Statistica Descrittiva

Cellesi Massimo

Statistica

- La **statistica** è una scienza che studia fenomeni collettivi con metodi matematici basati prevalentemente su tecniche di campionamento e sul calcolo della probabilità

STATISTICA DESCRITTIVA

Raccogliere i dati di una popolazione o di un campione e sintetizzare i dati attraverso **tabelle**, **indici** e **grafici** che descrivono il fenomeno oggetto di studio

STATISTICA INFERENZIALE

Metodi per stabilire, sulla base dei dati rilevati su un campione, le caratteristiche di una popolazione con una possibilità di errore predeterminata

Statistica inferenziale

Popolazione

Campione

Inferenza Statistica

Parametri della popolazione

Statistica descrittiva

Terminologia

- **Popolazione**: L'insieme degli elementi oggetto dello studio, cioè l'insieme delle unità (*unità statistiche*) sulle quali vengono rilevate le modalità con le quali il fenomeno si manifesta
- **Unità statistica**: ogni elemento della popolazione.
- **Campione**: un qualsiasi insieme di unità statistiche prese dalla popolazione sulle quali si misura realmente la variabile oggetto di studio
- **Dati grezzi**: dati del campione / popolazione non organizzati in tabelle, non sintetizzati o elaborati

Dati grezzi (esempi)

- Altezza di un campione di individui

  1,76 1,56 1,89 1,92 1,72 1,98 1,54 1,63 1,66 1,58 2,01 1,88 1,77 1,75
  1,71 1,88 1,91 1,67 1,52 1,62 1,60 1,63 1,77 1,86 1,89 1,68 1,90 1,65

- Numero di suini presenti in aziende suinicole

  123 198 54 89 65 190 220 70 67 88 67 120 143 78 87 50 120 64

- In ciascun esempio si osserva una **variabile casuale**
  - Altezza di un individuo
  - N° di suini

Variabili

possono assumere tutti i valori di R
o di un suo sottoinsieme

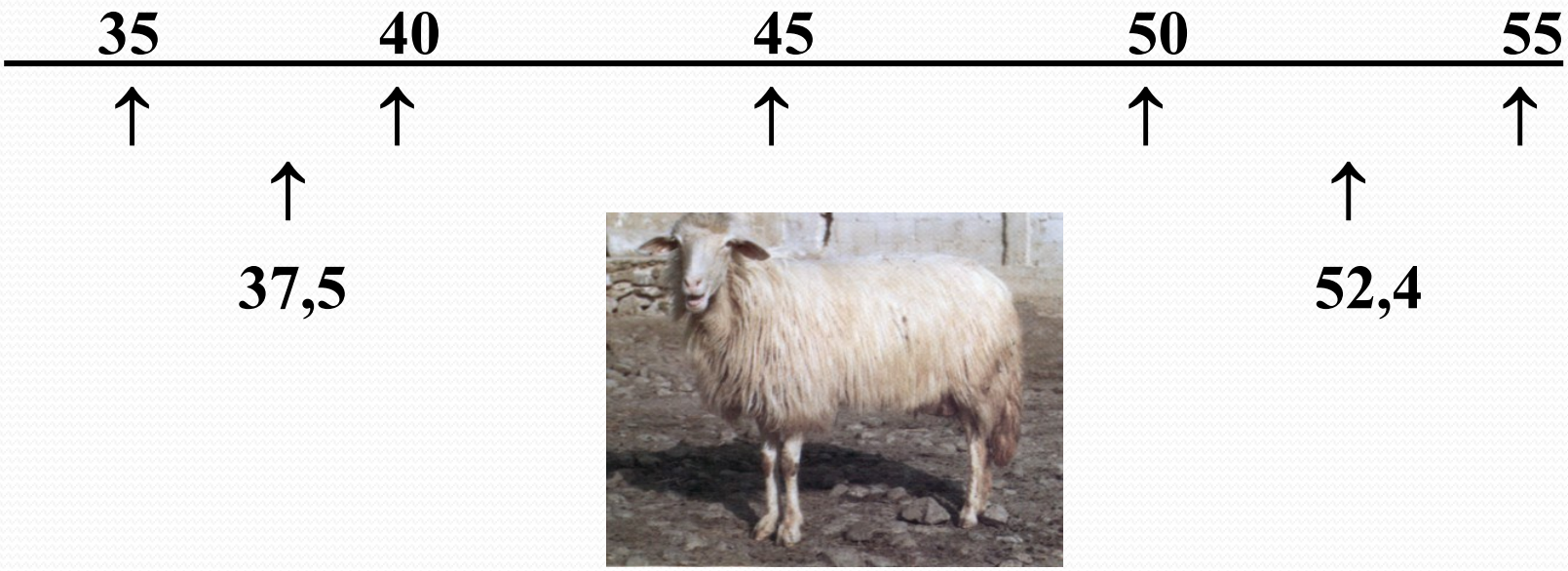
- variabili
  - quantitative (numeriche)
    - continue
    - discrete
  - qualitative (non numeriche)

l'insieme dei valori che la variabile può assumere (a priori) è finito

Maschio/Femmina; Alto/Medio/Basso; colore dei capelli

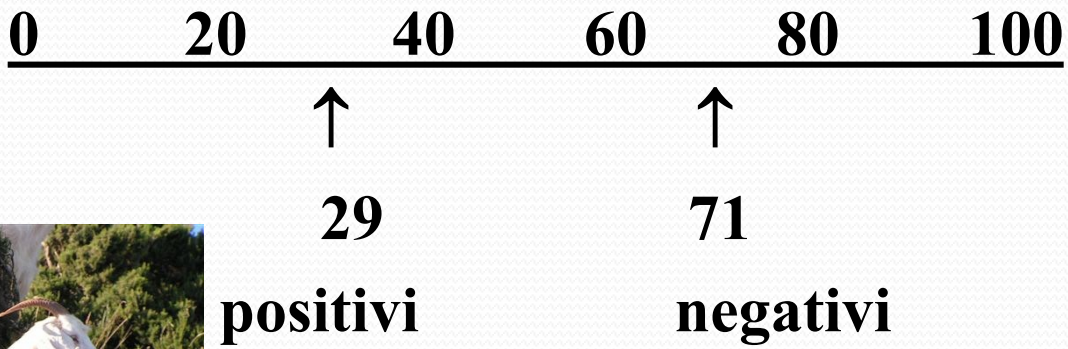
Variabili Continue

- Ex. Peso di una pecora



Variabili discrete

- Ad es. numero di capre di un allevamento (totale=100 animali) risultate positive ad una indagine sierologica



Raggruppamento dei dati

- Per descrivere i dati tramite una tabella (**tabella di distribuzione di frequenza**)dobbiamo:
  - dividere i dati in **classi**
  - determinare il numero di individui di ciascuna classe
    - **Frequenza assoluta** (o **frequenza della classe**).
    - **Frequenza percentuale**
    - **Frequenza cumulata**
- I dati ordinati e riassunti nella tabella di distribuzione di frequenza sono detti **dati raggruppati**.

Variabili Discrete

voto esame Matematica	Freq Ass	freq Rel	Freq Perc	Freq Cum
18	12	0,047	4,7%	12
19	10	0,039	3,9%	22
20	18	0,071	7,1%	40
21	20	0,079	7,9%	60
22	19	0,075	7,5%	79
23	22	0,087	8,7%	101
24	30	0,118	11,8%	131
25	26	0,102	10,2%	157
26	24	0,094	9,4%	181
27	23	0,091	9,1%	204
28	20	0,079	7,9%	224
29	17	0,067	6,7%	241
30	13	0,051	5,1%	254
totale	254	1	100%	

$$\text{freq rel} = \frac{\text{freq Ass}}{TOT}$$

$$\text{Freq Cum} = \sum \text{freq Ass prec}$$

$$\text{freq perc} = \text{freq rel} * 100$$

Variabili Continue

- **Esempio: accrescimento in peso giornaliero di pulcini**

3,7	4,4	4,3	4,4	4,9
4,2	4,2	4,6	4,3	4,8
4,7	4,2	4,5	4,1	3,9
4	4	4,4	4	3,8
4,9	4,3	3,9	4,7	4
4,3	4,1	4,6	4,6	4,9
4	4,5	3,8	4,3	4,5
4,2	3,8	4	4,1	4,7
4,7	4,8	4,3	4,2	4,4
4,1	4,3	4,4	4,6	4,6
4,2	4,4	4,2	4,8	4,4
3,8	4,4	3,9	4,5	3,9
4,2	4,8	3,6	4,3	4,2
4,2	4,5	4,1	4	4,6
3,8	4,3	3,8	3,9	4,2
4,7	4	4,4	4,4	4,4
3,9	4,3	4,1	4,2	4,4
4,7	4,5	4,2	4	4,1
4,1	4,1	4,7	4,1	4,8
4,9	4,4	4,3	4,5	4

- Quante e quali classi considerare?
- Campo di Variazione (CV)=Max-min
- N=numerosità del campione
- Numero classi (K): $k = \sqrt{N}$

$$(5 \leq k \leq 20)$$

- Ampiezza classi (A): $A = \frac{CV}{K}$

max	4,9	k=	10
min	3,6	CV=	1,3
N	100	A=	0,13

Variabili continue

min

Inf+A

inf	sup	freq Ass	freq Cum	freq rel	freq %	Freq cum %
3,6	3,73	2	2	0,02	2%	2%
3,73	3,86	6	8	0,06	6%	8%
3,86	3,99	6	14	0,06	6%	14%
3,99	4,12	20	34	0,2	20%	34%
4,12	4,25	13	47	0,13	13%	47%
4,25	4,38	11	58	0,11	11%	58%
4,38	4,51	20	78	0,2	20%	78%
4,51	4,64	6	84	0,06	6%	84%
4,64	4,77	7	91	0,07	7%	91%
4,77	4,9	9	100	0,09	9%	100%
Totale		100		1	100%	

Variabili qualitative

Casi	pecora1	pecora2	pecora3	pecora4	…..	pecora19	pecora20
Colore mantello	B	B	N	B		G	N

	freq. Ass	freq. Rel	Freq. %
Bianco	9	0,45	45 %
Nero	3	0,15	15 %
Grigio	8	0,40	40 %
Totale	20	1	100 %

Rappresentazione grafica

- Rappresentazione grafica evidenzia, in un modo semplice ma efficace, come si distribuiscono i dati
  - Può far notare irregolarità (errori di misurazione)

Variabili **discrete/qualit.**

- Diagramma circolare
- Diagramma a barre (x dati raggruppati, y freq Ass)

Variabili continue

- Istogrammi
- Grafici di Frequenze

Variabili qualitative/Discrete

Diagramma circolare/Aerogramma

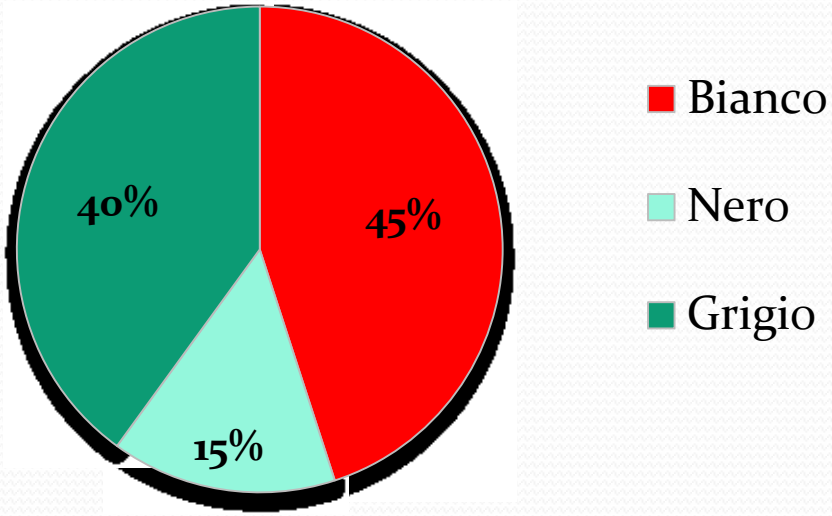
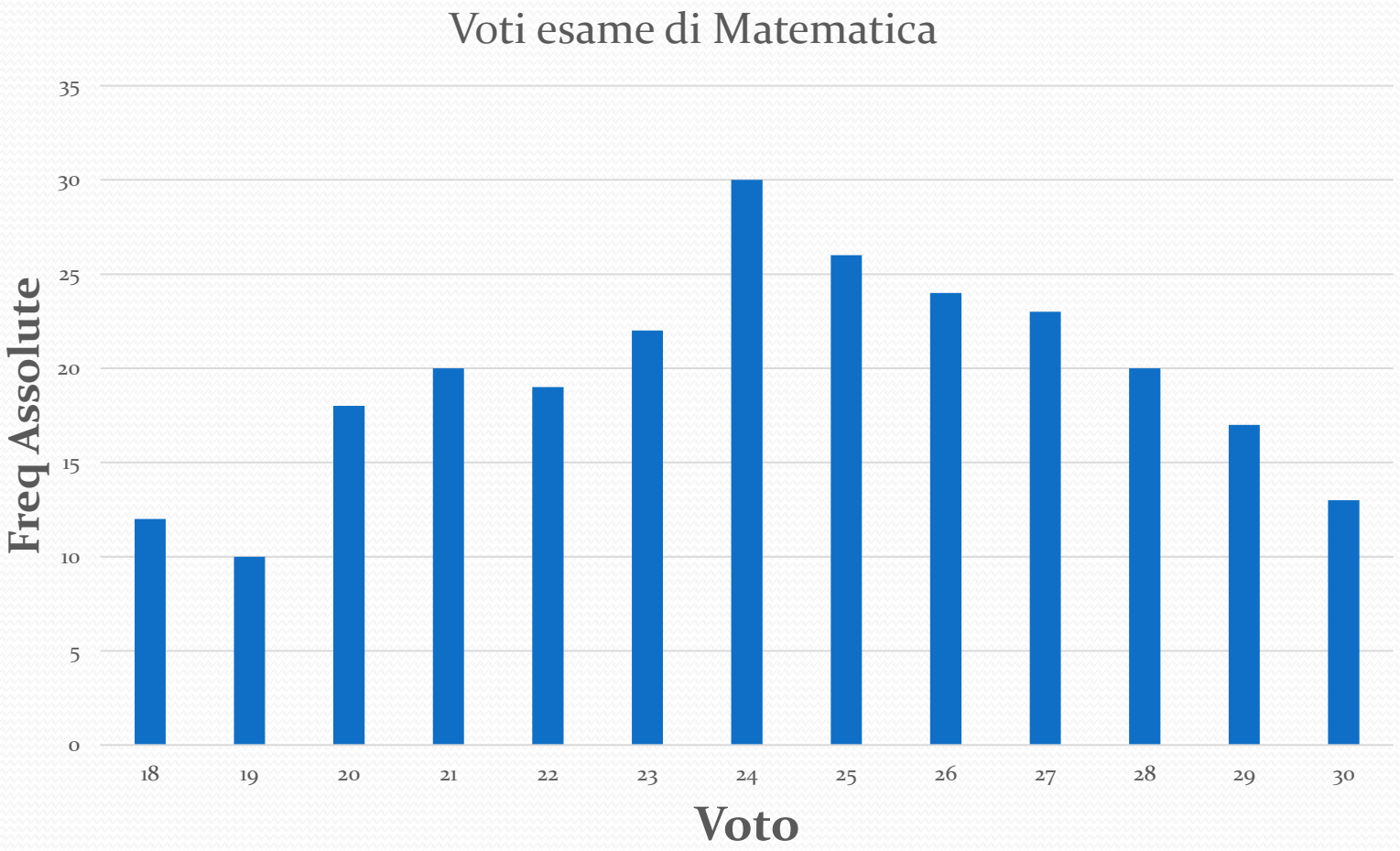
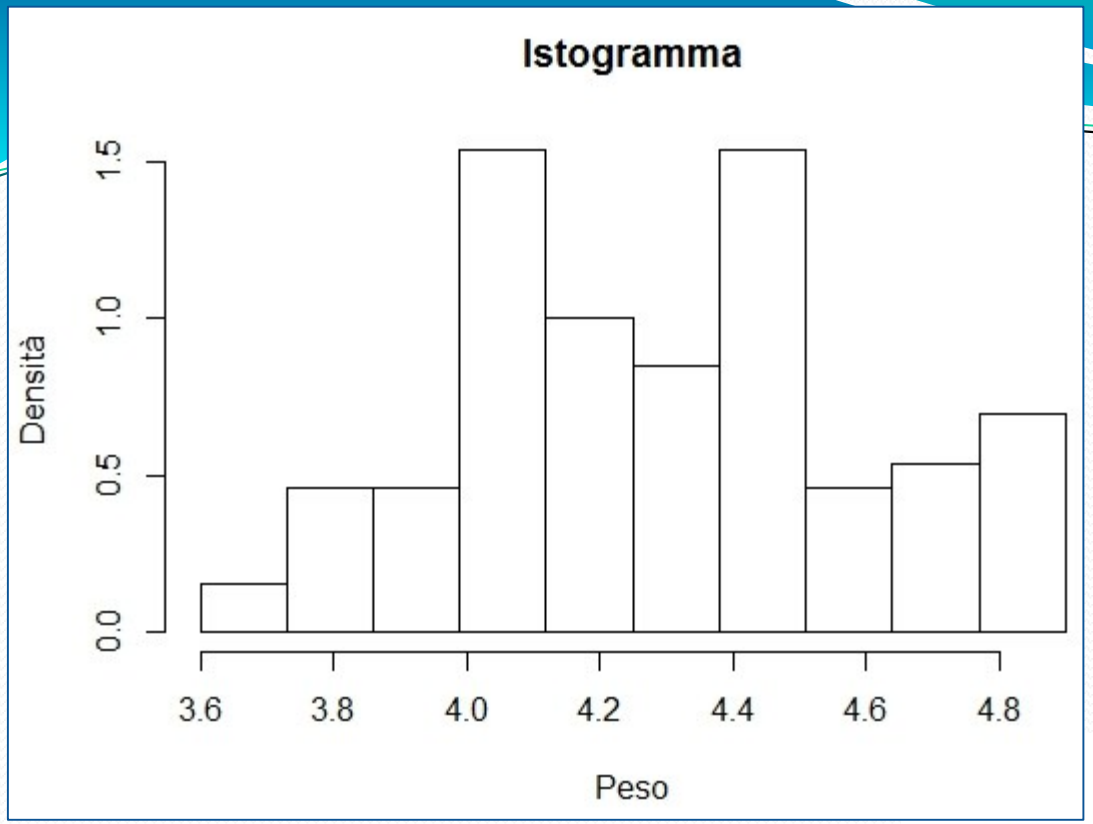


Diagramma a barre (qual / discrete)

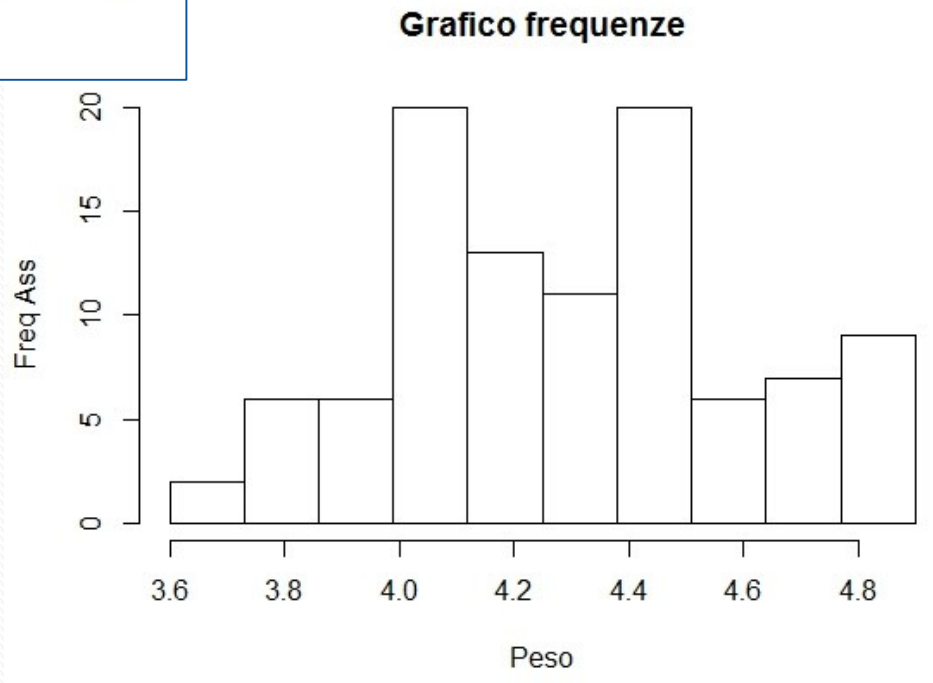


Istogrammi (variabili continue)

- Rettangoli adiacenti le cui basi sono allineate su un asse orientato dotato di unità di misura.
  - Adiacenza dei rettangoli -> continuità del carattere.
- I rettangoli hanno:
  - base di lunghezza pari all'ampiezza della classe;
  - l'altezza è calcolata come densità di frequenza, ovvero essa è pari al rapporto fra la frequenza (relativa) della classe e la sua ampiezza.
- L'area della superficie di ogni rettangolo coincide con la frequenza relativa della classe.



Istogramma:
Area rettangolo = freq Rel



Indici di posizione e dispersione

- Indici (numeri) detti **statistiche**, utili per descrivere dei dati numerici e la loro distribuzione di frequenza.
- Indici di posizione centrale
  - Media (aritmetica)
  - Moda
  - Mediana
- Indici di dispersione
  - Coefficiente di variazione
  - Varianza
  - Scarto quadratico medio (deviazione standard)

Osservazione

- Le statistiche riferite al campione sono dette Statistiche e sono indicate con l'alfabeto latino

$$\text{media} = \bar{x} \qquad \text{dev standard} = s_x$$

- Le statistiche riferite alla popolazione sono dette parametri e sono indicate con l'alfabeto greco

$$\text{media} = \mu \qquad \text{dev standard} = \sigma$$

Indici di posizione centrale: Media

- Un modo informativo di descrivere la collocazione di un insieme di dati è quello di riportarlo ad un valore centrale
- Media aritmetica (media campionaria) per dati disaggregati

$$x_1, x_2, x_3, \ldots\ldots\ldots, x_n$$

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

Media: dati raggruppati in classi

Inf	Sup	m_i (valore centrale)	Freq Ass
1	5	3	10
5	9	7	12
9	13	11	9
…	…	…	

$$\bar{x} = \frac{\sum_{i=1}^{n} m_i f_{ass}}{N} = \frac{3*10 + 7*12 + \ldots\ldots}{N}$$

Mediana

- La mediana M di un insieme di n dati disaggregati, **ordinati** in ordine crescente è il **valore centrale dei dati**, se il numero di dati è dispari, o **la media aritmetica dei due valori centrali**, se il numero dei dati è pari.

Mediana:
valore che divide la distribuzione dei dati in 2 parti uguali

1 5 **7** 12 15

3 8 9 **11 12** 15 16 18

$$\frac{11 + 12}{2} = 11,5$$

Moda

- La **moda** $\tilde{x}$ di un insieme di *n* dati è il valore corrispondente alla massima frequenza assoluta.
- La moda è per lo più utilizzata quando si trattano dati di tipo **qualitativo**, per i quali non è possibile calcolare media e mediana.
- La moda può non esistere A B C D E F
- Una distribuzione può avere più di un valore modale (bimodali = 2 mode)

1 2 2 2 3 3 4 4 4 5 6 9

Indici di posizione centrale

- La media è la misura di tendenza centrale più comunemente utilizzata.
- La media, tuttavia, è influenzata dai valori estremi, mentre la mediana e la moda non lo sono.
- Esempio: Altezze di 7 persone (in cm):

168, 178, 126, 181, 156, 161, 170 -> $\overline{X} = 163$ $M = 168$

Se l'altezza di 126 cm è da imputarsi ad un errore di battitura e l'altezza corretta è 162:

$\overline{X} = 168$ $M = 168$

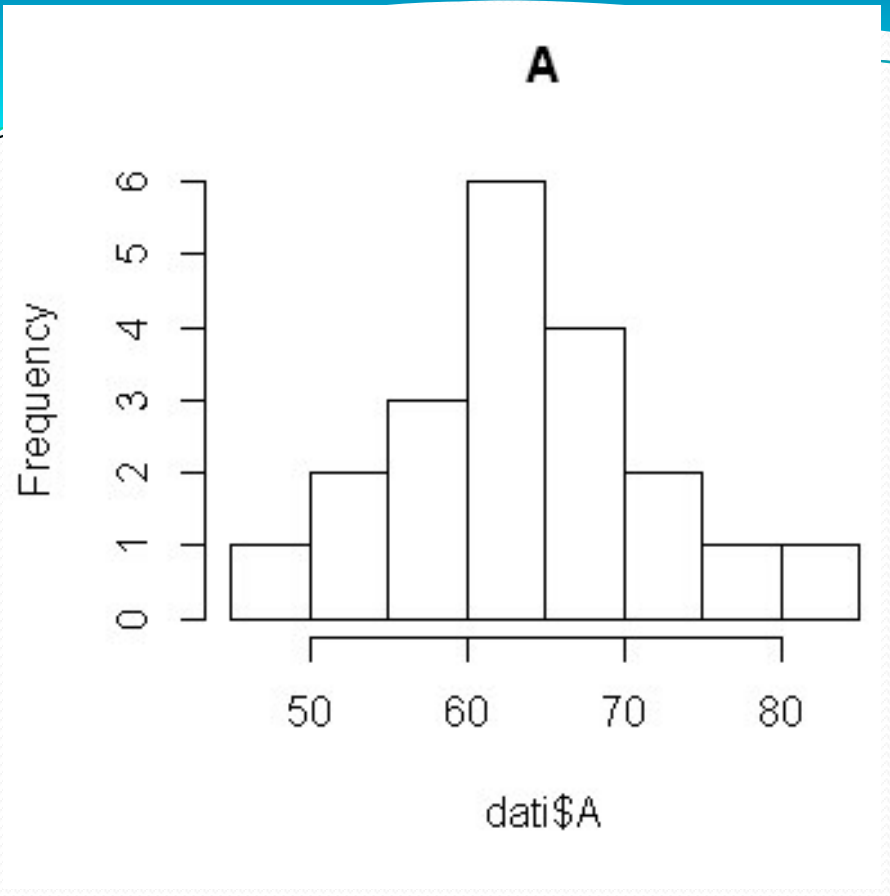
Indici di dispersione

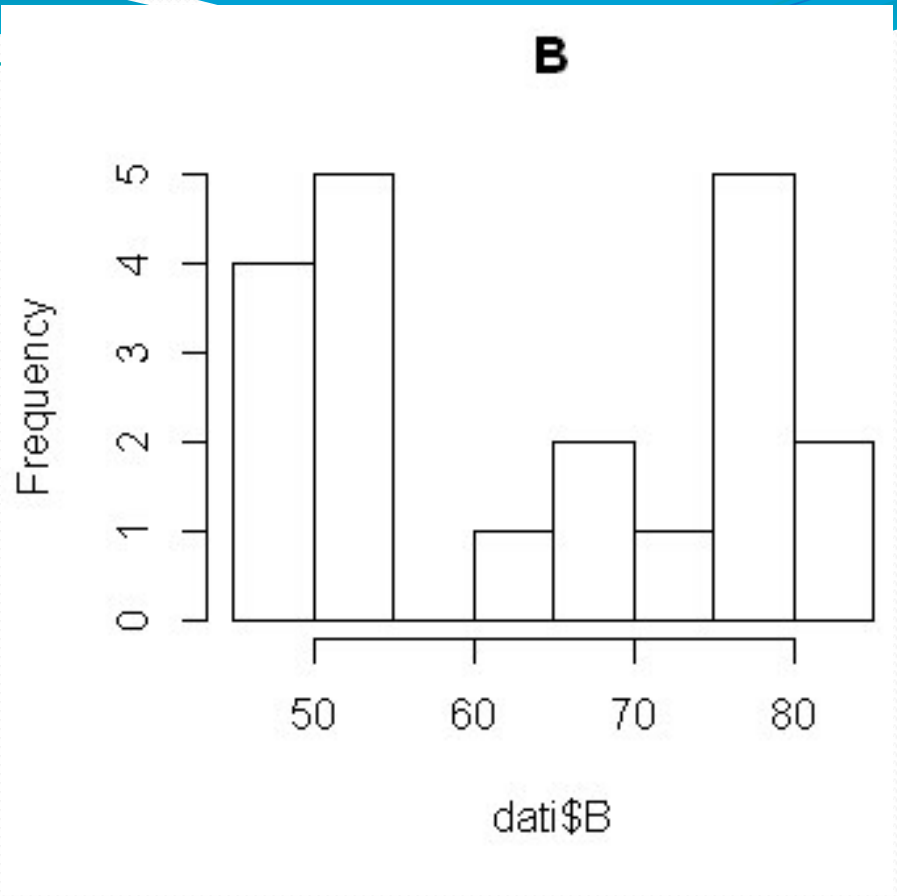
- Gli indici di posizione centrale non tengono conto della variabilità esistente fra i dati; vi sono distribuzioni che, pur avendo la stessa media, sono molto diverse fra loro.

A	60,1	52,6	64,6	68,8	67,7	59,5	74,9	64,2	60,2	54,3
	61,3	47,4	78,4	67,3	84,4	74,9	63,6	58,1	59,4	69,5

$$\overline{x}_A = \overline{x}_B = 64,56$$

B	77,3	67,8	79,1	64,4	78,5	53	83,6	69,2	54,1	82,4
	54,7	76,2	73	45,6	54,6	49,8	48,7	77,9	46,5	54,9





- Nonostante le medie siano uguali, i due insiemi di dati sono STRUTTURALMENTE DIVERSI

- I dati dei due insieme presentano una **DIVERSA DISTRIBUZIONE**

Varianza

- Si definisce **varianza** del campione/popolazione x_1, x_2, x_3, ……., x_N la quantità:

Campione

$$s^2 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})^2}{N - 1}$$

Popolazione

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

dove

$\overline{x}$ rappresenta la media <u>campionaria</u>

μ rappresenta la media della <u>popolazione</u>

Deviazione standard

- Si definisce **scarto quadratico medio** o **deviazione standard** la radice quadrata della varianza

$$s = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \overline{x} \right)^2}{N - 1}} \qquad \sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \mu \right)^2}{N}}$$

Varianza e scarto quadratico medio sono detti **indici di dispersione** o **indici di variabilità**, perché misurano la dispersione dei dati attorno alla media

Dati raggruppati

- Varianza

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{k}\left(m_i - \bar{x}\right)^2 f_i \qquad \sigma^2 = \frac{1}{N}\sum_{i=1}^{k}\left(m_i - \mu\right)^2 f_i$$

k numero di classi

m_i valore centrale della classe i-ma

f_i frequenza assoluta della classe i-ma

Coefficiente di variazione

- Il coefficiente di variazione (CV) è una misura di dispersione relativa e permette di confrontare la dispersione dei dati di fenomeni diversi tra loro.

$$CV = \frac{s}{\bar{x}} \qquad CV = \frac{\sigma}{\mu}$$

- CV è un numero adimensionale

Correlazione tra variabili

- Spesso, in una indagine statistica, si osservano più variabili per ogni unità statistica
- Problema tipico consiste nel chiedersi se esiste una **correlazione** tra le variabili osservate
  - Correlazione tra variabili: relazione che permette di determinare il probabile valore di una variabile noto il valore dell'altra.

- Una prima indagine qualitativa può essere svolta rappresentando le variabili in un piano cartesiano (grafico a dispersione o scatterplot)

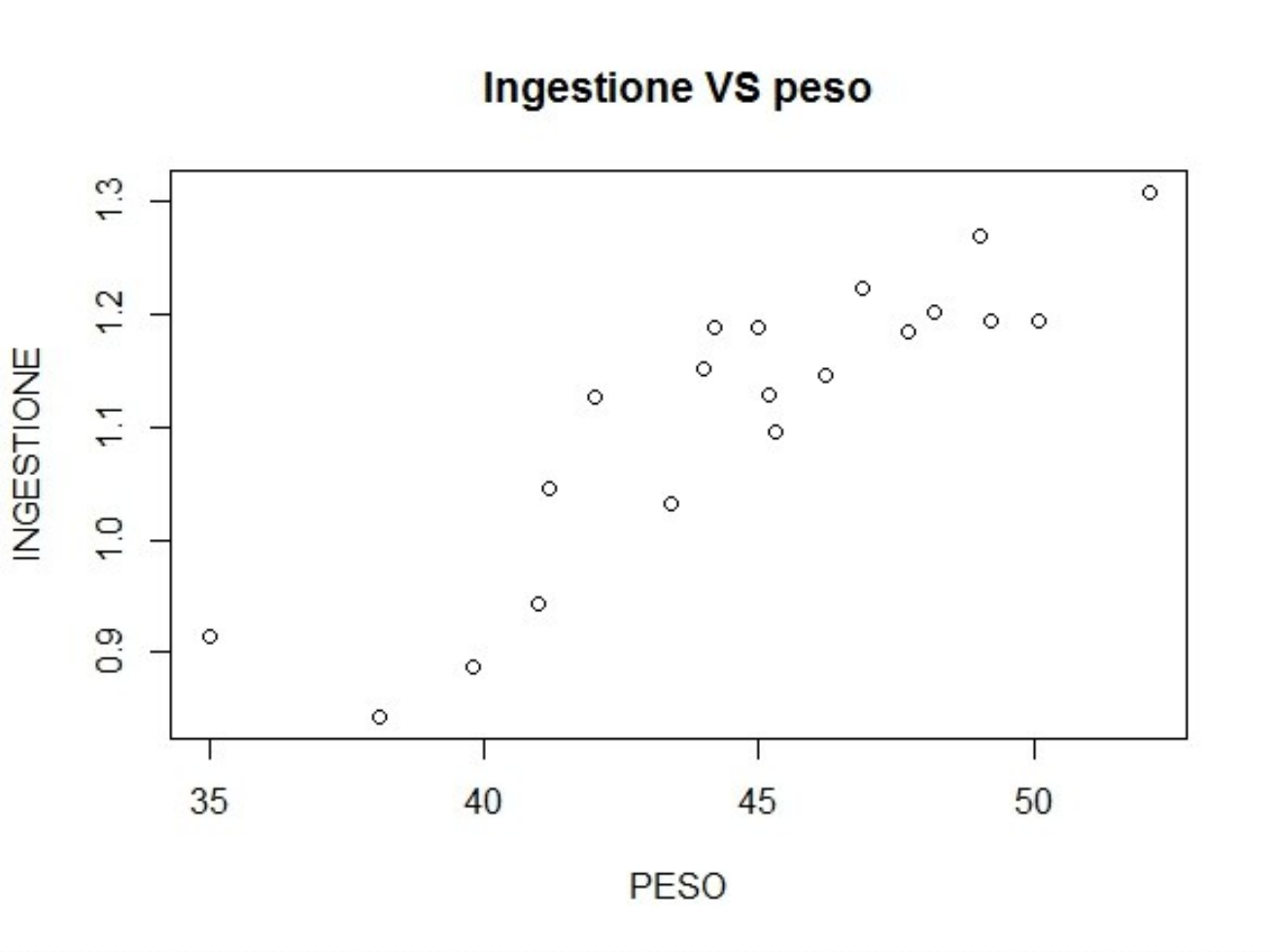
Dati multipli

ID	PESO	INGESTIONE
1	45	1,189
2	44	1,151
3	35	0,915
4	41	0,944
5	42	1,127
6	43,4	1,031
7	46,9	1,223
8	45,2	1,129
9	39,8	0,888
10	50,1	1,193

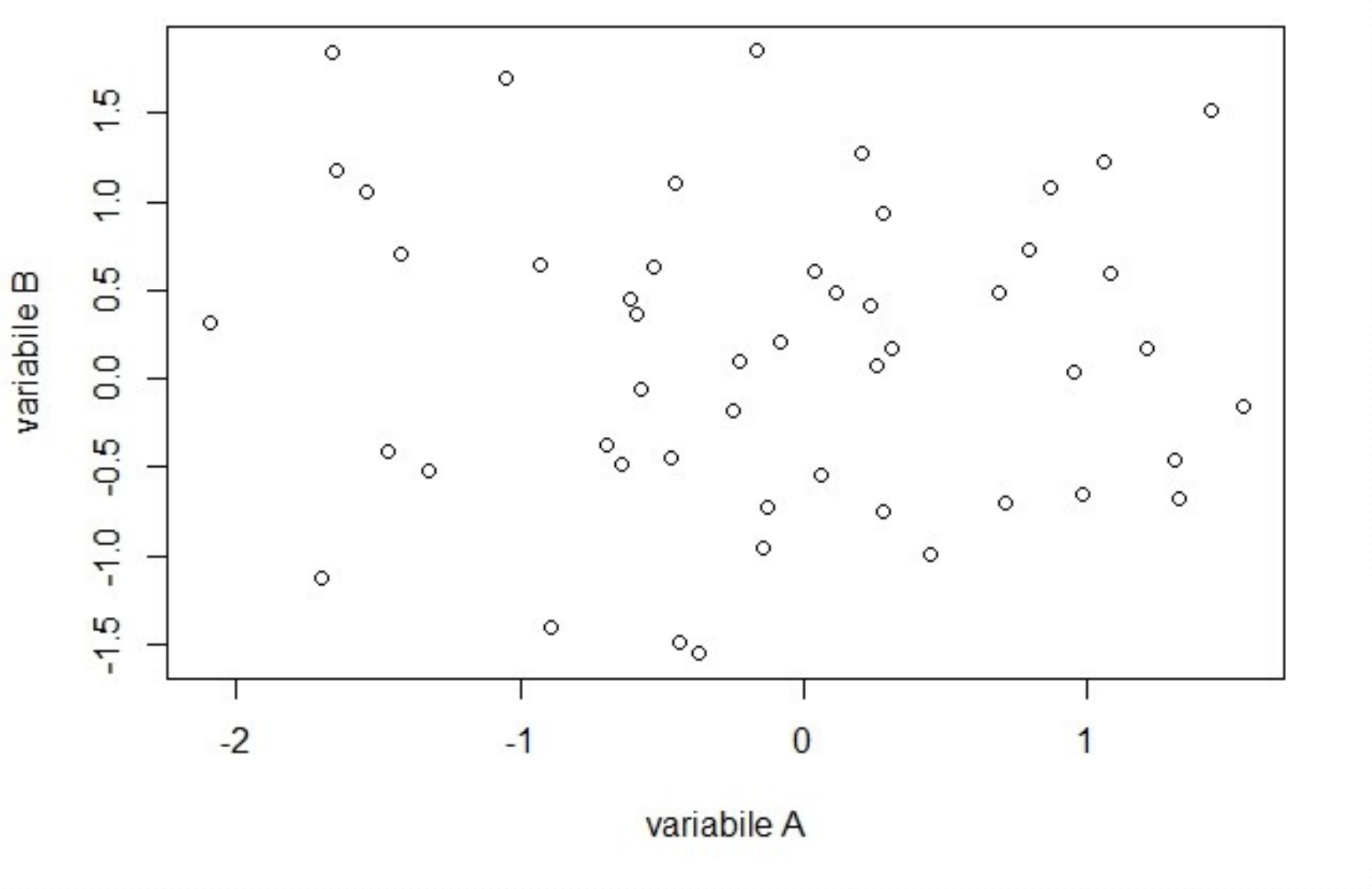
ID	PESO	INGESTIONE
11	49,2	1,194
12	45,3	1,095
13	46,2	1,146
14	44,2	1,189
15	41,2	1,045
16	52,1	1,308
17	47,7	1,185
18	48,2	1,202
19	38,1	0,843
20	49	1,27

Peso e ingestione media giornaliera di pecore di razza sarda

Scatterplot dei dati



Scatterplot fra variabile che non presentano correlazione



Covarianza

- N osservazioni congiunte

$$(x_1, y_1), (x_2, y_2), \ldots\ldots, (x_N, y_N)$$

$$S_{x,y} = \frac{1}{N-1}\left[\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})\right]$$

La covarianza è un numero che fornisce una misura di quanto le due variabili variano assieme, ovvero della loro dipendenza.

$S_{x,y}$>o correlazione positiva
$S_{x,y}$=o nessuna correlazione
$S_{x,y}$>o correlazione negativa

Coefficiente di correlazione

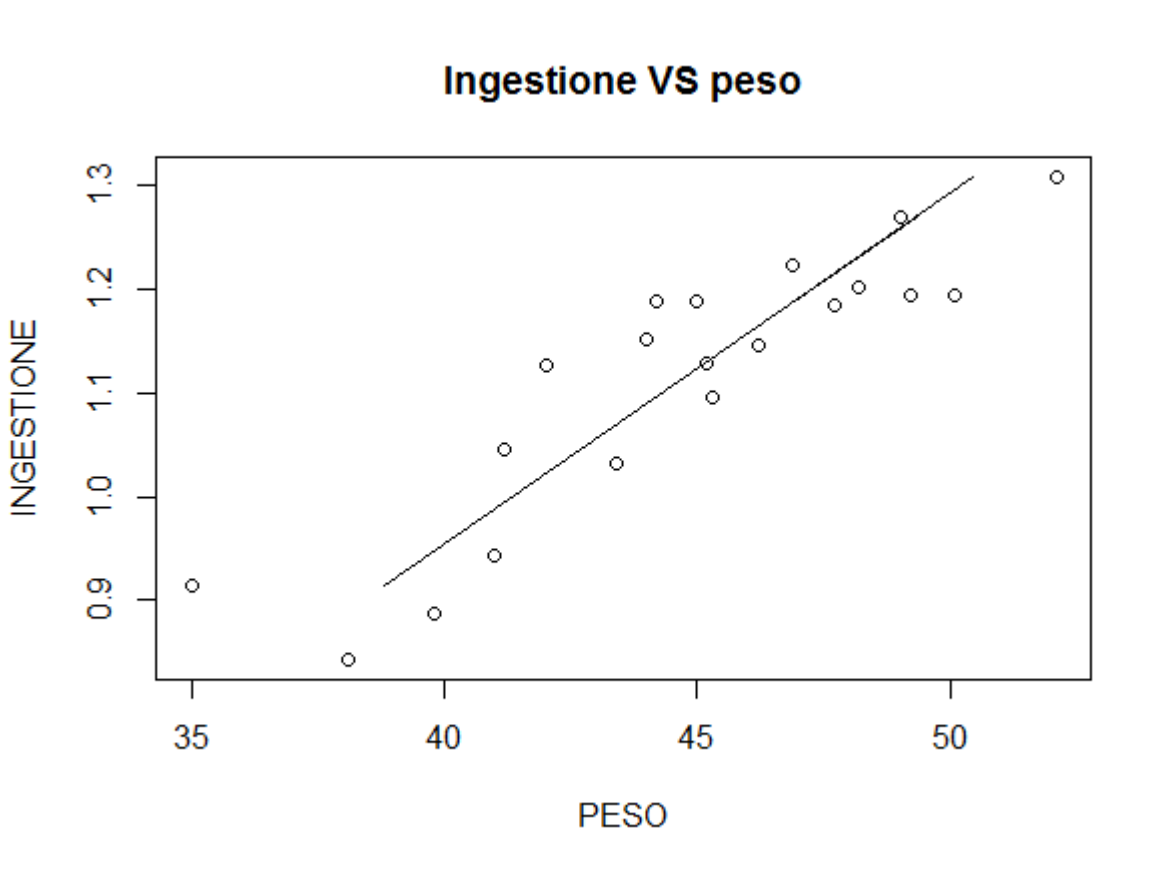
$$r_{x,y} = \frac{S_{x,y}}{S_x S_y}$$

Numero adimensionale che assume valori tra: $-1 \leq r_{x,y} \leq 1$

S_x S_y deviazione standard, $S_{x,y}$ covarianza

$\lvert r_{x,y} \rvert < 0,3$	correlazione debole
$0,3 < \lvert r_{x,y} \rvert < 0,7$	correlazione modereta
$\lvert r_{x,y} \rvert > 0,7$	correlazione forte

$S_{x,y}$ e $r_{x,y}$



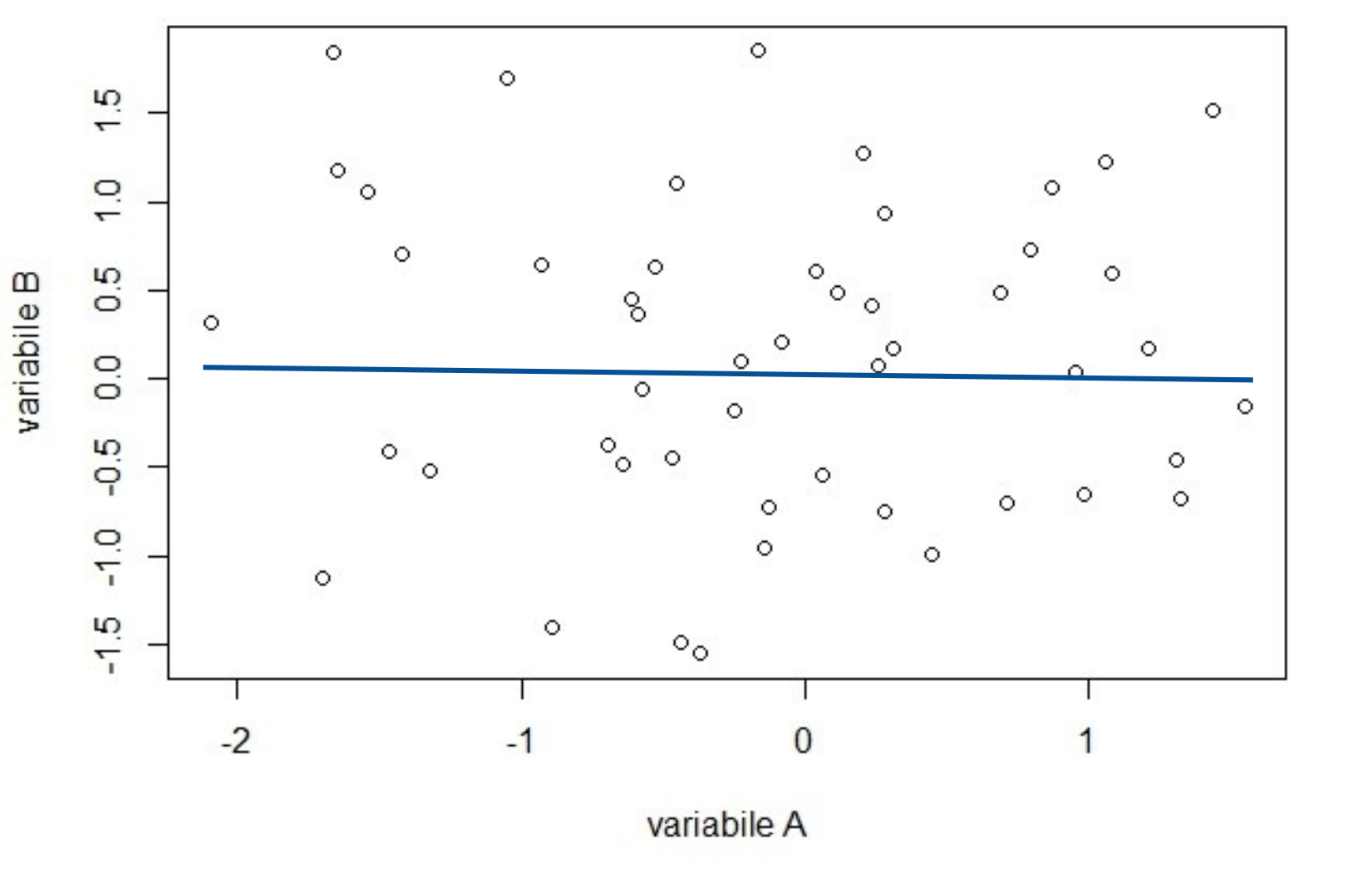
$$S_{x,y} = 0.496$$

$$S_x = 4.286$$

$$S_y = 0.129$$

$$r_{x,y} = 0.894$$

$$r_{x,y} = -0.032$$



Modello lineare

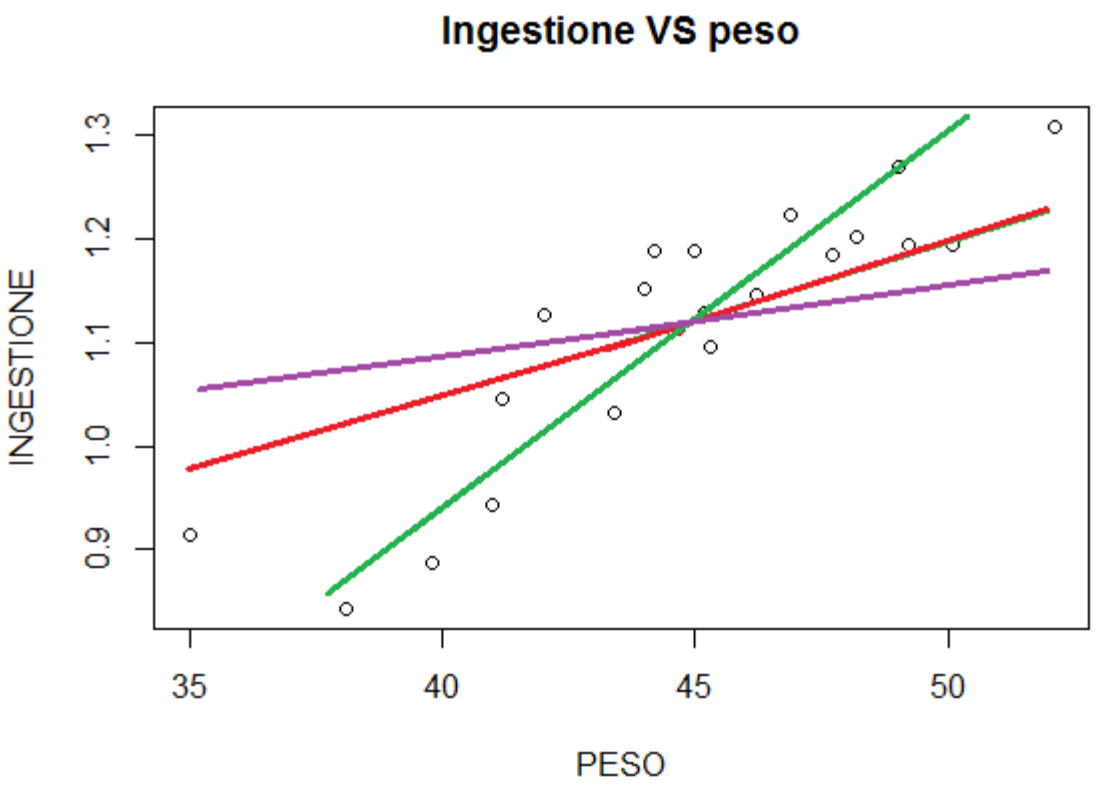
- Se due variabili risultano fortemente correlate, possiamo stimare il valore di una di esse (ingestione) dalla conoscenza dell'altra (peso)
- Il modello più semplice per effettuare tale stima è il modello lineare

$$y = A\,x + B$$

y=Ingestione *x=Peso*

A=coeff. angolare *B=intercetta*

Determinazione dei coefficienti



Come determinare i coefficienti della retta di regressione?

Quale retta devo considerare?

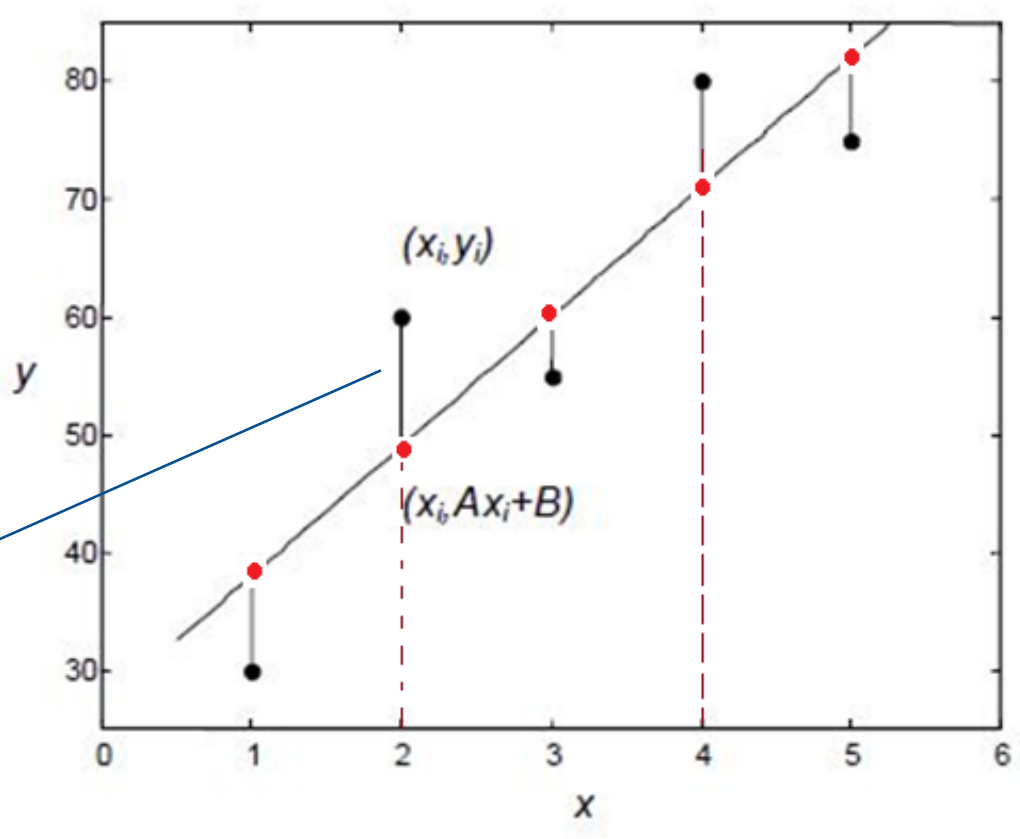
Errore

Valori stimati dal modello

Valori reali

Errore:
Valore stimato – valore reale

$$E_i = A x_i + B - y_i$$



Retta minimi quadrati

- Retta dei **minimi quadrati** (ordinary least squares)

$$y = A x + B$$

- È la retta per la quale è minima la somma dei quadrati degli errori

$$E = \sum_{i=1}^{N} \left(A x_i + B - y_i \right)^2$$

Retta minimi quadrati

$$y = Ax + B$$

I valori del coefficiente angolare A e dell'intercetta B si ottengono:

$$A = \frac{S_{x,y}}{S_x^2} \qquad B = \overline{y} - A\overline{x}$$