Linear Correlation

A correlation exists between two variables where one of the variables is related to the other in some way. **Linear correlation** is when the relationship that exists between the two variables is linear.

The degree of linear correlation is found by calculating the Pearson's Correlation Coefficient.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Where x and y are the variables whose relationship is in question. The value of r will always be between -1 and 1 inclusive.

Remember r is a measure of the strength of a linear association between two variables.

Example

Suppose the age and distance (to the nearest mile) moved was recorded for 25 adults who moved to Phoenix from outside the state of Arizona. Is there a significant linear correlation between the age of the adult and the distance moved?

Age	Distance (mi.)
25	1852
37	1603
72	450
47	975
72	373
41	1336
59	586
32	282
59	202
37	1400
80	287
63	1801
80	1013
31	356
38	375
31	816
45	749
36	2367

45	2368
22	2721
45	2341
44	2455
48	1001
31	1725
48	1075

Let the ages be X and the distances be Y. To calculate the Correlation Coefficient make a table.

	X=Age	Y=Distance (mi.)	XY	X^2	Y^2
	25	1852	46300	625	3429904
	37	1603	59311	1369	2569609
	72	450	32400	5184	202500
	47	975	45825	2209	950625
	72	373	26856	5184	139129
	41	1336	54776	1681	1784896
	59	586	34574	3481	343396
	32	282	9024	1024	79524
	59	202	11918	3481	40804
	37	1400	51800	1369	1960000
	80	287	22960	6400	82369
	63	1801	113463	3969	3243601
	80	1013	81040	6400	1026169
	31	356	11036	961	126736
	38	375	14250	1444	140625
	31	816	25296	961	665856
	45	749	33705	2025	561001
	36	2367	85212	1296	5602689
	45	2368	106560	2025	5607424
	22	2721	59862	484	7403841
	45	2341	105345	2025	5480281
	44	2455	108020	1936	6027025
	48	1001	48048	2304	1002001
	31	1725	53475	961	2975625
	48	1075	51600	2304	1155625
Sum	1168	30509	1292656	61102	52601255

We need a column for the X values, the Y values, X times Y, X squared, and Y squared. Find the sum of each of these columns. Then these values are plugged into the formula.

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

$$r = \frac{25(1292656) - (1168)(30509)}{\sqrt{25(61102) - (1168)^2}\sqrt{25(52601255) - (30509)^2}}$$

$$r = \frac{32316400 - 35634512}{\sqrt{1527550 - 1364224}\sqrt{1315031375 - 930799081}}$$

$$r = \frac{-3318112}{\sqrt{163326}\sqrt{384232294}} = \frac{-33318112}{(404.1361157)(19601.84415)}$$

$$r = \frac{-3318112}{7921813.155} = -.4188576447 \approx -.41886$$

This is Pearson's Correlation Coefficient, but to determine if there is a significant linear correlation present, we must have something to compare this to. Use the table of critical values for Pearson Correlation Coefficient.

For a significance level of 0.05, the critical value for a sample size of 25 is .396.

If the absolute value of r is greater than the critical value from the table, then we conclude there is significant linear correlation.

Since $|-.41886| = .41886 > .396$ = the critical value for 0.05 level of significance at a sample size of 25, then there is a significant linear correlation between the age and the distance moved.

Coefficient of Determination

Can the correlation coefficient be used to explain the variation?

Yes, the Coefficient of Determination is the proportion of variation in Y that is explained by the linear association between x and y.

The Coefficient of Determination is the square of the Pearson Correlation Coefficient.

$$\text{Coefficeint of Determination} = r^2$$

Where r is the Pearson Correlation Coefficient.

Example

Suppose the age and distance moved was recorded for 25 adults who moved to Phoenix from outside the state of Arizona. What proportion of variation in the distance moved can be explained by the linear relationship between the distance moved and the ages of those moving?

Pearson's Correlation Coefficient was calculated above to be $r = -.41886$

The Coefficient of Determination is $r^2 = (-.41886)^2 = .17544$

We can say that .17544 or 17.54% of the variation in the distance moved can be explained by the linear relationship between the distance moved and the age of the adult moving.