il centro di gravità dei punti f_j^i con una massa f_i è come la media della Nuvola N(I), ma una media dove ogni punto f_j^i gioca un ruolo proporzionale alla sua massa f_i.

Per determinare il centro di gravità o baricentro utilizzeremo la nozione di media ponderata[14] e applicheremo questa nozione al nostro caso delle nuvole N(I) e N(J). In questo caso il centro di gravità è una sorta di media spaziale, ogni punto gioca un ruolo proporzionale alla sua massa prendendo la media ponderata delle coordinate asse per asse.

[14] La media ponderata tra due nuemri x_1 e x_2 di massa rispettivamente m_1 e m_2 sarà definita da: $(m_1x_1+m_2x_2)/(m_1+m_2)$.

Per N(I), lo spazio ambiente è lo spazio dei profili su J; i punti di N(I) sono i profili su J, ciascuno munito di massa f_i. Il centro di gravità di N(I) è un profilo su J che chiameremo g_j, la sua j^{esima} coordinata g_j è la media ponderata della j^{esima} coordinata di punti di N(I):

$$g_j = \frac{\sum\left\{ f_i f_j^i / i \in I \right\}}{\sum\left\{ f_i / i \in I \right\}}$$

Ma siccome la somma degli f_i vale 1, la formula si può anche scrivere:

$$g_j = \sum\left\{ f_i f_j^i / i \in I \right\}$$

e rimpiazzando in questa espressione f_i e f_j^i con le espressioni in funzione di k(i,J), k(i), K(j) e k avremo:

$$g_j = f_j = \left\{ \frac{k(j)}{k} / j \in J \right\}$$

Analogamente avremo nello spazio dei profili di I la seguente formula:

$$g_I = f_I = \left\{ \frac{k(i)}{k} / i \in I \right\}$$

Dispersione della Nuvola e Varianza:

Una nuvola N(I) sarà più o meno dispersa attorno al suo centro di gravità f_j e la sua dispersione può essere calcolata prendendo le medie degli scarti dei diversi punti rispetto al centro di gravità. Verrà definito lo scarto di una nuvola N in rapporto ad un punto P qualunque con la somma $I_P(N)$:

Si consideri una nuvola N di n punti M^i con massa rispettivamente m_i : N{(M^i, m_i) / i=1,...,n}e P un punto qualunque situato nello stesso spazio di N.

$I_P(M^i,m_i)=m_i(d(P,M^i))^2=m_i d^2(P,M^i)$ Inerzia del punto (M^i,m_i) rispetto a P, dove $d^2(P,M^i)$ è la distanza tra P e M^i elevata al quadrato.

$I_P(N)=\sum\{m_i d^2(P,M^i) / i= 1,...,n\}$.

e rispetto al suo centro di gravità:

$$\mathbf{I_G(N)=\sum\{m_i d^2(G,M^i) \,/\, i= 1,...,n\}.}$$

Le inerzie di N rispetto a G e rispetto ad un punto P qualunque sono legate dalla seguente formula dovuta a Huyghens:

$$\mathbf{I_P(N)= I_G(N) + m_{tot} d^2(G,P)} \qquad (\mathbf{m_{tot}=m_1+m_2+...+m_n} \text{ massa totale di N}) \quad \mathbf{[5]}$$

Cioè l'inerzia della nuvola N rispetto ad un punto qualunque P è eguale all'inerzia della nuvola N rispetto al suo centro di gravità, aumentato della quantità positiva $m_{tot}d^2(G,P)$ *prodotto della massa totale per il quadrato della distanza tra G e P.*

8.4.3 Distanza distribuzionale.

Si definisce una distanza distribuzionale tra profili introducendo un principio di equivalenza distribuzionale.

Consideriamo il seguente esempio. Supponiamo che vi siano una lista di nomi I ed una lista di verbi J. Per ogni coppia (i,j) di un nome i e di un verbo j si calcola quante volte il nome i è soggetto del verbo j. Nella tabella dei dati, ad ogni nome i di I corrisponde una linea che chiameremo anche i, e ad ogni verbo j di J corrisponde una colonna j. All'incrocio della linea i e della colonna j è scritto il numero di volte, k(i,j), che il nome i è stato trovato soggetto del verbo j.

Che significato dobbiamo attribuire a due verbi che hanno lo stesso profilo ?

Essi hanno gli stessi soggetti con le stesse frequenze relative. I due verbi j e j' li chiameremo sinonimi distribuzionali.

In N(J) sono rappresentati dai punti f_I^j e $f_I^{j'}$ che coincidono rispettivamente con le masse f_j e $f_{j'}$. Nell'analisi è come se si considerasse un solo verbo j_0 che ingloba gli usi di j e j' con massa $f_{j0}=f+f_{j'}$.

Cosa avviene alla nuvola N(I) quando la tabella è modificata cumulando le due colonne in una sola?

La formula della distanza sarà allora del tipo: $\alpha_j=1/f_j$.

La formula della distanza distribuzionale tra due linee nello spazio dei profili su J (N(I)):

$$d^2(f_J^i, f_J^{i'}) = \sum\left\{\left(\frac{1}{f_j}\right)\left(f_j^i - f_j^{i'}\right)^2 / \; j \in J\right\} \qquad \mathbf{[6]}$$

La formula della distanza distribuzionale tra due colonne nello spazio dei profili su I (N(J)):

$$d^2\left(f_I^{\,j}, f_I^{\,j'}\right) = \sum\left\{\left(\frac{1}{f_i}\right)\left(f_i^{\,j} f_i^{\,j'}\right)^2 / \, i \in I\right\} \qquad \textbf{[7]}$$

8.4.4 Gli assi principali d'inerzia.

La ricerca degli assi principali d'inerzia di una nuvola di punti muniti di massa in uno spazio Euclideo è pregiudiziale per qualsiasi tipo di analisi Fattoriale. Oggi con il termine "Analisi Fattoriale" si intende l'Analisi delle Componenti Principali (ACP): ricerca degli assi principali d'inerzia. In questa trattazione seguiremo questa impostazione.

L'Analisi Fattoriale delle Corrispondenze (AFC) seguirà l'analisi delle componenti principali.

Dal punto di vista dell'analisi dei dati multidimensionale è importante ridurre la nuvola N ad una rappresentazione accessibile alla nostra visione (cioè di dimensioni accettabili ottenuta per proiezione su una retta o su un piano) e fedele alla complessità del reale (se la dispersione della nuvola proiettata è quasi uguale a quella di N stesso). E' per questo che bisogna considerare non solamente l'inerzia della nuvola N rispetto ad un punto ma anche l'inerzia trasversalmente ad un sotto spazio L (retta o piano), misura dello scarto della nuvola a L, e, correlativamente, l'inerzia lungo un sotto spazio L, misura della fedeltà della rappresentazione della nuvola dalla sua proiezione ortogonale su L. Il teorema di Huyghens (vedi formula **[5]**) gioca un ruolo fondamentale in quanto permette di restringere lo studio dell'inerzia ai sotto spazi passanti per il centro di gravità G di una nuvola N. In particolare alle rette passanti per G o ai piani passanti per G.

Le coordinate f_j^i sugli assi sono i fattori che si studieranno: l'Analisi Fattoriale si occupa del cambiamento delle coordinate e del cambiamento degli assi.

Generalmente per ottenere la massima percentuale d'inerzia totale è sufficiente considerare i primi 4 assi d'inerzia[15]:

- Δ_1 : Asse principale d'inerzia della nuvola N(I);
- Δ_2 : E' tra le rette perpendicolari a Δ_1, quella sulla quale la nuvola si proietta con la più grande dispersione (lungo la quale l'inerzia della nuvola è più grande);

[15] Teoricamente il numero degli assi fattoriali della nuvola N(I) non può superare il più piccolo dei due numeri (Card J - 1) e (Card I - 1).

- Δ_3 : E' tra le rette perpendicolari a Δ_1 e Δ_2, quella sulla quale la nuvola si proietta con la più grande dispersione;
- ecc..

Δ_α : Insieme di rette a due a due perpendicolari è utilizzato come sistema d'assi di coordinate che chiameremo fattori.

L'indipendenza dei fattori è assicurata dall'indipendenza degli assi ortogonali tra loro.

L'inerzia della nuvola lungo l'asse α è il valore λ_α relativo a questo asse ed è definita:

$\lambda_\alpha = \sum\{f_i \ F_\alpha^2 (i) / i \in I\}$, dove f_i = massa del punto f_j^i , e $F_\alpha^2(i)$ = quadrato della distanza da

f_j alla proiezione ortogonale di f_j^i sull'asse α.

La media ponderata del quadrato di una funzione centrata a zero è chiamata varianza.

Per covarianza si intende la media del prodotto di due funzioni centrate (di media nulla) $F_\alpha F_\beta$.

Due fattori distinti hanno covarianza nulla e quindi non sono correlati od anche le funzioni sono indipendenti (condizione necessaria).

Per fare l'analisi su N(I) o N(J) si seguono due processi analoghi e questi due processi conducono agli stessi risultati per due vie simmetriche.

8.5 Analisi delle Componenti Principali (ACP).

Consiste nel ricercare i piani principali determinati dagli assi d'inerzia ed il centro del centro di gravità della nuvola per rappresentare le migliori proiezioni.

Permette anche di precisare in quale misura le variabili sono correlate o legate tra loro.

Nell'interpretazione statistica gli assi principali saranno chiamati "fattori principali".

Nella ACP le correlazioni tra i soggetti sono più delicati da interpretare. Le variabili ed i soggetti vengono rappresentati in spazi differenti (Vedi paragrafi 8.7.2-8.7.3-8.7.4).

8.6 Analisi Fattoriale delle corrispondenze (AFC).

Segue lo stesso principio della ACP ma la distanza utilizzata è quella di X^2 (leggermente diversa da quella utilizzata nella ACP) e che permette di meglio trattare il caso di una matrice d'incidenza (variabili booleane).

La distanza utilizzata è la seguente:

$$d^2(V_i, V_j) = \sum_k \frac{S}{S_k}\left(\left|\frac{S_{ki}}{S_i} - \frac{S_{kj}}{S_j}\right|\right)^2 \qquad \textbf{[8]}$$

Dove S_k è la somma della linea k (valori corrispondenti al soggetto k), S_i e S_j le somme delle colonne corrispondenti alle variabili V_i e V_j, S_{ki} e S_{kj} i valori delle variabili V_i e V_j osservate per il soggetto k e S la somma dei valori della tabella di riferimento.

Inoltre i soggetti e le variabili possono essere messi nello stesso spazio. E' possibile allora interpretare le prossimità dei soggetti tra loro, quelle delle variabili tra loro, e quelle dei soggetti con le variabili.

Il "significato" da attribuire ai fattori è tutto a carico del ricercatore, il quale deve interpretare le informazioni che sono più nascoste e che discendono da questo significato. Ci si interesserà ai contributi di certi punti a questi fattori e alle posizioni relative dei sottogruppi di popolazione studiata. Risulta molto interessante l'introdurre delle variabili supplementari e/o dei soggetti supplementari.

Le analisi ACP e AFC sono implementate per PC da un programma che si chiama SPSS[16]. Introdotti i dati il programma consente l'uscita dei dati sia su schermo che su stampante.

[16]Il programma SPSS può essere richiesto al seguente indirizzo web:. http://www.spss.com/ e http://www.spss.it/

8.7 Un esempio con l'uso del computer: il programma SPSS.

Viene riportato un esempio di analisi statistica eseguita con il programma SPSS. I dati presentati sono irreali e pertanto non si potrà intervenire nel commento con considerazioni più approfondite, si cercherà di interpretare i dati con considerazioni di carattere generale.

In una situazione sperimentale giocano un ruolo importante l'introduzione di variabili supplementari o l'introduzione di profili di individui supplementari, secondo la natura del problema. In questo esempio non sono state introdotte né variabili supplementari né individui supplementari.

8.7.1 Una simulazione al computer.

Consideriamo la seguente situazione con 12 osservazioni (Allievi ad esempio) e 10 Variabili:

Tabella di contingenza										
	f1	**f2**	**f3**	**f4**	**f5**	**f6**	**f7**	**f8**	**f9**	**f10**
A1	1	0	1	0	0	1	1	0	1	0
A2	1	0	1	0	0	0	0	0	1	0
A3	1	1	1	1	1	1	1	1	1	1
A4	1	1	1	1	1	1	0	1	1	1
A5	1	1	1	1	1	1	0	1	1	1
A6	1	1	1	1	1	1	0	1	0	1
A7	0	1	0	1	0	1	0	0	0	0
A8	0	0	0	1	0	1	0	0	0	0
A9	0	0	0	0	1	1	0	1	0	1
A10	1	0	1	0	1	1	0	1	1	1
A11	0	1	0	1	0	1	0	0	1	0
A12	1	1	1	1	0	1	1	0	1	0

Fig. 1

Questa è una tabella logica, disgiunta, completa:

1. logica: I casi della tabella comprendono soltanto 0 e 1 rappresentanti di NO e SI;
2. Disgiunta: La modalità di 0 e 1 di una stessa questione si escludono vicendevolmente; Se la linea i contiene il numero 1 in una delle due colonne essa contiene necessariamente 0 nell'altra;
3. Completa: Ogni soggetto risponde effettivamente ad ogni quesito; la linea 1 contiene necessariamente il numero 1 in ogni gruppo di colonne afferenti alla stessa questione.

Inizieremo con l'Analisi delle Componenti Principali (ACP). Il numero degli Assi presi in considerazione sono **5**.

Statistiche Elementari:

Variabili	Medie	Scarti
f1	0.667	0.4714
f2	0.583	0.4930

f3	0.667	0.4714
f4	0.667	0.4714
f5	0.500	0.5000
f6	0.917	0.2764
f7	0.250	0.4330
f8	0.500	0.5000
f9	0.667	0.4714
f10	0.500	0.5000

Fig. 2

Correlazioni[17]:

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
f1	1.000									
f2	0.120	1.000								
f3	1.000	0.120	1.000							
f4	-0.125	0.837	-0.125	1.000						
f5	0.354	0.169	0.354	-0.000	1.000					
f6	-0.213	0.357	-0.213	0.426	0.302	1.000				
f7	0.408	0.098	0.408	-0.000	-0.192	0.174	1.000			
f8	0.354	0.169	0.354	-0.000	1.000	0.302	-0.192	1.000		
f9	0.625	0.120	0.625	-0.125	-0.000	-0.213	0.408	-0.000	1.000	
f10	0.354	0.169	0.354	-0.000	1.000	0.302	-0.192	1.000	-0.000	1.000

Fig. 3

8.7.2 L'analisi delle componenti principali

Varianza totale spiegata per 2 fattori

Autovalori iniziali(a)			Pesi dei fattori non ruotati			Pesi dei fattori ruotati		
Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata

[17] Ricordiamo che:

- il coefficiente di correlazione lineare si può esprimere in vari modi. Uno di essi è: $(\rho=cov(xy) / \rho_x\rho_y)$;
- il coefficiente di correlazione lineare ρ è compreso tra -1 e +1 $(-1\leq\rho\leq+1)$;
- il coefficiente di correlazione lineare ρ è uguale a ±1 solamente quando tutti dati sono allineati lungo una retta crescente o decrescente;
- Se si effettua sulle variabili x e y una trasformazione lineare, il coefficiente di correlazione lineare ρ non cambia;
- Se x e y sono indipendenti è nulla la loro correlazione (condizione necessaria ma non sufficiente).

,960	40,997	40,997	,960	40,997	40,997	,905	38,680	38,680
,602	25,733	66,731	,602	25,733	66,731	,657	28,051	66,731
,472	20,159	86,890						
,148	6,314	93,204						
,101	4,326	97,530						
,035	1,475	99,005						
,023	,995	100,000						
9,434E-17	4,030E-15	100,000						
2,418E-17	1,033E-15	100,000						
-5,323E-18	-2,274E-16	100,000						

Fig. 4

Grafico delle componenti 1 e 2.

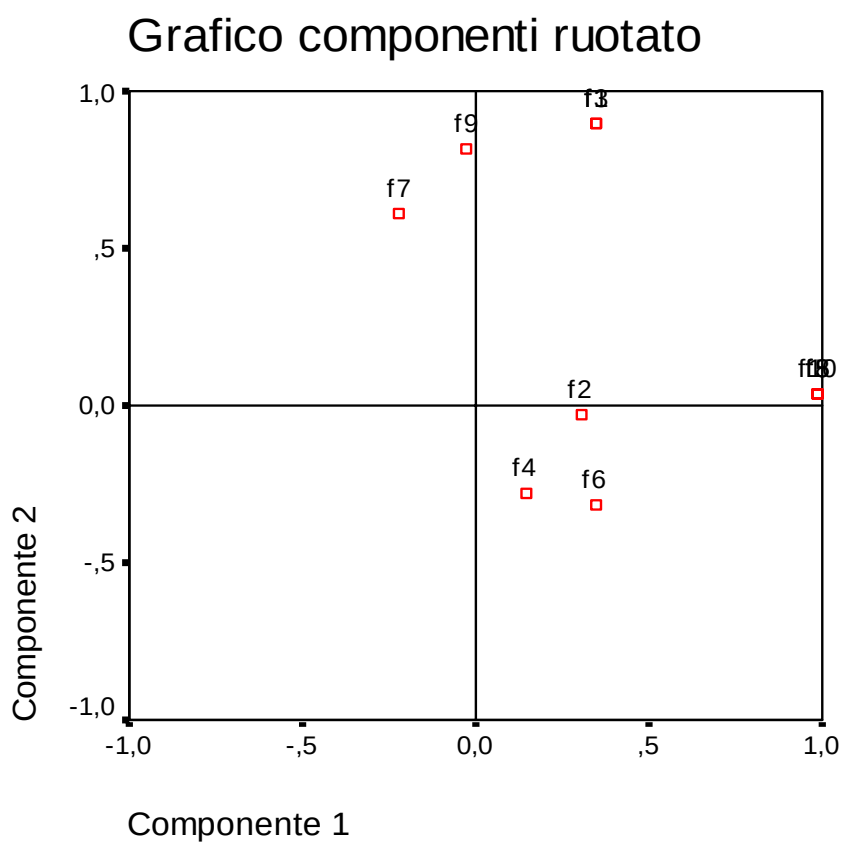


Fig. 5

Dalla Fig.5 si evidenzia che la varianza totale spiegata per i primi due fattori è del 66,71% che è un buon indicatore dell'informazione contenuta (Vedi Fig.4).

Possiamo dire che vi sono dei raggruppamenti di variabili e precisamente:

- 1° gruppo: f2, f4, f6;
- 2° gruppo: f7, f9, f3, f1(le ultime due variabili sovrapposte);
- 3° gruppo f5, f8, f10 (sovrapposti, vedi tabella delle coordinate).

	Coordinate	
	Iniziale	Estrazione
F1	,242	,225
F2	,265	,247
F3	,242	,225
F4	,242	,225
F5	,273	,271
F6	,083	,030
F7	,205	,101
F8	,273	,271
F9	,242	,167
F10	,273	,271

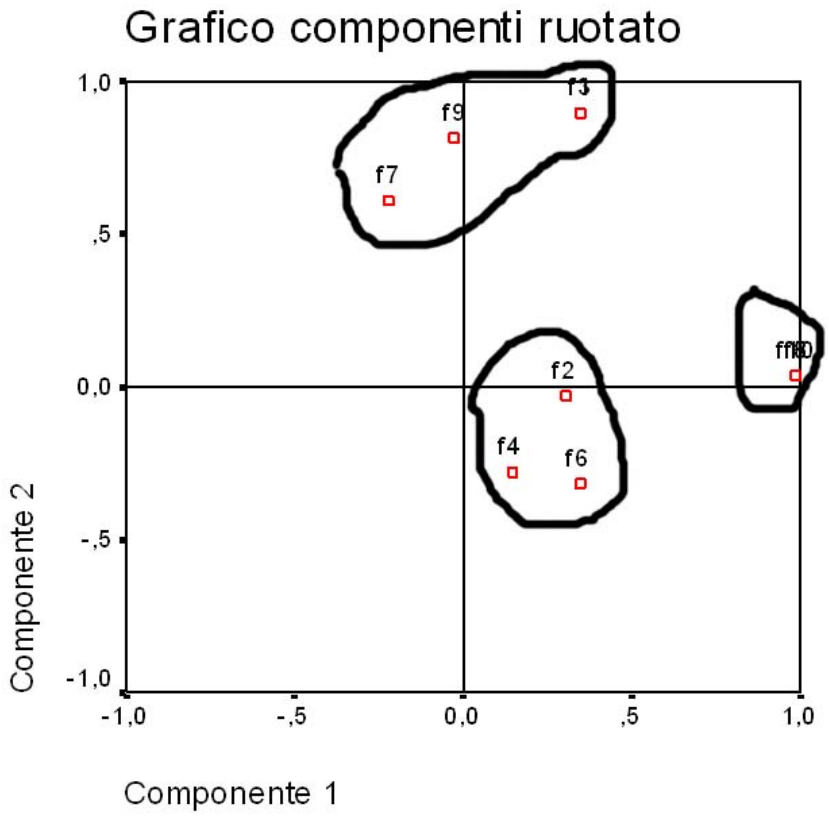


Fig. 6

Stessa immagine della precedente ma con le variabili messe in evidenza.

8.7.4 Le variabili supplementari.

Prendiamo in considerazione la tabella di contingenza della Fig.1 per riprendere la nostra simulazione.

Aggiungiamo alla tabella, secondo il nostro piano sperimentale e soprattutto secondo le nostre ipotesi ed analisi a priori delle variabili supplementari che mettono in evidenza alcuni profili di allievi particolari con i quali vorremmo confrontare la popolazione degli studenti.

Tabella di contingenza con variabili supplementari										
	f1	**f2**	**f3**	**f4**	**f5**	**f6**	**f7**	**f8**	**f9**	**f10**

A1	1	0	1	0	0	1	1	0	1	0
A2	1	0	1	0	0	0	0	0	1	0
A3	1	1	1	1	1	1	1	1	1	1
A4	1	1	1	1	1	1	0	1	1	1
A5	1	1	1	1	1	1	0	1	1	1
A6	1	1	1	1	1	1	0	1	0	1
A7	0	1	0	1	0	1	0	0	0	0
A8	0	0	0	1	0	1	0	0	0	0
A9	0	0	0	0	1	1	0	1	0	1
A10	1	0	1	0	1	1	0	1	1	1
A11	0	1	0	1	0	1	0	0	1	0
A12	1	1	1	1	0	1	1	0	1	0
S1	0	1	0	1	0	1	0	0	0	0
S2	1	0	1	0	0	0	1	0	1	0

Fig. 7

S1 e S2 rappresentato le nostre variabili supplementari che individuano le concezioni delle variabili del primo e secondo gruppo su citate:

1° gruppo: f2, f4, f6;

2° gruppo: f7, f9, f3, f1.

Adesso consideriamo la matrice trasposta utilizzando Excel (Copia ed incolla speciale in un file vuoto.

	A1	**A2**	**A3**	**A4**	**A5**	**A6**	**A7**	**A8**	**A9**	**A10**	**A11**	**A12**	**S1**	**S2**
f1	1	1	1	1	1	1	0	0	0	1	0	1	0	1
f2	0	0	1	1	1	1	1	0	0	0	1	1	1	0
f3	1	1	1	1	1	1	0	0	0	1	0	1	0	1
f4	0	0	1	1	1	1	1	1	0	0	1	1	1	0
f5	0	0	1	1	1	1	0	0	1	1	0	0	0	0
f6	1	0	1	1	1	1	1	1	1	1	1	1	1	0
f7	1	0	1	0	0	0	0	0	0	0	0	1	0	1
f8	0	0	1	1	1	1	0	0	1	1	0	0	0	0
f9	1	1	1	1	1	0	0	0	0	1	1	1	0	1
f10	0	0	1	1	1	1	0	0	1	1	0	0	0	0

Fig. 8

Adesso sono variabili gli allievi assieme alle nostre variabili supplementari. Quindi ci aspetteremo un grafo dell'analisi fattoriale che metta in evidenza il ruolo degli allievi rispetto alle variabili supplementari.

Varianza totale spiegata

	Componente	Autovalori iniziali(a)			Pesi dei fattori non ruotati			Pesi dei fattori ruotati		
		Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata
Semplice	1	1,096	39,146	39,146	1,096	39,146	39,146	1,024	36,577	36,577
	2	,916	32,719	71,865	,916	32,719	71,865	,988	35,288	71,865
	3	,408	14,554	86,419						
	4	,212	7,567	93,986						
	5	,128	4,561	98,547						
	6	,041	1,453	100,000						
	7	1,819E-16	6,495E-15	100,000						
	8	5,754E-17	2,055E-15	100,000						
	9	2,950E-17	1,053E-15	100,000						
	10	9,194E-18	3,284E-16	100,000						
	11	-7,117E-33	-2,542E-31	100,000						
	12	-1,217E-17	-4,346E-16	100,000						
	13	-2,272E-16	-8,113E-15	100,000						

Fig. 9

Da questa tabella vediamo che l'informazione totale è del 71,8%.

Matrice dei componenti ruotata(a)

	Coordinate	
	1	2
A1	,440	-,035
A2	,352	-,143
A4	-,138	,016
A5	-,138	,016
A6	-,280	,031
A7	-,125	,460
A8	-,091	,309
A9	-,416	-,176
A10	-,064	-,319
A11	,017	,445
A12	,365	,301

S1	-,125	,460
S2	,490	-,159

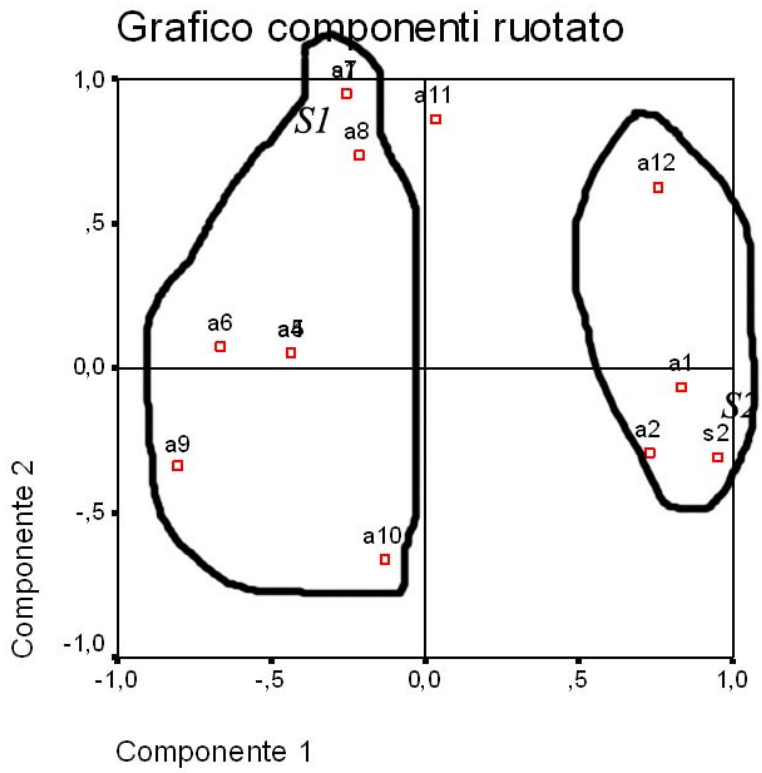


Fig. 10

Dalla Fig. 10 evidenziamo che gli studenti si sono distribuiti, rispetto al fattore individuato dalla y, attorno alle due variabili S1 ed S2. Per determinare S1 ci siamo serviti della tabella (Fig. 9) delle coordinate ed abbiamo visto che la variabile A7 ed S1 erano sovrapposte.
Il metodo delle variabili supplementari si può utilizzare sia con la similarità di Lermann che con l'analisi implicativa di Gras. Attraverso uno studio teorico sperimentale[18] sono state considerate le analogie e differenze tra analisi fattoriale ed analisi implicativa.

[18] Statistical Implicative Analysis : theory and applications, Series: Studies in Computational Intelligence , Vol. 127, Editors: R. Gras, E. Suzuki, F. Guillet, F. Spagnolo, 2008, XVI, 514 p. 147 illus., Hardcover, Springer, ISBN: 978-3-540-78982-6.

Appendice: Tavola dei valori critici del Chi Quadro.

Probabilità sotto Ho $\chi \leq$ Chi Quadro

	.99	.98	.95	.90	.80	.70	.50	.30	.20	.10	.05	.02	.01	.001
1	.00016	.00063	.0039	.0.16	.064	.15	.46	1.07	1.64	2.71	3.84	5.41	6.64	10.83
2	.02	.04	.10	.21	.45	.71	1.39	2.41	3.22	4.60	5.99	7.82	9.21	13.82
3	.12	.18	.35	.58	1.00	1.42	2.37	3.66	4.64	6.25	7.82	9.84	11.34	16.27
4	.30	.43	.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	11.67	13.28	18.46
5	.55	.75	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	13.39	15.09	20.52
6	.87	1.13	1.64	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	15.03	16.81	22.46
7	1.24	1.56	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	16.62	18.48	24.32
8	1.65	2.03	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	18.17	20.09	26.12
9	2.09	2.53	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	19.68	21.67	27.88
10	2.56	3.06	3.94	4.86	6.18	7.27	9.34	11.78	13.44	15.99	18.31	21.16	23.21	29.59
11	3.05	3.61	4.58	5.58	6.99	8.15	10.34	12.90	14.63	17.28	19.68	22.62	24.72	31.26
12	3.57	4.18	5.23	6.30	7.81	9.03	11.34	14.01	15.81	18.55	21.03	24.05	26.22	32.91
13	4.11	4.76	5.89	7.04	8.63	9.93	12.34	15.12	16.98	19.81	22.36	25.47	27.69	34.53
14	4.66	5.37	6.57	7.79	9.47	10.82	13.34	16.22	18.15	21.06	23.68	26.87	29.14	36.12
15	5.23	5.98	7.26	8.55	10.31	11.72	14.34	17.32	19.31	22.31	25.00	28.26	30.58	37.70
16	5.81	6.61	7.96	9.31	11.15	12.62	15.34	18.42	20.46	23.54	26.30	29.63	32.00	39.29
17	6.41	7.26	8.67	10.08	12.00	13.53	16.34	19.51	21.62	24.77	27.59	31.00	33.41	40.75
18	7.02	7.91	9.39	10.86	12.86	14.44	17.34	20.60	22.76	25.99	28.87	32.35	43.80	42.31
19	7.63	8.57	10.12	11.65	13.72	15.35	18.34	21.69	23.90	27.20	30.14	33.69	36.19	43.82
20	8.26	9.24	10.85	12.44	14.58	16.27	19.34	22.78	25.04	28.41	31.41	35.02	37.57	45.32
21	8.90	9.92	11.59	13.24	15.44	17.18	20.34	23.86	26.17	29.62	32.67	36.34	38.93	46.80
22	9.54	10.60	12.34	14.04	16.31	18.10	21.34	24.94	27.30	30.81	33.92	37.66	40.29	48.27
23	10.20	11.29	13.09	14.85	17.19	19.02	22.34	26.02	28.43	32.01	35.17	38.97	41.64	49.73
24	10.86	11.99	13.85	15.66	18.06	19.94	23.34	27.10	29.55	33.20	36.42	40.27	42.98	51.18
25	11.52	12.70	14.61	16.47	18.94	20.87	24.34	28.17	30.68	34.38	37.65	41.57	44.31	52.62
26	12.20	13.41	15.38	17.29	19.82	21.79	25.34	29.25	31.80	35.56	38.88	42.86	45.64	54.05
27	12.88	14.12	16.15	18.11	20.70	22.72	26.34	30.32	32.91	36.74	40.11	44.14	46.96	55.48
28	13.56	14.85	16.93	18.94	21.59	23.65	27.34	31.39	34.03	37.92	41.34	45.42	48.28	56.89
29	14.26	15.57	17.71	19.77	22.48	24.58	28.34	32.46	35.14	39.09	42.56	46.69	49.59	58.30
30	14.95	16.31	18.49	20.60	23.36	25.51	29.34	33.53	36.25	40.26	43.77	47.96	50.89	59.70

TAVOLA della Distribuzione Normale

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

x	*.00*	*.01*	*.02*	*.03*	*.04*	*.05*	*.06*	*.07*	*.08*	*.09*
..0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5310	.5359
.1	.5398	.5438	.5478	.5517	.5357	.5590	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6061	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6918	.6950	.6685	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7380	.7422	.7454	.7480	.7517	.7540
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8105	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8043	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9230	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9658	.9664	.9671	.9678	.9686	.9603	.9609	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9876	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

.