Chapter 10
Correlation and Regression

We deal with two variables, x and y.

Main goal: investigate how x and y are related, or *correlated*; how much they depend on each other.

Example

- x is the height of mother
- y is the height of daughter

Question: are the heights of daughters independent of the height of their mothers? Or is there a correlation between the heights of mothers and those of daughters? How strong is it?

Example:

This table includes a random sample of heights of mothers, fathers, and their daughters.

Heights of mothers and their daughters in this sample seem to be <u>strongly</u> *correlated*...

But heights of fathers and their daughters in this sample seem to be <u>weakly</u> correlated (if at all).

Height of Mother	Height of Father	Height of Daughter
63	64	58.6
67	65	64.7
64	67	65.3
60	72	61.0
65	72	65.4
67	72	67.4
59	67	60.9
60	71	63.1
58	66	60.0
72	75	71.1
63	69	62.2
67	70	67.2
62	69	63.4
69	62	68.4
63	66	62.2
64	76	64.7
63	69	59.6
64	68	61.0
60	66	64.0
65	68	65.4

Section 10-2

Correlation between two variables (x and y)

Definition

A correlation exists between two variables when the values of one are somehow associated with the values of the other in some way.

Key Concept

Linear correlation coefficient, *r*, is a numerical measure of the strength of the relationship between two variables representing quantitative data.

Using paired sample data (sometimes called bivariate data), we can find the value of *r*.

Then we use that value to conclude that *there is* (or *is not*) a linear correlation between the two variables.

Definition

The linear correlation coefficient r measures the strength of the linear relationship between the paired quantitative x- and y-values in a sample.

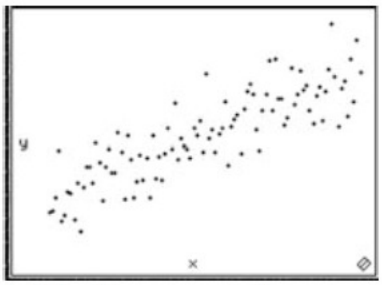
Also called Pearson correlation coefficient

Exploring the Data

We can often see a relationship between two variables by constructing a scatterplot.

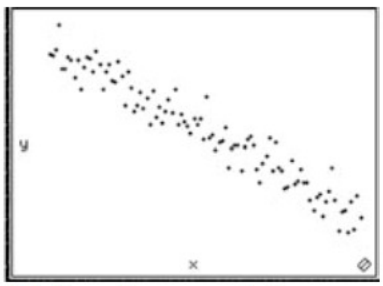
Scatterplots of Paired Data

ActivStats



(a) Positive correlation:
$r = 0.851$

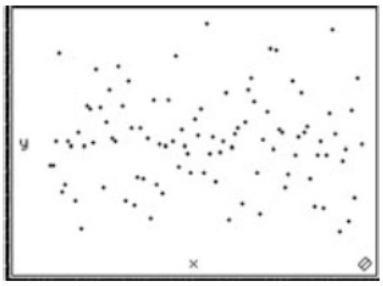
ActivStats



(b) Negative correlation:
$r = -0.965$

Scatterplots of Paired Data

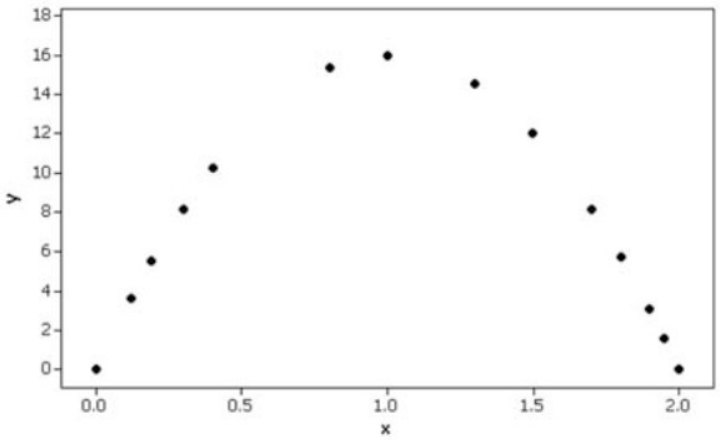
ActivStats



(c) No correlation: $r = 0$

Scatterplots of Paired Data

Minitab



(d) Nonlinear relationship: $r = -0.087$

Requirements

1. **The sample of paired (x, y) data is a random sample of quantitative data.**
2. **Visual examination of the scatterplot must confirm that the points approximate a straight-line pattern.**
3. **The outliers must be removed if they are known to be errors.**

Notation for the Linear Correlation Coefficient

n = number of pairs of sample data

Σ denotes the addition of the items indicated.

Σx denotes the sum of all x-values.

Σx^2 indicates that each x-value should be squared and then those squares added.

$(\Sigma x)^2$ indicates that the x-values should be added and then the total squared.

Notation for the Linear Correlation Coefficient

Σxy indicates that each x-value should be first multiplied by its corresponding y-value. After obtaining all such products, find their sum.

r = linear correlation coefficient for sample data.

ρ = linear correlation coefficient for population data.

Formula

The linear correlation coefficient r measures the strength of a linear relationship between the paired values in a sample.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{n(\Sigma x^2) - (\Sigma x)^2}\sqrt{n(\Sigma y^2) - (\Sigma y)^2}}$$

Interpreting r

Using Table A-6:

If the absolute value of the computed value of r, denoted $|r|$, exceeds the value in Table A-6, conclude that there is a linear correlation.

Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Note: In most cases we use the significance level $\alpha = 0.05$ (the middle column of Table A-6).

Caution

Know that the methods of this section apply only to a *linear* correlation.

If you conclude that there is no linear correlation, it is possible that there is some other association that is not linear.

Properties of the Linear Correlation Coefficient r

1. $-1 \le r \le 1$
2. **if all values of either variable are converted to a different scale, the value of r does not change.**
3. **The value of r is not affected by the choice of x and y. Interchange all x- and y-values and the value of r will not change.**
4. **r measures strength of a linear relationship.**
5. **r is very sensitive to outliers, they can dramatically affect its value.**