Lezione 6

La correlazione lineare

Argomenti della lezione:

- **Variabili a livello degli intervalli**
- **Variabili dicotomiche e ordinali**

Relazione tra variabili

Studio dei legami che intercorrono tra più variabili rilevate contemporaneamente

- **il tipo ovvero la forma della relazione tra X e Y**
- **l'intensità ovvero la forza del legame tra X e Y**
- **la direzione della relazione ovvero il verso positivo o negativo**

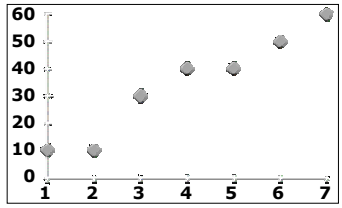
Forma della relazione tra due variabili

Rappresentazione grafica della distribuzione congiunta:

- **Diagramma di dispersione (Scatterplot)**

X = aggressività valutata dai compagni
Y = aggressività autovalutata

Sogg.	a	b	c	d	e	f	g
X	1	2	6	3	5	7	4
Y	10	10	50	30	40	60	40



Intensità e direzione della relazione

Coefficiente di correlazione lineare "r" di Pearson

$$r = \frac{\sum_{1}^{N} z_x z_y}{N}$$

Livello di misura: almeno intervalli equivalenti

Somma di prodotti di variabili standardizzate

$$Z_x = \frac{X - \bar{X}}{s_x} \qquad Z_y = \frac{Y - \bar{Y}}{s_y}$$

Concordanza tra X e Y:

- X maggiore della media e Y maggiore della media
- X minore della media e Y minore della media

Non concordanza tra X e Y:

- X maggiore della media e Y minore, X minore della media e Y maggiore

r di Pearson varia tra -1 e +1

-1 correlazione perfetta negativa

0 assenza di relazione lineare

+1 correlazione perfetta positiva

Valori intermedi: relazione negativa o positiva più o meno forte

Formule di calcolo:

Formula A)

$$r = \frac{\frac{\sum XY}{N} - \bar{X}\bar{Y}}{s_x s_y}$$

Relazione tra X e Y come rapporto tra covarianza e varianze di X e Y

Numeratore ⇒ Covarianza

$$\frac{\sum (X - \bar{X})(Y - \bar{Y})}{N}$$

Denominatore ⇒ $\sqrt{s_x^2 s_y^2}$

Coefficiente di Pearson ⇒ $\frac{Cov_{xy}}{\sqrt{s_x^2 s_y^2}}$

formula B)

utile per calcoli manuali

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]}\sqrt{\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

Verifica delle ipotesi sul coefficiente di r di Pearson

Distribuzione campionaria di ρ

H_0: $\rho = 0.00$
gdl = (n-2)

- **Ipotesi nulla e alternativa: H0: $\rho = 0$; H1: $\rho > 0$ (<, ≠)**
- **$\rho_{critico}$ per α =.05 e gdl (n-2)**
- **Calcolo di r sul campione**
- **Confronto valori, decisione**

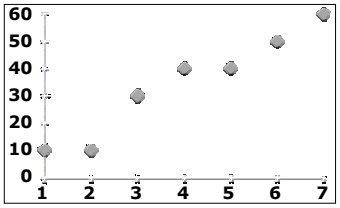
r di Pearson e t di Student:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

si distribuisce come t di Student con (n-2) gdl

X = aggressività valutata dai compagni
Y = aggressività autovalutata

Sogg.	a	b	c	d	e	f	g
X	1	2	6	3	5	7	4
Y	10	10	50	30	40	60	40



Sogg	X	Y	XY	X^2	Y^2
a	1	10	10	1	100
b	2	10	20	4	100
c	6	50	300	36	2500
d	3	30	90	9	900
e	5	40	200	25	1600
f	7	60	420	49	3600
g	4	40	160	16	1600
Tot	28	240	1200	140	10400

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left[\sum X^2 - \frac{(\sum X)^2}{N}\right]}\sqrt{\left[\sum Y^2 - \frac{(\sum Y)^2}{N}\right]}}$$

$$r = \frac{1200 - \frac{28*240}{7}}{\sqrt{(140 - \frac{28^2}{7})(10400 - \frac{240^2}{7})}}$$

$$= \frac{1200 - 960}{\sqrt{(140-112)(10400-8229)}}$$

r = 240/247 = 0.97

Verifica delle ipotesi:

- **H_0: $\rho = 0$ H_1: $\rho \neq 0$**
- **$\alpha = .05$, gdl = 5, $\rho_{critico} = .754$**
- **$r_{empirico} > \rho_{critico}$ respingiamo H_0**

Esiste una relazione significativa tra aggressività valutata dai compagni e autovalutata

Relazione tra due variabili dicotomiche

0 = risposta sbagliata;
1 = risposta giusta

Var	a	b	c	d	e	f	g	h	i
X	0	0	1	0	1	1	1	0	1
Y	0	0	1	0	0	1	1	1	0

Si costruisce una tabella a doppia entrata che incrocia le due variabili

		Dom Y		
		1	0	Tot
Dom X	1	f_a=3	f_b= 2	p=5
	0	f_c= 1	f_d= 4	q=5
		p'=4	q'=6	n=10

$$r_{phi} = \frac{f_a f_d - f_b f_c}{\sqrt{pp' qq'}}$$

$$r_{phi} = \frac{3*4 - 2*1}{\sqrt{5*5*4*6}} = \frac{12 - 2}{\sqrt{600}} = \frac{10}{24.5} = 0.41$$

Verifica delle ipotesi

$$r^2_{phi} = \frac{Chi^2}{n} \qquad n * r^2_{phi} = Chi^2$$

- **$chi^2_{critico}$ con 1 gdl**
- **se il chi^2 è significativo, anche r_{phi} è significativo**
- **confronto tra chi^2 empirico e critico, decisione**

Nel nostro esempio:

$n*r^2_{phi}$ = 0.17*10= 1.7

Chi^2 Critico 1 gdl = 3.84

Non si può rifiutare H_0

Non c'è una relazione significativa tra le due domande

Relazione tra una variabile dicotomica e una variabile continua

Coefficiente di correlazione "punto-biseriale"

$$r_{pb} = \frac{\bar{X}_a - \bar{X}_b}{s_x} * \sqrt{\frac{n_a}{n}\frac{n_b}{n}}$$

Domande in un test attitudinale cui si può rispondere giusto (Y=1) o errato (Y=0)

Verificare se l'item 3 va nello stesso "verso" del totale del test: quando un soggetto ha un punteggio alto nel test, risponde giusto alla D3 e viceversa

Sogg	D3 (Y)	Tot (X)
a	1	6
b	1	7
c	0	5
d	0	2
e	1	5
f	0	2
g	1	4
h	1	6
i	1	7
l	1	3
m	1	4

- Per i valori in cui D3 = 1 la media è: (6+7+5+4+6+7+3+4)/8 = 5.25
- Per i valori in cui D3 = 0 la media è: (5+2+2)/3=3.00

$\bar{X}_a$=5.25, $\bar{X}_b$=3, S_x=1.8, n_a=8, n_b=3

$$r_{pb} = \frac{\bar{X}_a - \bar{X}_b}{s_x} * \sqrt{\frac{n_a}{n}\frac{n_b}{n}} =$$

$$= \frac{5.25-3.00}{1.80}\sqrt{\frac{8*3}{11*11}} = 0.55$$

Verifica delle ipotesi:

si trasforma r_{pb} in t di Student

$$t = r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}}$$

Si confronta questa t con la t critica per α prescelto, gdl=(n-2), e H_1 mono o bi-direzionale

- H_0: $r_{pb} = 0$ H_1: $r_{pb} > 0$
- trasformiamo di r_{pb} in t:

$$t = r_{pb}\sqrt{\frac{n-2}{1-r_{pb}^2}} \qquad t = 0.55\sqrt{\frac{11-2}{1-0.55^2}} = 2.08$$

- $t_{critico}$ con 9 gdl per α =0.05 (monodirezionale) = 1.833
- decisione: t calcolato > $t_{critico}$ Rifiutiamo l'ipotesi nulla. C'è relazione tra le due variabili

Relazione tra variabili su scala ordinale

Due variabili misurate su scala ordinale, oppure una su scala ordinale e l'altra su scala a intervalli o rapporti

Coefficiente r_s (o rho) di Spearman: correlazione tra ranghi

X = graduatoria in italiano
Y = graduatoria in matematica

Sogg	X	Y	d=(X-Y)
a	1	2	-1
b	3	4	-1
c	4	5	-1
d	2	3	-1
e	5	1	+4

C'è concordanza tra le due graduatorie?

Coefficiente di correlazione tra ranghi:

$$r_s = 1 - \frac{6\sum_1^n d^2}{n(n^2 - 1)}$$

$\Sigma d^2 = (1+1+1+1+16) = 20$

$$r_s = 1 - \frac{6x20}{5(5^2 - 1)} = 1 - \frac{120}{120} = 1 - 1 = 0$$

Non c'è relazione tra le due graduatorie

Verifica delle ipotesi:

Se ≤ 30: valori tabulati di r_S in funzione di α e di n

Se n > 30: si trasforma r_S in t di Student con (n-2) gdl

$$t = r_s \sqrt{\frac{n - 2}{1 - r_s^2}}$$

→ $H_0: \rho_s = 0.00$ $H_1: \rho_s > 0.00$

→ per $\alpha = 0.05$ e $n = 5$ $\rho_{scritico} = .90$

$$r_s = 1 - \frac{6*20}{5(5^2-1)} = 1 - \frac{120}{120} = 1-1 = 0$$

Non posso respingere H_0
$r_s = 0 < r_{scritico} = 0.90$

CONCLUSIONE

Livello di misura	Indice di correlazione
2 var. almeno intervalli eq.	r di Bravais-Pearson

CONCLUSIONE

Livello di misura	Indice di correlazione
2 var. ordinali 1 ordinale 1 int. eq.	rho di Spearman (r_s)

CONCLUSIONE

Livello di misura	Indice di correlazione
1 Var. dicotomica 1 Var. almeno int. eq.	r punto-biseriale

CONCLUSIONE

Livello di misura	Indice di correlazione
2 var. dicotomiche	r_{phi}