FORMULARIO STATISTICA I

FREQUENZE ASSOLUTE

$$n_1+n_2+\dots+n_k=\sum_{j=1}^{k} n_j = n$$

$j=1,\dots,h \qquad 0 \le n_j \le h$

FREQUENZE RELATIVE

$$F_j = \frac{\text{FREQUENZE ASSOLUTE}}{\text{NUMEROSITÀ CAMPIONARIA}}$$

$$F_j = \frac{n_j}{n} \qquad j=1,\dots,k$$

FUNZIONE DI RIPARTIZIONE EMPIRICA

$$F(x)=\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \le x)$$

↓ FUNZIONE INDICATRICE

PUNTATORE CHE CONTA

- $x_i \le x = 1$
- $x_i \ge x = 0$

- Il massimo è 1
- si sommano le FREQUENZE RELATIVE SI CUMULANO

MEDIA ARITMETICA

$$\bar{x}=\frac{x_1+x_2+\dots+x_n}{n}=\frac{1}{n}\sum_{i=1}^{n} x_i$$

(numeratore: VALORI; denominatore: NUMEROSITÀ CAMPIONARIA; x_i: 1 per OGNI x)

$$\bar{x}=\frac{5+2+3+8}{4}=4{,}5$$

- compreso tra il MIN e MAX
- all'interno può esserci una media parziale dei dati → NON CAMBIA
- se $y_i = a + bx_i \quad i=1,\dots,h$
  allora $\bar{y}=\frac{1}{h}\sum_{i=1}^{h} y_i = \underbrace{a+b\bar{x}}_{\text{SONO 2 NUMERI QUALSIASI}}$

MEDIA PROPRIETÀ DEL BARICENTRO

$$\sum_{i=1}^{n}(x_i-\bar{x})=(x_1-\bar{x})+\dots+(x_n-\bar{x})=0$$

- (CI SARANNO VALORI NEGATIVI SOMMATI A POSITIVI) = 0
- la somma degli SCARTI È = 0

MEDIA PROPRIETÀ SCARTI QUADRATICI

- MINIMIZZA LA SOMMA DEGLI SCARTI QUADRATICI

$$\bar{x}=\arg\min_{z}\sum_{i=1}^{n}(x_i-z)^2$$

NON OTTERRÒ MAI UN RISULTATO + PICCOLO

ESERCIZIO ESEMPIO

distribuzione semplice = 2, 1, 1, 4

$\text{MEDIA} = \frac{2+1+1+4}{4} = 2$

SOMMA SCARTI^2 → $(2-\underbrace{2}_{\text{MEDIA}})^2+(1-2)^2+(1-2)^2+(4-2)^2=6$

CALCOLO RICORSIVO

- SE VA AGGIUNTO ALLA MEDIA UN NUOVO DATO

$$\bar{X}_{n+1} = \frac{n}{n+1}\bar{x}_n + \frac{1}{n+1}\bar{X}_{n+1}$$

ESERCIZIO

$\bar{x}_9 = 26$
$x_{10} = 30$
$\bar{x}_{10} = ?$

$$\bar{X}_{10} = \frac{9}{9+1}26 + \frac{1}{9+1}30 = 26.4$$

(n sopra 9; $26 = \bar{x}_9$; $30 = \bar{x}_{9+1}$)

CONVESSITÁ / CONCAVITÁ

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \text{CONVESSA} \qquad \forall x, y \in (a,b) \text{ e } 0<\lambda<1$$
$$\geq \quad " \quad \text{CONCAVA}$$

TEOREMA DISUGUAGLIANZA DI JENSEN

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \geq f\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)$$

APPROSSIMAZIONE DELLA MEDIA

$$\bar{x} \approx \frac{1}{n}\sum_{i=1}^{n} n_i m_i$$

(m_i: PUNTO CENTRALE DELL'INTERVALLO)

MEDIA ARITMETICA PONDERATA

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} W_i}$$

(w_i: PESO NUMERICO; W_i: PESI STANDARDIZZATI)

MEDIANA

$$Me = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{se } n \text{ è dispari} \\ \left(x_{(n/2)} + x_{(n/2+1)}\right)/2 & \text{se è pari} \end{cases}$$

$$F(me) \approx \frac{1}{2}$$

APPROSSIMAZIONE MEDIANA

$$Me \approx z_{j-1} + (z_j - z_{j-1})\frac{1/2 - F(z_{j-1})}{F(z_j) - F(z_{j-1})}$$

- IN CASO DI DATI RAGGRUPPATI
- z_{j-1} e z_j = ESTREMI INTERVALLO
- $F(x)$ = FUNZIONE DI RIPARTIZIONE

ESERCIZIO

CLASSI	$(0,1]$	$(1,2]$	$(2,3]$	$(3,4]$	$(4,5]$
FREQ. AS.	1	4	4	2	1

$$Me \approx z_{j-1} + (z_j - z_{j-1})\frac{1/2 - F(z_{j-1})}{F(z_j) - F(z_{j-1})} = 2 + (3-2)\frac{1/2 - 5/12}{9/12 - 5/12} = 2.25$$

QUANTILE

- GENERALIZZANO LA MEDIANA

0.25, 0.5, 0.75 → 1°, 2°, 3° QUARTILE
(0.5 ↳ MEDIANA)

0.1, 0.2 ... 0.9 → DECILI

0.01, ..., 0.99 → PERCENTILI

$Q_p = \inf\{x : F(x) \geq p\}$ → p è il più piccolo valore che è più grande di almeno $100 \times p\%$ dei dati

$$Q_{0.5} = \begin{cases} X_{(\frac{n+1}{2})} & n \text{ dispari} \\ X_{(n/2)} & n \text{ PARI} \end{cases}$$

MODE DI BONFERRONI

$$M = f^{-1}\left(\frac{1}{B}\sum_{i=1}^{n} f(x_i)\right)$$

↳ funzione INVERSA

MODE DI CHISINI

$$\sum_{i=1}^{n} f(x_i)$$

MODE DI WALD

$$g(l(x_1, M), \ldots, l(x_n; 2))$$

VARIANZA (SIGMA)

$$\sigma^2 = \frac{(x_1-\bar{x})^2 + (x_2-\bar{x})^2 + \ldots + (x_n-\bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

- É UNA MISURA DI QUANTO I DATI SIANO DISTANTI DALLA MEDIA ARITMETICA
- volutata usando i quadrati delle differenze

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \;(1) = \frac{1}{n^2}\sum_{j<i}^{n}(x_i-x_j)^2 \;(3) = \left(\frac{1}{n}\sum_{i=1}^{n}x_i^2\right) - \bar{x}^2 = \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}(x_i-x_j)^2 \;(2)$$

ESEMPIO DI CALCOLO DELLA VARIANZA

DATI 1, 3, 2, 5

- MEDIA ARITMETICA (MOMENTO PRIMO) $\bar{x} = (1+3+2+5)/4 = 2.75$ M. 1°
- MEDIA DEI QUADRATI (MOMENTO SECONDO) $\frac{1}{n}\sum_{i=1}^{n}x^2 = \frac{1^2+3^2+2^2+5^2}{4} = 9.75$ M 2°
- VARIANZA $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \bar{x}^2 = 9.75 - 2.75^2 = 2.19$

(9.75 ↳ MOMENTO SECONDO; 2.75 ↳ MOMENTO PRIMO)

ESEMPIO VARIANZA

$k = 3$ (MODALITÀ)

MODALITÀ c_j	4	6	7
FR. ASSOL. h_j	2	8	3

- MEDIA $\overline{x} = (4\times2 + 6\times3 + 7\times3)/13 = 5,9231$
- MEDIA AL QUADRATO $\overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{n}\sum_{j=1}^{k} h_j c_j^2 = (2\times4^2 + 8\times6^2 + 3\times7^2)/13 =$

$= 35.9230$

- VARIANZA $\sigma^2 = 35.9230 - 5,9231^2 = 0.840$

PROPRIETÀ

La varianza σ^2 è sempre MAGGIORE O UGUALE a 0

$$\sigma^2 \geq 0$$

La varianza $\sigma^2 = 0$ se i dati sono UGUALI TRA LORO

SCARTO QUADRATICO MEDIO

- è la radice quadrata della varianza

(SCARTO QUADRATICO MEDIO) $= \sigma = \sqrt{\sigma^2}$ (coincide con l'unità di misura dei dati)

CAMPO DI VARIAZIONE

differenza tra MAX e MIN

- CAMPO DI VARIAZIONE $= x_{(n)} - x_{(1)}$

SCARTO INTERQUARTILE

$S.I. = Q_{0.75} - Q_{0.25}$

MAD median absolute deviation

$MAD = \text{Mediana}(|x_1 - Me_x|, \ldots, |x_n - Me_x|)$ $\quad Me_x = \text{MEDIANA}(x_1, \ldots x_n)$

COEFFICIENTE DI VARIAZIONE

$C.V. = \frac{\sigma}{\overline{x}}$ → SCARTO QUAD MEDIO / → MEDIA ARITMETICA $= \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{\overline{x}}\right)^2}$

STANDARDIZZAZIONE DEI DATI

trasformare un insieme di dati in modo tale che abbia MEDIA NULLA e VARIANZA UNITARIA (↳ VAR(z) = 1)

$z_i = \frac{x_i - (\overline{x})}{\sigma}$ $\quad i = 1, \ldots, n$ (N. DEI DATI) $\quad$ STANDARDIZZATI

DISUGUAGLIANZA DI CHEBYSHEV

Siano $x_1 \dots x_n$ con $\bar{x}$ e σ. per qualsiasi valore di $K \geq 1$ vale

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(|x_i - \bar{x}| > k\sigma) \leq \frac{1}{k^2}$$

DIMOSTRAZIONE

$$\underbrace{n\sigma^2}_{\text{VAR}(x)\cdot\text{N. DI OSSERVAZ.}} = \underbrace{\sum_{i=1}^{n}(x_i - \bar{x})^2}_{\text{VARIANZA}} = \sum_{i:x_i-\bar{x}<-k\sigma}(x_i-\bar{x})^2 + \sum_{i:|x_i-\bar{x}|\leq k\sigma}(x_i-\bar{x})^2 + \sum_{i:x_i-\bar{x}>k\sigma}(x_i-\bar{x})^2$$

$$\geq \sum_{i:x_i-\bar{x}<-k\sigma}(x_i-\bar{x})^2 + \sum_{i:x_i-\bar{x}>k\sigma}(x_i-\bar{x})^2$$

$$\geq \sum_{i:x_i-\bar{x}<-k\sigma}(k\sigma)^2 + \sum_{i:x_i-\bar{x}>k\sigma}(k\sigma)^2 =$$

$$= \sum_{i:|x_i-\bar{x}|>k\sigma}(k\sigma)^2 = k^2\sigma^2 \sum_{i=1}^{n}\mathbb{1}(|x_i-\bar{x}|>k\sigma)$$

STURGES

$$n^\circ \text{ intervalli} = 1 + \log_2 n$$

DIACONIS & FREEDMAN

$$n^\circ \text{ intervalli} = \frac{x_n - x_1}{2(Q_{0.75} - Q_{0.25})}\, n^{1/3}$$

DENSITÀ DI UN INTERVALLO

$$d_j = \frac{n_j}{z_j - z_{j-1}} \quad j = 1, \dots, k$$

n_j ← FREQ ASSOLUTE

$z_j - z_{j-1}$ ↳ lunghezza intervalli

COSTRUZIONE ISTOGRAMMA

(base rettangoli) = (lunghezza intervalli)

(altezza rettang.) = $\lambda \cdot d_j$ ← lambda, $\lambda > 0$, num. qualsiasi

$\lambda = \frac{1}{n}$

BAFFI BOXPLOT (lunghezza)

$\lambda \times (Q_{0.75} - Q_{0.25})$ dalla scatola

↳ 1.5 costante arbitraria

DISTRIBUZIONE SIMMETRICA

$\bar{x} = Me$ MEDIA MEDIANA

DISTRIBUZIONE ASIMMETRICA

ASIM. POSITIVA $\bar{x} - Me > 0$

ASIM. NEGATIVA $\bar{x} - Me < 0$

INDICE DI ASIMMETRIA DI PEARSON

$$\gamma = \frac{1}{\text{sqm}(x)^3}\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{\sigma}\right)^3$$

- se $\bar{x} = Me$ → SIMMETRICO $\gamma = 0$
- se $\bar{x} > Me$ → ASIMMETRICA POSITIVA $\gamma > 0$
- se $\bar{x} < Me$ → ASIMMETRICA NEGATIVA $\gamma < 0$

INDICE DI ASIMMETRIA DI BOWLEY

$$B = \frac{(Q_{0.75} - Me) + (Q_{0.25} - Me)}{Q_{0.75} - Q_{0.25}}$$

$B = 0$ → simmetrica: stanno a Me
$B > 0$ → asimmetria positiva
$B < 0$ → asimmetria negativa

INDICE DI CURTOSI DI PEARSON

$$K = \frac{1}{\Delta a H(x)^4} \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^4 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma} \right)^4$$

↳ rapporto tra due indici di variabilità

$K \geq 0$
$K = 0$ se i dati sono costanti

INDICE DI MUTABILITÀ DI GINI

$$G = \sum_{j=1}^{k} f_j (1 - f_j) = 1 - \sum_{j=1}^{k} f_j^2$$

↳ frequenze relative

MINIMA MUTABILITÀ

$$G = 1 - \sum_{j=1}^{k} f_j^2 = 1 - (0^2 + \ldots + 1^2 + \ldots + 0^2) = 1 - 1 = 0$$

MASSIMA MUTABILITÀ

$$G = 1 - \sum_{j=1}^{k} \frac{1}{k^2} = 1 - \frac{k}{k^2} = 1 - \frac{1}{k} = \frac{k-1}{k}$$

DISTANZA DI HAMMING

(fra x_i e x_j)

$$\mathbb{1}(x_i \neq x_j) = \begin{cases} 0 & \text{se } x_i = x_j \\ 1 & \text{se } x_i \neq x_j \end{cases}$$

DIMOSTRAZIONE INDICE DI GINI

$$G = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{1}(x_i \neq x_j) = \frac{1}{n^2} \sum_{i=1}^{h} \sum_{j=1}^{K} n_j \mathbb{1}(x_i \neq c_j) = \frac{1}{n^2} \sum_{i=1}^{K} \sum_{j=1}^{K} n_i n_j \mathbb{1}(c_i \neq c_j) =$$

$$= \frac{1}{n^2} (0 \times n_1 n_1 + n_1 n_2 + \ldots + 1 \times n_1 n_K + n_2 n_1 + 0 \times n_2 n_2 + \ldots + n_2 n_K + \ldots + 0 \times n_K n_K) =$$

$$= \frac{1}{n^2} [n_1 (n - n_1) + n_2 (n - n_2) + \ldots + n_K (n - n_K)] =$$

$$= \frac{1}{n^2} \sum_{j=1}^{k} n_j (n - n_j) = 1 - \frac{1}{n^2} \sum_{j=1}^{k} n_j^2 = 1 - \sum_{j=1}^{k} f_j^2$$

GINI ALTERNATIVO

$$G = 1 - \frac{1}{n^2} \sum_{j=1}^{k} n_j^2$$

GINI NORMALIZZATO

$$G_{NORM} = \frac{K}{K-1} G$$

ENTROPIA DI SHANNON

$$H = -\sum_{j=1}^{K} f_j \log f_j$$

↳ freq. relative

SHANNON NORMALIZZATA

$$H_{NORM} = \frac{H}{\log K}$$

COVARIANZA

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

TRASFORMAZIONE LINEARE COV

$$cov(v,w) = b_x b_y (x,y)$$

VARIANZA DI UNA SOMMA DI DUE VARIABILI

$$Var(w) = Var(x) + Var(y) + 2cov(x,y)$$

VARIANZA CON COVARIANZA

$$Var\left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y}\right) = 2\left(1 + \frac{1}{\sigma_x \sigma_y} cov(x,y\right.$$

COEFFICIENTE DI CORRELAZIONE

$$\rho = cov(x,y) = \frac{cov(x,y)}{sqm(x)\,sqm(y)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right)$$

RO

È LA COVARIANZA STANDARDIZZATA

FUNZIONE CHE MINIMIZZA LA PIANTA

$$\ell(\alpha,\beta) = \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)$$

REGRESSIONE LINEARE

$y \approx \alpha + \beta x + \varepsilon i$ $\quad (\hat{\alpha},\hat{\beta})$ = valori che minimizzano la funzione di perd

Se COSE PRESE SONO IL MENO POSSIBILE

$$(\hat{\alpha},\hat{\beta}) = \underset{\alpha\beta}{\operatorname{argmin}} \sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2 = \begin{matrix} \alpha = \bar{y} - \hat{\beta}\bar{x} \\ \beta = \dfrac{cov(x,y)}{Var(x)} \end{matrix}$$

PUNTO MINIMO

↳ VALORE = $\alpha + \beta$(diametro) + (errore)

RESIDUI

$$r_i = y_i - (\hat{\alpha} + \hat{\beta} x_i) \qquad i = 1 \ldots n$$

- VARIANZA DEI RESIDUI UTILIZZATA X VALUTARE LA BONTÀ DI ADATTAMENTO DEL MODELLO AI DATI
- LA MEDIA DEI RESIDUI È NULLA

COEFFICIENTE DI DETERMINAZIONE 2

$$R^2 = \frac{cov(x,y)^2}{VAR(x)\,VAR(y)} = cor(x,y)^2$$

COEFFICIENTE DI DETERMINAZIONE R^2

$$R^2 = 1 - \frac{VAR(x)}{VAR(y)} = cor(x,y)^2$$

$$0 \le R^2 \le 1$$

CORRELAZIONE PARZIALE

$$\tilde{x} = x,z \rightarrow x_i - \bar{x} - \frac{Cov(x,z)}{Var(z)}(z_i - \bar{z})$$

$$\tilde{y} = y,z \rightarrow y_i - \bar{y} - \frac{Cov(y,z)}{Var(z)} z_i - \bar{z}$$

MEDIE DEI GRUPPI

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} \qquad j = 1, \dots, k$$

MEDIA TOTALE

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{k} n_j \bar{x}_j = \frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$$

1° GRUPPO, 2° GRUPPO

DEVIANZA

$$D_{tr}^2 = \sum_{j=1}^{k} n_j (\bar{x}_j - \bar{x})^2$$

SE = 0 ALORA DEVIANZA ZERO

DEVIANZA INTRO I GRUPPI

$$d_j^2 = \sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2 \qquad j = 1, \dots, k$$

$$D_{en}^2 = \sum_{j=1}^{k} d_j^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{x}_j)^2$$

DEVIANZA TOTALE

$$D^2 = \sum_{j=1}^{k}\sum_{i=1}^{n_j} (x_{ij} - \bar{x})^2$$

→ NON CI'

- SCARTI DI TUTTE LE OSSERVAZIONI DALLA MEDIA TOTALE

TRONCA SCOMPOSIZIONE DEVIANZA

$$D^2 = D_{tr}^2 + D_{en}^2$$

D_{tr}^2 → TRA I GRUPPI

D_{en}^2 ← INTRO I GRUPPI

$$0 \le D_{tr}^2 \le D^2$$

MASSIMO TRONCO

RAPPORTO DI CORRELAZIONE

$$\eta^2 = \frac{(\text{DEVIANZA TRA I GRUPPI})}{(\text{DEVIANZA TOTALE})} = 1 - \frac{(\text{DEVIANZA INTRO I GRUPPI})}{(\text{DEVIANZA TOTALE})}$$

$$\eta^2 = \frac{D_{tr}^2}{D^2} = 1 - \frac{D_{en}^2}{D^2}$$

→ MISURA LA FORZA DELLA DIPENDENZA IN MEDIA

INDICE NORMALIZZATO POICHÉ

$$0 \le \eta^2 \le 1$$

INDICE X^2 DI PEARSON

$$X^2 = \sum_{i=1}^{h}\sum_{j=1}^{k} \frac{(n_{ij} - \hat{n}_{ij})^2}{\hat{n}_{ij}} = n\left(\sum_{i=1}^{h}\sum_{j=1}^{k} \frac{F_{ij}^2}{F_{i+}F_{+j}} - 1\right)$$

MEDIE CONDIZIONATE

$$\bar{y}_i = \frac{1}{n_{i+}} \sum_{j=1}^{k} n_{ij} d_j \quad i = 1..h$$

DIFFERENZA SEMPLICE MEDIA

$$\Delta = \frac{1}{n(n-1)} \sum_{i=1}^{n}\sum_{j=1}^{n} |x_i - x_j|$$

OPPURE $\Delta = \frac{4}{n(n-1)}\left(\sum_{i=1}^{n} i\, x_{(i)}\right) - 2\bar{x}\frac{n+1}{n-1}$

TRASFORMAZ LINEARE Δ

$$\Delta_y = |b| \Delta_x$$

MAX e MIN Δ

$$\Delta_{MAX} = 2\bar{x}$$
$$\Delta_{MIN} = 0$$

Δ NORM

$$R = \frac{\Delta}{2\bar{x}}$$

$0 \le R \le 1$

TRASFORMAZIONE LINEARE R

$$R_x = R_y$$

CURVA DI LORENZ

$$q_i = \frac{1}{S} \sum_{j=1}^{i} x_j$$

RICCHEZZA

$$R = \frac{n-1}{n}$$