

STATISTICA DESCRITTIVA:

MODELLI STATISTICI

Campione Casuale, Distribuzione Modello e DATASET

- UN CAMPIONE CASUALE È UNA COLLEZIONE DI V.A. X_1, X_2, \dots, X_m INDIPENDENTI E IDENTICAMENTE DISTRIBUITE (i.i.d) COME X OVVERO AVENTI FUNZIONE DI DISTRIBUZIONE F_X . m È L'AMPIZZA DEL CAMPIONE.
- LA DISTRIBUZIONE DI PROBABILITÀ F_X DI X , E QUINDI DI OGUNA DELLE V.A. DEL CAMPIONE CASUALE È CHIAMATA DISTRIBUZIONE MODELLO.
- UN DATASET (o CAMPIONE OSSERVATO) CONSISTE IN RIPETUTE MISURAZIONI DI UNA CERTA QUANTITÀ X_1, X_2, \dots, X_m CHE VENGONO INTERPRETATE COME REALIZZAZIONI DI UN CAMPIONE CASUALE X_1, X_2, \dots, X_m .

ESEMPIO:

CAMPIONE CASUALE: $X_1, X_2, X_3, X_4, X_5, X_6, \dots, X_m$ VETTORE DI V.A. (i.i.d). $X \sim \text{BER}(p)$.

POSSIBILE REALIZZAZIONE: $X_1, X_2, X_3, X_4, X_5, X_6, \dots, X_m$ DATASET: È UNA REALIZZAZIONE DI QUESTO VETTORE.

0 1 1 0 1 0, ..., 1

UN DETERMINATO ESITO DI QUESTI m ESPERIMENTI INIDPENDENTI.

CONSIDERAZIONI: COME PRIMA COSA SI RIORDINA IL DATASET, E OTENIAMO:

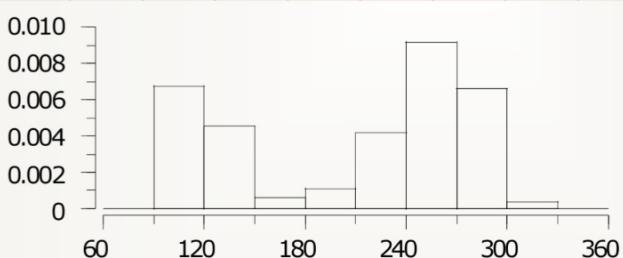
- DATASET ORIGINALE: X_1, X_2, \dots, X_m .
- DATASET ORDINATO: $X_{(1)}, X_{(2)}, \dots, X_{(m)}$.
- MINIMO $X_{(1)}$ E MASSIMO $X_{(m)}$: GLI ESTREMI DEL DATASET RIORDINATO.
- RANGE DEL DATASET $[X_{(1)}, X_{(m)}] \subseteq \mathbb{R}$.

SINGIFICATO STATISTICO: "PARTO DAI DATI E VOGLIO TROVARE IL MODELLO CHE È PIÙ ADATTO A DESCRIVERE QUEI DATI."

Costruzione di un Istogramma

(ORDINATO)
DATO UN DATASET $X_{(1)}, X_{(2)}, \dots, X_{(m)}$ SUDDIVIDIAMO IL RANGE DEI DATI IN INTERVALLI B_1, \dots, B_m CHIAMATI BINS, ABBIAMO:

- AREA RETTANGOLO = FRAZIONE DEI DATI NEL BIN CORRISPONDENTE.
- AMPIZZA $B_i = |B_i|$
- ALTEZZA $B_i = \frac{|x_i \in B_i|}{m \cdot |B_i|}$



ESERCIZIO: 1 2 3 3 3 5 6 9 10, $m = 9$, $|B_i| = 2$

• $B_1 = [0, 2)$, $B_3 = [4, 6)$, $B_5 = [6, 8)$, $B_7 = [8, 10)$,

$B_6 = [10, 12)$: ALTEZZA = $\frac{1}{9} \cdot \frac{1}{2} = \frac{1}{18}$.

• $B_2 = [2, 4)$: ALTEZZA = $\frac{2}{9} \cdot \frac{1}{2} = \frac{2}{18}$.



Funzione di Distribuzione Empirica

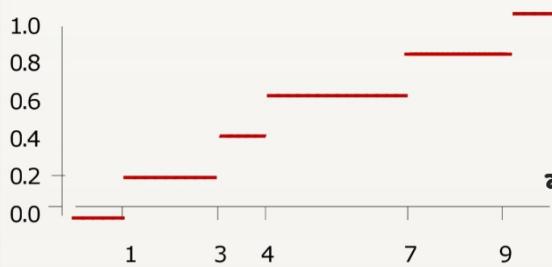
DATO UN DATASET x_1, x_2, \dots, x_m LA FUNZIONE DI DISTRIBUZIONE EMPIRICA F_m È DEFINITA DA:

$$F_m(\alpha) = \frac{\text{NUMERO ELEMENTI DEL DATASET} \leq \alpha}{m} = \frac{|x_i \leq \alpha|}{m}$$

SI TRATTA DELL'ANALOGO EMPIRICO DELLA FUNZIONE DI DISTRIBUZIONE: $F_x(x) = P(X \leq x)$.

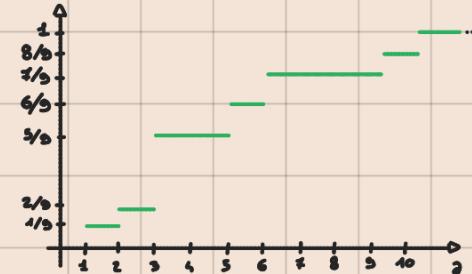
Esempi calcolo della Fn

ESEMPIO CALCOLO F_m : 1 3 4 7 9



ESEMPIO CALCOLO F_m : 1 2 3 3 3 5 6 9 10

$$F_m(\alpha) = \begin{cases} 0 & \text{SE } \alpha < 1 \\ \frac{1}{9} & \text{SE } 1 \leq \alpha < 2 \\ \frac{2}{9} & \text{SE } 2 \leq \alpha < 3 \\ \frac{5}{9} & \text{SE } 3 \leq \alpha < 5 \\ \frac{6}{9} & \text{SE } 5 \leq \alpha < 6 \\ \frac{7}{9} & \text{SE } 6 \leq \alpha < 9 \\ \frac{8}{9} & \text{SE } 9 \leq \alpha < 10 \\ 1 & \text{SE } \alpha \geq 10 \end{cases}$$



Analisi dei dati esplorativa: Osservabili Numeriche

INDICI DI POSIZIONE

MEDIA EMPIRICA:

$$\bar{x}_m = \langle x \rangle = \frac{1}{m} \cdot \sum_{i=1}^m x_i$$

ANALOGO IN PROBABILITÀ

$$E[X]$$

MOMENTI EMPIRICI:

$$\langle x^K \rangle = \frac{1}{m} \cdot \sum_{i=1}^m x_i^K$$

$$E[X^K]$$

MEDIANA EMPIRICA:

$$\text{MED}_m = \text{MED}(x_1, \dots, x_m)$$

INDICI DI DISPERSIONE

VARIANZA EMPIRICA:

$$d_m^2 = \frac{1}{m} \cdot \sum_{i=1}^m (x_i - \bar{x}_m)^2$$

$$E[(X - E[X])^2]$$

DEVIAZIONE STANDARD EMPIRICA:

$$d_m = \sqrt{d_m^2}$$

$$\sqrt{\text{VAR}(X)}$$

DEVIAZIONE MEDIANA ASSOLUTA:

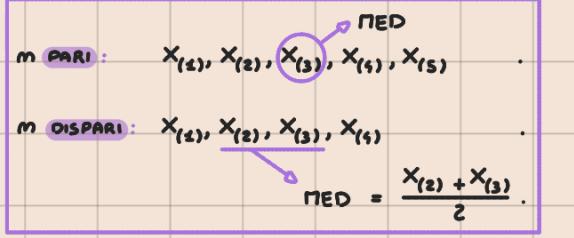
$$\text{MAD}(x_1, \dots, x_m) = \text{MED}(|x_1 - \text{MED}_m|, \dots, |x_m - \text{MED}_m|)$$

ANALOGO IN PROBABILITÀ

Esempio esercizio sulle osservabili numeriche

DATI: 2 8 3 16 , ORDINATO: 1 2 3 6 8 , $m = 5$

$$\bullet \quad x_{(1)} = 1, \quad x_{(5)} = 8, \quad \text{MED}_5 = 3.$$



$$\bullet \quad \bar{x}_m = \frac{1}{5} \cdot (1+2+3+6+8) = \frac{19}{5} = 3.8.$$

$$\bullet \quad d_m^2 = \frac{1}{5} \cdot [(1-3.8)^2 + (2-3.8)^2 + (3-3.8)^2 + (6-3.8)^2 + (8-3.8)^2] = \frac{1}{5} \cdot (7.84 + 3.24 + 0.64 + 4.84 + 17.64) = 6.84$$

$$\bullet \quad \text{MAD}(1, 2, 3, 6, 8) = \text{MED}(1-31, 12-31, 13-31, 16-31, 18-31) = \text{MED}(0, 1, 2, 3, 5) = 2.$$

RIORDINO

Quantili Empirici

SIA x_1, x_2, \dots, x_m UN DATASET, E $p \in [0, 1]$, ALLORA p -QUANTILE EMPIRICO IL VALORE $q_m(p)$ TALE CHE:

- UNA PROPORZIONE p DEL DATASET È MINORE DI $q_m(p)$.
- UNA PROPORZIONE $1-p$ DEL DATASET È MAGGIORI DI $q_m(p)$.

$q_m(0.5)$	QUARTILE INFERIORE
$q_m(0.25)$	MEDIANA EMPIRICA
$q_m(0.75)$	QUARTILE SUPERIORE

PROCEDURA CALCOLO DEI QUANTILI: SIA $x_{(1)}, x_{(2)}, \dots, x_{(m)}$ IL DATASET ORDINATO.

DATI: 0 1 4 4 4 5 5 6 6 8 12 , $m = 11$, $p = 0.2$, $q_m(p) = ?$

1) CALCOLO: $p \cdot (m+1) = 0.2 \cdot 12 = 2.4$

$$p(m+1) = K + r$$

2) OTTENGO: $K = \lfloor p \cdot (m+1) \rfloor = 2$ E $r = 0.4$

$$q_m(p) = x_{(K)} + r \cdot (x_{(K+1)} - x_{(K)})$$

3) CALCOLO: $q_m(p) = x_{(K)} + r \cdot (x_{(K+1)} - x_{(K)}) = 1 + 0.4(4-1) = 1 + 1.2 = 2.2$

DOVE K È LA PARTE INTERA DI $p(m+1)$ MENTRE r È LA PARTE DECIMALE.

CALCOLO DELLA MEDIANA CIOÈ QUANTILE DI 0.5:

• MEDIANA CON m DISPARI: $\text{MED}_m = q_m(0.5) = x_{\left(\frac{m+1}{2}\right)} + 0$

• MEDIANA CON m PARI: $\text{MED}_m = q_m(0.5) = x_{\left(\frac{m}{2}\right)} + \frac{1}{2} \cdot (x_{\left(\frac{m+1}{2}\right)} - x_{\left(\frac{m}{2}\right)}) = \frac{x_{\left(\frac{m+1}{2}\right)} + x_{\left(\frac{m}{2}\right)}}{2}$

Sintesi a cinque numeri e visualizzazione a Boxplot

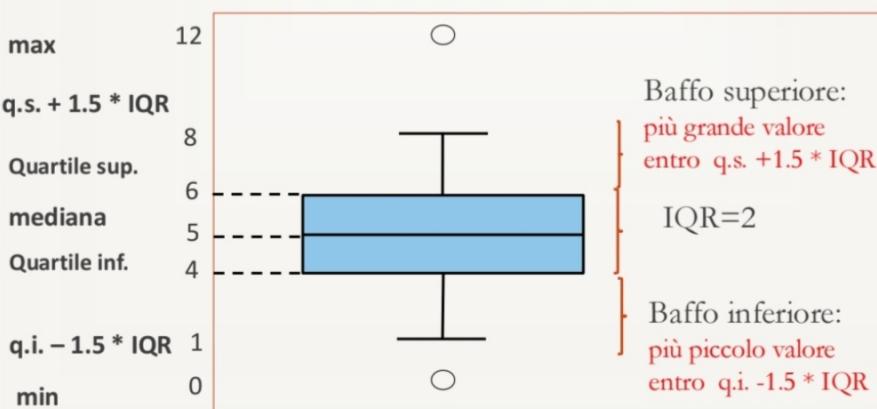
LA SINTESI A 5 NUMERI CONSISTE NELLE SEGUENTI CINQUE OSSERVABILI:

- | | | |
|-----------------------|-----------------------|------------|
| 1) Minimo | 3) Mediana | 5) Massimo |
| 2) Quantile inferiore | 4) Quantile superiore | |

Boxplot: visualizzare la sintesi a cinque numeri

Per il dataset 0,1,4,4,4,5,5,6,6,8,12 :

Min=0, Quartile Inf.=4, Mediana=5, Quartile Sup.=6, Max=12



CALCOLI SUI QUANTILI: $m = 11$

$$\begin{aligned} q.i. &= q_m(0.25) = x_{(3)} = 4 \\ (m+1) \cdot p &= 12 \cdot 0.25 = 3 \quad K = 3 \quad r = 0 \end{aligned}$$

$$\begin{aligned} q.o. &= q_m(0.75) = x_{(9)} = 6 \\ (m+1) \cdot p &= 12 \cdot 0.75 = 9 \quad K = 9 \quad r = 0 \end{aligned}$$

INTER-QUANTILE-RANGE (I.Q.R.) = $q.o. - q.i. = 6 - 4 = 2$

- I.Q.R. = $q.o. - q.i. = 6 - 4 = 2$
- $q.o + 1.5 \cdot \text{IQR} = 6 + 1.5 \cdot 2 = 9 = k$
- $q.i. - 1.5 \cdot \text{IQR} = 4 + 1.5 \cdot 2 = 3 = k$