EXPLORATION 5 Scatterplots & Correlation

The students in Mr. Santos's class were asked to record the number of hours spent studying for their mathematics test. For each student, Mr. Santos wrote an ordered pair (x, y). The value of x represented the number of hours of study and the value of y was the student's mark on the test. Marsha's ordered pair was (3, 82) because she spent 3 hours studying and received a final mark of 82 out of 100. The set of ordered pairs that Mr. Santos recorded are shown on the right.

Ordered Pairs

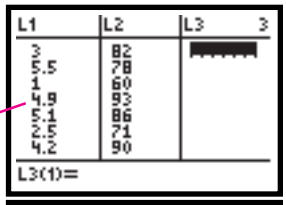
(3.0, 82)	(5.5, 78)	(1.0, 60)	(4.9, 93)	(5.1, 86)
(2.5, 71)	(4.2, 90)	(0.5, 40)	(3.5, 88)	(7.0, 96)
(1.5, 73)	(2.4, 82)	(2.0, 53)	(6.2, 87)	(8.4, 100)
(2.6, 75)	(3.7, 85)	(5.4, 70)	(9.3, 89)	(7.6, 87)
(1.4, 48)	(0.5, 56)	(6.5, 85)	(2.3, 61)	(5.2, 74)
(1.0, 47)	(3.5, 87)	(8.2, 94)	(3.0, 86)	(5.4, 92)

WORKED EXAMPLE 1

Use the data given above to determine whether there is a relationship between the number of hours of study and the test mark.

SOLUTION

To enter these data, we access the STAT EDIT table by pressing: **STAT** **ENTER**
We clear the table and enter in L_1 the first coordinates of all the ordered pairs and in L_2, all the second coordinates, taking care to check that opposite each value in L_1 is the corresponding second coordinate in L_2.



Row 4 corresponds to the ordered pair (4.9, 93)

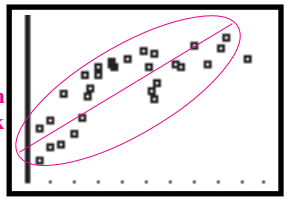
The display shows the first 7 rows of the table.

To determine whether there is a relationship between the number of hours of study and the test mark we plot the 30 data points corresponding to the 30 ordered pairs. To do this, we first turn off all statistical plots by entering:

2nd [STAT PLOT] **4** **ENTER** followed by **2nd** [STAT PLOT] **ENTER**

We define Plot 1 by selecting the options as shown in the display. That is, we choose the first icon in row "Type", which represents a type of plot called a *scatter plot*. We enter L_1 as the Xlist and L_2 as the Ylist.

To graph this plot, we press: **ZOOM** **9** to get the display shown here.



The display shows that almost all the data points in the scatter plot can be contained in an oval which is inclined along an axis with positive slope. In such a case we say that the variables plotted on both axes are *positively correlated.* The scatterplot shows that there is a positive correlation between preparation time for a test and the test mark. When the data points cluster close to the axis of the oval, we say the variables have a *strong positive correlation.* This scatterplot shows a fairly strong positive correlation.

IS A CAR AN INVESTMENT OR AN EXPENDITURE?

When an item that you purchase tends to increase in value, we usually refer to it as an *investment*. Conversely, an item that decreases in value, is called an *expenditure*. Understanding the difference between an investment and an expenditure is an important key to managing your personal finances. To determine whether a car is an expenditure or an investment, we must find out whether its value increases or decreases with its age.

The table displayed here shows for a popular model car, the resale value in dollars for each age between 1 and 10 years.

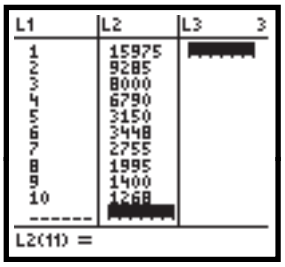
Age	Value	Age	Value
1	$15,975	6	$3,448
2	$ 9,285	7	$2,755
3	$ 8,000	8	$1,995
4	$ 6,790	9	$1,400
5	$ 3,150	10	$1,268

WORKED EXAMPLE 2

Use the data given in the table to determine whether there is a correlation between the age of a car and its resale value.

SOLUTION

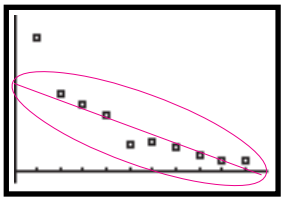
By scanning the table, we see that the resale value of the car tends to decrease with increasing age, so there seems to be a relationship, i.e. a *correlation*, between these two variables. To determine whether it is a strong or a weak correlation, we proceed as in *Worked Example* 1.



To enter these data, we access the STAT EDIT table by pressing: **STAT** **ENTER**
We clear the table and enter in L_1 the ages from 1 to 10 years and in L_2 all the corresponding resale values.

To determine whether there is a relationship between the age of that particular model car and the resale value, we plot the 10 data points as 10 ordered pairs. To do this, we first turn off all statistical plots by entering:

2nd [STAT PLOT] **4** **ENTER** followed by **2nd** [STAT PLOT] **ENTER**



We define Plot 1 choosing the same options as in *Worked Example* 1, and press:
ZOOM **9** to get the display shown here.

The display shows that almost all the data points in the scatter plot can be contained in an oval that is inclined along an axis with negative slope. In such a case we say that the variables plotted on both axes are *negatively correlated.* The scatterplot shows that there is a negative correlation between the age of a car and its resale value. This scatterplot shows a very strong negative correlation.

Note: It is important to understand that if two variables are strongly correlated, it is not necessarily true that a change in one variable "causes" a change in the other. Often changes in both variables result from a common "cause". For example, shoe size and reading ability of school children are positively correlated because increases in both these variables come with increased age. However, having large feet does not "cause" increased reading ability.

Exercises

1. Write in your own words the meanings of the following phrases:

- positive correlation
- negative correlation
- no correlation
- strong positive correlation
- strong negative correlation

2. Use the phrases above to identify the kind of correlation which is characterized by each of the following scatterplots.

3. The table below shows 30 ordered pairs. The first coordinate in each ordered pair is the age of a randomly selected married woman. The second component is the age of her spouse. Use the procedure in the worked examples to create a scatterplot that will reveal whether there is a correlation between the age of a woman and her spouse.

Ages of Women and their Husbands

(29, 34)	(37, 38)	(19, 20)	(57, 57)	(34, 32)
(23, 25)	(51, 54)	(72, 81)	(29, 23)	(70, 70)
(45, 54)	(39, 37)	(58, 56)	(64, 71)	(35, 35)
(42, 50)	(25, 24)	(37, 48)	(36, 36)	(27, 32)
(56, 42)	(17, 24)	(28, 28)	(57, 26)	(41, 39)
(47, 47)	(16, 18)	(24, 27)	(84, 87)	(55, 59)

If you discover a correlation, indicate whether it is positive or negative and describe it as strong or weak.

4. A biologist who was studying crickets hypothesized that the number of chirps per minute made by a cricket is strongly correlated to the outside temperature increases. He recorded the data shown in the table. Create a scatter plot to test his hypothesis.

Temperature in ° C	Chirps per min
17	105
18	110
19	110
20	126
21	126
22	130
23	130
24	152
24	156
25	160
26	170
27	171
28	175
29	196
30	212

5. For each of the following pairs of variables, indicate what kind of correlation (if any) you would expect to exist. Explain why.

a) height and weight
b) age and personal savings
c) level of education and income
d) latitude of a U. S. city and its mean temperature in winter
e) A golfer's age and golf score
f) travel time and speed of travel for a particular distance
g) I. Q. and number of olives eaten per month

Investigations

6. In order to quantify the "degree" of correlation between two variables, Karl Pearson, one of the pioneers of statistics, defined the *correlation coefficient*, *r,* between two variables x and y by the equation:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where y_i is the value of y corresponding to $x = x_i$. $i = 1 \ldots n$

a) Suppose that x and y are so strongly correlated that y is a linear function of x; i.e. $y = ax + b$ for some constants a, b. Prove algebraically that $|r| = 1$. How are x and y related when $r = 1$? How are they related when $r = -1$?

b) What range of values of r would indicate:
a strong correlation? a weak correlation? no correlation?

c) Enter into your STAT EDIT table the temperatures and corresponding chirps per minute given in Exercise 4.
To find the correlation coefficient r for cricket chirps per minute and temperature, select the command **DiagnosticOn** by pressing 2nd CATALOG and moving the cursor through the catalog menu to this command. Then press ENTER twice.
To obtain both r and r^2 press:

STAT ▶ 4 ENTER

Record the values of r and r^2 shown on the display.

7. a) Calculate the correlation coefficient between the ages of a woman and her spouse (See Exercises 3 and 6.)
b) Calculate the correlation coefficient between hours of study and test mark presented in *Worked Example* 1.

Drawing Inferences

A common error is the assumption that if two variables have a high correlation, then one of the variables is the "cause" of the other. This recent newspaper article revealed an interesting trend that resulted from such a false inference. Give a possible reason for the correlation.

Latin's Remarkable Resurrection

(New York)- A real resurrection of Latin is taking place. The upsurge is at the high school level and is very much a result of student demand.… Why are students flocking to Latin, and what do they expect to get out of it?
…There is, allegedly, a clear correlation between taking Latin at school and doing well on scholastic aptitude exams, and the students are well aware, as they always are, of the statistics. They may not see why Latin, in the high-tech age should help them reach their goal, but they are clear on the fact that it does, and plausibly invoking the national adage, "If it works, don't fix it," queue up and sign on.