

LA STATISTICA DESCRITTIVA SI OCCUPA DI RACCOLGERE E ORGANIZZARE DATI, RIASSUMENDONE LE PROPRIETÀ SALIENTI ATTRAVERSO INDICATORI SINTETICI.

DIAMO SUBITO QUALCHE DEFINIZIONE

- SI DICE UNITÀ STATISTICA LA MINIMA UNITÀ DELLA QUALE SI RACCOLGONO I DATI
- SI DICE POPOLAZIONE L'INSIEME DELLE UNITÀ STATISTICHE OGGETTO DI STUDIO
- SI DICE CAMPIONE UNA PORZIONE DELLA POPOLAZIONE
- SI DICONO CARATTERI LE PROPRIETÀ CHE SONO OGGETTO DI RILEVAZIONE.

I CARATTERI POSSONO ESSERE QUALITATIVI O QUANTITATIVI

I CARATTERI QUALITATIVI VENGONO INDICATI ATTRAVERSO ESPRESSIONI VERBALI. VENGONO SUDDIVISI IN CATEGORIE.

PER ESEMPIO LO STATO CIVILE (CELIBELNUBILE, CONIUGATO/A), IL SESSO (MASCHIO, FEMMINA), IL COLORE DEGLI OCCHI (VERDI, AZZURRI, CASTANI ECC).

I CARATTERI QUANTITATIVI SONO ESPRIMIBILI NUMERICAMENTE E SI DIVIDONO IN DISCRETI E CONTINUÏ. I CARATTERI DISCRETI ASSUMONO SOLO UNA QUANTITÀ AL PIÙ NUMERABILE DI VALORI,

DI SOLITO NUMERI INTERI. PER ESEMPIO IL NUMERO DI STUDENTI IN UNA CLASSE, I PUNTI EFFETTUATI IN UNA PARTITA, IL NUMERO DI ABITANTI. I CARATTERI CONTINUI POSSONO INVECE ASSUMERE QUALSIASI VALORE REALI IN UN DATO INTERVALLO. PER ESEMPIO IL PESO, L'ALTEZZA.

ESEMPIO SUPPONIAMO DI AVERE UN GRUPPO DI 200 FAMIGLIE (CAMPIONE IN ESAME), DI CUI RILEVIAMO IL SEGUENTE CARATTERE: "TITOLO DI STUDIO DEL CAPOFAMIGLIA". TALE CARATTERE È DI TIPO QUALITATIVO. LO SUDDIVIDIAMO IN 5 CATEGORIE. QUESTO IL RISULTATO

Nessun titolo	18	0.090
Licenza elementare	52	0.260
Diploma scuola media inferiore	74	0.370
Diploma scuola media superiore	49	0.245
Laurea	7	0.035
	200	1.000

NELLA COLONNA DI DESTRA LA FRAZIONE SUL TOTALE.

ESEMPIO In uno stabilimento vengono registrati gli episodi di malfunzionamento di un macchinario, insieme alle cause. I dati annuali sono

fluttuazioni di tensione:	6
instabilità del sistema di controllo:	22
errore dell'operatore:	13
strumento consumato e non sostituito:	2
altre cause:	5
Totale:	48

L'unità statistica è il singolo malfunzionamento, il campione i malfunzionamenti avvenuti durante l'anno.

Il carattere in esame è "episodi di malfunzionamento". È di tipo qualitativo ed è stato diviso in 5 categorie.

ESEMPIO I diametri di 20 sfere prodotte da una linea produttiva sono stati misurati. Le misure, espresse in cm, sono

2.08, 1.72, 1.90, 2.11, 1.79, 1.86, 1.80, 1.91, 1.82, 1.84,

2.04, 1.86, 2.04, 1.80, 1.82, 2.08, 2.04, 1.85, 2.07, 2.03.

Il carattere in esame, "diametro", è di tipo quantitativo continuo.

OLTRE A RACCOLGERE I DATI, DOBBIAMO ANCHE ORGANIZZARLI. UNO DEI MODI CONSISTE NEL FORNIRE LA TABELLA DI DISTRIBUZIONE DI FREQUENZA, CIÒ È DIVIDIAMO L'INSIEME DEI DATI IN CLASSI E CONTIAMO IL NUMERO DI OSSERVAZIONI CHE CADONO IN CIASCUNA DI ESSI. QUESTA VIENE CHIAMATA FREQUENZA ASSOLUTA. DETTO m IL NUMERO DI DATI, N IL NUMERO DI CLASSI ED $f_a(k)$ LA FREQUENZA ASSOLUTA NELLA k -ESIMA CLASSE, OVVIAMENTE

$$\sum_{k=1}^N f_a(k) = m$$

CHIAMIAMO INVECE FREQUENZA RELATIVA IL RAPPORTO $f_r(k) = f_a(k)/m$. OVVIAMENTE $\sum_{k=1}^N f_r(k) = 1$.

SIMILMENTE, CHIAMIAMO FREQUENZA PERCENTUALE LA QUANTITÀ $f_p(k) = f_r(k) \cdot 100$.

LA FREQUENZA ASSOLUTA CUMULATIVA $F_a(k)$ DELLA k -ESIMA CLASSE È IL NUMERO TOTALE DI OSSERVAZIONI CHE RICADONO NELLE CLASSI FINO ALLA k -ESIMA COMPRESA:

$$F_a(k) = \sum_{j=1}^k f_a(j)$$

OVVIAMENTE F_a È NON DECRESCENTE E $F_a(N) = m$.

NEL CASO DI CARATTERI QUALITATIVI, LE CLASSI COINCIDONO DI SOLITO CON LE CATEGORIE. TORNANDO AL PRIMO ESEMPIO, $N=5$ E

$$f_r(1) = 0,09 \quad f_r(2) = 0,26 \quad f_r(3) = 0,32 \quad f_r(4) = 0,245 \quad f_r(6) = 0,035$$

ANCHE NEL SECONDO ESEMPIO, $N=5$ E

$$f_r(1) = 0,125 \quad f_r(2) = 0,458 \quad f_r(3) = 0,271 \quad f_r(4) = 0,042 \quad f_r(6) = 0,104$$

SE IL CARATTERE \bar{x} DI TIPO QUANTITATIVO E DISCRETO, DETTI x_1, \dots, x_m, \dots I VALORI CHE PUO' ASSUMERE, E' NATURALE SCEGLIERE COME CLASSI GLI INSIEMI

$$A_k := \{ \text{UNITA' CON VALORE } x_k \}$$

IN QUESTO CASO $f_a(k) = |A_k|$

ESEMPIO SU UN GRUPPO DI 10 STUDENTI DI INGEGNERIA AL PRIMO ANNO, I RISULTATI ALL'ESAME DI PROBABILITA' SONO, ESPRESSI IN 30-ESIMI,

STUD 1	18	STUD 6	19
STUD 2	18	STUD 7	26
STUD 3	20	STUD 8	18
STUD 4	30	STUD 9	27
STUD 5	18	STUD 10	18

Quindi i valori assumibili sono gli interi da 18 a 30.

Abbiamo $A_{18} = \{\text{stud } 1, 2, 5, 8, 18\}$, $A_{19} = \{\text{stud } 6\}$, ...

$A_{29} = \emptyset$, $A_{30} = \{\text{stud } 4\}$.

Inoltre $f_a(18) = 5$, $f_a(19) = 1$, ..., $f_a(29) = 0$, $f_a(30) = 1$.

Nel terzo esempio il carattere \bar{x} è di tipo continuo (può assumere a priori tutti i valori compresi in un certo intervallo di numeri reali). In questo caso, i valori assunti sono compresi tra 1.70 e 2.10. Possiamo allora, per esempio, suddividere questo intervallo in sottointervalli di ampiezza 0.05 e considerare come classi i sottointervalli del campione che cadono rispettivamente in $[1.70, 1.75]$, $(1.75, 1.80]$, ..., $(2.05, 2.10]$

	f_a	f_r	f_p	F_a
1.70 - 1.75	1	0.05	5	1
1.75 - 1.80	3	0.15	15	4
1.80 - 1.85	3	0.15	15	7
1.85 - 1.90	4	0.20	20	11
1.90 - 1.95	1	0.05	5	12
1.95 - 2.00	0	0	0	12
2.00 - 2.05	4	0.20	20	16
2.05 - 2.10	3	0.15	15	19
2.10 - 2.15	1	0.05	5	20
Totale	20	1	100	

T
A
B
E
L
L
A

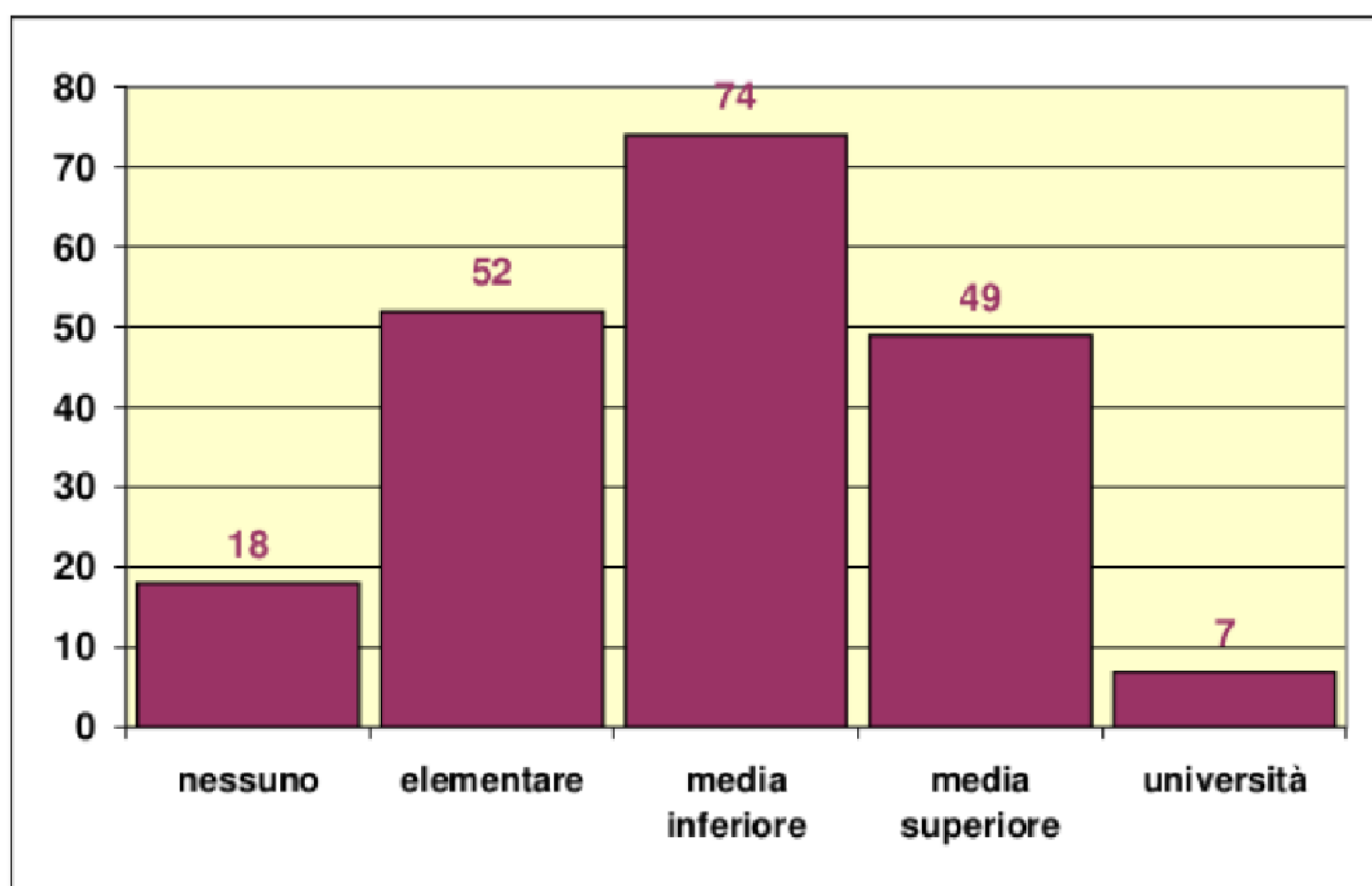
QUANTO VISTO NELL'ESEMPIO È IL NODO STANDARD DI PROCEDERE NEL CASO DI CARATTERI CONTINUI. VANNO FATTE DUE OSSERVAZIONI SULLA PROCEDURA.

- LA SCELTA DELL'AMPIEZZA DEI SOTTOINTERVALLI, E QUINDI DELLE CLASSI, NON È UNIVOCA. NELL'ESEMPIO AURENNO POTUTO PRENDERE AMPIEZZA 0,001 OPPURE 0,1. L'IMPORTANTE È CHE OGNI UNITÀ STATISTICA CADDA IN UNA E UNA SOLA CLASSE. DAL PUNTO DI VISTA PRATICO, TROPPE CLASSI RENDONO LA TABELLA POCO LEGGIBILE (PENSATE A MILIONI DI DATI E 10000 SOTTOGRUPPI). D'ALTRA PARTE POCHE CLASSI RENDONO LA TABELLA POCO UTILE (PENSATE A MILIONI DI DATI E 2 SOTTOGRUPPI). CI VUOLE EQUILIBRIO.
- NEL SUDDIVIDERE I DATI IN CLASSI PERDIAMO PARTE DELL'INFORMAZIONE: FACENDO SEMPRE RIFERIMENTO ALL'ESEMPIO, UN CONTO È SAPERE CHE TRE SFERETTE HANNO DIAMETRO 1.79, 1.80, 1.80, UN CONTO È SAPERE CHE TRE SFERETTE HANNO DIAMETRO CHE CADDE NELL'INTERVALLO $(1.75, 1.80]$. D'ALTRA PARTE, CI SI GUADAGNA IN FACILITÀ DI LETTURA E CHIAREZZA.

LE INFORMAZIONI CONTENUTE NELLA TABELLA DELLA DISTRIBUZIONE DELLE FREQUENZE POSSONO ESSERE RAPPRESENTATE GRAFICAMENTE ATTRAVERSO UN ISTOGRAMMA.

È UNA SORTA DI GRAFICO COSTITUITO DA RETTANGOLI ADIACENTI, LE CUI BASI RAPPRESENTANO LE CLASSI, MENTRE LE ALTEZZE RAPPRESENTANO LE RISPETTIVE FREQUENZE (ASSOLUTE O RELATIVE, DIPENDE DALLA SCALA SCELTA).

PER ESEMPIO, FACENDO RIFERIMENTO AL PRIMO ESEMPIO



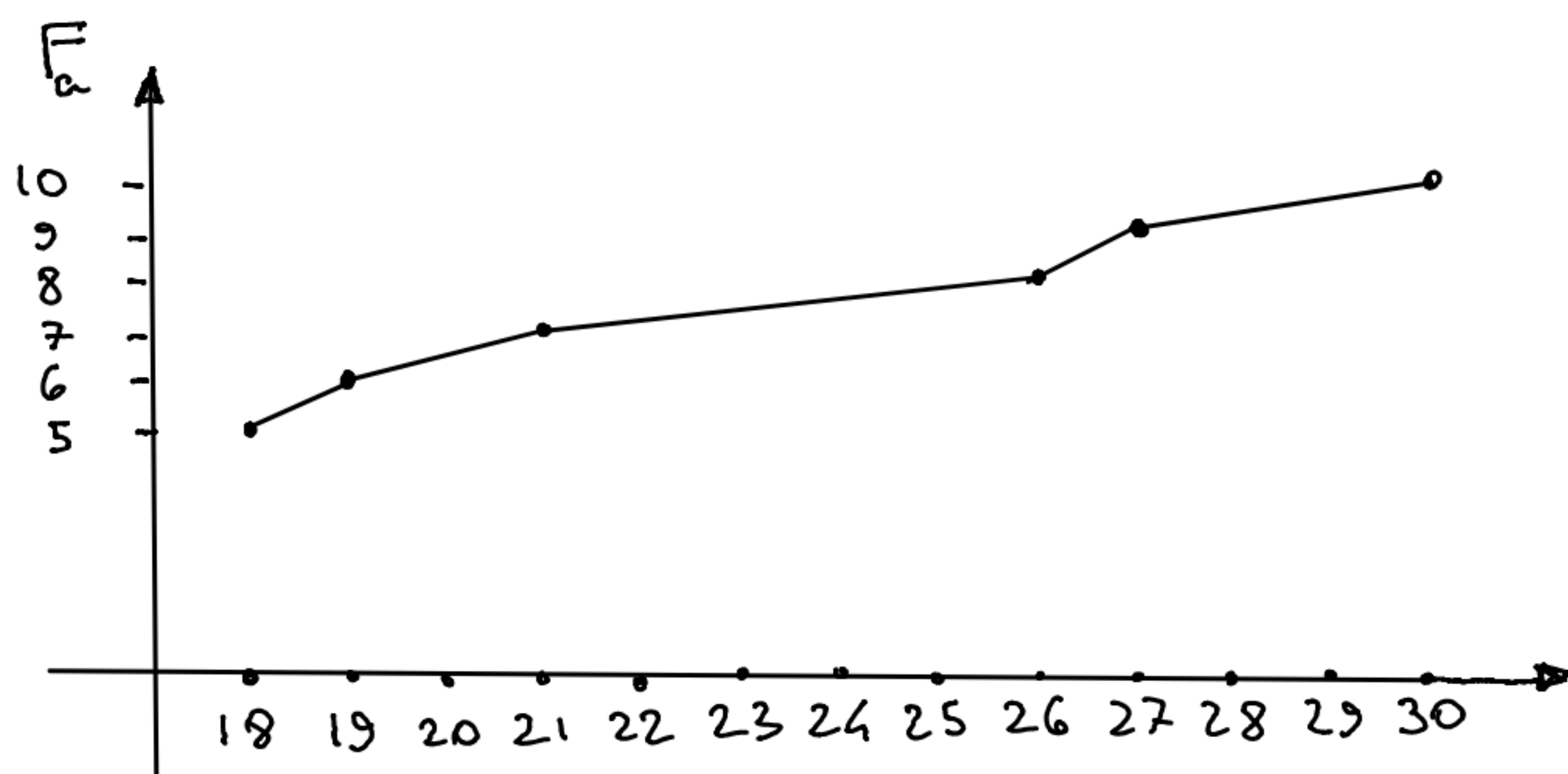
SI È USATA LA FREQUENZA ASSOLUTA IN QUESTO ISTOGRAMMA.

NEL CASO DI CARATTERI QUALITATIVI OPPURE QUANTITATIVI DISCRETI LE BASI DEI RETTANGOLI SONO TUTTE UGUALI.

NEL CASO DI CARATTERI QUANTITATIVI CONTINUI LE BASI SONO

COSTITUITE DAGLI INTERVALLINI IN CUI SUDDIVIDIANO I DATI.
(E ANCHE QUI, DI SOLITO, SONO UGUALI).

LA FREQUENZA CUMULATIVA DI UN CARATTERE QUANTITATIVO (DISCRETO O CONTINUO) PUÒ ESSERE RAPPRESENTATO CON UN GRAFICO, DETTO OGIVA. SULL'ASCISSE SI PONGONO NEL CASO DISCRETO I VALORI OSSERVATI (DI SOLITO INTERI), MENTRE NEL CASO CONTINUO SI PONGONO GLI ESTREMI DEGLI INTERVALLINI USATI. IN ORDINATA SI RIPORTANO LE FREQUENZE CUMULATIVE CORRISPONDENTI UNITE ATTRAVERSO UNA SPEZZATA.



NELL'ESEMPIO DEI VOTI VISTO PRECEDENTEMENTE, AVEVANO

$$F_a(18)=5, F_a(19)=6, F_a(21)=7, F_a(26)=8, F_a(27)=9, F_a(30)=10$$

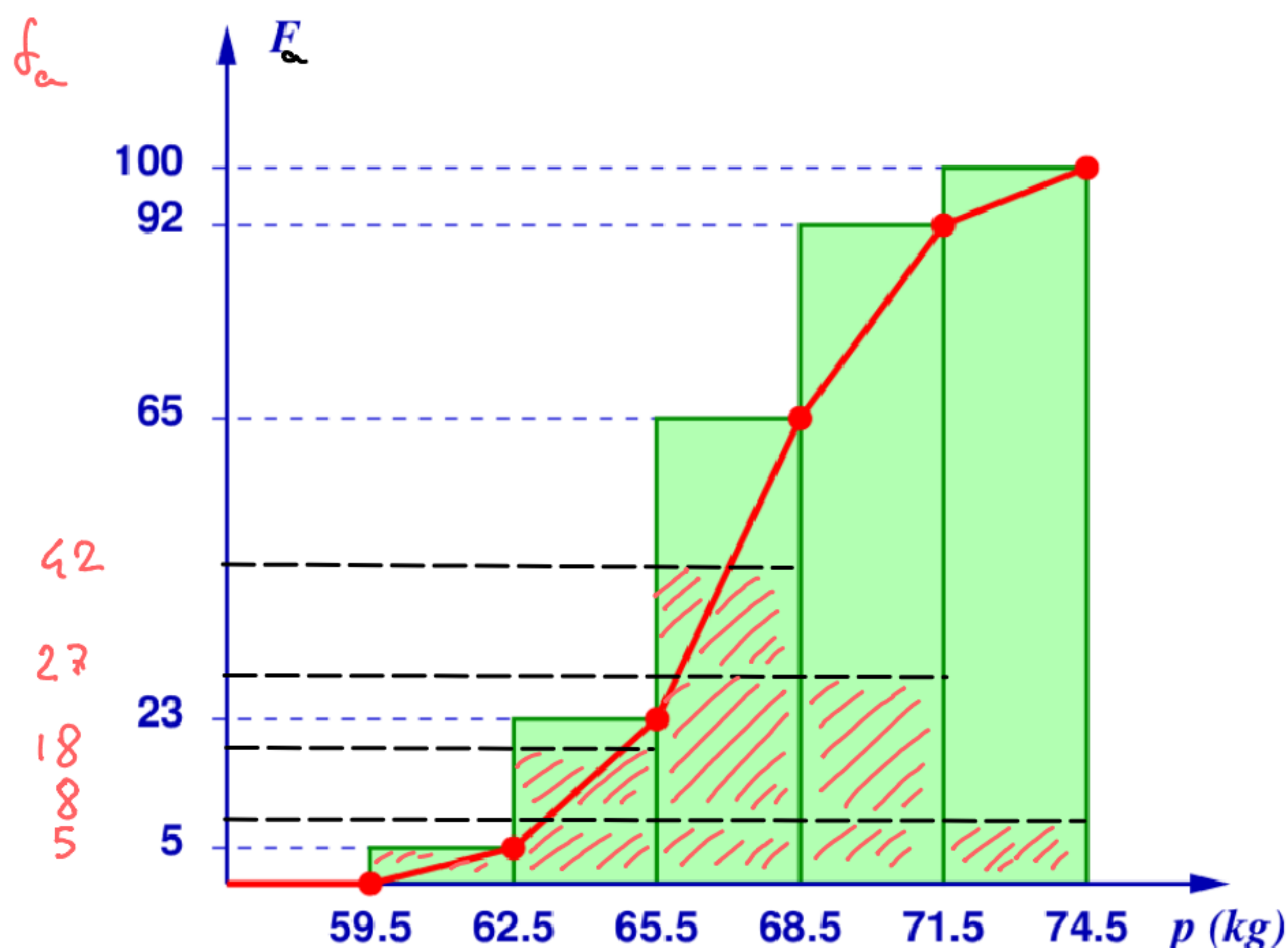
L'OGIVA RELATIVA È QUELLA RAPPRESENTATA IN FIGURA.


ESEMPIO Un'indagine sul peso, condotta su un campione di $n=100$ studenti, ha prodotto il seguente risultato. I pesi p sono espressi in kg e sono stati raggruppati in $N=5$ classi di peso.

CLASSE	f_a	f_r	F_a
$59.5 < p \leq 62.5$	5	0,05	5
$62.5 < p \leq 65.5$	18	0,18	23
$65.5 < p \leq 68.5$	42	0,42	65
$68.5 < p \leq 71.5$	27	0,27	92
$71.5 < p \leq 74.5$	8	0,08	100

T
A
B
E
L
L
A

Si tratta di carattere quantitativo continuo. Sono stati scelti intervalli di ampiezza $3 kg$.



Istogramma 
e OGIVA.

UNA VOLTA CHE ABBIAMO RACCOLTO E ARGANIZZATO I DATI, DOBBIAMO ANCHE FORNIRE ALCUNI INDICATORI SINTETICI CHE FORNISCONO UN'IDEA DI MASSIMA DELLA SITUAZIONE, CIOÈ DI DOVE (INDICI DI POSIZIONE) E COME (INDICI DI DISPERSIONE) I DATI SONO DISTRIBUITI. GLI INDICI DI POSIZIONE PIÙ USATI SONO LA MEDIA, LA MEDIANA E LA MODA. CI DICONO ATTORNO A QUALE VALORE IL CAMPIONE È "CENTRATO".

- SI DICE MODA LA CATEGORIA/CLASSE CUI CORRISPONDE LA MASSIMA FREQUENZA.

NEGLI ESEMPI PRECEDENTI, RISPETTIVAMENTE,

- DIPLOMA SCUOLA MEDIA INFERIORE
- INSTABILITÀ DEL SISTEMA DI CONTROLLO
- INTERVALLI $(1.85, 1.90]$ E $(2.00, 2.05]$ (DIAMETRI SFERETTE)
- IL VOTO 18
- INTERVALLO $[65.5, 68.5]$ (PESO STUDENTI)

ESEMPIO SI RILEVA IL NUMERO DI STANZA DI CIASCUN APPARTAMENTO DI UN CONDOMINIO

NUMERO STANZE	2	3	4	5	6	7
FREQUENZA ASSOLUTA	1	3	8	2	1	1

LA MODA È 4.

- Si dice MEDIANA il valore che occupa il posto di mezzo, quando i dati sono disposti in ordine crescente e sono in numero dispari, oppure la media aritmetica dei due valori in posizione centrale quando sono in numero pari

ESEMPIO LE MISURE OTTENUTE SU UN CAMPIONE SONO

18, 6, 31, 71, 84, 17, 23, 1, 9, 43

LE ORDINIAMO

1, 6, 9, 17, 18, 23, 31, 43, 71, 84

POICHÉ $n=10$, QUINDI PARI, LA MEDIANA È $\frac{18+23}{2} = 20.5$

SE INVECE LE MISURE OTTENUTE SONO

4, 5, 5, 6, 7, 8, 9

ALLORA LA MEDIANA È 6 (IN QUESTO CASO $n=7$, DISPARI)

MEDIANA PER DATI RAGGRUPPATI. SE I DATI SONO STATI

GIÀ SUDDIVISI IN CLASSI E SI CONOSCE SOLO LA FREQUENZA,

ALLORA SI PUÒ DEFINIRE MEDIANA COME QUEL VALORE CHE

DIVIDE L'INSIEME DEI DATI IN DUE GRUPPI UGUALMENTE

NUMEROSI. IN PRATICA SI PRENDE LA CLASSE k -ESIMA

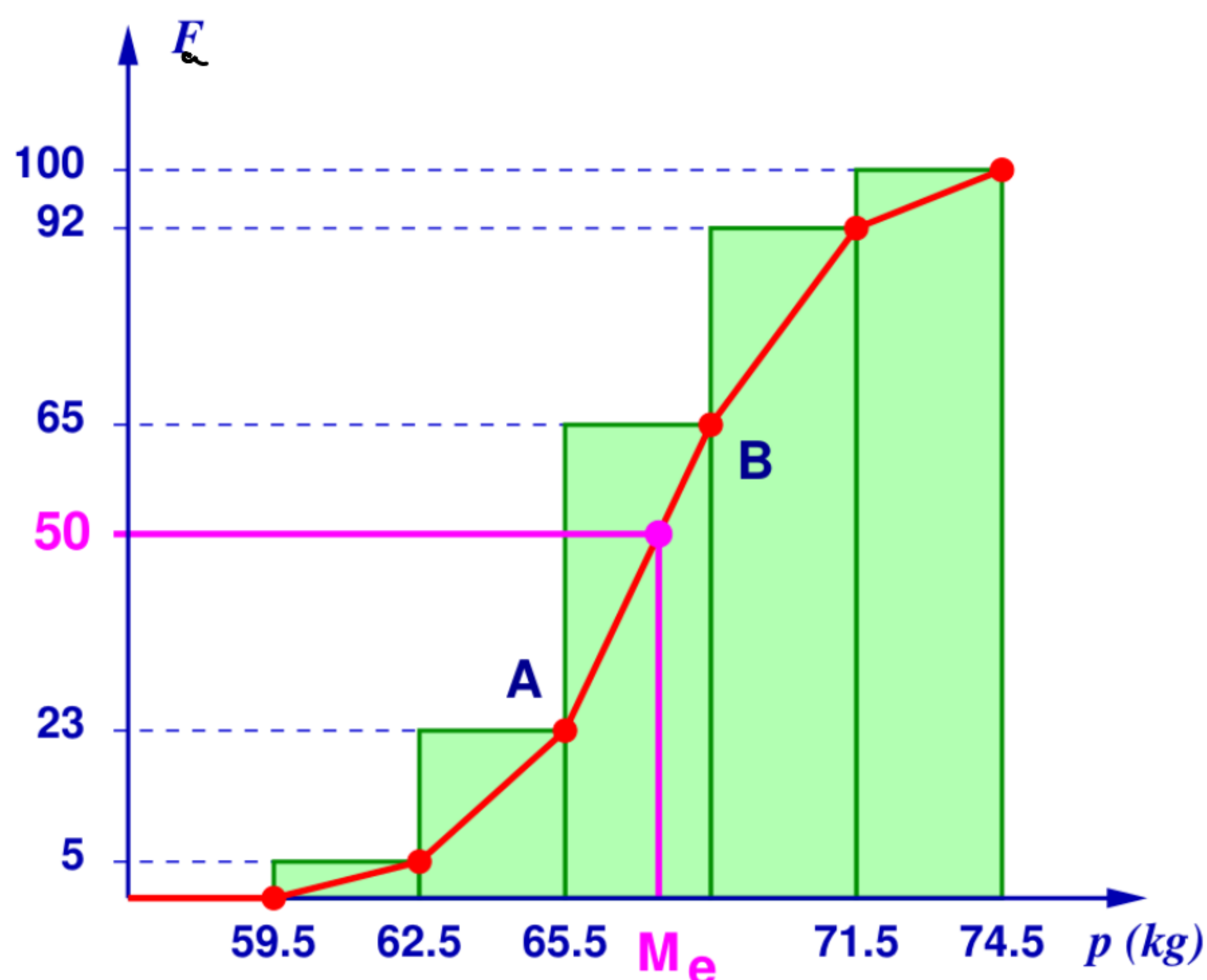
PER LA QUALE $F_a(k-1) < M_{\frac{1}{2}} \leq F_a(k)$. LA MEDIANA M_e È IL VALORE PER CUI

$$F_a(k-1) + \frac{(M_e - m_k)}{(M_k - m_k)} \underbrace{(F_a(k) - F_a(k-1))}_{f_a(k)} = \frac{M}{2}$$

DOVE M_k E m_k SONO, RISPETTIVAMENTE, I VALORI SUP E INF DELLA CLASSE k -ESIMA (QUINDI $M_k - m_k$ È LA LUNGHEZZA DELLA BASE DEI RETTANGOLI NELL'ISTOGRAMMA). GEOMETRICAMENTE, SIGNIFICA TROVARE IL VALORE M_e SULL'ASSE DELLE ASCISSE CHE SUL GRAFICO DELL'OGIVA SULL'ORDINATA RESTITUISCE $M/2$.

TORNANDO ALL'ESEMPIO CON IL PESO DEGLI STUDENTI,

$$M/2 = 50, \quad k=3, \quad m_k=65.5, \quad M_k - m_k = 3, \quad F_a(3) - F_a(2) = 42.$$



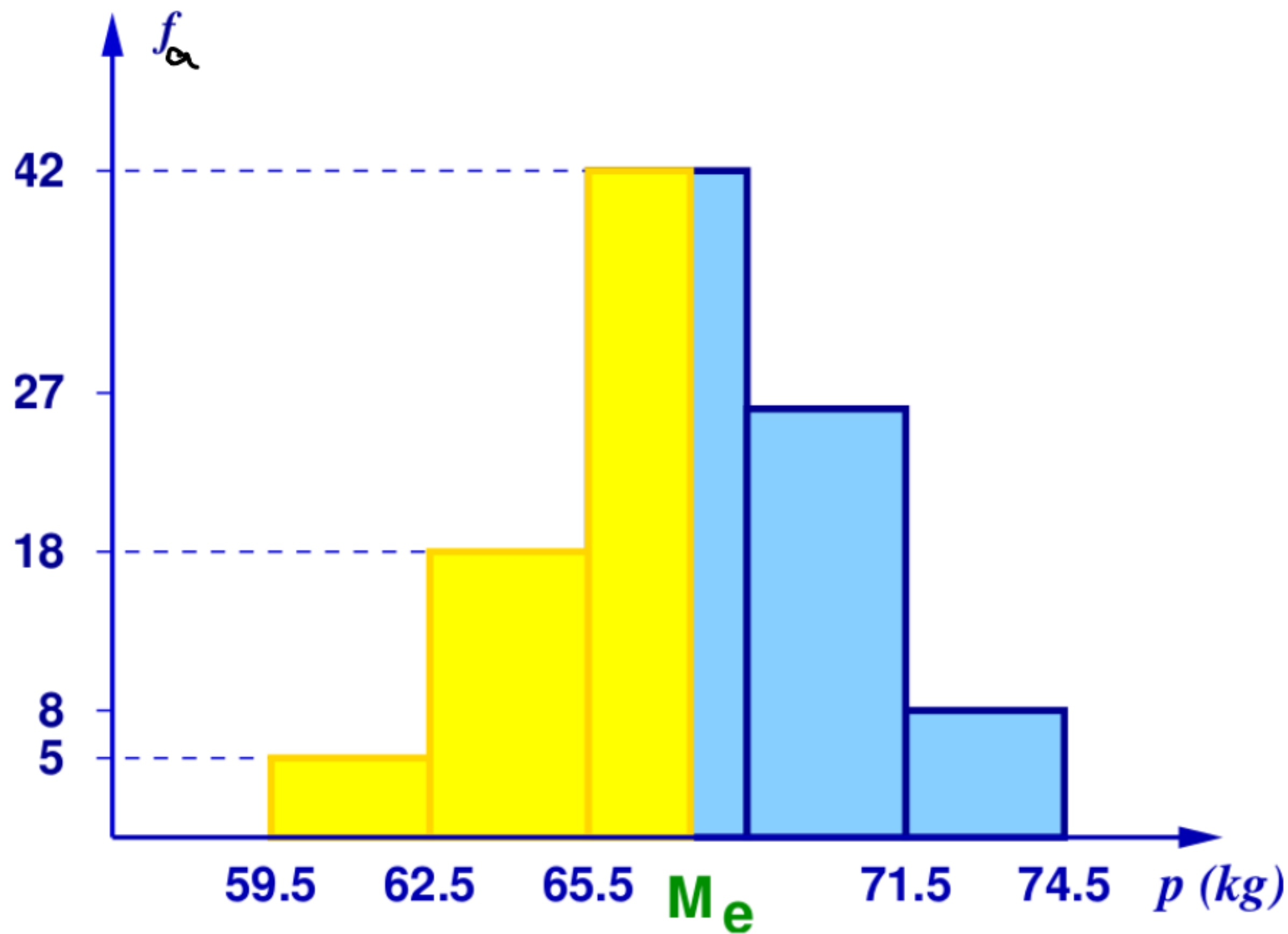
DALL'EQUAZIONE

$$\frac{100}{2} = \frac{M_e - 65,5}{3} \cdot 42 + 23$$

OTTENIAMO

$$M_e \approx 67,43 \text{ kg}$$

È EQUIVALENTEMENTE, DOBBIAMO TROVARE IL PUNTO M_e TALE CHE L'AREA IN GIALLO SIA IL 50% DELL'AREA TOTALE SOTTESA ALL'ISTOGRAMMA.



L'AREA TOTALE È

$$n \cdot 3 = 300$$

(3 È LA BASE DEI RETTANGOLI)

DALL'EQUAZIONE $3 \cdot 5 + 3 \cdot 18 + (M_e - 65,5) \cdot 42 = 150$
TROVIANO $M_e \approx 67,43 \text{ kg}$

VA NOTATO CHE LA MEDIANA PUÒ ESSERE USATA ANCHE QUANDO I DATI NON HANNO CARATTERE NUMERICO: È SUFFICIENTE CHE SIANO ORDINABILI. PER ESEMPIO, SUPPONGIAMO DI AVERE I SEGUENTI VOTI:
GRAVE, INSUFF., INSUFF., INSUFF., SUFF., DISCRETO, BUONO, OTTIMO.
LA MEDIANA È "SUFFICIENTE".