Cenni di Statistica Descrittiva Bivariata

Data un'unità statistica, con *carattere* dell'unità statistica *(o variabile statistica)* si intende la grandezza che viene rilevata in relazione a tale unità: se l'unità statistica è costituita da un individuo un carattere soggetto all'indagine statistica, ad esempio, potrebbe essere il colore degli occhi.

Per *modalità* di un carattere si intende una delle possibili espressioni con cui un carattere può manifestarsi: ad esempio se il carattere considerato è il colore degli occhi una modalità potrebbe essere marrone piuttosto che azzurro, etc.

Consideriamo il caso di una o più caratteristiche qualitative e quantitative riferite ad una determinata unità statistica. Ad esempio, prendiamo un individuo estratto da una certa popolazione e consideriamone l'altezza, il colore dei capelli, il peso, l'orientamento religioso, etc. In generale le relazioni che intercorrono tra i vari caratteri citati sarebbero argomento della statistica multivariata, noi ci accontenteremo di occuparci di due soli caratteri, che potranno essere entrambi qualitativi, entrambi quantitativi oppure uno qualitativo ed uno quantitativo.

La **Statistica Bivariata** si occupa delle caratteristiche di *indipendenza* o dipendenza (*connessione*) reciproca tra i due caratteri considerati.

Avremo perciò a che fare con due caratteri relativi alla stessa unità statistica che indicheremo convenzionalmente con X e Y.

X e Y potranno essere *variabili statistiche* (dati quantitativi) oppure *mutabili statistiche* (dati qualitativi): potremmo incrociare due caratteri qualitativi (due *mutabili*), due caratteri quantitativi (due *variabili*) oppure un carattere qualitativo ed uno quantitativo (una *mutabile* ed una *variabile*).

Le modalità con cui si esprime la grandezza X saranno:

$$x_1, x_2, \ldots\ldots, x_r$$

Le modalità con cui si esprime la grandezza Y saranno:

$$y_1, y_2, \ldots\ldots, y_c$$

Intendiamo quindi dire che, relativamente all'unità statistica data, sono state effettuate le osservazioni di r modalità del carattere X congiuntamente a c modalità del carattere Y.

Per l'osservazione *congiunta* di X nella modalità x_i e di Y nella modalità y_j utilizziamo la notazione seguente per indicarne la frequenza:

$$n_{ij}$$

Le osservazioni possono poi essere riassunte poi, schematicamente, nella seguente tabella a doppia entrata:

	y_1	y_2	....	y_j	....	y_c	Tot
x_1	n_{11}	n_{12}	....	n_{1j}	....	n_{1c}	n_{10}
x_2	n_{21}	n_{22}	....	n_{2j}	....	n_{2c}	n_{20}
....	....	....	....	....	....	....	....
x_i	n_{i1}	n_{i2}	....	n_{ij}	....	n_{ic}	n_{i0}
....	....	....	....	....	....	....	....
x_r	n_{r1}	n_{r2}	....	n_{rj}	....	n_{rc}	n_{r0}
Tot	n_{01}	n_{02}	....	n_{0j}	....	n_{0c}	n

Def. Frequenza Marginale X

Sono le n_{i0} (frequenza marginale della modalità x_i)

$$n_{i0} = \sum_{j=1}^{c} n_{ij}$$

Def. Frequenza Marginale Y

Sono le n_{0j} (frequenza marginale della modalità y_j)

$$n_{0j} = \sum_{i=1}^{r} n_{ij}$$

Dalla tabelle possiamo ottenere *n+c+2* distribuzione univariate:

Distribuzioni di Y condizionate ad X (sono in numero di r) :

$$Y\left| x_i = \begin{Bmatrix} y_1 ..... y_c \\ n_{i1} .... n_{ic} \end{Bmatrix} n_{i0} \quad (\mathrm{i} = 1,...,\mathrm{r}) \right.$$

Distribuzioni di X condizionate ad Y (sono in numero di c) :

$$X\left| y_j = \begin{Bmatrix} x_1 ..... x_r \\ n_{1j} .... n_{rj} \end{Bmatrix} n_{0j} \quad (\mathrm{j} = 1,...,\mathrm{c}) \right.$$

Distribuzione marginale della variabile Y :

$$Y = \begin{Bmatrix} y_1 ..... y_c \\ n_{01} .... n_{0c} \end{Bmatrix} n$$

Distribuzione marginale della variabile X :

$$X = \begin{Bmatrix} x_1 ..... x_r \\ n_{10} .... n_{r0} \end{Bmatrix} n$$

Le due distribuzioni marginali (di X e Y) sono costituite dalle frequenze marginali.

01 - Indipendenza e Connessione

Introduciamo il concetto di indipendenza delle due variabili/mutabili X e Y partendo dalle seguenti tabelle:

INDIPENDENZA di X ad Y

Frequenze assolute

		Carattere Y			
Carattere X		Y_1	Y_2	Y_3	Tot.
	X_1	10	5	2	17
	X_2	30	15	6	51
	X_3	40	20	8	68
	X_4	60	30	12	102
	Tot.	140	70	28	238

Frequenze relative del carattere X condizionato ad Y

		Carattere Y			
Carattere X		Y_1	Y_2	Y_3	
	X_1	0,0714	0,0714	0,0714	0,0714
	X_2	0,2143	0,2143	0,2143	0,2143
	X_3	0,2857	0,2857	0,2857	0,2857
	X_4	0,4286	0,4286	0,4286	0,4286
	Tot.	1,0000	1,0000	1,0000	1,0000

Il carattere X si dirà indipendente dal carattere Y se tutte le distribuzioni relative di X condizionate ad Y risultano uguali tra loro e uguali alla distribuzione marginale (e dunque, al variare della modalità Y la distribuzione relativa di X è la medesima).

INDIPENDENZA di Y ad X

Frequenze assolute

		Carattere Y			
Carattere X		Y_1	Y_2	Y_3	Tot.
	X_1	10	5	2	17
	X_2	30	15	6	51
	X_3	40	20	8	68
	X_4	60	30	12	102
	Tot.	140	70	28	238

Frequenze relative del carattere Y condizionato ad X

		Carattere Y			
Carattere X		Y_1	Y_2	Y_3	Tot.
	X_1	0,5882	0,2941	0,1176	1,0000
	X_2	0,5882	0,2941	0,1176	1,0000
	X_3	0,5882	0,2941	0,1176	1,0000
	X_4	0,5882	0,2941	0,1176	1,0000
		0,5882	0,2941	0,1176	1,0000

Il carattere Y sarà detto indipendente dal carattere X se tutte le distribuzioni relative di Y condizionate ad X risultano uguali tra loro e uguali alla distribuzione marginale (e dunque, al variare della modalità X la distribuzione relativa di Y è la medesima)

E' possibile dimostrare che se il carattere X è indipendente dal carattere Y, allora vale anche la relazione contraria: anche il carattere Y sarà indipendente dal carattere X.

Pertanto: due caratteri X ed Y si diranno indipendenti se le distribuzioni relative condizionate di un carattere rispetto alle modalità dell'altro sono uguali.

Def. Frequenza Teorica

$$n_{ij}^{*} = \frac{n_{i0} \cdot n_{0j}}{n}$$

Def. Due variabili aleatorie X e Y si dicono **indipendenti** se le frequenze relative di ogni modalità x_i condizionate ad y_j sono le stesse per ogni modalità di Y e sono uguali alla frequenza relativa marginale della modalità x_i.(*)

In altri termini:

$$\frac{n_{ij}}{n_{0j}} = \frac{n_{i0}}{n} \ \forall i, j \ \Leftrightarrow n_{ij} = \frac{n_{i0} \cdot n_{0j}}{n} = n_{ij}^* \quad \forall i, j$$

In pratica le due variabili aleatorie X e Y sono indipendenti se tutte le frequenze rilevate nella tabella di indagine sono uguali alle frequenze teoriche sopra definite.

(*) Nota: evidentemente vale una analoga definizione invertiti x_i ed y_j, X ed Y, con la relazione scritta come:

$$\frac{n_{ij}}{n_{i0}} = \frac{n_{0j}}{n} \ \forall i, j \ \Leftrightarrow n_{ij} = \frac{n_{i0} \cdot n_{0j}}{n} = n_{ij}^* \quad \forall i, j$$

Si definisce *contingenza* relativa alla modalità x_i della variabile X ed alla modalità y_j della variabile Y la differenza tra frequenza osservata (per i due caratteri congiunti) e frequenza teorica

Def. Contingenza

$$C_{ij} = n_{ij} - n_{ij}^*$$

Notiamo che nel caso di variabili aleatorie indipendenti le contingenze sono nulle.

Aggiungiamo i seguenti esempi:

Esempio 1 :

Consideriamo la seguente distribuzione bivariata, si può calcolare facilmente che le due variabili/mutabili sono (perfettamente) indipendenti:

PERFETTA INDIPENDENZA DI X E Y

	Y_1	Y_2	Y_3	n_{i0}
X_1	1	3	2	6
X_2	2	6	4	12
X_3	3	9	6	18
n_{0j}	6	18	12	36

Osserviamo che le frequenze riportate nella tabella coincidono con le frequenze teoriche, tutte le contingenze sono nulle, e quindi le due variabili sono perfettamente indipendenti:

Inoltre:

- La condizione di indipendenza è molto stringente: basta alterare il valore di una cella della tabella affinché non si verifichi più
- E' una condizione molto difficile da ottenere anche per variabili statistiche molto "lontane" dal punto di vista logico
- Il grado di allontanamento dalla condizione di indipendenza può essere valutato, in prima istanza, dalle contingenze.

In seconda istanza il grado di allontanamento dalla condizione di indipendenza é valutato dagli indici di connessione che tratteremo successivamente

Dalla parte opposta, nel comportamento atteso di due caratteri congiunti delle variabili X e Y, rispetto al concetto di indipendenza abbiamo il concetto di **connessione**:

Def. **Connessione :** Due variabili aleatorie si dicono perfettamente **connesse** (con connessione massima) quando fissata una modalità della prima si può risalire in modo univoco alla modalità della seconda.

Esempio 2

PERFETTA CONNESSIONE DI X E Y

	Y_1	Y_2	Y_3	n_{i0}
X_1	0	0	7	7
X_2	5	0	0	5
X_3	0	3	0	3
n_{0j}	5	3	7	15

Notiamo che per la perfetta connessione tra le due variabili/mutabili X e Y è necessario che esse abbiamo un eguale numero di modalità nella distruzione statistica considerata.

Nel seguente esempio, essendo diverse il numero delle modalità di X e di Y, esse non potranno essere perfettamente connesse:

Esempio 3

CONNESSIONE DI X E Y

	Y_1	Y_2	Y_3	Y_4	n_{i0}
X_1	0	1	0	0	1
X_2	1	0	0	1	2
X_3	0	0	1	0	1
n_{0j}	1	1	1	1	4

Osservazioni:

- Se il numero di righe ed il numero di colonne non sono uguali non è possibile stabilire una perfetta (connessione) dipendenza di Y da X (o di X da Y).
- Nella tabella sopra riportata se conosciamo la modalità della e Y possiamo determinare univocamente la modalità della X:
  - $y_1 \leftarrow\rightarrow x_2$, $y_2 \leftarrow\rightarrow$ x_1 , $y_3 \leftarrow\rightarrow x_3$, $y_4 \leftarrow\rightarrow x_2$
- Tuttavia partendo dalle modalità delle X non si può risalire in modo univoco alla modalità Y corrispondente:
  - $x_1 \leftarrow\rightarrow y_2$, $x_3 \leftarrow\rightarrow y_3$, $x_2 \leftarrow\rightarrow$??
- Ne concludiamo che mentre X è perfettamente connessa a Y (connessione unilaterale di X da Y) , il viceversa non vale, e quindi Y non è perfettamente connessa a X.

02 – Tabella Contingenze

Attraverso le contingenze precedentemente definite è possibile creare la tabella delle contingenze il cui termine generale è ovviamente:

$$\left[C_{ij}\right] = \left[n_{ij} - n_{ij}^{*}\right]$$

Per essa vale il seguente teorema:

Teorema (*) : La somma delle contingenze su una riga o su una colonna è uguale a zero

Esempio 4

Si consideri la seguente tabella:

	Y_1	Y_2	Y_3	n_{i0}
X_1	2	7	11	20
X_2	3	7	20	30
X_3	5	16	29	50
n_{0j}	10	30	60	100

Da essa calcoliamo la tabella della frequenze teoriche:

	Y_1	Y_2	Y_3	n_{i0}
X_1	2	6	12	20
X_2	3	9	18	30
X_3	5	15	30	50
n_{0j}	10	30	60	100

Quindi la tabella della contingenze:

	Y_1	Y_2	Y_3	n_{i0}
X_1	0	1	-1	0
X_2	0	-2	2	0
X_3	0	1	-1	0
n_{0j}	0	0	0	0

Come si può notare le somme dei termini della tabella delle contingenze sulle righe e sulle colonne è sempre zero!

(*) Per dimostrazione Vedi Appendice A1

Infine ricordiamo che:

Def. Due variabili/mutabili sono dette **indipendenti** se e solo se:

$$n_{ij} = n_{ij}^{*} \Leftrightarrow C_{ij} = 0$$

Def. Due variabili/mutabili sono dette **positivamente connesse** se e solo se:

$$n_{ij} > n_{ij}^{*} \Leftrightarrow C_{ij} > 0$$

Si osserva un aumento di frequenze rispetto alla condizione di indipendenza

Def. Due variabili/mutabili sono dette **negativamente connesse** se:

$$n_{ij} < n_{ij}^{*} \Leftrightarrow C_{ij} < 0$$

Si osserva una diminuzione di frequenze rispetto alla condizione di indipendenza

03 – Chi – quadrato, Indice di Contingenza Quadratico Medio, Indice di Cramer e di Tschuprow

Iniziamo ora la costruzione di un indice normalizzato che ci porti a considerare l'indipendenza o la connessione di due variabili/mutabili X e Y.

Iniziamo ad introdurre l'indice **Chi-quadrato** (a volte detto anche Chi-Quadro) di Pearson:

Def. Chi-Quadrato:

$$\chi^2 := \sum_{i,j} \frac{C_{ij}^2}{n_{ij}^{*}} \left(= \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{C_{ij}^2}{n_{ij}^{*}} \right)$$

Come vediamo è la somma estesa a tutti gli elementi della tabella della statistica bivariata delle contingenze al quadrato divise per le frequenze teoriche corrispondenti.

Def. Indice di Contingenza Quadratico Medio

$$\Phi = \sqrt{\frac{\chi^2}{n}}$$

E' possibile dimostrare che:

$$0 \leq \frac{\chi^2}{n} = \Phi^2 \leq \min(r-1, c-1)$$

Inoltre.

$\Phi^2 = 0 \Rightarrow$ Indipendenza

$\Phi^2 = r-1 \Rightarrow$ Connessione Perfetta di X da Y

$\Phi^2 = c-1 \Rightarrow$ Connessione Perfetta di Y da X

Da queste osservazioni e dalle diseguaglianze a cui è sottoposto l'indice di contingenza quadratico medio è possibile dedurre un indice normalizzato:

Def. Indice di Cramer

$$\varphi_C = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}} \qquad 0 \le \varphi_C \le 1$$

Osservazioni:

- Per tale indice vale la disuguaglianza $0 \le \varphi_C \le 1$.
- Vale 0 nel caso di indipendenza
- Valori prossimi a 1 denotano una forte connessione tra le due variabili ma non danno indicazioni sulle modalità di associazione
- Nel caso di perfetta connessione con r=c abbiamo φ=1.
- Nel caso in cui:
  X perfettamente connessa a Y (connessione unilaterale di X da Y) e r<c , abbiamo φ =1 (con r ≠ c)
  Y perfettamente connessa a X (connessione unilaterale di Y da X)e c<r , abbiamo φ =1 (con r ≠ c)
  (vedi esempio 5)

L'ultima indicazione che diamo riguarda l'indice di Tschuprow. Anch'esso è normalizzato ed assume quindi valori tra 0 ed 1. Al contrario dell'indice di Cramer, esso assume però il valore 1 solo in caso di perfetta connessione di X ed Y e quindi con r=c (*connessione bilaterale*). In caso di *connessione unilaterale* (cioè perfetta connessione di X da Y o di Y da X ma con r≠c) mentre l'indice di Cramer vale 1, l'indice di Tschuprow assume valore minori di 1 (vedi esempi 5 e 6).

Def. Indice di Tschuprow

$$T = \sqrt{\frac{\left(\frac{\chi^2}{n}\right)}{\sqrt{(c-1)(r-1)}}}$$

Esempio 5 (Perfetta Connessione di X e Y: Connessione Bilaterale)

Si consideri la seguente distribuzione bivariata:

	Y_1	Y_2	Y_3	n_{i0}
X_1	0	0	4	4
X_2	2	0	0	2
X_3	0	4	0	4
n_{0j}	2	4	4	10

Calcoliamo la tabella delle frequenze teoriche:

	Y_1	Y_2	Y_3	n_{i0}
X_1	0,8	1,6	1,6	4
X_2	0,4	0,8	0,8	2
X_3	0,8	1,6	1,6	4
n_{0j}	2	4	4	10

E quindi la tabella delle contingenze:

	Y_1	Y_2	Y_3	n_{i0}
X_1	-0,8	-1,6	2,4	0
X_2	1,6	-0,8	-0,8	0
X_3	-0,8	2,4	-1,6	0
n_{0j}	0	0	0	0

Da essa si ottiene la tabella che permette il calcolo del Chi-quadrato (ogni cella contiene il quadrato della rispettiva contingenza divisa per la corrispondente frequenza teorica):

0,8	1,6	3,6		
6,4	0,8	0,8		
0,8	3,6	1,6		
8	6	6	20	**Chi-quadrato**

Da quest'ultima tabella si possono ottenere agevolmente sia l'indice di Cramer che quello di Tschuprow:

Indice (Cramer)	**1**
Indice (Tschuprow)	**1**

Esempio 6 (Perfetta Connessione di X da Y: Connessione Unilaterale di X da Y)

Si consideri la seguente distribuzione bivariata:

	Y_1	Y_2	Y_3	Y_4	n_{i0}
X_1	0	2	0	0	2
X_2	2	0	2	0	4
X_3	0	0	0	2	2
n_{0j}	2	2	2	2	8

Calcoliamo la tabella delle frequenze teoriche:

	Y_1	Y_2	Y_3	Y_4	n_{i0}
X_1	0,5	0,5	0,5	0,5	2
X_2	1	1	1	1	4
X_3	0,5	0,5	0,5	0,5	2
n_{0j}	2	2	2	2	8

E quindi la tabella delle contingenze:

	Y_1	Y_2	Y_3	Y_4	**totali**
X_1	-0,5	1,5	-0,5	-0,5	0
X_2	1	-1	1	-1	0
X_3	-0,5	-0,5	-0,5	1,5	0
totali	0	0	0	0	0

Da essa si ottiene la tabella che permette il calcolo del Chi-quadrato (ogni cella contiene il quadrato della rispettiva contingenza divisa per la corrispondente frequenza teorica):

	Y_1	Y_2	Y_3	Y_4	n_{i0}	
X_1	0,5	4,5	0,5	0,5		
X_2	1	1	1	1		
X_3	0,5	0,5	0,5	4,5		
n_{0j}	2	6	2	6	**16**	**Chi-quadrato**

Da quest'ultima tabella si possono ottenere agevolmente sia l'indice di Cramer che quello di Tschuprow;

Indice (Cramer)	**1**
Indice (Tschuprow)	**0,904**

04 – Esempio di Verifica di Ipotesi e Test Chi-Quadrato

In generale per la valutazione di ipotesi di dipendenza od indipendenza statistica (in medicina e biologia soprattutto), si ipotizza la distribuzione della variabile Chi-Quadrato definita come una somma quadratica di variabili statistiche normalizzate (gaussiane) secondo la seguente relazione:

$$\chi_n^2 = Z_1^2 + ... + Z_n^2$$

Alla distribuzione stessa vengono attribuiti i seguenti gradi di libertà (per noi r è, come al solito, numero di righe della tabella e c il numero delle colonne):

$$n = (r-1)\cdot(c-1)$$

I *gradi di libertà* costituiscono un parametro essenziale per stabilire la forma analitica della distribuzione della variabile Chi-Quadrato e quindi il modo in cui la relativa tabella deve essere consultata (vedi Appendice A2).

Esempio

Due gruppi di pazienti che indichiamo con A e B, sono composti da 100 individui, tutti soggetti ad una certa malattia. Al gruppo A viene somministrato un siero che non viene invece somministrato al gruppo B. Si vuole valutare se c'è dipendenza tra guarigione e somministrazione del siero.

Consideriamo allora le consuete tabelle e procediamo al calcolo del Chi-Quadrato della tabella:

Frequenze OSSERVATE			
	Guariti	**Non Guariti**	**Totali**
A- con Siero	75	25	100
B-senza Siero	65	35	100
Totali	140	60	200

Frequenze TEORICHE			
	Guariti	**Non Guariti**	**Totali**
A- con Siero	70	30	100
B-senza Siero	70	30	100
Totali	140	60	200

CONTINGENZE			
	Guariti	Non Guariti	Totali
A- con Siero	5	-5	0
B-senza Siero	-5	5	0
Totali	0	0	0

CALCOLO CHI_2			
	Guariti	Non Guariti	
A- con Siero	0,357	0,833	
B-senza Siero	0,357	0,833	
Totali	0,714	1,667	2,381

$$\chi^2_{n=1} = 2{,}381$$

Notiamo che i gradi di libertà sono (2-1)*(2-1)=1

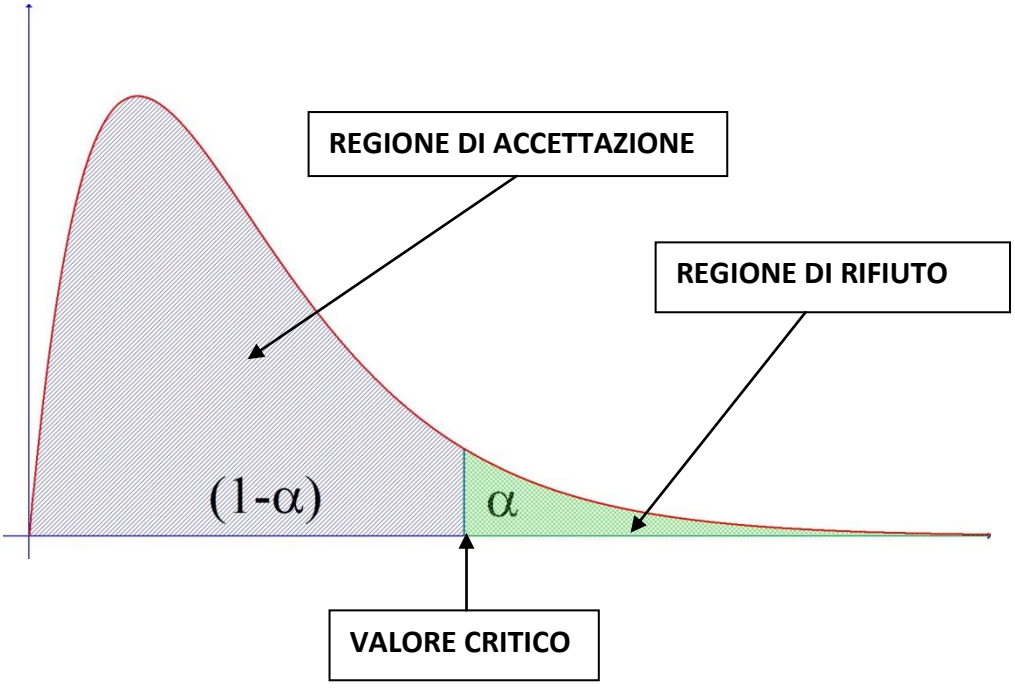
Il Test Chi -Quadrato

Supponiamo che sia da verificare l'ipotesi se il siero utilizzato sia o meno efficace come cura con una probabilità del 95% [di solito indicata con 1-α (con α=0.05=5%) individuato come *livello di significatività*]. Si tratta di verificare l'ipotesi detta H_0 (*ipotesi nulla*) di indipendenza delle due mutabili. Si usa esplicitare anche l'*ipotesi alternativa*, di solito indicata con H_1

H_0 indipendenza **H1 dipendenza**

Sostanzialmente se le due mutabili sono indipendenti allora il siero non è ritenuto efficace, al livello di significatività assegnato, altrimenti è ritenuto efficace. Nel primo caso H_0 è vera (ed H_1 è falsa), nel secondo H_0 è falsa (H_1 è vera), nel prima caso si dice anche che *H_0 è accettata*, nel secondo che *H_0 è rifiutata*,

Se mettiamo in un grafico la distribuzione Chi-Quadrato, avremo la seguente situazione:



Fissato α, l'ipotesi nulla dovrà essere rifiutata se il valore osservato della statistica Chi-quadrato è maggiore del valore critico di una distribuzione Chi-quadrato con 1 grado di libertà.

Per stabilire l'indipendenza delle due mutabili si confronta il valore del Chi-Quadrato misurato con quello teorico (tabulato) e si procede secondo il seguente schema:

$$\chi^2\big|_{osservato} \leq \chi^2_{1-\alpha,n}\big|_{critico} \Rightarrow$$ **H_0 vera (H_1 falsa, H_0 ACCETTATA)**

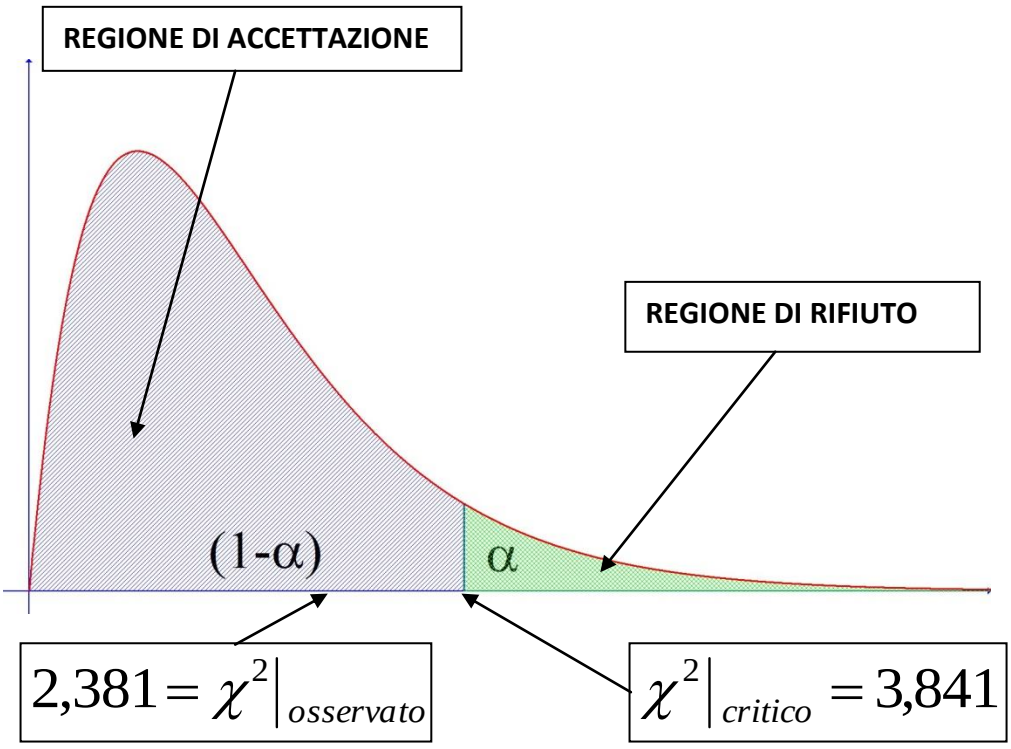
$$\chi^2\big|_{osservato} < \chi^2_{1-\alpha,n}\big|_{critico} \Rightarrow$$ **H_0 falsa (H_1 vera, H_0 RIFIUTATA)**

L'idea, sostanzialmente, è quella che essendo i valori del Chi-Quadrato più piccoli essi sono indicativi della indipendenza delle mutabili, un valore misurato più piccolo di quello teorico indica una maggiore indipendenza rispetto a quella stabilita mediante il livello di significatività 1-α.

Avendo calcolato il valore del Chi-Quadrato per n=(2-1)*(2-1)=1 (gradi di libertà) possiamo confrontare questo valore con la tabella del Chi-Quadrato a 1 grado di libertà con 1-α =0.95 (vedi tabella della distribuzione del Chi-Quadrato, Appendice A2):

Troviamo:

$$\chi^2\big|_{1-\alpha=0.95,n=1} = 3.841 = \chi^2\big|_{critico}$$



Siccome:

$$2{,}381 = \chi^2\big|_{osservato} < \chi^2\big|_{critico} = 3{,}841$$

L'ipotesi nulla H_0 è accettata: possiamo affermare che con una "probabilità del 95%" le due mutabili statistiche considerate sono indipendenti e quindi che **il siero non è efficace per la guarigione**.

05 - Distribuzioni Bivariate tra una variabile ed una mutabile

Consideriamo ora una distribuzione bivariata costituita da una variabile statistica e da una mutabile statistica. Anche per questo tipo di distribuzioni si parla di connessione tra i dati statistici acquisiti. Si supponga che la X sia la mutabile con modalità (attributi) $x_{1,}..,x_r$ (righe) e che la Y rappresenti la variabile di valori $y_1,\ldots,y_c$

Le distribuzioni semplici che possono essere ricavate da una distribuzione doppia come la precedente sono r (una per ogni riga) + c (una per ogni colonna) + 2 (la distribuzione marginale delle X e la distribuzione marginale delle Y). Per le distribuzioni della Y condizionate alla X è possibile ottenere una media semplice condizionata e la media semplice marginale secondo le seguenti formule:

Media Semplice Condizionata Y|X

$$M_{Y|x_i} = \frac{\sum_{j=1}^{c} y_j \cdot n_{ij}}{\sum_{j=1}^{c} n_{ij}} = \frac{\sum_{j=1}^{c} y_j \cdot n_{ij}}{n_{i0}} \qquad i = 1,\ldots,r$$

Media Y Semplice Marginale

$$M_Y = \frac{\sum_{j=1}^{c} y_j \cdot n_{0j}}{\sum_{j=1}^{c} n_{0j}} = \frac{\sum_{j=1}^{c} y_j \cdot n_{0j}}{n}$$

Per stabilire il grado di connessione tra la variabile statistica (v.s.) Y e la mutabile statistica (m.s.) X si introduce l'indice di connessione η di Pearson:

Indice $\eta_{Y|X}$ di Pearson:

$$\eta_{Y|X} = \sqrt{\frac{\sum_{i=1}^{r} \left(M_{Y|x_i} - M_Y\right)^2 \cdot n_{i0}}{\sum_{j=1}^{c} \left(y_j - M_Y\right)^2 \cdot n_{0j}}}$$

L'indice $\eta_{Y|X}$ di Pearson assume valori compresi tra 0 e 1: $0 \le \eta \le 1$.

$\eta_{Y|X} = 0$ nel caso di connessione nulla (perfetta indipendenza in media di Y da X)

$\eta_{Y|X} = 1$ nel caso di connessione massima (perfetta dipendenza di Y da X)

Esempio 7 – Connessione tra una mutabile (X) ed una variabile (Y)

Su un gruppo di 1000 persone è stata condotta un'indagine per accertare l'esistenza di una qualche relazione tra il grado di istruzione di ciascuna persona e l'entità della popolazione residente nella città dalla quale proviene la persona intervistata

Frequenze Rilevate			POPOLAZIONE (MIGLIAIA)				
			Y_1	Y_2	Y_3	Y_4	
			50	75	150	300	n_{i0}
Grado Istruzione	X_1	Elem.	10	20	30	40	100
	X_2	Medie	100	40	190	70	400
	X_3	Sup.	50	20	40	90	200
	X_4	Univ.	90	70	40	100	300
		n_{0j}	250	150	300	300	1000

Riportiamo i valori delle medie semplici condizionate

$$M_{Y|x_1} = 185$$
$$M_{Y|x_2} = 143{,}75$$
$$M_{Y|x_3} = 185$$
$$M_{Y|x_4} = 152{,}5$$

Il valore della media semplice marginale: $M_Y = 158{,}75$

Ed il valore dell'indice η di pearson: $\eta_{Y|X} = 0{,}175$

Esempio 8 – Indipendenza tra una mutabile (X) ed una variabile (Y)

Dall'esempio 7 con dati diversi:

Frequenze Rilevate			POPOLAZIONE (MIGLIAIA)				
			Y_1	Y_2	Y_3	Y_4	
			50	75	150	300	n_{i0}
Grado Istruzione	X_1	Elem.	1	2	3	4	10
	X_2	Medie	2	4	6	8	20
	X_3	Sup.	3	6	9	12	30
	X_4	Univ.	4	8	12	16	40
		n_{0j}	10	20	30	40	100

Riportiamo i valori delle medie semplici condizionate

$$M_{Y|x_1} = 185$$
$$M_{Y|x_2} = 185$$
$$M_{Y|x_3} = 185$$
$$M_{Y|x_4} = 185$$

Il valore della media semplice marginale $M_Y = 185$

Ed il valore dell'indice η di pearson: $\eta_{Y|X} = 0$

Esempio9 – Perfetta Connessione tra una mutabile (X) ed una variabile (Y)

Dall'esempio 7 con dati diversi:

Frequenze Rilevate			POPOLAZIONE (MIGLIAIA)				
			Y_1	Y_2	Y_3	Y_4	
			50	75	150	300	n_{i0}
Grado Istruzione	X_1	Elem.	1	0	0	0	1
	X_2	Medie	0	1	0	0	1
	X_3	Sup.	0	0	1	0	1
	X_4	Univ.	0	0	0	1	1
		n_{0j}	1	1	1	1	4

Riportiamo i valori delle medie semplici condizionate

$$M_{Y|x_1} = 50$$
$$M_{Y|x_2} = 75$$
$$M_{Y|x_3} = 150$$
$$M_{Y|x_4} = 300$$

Il valore della media semplice marginale: $M_Y = 143{,}75$

Ed il valore dell'indice η di pearson . $\eta_{Y|X} = 1$

06 - Distribuzioni Bivariate tra due variabili quantitative

Consideriamo infine una distribuzione bivariata costituita da due variabili statistiche.

Possiamo definire, rispetto al solito schema, le seguenti medie parziali (essendo X e Y variabili statistiche, tutte le modalità ad esse relative sono quantitative).

Media Condizionata (o parziale) x_j

$$\overline{x_j} = M_{X|y_j} = \frac{\sum_{i=1}^{r} x_i \cdot n_{ij}}{n_{0j}} \qquad j = 1,\ldots,c$$

Media Condizionata (o parziale) y_i

$$\overline{y_i} = M_{Y|x_i} = \frac{\sum_{j=1}^{c} y_j \cdot n_{ij}}{n_{i0}} \qquad i = 1,\ldots,r$$

Media Marginale X

$$\overline{X} = \frac{\sum_{i=1}^{r} x_i \cdot n_{i0}}{n}$$

Media Marginale Y

$$\overline{Y} = \frac{\sum_{j=1}^{c} y_j \cdot n_{0j}}{n}$$

La dipendenza tra le due variabili aleatorie X ed Y viene studiata, in prima istanza, attraverso l'indice η di Pearson:

$$\eta_{Y|X} = \sqrt{\frac{\sum_{i=1}^{r} \left(M_{Y|x_i} - M_Y\right)^2 \cdot n_{i0}}{\sum_{j=1}^{c} \left(y_j - M_Y\right)^2 \cdot n_{0j}}}$$

Indica la dipendenza di Y da X

Mentre:

$$\eta_{X|Y} = \sqrt{\frac{\sum_{j=1}^{c}\left(M_{X|y_j} - M_X\right)^2 \cdot n_{0j}}{\sum_{i=1}^{r}\left(x_i - M_X\right)^2 \cdot n_{i0}}}$$

Indica la dipendenza di X da Y

Nota: Se $\eta_{X|Y} \neq \eta_{Y|X}$ significa che la rilevazione non è simmetrica

Esempio 10

		Y_1	Y_2	Y_3	Y_4	
		2	**4**	**6**	**8**	n_{i0}
X_1	3	12	7	2	3	24
X_2	5	2	3	6	4	15
X_3	7	23	16	8	14	61
	n_{0j}	37	26	16	21	100

$$M_{X|y_1} = 5{,}59 \qquad M_{Y|x_1} = 3{,}67 \qquad M_X = 5{,}74$$
$$M_{X|y_2} = 5{,}69 \qquad M_{Y|x_2} = 5{,}60 \qquad M_Y = 4{,}42$$
$$M_{X|y_3} = 5{,}75 \qquad M_{Y|x_3} = 4{,}43$$
$$M_{X|y_3} = 6{,}05$$

$$\eta_{X|Y} = 0{,}100 \qquad \eta_{Y|X} = 0{,}255$$

Appendice A1

Teorema: La somma delle contingenze su una riga o su una colonna è uguale a zero

Dim.: Consideriamo una colonna della tabella delle contingenze: sommando gli elementi della colonna j-esima:

$$\sum_{i=1}^{r} C_{ij} = \sum_{i=1}^{r}\left(n_{ij} - \frac{n_{i0} \cdot n_{0j}}{n} \right) = n_{0j} - n_{0j} \frac{1}{n} \sum_{i=1}^{r} n_{i0} = 0$$

Poiché: $\sum_{i=1}^{r} n_{ij} = n_{0j}$ $\sum_{i=1}^{r} n_{i0} = n$

Consideriamo, infine, una riga della tabella delle contingenze: sommando gli elementi della i-esima riga:

$$\sum_{j=1}^{c} C_{ij} = \sum_{j=1}^{c}\left(n_{ij} - \frac{n_{i0} \cdot n_{0j}}{n} \right) = n_{i0} - n_{i0} \sum_{j=1}^{c} \frac{n_{0j}}{n} = 0$$

Poiché: $\sum_{j=1}^{c} n_{ij} = n_{i0}$ $\sum_{j=1}^{c} n_{0j} = n$

Appendice A2 - Tabella Chi-Quadrato

n\ (1-α)	0.001	0.002	0.005	0.01	0.02	0.05	0.1	0.2	0.5	0.75	0.8	0.9	0.95	0.98	0.99	0.995	0.998	0.999
1	0.000	0.000	0.000	0.000	0.001	0.004	0.016	0.064	0.455	1.323	1.642	2.706	3.841	5.412	6.635	7.879	9.550	10.828
2	0.002	0.004	0.010	0.020	0.040	0.103	0.211	0.446	1.386	2.773	3.219	4.605	5.991	7.824	9.210	10.597	12.429	13.816
3	0.024	0.039	0.072	0.115	0.185	0.352	0.584	1.005	2.366	4.108	4.642	6.251	7.815	9.837	11.345	12.838	14.796	16.266
4	0.091	0.129	0.207	0.297	0.429	0.711	1.064	1.649	3.357	5.385	5.989	7.779	9.488	11.668	13.277	14.860	16.924	18.467
5	0.210	0.280	0.412	0.554	0.752	1.145	1.610	2.343	4.351	6.626	7.289	9.236	11.070	13.388	15.086	16.750	18.907	20.515
6	0.381	0.486	0.676	0.872	1.134	1.635	2.204	3.070	5.348	7.841	8.558	10.645	12.592	15.033	16.812	18.548	20.791	22.458
7	0.598	0.741	0.989	1.239	1.564	2.167	2.833	3.822	6.346	9.037	9.803	12.017	14.067	16.622	18.475	20.278	22.601	24.322
8	0.857	1.038	1.344	1.646	2.032	2.733	3.490	4.594	7.344	10.219	11.030	13.362	15.507	18.168	20.090	21.955	24.352	26.124
9	1.152	1.370	1.735	2.088	2.532	3.325	4.168	5.380	8.343	11.389	12.242	14.684	16.919	19.679	21.666	23.589	26.056	27.877
10	1.479	1.734	2.156	2.558	3.059	3.940	4.865	6.179	9.342	12.549	13.442	15.987	18.307	21.161	23.209	25.188	27.722	29.588
11	1.834	2.126	2.603	3.053	3.609	4.575	5.578	6.989	10.341	13.701	14.631	17.275	19.675	22.618	24.725	26.757	29.354	31.264
12	2.214	2.543	3.074	3.571	4.178	5.226	6.304	7.807	11.340	14.845	15.812	18.549	21.026	24.054	26.217	28.300	30.957	32.909
13	2.617	2.982	3.565	4.107	4.765	5.892	7.042	8.634	12.340	15.984	16.985	19.812	22.362	25.472	27.688	29.819	32.535	34.528
14	3.041	3.440	4.075	4.660	5.368	6.571	7.790	9.467	13.339	17.117	18.151	21.064	23.685	26.873	29.141	31.319	34.091	36.123
15	3.483	3.916	4.601	5.229	5.985	7.261	8.547	10.307	14.339	18.245	19.311	22.307	24.996	28.259	30.578	32.801	35.628	37.697
16	3.942	4.408	5.142	5.812	6.614	7.962	9.312	11.152	15.338	19.369	20.465	23.542	26.296	29.633	32.000	34.267	37.146	39.252
17	4.416	4.915	5.697	6.408	7.255	8.672	10.085	12.002	16.338	20.489	21.615	24.769	27.587	30.995	33.409	35.718	38.648	40.790
18	4.905	5.436	6.265	7.015	7.906	9.390	10.865	12.857	17.338	21.605	22.760	25.989	28.869	32.346	34.805	37.156	40.136	42.312
19	5.407	5.969	6.844	7.633	8.567	10.117	11.651	13.716	18.338	22.718	23.900	27.204	30.144	33.687	36.191	38.582	41.610	43.820
20	5.921	6.514	7.434	8.260	9.237	10.851	12.443	14.578	19.337	23.828	25.038	28.412	31.410	35.020	37.566	39.997	43.072	45.315
21	6.447	7.070	8.034	8.897	9.915	11.591	13.240	15.445	20.337	24.935	26.171	29.615	32.671	36.343	38.932	41.401	44.522	46.797
22	6.983	7.636	8.643	9.542	10.600	12.338	14.041	16.314	21.337	26.039	27.301	30.813	33.924	37.659	40.289	42.796	45.962	48.268
23	7.529	8.212	9.260	10.196	11.293	13.091	14.848	17.187	22.337	27.141	28.429	32.007	35.172	38.968	41.638	44.181	47.391	49.728
24	8.085	8.796	9.886	10.856	11.992	13.848	15.659	18.062	23.337	28.241	29.553	33.196	36.415	40.270	42.980	45.559	48.812	51.179
25	8.649	9.389	10.520	11.524	12.697	14.611	16.473	18.940	24.337	29.339	30.675	34.382	37.652	41.566	44.314	46.928	50.223	52.620
26	9.222	9.989	11.160	12.198	13.409	15.379	17.292	19.820	25.336	30.435	31.795	35.563	38.885	42.856	45.642	48.290	51.627	54.052
27	9.803	10.597	11.808	12.879	14.125	16.151	18.114	20.703	26.336	31.528	32.912	36.741	40.113	44.140	46.963	49.645	53.023	55.476
28	10.391	11.212	12.461	13.565	14.847	16.928	18.939	21.588	27.336	32.620	34.027	37.916	41.337	45.419	48.278	50.993	54.411	56.892
29	10.986	11.833	13.121	14.256	15.574	17.708	19.768	22.475	28.336	33.711	35.139	39.087	42.557	46.693	49.588	52.336	55.792	58.301
30	11.588	12.461	13.787	14.953	16.306	18.493	20.599	23.364	29.336	34.800	36.250	40.256	43.773	47.962	50.892	53.672	57.167	59.703
35	14.688	15.686	17.192	18.509	20.027	22.465	24.797	27.836	34.336	40.223	41.778	46.059	49.802	54.244	57.342	60.275	63.955	66.619
40	17.916	19.032	20.707	22.164	23.838	26.509	29.051	32.345	39.335	45.616	47.269	51.805	55.758	60.436	63.691	66.766	70.618	73.402
45	21.251	22.477	24.311	25.901	27.720	30.612	33.350	36.884	44.335	50.985	52.729	57.505	61.656	66.555	69.957	73.166	77.179	80.077
50	24.674	26.006	27.991	29.707	31.664	34.764	37.689	41.449	49.335	56.334	58.164	63.167	67.505	72.613	76.154	79.490	83.657	86.661