Il **Coefficiente di Correlazione di Pearson** è una misura statistica che indica la **forza** e la **direzione** della relazione lineare tra due variabili numeriche. È uno degli indici di correlazione più utilizzati in statistica e data analysis.

Definizione del Coefficiente di Correlazione di Pearson

Il coefficiente di Pearson, indicato con r, varia tra **-1** e **+1**:

- $r = +1$ indica una **correlazione positiva perfetta**: al crescere di una variabile, anche l'altra cresce in modo proporzionale.
- $r = -1$ indica una **correlazione negativa perfetta**: al crescere di una variabile, l'altra diminuisce in modo proporzionale.
- $r = 0$ indica **assenza di correlazione lineare**: le due variabili non mostrano una relazione lineare; possono comunque esserci relazioni non lineari.

Formula del Coefficiente di Pearson

La formula per calcolare r è:

$$r = \frac{\sum(X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum(X - \overline{X})^2 \cdot \sum(Y - \overline{Y})^2}}$$

dove:

- X e Y sono le due variabili,
- $\overline{X}$ e $\overline{Y}$ sono le medie di X e Y,
- $(X - \overline{X})$ e $(Y - \overline{Y})$ rappresentano gli scarti di ciascun valore dalla propria media,
- $\sum$ indica la somma su tutti i valori.

La formula può essere interpretata come il **prodotto tra gli scarti** di ogni valore e la **deviazione standard** di ciascuna variabile. Questo rapporto fa sì che la correlazione sia **standardizzata** tra -1 e +1.

Interpretazione del Coefficiente di Pearson

1. $r \approx +1$: forte correlazione positiva, le variabili crescono o decrescono insieme.
2. $r \approx -1$: forte correlazione negativa, all'aumentare di una variabile, l'altra tende a diminuire.
3. $r \approx 0$: nessuna correlazione lineare, le variabili non sono linearmente correlate (ma potrebbero esserci relazioni non lineari).

Esempio di Calcolo

Immaginiamo di avere due variabili, **X** (ore di studio) e **Y** (punteggio su un test). I valori sono:

Ore di Studio (X)	Punteggio Test (Y)
1	50
2	55
3	65
4	70
5	80

Step 1: Calcoliamo le medie di X e Y:

$$\overline{X} = \frac{1+2+3+4+5}{5} = 3$$

$$\overline{Y} = \frac{50+55+65+70+80}{5} = 64$$

Step 2: Calcoliamo lo scarto dalla media per ogni valore, moltiplichiamo gli scarti di X e Y e sommiamo i prodotti:

$$\sum (X-\overline{X})(Y-\overline{Y}) = (1-3)(50-64) + (2-3)(55-64) + (3-3)(65-64) + (4-3)(70-64)$$

$$= 28 + 9 + 0 + 6 + 32 = 75$$

Step 3: Calcoliamo la deviazione standard di X e Y:

$$\sqrt{\sum (X-\overline{X})^2} = \sqrt{(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2} = \sqrt{10}$$

$$\sqrt{\sum (Y-\overline{Y})^2} = \sqrt{(-14)^2 + (-9)^2 + 1^2 + 6^2 + 16^2} = \sqrt{498}$$

Step 4: Inseriamo tutto nella formula:

$$r = \frac{75}{\sqrt{10 \cdot 498}} \approx 0.95$$

Questo valore vicino a **+1** indica una **forte correlazione positiva** tra le ore di studio e il punteggio ottenuto nel test: al crescere delle ore di studio, anche il punteggio tende a crescere.

Limitazioni del Coefficiente di Pearson

- **Sensibile agli outlier**: Un valore anomalo può distorcere il risultato.
- **Valido solo per relazioni lineari**: Se la relazione tra le variabili è non lineare, Pearson potrebbe non rilevarla.
- **Non applicabile a variabili categoriche**: Funziona solo con variabili quantitative.

Analisi della correlazione di Spearman o di Kendall

Supponiamo di avere un dataset di esempio con due variabili: **Ore di studio** e **Punteggio ottenuto in un test**.

Esempio di Dataset

A	B
Ore di studio	**Punteggio test**
2	50
4	55
6	65
8	70
10	80
3	52
5	60
7	68
9	75
1	45

Passaggi per Calcolare la Correlazione di Spearman

1. **Calcola i ranghi** per ciascuna variabile:
    - Assegna un rango a ogni valore delle variabili **Ore di studio** e **Punteggio test**, dove il valore più basso ha rango 1, il successivo 2, e così via.

Ore di studio	Rango Ore	Punteggio test	Rango Punteggio
2	2	50	3
4	4	55	4
6	6	65	7
8	8	70	8
10	10	80	10
3	3	52	2
5	5	60	5
7	7	68	9
9	9	75	6
1	1	45	1

2. **Calcola la differenza tra i ranghi** (D) e il quadrato di tale differenza (D^2):
   - Differenza tra i ranghi: $D = \text{Rango Ore} - \text{Rango Punteggio}$
   - Quadrato della differenza: D^2

Rango Ore	Rango Punteggio	D	D^2
2	3	-1	1
4	4	0	0
6	7	-1	1
8	8	0	0
10	10	0	0
3	2	1	1
5	5	0	0
7	9	-2	4
9	6	3	9
1	1	0	0

3. **Somma** D^2:

$$\sum D^2 = 1 + 0 + 1 + 0 + 0 + 1 + 0 + 4 + 9 + 0 = 16$$

4. **Calcola il coefficiente di correlazione di Spearman** (r_s):

$$r_s = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

dove n è il numero di osservazioni (in questo caso, 10).

Inseriamo i valori:

$$r_s = 1 - \frac{6 \times 16}{10(10^2 - 1)} = 1 - \frac{96}{990} \approx 0.903$$

Interpretazione del Risultato

Il valore di **0.903** indica una **forte correlazione positiva** tra le **Ore di studio** e il **Punteggio del test**, anche se non necessariamente lineare.

Come Calcolare in Excel

Per eseguire questo calcolo in Excel:

1. **Ordina i dati** e assegna manualmente i ranghi o usa la funzione `RANGO` per calcolarli automaticamente:

```excel
=RANGO(A2; $A$2:$A$11; 1)  // Per calcolare il rango di una cella
```

2. Calcola D e D^2 usando le formule di sottrazione e potenza:

```excel
=B2 - C2  // Calcolo di D
=D2^2     // Calcolo di D^2
```

3. Somma i D^2 con `SOMMA` e applica la formula per r_s.

Esempio di Correlazione Negativa

Immaginiamo di avere un dataset che mostra il numero di ore di allenamento fisico al giorno e il peso corporeo in kg:

Ore di allenamento al giorno (X)	Peso corporeo (kg) (Y)
1	90
2	85
3	80
4	75
5	70

Analisi: Se calcoliamo il coefficiente di correlazione di Pearson, troveremo un valore **negativo** (ad esempio, -0.95), che indica una **forte correlazione negativa**. Questo significa che all'aumentare delle ore di allenamento, il peso corporeo tende a diminuire in modo proporzionale.

Interpretazione: La correlazione negativa indica che esiste una relazione inversa tra le due variabili: se una aumenta, l'altra tende a diminuire.

Esempio di Correlazione Indecidibile (Vicina a Zero)

Immaginiamo di avere un dataset con il numero di libri letti al mese e la temperatura esterna in gradi Celsius:

Libri letti al mese (X)	Temperatura esterna (°C) (Y)
3	15
4	20
2	18
5	10
1	25

Analisi: Se calcoliamo il coefficiente di correlazione di Pearson per queste due variabili, potremmo ottenere un valore molto vicino a **0** (ad esempio, **0.1** o **-0.05**). Questo indica che non c'è una correlazione lineare evidente tra i libri letti e la temperatura esterna.

Interpretazione: La correlazione vicina a zero suggerisce che non esiste una relazione chiara tra le due variabili, o che la relazione è talmente debole da essere considerata **indecisa** o insignificante per l'analisi lineare.