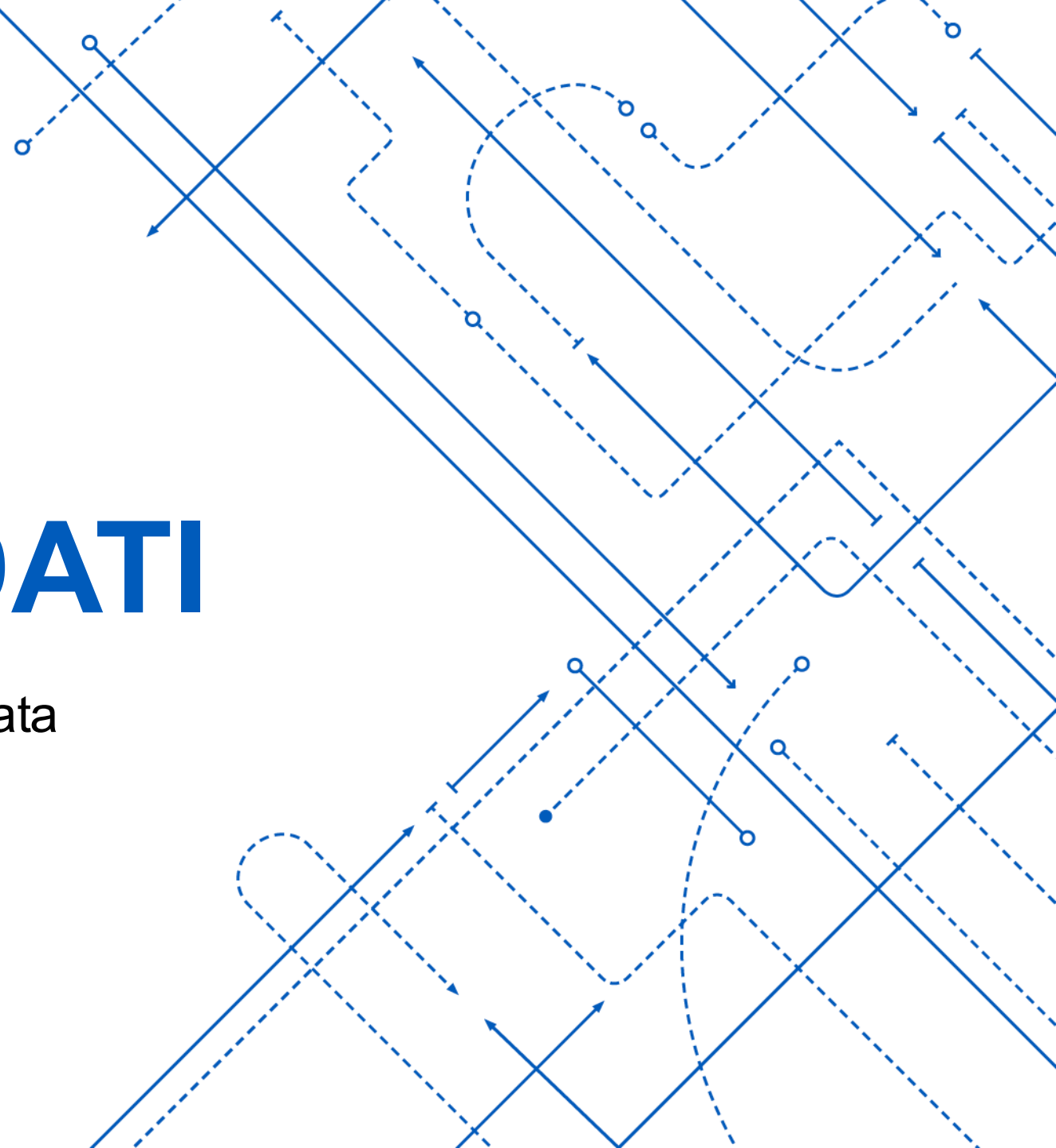
STATISTICA E ANALISI DEI DATI

Capitolo 5 - Statistica descrittiva bivariata

Dott. Stefano Cirillo
Dott. Luigi Di Biasi

a.a. 2025-2026



Cosa si intende con statistica descrittiva bivariata?

Con il termine “Bivariata” indichiamo il ramo della statistica che si occupa dei **metodi grafici** e **statistici** per descrivere eventuali **relazioni tra coppie di variabili.**

Siano X e Y **due** variabili di tipo quantitativo.

(X,Y) è la coppia di variabili per le quali intendiamo studiare eventuali relazioni.

Data una **coppia (X,Y) di variabili**, possiamo definire un campione C di n **osservazioni** del valori assunti dalla coppia (X,Y).

$C = (x_1, y_1,), (x_2, y_2), \ldots, (x_n, y_n)$ con $n = |C|$

Relazioni tra variabili

Non è detto che la relazione esista.

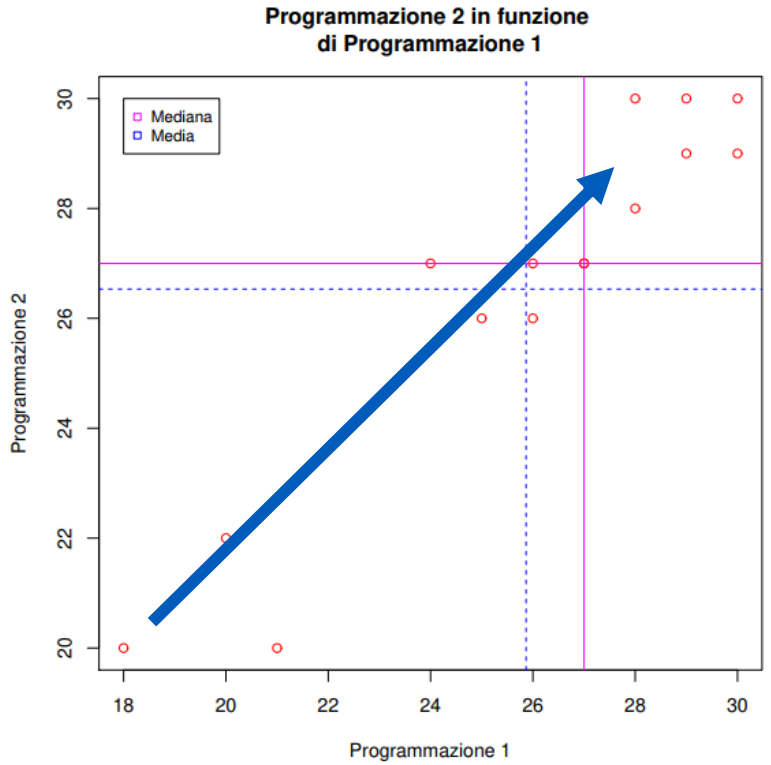
Quando parliamo di ***"relazione"*** tra due variabili indichiamo una qualche ***"forma di regolarità"*** tra i valori assunti da X e da Y.

Nello specifico:

- Definiamo X variabile indipendente;
- Definiamo Y variabile dipendente;
- Cerchiamo di capire se i valori assunti da Y **sono in qualche modo "guidati"** (correlati) dai valori assunti da X.

ESEMPIO: Quando il valore assunto da X cresce, **"sembra che"** anche il valore di "Y" cresca.

La statistica descrittiva bivariata fornisce gli strumenti per determinare il **«sembra che» è vero oppure no**.

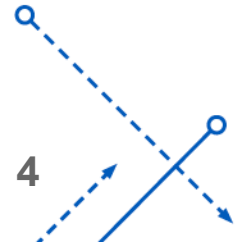


Scatterplot

Le **relazioni** tra variabili quantitative (X,Y) possono essere rappresentate graficamente con i **diagrammi di dispersione**.

Nello specifico:

- Ogni coppia di osservazioni è rappresentata da un simbolo;
- La variabile **indipendente** viene posta sulle **ascisse**;
- La variabile **dipedente** sulle **ordinate**;



Un primo esempio /1

Consideriamo, ad esempio, i voti conseguiti da 15 studenti negli esami di Programmazione 1 e di Programmazione 2.

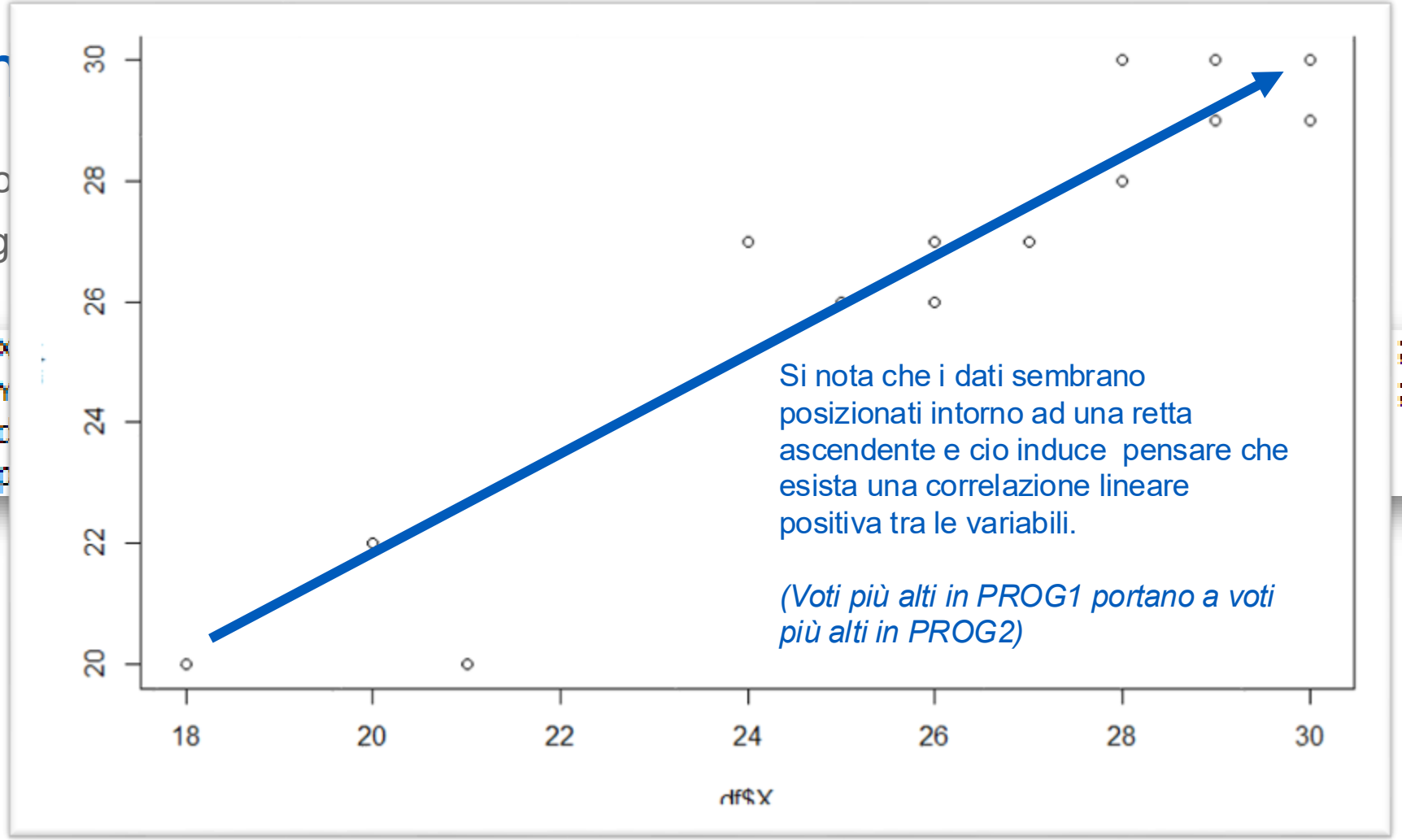
- (X,Y) è la nostra coppia di variabili;
- **X è la variabile indipendente** che assume i valori relativi ai voti raggiunti dagli studenti per PROG1; Y è la **variabile dipendente** che assume valori relativi ai voti raggiunti dagli studenti per PROG2.
  - $\mathbf{X_i}$ indica il voto in PROG1 dell'i-esimo studente;
  - $\mathbf{Y_i}$ indica il voto in PROG2 dell'i-esimo studente;
  - $C = ((X_1,Y_1),(X_2,Y_2), \ldots, (X_{15},Y_{15}))$ è il nostro campione.

Ci piacerebbe «scoprire» **se** per un dato studente S_i, il voto raggiunto all'esame di Programmazione 2 (Y_i) è in qualche modo in relazione con il voto raggiunto all'esame di Programmazione 1 (X_i).

Statistica e analisi dei dati

Un prin

Consideriamo
esami di Prog



Un primo esempio…/3

Oltre ad usare lo scatterplot, possiamo calcolare anche gli indici di posizione e dispersione.

```
x <- c(24 ,26 ,30 ,25 ,29 ,27 ,20 ,29 ,27 ,28 ,18 ,21 ,26 ,30 ,28);
Y <- c(27 ,26 ,29 ,26 ,30 ,27 ,22 ,29 ,27 ,28 ,20 ,20 ,27 ,30 ,30);
df <-data.frame (x,Y);
#plot(df$x,df$Y);
median(df$x)
median(df$x)
sd(df$x)
median(df$Y)
median(df$Y)
sd(df$Y)
```

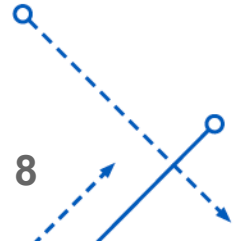
	progr1	progr2
Mediana campionaria	27	27
Media campionaria	25.86667	26.53333
Deviazione standard	3.681356	3.356586

Relazioni e “Correlazioni”

Nell’esempio precedente abbiamo usato uno **scatterplot per indagare graficamente** l’eventuale presenza di una relazione tra due variabili quantitative (X,Y).

Tuttavia, è possibile ottenere **anche una misura quantitativa (un numero) della correlazione** tra più variabili quantitative osservate su uno stesso gruppo di individui.

Questa misura quantitative prende il nome di **covarianza campionaria** e ci fornisce un **indice che descrive la dipendenza (se esiste)** tra X e Y.



Covarianza campionaria

Data una **coppia (X,Y) di variabili quantitative**:

```
X <- c(24 ,26 ,30 ,25 ,29 ,27 ,20 ,29 ,27 ,28 ,18 ,21 ,26 ,30 ,28);
Y <- c(27 ,26 ,29 ,26 ,30 ,27 ,22 ,29 ,27 ,28 ,20 ,20 ,27 ,30 ,30);
```

C =$(x_1,y_1,)$, (x_2,y_2), …, (x_n,y_n) con n = |C| è il nostro **campione** di n **osservazioni.**

Utilizziamo la definizione di **media campionaria** e calcoliamola per le variabili quantitative del campione.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Definiamo la covarianza campionaria Cxy tra X e Y del campione C con la formula seguente.

$$C_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y})$$

Covarianza campionaria /1

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Osservazioni:

- Con X_i indichiamo i valori assunti da X;
- Se $X_i > \overline{x}$ allora la differenza sarà **positiva**;
- Se $X_i < \overline{x}$ allora la differenza sarà **negativa**;
- Lo stesso ragionamento si applica ai valori assunti da Y.

Covarianza campionaria /2

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Come si comporta il prodotto $(x_i - \overline{x})\,(y_i - \overline{y})$?

- Assumerà un **valore positivo** se **$X_i > \overline{x}$ e $Y_i > \overline{y}$** oppure se **$X_i < \overline{x}$ e $Y_i < y$** – definiamo questo comportamento **correlazione positiva**.
- Assumerà un **valore positivo** se **$X_i > \overline{x}$ e $Y_i < \overline{y}$** oppure se **$X_i < \overline{x}$ e $Y_i > y$** – definiamo questo comportamento **correlazione negativa**.

Notate sempre che è una sorta di **«forma di regolarità»**

Covarianza campionaria /3

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})\,(y_i - \overline{y})$$

Come si comporta il prodotto $(x_i - \overline{x})\,(y_i - \overline{y})$?

- **Se** l'intero campione **presenta una forte correlazione** (negativa o positiva) **«ci aspettiamo»** che la sommatoria di tutti gli scarti produca un valore **«molto positivo» oppure «molto negativo».**
- Se **$C_{xy} > 0$** consideriamo X e Y **correlate positivamente**;
- Se **$C_{xy} < 0$** le consideriamo **correlate negativamente**;
- Se **$C_{xy} = 0$** le consideriamo **NON CORRELATE**.

Covarianza campionaria /3

$$C_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

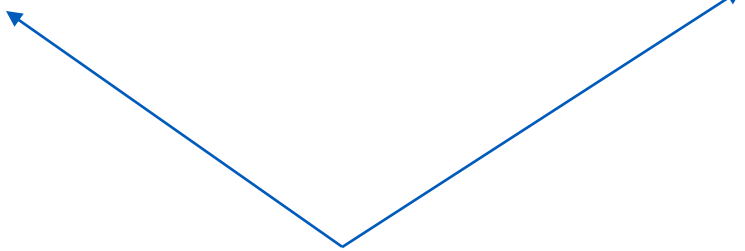
- Si normalizza la sommatoria in modo che **C_{xy} assuma lo stesso valore della varianza campionaria** nel caso in cui $X_i = Y_i$
- **Se** $X_i = Y_i \rightarrow$ **$C_{xy}=S^2$**

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

Covarianza campionaria /4

Un **metodo alternativo** per **calcolare la covarianza campionaria** è il seguente:

$$C_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) \quad = \frac{n}{n-1}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\,y_i - \overline{x}\,\overline{y}\right],$$



Le due **formule sono equivalenti**. Per esercitarsi e convincersi della veridicità di questa uguaglianza, provate a passare dalla formula di sinistra a quella di destra.

Coefficiente di Correlazione Campionario /1

E' un metodo alternativo per ottenere una misura **quantitativa** della **correlazione** tra le variabili.

Data la solita **coppia (X,Y) di variabili quantitative**:
C =(x_1,y_1,), (x_2,y_2), …, (x_n,y_n) con n = |C| è il nostro **campione** di n **osservazioni.**

```
X <- c(24 ,26 ,30 ,25 ,29 ,27 ,20 ,29 ,27 ,28 ,18 ,21 ,26 ,30 ,28);
Y <- c(27 ,26 ,29 ,26 ,30 ,27 ,22 ,29 ,27 ,28 ,20 ,20 ,27 ,30 ,30);
```

Sia $\overline{x}$ la **media campionaria dei valori assunti da X in C (x_1, x_2, …, x_n)** $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sia **S_X** la **deviazione standard campionaria** di **(x_1, x_2, …, x_n)** $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}$

NOTA: E' la radice quadrata della varianza campionaria (S^2)

Sia $\overline{y}$ la **media campionaria** dei valori assunti da Y in C **(y_1, y_2, …, y_n)** $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Sia **S_Y** la **deviazione standard campionaria** di **(y_1, y_2, …, y_n)**

Coefficiente di Correlazione Campionario /2

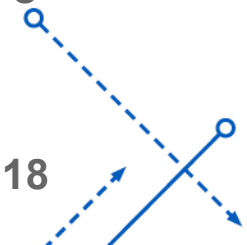
E' un metodo alternativo per ottenere una misura **quantitativa** della **correlazione** tra le variabili.

$$r_{xy} = \frac{C_{xy}}{s_x \, s_y}.$$

- Se Rxy > 0 → **correlazione positiva**.
- Se Rxy < 0 → **correlazione negativa**.
- Se Rxy = 0 → NO correlazione.

Proprietà:

- E' **adimensionale**;
- **Ha sempre valore nell'intervallo [-1, 1];**
- Non fa distinzione tra variabile indipendente e dipendente;
- Esiste solo se le variabili sono quantitative;
- **Non risente dell'unità di misura delle variabili misurate**;
- E' fortemente influenzato (usa la media campionaria) dagli outlier (valori anomali);
- Ha lo stesso **segno della covarianza**.



Coefficiente di Correlazione Campionario /2

E' un metodo alternativo per ottenere una misura **quantitativa** della **correlazione** tra le variabili.

$$r_{xy} = \frac{C_{xy}}{s_x \, s_y}.$$

- Se Rxy > 0 → **correlazione positiva**.
- Se Rxy < 0 → **correlazione negativa**.
- Se Rxy = 0 → NO correlazione.

Proprietà:

- E' **adimensionale**;
- **Ha sempre valore nell'intervallo [-1, 1];**
- Non fa distinzione tra variabile indipendente e dipendente;
- Esiste solo se le variabili sono quantitative;
- **Non risente dell'unità di misura delle variabili misurate**;
- E' fortemente influenzato (usa la media campionaria) dagli outlier (valori anomali);
- Ha lo stesso **segno della covarianza**.

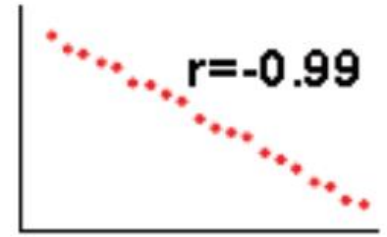
Coefficiente di Correlazione Campionario /3

Ulteriori importanti proprietà del Coeff. Di correlazione Campionario:

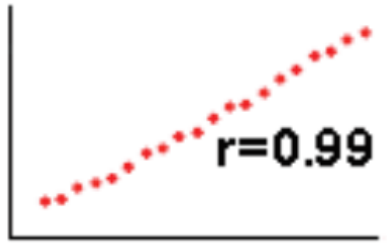
(1) $-1 \leq r_{xy} \leq 1;$

(2) (3) Queste due proprietà **mostrano** che i valori limite −1 e +1 sono effettivamente raggiunti **solo quando tra X e Y sussiste una relazione lineare**, ossia quando i punti dello scatterplot giacciono tutti su di una retta.

(4) *se esistono quattro numeri reali* a, b, c, d *e se risulta* $z_i = a\,x_i + b$ *e* $w_i = c\,y_i + d$ *per* $i = 1, 2, \ldots, n$, *allora* $r_{zw} = r_{xy}$ *se* $ac > 0$ *e* $r_{zw} = -r_{xy}$ *se invece* $ac < 0$.



max correlazione negativa



max correlazione positiva

Coefficiente di Correlazione Campionario /4

Ulteriori importanti proprietà del Coeff. Di correlazione Campionario:

(1) $-1 \leq r_{xy} \leq 1$;

(2) *se esistono due numeri reali* a *e* b, *con* $a > 0$, *tali che* $y_i = a\,x_i + b$ *per ogni* $i = 1, 2, \ldots, n$, *allora* $r_{xy} = 1$;

(3) *se esistono due numeri reali* a *e* b, *con* $a < 0$, *tali che* $y_i = a\,x_i + b$ *per ogni* $i = 1, 2, \ldots, n$, *allora* $r_{xy} = -1$;

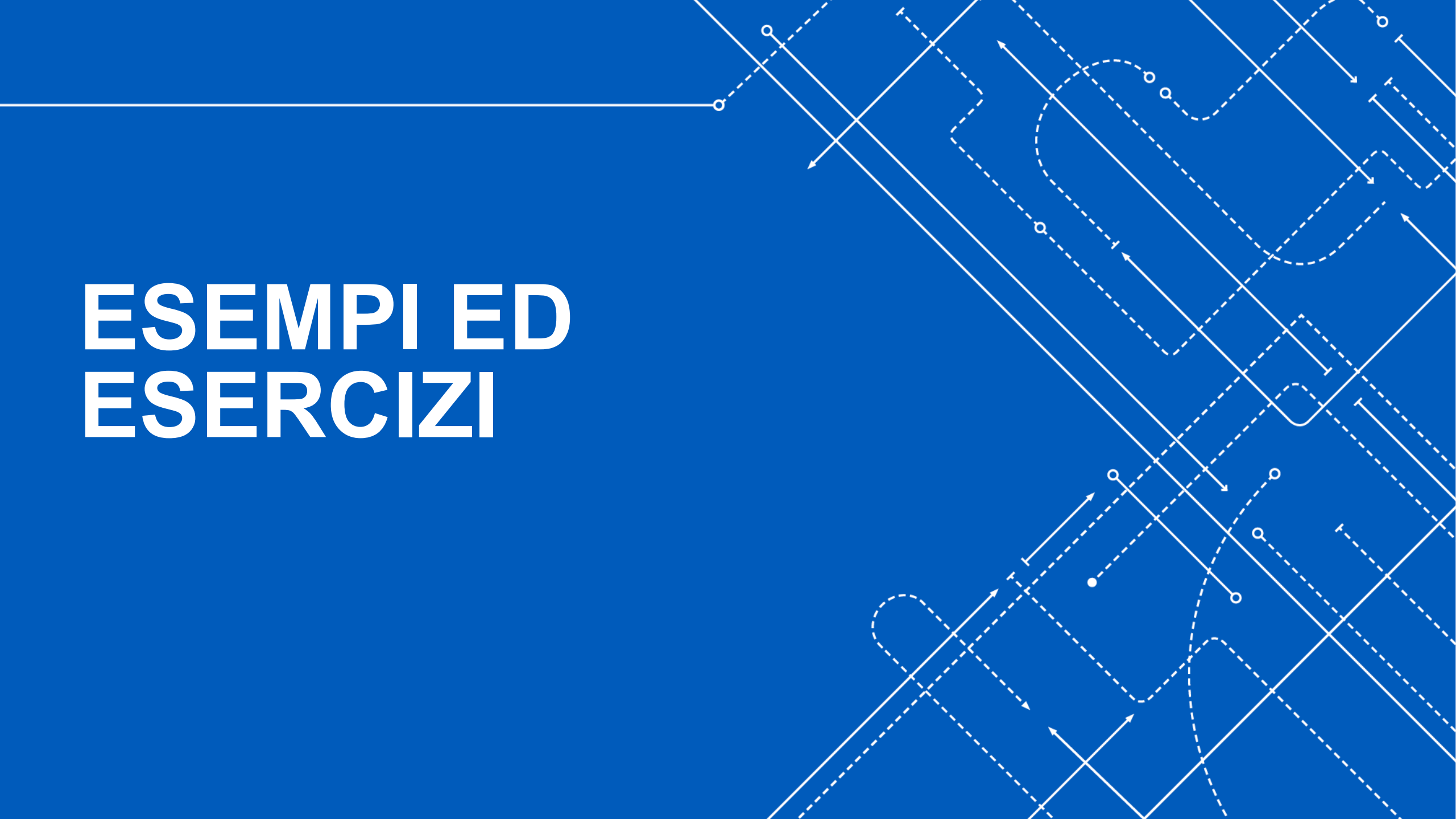
(4) Questa proprietà afferma che il quadrato del coefficiente di correlazione non cambia se sommiamo costanti o moltiplichiamo per costanti tutti i valori di X e/o di Y → **il coefficiente di correlazione non dipende dalle unità di misura** scelte per rappresentare i dati.

Coefficiente di Correlazione Campionario /2

Occorre ricordare che il coefficiente di correlazione campionario rxy misura la forza del legame di natura lineare esistente tra due variabili quantitative. Eventuali relazioni tra le variabili **che assumono una forma curvilinea non possono pertanto essere individuati** con tale coefficiente.

$$r_{xy} = \frac{C_{xy}}{s_x \, s_y}.$$

ESEMPI ED ESERCIZI



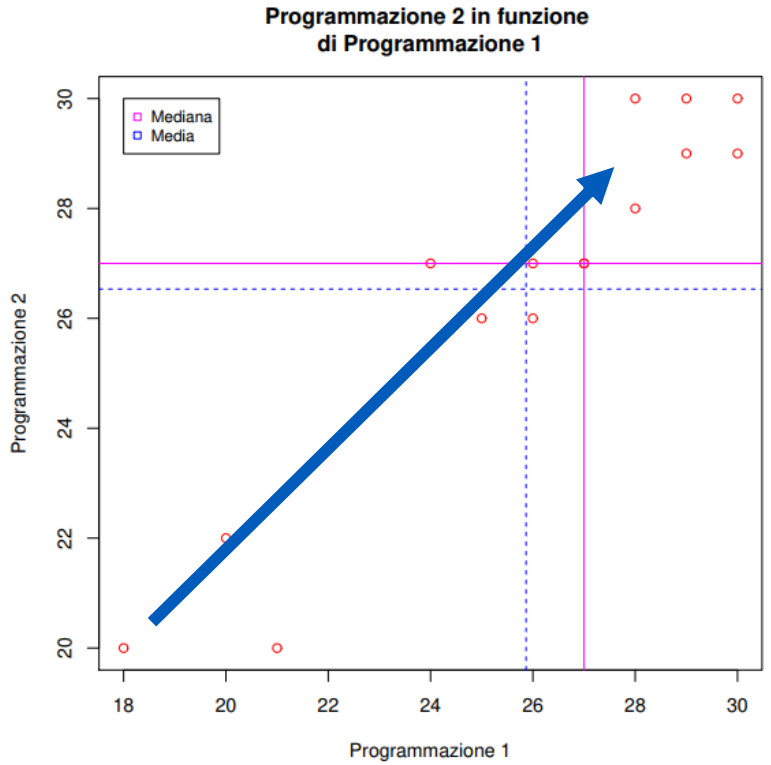
Esempio 1: Breve analisi esplorativa

Il nostro campione C è composto da 15 coppie di osservazioni (Xi,Yi) sui voti di esami di Programmazione 1 (X) e di Programmazione 2 (Y).

In precedenza, usando uno scatterplot abbiamo avuto osservato un andamento che ci suggeriva una «relazione» tra X e Y

Analizziamo ora **quantitativamente** cosa accade tra X e Y:

- In R è possibile usare la funzione **cov** per calcolare la **covarianza** tra X e Y e la funzione **cor** per calcolare il **coefficiente di correlazione** tra X e Y



Esempio 1: Breve analisi esplorativa /2

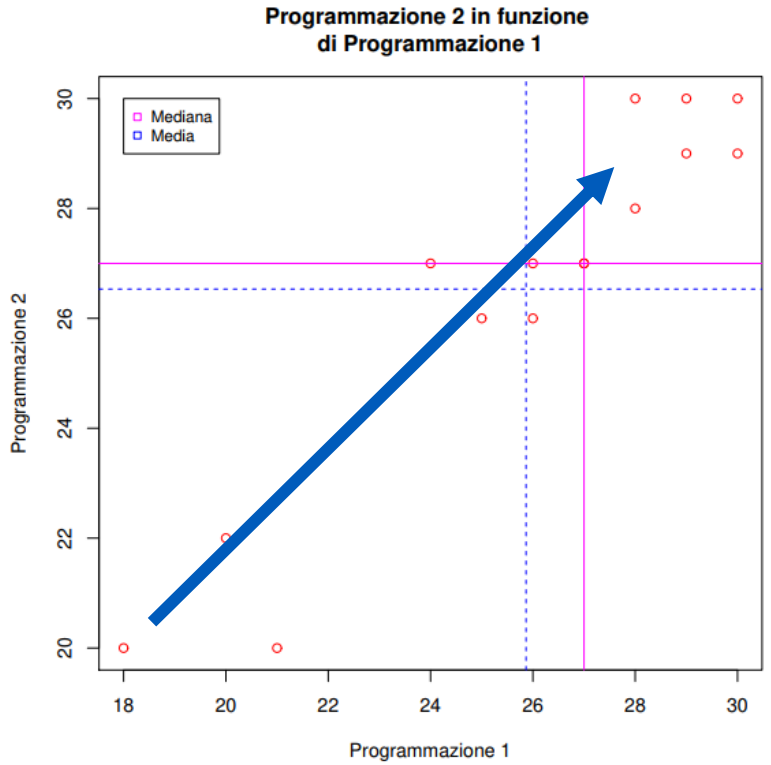
Osservando lo scatterplot:

- **Che valore vi aspettate assumeranno cov(X,Y) e cor(X,Y)?**

```
X <- c(24 ,26 ,30 ,25 ,29 ,27 ,20 ,29 ,27 ,28 ,18 ,21 ,26 ,30 ,28);
Y <- c(27 ,26 ,29 ,26 ,30 ,27 ,22 ,29 ,27 ,28 ,20 ,20 ,27 ,30 ,30);
df <-data.frame (X,Y);
#plot(df$X,df$Y);
median(df$X)
median(df$X)
sd(df$X)
median(df$Y)
median(df$Y)
sd(df$Y)

cov(df$X,df$Y);
cor(df$X,df$Y);
```

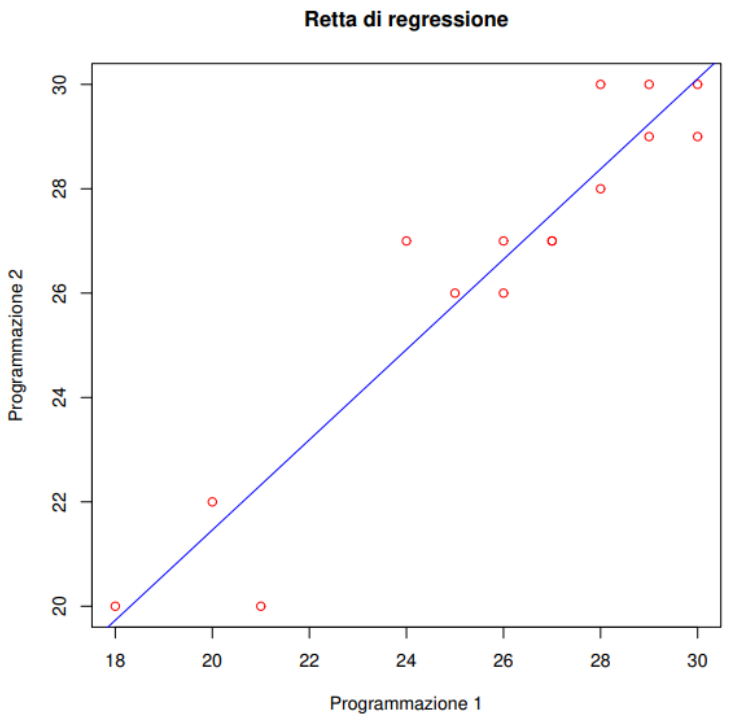
← 11.71905
←0.9483896 (forte correlazione lineare positiva)



Esempio 1: Breve analisi esplorativa /3

E' possibile costruire uno **scatterplot** mostrando anche la linea **interpolante** stimata.

```
plot(df$X ,df$Y ,main =" Retta di regressione ", xlab=" Programmazione 1",ylab=" Programmazione 2", col ="red ")
abline (lm(df$Y~df$X ), col =" blue")
```



Alcune osservazioni importanti:

- Gli scatterplot permettono di **visualizzare** le eventuali **relazioni** che possono intercorrere tra variabili quantitative;
- Covarianza e Coeff. Di correlazione ci permettono di ragionare quantitativamente sulle eventuali relazioni tra variabili;
- Tuttavia per studiare in modo accurato le relazioni è necessario utilizzare **altre tecniche statistiche** in grado di misurare con maggiore precisione questo legame.

Esercizio (per casa)

Identificate un **fenomeno multimodale misurabile** ed eseguitene un'analisi statistica esplorativa utilizzando in R le funzioni studiate fino a questo punto.

Discutete in modo libero dei risultati ottenuti.