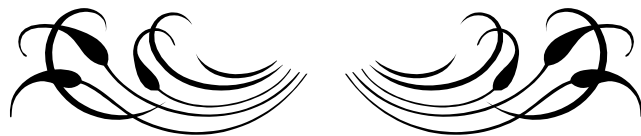


IMPLICAZIONE TRA VARIABILI BINARIE: L'Implicazione di Gras



“Date due variabili binarie a e b, in quale misura posso assicurare che in una popolazione da ogni osservazione di a segue necessariamente quella di b?”

“E’ vero che se a allora b?”

Siano A e B due sotto-popolazioni di aventi rispettivamente le variabili a e b .

Se $A \subset B$ la proposizione è verificata, ma generalmente la intersezione $A \cap \bar{B}$ è non vuota.

$$\varphi(a, \bar{b}) = 1 - \text{Prob}[\text{card}(X \cap \bar{Y}) \leq \text{card}(A \cap \bar{B})]$$

e si dirà:

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

Dove:

$$n_a = \text{card}A, n_b = \text{card}B, n_{a\bar{b}} = \text{card}(A \cap \bar{B})$$

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

q si chiama indice di implicazione.

L'intensità di implicazione é data da:

Nel caso in cui $n_b \neq n$, l'intensità d'implicazione di a su b é

$$\varphi(a, \bar{b}) = 1 - \text{Prob}[Q(a, \bar{b}) \leq q(a, \bar{b})] = \frac{1}{\sqrt{2\pi}} \int_{q(a, \bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Se $q(a, \bar{b}) = -1,816$

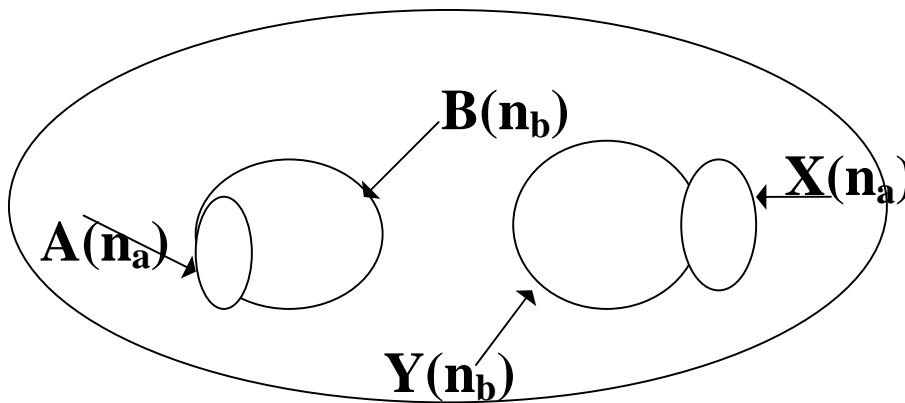
$$\varphi(a, \bar{b}) = 1 - \frac{1}{\sqrt{2\pi}} \int_{q(a, \bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt = 0,9653$$

si dirà che $a \Rightarrow b$ al livello di significatività del 96,5%

- **A e B sono gli insiemi con i dati sperimentali.**
- **X e Y due parti qualunque di E (Insieme comprendente A e B) scelti aleatoriamente e indipendentemente e della stessa cardinalità di A e B. La variabile Aleatoria Card($X \cap \bar{Y}$) segue**

la legge di Poisson di parametro n $p(a)p(\bar{b})$ (di media e varianza $n\pi$).

Intuitivamente l'implicazione $a \Rightarrow b$ sarà ammissibile al termine di una esperienza se il numero di individui di E che la contraddice nell'esperienza é molto piccolo in rapporto al numero di individui attesi sotto l'ipotesi di assenza di legami.



Esempio:

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

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

Esempio di calcolo dell'indice di intensità di implicazione

$a \Rightarrow b$:

$$n_a = 7$$

$$n_{\bar{b}} = 4$$

$$n_{a \wedge \bar{b}} = 5$$

$$q(a, \bar{b}) = \frac{5 - \frac{7 \cdot 4}{10}}{\sqrt{\frac{7 \cdot 4}{10}}} = \frac{5 - 2,8}{1,67} = 1,314$$

10.	1	1
	7/10	6/10

Se l'intensità d'implicazione è abbastanza piccola le due variabili non saranno legate.

Se $\varphi(a,b)=0,965$ è un livello di confidenza del 96.5% per l'implicazione in quanto la probabilità che $\text{card}(X \cap Y)$ sia più piccola di 1 è 0,035 (In generale se il livello di confidenza è superiore al 95% si ritiene accettabile). Il livello di confidenza va calcolato con l'integrale relativo alla distribuzione di Poisson.



Presentazione di "CHIC"

E' stato messo a punto dal gruppo I.R.MA.R.¹ un programma su PC che consente di fare l'analisi implicativa abbastanza celermente. Il programma si chiama CHIC (Compilato in Pascal VI) ed il suo menu (sino alla edizione del 1994 curata da Ratsimba-Rajohn) è:

- 1.Edizione;
- 2.Classificazione gerarchica delle variabili;
- 3.Implicazioni tra variabili;

¹Istituto di ricerca Matematica di Rennes, Università di Rennes I, prof. Regis Gras. Il programma CHIC è disponibile al seguente indirizzo: Prof.R.Gras IRMAR Institut Mathématique Campus de Beaulieu, 35042 Rennes Cedex, Francia.

4. Calcolo dei parametri;
5. Costruzione del grafo implicativo;
6. Gerarchia implicativa di classi;
7. Livelli significativi;
8. Individui e/o categorie determinanti;
9. Conversioni di vecchi file ____ .DTA.
10. Abbandonare.

1. La prima operazione riguarda l'Edizione che contiene un sotto menu riguardante l'immissione di dati, la consultazione dei file che sono il estensione .DTA, concatenazione di file o trasposizione, correzione. Si ricorda che le varianili possono essere di tre tipi: binarie, frequenziali, modali.

2. La seconda operazione riguarda la classificazione gerarchica delle variabili. 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 legat 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. La rappresentazione grafica di CHIC

della classificazione gerarchica tra variabili ci consente di avere delle informazioni sul raggruppamento di classi di variabili.

3. Il calcolo dell'implicazione tra variabili viene eseguito previa chiarificazione del tipo di variabile che si utilizza. Se non vi è coerenza tra il tipo di variabile ed il valore massimo che questa può assumere il programma si arresta (variabile binaria, max=1, ad esempio).

Dopo aver calcolato l'implicazione tra le variabili viene chiesto di salvare il file con estensioni diverse per poter fare le ulteriori analisi statistiche.

In particolare se al file di origine era stato dato il nome "piro" con estensione .DTA per la "Gerarchia implicativa di classi" l'estensione sarà "piro.VAL". Mentre per la "Costruzione del grafo implicativo" sarà "piro.GRI". Se invece si vuole conservare il file riguardante le somme delle risposte giuste per variabile (occorrenze) l'estensione sarà "piro.OCC"².

4. Il calcolo dei parametri consente di avere le medie per variabile ed individuo, varianza e correlazioni tra variabili.

5. Il grafo implicativo dopo aver digitato il file con l'estensione ".GRI" viene rappresentato nello schermo chiedendo prima a quale soglia di confidenza si vuole l'implicazione.

6. La gerarchia implicativa delle classi dopo aver digitato il file con l'estensione ".VAL" viene rappresentato nello schermo. Questo grafo ci fornisce delle informazioni sulla implicazione tra classi di variabili. Per poter costruire un grafo per la gerarchia

²L'operazione di salvare il file con estensioni diverse è importante per il programma, nel senso che le scelte 6 o 5 non possono essere fatte senza la scelta 3, le scelte 7 o 8 non possono essere fatte senza le scelte 3 e poi 6, la scelta 8 si può fare attualmente con i dati binari.

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 $(Q(a, \bar{b}) > q(a, \bar{b}))$ sarà E:

$$E = -p \log_2(p) - (1-p) \log_2(1-p)$$

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

E é massima quando $p = 0.5$

E é minima quando $p = 1$ o $p = 0$ ($0 \log_2(0) = 0$)

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 é 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 variabili ed a quale livello.