Lezione 30

13/12/24

TEST χ^2

- DI ADATTAMENTO
- DI INDIPENDENZA

H_0 : LA DISTRIBUZIONE COINCIDE CON QUELLA TEORICA

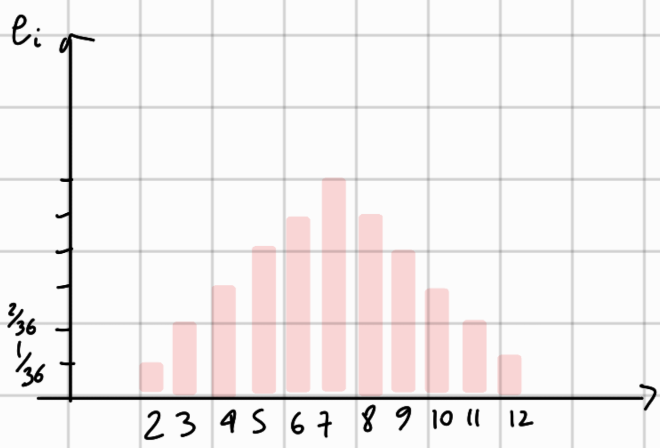
H_1 : LA DISTRIBUZIONE NON COINCIDE

Se abbiamo una distribuzione discreta o qualitativa

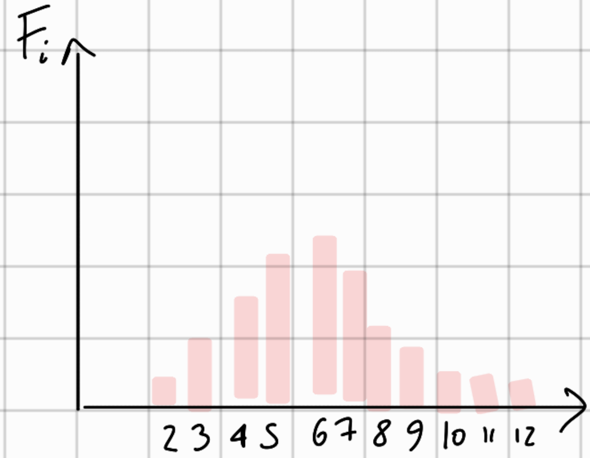
CONDIZIONE: $F_i \geq 5$ →

TESTIAMO SE UNA VARIABILE è VEROSIMILE CHE SIA STATA ESTRATTA SECONDO UNA DISTRIBUZIONE DI PROBABILITÀ NOTA.

ES. LANCIO DI DUE DADI $n = 250$ VOLTE



DISTRIBUZIONE DI PROBABILITÀ



DISTRIBUZIONE DI FREQUENZE

SE 10 LANCIO 2 DADI

2	3	4	5	6	7	8	9	10	11	12
9	15	22	28	39	43	34	25	17	13	5

$E_i = n e_i$ ← probabilità teorica

F_i

FREQUENZA ASSOLUTA ATTESA

FREQUENZE ASSOLUTE REALI

$\boxed{F_i \geq 5}$

DEVO RIPETERE L'ESPERIMENTO ABBASTANZA VOLTE DA AVERE CHE OGNI NUMERO È USCITO ALMENO 5 VOLTE.

$K = \#$ DI CLASSI

NELL'ESEMPIO $K = 11$

$$\boxed{\chi^{2*}_{K-1} = \sum_{i=1}^{m} \frac{(F_i - E_i)^2}{E_i}}$$

$$E_i = \begin{array}{cccccc} 2 & 3 & 4 & & & 12 \\ \frac{250}{36} & \frac{2\cdot 250}{36} & \cdot & \cdot & - \; - & \frac{1}{36}\cdot 250 \\ \| & \| & \| & & & \| \\ 6.94 & 13.89 & 20.83 & & & 6.94 \end{array}$$

$$\chi^2_{10} = \frac{(9-6.94)^2}{6.94} + \frac{(15-13.89)^2}{13.89} + \frac{(22-20.83)^2}{20.83} + \cdots + \frac{(5-6.94)^2}{6.94} =$$

$$= 2.93 < 15.987$$

$\Rightarrow$ ACCETTIAMO L'IPOTESI NULLA $\rightarrow$ I DADI CHE HO LANCIATO NON SONO TRUCCATI.

(ESERCIZIO 4 DEL 6 SETTEMBRE 2024)

$e_1 = \frac{5}{10}$ $\quad E_1 = \frac{5}{\cancel{10}} \cdot 10\cancel{0} = 50$ $\quad 10 = 5 + 3 + 2$

$e_2 = \frac{3}{10}$ $\quad E_2 = \frac{3}{10} \cdot 100 = 30$

$e_3 = \frac{2}{10}$ $\quad E_3 = \frac{2}{10} \cdot 100 = 20$

TEST χ^2 DI INDIPENDENZA (È UN CASO PARTICOLARE)

SUPPONIAMO CHE DI OGNI UNITÀ STATISTICA CONOSCIAMO DUE VARIABILI QUALITATIVE.

ESEMPIO HO UN FARMACO E LO TESTO CON 3 POSSIBILITÀ

$K = 3$ CATEGORIE

- DOSE DOPPIA
- DOSE SINGOLA
- PLACEBO

DI OGNI PERSONA VERIFICO SE È MIGLIORATO OPPURE NO

$h = 2$ CATEGORIE

TABELLA FREQUENZE

FREQUENZE MARGINALI

		D_D	D_S	P		
MIGLIORATO	SI	$M_{11}=29$	$M_{12}=5$	$M_{13}=29$	46	$= M_{1\cdot}$
	NO	90	80	114	284	$= M_{2\cdot}$
		119	85	126	330	
		$M_{\cdot 1}$	$M_{\cdot 2}$	$M_{\cdot 3}$		

NUMERO TOTALE n

SE LE DUE VARIABILI FOSSERO SLEGATE AVREMMO LE SEGUENTI FREQUENZE ATTESE: $\forall i = 1..k$ $\forall j = 1...h$

$$E_{i,j} = \frac{M_{i\cdot} \times M_{\cdot j}}{n}$$

$E_{1,1}$

ESEMPIO

$$E_{12} = \frac{M_{1\cdot} \times M_{\cdot 2}}{n} = \frac{46 \cdot 85}{330}$$

	D_D	D_S	P		
SI	$\frac{46 \cdot 119}{330} = 16.6$	$\frac{85 \cdot 46}{330} = 11.8$	$\frac{46 \cdot 126}{330} = 17.6$	46	$= M_{1\cdot}$
NO	102.4	73.2	108.4	284	$= M_{2\cdot}$
	119	85	126	330	n
	$M_{\cdot 1}$	$M_{\cdot 2}$	$M_{\cdot 3}$		

$$\chi^{2*}_{(k-1)(h-1)} = \sum_{i,j} \frac{(F_{ij} - E_{ij})^2}{E_{ij}} = \frac{(29-16.6)^2}{16.6} + \frac{(5-11.8)^2}{11.8} + \dots + \frac{(114-108.4)^2}{108.4}$$

$$\chi^{2*}_{2\cdot 1=2} = 17.4 > 13.816 \Rightarrow$$ RIFIUTO L'IPOTESI NULLA CHE LE DUE VARIABILI SIANO INDIPENDENTI

IL FARMACO INFLUISCE SULLA GUARIGIONE AL 99.9 %

DIAMO UN FARMAC CONTIAMO LE GUARIGIONI

FARMACO

TUTTI I NUMERI SONO ≥ 5 ok

GUARIGIONE

F_{ij}	SI	NO	
SI	52	28	80
NO	8	12	20
	60	40	100

$k = 2$

$h = 2$

E_{ij}	SI	NO	
SI	48	32	80
NO	12	8	20
	60	40	100

$$\nu = (2-1)(2-1) = 1$$

$$\chi^{2*}_{1} = \frac{(52-48)^2}{48} + \frac{(28-32)^2}{32} + \frac{(8-12)^2}{12} + \frac{(12-8)^2}{8} = 4.17$$

IL FARMAC FUNZIONA CON CONFIDENZA 95 %

STATISTICA BIVARIATA

Di ogni unità statistica misuriamo due variabili numeriche continue:
HO N COPPIE DI DATI

$$(x_1, y_1),\ (x_2, y_2),\ \dots,\ (x_N, y_N)$$

Solitamente UNA DELLE DUE è FACILE DA MISURARE e l'ALTRA NO
Se riusciamo a trovare un legame tra le due allora potremo usarlo

ESEMPIO [ALBERI]

	MEDIA	VARIANZA
x_i = diametro	$\bar{x}$,	σ_x^2
y_i = età	$\bar{y}$,	σ_y^2

Si definisce COVARIANZA tra X e Y la quantità

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$$

OSSERVO CHE

$\sigma_{xy} > 0$ X CRESCE Y CRESCE e VICEVERSA → VARIABILI CONCORDI

$\sigma_{xy} < 0$ X CRESCE Y DECRESCE e VICEVERSA → VARIABILI DISCORDI

$$\boxed{\overline{xy} = \frac{1}{N} \sum_{i=1}^{N} x_i y_i}$$

PROPOSIZIONE

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y}) = (\overline{xy}) - \bar{x} \cdot \bar{y}$$

DIM.

$$\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y}) = \frac{1}{N}\sum_{i=1}^{N}\left(x_iy_i-\bar{x}y_i-x_i\bar{y}+\bar{x}\bar{y}\right) =$$

$$= \frac{1}{N}\sum_{i=1}^{N}x_iy_i - \frac{1}{N}\sum_{i=1}^{N}\bar{x}y_i - \frac{1}{N}\sum_{i=1}^{N}x_i\bar{y} + \frac{1}{N}\sum_{i=1}^{N}\bar{x}\bar{y}$$

$$= (\overline{xy}) - \bar{x}\boxed{\frac{1}{N}\sum_{i=1}^{N}y_i} - \bar{y}\boxed{\frac{1}{N}\sum_{i=1}^{N}x_i} + \frac{1}{\cancel{N}}\cancel{N}\bar{x}\bar{y}$$

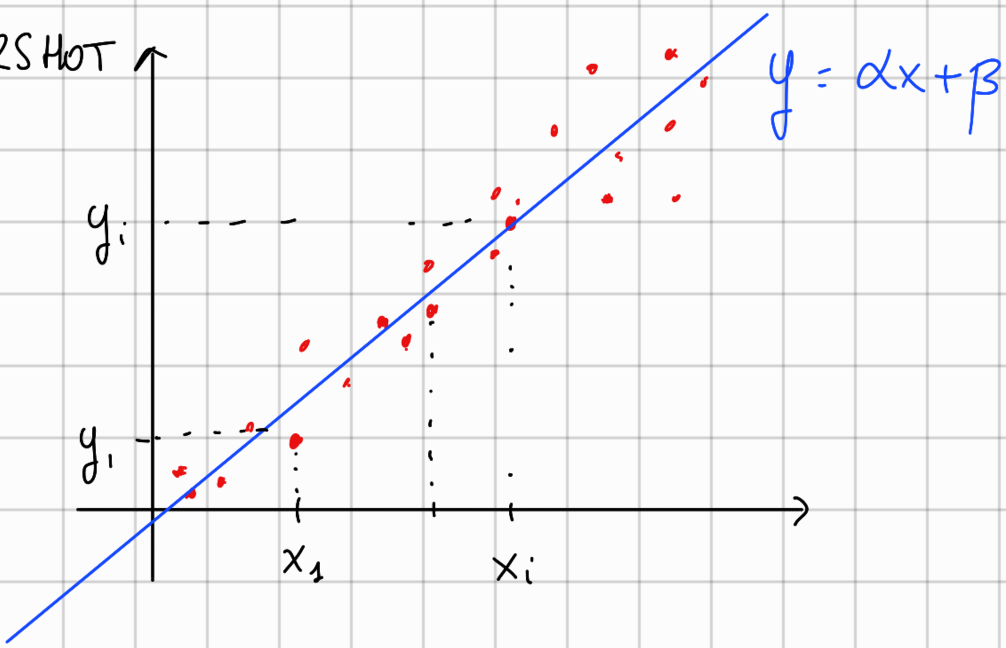
$$= (\overline{xy}) - \bar{x}\bar{y} \cancel{-\bar{x}\bar{y}} + \cancel{\bar{x}\bar{y}} \qquad \square$$

DEFINIZIONE : COEFFICIENTE DI CORRELAZIONE LINEARE

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y}$$ è UN NUMERO ADIMENSIONALE

$$\boxed{-1 \leq \rho_{xy} \leq 1}$$

SCATTERSHOT



LA RETTA CHE APPROSSIMA MEGLIO QUESTI PUNTI SI CHIAMA RETTA DI REGRESSIONE LINEARE.