Warm-UP

- What is a z score?
- What is the z-score associated with the Standard normal probability of .8577?
- What is bivariate data?

Warm-UP

- What is a z score?

A *z* score gives the relationship between an observation (x_i) and the mean ($\bar{x}$) of a distribution in terms of some number of standard deviations. It is positive when it's greater than the mean, and negative when it's less than the mean.

- What is the z-score associated with the Standard normal probability of .8577?

z score = 1.07

Warm-UP

- What is bivariate data?

Bivariate data are data for two different variables. Usually the variables are related and often are taken from the same observations of a sample or a population.

Chapter 5: Summarizing Bivariate Data

Section 5.1

Scatterplots and Correlation

Statistics & Data Analysis, 5th edition – For AP*

PECK, OLSEN, DEVORE

Chapter 5
Describing Relationships

- 3.4 (Rewind to **Scatterplots**)
- 5.1 Correlation
- 5.2 **Linear Regression** (Least-Squares Regression)

Section 3.4 & 5.1
Scatterplots and Correlation

Learning Targets

After this section, you should be able to…

- ✓ IDENTIFY explanatory and response variables
- ✓ CONSTRUCT scatterplots to display relationships
- ✓ INTERPRET scatterplots
- ✓ MEASURE linear association using correlation
- ✓ INTERPRET correlation

Explanatory and Response Variables

Most statistical studies examine data on more than one variable. In many of these settings, the two variables play different roles.

Definition:

A **response variable** (dependent variable) measures an outcome of a study.

An **explanatory variable (or predictor/factor** indep. var.**)** may help explain or influence changes in a response variable.

Note: In many studies, the goal is to show that changes in one or more explanatory variables actually *cause* changes in a response variable. However, many explanatory-response relationships don't involve *direct causation*.

Displaying Relationships: Scatterplots

The most useful graph for displaying the relationship between two quantitative variables is a **scatterplot**.

Definition:

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point on the graph.

How to Make a Scatterplot

1. Decide which variable should go on each axis.
   - *Remember, the eXplanatory variable goes on the X-axis!*
2. Label and scale your axes.
3. Plot individual data values.

Displaying Relationships: Scatterplots

Make a scatterplot of the relationship between body weight and pack weight.

Since Body weight is our eXplanatory variable, be sure to place it on the X-axis!

Body weight (lb)	120	187	109	103	131	165	158	116
Backpack weight (lb)	26	30	26	24	29	35	31	28

Interpreting Scatterplots

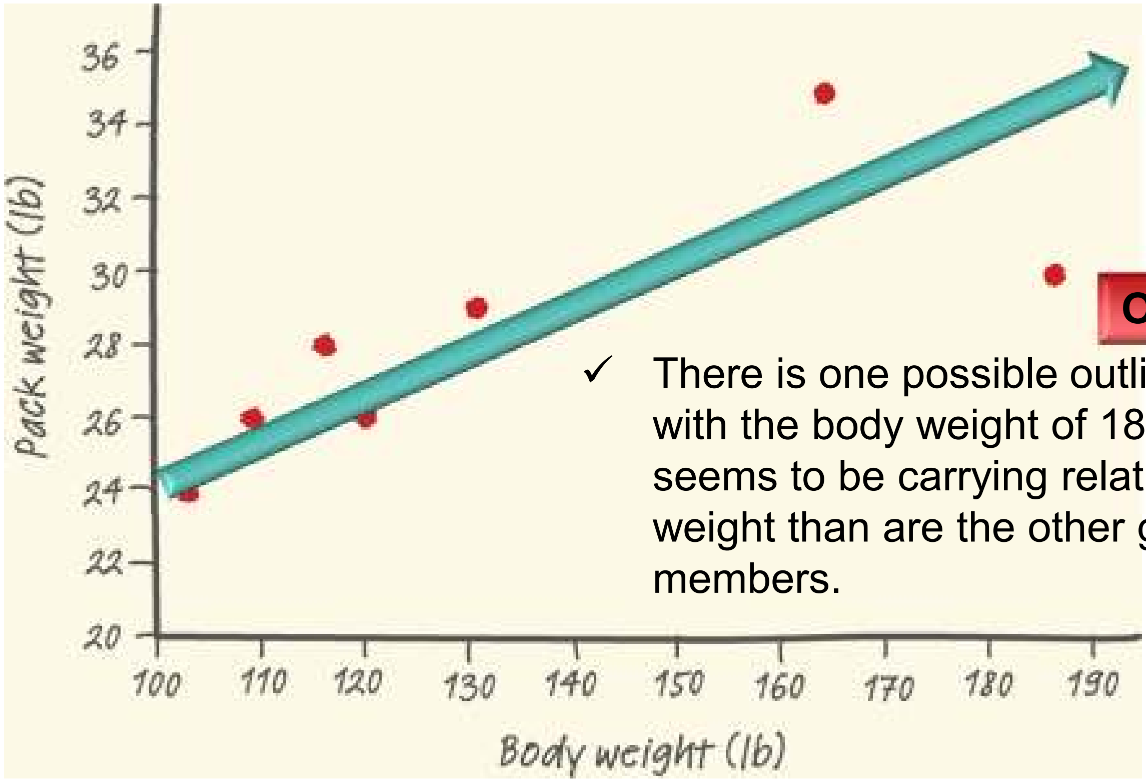
To interpret a scatterplot, follow the basic strategy of data analysis from previous chapters . Look for patterns and important departures from those patterns.

How to Examine a Scatterplot

As in any graph of data, look for the *overall pattern* and for striking *departures* from that pattern.

- You can describe the overall pattern of a scatterplot by the **direction**, **form**, and **strength** of the relationship.
- An important kind of departure is an **outlier**, an individual value that falls outside the overall pattern of the relationship.

Interpreting Scatterplots



Outlier

- ✓ There is one possible outlier, the hiker with the body weight of 187 pounds seems to be carrying relatively less weight than are the other group members.

Strength **Direction** **Form**

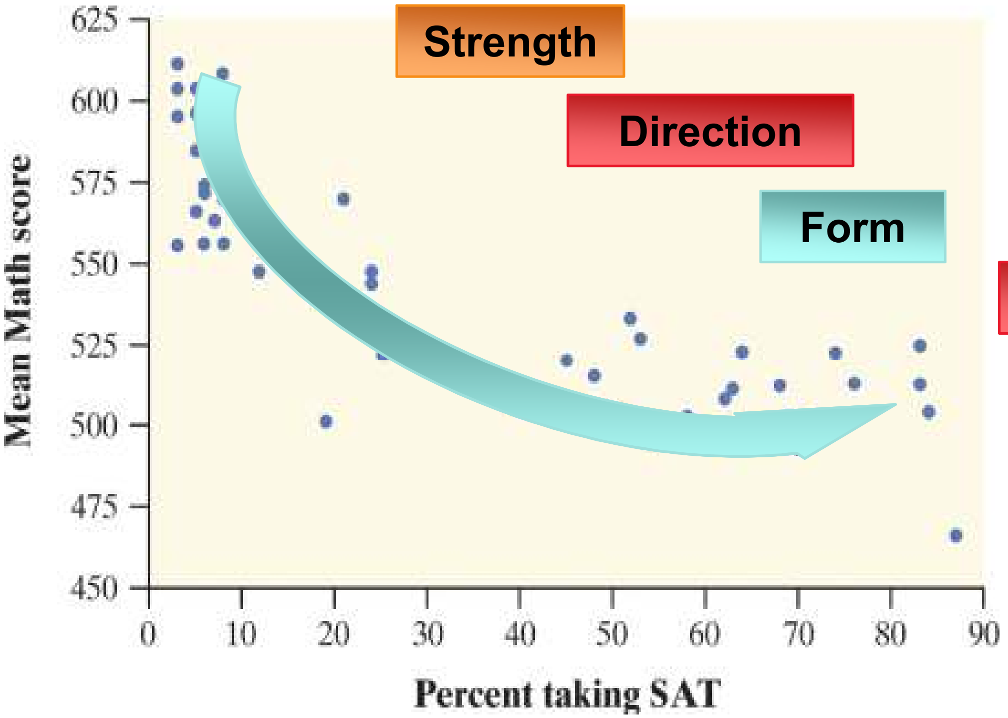
- ✓ There is a moderately strong, positive, linear relationship between body weight and pack weight.
- ✓ It appears that lighter students are carrying lighter backpacks.

Interpreting Scatterplots

Definition:

Two variables have a **positive association** when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.

Two variables have a **negative association** when above-average values of one tend to accompany below-average values of the other.



Strength

Direction

Form

Consider the SAT example from page 144. Interpret the scatterplot.

There is a moderately strong, negative, curved relationship between the percent of students in a state who take the SAT and the mean SAT math score.
Further, there are two distinct clusters of states and two possible outliers that fall outside the overall pattern.

Measuring Linear Association: Correlation

A scatterplot displays the strength, direction, and form of the relationship between two quantitative variables.

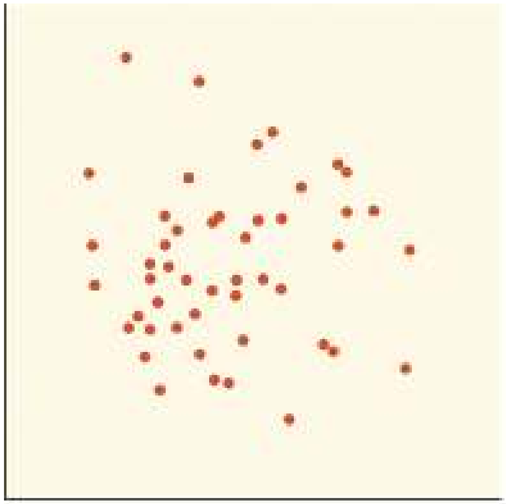
Linear relationships are important because a straight line is a simple pattern that is quite common. Unfortunately, our eyes are not good judges of how strong a linear relationship is.

Definition:

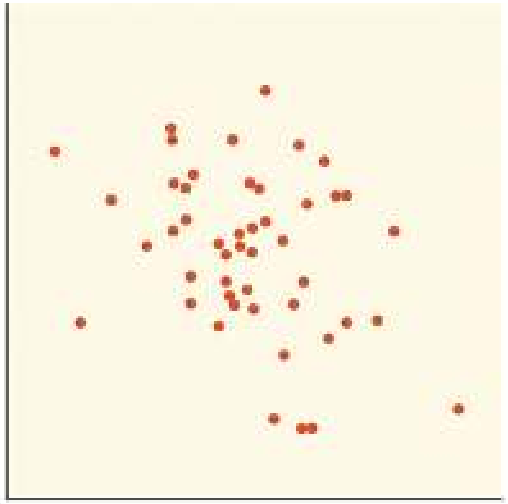
The **correlation *r*** measures the strength of the linear relationship between two quantitative variables.

- *r* is always a number between -1 and 1
- *r* > 0 indicates a positive association.
- *r* < 0 indicates a negative association.
- Values of *r* near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as *r* moves away from 0 towards -1 or 1.
- The extreme values *r* = -1 and r = 1 occur only in the case of a perfect linear relationship.

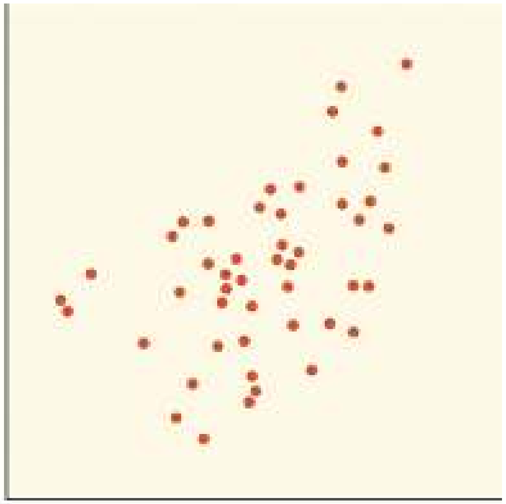
Measuring Linear Association: Correlation



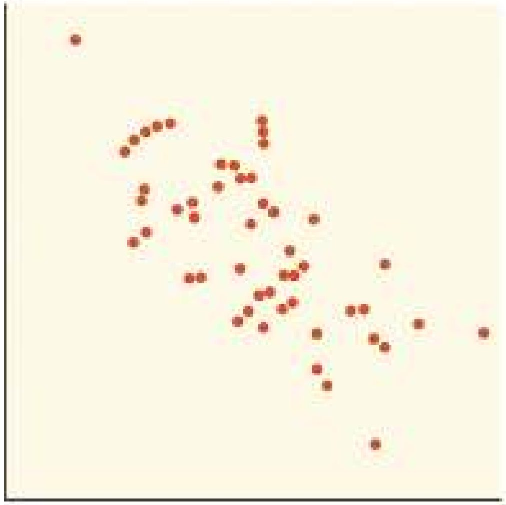
Correlation $r = 0$



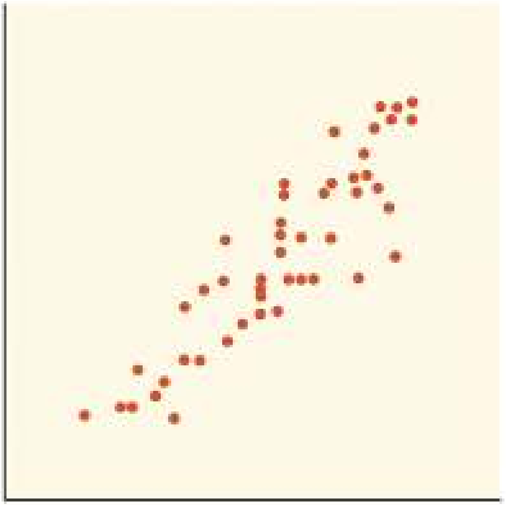
Correlation $r = -0.3$



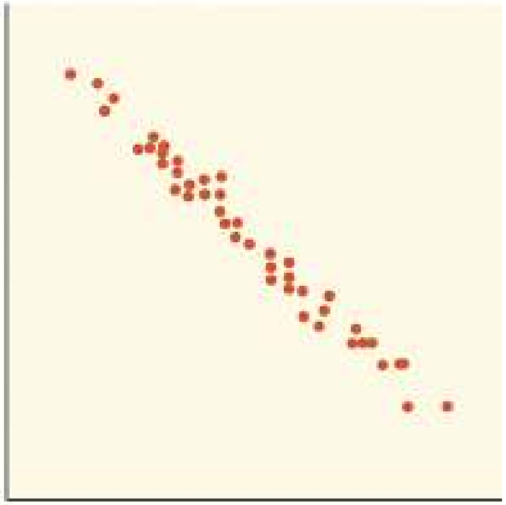
Correlation $r = 0.5$



Correlation $r = -0.7$



Correlation $r = 0.9$



Correlation $r = -0.99$

Correlation

The formula for *r* is a bit complex. It helps us to see what correlation is, but in practice, you should use your calculator or software to find *r*.

How to Calculate the Correlation *r*

Suppose that we have data on variables *x* and *y* for *n* individuals.

The values for the first individual are x_1 and y_1, the values for the second individual are x_2 and y_2, and so on.

The means and standard deviations of the two variables are *x-bar* and s_x for the *x*-values and *y-bar* and s_y for the *y*-values.

The correlation *r* between *x* and *y* is:

$$r = \frac{1}{n-1}\left[\left(\frac{x_1-\bar{x}}{s_x}\right)\left(\frac{y_1-\bar{y}}{s_y}\right)+\left(\frac{x_2-\bar{x}}{s_x}\right)\left(\frac{y_2-\bar{y}}{s_y}\right)+\ldots+\left(\frac{x_n-\bar{x}}{s_x}\right)\left(\frac{y_n-\bar{y}}{s_y}\right)\right]$$

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

Facts about Correlation

How correlation behaves is more important than the details of the formula. Here are some important facts about *r*.

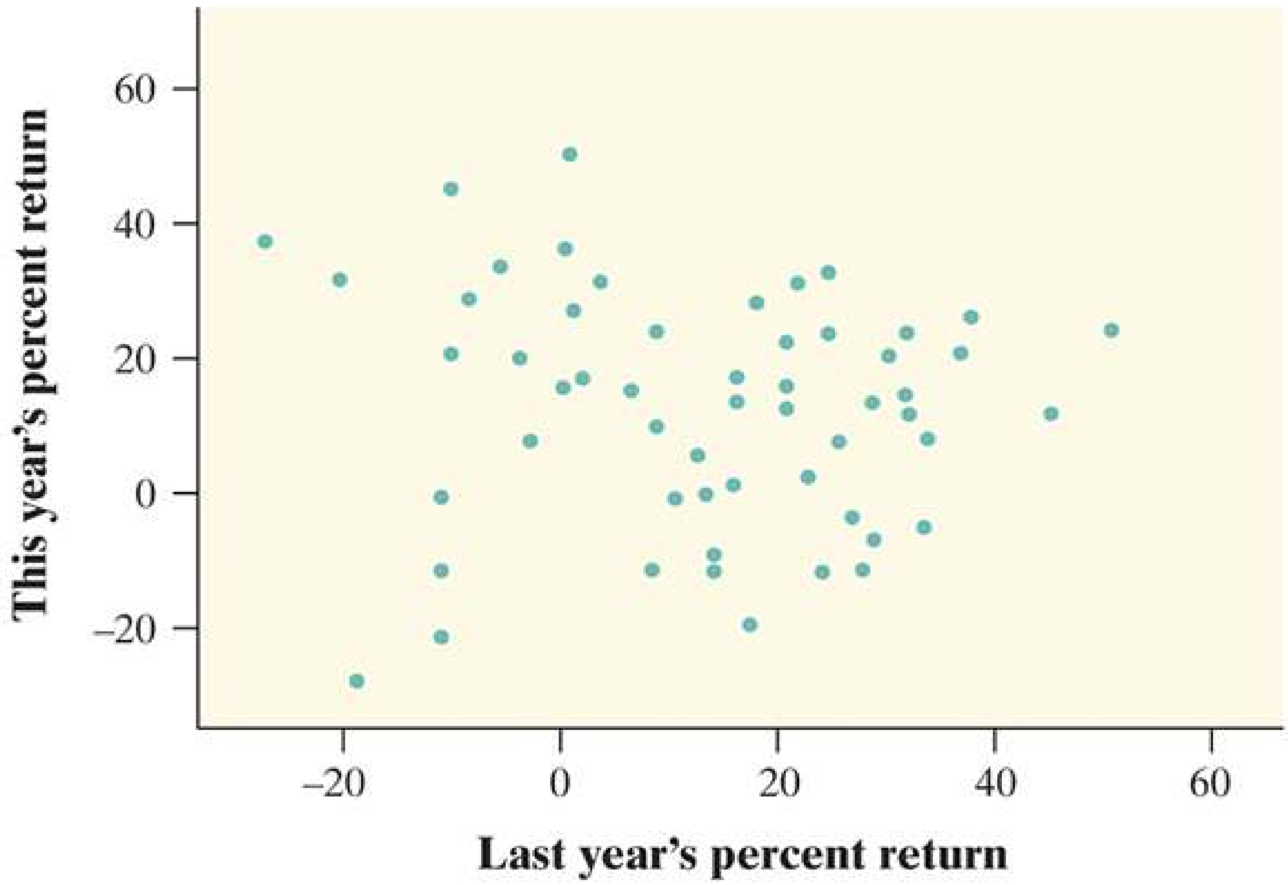
1. **Correlation makes no distinction between explanatory and response variables.**
2. ***r* does not change when we change the units of measurement of *x*, *y*, or both.**
3. **The correlation *r* itself has no unit of measurement.**

Cautions:

- Correlation requires that both variables be quantitative.
- Correlation does not describe curved relationships between variables, no matter how strong the relationship is.
- Correlation is not resistant. *r* is strongly affected by a few outlying observations (Beware of outliers!).
- Correlation is not a complete summary of two-variable data.

Correlation Practice

For each graph, estimate the correlation *r* and interpret it in context.



(d)

+

Practice Problem (p. 212) Problem 5.5

- Input the data for Quality rating in List 1 and Satisfaction rating in List 2
- Construct a scatter plot
- Compute and interpret the correlation coefficient

Section 5.1
Scatterplots and Correlation

Summary

In this section, we learned that…

- ✓ A **scatterplot** displays the relationship between two quantitative variables.
- ✓ An **explanatory variable** may help explain, predict, or cause changes in a **response variable.**
- ✓ When examining a scatterplot, look for an overall pattern showing the **direction**, **form**, and **strength** of the relationship and then look for **outliers** or other departures from the pattern.
- ✓ The **correlation** ***r*** measures the strength and direction of the linear relationship between two quantitative variables.

Looking Ahead…

In the next Section…

We'll learn how to describe linear relationships between two quantitative variables.

We'll learn

- ✓**Least-squares Regression line**
- ✓**Prediction**
- ✓**Residuals and residual plots**
- ✓**The Role of r^2 in Regression**
- ✓**Correlation and Regression Wisdom**