Piano Nazionale Lauree Scientifiche (PLS 2016-2017)

Statistica Laboratorio di Statistica

UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

DIPARTIMENTO DI SCIENZE ECONOMICHE E METODI MATEMATICI

Le relazioni tra variabili. L'indipendenza

prof.ssa Angela Maria D'Uggento angelamaria.duggento@uniba.it

materiale didattico II incontro

La distribuzione di una variabile qualitativa: le tabelle di contingenza

Le rappresentazioni più opportune per due variabili qualitative sono le tabelle di contingenza; graficamente si ricorre ai diagrammi a barre (o istogrammi).

Nelle tabelle di contingenza (X,Y) ad ogni valore *i* della variabile X è associato il numero n_i delle volte in cui tale valore si riscontra nelle n osservazioni oppure la sua frequenza relativa (n_i /n); ad ogni valore *j* della variabile Y è associato il numero n_j delle volte in cui tale valore si riscontra nelle n osservazioni oppure la sua frequenza relativa (n_j /n); l'osservazione che possiede contemporaneamente la modalità i e la modalità j è associata alla frequenza n_{ij}.

Per rappresentare graficamente il diagramma a barre si pongono sull'asse delle ascisse i valori assunti dalla variabile e in ordinata i conteggi o le frequenze; nel caso di due o più variabili si possono affiancare le colonne.

La tabella di contingenza

	Disciplina (Y)		
Genere (X)	Statistica	Ingegneria	Totale riga
Maschi	1 ($n_{1;1}$)	3 ($n_{1;2}$)	4 ($N_{1;0}$)
Femmine	5	1	6 ($N_{2;0}$)
Totale colonna	6 ($N_{0;1}$)	4 ($N_{0;2}$)	10 (N)

Distribuzioni marginali di X (Totale riga)e di Y (Totale colonna)

L'Indipendenza, ovvero alla ricerca di una relazione tra due variabili qualitative

Con un test statistico è possibile verificare se esiste una relazione (associazione) tra variabili qualitative (nominali o ordinali) o confrontare la significatività della differenza tra due percentuali.

Tale test consente anche di misurare la «distanza» tra quello che ci si aspettava e la realtà osservata attraverso i dati e di capire se questa «differenza» è maggiore di quella che potrebbe verificarsi per il solo effetto del caso.

Esempio: nella tabella precedente verificare l'associazione tra genere e scelta del corso universitario.

Domanda: la scelta del corso universitario è influenzata/dipende dal genere o è indipendente?

Per rispondere alla domanda bisogna

1. costruire la tabella doppia
2. testare la dipendenza tra le due variabili, utilizzando il test del Chi quadrato (χ^2).

Costruire la tabella di frequenza doppia

Soggetti	Genere	Corso
1	Maschio	Statistica
2	Femmina	Statistica
3	Femmina	Statistica
4	Femmina	Statistica
5	Maschio	Ingegneria
6	Femmina	Statistica
7	Femmina	Statistica
8	Maschio	Ingegneria
9	Femmina	Ingegneria
10	Maschio	Ingegneria

Tabella di frequenza doppia

	Statistica	Ingegneria	Totale riga
Maschi	1	3	4
Femmine	5	1	6
Tot colonna	6	4	**10**

Informazioni della tabella di frequenza doppia:

Esprime la relazione tra 2 variabili, in questo caso entrambe qualitative.

I numeri all'interno di ciascuna cella indicano le frequenze osservate (f_o) sul campione o collettivo e corrispondono al «conteggio» delle unità statistiche che presentano contemporaneamente le due modalità.

Frequenze attese

Per verificare l'eventuale dipendenza è necessario calcolare la frequenza attesa (f_e; *expected frequency*) per ciascuna cella (frequenza osservata).

Con riferimento alla precedente tabella, la domanda potrebbe essere: «Sapendo che gli iscritti a Statistica sono 6 e che i maschi sono 4 in totale, quanti maschi "mi aspetto" che si iscrivano a Statistica? E quante femmine? E ad Ingegneria?»

Tabella doppia con f_o ed f_e

	Statistica	Ingegneria	Totale riga
Maschi	1 (f_e=6/10*4=2,4)	3 (fe=4/10*4=1,6)	4
Femmine	5 (f_e=6/10*6=3,6)	1 (f_e=4/10*6=2,4)	6
Totale colonna	6	4	10

La somma delle frequenze attese per riga deve essere uguale al totale di riga e si ottengono moltiplicando il peso teorico (tot col/N) per la freq. osservata totale (tot riga).Esempio fe=2,4 deriva da (6*4)/10.
La somma delle frequenze attese per colonna deve essere uguale al totale di colonna.
La somma di tutte le frequenze attese deve essere uguale a N.

Indice di dipendenza χ^2

Dopo aver calcolato le f_e si può calcolare il valore di χ^2

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$

- Il campo di variazione del χ^2 è $(0 < \chi^2 < \infty)$.

- se tutte le f_o e le f_e sono uguali, il valore del χ^2 sarà uguale a 0, dunque le due variabili sono **indipendenti**. Se almeno una è diversa, allora son dipendenti.
- ne consegue che, maggiore è la differenza tra f_o ed f_e, e maggiore sarà il valore del χ^2.

Calcolo del χ^2

	Statistica	Ingegneria	Totale riga
Maschi	1 f_e=2,4	3 f_e=1,6	4
Femmine	5 f_e=3,6	1 f_e=2,4	6
Tot colonna	6	4	10

$$\chi^2 = \frac{(1-2,4)^2}{2,4} + \frac{(3-1,6)^2}{1,6} + \frac{(5-3,6)^2}{3,6} + \frac{(1-2,4)^2}{2,4} =$$

$$\chi^2 = 0,82 + 1,23 + 0,54 + 0,82 = 3,41$$

Caratteristiche del χ^2

- L'indice χ^2 è sempre positivo
- Può assumere valori che variano tra 0 (massima indipendenza) e ∞ (con N = massima dipendenza solo nel caso di tabelle 2x2)
- Nelle tabelle 2x2, il χ^2 risulta più facilmente interpretabile con una «regola empirica» per la quale è necessario calcolare il χ^2 **relativo** ossia χ^2/N, che varia tra 0 ed 1, e si interpreta come segue:

Regola empirica

X^2 /N compreso tra 0 e 0,5	X^2 /N compreso tra 0,5 e 1
Indipendenza tra le variabili	**Dipendenza tra le variabili**

Calcolo del χ^2 relativo: interpretazione

$\chi^2/N = 3,41/10=0,34$

Interpretazione: le due variabili non risultano associate, ossia il genere e la scelta del corso di studi universitario sono indipendenti (o più precisamente, la scelta universitaria non dipende dal genere).

Esercizio 1

Verificare l'esistenza di una relazione tre le variabili "genere" ed "esito dell'esame di Statistica".

Soggetti	Genere	Esito
1	Maschio	Bocciato
2	Maschio	Bocciato
3	Maschio	Promosso
4	Femmina	Promosso
5	Femmina	Promosso
6	Femmina	Promosso
7	Femmina	Bocciato
8	Femmina	Bocciato
9	Femmina	Bocciato
10	Femmina	Bocciato
11	Femmina	Bocciato
12	Femmina	Bocciato

Caso di massima indipendenza

	Promozione	Bocciatura	Totale riga
Maschi	1 f_e=1	2 f_e=2	3
Femmine	3 f_e=3	6 f_e=6	9
Tot colonna	4	8	12

$$\chi^2 = \frac{(1-1)^2}{1} + \frac{(2-2)^2}{2} + \frac{(3-3)^2}{3} + \frac{(6-6)^2}{6} = 0$$

Esercizio 2

Verificare l'esistenza di una relazione tra le variabili "genere" e "professione".

Soggetti	Genere	Professione
1	Maschio	Meccanico
2	Maschio	Meccanico
3	Maschio	Meccanico
4	Maschio	Meccanico
5	Maschio	Meccanico
6	Maschio	Meccanico
7	Maschio	Estetista
8	Femmina	Estetista
9	Femmina	Estetista
10	Femmina	Estetista
11	Femmina	Estetista
12	Femmina	Estetista

Caso di dipendenza

	Estetista	Meccanico	Totale riga
Maschi	1 f_e=3,5	6 f_e=3,5	7
Femmine	5 f_e=2,5	0 f_e=2,5	5
Tot colonna	6	6	12

$$\chi^2 = \frac{(1-3{,}5)^2}{3{,}5} + \frac{(6-3{,}5)^2}{3{,}5} + \frac{(5-2{,}5)^2}{2{,}5} + \frac{(0-2{,}5)^2}{2{,}5} = 1{,}79 + 1{,}79 + 2{,}5 + 2{,}5 = 8{,}58$$

8,58/12=0,72 quindi?

Tra le due variabili c'è dipendenza (la professione dipende dal genere): i maschi scelgono di fare il meccanico, mentre le femmine di fare l'estetista.

Una applicazione di χ^2 in ambito medico: confronto «atteso» vs «osservato»

In un ospedale americano, alcuni pazienti dializzati si sono sottoposti ad una sperimentazione per valutare l'ipotesi che l'assunzione di aspirina potesse inibire la formazione di trombi. I 44 pazienti furono assegnati a caso ad un gruppo a cui era somministrata l'aspirina o ad un altro gruppo trattato con il placebo. Per evitare autosuggestione nei pazienti e nei ricercatori lo studio fu condotto in doppio cieco, ossia nessuno dei partecipanti sapeva se i pazienti venivano trattati con aspirina o con placebo. La ricerca continuò finché 24 dei 44 pazienti ebbero sviluppato dei trombi e, a questo punto, i ricercatori identificarono le compresse somministrate e valutarono statisticamente i risultati delle differenti terapie: 19 pazienti erano stati trattati con aspirina e 25 con il placebo. I due gruppi non presentavano differenze significative per le variabili età, sesso, tempo in dialisi ed altro.Riepiloghiamo i risultati della sperimentazione:

nel gruppo trattato con aspirina (n_1=19) 6 pazienti avevano sviluppato trombi

nel gruppo trattato con placebo (n_2=25) 18 pazienti avevano sviluppato trombi

Esprimendo tali dati in termini relativi avremo

6/19=0,32 e 18/25=0,72.

La differenza tra 0,32 e 0,72

è statisticamente significativa?

Gruppi	Presenza trombi	Assenza trombi	**Totali**
Placebo	**18**	**7**	**25**
Aspirina	**6**	**13**	**19**
	24	**20**	**44**

Procediamo nella costruzione della tabella teorica ipotizzando che l'aspirina NON abbia un effetto inibitorio sulla formazione di trombi.

Su 44 pazienti, 24 hanno sviluppato trombi (il 55%) mentre 20 no (il 45%).

Quindi, se l'aspirina producesse gli stessi effetti del placebo, dovremmo assistere alla formazione di trombi nel 55% dei casi e nel 45% no.

In altri termini, 24 pazienti (cioè il 55% di 44) hanno la probabilità di sviluppare trombi indipendentemente dal gruppo a cui appartengono.

Tabella Frequenze teoriche

Gruppi	Presenza trombi	Assenza trombi	Totali
Placebo	55% di 25= 13,64	45% di 25= 11,36	25
Aspirina	55% di 19= 10,36	45% di 19= 8,64	19
	24	20	44

La v.c. Chi misura quanto le differenze osservate differiscono da quelle attese o teoriche in ogni casella. Calcolando

$$\chi^2 = \sum \frac{(\text{Freq. osservate} - \text{Freq. attese})^2}{\text{Freq. attese}} = \frac{(18-13{,}64)^2}{13{,}64} + \frac{(7-11{,}36)^2}{11{,}36} + \frac{(6-10{,}36)^2}{10{,}36} + \frac{(13-8{,}64)^2}{8{,}64} = 7{,}10$$

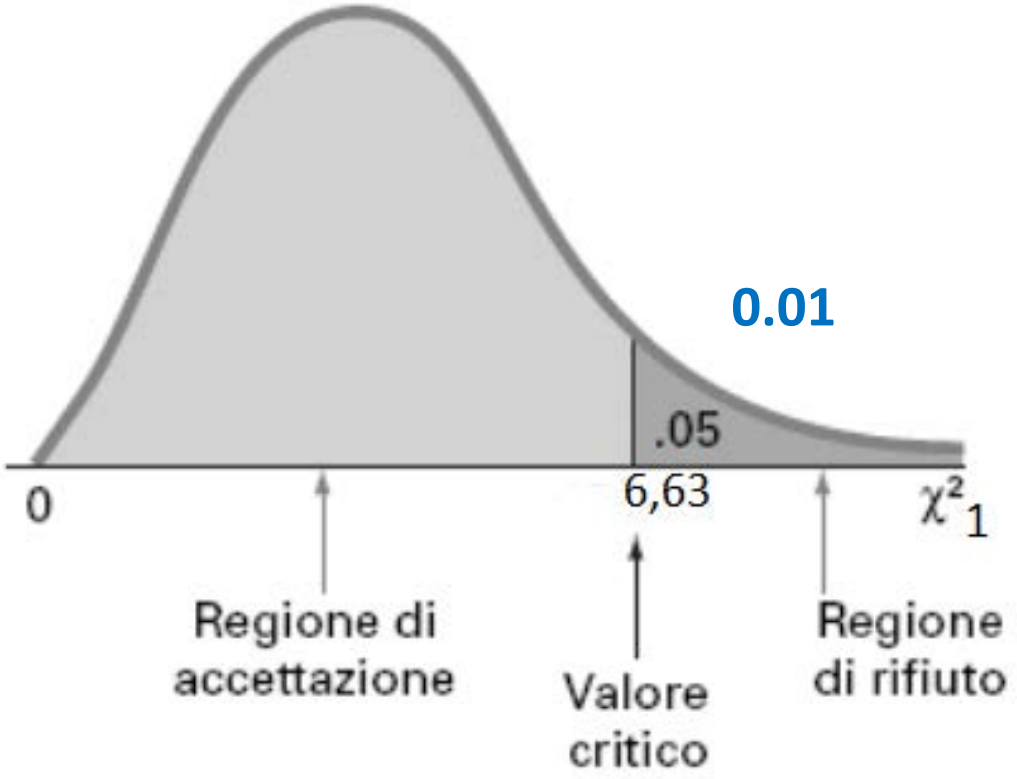
gradi di libertà = (righe–1)(colonne-1)=(2-1)(2-1)=1.

Nella tavola apposita, il valore soglia in corrispondenza dei g.d.l.=1 e α=0,01 è χ2=6,635.

Quindi, rifiutiamo H_0 e concludiamo che l'aspirina riduce i trombi nel 99,9% dei casi

La verifica di ipotesi*

Regola di decisione: rifiutare H_0 se il valore osservato della statistica χ2 è maggiore del valore critico; altrimenti, accettare H_0 (indipendenza tra variabili)



Nel nostro esempio :
Valore test
7,10 > 6,63 quindi
ACCETTIAMO ipotesi alternativa con α=0,01 (H_1 : dipendenza)

La verifica di ipotesi è più ampiamente trattata nelle dispense della Prof. Ribecco, alle quali si rimanda

Alcune tabelle per esercitazione

L’analisi della dipendenza o regressione

Investimento pubblicitario (migliaia di euro)	Durata spot (secondi)
100	257
102	264
103	274
101	266
105	277
100	263
99	258
105	275

L’analisi della indipendenza

Gruppi	Presenza trombi	Assenza trombi	Totale
Placebo	18	7	25
Aspirina	6	13	19
Totale	24	20	44