

10.1 Paired Data and Scatter Diagrams

Linear Equations

Linear equations (or linear functions) graph as straight lines and can be written in the form:

$$y = bx + a$$

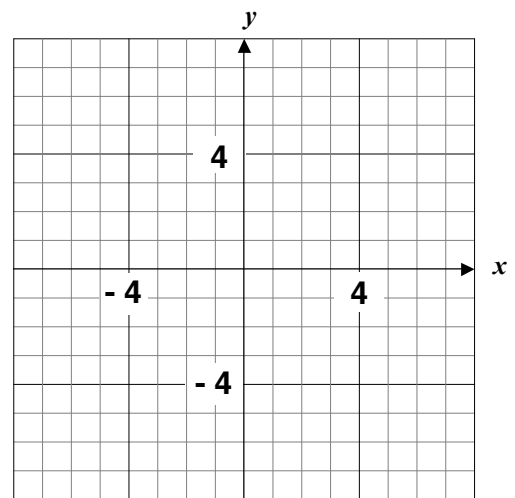
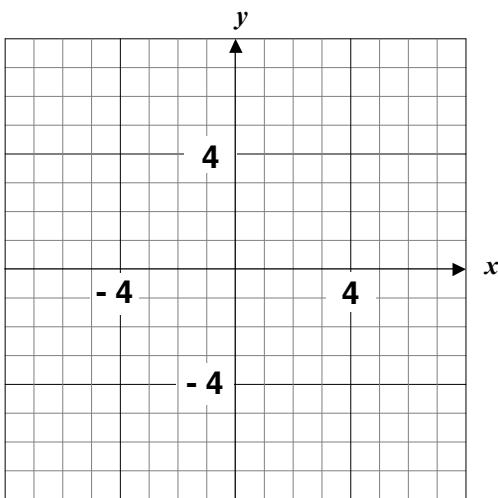
where $(0, a)$ is the y -intercept, and

$$b = \text{slope of the line} = \frac{\text{rise}}{\text{run}} = \frac{y_2 - y_1}{x_2 - x_1}$$

Example A

- Identify the slope and y -intercept in each of the equations.
- Graph each equation using the y -intercept and the slope.

Equation	Slope	y -intercept
$y = bx + a$	b	$(0, a)$
$y = 2x - 5$		
$y = -\frac{2}{3}x + 3$		
$y = 5$		
$y = -x$		



- Graph each equation using the TI-83.

Scatter Diagram; Explanatory & Response Variables

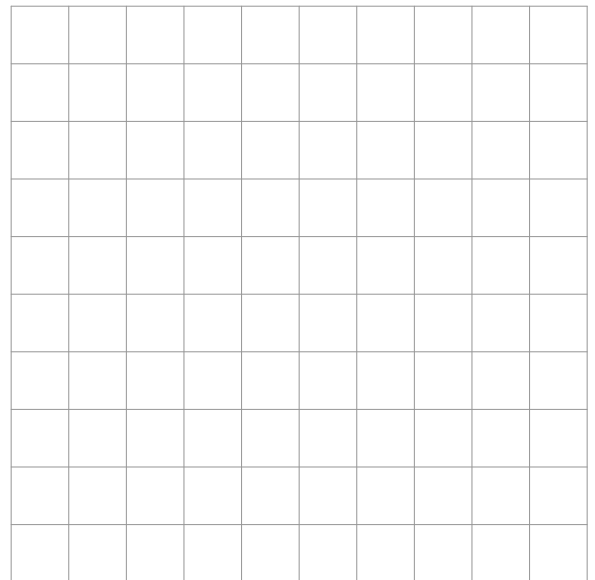
A **scatter diagram** is a plot of ordered-pair (x, y) data. We call x the **explanatory variable** and y the **response variable**.

Example 1

Phosphorous, a chemical in many household and industrial cleaning compounds, often finds its way into surface water. A random sample of eight sites in California wetlands gave the following information about phosphorous reduction in drainage water. In this study, x represents phosphorous concentration (in 100 mg/l) at the inlet of a bio-treatment facility and y represents the phosphorous concentration at the outlet of the facility.

x	5.2	7.3	6.7	5.9	6.1	8.3	5.5	7.0
y	3.3	5.9	4.8	4.5	4.0	7.1	3.6	6.1

- Make a scatter diagram of the data. Label and scale the axes. Then draw a “best fit” linear model through the data.
- Do x and y appear to be linearly related?
- Use the linear model (line) to predict the outlet concentration of phosphorous if the inlet concentration is 700 mg/l.
- Use the linear model to predict the inlet concentration of phosphorous if the outlet concentration is 200 mg/l.



Linear Correlation

If the scatter plot of ordered pair data, represented by variables x and y , trends roughly into a straight line, then we say that x and y are **linearly correlated**.

Linear correlation is classified in two general ways:

1. **Degree:** none, low-moderate, high, perfect

Perfect linear correlation means that all (x, y) ordered pairs of data lie on the same straight line.

2. **Sign or slope:** positive or negative

- i. Positive linear correlation means that high values of x correlate with high values of y , and low values of x correlate with low values of y . The graph has a positive slope (and trends upward from left to right).

- ii. Negative linear correlation means that high values of x correlate with low values of y , and low values of x correlate with high values of y . The graph has a negative slope (and trends downward from left to right).

See figures 10-1, 10-2, 10-4 Guided Exercise 2, Table 2, and Exercises 1 & 2

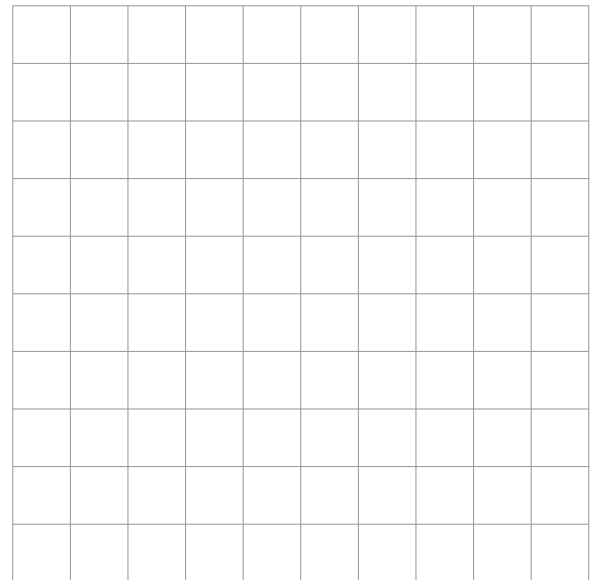
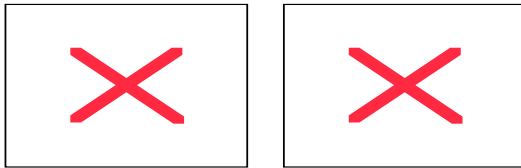
Guided Exercise 1

An industrial plant has 7 divisions that do the same type of work. A safety inspector tracks x = “the number of work-hours devoted to safety training” and y = “the number of work-hours lost due to accidents.” The results are shown.

Division	x	y
1	10.0	80
2	19.5	65
3	30.0	68
4	45.0	55
5	50.0	35
6	65.0	10
7	80.0	12

(a) Make a scatter diagram for the data.

(b) Make a scatter diagram on your calculator. Enter the x -values into L_1 and the y -values into L_2 . Turn the STAT PLOT on and adjust the window settings to



(c) Draw a “best fit” line through the data and classify the linear correlation as (i) none, low-moderate, high, or perfect, and (ii) positive or negative.

(d) Use your linear model (i.e. read from the line) to predict the number of safety training hours needed so that 20 work-hours are lost due to accidents.

(e) Use your linear model (i.e. read from the line) to predict the number of work-hours are lost due to accidents when 30 hours are spent on safety training.

Sample Correlation Coefficient r

The **correlation coefficient** r is a unit-less numerical measure that assesses the strength of linear relationship between two variables x and y . See table 10-2.

1. $-1 \leq r \leq 1$
2. If $r = 1$, there is perfect positive linear correlation.
3. If $r = -1$, there is perfect negative linear correlation.
4. The closer r is to 1 or -1 , the better a line describes the relationship between x and y .
5. If r is positive, then as x increases, y increases.
6. If r is negative, then as x increases, y decreases.
7. The value of r is the same regardless of which variable is the explanatory and which is the response variable. Data plotted as (x, y) and (y, x) will have the same value for r .

Computation of r

1. Turn `DiagnosticOn` in the CATALOG menu.
2. Enter the x -values into L_1 and the y -values into L_2 .
3. `STAT / CALC / 4: LinReg(ax+b) Lx, Ly, Y1`
4. Highlight `Calculate` and press `ENTER`. Then scroll down to find the value of r .

Guided Exercise 1

- (f) Find the sample correlation coefficient for the data in guided exercise 1.

Sample versus Population Correlation Coefficient r

$r =$ **sample correlation coefficient** computed from a random sample of (x, y) data pairs.

$\rho =$ **population correlation coefficient** computed from all population data pairs (x, y) .

Lurking Variables

In ordered pairs (x, y) , x is called the **explanatory variable** and y is called the **response variable**. When r indicates a linear correlation between x and y , a change in the values of y tends to respond to changes in values of x according to a linear model. A **lurking variable** is a variable that is neither an explanatory nor a response variable. Yet, a lurking variable may be responsible for changes in both x and y . Correlation does not necessarily mean causation.

Example 3

It has been observed in a certain community that over the years the correlation between x , the number of people going to church, and y , the number of people in jail, was $r = 0.90$. Does going to church cause people to go to jail, or visa versa? Explain.

10.2 Linear Regression and the Coefficient of Determination

Least-Squares Linear Regression Line

The **least-squares linear regression line** is the line that fits the (x, y) data points in such a manner that the sum of the squares of all the vertical distances from the data points to the line is as small as possible. The point (\bar{x}, \bar{y}) is always on the least-squares regression line.

Computing the Linear Regression Line on the TI-83/84

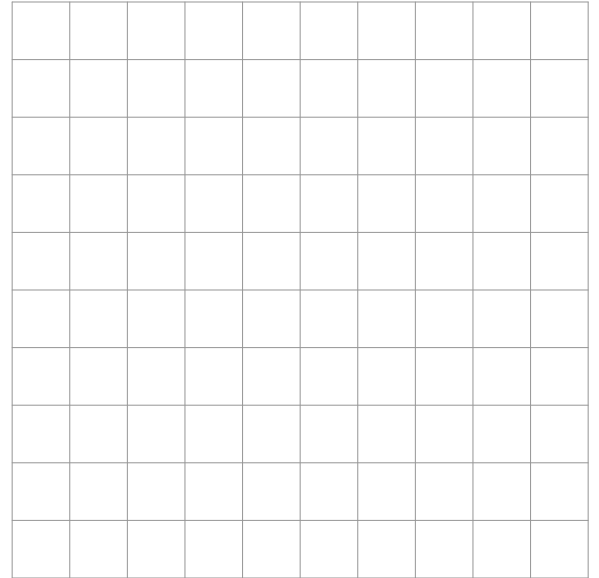
STAT / CALC / 4: LinReg(ax+b) L_x, L_y, Y₁

Output: a and b in $y = bx + a$

Example 4

In Denali national Park, Alaska, the wolf population is dependent on the caribou population. Let x represent the caribou population (in hundreds) and y represent the wolf population. A random sample in recent gave the following information.

x	30	34	27	25	17	23	20
y	66	79	70	60	48	55	60



- Identify the explanatory and response variables.
- Make a scatter diagram of the data.
- Find the linear regression line for the data. Graph the LSRL – write at least 2 points on the line.
- Interpret the slope of the line in this application.
- Predict the size of the wolf population when the caribou population is 2100. Is this interpolation or extrapolation?
- Predict the size of the wolf population when the caribou population is 4000. Is this interpolation or extrapolation?

Coefficient of Determination r^2

If r is the correlation coefficient, then r^2 is called the **coefficient of determination** and

$$r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}}.$$

1. The value of r^2 is the ratio of explained variation over total variation. That is, r^2 is the fractional amount of the total variation in y that can be explained by using the linear model $y = bx + a$ with x as the explanatory variable.
2. $1 - r^2$ is the fractional amount of the total variation in y that is due to random chance or to the possibility of lurking variables that influence y .

Example 4A

- (a) Find r and r^2 for example 4.
- (b) Explain the value of r^2 in this application.
- (c) Explain the value of $1 - r^2$ in this application.

Change on Directions for 10.2 Exercises

Do the following in problems 7-18:

- (a) View the scatter diagram of the data on your calculator to verify a linear model is appropriate.
- (b) Find a and b for the least-squares regression line $y = bx + a$. Then find r , r^2 , \bar{x} and \bar{y} .
- (c) Graph the regression line from part (b). Be sure the point (\bar{x}, \bar{y}) is on the graph.
- (d) Interpret the values of r^2 in one sentence relevant to the application. Interpret the values of $1 - r^2$ in one sentence relevant to the application.

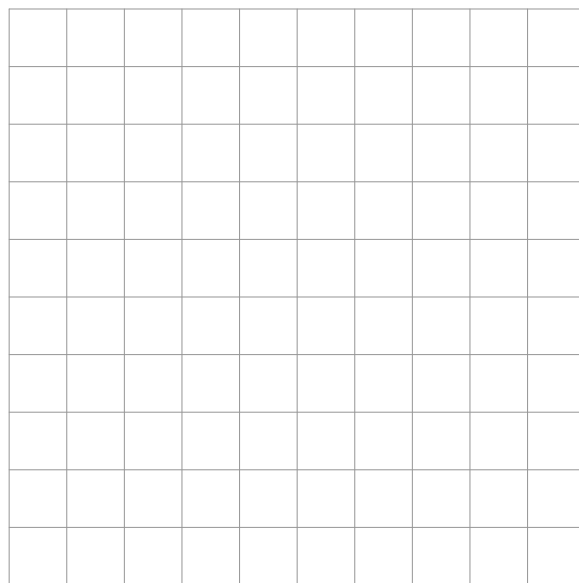
Guided Exercise 3

Quick Sell car dealership runs 1-minute TV advertisements and tracks x = “the number of ads that week.” and y = “the number of cars sold that week.”

x	6	20	0	14	25	16	28	18	10	8
y	15	31	10	16	28	20	40	25	12	15

Complete steps (a) – (d). Then find the predicted number of cars sold per week if the budget only allows 12 ads to be run per week.

- (a) View the scatter diagram of the data on your calculator to verify a linear model is appropriate.
- (b) Find a and b for the least-squares regression line $y = bx + a$. Then find r , r^2 , \bar{x} and \bar{y} .



- (c) Graph the regression line from part (b). Be sure the point (\bar{x}, \bar{y}) is on the graph.
- (d) Interpret the values of r^2 in one sentence relevant to the application. Interpret the values of $1 - r^2$ in one sentence relevant to the application.

10.3 Testing the Correlation Coefficient

The **population correlation coefficient** ρ (rho, read “row”) is estimated by the statistic r .

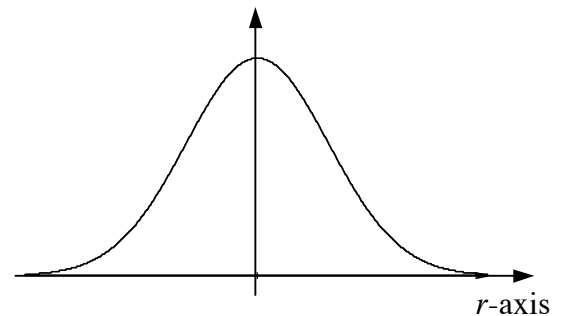
If we assume the variables x and y are normally distributed and want to test if they are correlated in the population, then we set the null hypothesis to say they are not correlated:

$H_0: \rho = 0$ x and y are not correlated at the given level of significance.

Theorem

Let random variables x and y be normally distributed. If $\rho = 0$ (as assumed in the null hypothesis), then the distribution of sample correlation coefficients (the r values) is normally distributed about $r = 0$.

Distribution of r – values when $\rho = 0$



Example 6

Do college graduates have an improved chance of a better income?

Let x = percentage of the population 25 or older with at least 4 years of college and y = percentage growth in per capita income over the past seven years. A random sample of six communities in Ohio gave the information in the table.

x	9.9	11.4	8.1	14.7	8.5	12.6
y	37.1	43	33.4	47.1	26.5	40.2

- (a) Find the correlation coefficient r .
- (b) Test to see if the correlation coefficient is positive at a 1% level of significance.
- (c) Summarize your conclusions in one sentence relevant to this application.

10.3 Homework

Do Exercises 7-12 parts (b) and (d) only.

Exercise 10.3 #3

What is the optimal time for a scuba diver to be at the bottom of the ocean? The navy defines optimal time to be the time at each depth for the best balance between length of work period and decompression time after surfacing. Let x = depth of a dive in meters, and y = optimal time in hours. A random sample of divers gave the following data.

x	14.1	24.3	30.2	38.3	51.3	20.5	22.7
y	2.58	2.08	1.58	1.03	0.75	2.38	2.20

- (b) Use a 1% level of significance to test the claim that $\rho < 0$.
- (d) Find the predicted optimal time for a dive depth of 18 meters.