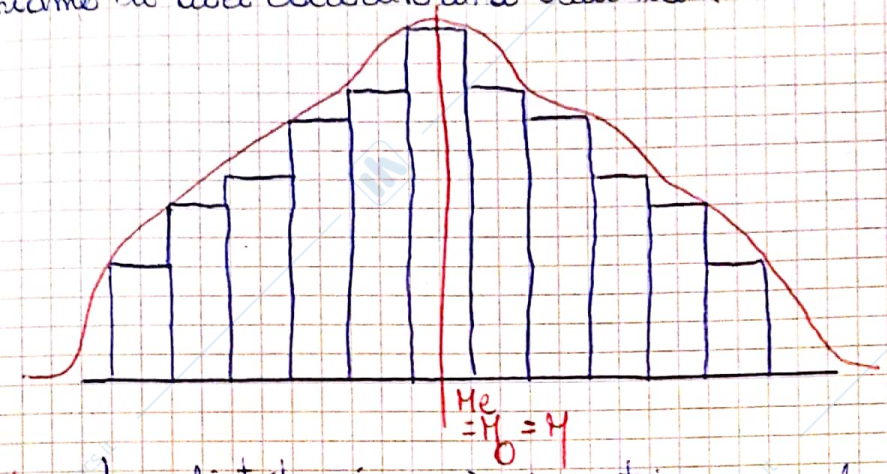
STATISTICA

Come descrivere una variabile statistica:

- La forma della distribuzione

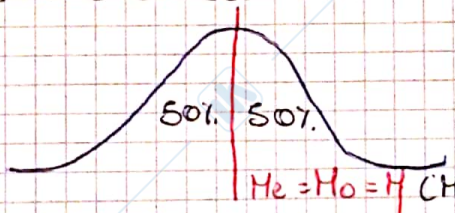
Immaginiamo di aver calcolato una variabile x e abbiamo questo istogramma:



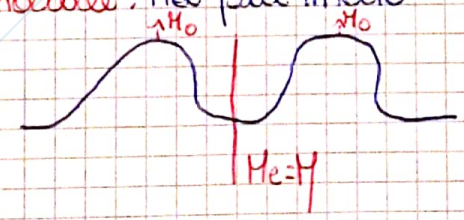
Simmetrica: La distribuzione è simmetrica quando dividendo la mia distribuzione a metà (in corrispondenza della mediana) le due metà saranno esattamente uguali.

Avremo così una variabile **Simmetrica unimodale** (= ha una sola moda, perché il valore più alto è solo uno)

FORMA TIPICA:

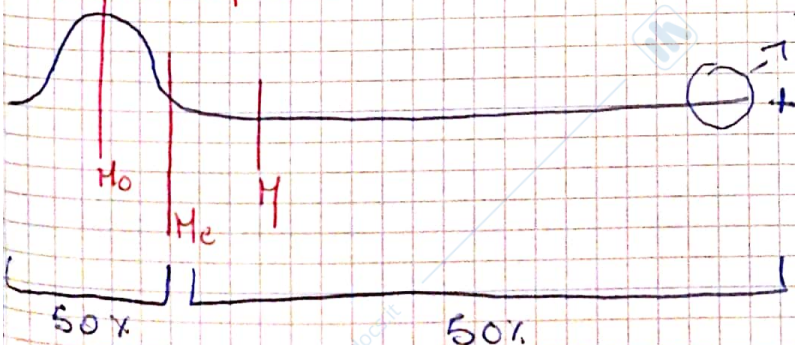


Simmetrica plurimodale: Ha più mode



Se la nostra variabile non è simmetrica è asimmetrica:

Asimmetrica positiva: Tende a $+\infty$ valori, avendo hanno valori anomali verso $+\infty$.

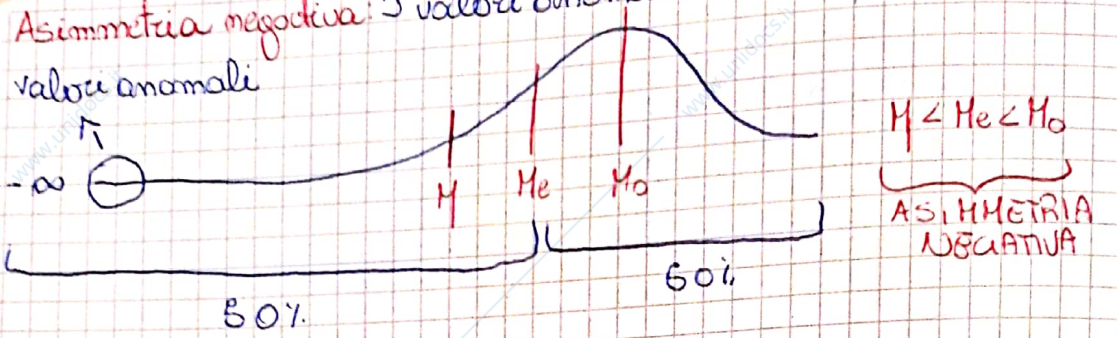


$M > Me > Mo$

ASIMMETRIA POSITIVA

Avendo questo non solo capisco che c'è asimmetria positiva, ma anche che c'è mancanza di robustezza e anche che ci sono valori anomali.

Asimmetria negativa: i valori anomali tendono verso $-\infty$.



In breve:
Quando abbiamo la simmetria unimodale avremo: $M = Me = Mo$
Quando abbiamo l'asimmetria positiva avremo: $M > Me > Mo$
Quando abbiamo l'asimmetria negativa avremo: $M < Me < Mo$
Quando abbiamo la simmetria plurimodale avremo: $M = Me$.

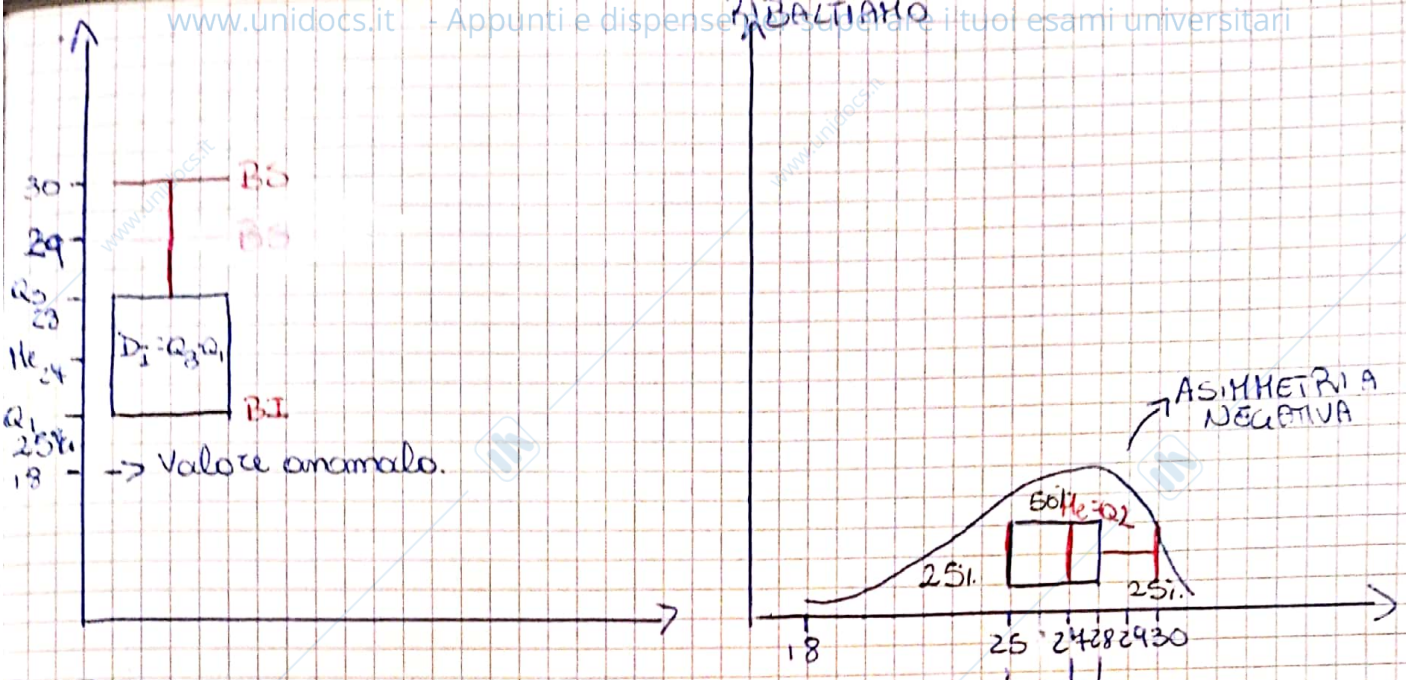
Variabilità:
Nel calcolo della varianza, gli scarti più elevati acquistano molta più importanza rispetto agli scarti più piccoli.

SQM = deviazione standard.

Box Plot

È un grafico che si utilizza per le variabili numeriche e serve per confrontare due variabili insieme. Nel box plot inseriamo tutti i dati studiati.
Esempio voti d'esame:

x	n	%	F	
18	1	5%	5%	
25	5	25%	30%	$= Q_1$
27	7	35%	65%	$= Me$
28	5	25%	90%	$= Q_3$
29	1	5%	95%	
30	1	5%	100%	
	20	100%		



Box Plot:

Baffo superiore = $BS = Q_3 + 1,5 \cdot D_I$

Baffo inferiore = $BI = Q_1 - 1,5 \cdot D_I$

} Ci daranno una banda, se le variabili superano questi valori avremo dei valori anomali.

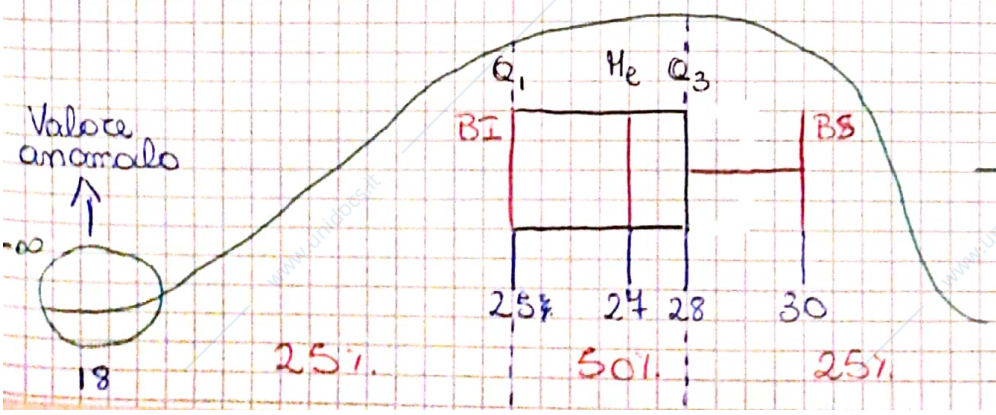
$D_I = Q_3 - Q_1 = 28 - 25 = 3.$

$BS = 28 + (1,5 \cdot 3) = 28 + 4,5 = 32,5.$ -> Il baffo lo disegniamo a 30

$BI = 25 + (1,5 \cdot 3) = 25 + 4,5 = 20,5$ -> Il valore più vicino interno sarà 25

Siccome non ho valori uguali a 32,5, prendo valori interni più vicini, quindi 30 (BS) e non avendo valori uguali 20,5, prendo i valori interni, quindi 25 (BI).

} Tutti i valori esterni a questi valori saranno valori anomali.
Non abbiamo variabili superiori, quindi esterni, a 32,5, ma abbiamo valori inferiori, quindi esterni, a 20,5, che sarebbe 18 (È oltre la banda inferiore)



ESERCITAZIONE:

L'indice di Gini, è un indice che misura il grado di eterogeneità (omogeneità) in una distribuzione statistica di dati suddivise in k categorie (le variabili qualitative) ciascuna delle quali con frequenza relativa al quadrato.

Date k=4 (4 modalità) della variabile soggiorno, l'eterogeneità omologata si calcola mediante l'indice di Gini, così definito:

$$G = 1 - \sum f_i^2$$

$$G = 1 - (0,0413 + 0,0502 + 0,0182 + 0,0146 + 0,0128 + 0,0076 + 0,0028) = 0,8225$$

Poichè si ha sempre $0 \leq G \leq \frac{k-1}{k}$ → $0 \leq a \leq \frac{6}{4} = 0,8541$

(Valore minimo ↑ — Valore massimo ↗)

possiamo dire che i dati sono distribuiti in modo altamente eterogeneo sulle 4 scelte (cioè le scelte relative alla tipologia di soggiorno sono molto simili tra loro).

↓

Questa ricerca si basa sulle tipologie di soggiorno scelte dagli italiani per trascorrere le proprie vacanze estive nel 2016 in Italia.

x	f_i	f_i^2
Albergo	0,264	0,0413
Parenti/amici	0,224	0,0502
Casa proprietà	0,135	0,0182
Villaggi turistici	0,121	0,0146
In affitto	0,113	0,0128
Residence	0,087	0,0076
Campeggio	0,053	0,0028

Se vogliamo una misura relativa del grado di eterogeneità, possiamo calcolare l'indice di Gini relativo o normalizzato che ci dà più l'idea di livello di eterogeneità:

$$G_N = \frac{G}{\frac{k-1}{k}}$$

$$a_N = \frac{0,8225}{0,8541} = 0,9590.$$

Possiamo vederlo anche in percentuale:

$a_N \cdot 100 = a_N\%$

$0,9596 \cdot 100 = 95,96\% \rightarrow$ (95,96% di eterogeneità = elevata eterogeneità)

a_N è chiaramente compreso tra 0 e 1, valori corrispondenti rispettivamente al minimo grado di eterogeneità e massimo grado. Nel nostro caso conferma l'elevata eterogeneità tra i dati della distribuzione.

Riepilogo 1:
In un certo collettivo, dove n=5, la variabile x rappresenta il numero di componenti delle famiglie che presentano i seguenti valori 3, 1, 2, 4, 2.
Calcolare la devianza, la varianza e lo sqm e il CV.

	x_i	$(x_i)^2$	$(x_i - M)$	$(x_i - M)^2$
	3	9	0,6	0,36
	1	1	-1,4	1,96
	2	4	-0,4	0,16
	4	16	1,6	2,56
	2	4	-0,4	0,16
	12	34		5,20

$M = \frac{12}{5} = 2,4$

$Dev(x = 5,20)$

$Var(x) = \sigma^2 = \frac{1}{N} \sum (x_i - M)^2 = \frac{Dev}{N} = \frac{5,20}{5} = 1,04$

$\sigma = \sqrt{VAR(x)} = \sqrt{1,04} = 1,02$

$CV = \frac{1,02}{2,4} = 0,4249 \cdot 100 = 42,49$

Riepilogo 2:
La distribuzione del reddito annuo in euro dei 1.000 abitanti di un comune è la seguente. Calcolare dev(x), var(x), SQM, CV.

CLASSI DI REDDITO	VALORE CENTRALE	n_i	$c_i \cdot n_i$	$(c_i - M)$	$(c_i - M)^2$	$(c_i - M)^2 \cdot n_i$
1.000 ⊢ 5.000	3.000	100	300.000	19800	392040000	39204000000
5.000 ⊢ 15.000	10.000	400	4000000	12800	163840000	65536000000
15.000 ⊢ 35.000	25.000	300	7500000	2200	4840000	1452000000
35.000 ⊢ 75.000	55.000	200	11000000	32200	1036840000	207368000000
TOTALE		1000	22800000			313560000000

$M = \frac{22800000}{1000} = 22.800$ $\quad Dev(x) = 313560000000$

$\sigma = \sqrt{VAR(x)} = \sqrt{\quad}$

$VAR(x) = \sigma^2 = \frac{1}{N} \sum (c_i - \mu)^2 \cdot m_i = \frac{Dev(x)}{N} = \frac{313560000000}{1000} = 313560000$

$\sigma = \sqrt{VAR(x)} = \sqrt{313560000} = 17707,625$

$CV = \frac{\sigma}{\mu} \cdot 100 = \frac{17707,625}{22800} = 0,7767 \cdot 100 = 77,67$

Riepilogo 3.

La seguente tabella fornisce la distribuzione delle 100 famiglie di un quartiere secondo il carattere x = "numero dei figli"

x_i	m_i	N_i
0	30	30
1	15	45
2	20	65
3	12	77
4	10	87
5	9	96
6	4	100
TOTALE	100	
	N	

Range $= x_{max} - x_{min} = 6 - 0 = 6$

$IQR = Q_3 - Q_1 = x\left(3 \cdot \frac{N+1}{4}\right) - x\left(\frac{N+1}{4}\right) = x_{(75,75)} - x_{(25)}$

Differenza interquartile $= 3 - 0 = 3$.

Tale valore sta ad indicare che il 50% delle famiglie analizzate hanno un numero di figli in un intervallo di ampiezza 3.

Gli indici che calcolano la dispersione tra le unità sono: Range e differenza interquartile.

Gli indici che calcolano la dispersione delle unità da un valore centrale sono: scarto, deviazione, varianza, coefficiente di variazione

Riepilogo 4 (Esercizio al 3)

x_i	$x_i n_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$
0	0	-2	4	120
1	15	-1	1	15
2	40	0	0	0
3	36	1	1	12
4	40	2	4	40
5	45	3	9	81
6	24	4	16	64
TOTALE	200			332

M

D

Va

σ

CV

STATISTICA

BIVARIABILE

X (Titolo di studio)	Y (Reddito)
A	H
B	B
H	A
A	A
A	A
H	H
B	H

A = Alto
B = Basso
H = Medio

Nella distribuzione unitaria abbiamo due variabili statistiche, ciascuna con informazioni sul nostro individuo (Titolo di studio). Oltre a studiare le due variabili singolarmente, possiamo studiare la dipendenza delle due variabili, ossia la relazione tra le due variabili.
A tale scopo useremo, prima di tutto la distribuzione in frequenza

Variazione p...

DISTRIBUZIONE IN FREQUENZA DOPPIA:

X	Y	y_1 B	y_2 H	y_3 A
x_1 B		n_{11}	n_{12}	n_{13}
x_2 H		n_{21}	n_{22}	n_{23}
x_3 A		n_{31}	n_{32}	n_{33}
	TOTALE			

-> Andrò a vedere, per esempio, quante persone che hanno un titolo di studio hanno avranno anche un reddito hanno.

-> È per colonne esempio (Andrò ad aggiungere anche un'altra tabella dove faccio è dei marginali di riga

n_{ij} = frequenza congiunta (Misura congiuntamente sia le modalità di x e le modalità di y)
$n_{i\cdot}$ = TOTALE FREQUENZE ASSOLUTE
N = Numero popolazione
i = X
j = Y

ESEMPIO:

TITOLO DI STUDIO (X)	REDDITO: BASSO	MEDIO	ALTO	
BASSO	5	3	1	9
MEDIO	1	18	1	20
ALTO	1	4	16	21
	7	25	18	50

TOTALE MARGINALI X (È per riga)

TOTALE MARGINALI Y (È per colonna)

Il totale dei marginali di y e di x sarà sempre uguale, esso ci darà il numero della popolazione.
Se andremo a prendere una singola riga prenderemo una sotto popolazione e ognuna delle righe sarà distribuzione condizionata, per esempio se ci soffermaniamo sulla prima riga analizzeremo come varia la variabile y condizionata da x = BASSA ($Y|X=B$) Lo stesso discorso vale per le colonne. Quindi avremo:
- Distribuzione condizionata per x;
- Distribuzione condizionata per y.

Più le distribuzioni condizionate sono diverse tra loro, più c'è dipendenza, più sono uguali tra loro, più non c'è dipendenza.

GRAFICA:

$Y|X=B$ -> ASIMMETRICA POSITIVA

$Y|X=H$ -> SIMMETRICO

$Y|X=A$ -> ASIMMETRICO NEGATIVO

DISTRIBUZIONE CONDIZIONATA PER Y: 

Dai grafici vediamo, quindi, che le distribuzioni sono diverse tra loro, quindi, hanno un rapporto di dipendenza, se fossero stati uguali non c'era dipendenza.
PER COSTRUIRE IL GRAFICO DOBBIAMO FARE LE FREQUENZE RELATIVE, NON ASSOLUTA
$(n_{i \cdot j} \cdot n)$ -> FREQUENZA RELATIVA.

FREQUENZE RELATIVE OSSERVATE:

	B	H	A	
B	0,56	0,33	0,11	1
H	0,05	0,90	0,05	1
A	0,05	0,19	0,76	1
	4/50	25/50	18/50	-> PROFILO MEDIO RIGA

} TOTALE FREQUENZE RELATIVE.

} le frequenze relative per riga si chiamano PROFILI DI RIGA

la frequenza relativa la puoi calcolare sia per righe che per colonne, è uguale. (PROFILO DI COLONNA)

Se i profili di riga (o profili di colonna) sono uguali al profilo medio, abbiamo un rapporto di indipendenza tra x e y.

$$\frac{n_{ij}}{n_{i \cdot}} = \frac{n_{\cdot j}}{N} \rightarrow n_{ij} = \frac{n_{i \cdot} \cdot n_{\cdot j}}{N}$$

-> FREQUENZA CONGIUNTA OSSERVATA.

↓

TABELLA INIZIALE (Distribuzione in frequenze)

PROFILO MEDIO RIGA OSSERVATO

0,14	0,50	0,36	1

FREQUENZA CONGIUNTA TEORICA: (n^*_{ij} = FREQUENZA TEORICA)

	B	H	A
B	$\frac{4,9}{50} = 1,26$	4,50	3,24
H	2,80	10,00	7,20
A	2,94	10,50	7,56

Ci sarà dipendenza quando:
$n_{ij} \neq n^*_{ij}$
(la frequenza congiunta osservata è diversa alla frequenza teorica)
PIÙ C'È DIFFERENZA PIÙ C'È DIPENDENZA

Ci sarà indipendenza quando
$n_{ij} = n^*_{ij}$

Per vedere quanta differenza abbiamo, calcoliamo la contingenza, è come se fosse uno scarto. Facciamo la differenza tra n_{ij} e n^*_{ij} $(n_{ij} - n^*_{ij})$

Variabili

Se la differenza è uguale a 0, avremo indipendenza

la contingenza la indichiamo con C_{ij}

CONTINGENZA ESEMPIO:

	B	H	A	
B	3,74	-1,50	-2,24	0
H	-1,80	3	-6,20	0
A	-1,34	-6,50	8,44	0
	0	0	0	

la somma delle contingenze (per riga / per colonna) ci d

Dobbiamo stabilire quanto dipendenza c'è.

$\chi^2_{MAX} = N \cdot \min(I-1 \text{ e } J-1)$

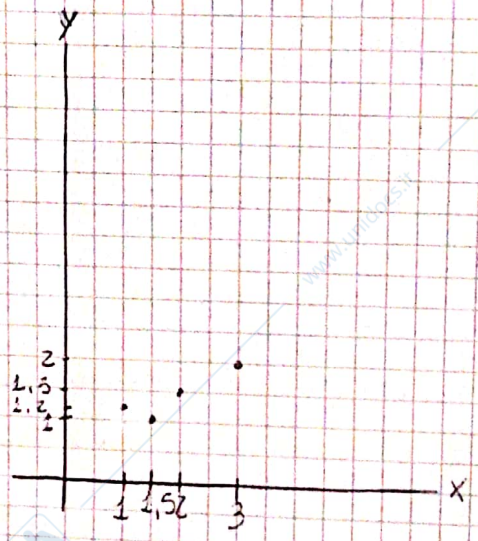
$465 \cdot 1 = 465$

$V = \frac{4,08}{465} = 0,01$ Quasi nulla dipendenza

Il diagramma a dispersione o scatter plot, ci permette di rappresentare sul grafico due variabili

ESEMPIO:

X	Y
1	1,2
1,5	1
2	1,5
3	2



Se la differenza è uguale a 0, avremo indipendenza. ($C_{ij} = 0$)

La contingenza la indichiamo con C_{ij}

CONTINGENZA ESERCIZIO:

	B	H	A	
B	3,74	-1,50	-2,24	0
H	-1,80	8	-6,20	0
A	-1,34	-6,50	8,44	0
	0	0	0	

Il valore di una variabile non cambia al variare dell'altra variabile

La somma delle contingenze (per riga / per colonna) ci darà sempre 0.

Dobbiamo stabilire quanto dipendenza c'è.

Useremo il quadrato della contingenza, in quanto elimina i valori negativi e semplifica la contingenza.

Avremo quindi $(n_{ij} - n_{ij}^*)^2 = C_{ij}^2$

Dopo possiamo scostamento relativo, in questo modo $\left(\frac{C_{ij}^2}{n_{ij}^*}\right)$ rapporto tra la contingenza al quadrato e la frequenza teorica.

SCOSTAMENTO ESERCIZIO (INDICE DI PEARSON)

	B	H	A	
B	11,10	0,60	1,55	13,15
H	1,16	6,40	5,34	12,90 → PEARSON
A	1,28	4,02	9,42	14,72
	13,54	10,92	16,31	40 → TOTALE

↓ PEARSON

A questo punto misuriamo la dipendenza in una tabella a doppia entrata con l'Indice di associazione di Pearson.

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij} - x_{ij}^*)^2}{n_{ij}^*}$$

= 0 → INDIPENDENZA ($C_{ij} = 0 \ \forall_{i,j}$) ($n_{ij} = n_{ij}^* \ \forall_{i,j}$)

> 0 → DIPENDENZA (anche detta INTERDIPENDENZA)

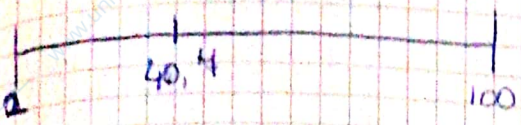
Dobbiamo quindi, stabilire il valore massimo, con questa formula:

$$\chi^2_{MAX} = N \cdot \min(I - 1 \text{ e } J - 1)$$

↓ Numero di righe — ↓ Numero di colonne

$\chi^2_{max} = 50 \cdot 2 = 200$

Quindi:



Per capire con valori relativi l'associazione, dipendenza usiamo la V di Cramer

$$V = \frac{\chi^2}{\chi^2_{MAX}}$$

$$V = \frac{40,4}{100} = 0,41 = 41\%$$

Indice di associazione di Pearson.

L'indice di associazione di Pearson è un indice simmetrico, cambiando la disposizione delle variabili, l'indice non cambia. X ed Y si influenzano tra loro.

Indipendenza

Avremo un rapporto di indipendenza quando i rapporti tra i profili riga è uguale al rapporto tra i profili medi di riga.

Interdipendenza

È il valore più forte di dipendenza tra due variabili.

Esempio

	n_{ij}		
			n_j
H	70	110	180
F	85	200	285
	155	310	465
	n_i		

	n^*_{ij}		
H	60	120	180
F	95	190	285
	155	310	465

$$n^*_{ij} = \frac{n_i \cdot n_j}{N}$$

	C_{ij}^2	
H	1,67	0,83
F	1,05	0,05

$$C_{ij}^2 = \frac{(n_{ij} - n^*_{ij})^2}{n^*_{ij}}$$

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij} - x_{i,j})^2}{n_{ij}} = 4,08$$