4.1: Scatter Diagrams and Correlation

Graphing with Two Variables

The **response variable** is the variable whose value can be explained by the value of the **explanatory** or **predictor** variable.

A **scatter diagram** is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The explanatory variable is plotted on the horizontal axis, and the response variable is plotted on the vertical axis.

Example

The data shown to the right are based on a study for drilling rock. The researchers wanted to determine whether the time it takes to dry drill a distance of 5 feet in rock increases with the depth at which the drilling begins. So, depth at which drilling begins is the explanatory variable, *x*, and time (in minutes) to drill five feet is the response variable, *y*.

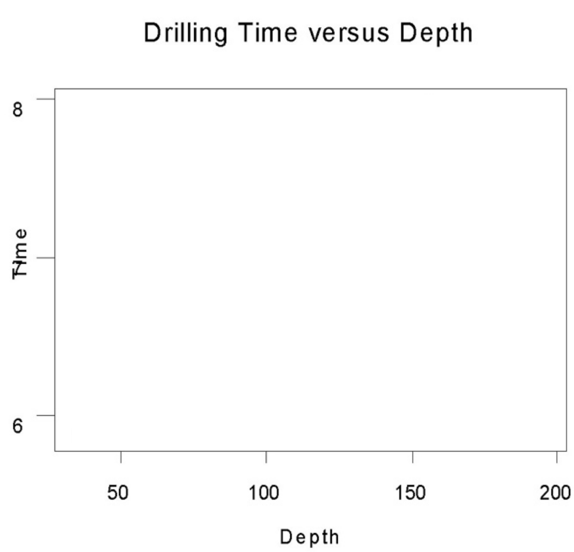
Source: Penner, R., and Watts, D.G. "Mining Information." *The American Statistician,* Vol. 45, No. 1, Feb. 1991, p. 6.

Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

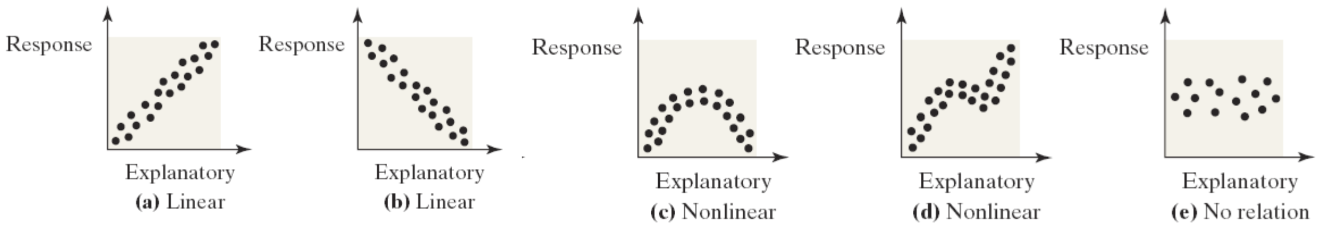
Draw a scatter diagram of the data using your TI-84

Key Steps:

- [STAT] → [Edit]
- Enter data into L1 and L2
- [2nd]→[StatPlot]
- Choose Scatterplot
- [Zoom]→[ZoomStat]
-



Types of Relations in Scatter Diagrams



Positive Association:
whenever the value of one variable increases, the value of the other variable also increases

Negative Association: whenever the value of one variable increases, the value of the other variable decreases.

Properties of the Linear Correlation Coefficient

The **linear correlation coefficient** - a measure of the strength and direction of the linear relation between two quantitative variables. The Greek letter ρ (rho) represents the population correlation coefficient, and r represents the sample correlation coefficient.

Try It:

Determine the linear correlation coefficient of the drilling data using your TI-84.

Key Steps:

- [STAT] → [Edit]
- Enter data into L1 and L2
- [STAT]→[CALC]
- Choose LinReg{ax+b}
- Or Choose 2-Var Stats

Depth at Which Drilling Begins, x (in feet)	Time to Drill 5 Feet, y (in minutes)
35	5.88
50	5.99
75	6.74
95	6.1
120	7.47
130	6.93
145	6.42
155	7.97
160	7.92
175	7.62
185	6.89
190	7.9

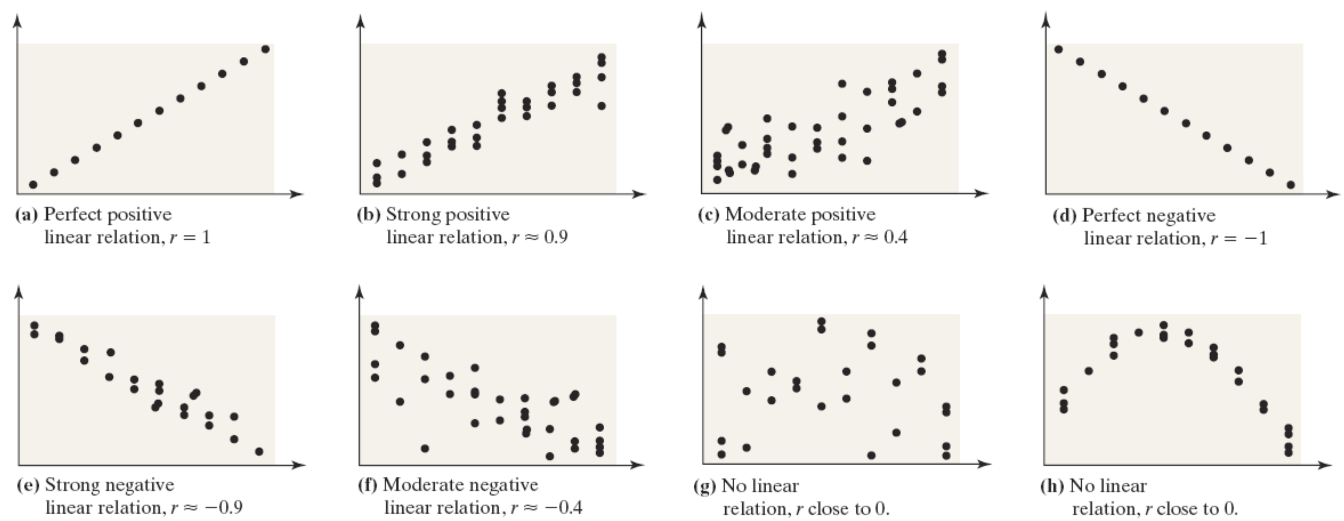
$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n-1}$$

$$= \frac{8.501037}{12-1}$$

$$= 0.773$$

Properties of the Linear Correlation Coefficient

1. The linear correlation coefficient is always between –1 and 1, inclusive. That is, $-1 \le r \le 1$.
2. If $r = +1$, then there is a perfect positive linear relation
3. If $r = -1$, then there is a perfect negative linear relation
4. The closer r is to +1, the stronger is the evidence of positive association between the two variables.
5. The closer r is to –1, the stronger is the evidence of negative association between the two variables.
6. If r is close to 0, then little or no evidence exists of a linear relation between the two variables. So r close to 0 does not imply no relation, just no linear relation.
7. The linear correlation coefficient is a unitless measure of association. So the unit of measure for x and y plays no role in the interpretation of r.
8. The correlation coefficient is not resistant. Therefore, an observation that does not follow the overall pattern of the data could affect the value of the linear correlation coefficient.



(a) Perfect positive linear relation, $r = 1$
(b) Strong positive linear relation, $r \approx 0.9$
(c) Moderate positive linear relation, $r \approx 0.4$
(d) Perfect negative linear relation, $r = -1$
(e) Strong negative linear relation, $r \approx -0.9$
(f) Moderate negative linear relation, $r \approx -0.4$
(g) No linear relation, r close to 0.
(h) No linear relation, r close to 0.

Where it comes from: Sample Linear Correlation Coefficient

$$r = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)}{n-1}$$

$\bar{x}$ is the sample mean of the explanatory variable
s_x is the sample standard deviation of the explanatory variable
$\bar{y}$ is the sample mean of the response variable
s_y is the sample standard deviation of the response variable
n is the number of individuals in the sample