Statistica Descrittiva: Descrive ciò che si osserva nei DATI. Fornisce un quadro chiaro e conciso della situazione attuale. Utilizza strumenti come la MEDIA e MEDIANA.

Statistica Inferenziale: Va oltre la descrizione dei DATI, si cercano di trarre conclusioni, formulare ipotesi e verificare la significatività dei risultati. Permette di generalizzare i risultati da un CAMPIONE all'intera POPOLAZIONE

VARIABILI ALEATORIE: Una VARIABILE ALEATORIA è una funzione che associa ad ogni esperimento casuale un numero reale, viene usata per modellare fenomeni incerti o casuali

VARIABILI ALEATORIE → Qualitative (factor) → Etichetta
VARIABILI ALEATORIE → Quantitative (numeric) → Numerica

DESCRIZIONE CAMPIONE

1) DISTRIBUZIONE UNITARIA → Elenco di tutte le osservazioni

$x_1, x_2, \dots, x_n$ → Ampiezza Campione

2) DISTRIBUZIONE DI FREQUENZE

x_1	m_1
⋮	⋮
x_m	m_i
	⋮
	m_K

FREQUENZE → ASSOLUTA → RELATIVA → CUMULATA

ASSOLUTA	RELATIVA	CUMULATA
m_1	$f_1 = \frac{m_1}{n}$	$F_1 = f_1$
mi	$f_i = \frac{m_i}{n}$	$F_i = F_{i-1} + f_i$
⋮	⋮	⋮
m_K	$f_K = \frac{m_K}{n}$	$F_K = F_{K-1} + f_K$

FREQUENZA ASSOLUTA: Numero di tutte le unità statistiche che assumono un certo valore o modalità in relazione ad un carattere. Sono tutti quei dati che assumono quel determinato valore o modalità

Esempio:

			Freq assolute
25	occhi	marroni	25
14	//	verdi	14
8	//	azzurri	8

FREQUENZA RELATIVA: Rapporto tra la FREQUENZA ASSOLUTA e la numerosità della popolazione o del campione statistico. La somma di tutte le freq. relative deve essere $=1$

$$\text{FREQ. RELATIVA} = \frac{\text{FREQUENZA}}{\text{N°tot. Dati}}$$

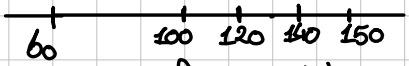
oss: $\sum_{i=1}^{k} n_i = n \quad \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^{k} n_i = \frac{1}{n} \cdot n = 1$

FREQUENZA CUMULATA → su R cumsum()

- plot() diagramma a linee
- barplot() diagramma a barre ≠ istogramma

Es. distribuzione classi

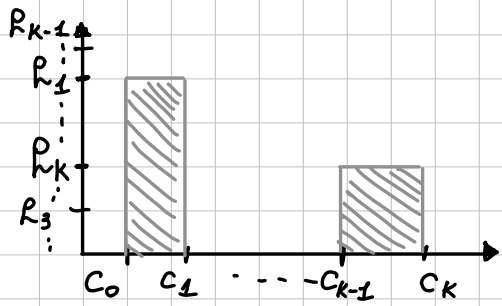
$C_0 = 60 \quad C_1 = 100$
$C_2 = 120 \quad C_3 = 140$
$C_K = C_4 = 150$



	freq. assoluta	freq. Relativa	Densità di Frequenza
$[C_0, C_1)$	n_1	$f_1 = \frac{n_1}{k}$	$p_1 = f_1 / C_1 - C_0$
$[C_1, C_2)$	n_2	$f_2 = \frac{n_2}{k}$	$p_2 = f_2 / C_2 - C_1$
⋮	⋮	⋮	⋮
$[C_{k-1}, C_k]$	n_k	$f_k = \frac{n_k}{k}$	$p_k = f_k / C_k - C_{k-1}$

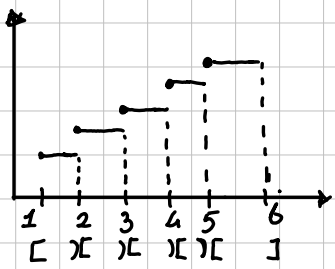
$$\text{DENSITÀ DI FREQUENZA} = \frac{\text{FREQ}}{\text{AMPIEZZA CLASSE}}$$

Se utilizzo freq. assoluta ho densità assoluta
se utilizzo freq. relativa ho densità Relativa



FUNZIONE DI RIPARTIZIONE

		f_i	F_i
m_1	1	0,1	0,1
m_2	2	0,3	0,4
m_3	3	0,1	0,5
m_4	5	0,2	0,7
m_5	8	0,2	0,9
m_6	11	0,1	1.0



f_i = frazioni di osservazioni uguali a m_i

F_i = frazioni di osservazioni f_i e m_i

Se abbiamo una V.A ordinata ci possiamo chiedere quale frazione di osservazioni sia più piccola o uguale ad un valore x

$F_m(x)$ = "frazioni di osservazioni del campione $\leq x$

↳ FORMULA DI RIPARTIZIONE EMPIRICA:

- Costante a tratti
- Continua a destra → gli intervalli sono chiusi a Sx

  Cioè non DECRESCENTE e $\lim_{x \to +\infty} F(x) = 1$

MODA: Modalità più rappresentativa (Frequenza maggiore Relativa) può essere univoca, biunivoca o non esistere:

m_i	f_i
3	0.1
7	0.3
11	0.2
15	0.3
18	0.1

Moda = {7, 15}

Possiamo definire la moda delle distribuzione in classi

	m_i	f_i	f_i
[1, 5)	3	3/10	15/200
[5, 10)	4	4/10	16/200
[10, 12)	2	2/10	20/200
[12, 17)	1	1/10	4/200

La MODA è la classe con la DENSITÀ DI FREQUENZA maggiore (il picco dell'istogramma)

$f_i : C_i - C_{i-1}$

$\frac{3}{40} + \frac{4}{50} + \frac{4}{10} + \frac{1}{50} =$

$\frac{15 + 16 + 20 + 4}{200}$

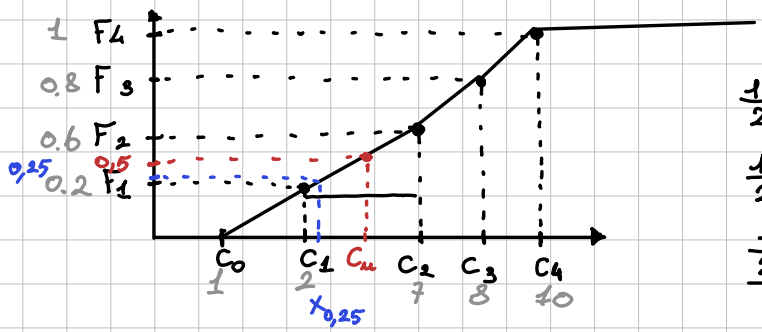
MEDIANA: É un valore $\bar{m}$ tale che almeno la metà delle osservazioni è minore o uguale a $\bar{m}$ o maggiore o uguale a $\bar{m}$

1 2 (3) 5 7 → Mediana

9 (3) (7) 11 → Faccio la media = (5) → Mediana

Oss: · La mediana può essere un valore che ∉ al campione.
· É un indicatore ROBUSTO → non cambia se variano gli estremi

Mediana a partire dalla funzione di ripartizione della distribuzione in classi di una v_A continua



i^* = classe in cui si trova la mediana

$$\frac{1}{2} = (\bar{m} - C_1) \cdot p_{i^*} + F_1$$

$$\frac{1}{2} = (\bar{m} - C_{i^*-1}) p_{i^*} + F_{i^*-1}$$

$$\frac{\frac{1}{2} - F_{i^*-1}}{p_{i^*}} + C_{i^*-1} = \bar{m}$$

QUANTILI ⇝ "Generalizzazioni della mediana

x_p valore tale per cui $p \cdot n$ osservazioni $\leq x_p$ $\quad (1-p)n$ osservazioni $\geq x_p$

Intuitivamente $x_{0,25}$ valore tale che il 25% delle osservazioni sono $\leq x_{0,25}$

MEDIA CAMPIONARIA

Es: Tempi di attese ad un servizio
1 3 8 11 8 8 5 6 6 8

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \Rightarrow \bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10} \cdot 64 = 6.4$$

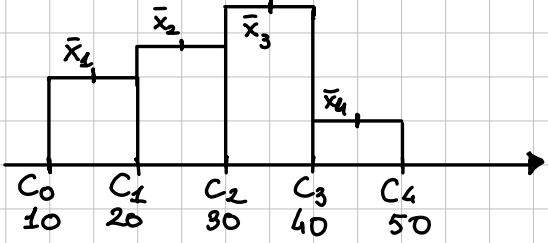
Es: calcoliamo le medie a partire dalla distribuzione delle frequenze Relative

m_i	n_i	f_i
1	10	10/38
2	4	4/38
3	8	8/38
4	16	16/38

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{k} m_i \cdot n_i = \sum_{i=1}^{k} \frac{n_i}{n} \cdot m_i = \sum_{i=1}^{k} f_i \cdot m_i$$

Oss: $\sum_{i=1}^{k} f_i = 1$

Supponiamo di avere una distribuzione in classi



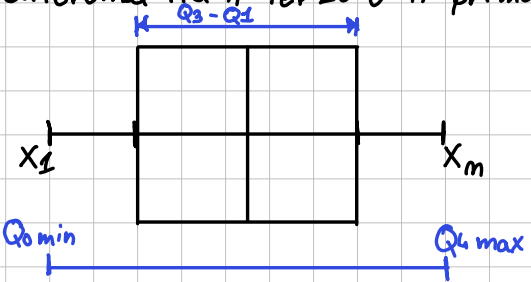
$$\bar{x} = \frac{\sum_{i=1}^{k} \bar{x}_i \cdot n_i}{n} = \sum_{i=1}^{k} f_i \cdot \bar{x}_i \qquad \bar{x}_i = \frac{c_i - c_{i-1}}{2}$$

Caratterizzare la dispersione dei dati

- CAMPO DI VARIAZIONE (range) $x_1 \ldots x_m$ $x_{(m)} - x_{(1)}$ (più grande, il più piccolo)

- SCARTO INTERQUARTILE: differenza tra il terzo e il primo quartile

$Q_3 - Q_1$ (3° quartile 75%, 1° quartile (25%))



Vogliamo descrivere le distanze tipica dei valori del campione da $\bar{x}$

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x}) = \frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x} = \bar{x} - \frac{1}{n}\cdot n\bar{x} = 0$$

(Scarti)

$\sigma_x^2 = \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$: Varianza della popolazione

$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$: Varianza Campionaria

$S_n^2 = \frac{n}{n-1}\cdot\sigma^2$

Ossi

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x}+\bar{x}^2) =$$

$$= \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2}{n}\sum_{i=1}^{n}x_i\bar{x} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{2}{n}\bar{x}\underbrace{\sum_{i=1}^{n}x_i}_{n\bar{x}} + \frac{1}{n}n\bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - 2\bar{x}^2 + \bar{x}^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 \geq 0$$

$(\geq 0,\ = 0 \iff$ tutti i $x_i = \bar{x})$

Esempio: Calcolare la varianza a partire dalla distr. di Frequenza

m_1	n_1	f_1
m_2	n_2	f_2
⋮	⋮	⋮
m_K	n_K	f_K

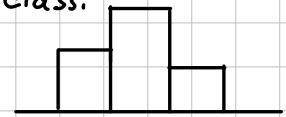
$$S_n^2 = \frac{n}{n-1}\sum_{i=1}^{K}(m_i-\bar{x})^2\cdot n_i$$

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{K}(m_i-\bar{x})^2\cdot n_i = \sum_{i=1}^{K}(m_i-\bar{x})^2\cdot f_i$$

$f_i = \frac{n_i}{n}$

$$S_n^2 = \frac{n}{n-1}\sigma^2$$

Varianza a partire dalla distribuzione in classi:



$$\frac{\sum_{i=1}^{K}(\bar{x}_i-\bar{x})\cdot n_i}{n-1} = S_n^2$$

($\bar{x}_i$: "medio della distribuzione in classi"; "medio dell'i-esima class")

(è una scelta convenzionale)

Esempio

Tempi di attesa in secondi: 1s, 2s, 3s

Calcoliamo S_n^2 $\quad n=3 \quad \bar{x}=2$

$$S_n^2 = \frac{(1-\bar{x})^2+(2-\bar{x})^2+(3-\bar{x})^2}{3-1} = 1s^2$$

Deviazione Standard (scarto quadratico medio)

$\sigma = \sqrt{\sigma^2}$ $\quad$ $s_m = \sqrt{s_m^2}$ $\quad$ $\frac{1}{m}\sum_{i=1}^{m}|x_i-\bar{x}|$ vs $\sqrt{\frac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2}$ $\quad$ $\frac{1}{m}\sum_{i=1}^{m}|x_i-\bar{x}|^2 \geq \left|\frac{1}{m}\sum_{i=1}^{m}|x_i-\bar{x}|\right|^2$ $\quad$ $|x_i-\bar{x}|^2 = (x_i-\bar{x})^2$

$$s_m \geq \frac{1}{m}\sum_{i=1}^{m}|x_i-\bar{x}|$$

Coefficente di Variazione

É un indice di dispersione, permette di confrontare misure di fenomeni riferite a unità di misura differenti, in quanto si tratta di una grandezza adimensionale

$$CV(x) = \frac{\sigma_x}{|\bar{x}|} \qquad Cv = \frac{\sigma}{\mu} \begin{matrix} \rightarrow \text{Deviazione standard} \\ \rightarrow \text{Media Aritmetica} \end{matrix}$$

1) Non dipende da unità di misura

x = secondi
y = minuti

$y = ax$ $\quad$ $\frac{\sigma_y}{|\bar{y}|} = \frac{|a|\,\sigma_x}{|a\bar{x}|} = \frac{\sigma_x}{|\bar{x}|}$

Relazioni fra più variabili aleatorie (Cerchiamo di descrivere le connessioni/relazioni fra i fenomeni aleatorie)

"m" osservazioni di coppie di variabili aleatorie

$(\bar{x}_1, \bar{y}_1)$
$(\bar{x}_2, \bar{y}_2)$
$\vdots$
$(\bar{x}_m, \bar{y}_m)$

x \ y	y_1	y_2	$\cdots$	y_j	$\cdots$	y_k	
x_1	$m_{1,1}$	$m_{1,2}$	$\cdots$	$m_{1,j}$	$\cdots$	$m_{1,k}$	m_1
x_2	$m_{2,1}$	$m_{2,2}$	$\cdots$	$m_{2,j}$	$\cdots$	$m_{2,k}$	m_2
$\vdots$							
x_i	$m_{i,1}$	$m_{i,2}$	$\cdots$	$m_{i,j}$	$\cdots$	$m_{i,k}$	m_i
$\vdots$							
x_f	$m_{f,1}$	$m_{f,2}$	$\cdots$	$m_{f,j}$	$\cdots$	$m_{f,k}$	$m_{f,\cdot}$
	$m_{\cdot,1}$	$m_{\cdot,2}$	$\cdots$	$m_{\cdot,j}$	$\cdots$	$m_{\cdot,k}$	

modalità delle y

modalità delle x

Frequenze congiunte assolute

Frequenze assolute MARGINALI

$$m_{i,\cdot} = \sum_{j=1}^{k} m_{i,j} \text{ (delle x)}$$

$$m_{\cdot,j} = \sum_{i=1}^{k} m_{i,j} \text{ (delle y)}$$

Esempio: x: reddito y: quartiere di residenza (A, B, C)

	A	B	C	
ALTO	7	4	1	12
MEDIO	3	5	11	19
BASSO	2	21	3	26
	12	30	15	

$$f_{i,\cdot} = \sum_{j=1}^{k} f_{i,j} = \sum_{j=1}^{k} \frac{m_{i,j}}{m}$$

$$f_{\cdot,j} = \sum_{i=1}^{R} f_{i,j} = \sum_{i=1}^{R} \frac{m_{i,j}}{m}$$

Qual è la proporzione di persone del quartiere A se sappiamo che le persone reddito alto? $\frac{7}{12}$

Frequenze Condizionate

$$\begin{cases} f_{x=i|y=j} := \frac{m_{i,j}}{m_{\cdot,j}} \\ f_{y=j|x=i} := \frac{m_{ij}}{m_{i,\cdot}} \end{cases}$$

	BLU	VERDE	MARRONE	
ALTO	2	4	14	20
MEDIO	3	6	21	30
BASSO	5	10	35	50
	10	20	70	

→

	BLU	VERDE	MARRONE	
ALTO	2/100	4/100	14/100	2/10
MEDIO	3/100	6/100	21/100	3/10
BASSO	5/100	10/100	35/100	5/10
	1/10	2/10	7/10	

CALCOLI LE CONDIZIONATE RISPETTO AL COLORE DEGLI OCCHI

	BLU	VERDE	MARRONE
ALTO	2/10	4/20	14/70
MEDIO	3/10	6/20	21/70
BASSO	5/10	10/20	35/70

OSS: Le condizionate coincidono con le marginali:

PROVIAMO A CALCOLARE LE CONDIZIONATE RISPETTO AL REDDITO

	BLU	VERDE	MARRONE
ALTO	2/20	4/20	14/20
MEDIO	3/30	6/30	21/30
BASSO	5/50	10/50	35/50

Anche in questo caso le frequenze condizionate coincidono con le frequenze marginali:

⇒ Le variabili sono dette INDIPENDENTI

Consideriamo due variabili indipendenti:

$$f_{x=i|y=j} = \frac{n_{ij}}{n_{\cdot,j}} = \frac{n_{i,\cdot}}{n} = f_{i,\cdot}$$

$$f_{y=j|x=i} = \frac{n_{ij}}{n_{i,\cdot}} = \frac{n_{\cdot,j}}{n} = f_{\cdot,j}$$

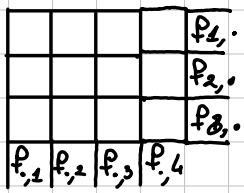
in entrambi i casi abbiamo $n_{ij} = \frac{n_{i,\cdot} \times n_{\cdot,j}}{n}$

$$\boxed{f_{ij} = f_{i,\cdot} \times f_{\cdot,j}}$$

Prendiamo questo come DEFINIZIONE DI INDIPENDENZA

Prendiamo due V.A X, Y con frequenze marginali $f_{i,\cdot}, f_{\cdot,j}$ con $i = 1, \dots, R$, $j = 1, \dots, K$

misuriamo il "grado di dipendenza"



Se fossero indipendenti la distribuzione congiunta $f^*_{ij} = f_{i,\cdot} \times f_{\cdot,j} \Leftrightarrow m^*_{ij} = \frac{m_{i,\cdot} \times m_{\cdot,j}}{n}$

Freq. Effettiva

$C_{i,j} = m_{ij} - m^*_{ij}$

Freq. Attesa

Contingenza

$$\chi^2 = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{(m_{ij} - m^*_{ij})^2}{m^*_{ij}} \qquad m^*_{ij} = \frac{m_{i,\cdot} \times m_{\cdot,j}}{n}$$

STATISTICA CHI-QUADRO

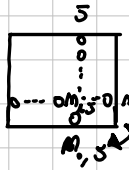
$$\chi^2 = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{(m_{ij} - m^*_{ij})^2}{m^*_{ij}} = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m_{ij}^2 - 2 m_{ij} \times m^*_{ij} + m^{*2}_{ij}}{m^*_{ij}} = \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m_{ij}^2}{m^*_{ij}} - \underbrace{2 \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m_{ij} \cdot m^*_{ij}}{m^*_{ij}}}_{n} + \underbrace{\sum_{i=1}^{K} \sum_{j=1}^{K} m^*_{ij}}_{n}$$

$$= n \sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m_{ij}^2}{m_{i,\cdot} \times m_{\cdot,j}} - \cancel{2n} + \cancel{n} = n \left(\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m_{ij}^2}{m_{i,\cdot} \times m_{\cdot,j}} - 1 \right) \geq 0$$

Def: Una statistica è una qualsiasi funzione del campione

Vogliamo confrontare il valore di χ^2 con un valore di riferimento. Determiniamo il valore massimo χ^2_{max} che la statistica χ^2 può assumere.

Si ha dipendenza massima se conoscere x ci dice anche il valore di y e viceversa



Calcoliamo χ^2_{max} in un caso di questo tipo

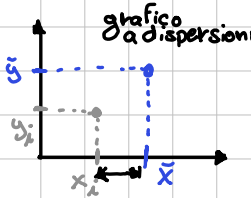
$$n \left(\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{m_{ij}^2}{m_{i,\cdot} \times m_{\cdot,j}} - 1 \right) \quad \text{poichè } m_{\cdot,j} = m_{i,\cdot} \Rightarrow n \left(\sum_{i=1}^{K} \sum_{j=1}^{K} \frac{\cancel{m_{ij}^2}}{\cancel{m_{ij}^2}} - 1 \right)$$

Al massimo ci sono $\min\{R, K\}$ elementi non nulli: $\chi^2 = n(\min\{R, K\} - 1)$

$\frac{\chi^2}{\chi^2_{max}}$ ← Chi-quadro RELATIVO $[0, 1]$

Più vicino a 1 e più le variabili sono dipendenti

Vogliamo misurare il grado di dipendenza lineare di due variabili aleatorie



$\mathrm{Cov}(x,y)\quad \sigma_{xy} = \dfrac{\sum_{i=1}^{m}(x_i-\bar{x})\cdot(y_i-\bar{y})}{m} = \dfrac{1}{m}\sum_{i=1}^{m}(x_iy_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y}) =$

$= \dfrac{1}{m}\sum_{i=1}^{m}x_iy_i - \underbrace{\dfrac{1}{m}\sum_{i=1}^{m}x_i\bar{y}}_{-\bar{x}\bar{y}} - \underbrace{\dfrac{1}{m}\sum_{i=1}^{m}\bar{x}y_i}_{-\bar{x}\bar{y}} + \underbrace{\dfrac{1}{m}\sum_{i=1}^{m}\bar{x}\bar{y}}_{\bar{x}\bar{y}} = \sum_{i=1}^{m}\dfrac{x_iy_i}{m} - \bar{x}\bar{y}$ Ci aspettiamo $\mathrm{Cov}(x,y)>0$

$X \rightsquigarrow Z = aX+b \quad z_i = ax_i+b$

$Y \rightsquigarrow W = cY+d \quad w_i = cy_i+d$ $\quad \mathrm{Cov}(Z,W) = \dfrac{1}{m}\sum_{i=1}^{m}(z_i-\bar{z})(w_i-\bar{w}) = \dfrac{1}{m}\sum_{i=1}^{m}(ax_i+b-a\bar{x}-b)(cy_i+b-c\bar{y}-d)$

$= \dfrac{1}{m}\sum_{i=1}^{m}a(x_i-\bar{x})c(y_i-\bar{y}) = a\cdot c\cdot\dfrac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y}) = a\cdot c\cdot\mathrm{Cov}(x,y)$

Esercizio: Siano X e Y t.c $\mathrm{Cov}(x,y)=48$ sia $\sigma_x=6$ e sia $\sigma_y=8$ quanto vale $\mathrm{Cov}\left(\frac{1}{6}x, -\frac{1}{8}y\right)$?

$\mathrm{Cov}\left(\frac{1}{6}x, -\frac{1}{8}y\right) = \frac{1}{6}\cdot\left(-\frac{1}{8}\right)\cdot\mathrm{Cov}(x,y) = -1$ $\quad \mathrm{Cov}(x,y) \rightsquigarrow \rho_{xy} = \dfrac{\mathrm{Cov}(x,y)}{\sigma_x\,\sigma_y}$

↑ COEFFICENTE DI CORRELAZIONE

oss: $\mathrm{Var}(x) = \dfrac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})^2 = \dfrac{1}{m}\sum_{i=1}^{m}(x_i-\bar{x})(x_i-\bar{x}) = \mathrm{Cov}(x,x)$

oss: In generale $\mathrm{Var}\left(\dfrac{x}{\sigma_x}\right) = \dfrac{1}{\sigma_x^2}\cdot\sigma_x^2 = 1$ $\quad \left[\mathrm{Var}(ax) = a^2\mathrm{Var}(x)\right]$

Siano $Z = aX+b$ e $W = cY+d$

calcoliamo il coeff. di correlazione di Z,W $\quad \rho_{zw} = \dfrac{\mathrm{Cov}(z,w)}{\sigma_z\,\sigma_w} = \dfrac{\mathrm{Cov}(aX+b, cY+d)}{\sigma_{aX+b}\;\sigma_{cY+d}} = \dfrac{a\cdot c\cdot\mathrm{Cov}(x,y)}{|a||c|\,\sigma_x\,\sigma_y}$

($\sigma_{aX+b} = |a|\sigma_x$, $\sigma_{cY+d} = |c|\sigma_y$; $\mathrm{Cov}(x,y) \rightsquigarrow \rho_{xy}$)

$= \mathrm{sign}(a\cdot c)\cdot\rho_{xy}$

+ o −

il valore assoluto del coeff. di corr. non cambia se le variabili aleatorie sono sottoposte ad una trasformazione affine

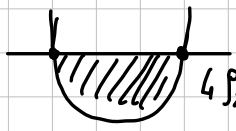
CLAIM: $-1 \le \rho_{xy} \le 1$

PROOF:

Siano x, y t.c $\bar{x}=0$ $\bar{y}=0$, $\sigma_x^2=1$ $\sigma_y^2=1$

$0 \le \dfrac{1}{m}\sum_{i=1}^{m}(x_i - t y_i)^2 = \underbrace{\dfrac{1}{m}\sum_{i=1}^{m}x_i^2}_{\mathrm{Var}(x)=1} - \underbrace{\dfrac{2t}{m}\sum_{i=1}^{m}x_iy_i}_{-2t\,\rho_{xy}} + \underbrace{\dfrac{1}{m}\sum_{i=1}^{m}t^2y_i^2}_{t^2\frac{1}{m}\sum_{i=1}^{m}y_i^2 = t^2\cdot 1} = 1 - 2t\rho_{xy} + t^2$

$\Delta = 4\rho_{xy}^2 - 4$



$4\rho_{xy}^2 - 4 \le 0 \iff \rho_{xy}^2 \le 1$ $\quad -1 \le \rho_{xy} \le 1$ ■

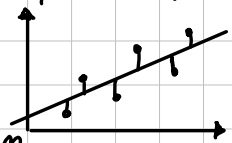
il $|\rho_{xy}| = 1 \iff y = ax+b$ e $x = \frac{y}{a} - \frac{b}{a}$

oss: Se due variabili aleatorie sono indipendenti: $\Rightarrow Cov(x,y)=0$ $\quad X \perp Y \Rightarrow Cov(x,y)=0$ (*)

Regressione Lineare

$Y = aX + b + Z$ → Termine di errore

$y_i = a x_i + b + z_i$



$\bar{z}=0$

Omoschedasticità la varianza della z non dipende dalla x

Trovare a e b in modo tale che sia minimo $\sum_{i=1}^{m}(y_i - a - bx_i)^2$: per trovarli poniamo le derivate parziali pari a 0

($y_i - a - bx_i$ = y_i^*)

$$\begin{cases} \frac{d}{da}\sum_{i=1}^{m}(y_i - a - bx_i)^2 = 0 \\ \frac{d}{db}\sum_{i=1}^{m}(y_i - a - bx_i)^2 = 0 \end{cases} \quad \begin{cases} 2\sum_{i=1}^{m} y_i - a - bx_i = 0 \\ -2\sum_{i=1}^{m}(y_i - a - bx_i)x_i = 0 \end{cases}$$

a) $\sum_{i=1}^{m} y_i - a - bx_i = 0 \Leftrightarrow m\bar{y} - ma - bm\bar{x} = 0 \Leftrightarrow \bar{y} - a - b\bar{x} = 0 \Rightarrow \hat{a} = \bar{y} - b\bar{x}$

b) $\sum_{i=1}^{m} y_i x_i - a x_i - b x_i^2 = \sum_{i=1}^{m} y_i x_i - \bar{y} x_i + b x_i \bar{x} - b x_i^2 = \sum_{i=1}^{m} y_i x_i - m\bar{x}\bar{y} + mb\bar{x}^2 - b\sum_{i=1}^{m} x_i^2 = 0 \Rightarrow \hat{b} = \frac{\sum_{i}^{m} x_i y_i - m\bar{x}\bar{y}}{\sum x_i^2 - m\bar{x}^2} =$

($a = \bar{y} - b\bar{x}$)

dividiamo sopra e sotto per m

$$\hat{b} = \frac{\frac{1}{m}\sum_{i}^{m} x_i y_i - \bar{x}\bar{y}}{\frac{1}{m}\sum_{i}^{m} x_i^2 - \bar{x}^2} = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\rho_{xy}\cdot\sigma_x\sigma_y}{\sigma_x^2} = \boxed{\rho_{xy}\cdot\frac{\sigma_y}{\sigma_x}}$$

sostituisco

con $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x\sigma_y} \Rightarrow \sigma_{xy} = \rho_{xy}\cdot\sigma_x\sigma_y$

$[\rho_{xy}]^2 > 0.7 \Rightarrow$ modello "Buono"

CAMBIAMENTI DI SCALA $\quad y = e^{(a+bX)} \Rightarrow \log y = a + bX$

$y = (a + bX)^K \Rightarrow (y)^{1/K} = a + bX$