Xi	ni	Xi- media aritmetica (scarti)	Scarti al quadrato	Scarti al quadrato moltiplicati per le frequenze assolute
1	1	1-2,43=-1,43	(-1,43)^2=2,04	2,05*1=2,04
2	2	2-2,43=-0,43	(-0,43)^2=0,18	0,18*2=0,36
3	4	3-2,43=0,57	0,57^2=0,32	0,32*4=1,28
4	0	4-2,43=1,57	2,46	2,46*0=0

Sappiamo che per una tabella di frequenza la media aritmetica è data dalla somma dei prodotti tra modalità e frequenze assolute, il tutto diviso il numero totale di unità statistiche. Nel nostro caso:

Media aritmetica= (1*1+2*2+3*4+4*0)/7= 17/7=2,43

Moda=3 in quanto è la modalità che si presenta con maggiore frequenza (4 volte)

Per il calcolo della varianza occorre prima calcolare gli scarti dalla media aritmetica, ovvero le differenze tra ciascuna modalità e la media aritmetica, poi elevare tali scarti al quadrato. Gli scarti al quadrato vanno moltiplicati per le frequenze assolute. Tali prodotti vanno sommati e l'ammontare ottenuto va diviso per il numero di unità statistiche.

Varianza= (2,04+0,36+1,28)/7= 0,36

Mediana: occorre ordinare i dati in misura non decrescente: 1 2 2 **3** 3 3 3

Poiché abbiamo un numero dispari di osservazioni (7 in totale), la mediana va individuata nella modalità che occupa la posizione (n+1)/2= (7+1)/2=4, per cui corrisponde alla modalità 3, evidenziata in grassetto.

In alternativa, soprattutto quando abbiamo una tabella di frequenza con un numero di osservazioni molto elevato, per il calcolo della mediana si può ricorrere alla frequenza cumulata. Nel nostro caso:

Xi	Ni (Frequenza cumulata)	Posizioni occupate
1	1	1°
2	2+1=3	2°-3°
3	3+4=7	4°-7°

E' inutile inserire la modalità 4, visto che non si presenta mai.

In base a tale tabella, la modalità 1 occupa la prima posizione, la modalità 2 la seconda e la terza, la modalità 3 dalla quarta alla settima. Siccome la mediana occupa la quarta posizione, essa corrisponde alla modalità 3.

Se avessimo avuto una numerosità n pari, la mediana l'avremmo dovuta calcolare facendo la semisomma tra la modalità che occupava la posizione n/2 e quella che occupava la posizione immediatamente successiva. Nel caso, ad esempio di n=8, la mediana è la semisomma tra la modalità che occupa la quarta posizione e quella che occupa la quinta posizione.

Xi	Yi	Xi – Media di X (Scarti di X)	Yi – Media di Y (Scarti di Y)	Prodotto degli scarti di X ed Y
0	0	0-1,2= -1,2	0-1,2= -1,2	(-1,2)*(-1,2)= 1,44
1	1	1-1,2= -0,2	1-1,2= -0,2	(-0,2)*(-0,2)= 0,04
1	2	1-1,2= -0,2	2-1,2= 0,8	(-0,2)*0,8= -0,16
2	2	2-1,2= 0,8	2-1,2= 0,8	0,8*0,8= 0,64
2	1	2-1,2= 0,8	1-1,2= -0,2	0,8*(-0,2)= -0,16

Media di X = (0+1+1+2+2)/5= 6/5=**1,2**

Media di Y = (0+1+2+2+1)/5= 6/5=**1,2**

Il coefficiente di correlazione si ottiene dal rapporto tra la covarianza di X e Y e il prodotto dei due scarti quadratici medi di X e Y, ovvero le radici quadrate delle rispettive varianze.

La covarianza è la contemporanea variazione di X e Y. Essa si ottiene sommando il prodotto degli scarti di X e Y e dividendo il totale ottenuto per il numero di unità statistiche. Quindi occorre calcolare prima gli scarti delle due variabili, moltiplicarli tra loro e sommare i prodotti. Alla fine si divide l'ammontare ottenuto per il numero di unità statistiche.

Nel nostro esercizio, la sommatoria del prodotto degli scarti è la seguente: 1,44+0,04-0,16+0,64-0,16= 1,8. Il numero di unità statistiche è pari a 5, per cui la covarianza è pari a 1,8/5=**0,36**.

Abbiamo calcolato il numeratore del coefficiente di correlazione. Per calcolare il denominatore ci servono i due scarti quadratici medi, quindi occorrono le due varianze.

Procediamo.

Eleviamo al quadrato gli scarti sia di X che di Y:

Scarti di X al quadrato	Scarti di Y al quadrato
(-1,2)^2 = 1,44	(-1,2)^2 = 1,44
(-0,2)^2 = 0,04	(-0,2)^2 = 0,04
(-0,2)^2 = 0,04	0,8^2 = 0,64
0,8^2 = 0,64	0,8^2 = 0,64
0,8^2= 0,64	(-0,2)^2 = 0,04

Varianza di X = (1,44+0,04+0,04+0,64+0,64)/5 = 2,8/5 = **0,56**

Varianza di Y = (1,44+0,04+0,64+0,64+0,04)/5 = 2,8/5 = **0,56**

Lo scarto quadratico medio di X è la radice quadrata di 0,56 ed è uguale a **0,75**. Lo stesso dicasi per

Y.

In conclusione, quindi, il coefficiente di correlazione è dato da: 0,36/(0,75*0,75) = **0,64**.
Il coefficiente di correlazione è sempre compreso tra -1 e 1. Quanto più si avvicina a -1 tanto più la correlazione è forte e negativa, quanto più si avvicina a 1 tanto più la correlazione è forte e positiva. Quando tale indice si avvicina a 0 la correlazione è debole.

Solitamente gli intervalli da considerare sono i seguenti:

tra -1 e -0,7 correlazione forte negativa

tra -0,7 e -0,3 correlazione moderata negativa

tra -0,3 e 0 correlazione debole negativa

tra 0 e 0,3 correlazione debole positiva

tra 0,3 e 0,7 correlazione moderata positiva

tra 0,7 e 1 correlazione forte positiva

Nel nostro caso l'indice di correlazione è uguale a 0,64, per cui la correlazione tra X e Y è moderata e positiva.

Un modello di regressione lineare è espresso dalla seguente relazione:

Y= a +bX + e,

dove Y rappresenta la variabile dipendente, ovvero il fenomeno che intendiamo spiegare, X la variabile indipendente, ovvero la causa che determina il comportamento di Y, a e b sono i coefficienti della regressione ed e rappresenta l'errore, ovvero ciò che non riusciamo a spiegare di Y.

L'obiettivo del modello è fare in modo di spiegare quanto più è possibile Y attraverso X e di ridurre al minimo gli errori, ovvero la componente di Y non spiegata.

In tale ottica assumono una grande importanza i parametri a e b, che vengono stimati con il metodo dei minimi quadrati, finalizzato proprio a minimizzare gli errori. In particolare, è il coefficiente b quello propriamente detto "di regressione", in quanto **rappresenta il legame che unisce X ad Y.** Se b è positivo infatti vuol dire che al crescere di X cresce anche Y, se b è negativo allora al crescere di X Y decresce.

Se consideriamo per semplificare un'equazione del tipo Y=2X e ad X assegniamo valore 1, Y vale 2, se X vale 2 Y vale 4, se X vale 3 Y vale 6 e così via. In questo caso, al crescere di un'unità di X, Y cresce di 2 unità.

L'indice di determinazione lineare R^2 invece esprime una misura della bontà di adattamento del modello al fenomeno reale. Esso è dato dal rapporto tra la varianza delle stime di Y (ovvero a + b*X) e la varianza dei valori osservati di Y. Infatti, l'errore altro non è che la differenza tra ciò che realmente si osserva di Y e il valore teorico assunto dalla stessa variabile. In quanto rapporto tra due varianze (quindi tra due valori positivi), di cui quella al numeratore ovviamente più piccola di quella al denominatore (basti pensare che Y = Y stimato + l'errore) **R^2 è sempre compreso tra 0 ed 1.** Quanto più R^2 si avvicina ad 1 tanto più il modello è valido. Se ad esempio R^2 è uguale a 0,97, vuol dire che il modello di regressione spiega per il 97% il comportamento di Y e c'è solo una componente del 3% non spiegata. E' chiaro quindi che l'indice R^2 ci indica anche la bontà della scelta di X quale variabile indipendente per spiegare il comportamento di Y.

Un'ultima osservazione. **La distribuzione teorica di Y rappresenta l'insieme dei valori che la variabile Y assumerebbe se essa dipendesse esclusivamente da X**, cioè se non contemplassimo la presenza degli errori. Gli errori sono infatti le differenze tra i valori osservati di Y e i valori teorici di Y.