

Activity

Regression & Correlation

Instructional Objectives:

1. For bivariate measurement data, be able to display a scatterplot, describe its shape, and describe, determine, and make inferences using regression coefficients, regression equations, and correlation coefficients with and without using technological tools (PSSM, 2000)
2. Determine the appropriate statistical methods, which is necessary to analyze data.

The Situation:

An elementary teacher is curious about the amount of classroom participation of her fourth grade students at the beginning of the year versus after they have become acquainted with the classroom environment. Specifically, she wants to know if there is a relationship between the participatory habits of her students during the first week of school and the tenth week and she wants to be able to predict their participatory habits.

She decided to tally the number of times that six children spoke up during the first week of the school year (X) and during the tenth week (Y). The following is her raw data:

X	Y
3	7
1	2
1	4
2	5
2	6
3	5

The Lesson:

1.
 - a. Have students create a scatterplot of the raw data. Ask the question: As we examine the data from left to right, does the data appear to move downward, upward, or in no direction?
 - b. Ask the students to visualize a straight line that would represent the data in the scatterplot. The line that fits the data should have the same characteristics as the set of points; it should actually summarize the data. Have students draw the line that seems to “fit” the data.

- c. Have students calculate the slope of the line they drew. Also, have students write a sentence or two describing the significance of the slope, including the sign. Their answer should be in terms of the wording of the original problem.

Teacher Note: After the students have calculated the slopes for their lines of best fit, have them convert them to decimal points and use the decimal representation to record the class' slope in a stem-and-leaf plot. Next, have the class decide on a "consensus" slope.

- d. Have students calculate the y-intercept. Ask students what this value means using the context of the problem.

Teacher Note: After each student has recorded his or her y-intercept, record each student's y-intercept in another stem-and-leaf plot. Next, have the class decide on a "consensus" y-intercept.

- e. Have the students draw the line that represents the "consensus" slope and y-intercept. Ask the students the following question: How much or to what degree are the two variables linearly-related? Give a percentage. How well do you think your line and the "consensus" line fits the data? Also, give a percentage.

Teacher Note: It is recommended that at this point to discuss the terms correlation coefficient, coefficient of determination, and regression equation as it relates to the exercise and in general. It is also recommended to have students, all depending upon their level, calculate the correlation coefficient, coefficient of determination, and the median fit line and/or least squares line. This will assist in understanding how FISH calculates these items. Please consult a statistics text for these procedures.

2. Have students open FISH → Set N= 6 → Select 2 Variables → Enter the data from problem (1) into FISH. Next, under the options tab, select "Use N for Calculations (population)". Last, have FISH perform all steps. (Calculate Means for X and Y through Calculate Y-Intercept). Have students record their findings into Table 1.

Teacher Note: Have students compare FISH's regression slope and y-intercept and their individual answers as well as the class' consensus answers. The answers should be reasonably the same, especially the consensus slope and y-intercept.

3. Have students click the reset button and under the options tab, select "Use N-1 for Calculations (sample)". Now enter the data again into FISH, perform steps 2-10 as the class did in problem 2.

Teacher Note: Have students record their answers from FISH in Table 2. Next, have students compare the answers from Table 1 and 2. Ask them if there are any differences between the two tables and why they are different. (Having students compare the two tables and noticing a difference in the variance, standard deviation, and covariance values might be a good review of the formulas and the use of N versus N-1.) Does using N – 1 or N make a difference for the

correlation coefficient, coefficient of determination, slope, and/or y-intercept? Should it make a difference? Why or why not? Which is appropriate for this problem and why?

4. Have the students review the elementary teachers questions.
 - a. Ask students what (and why) will be used to determine the relationship between the participatory habits of her students during the first week of school and the tenth week. Have students report what they are using and FISH's calculation of this item.

Teacher Note: *All depending upon the level of the students, it is recommended that at this point to determine if this value is significant at the $\alpha = .05$ level. Please consult a statistics text if needed.*

- b. Ask students what (and why) will be used to predict the elementary teacher's students' tenth week participatory habits. Have students report what they are using and FISH's calculation of this item.

Teacher Note: *All depending upon the level of the students, it is recommended that at this point to determine if this value is significant at the $\alpha = .05$ level. Students should be able to determine what statistic they will use to determine the significance of the regression equation. Please consult a statistics text if needed.*

- c. Ask students to view scatterplot (located in upper right corner of FISH software) and compare it to their scatterplot. Also, ask students to view some of the other features on this screen such as show regression line, show y-intercept, show mean lines, etc. Ask students what the relevance of each tab means.

Table 1

X	Y	X*Y	X-M	Y-M	(X-M)(Y-M)
3	7				
1	2				
1	4				
2	5				
2	6				
3	5				
Sum(X)=	Sum(Y)=				
Mean(X)=	Mean(Y)=	Sum=	Sum =	Sum =	Sum =
Std Dev(X)=	Std Dev(Y)=		Std Dev (X-M) =	Std Dev (Y-M) =	Covariance =
Var(X)=	Var(Y)=	Pearson r =	r² =	Slope =	Y-Intercept

Table 2

X	Y	X*Y	X-M	Y-M	(X-M)(Y-M)
3	7				
1	2				
1	4				
2	5				
2	6				
3	5				
Sum(X)=	Sum(Y)=				
Mean(X)=	Mean(Y)=	Sum=	Sum =	Sum =	Sum =
Std Dev(X)=	Std Dev(Y)=		Std Dev (X-M) =	Std Dev (Y-M) =	Covariance =
Var(X)=	Var(Y)=	Pearson r =	r² =	Slope =	Y-Intercept