Pearson Product-Moment Correlation Coefficient

The **Pearson Product - Moment Correlation Coefficient** or the Pearson Product - Moment Test is a statistical test that assesses the degree to which two sets of paired data are **correlated** and the **strength** of that correlation. It is often used as a more precise test than the Spearman's Rank Correlation Coefficient as it uses real data rather than the rankings of the latter.

Note: You can use the **Pearson Product - Moment Test** for data that is

- paired together
- normally distributed

Worked example:

A geographer wanted to test the theory that with distance from the source of a river, the size of the bedload will decrease. They worked with the following hypotheses to be tested through the Pearson Product - Moment Test:

H_1 There is a correlation between the size of the bedload and the distance from the source of the river.

H_0 There is no correlation between the size of the bedload found and the distance from the source of the river.

The following data was collected:

Distance from source (m)	Average bedload size (mm)
10	112
250	86
840	91
1570	45
2110	34
2750	10
3540	10
4890	3

First the mean distance from source ($\bar{x}$) and the mean bedload size ($\bar{y}$) are calculated, followed by the means minus the data (i.e. $\bar{x}$-x). This latter figure is known as the deviation (dx or dy).

Distance (x)	dx	dx^2	Size (y)	dy	dy^2	dx x dy
10	1985		112	-63		
250	1745		86	-37		
840	1155		91	-42		
1570	425		45	4		
2110	-115		34	15		
2750	-755		10	39		
3540	-1545		10	39		
4890	-2895		3	46		
$\bar{x}$ = 1995			$\bar{y}$ = 49			

For each piece of data, dx and dy are squared.

Distance (x)	dx	dx^2	Size (y)	dy	dy^2	dx x dy
10	1985	3940225	112	-63	3969	
250	1745	3045025	86	-37	1369	
840	1155	1334025	91	-42	1764	
1570	425	180625	45	4	16	
2110	-115	13225	34	15	225	
2750	-755	570025	10	39	1521	
3540	-1545	2387025	10	39	1521	
4890	-2895	8381025	3	46	2116	
$\bar{x}$ = 1995		Σ = 19851200	$\bar{y}$ = 49		Σ = 12501	

Finally, dx and dy are multiplied to give data in the final column.

Distance (x)	dx	dx^2	Size (y)	dy	dy^2	dx x dy
10	1985	3940225	112	-63	3969	-125055
250	1745	3045025	86	-37	1369	-64565
840	1155	1334025	91	-42	1764	-48510
1570	425	180625	45	4	16	1700
2110	-115	13225	34	15	225	-1725
2750	-755	570025	10	39	1521	-29445
3540	-1545	2387025	10	39	1521	-60255
4890	-2895	8381025	3	46	2116	-133170
$\bar{x}$ = 1995		Σ = 19851200	$\bar{y}$ = 49		Σ = 12501	Σ = -461025

Using the following formula, the Pearson Product - Moment Correlation Coefficient (r) can then be calculated:

$$r = \frac{\Sigma\,(dx \times dy)}{\sqrt{\Sigma\,(dx)^2 \times \Sigma\,(dy)^2}}$$

$$r = \frac{-461025}{\sqrt{19851200 \times 12501}}$$

$$r = \frac{-461025}{\sqrt{248159851200}}$$

$$r = \frac{-461025}{498156}$$

$$r = -0.92$$

The negative calculated r value indicates a **negative** correlation between the two variables. Therefore it appears that as distance from the source increases, the size of the bedload decreases. The calculated value also tells the geographer that they are looking at a **strong** negative correlation.

The geographer then wishes to know if their result simply occurred by chance or whether the same result would be found if repeated. To find this, the researcher compares the calculated value (r) with the critical value for the appropriate number of sets of paired data. This can be found in the significance table overleaf.

Degrees of Freedom	Significance level	
	0.05	0.01
1	0.9877	0.995
2	0.900	0.980
3	0.805	0.934
4	0.729	0.882
5	0.669	0.833
6	0.622	0.789
7	**0.582**	0.750
8	0.549	0.716
9	0.521	0.685
10	0.497	0.658
11	0.476	0.634
12	0.458	0.612
13	0.441	0.592
14	0.426	0.574
15	0.412	0.558
16	0.400	0.543
17	0.389	0.529
18	0.378	0.516
19	0.369	0.503
20	0.360	0.492

Degrees of Freedom = n - 1, where n is the number of paired observations.

Therefore in this case, the degrees of freedom is **7**.

From the significance table, we can read off a critical value of **0.582** at 95% significance that the results are meaningful and did not occur by chance. As the **calculated value** (regardless of correlation direction) **is greater than the critical value**, the null hypothesis can be rejected.

Therefore, there is a significant and strong negative correlation between the distance from the river's source and bedload size in this particular case.