The Pearson's Correlation -- Analysis of the Linear Relationship Between Two Quantitative Variables

Application: To test for a linear relationship between two quantitative variables. It is important to remember that Pearson's correlation only provides information about the direction and strength of the ***linear relationship*** between the two variables. If the research hypothesis involves some other pattern of relationship (i.e., curvilinear), then some other statistical analysis will be necessary. Fortunately, researchers are usually interested in linear relationships between variables, so this is a very useful statistical test.

H0: The variables do not have a linear relationship in the population represented by the sample.

To reject H0: is to say that there is a linear relationship between the variables in the population.

The data: The quantitative variables for this analysis are **fishnum** (number of fish displayed) and **fishgood** (rating of fish quality on a 1-10 scale).

32,6 41,5 31,3 38,3 21,7 13,9 17,9 22,8 24,6 11,9 17,7 20,8

Research Hypothesis: Knowing that store owners are often over-worked, the researcher hypothesized that stores with fewer fish would have healthier fish (thus predicting a negative or inverse relationship between these variables in this population).

H0: for this analysis: There is no linear relationship between the number of fish displayed in pet stores and the quality rating of the fish.

Assemble the data for analysis. Rearrange the data so that scores from each subject are in the appropriate columns, one for each variable. One of these variables is labeled X and one Y, to simplify the presentation and use of the formulas below.

fishnum	fishgood
X	Y
32	6
41	5
31	3
38	3
21	7
13	9
17	9
22	8
24	6
11	9
17	7
20	8

Compute the square of each score and place it in an adjacent column.

fishnum		fishgood	
X	X^2	Y	Y^2
32	1024	6	36
41	1681	5	25
31	961	3	9
38	1444	3	9
21	441	7	49
13	169	9	81
17	289	9	81
22	484	8	64
24	576	6	36
11	121	9	81
17	289	7	49
20	400	8	64

Compute the cross product of the variables for each subject (X*Y) and make a column for these values.

fishnum		fishgood		
X	X^2	Y	Y^2	XY
32	1024	6	36	192
41	1681	5	25	205
31	961	3	9	93
38	1444	3	9	114
21	441	7	49	147
13	169	9	81	117
17	289	9	81	153
22	484	8	64	176
24	576	6	36	144
11	121	9	81	99
17	289	7	49	119
20	400	8	64	160

Compute Σ**X,** Σ**X²,** Σ**Y,** Σ**Y²** and Σ**XY** and determine the **N** (sample size)**.**

fishnum			fishgood					
ΣX	=	287	ΣY	=	80	ΣXY	=	1719
ΣX^2	=	7879	ΣY^2	=	584	N	=	12

Compute the mean and std for each variable.

Computational examples using data from fishnum:

$$\text{Mean} = \frac{\Sigma X_{k1}}{N} = \frac{287}{12} = 23.92$$

$$\text{Standard deviation} = \sqrt{\left[\frac{\Sigma X_{k1}^2 - [(\Sigma X_{k1})^2/N]}{N-1}\right]} = \sqrt{\left[\frac{7879 - [(287)^2/12]}{12-1}\right]} = 9.61$$

Compute the index of covariation (the extent to which X and Y are related)

$$(N * \Sigma XY) - (\Sigma X * \Sigma Y) = (12 * 1719) - (287 * 80) = 20628 - 22960 = -2332$$

Compute the variation of X

$$\sqrt{(N * \Sigma X^2) - (\Sigma X)^2} = \sqrt{(12 * 7879) - (287)^2} = \sqrt{94548 - 82369} = \sqrt{12179} = 110.36$$

Compute the variation of Y

$$\sqrt{(N * \Sigma Y^2) - (\Sigma Y)^2} = \sqrt{(12 * 584) - (80)^2} = \sqrt{7008 - 6400} = \sqrt{608} = 24.66$$

Compute the correlation coefficient (r)

$$r = \frac{\text{index of covariation}}{\text{variation of X * variation of Y}} = \frac{-2332}{110.36 * 24.66} = -.86$$

Compute the degrees of freedom for a Pearson's correlation

$$df = N - 2 = 12 - 2 = 10$$

Look up the r-critical for α = .05 and the appropriate degrees of freedom using the r-table.

$$r\text{-critical } (\alpha = .05, df = 10) = .576$$

Determine whether to retain or reject H0:

Remember correlation values can be positive or negative, and so we will compare the **absolute value** of the r to the r-critical.

-- if the absolute value of the obtained r is less than the r-critical, then retain the null hypothesis and conclude that there is no linear relationship between the two variables, in the population represented by the sample.

-- if the absolute value of the obtained r is greater than the r-critical, then reject the null hypothesis and conclude that there is a linear relationship between the variables in the population represented in the sample.

For the example data, we would decide to reject the null hypothesis, because the absolute value of the obtained r is larger than the r-critical -- |-.86| >.576.

Determine whether or not the results support the research hypohtesis.

- Usually the researcher hypothesizes that there is a correlation between the conditions, either positive or negative. If so, then to support the research hypothesis will require:
    - Reject H0: that there is no linear relationship
    - The sign of the correlation must be in the same direction as that specified in the research hypothesis

- Sometimes, however, the research hypothesis is that there is **no** correlation between the variables. If so, the research hypothesis and H0: are the same!
    - When this is the case, retaining H0: provides support for the research hypothesis, whereas rejecting H0: provides evidence that research hypothesis is incorrect.

- **Please note:** When you have decided to retain H0: (because |r| < r-critical), then don't talk about there being a positive or a negative relationship between the variables -- retaining H0: is saying that the population correlation is 0.00 (and any apparent relationship is probably due to sampling variation, chance, etc.)

For the example data, we would decide that the research hypothesis is supported, because we rejected the null hypothesis, and the negative obtained r value agrees with the negative linear relationship hypothesized by the researcher.

Describe the results of the correlation analysis -- be sure to include the following

- Name and tell the univariate statistics (mean and standard deviation) of each variable.
- Report the r-value, df (in parentheses) and p-value ($p < .05$ or $p > .05$).
- If you reject H0:, describe the direction of the correlation between the variables
    - If you retain the H0:, then say that there is no significant correlation between the variables
- Tell whether or not the results support the research hypothesis

Please note: Reporting correlation results is not a form of "creative writing". The idea is to be succinct, clear, and follow the prescribed format -- it is really a lot like completing a fill-in-the-blanks sentence. After you write and read enough of these you'll develop some "style", but for now just follow the format.

Here are two write-ups of these results that say the same thing. The 1st reports the univariates and then the significance test. The 2nd combines them into a single sentence -- either is fine.

The mean number of fish at these stores was 23.92 (S = 9.61) and the fish had a mean quality rating of 6.67 (S = 2.15). Pearson's correlation supported the research hypothesis that those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

Pearson's correlation between the number of fish displayed in these stores (M = 23.92, S = 9.61) and the quality rating for the fish (M = 6.67, S = 2.15) supported the research hypothesis that those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

"Correlation Write-up Examples for Other Occasions"

Here's are examples (not all using the same data as the computational example above) of what we would write for different combinations of RH:, r & significance test results.

For ..

- RH: there would be a positive linear relationship
- retained H0:

The mean number of fish at these stores was 23.92 (S = 9.61) and the fish had a mean quality rating of 6.67 (S = 2.15). Contrary to the research hypothesis Pearson's correlation showed no linear relationship between these two variables, r(10) = -.26, p > .05.

Pearson's correlation between the number of fish displayed in these stores (M = 23.92, S = 9.61) and the quality rating for the fish (M = 6.67, S = 2.15) did not support the research hypothesis that those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.26, p > .05.

For…

- RH: there would be no correlation between the conditions (RH: = H0:)
- retained H0:

The mean number of fish at these stores was 23.92 (S = 9.61) and the fish had a mean quality rating of 6.67 (S = 2.15). Pearson's correlation supported the hypothesis that there would be no linear relationship between these two variables, r(10) = -.26, p > .05.

Pearson's correlation between the number of fish displayed in these stores (M = 23.92, S = 9.61) and the quality rating for the fish (M = 6.67, S = 2.15) supported the research hypothesis that there is no correlation between fish number and fish quality, r(10) = -.26, p > .05.

For…

- RH: there would be a positive correlation
- rejected H0:
- but found a negative correlation

The mean number of fish at these stores was 23.92 (S = 9.61) and the fish had a mean quality rating of 6.67 (S = 2.15). Pearson's correlation revealed that contrary to the research hypothesis those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

Pearson's correlation between the number of fish displayed in these stores (M = 23.92, S = 9.61) and the quality rating for the fish (M = 6.67, S = 2.15) revealed that contrary to the research hypothesis those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

For…

- RH: there would be no mean difference between the conditions (RH: = H0:)
- Rejected H0: and found a negative relationship

The mean number of fish at these stores was 23.92 (S = 9.61) and the fish had a mean quality rating of 6.67 (S = 2.15). Pearson's correlation revealed that contrary to the research hypothesis those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

Pearson's correlation between the number of fish displayed in these stores (M = 23.92, S = 9.61) and the quality rating for the fish (M = 6.67, S = 2.15) revealed that contrary to the research hypothesis those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

Here is another write-up of the example analysis using a Table to present the univariate statistics. Tables reduce the about of parenthetical information that can clutter a write-up,.

The number and quality of fish is summarized in Table 1. Pearson's correlation supported the research hypothesis that those stores with fewer fish tended to have healthier fish, whereas those stores with more fish would tend to have fish with lower health quality, r(10) = -.86, p < .05.

Table 1
Number and Quality of Fish (N = 12)

Variable	M	S
Number of Fish	23.92	9.61
Fish Quality	6.67	2.15