**LABORATORIO DI CALCOLO**

# DIPENDENZA TRA CARATTERI

La statistica bidimensionale si occupa dello studio congiunto di due fenomeni X e Y. Può essere quindi interessante verificare se esista una qualche relazione tra di essi, cioè se i due caratteri osservati siano dipendenti.

La dipendenza tra caratteri quantitativi, o variabili, viene identificata col nome di CORRELAZIONE. In questo caso è possibile un primo approccio grafico rappresentando le coppie (x,y) ottenute dalla rilevazione congiunta dei due fenomeni X e Y in un diagramma scatter e osservando la nuvola dei punti ottenuta identificando se sussiste una qualche relazione tra X e Y.

Più in generale, la dipendenza tra caratteri di cui almeno uno qualitativo, o mutabile, viene identificata col nome di CONNESSIONE. In tal caso, non potendo operare matematicamente sulle modalità dei caratteri, sarà necessario individuare una procedura che ci permetta di valutare in altro modo l'eventuale dipendenza tra X e Y.
Sappiamo che è possibile riassumere congiuntamente i due fenomeni mediante la tabella doppia XY. Ad esempio:

| X \ Y | a | b | Totale |
|---|---|---|---|
| 1 | 0 | 2 | 2 |
| 2 | 3 | 0 | 3 |
| 3 | 2 | 2 | 4 |
| Totale | 5 | 4 | 9 |

Il numero delle modalità di X viene identificato con **h** e il numero di modalità di Y con **k** (nell'esempio h=3 e k=2).

Le frequenze congiunte vengono identificate con con $n_{i,j}$ (nell'esempio $n_{21} = 3$).

Le frequenze marginali del carattere X si identificano con $n_{i\bullet}$ e le frequenze marginali del carattere Y con $n_{\bullet j}$ (nell'esempio dove i=1,...,3 e j=1,...,2)
**N** rappresenta la totalità delle unità statistiche (nell'esempio N = 12).

Se tutte le frequenze congiunte soddisfano la condizione $n_{ij} = \frac{ni\bullet \quad n\bullet j}{N}$ allora si è in presenza di perfetta indipendenza; se almeno una delle frequenze congiunte non soddisfa questa relazione allora si è in presenza di dipendenza tra i caratteri.

Per valutare se esiste dipendenza tra X e Y è quindi necessario realizzare i seguenti passaggi.

- determinare la tabella in condizioni di indipendenza (tabella teorica) calcolando le frequenze teoriche congiunte mediante la relazione $\overline{n_{ij}} = \frac{ni\bullet \quad n\bullet j}{N}$;
- calcolare la tabella delle contingenze come differenza tra le frequenze congiunte della tabella iniziale (empirica) e di quella teorica $c_{ij} = n_{ij} - \overline{n_{ij}}$ ;
- verificare se $c_{ij} = 0 \ \forall$ i,j (cioè $n_{i,j} = \overline{n_{ij}} \ \forall$ i,j) , se anche una sola delle contingenze risulta diversa da zero allora i due caratteri X e Y sono dipendenti.

(si ricordano brevemente le proprietà delle contingenze: $\sum_i c_{ij} = \sum_j c_{ij} = \sum_{i,j} c_{ij} = 0$ )

Se si osserva l'esistenza di dipendenza tra i due caratteri può essere interessante valutarne il grado nel seguente modo:

- Calcolare il coefficiente quadratico medio di contingenza Ic = $\sqrt{\frac{\chi^2}{\chi^2 + N}}$ , dove

$\chi^2 = \sum_{i,j} \frac{c_{ij}^2}{n_{ij}}$ (si legge chi quadro) è l'indice di contingenza di Pearson.

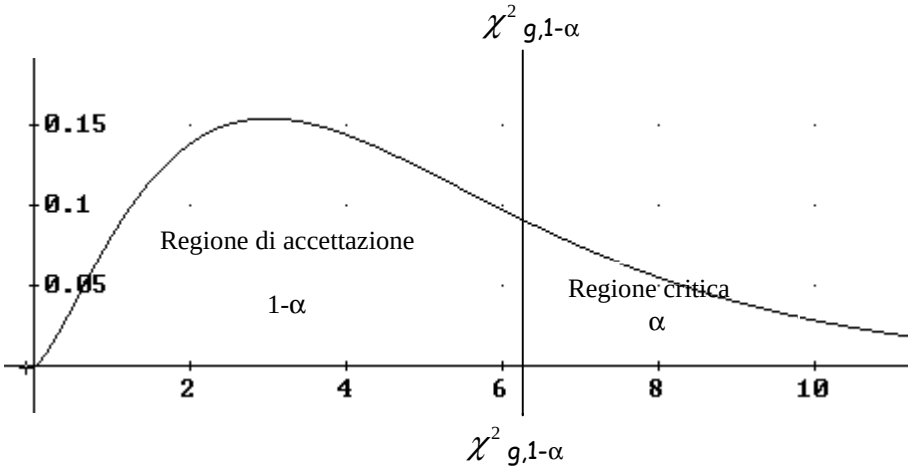
Nel caso di perfetta indipendenza Ic = 0 poiché tutte le contingenze risultano nulle; altrimenti si è in presenza di dipendenza tra i caratteri e all'aumentare del suo grado il valore dell'indice Ic tende a 1.

## L'applicazione del test $\chi^2$

La verifica dell'esistenza di dipendenza o meno tra due caratteri può essere realizzata anche mediante l'applicazione del test $\chi^2$ il quale mette a confronto le frequenze congiunte della tabella empirica con le frequenze teoriche della tabella in condizioni di indipendenza per valutare la bontà dell'accordo tra i due insiemi di valori.

La procedura per la sua applicazione è la seguente:

- si fissano le due ipotesi iniziali H0: caratteri indipendenti H1: caratteri dipendenti.
- si calcolano i gradi di libertà g=(h-1)(k-1) che permettono di individuare la variabile $\chi^2_g$ di confronto la quale, ad esempio con g=5, ha il seguente andamento:




- si sceglie $\alpha$ cioè il livello di significatività del test il quale permette di separare l'area sottesa dalla variabile $\chi^2_g$ in due regioni: quella di accettazione (con area pari $1 - \alpha$ ) e quella critica (con area pari $\alpha$ ); generalmente $\alpha$ è scelto come segue:
  - $\alpha = 0.05$ => $1 - \alpha = 0.95$ e si dice che il test è significativo
  - $\alpha = 0.01$ => $1 - \alpha = 0.99$ e si dice che il test è molto significativo

- si determina sulle tavole della variabile $\chi^2$ il <u>valore critico</u> $\chi^2_{g,1-\alpha}$ , cioè un valore della variabile tale che $P(\chi^2 < \chi^2_{g,1-\alpha}) = 1 - \alpha$
- si confronta il valore calcolato dell'indice $\chi^2$ di Pearson con il valore critico letto sulle tavole

  e se $\chi^2 < \chi^2_{g,1-\alpha}$ allora si accetta l'ipotesi H0 di caratteri indipendenti;

  altrimenti si accetta l'ipotesi H1 di caratteri dipendenti.