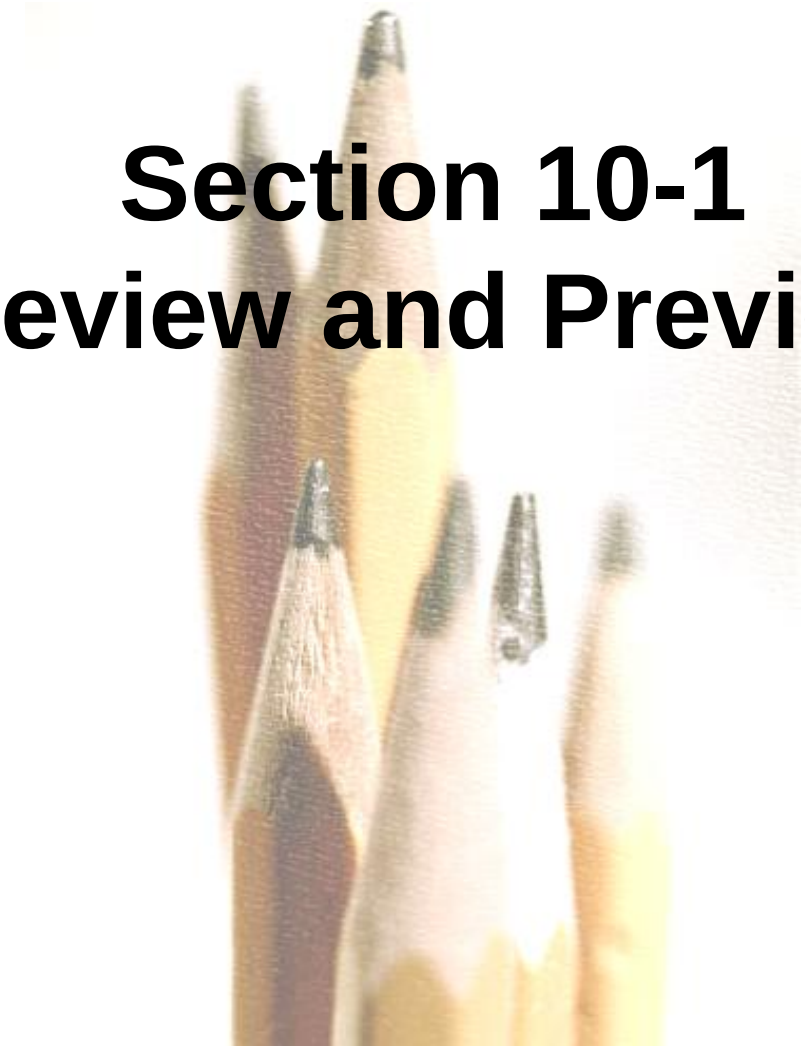
Chapter 10
Correlation and Regression

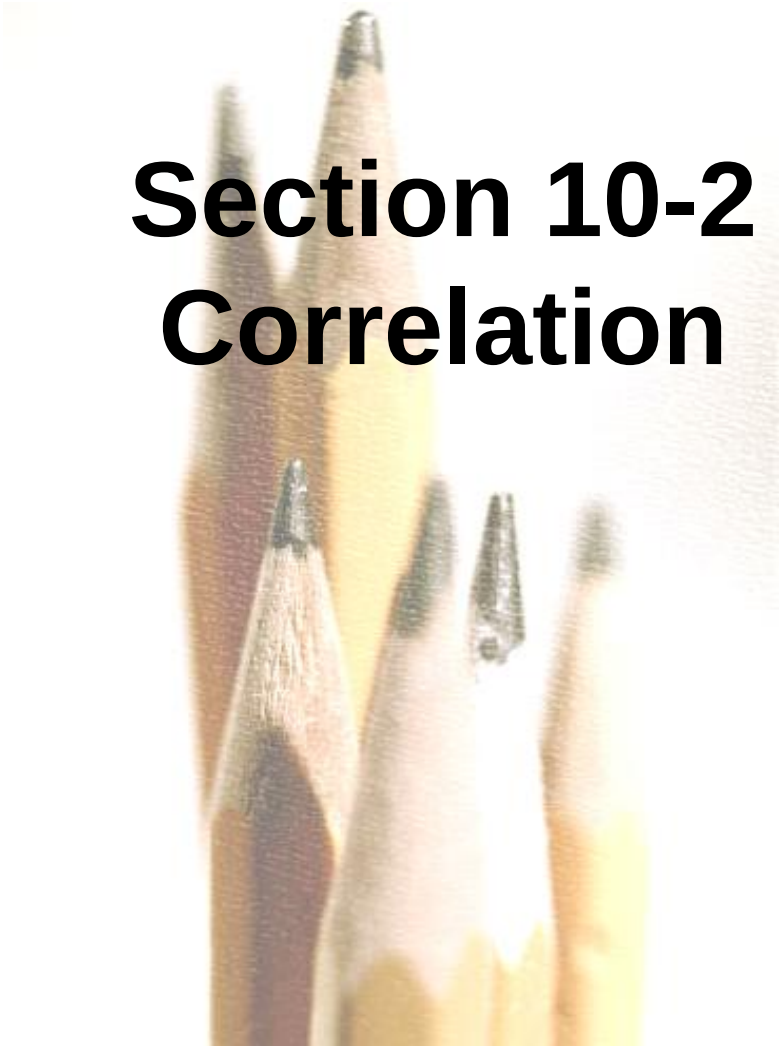
Section 10-1
Review and Preview



Preview

In this chapter we introduce methods for determining whether a correlation, or association, between two variables exists and whether the correlation is linear. For linear correlations, we can identify an equation that best fits the data and we can use that equation to predict the value of one variable given the value of the other variable. In this chapter, we also present methods for analyzing differences between predicted values and actual values.

Section 10-2
Correlation



Key Concept

In part 1 of this section introduces the linear correlation coefficient *r*, which 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 find the value of r (usually using technology), then we use that value to conclude that there is (or is not) a linear correlation between the two variables.

Part 1: Basic Concepts of Correlation

Definition

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

Definition

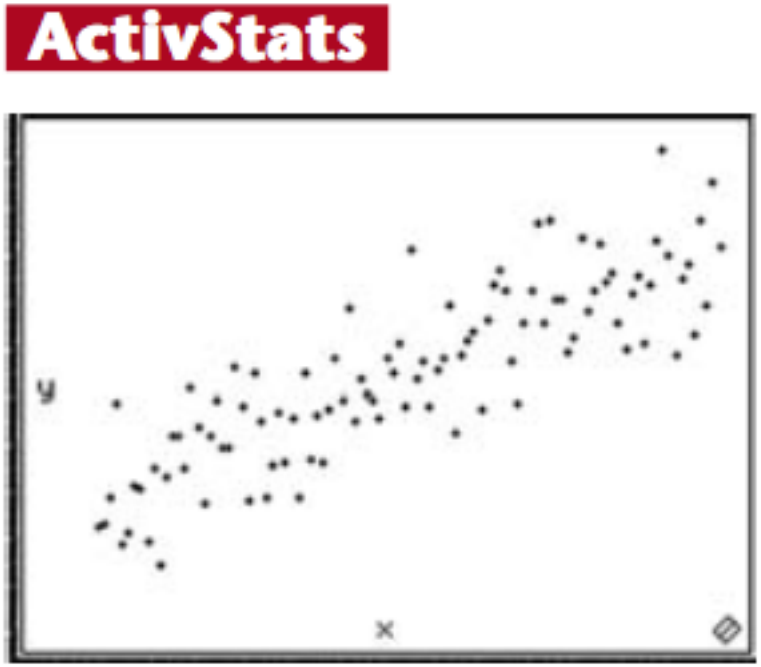
The linear correlation coefficient r measures the strength of the linear relationship between the paired quantitative x- and y-values in a sample.

Exploring the Data

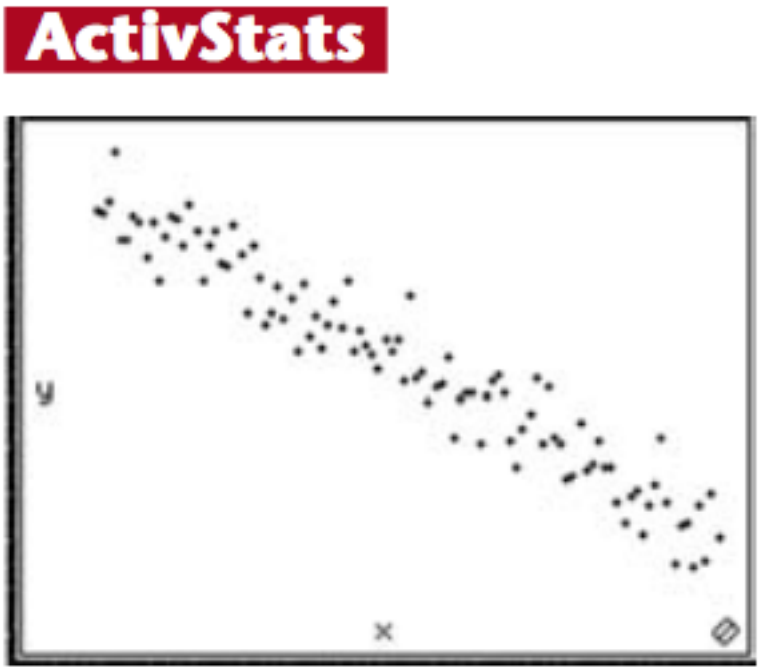
We can often see a relationship between two variables by constructing a scatterplot.

Figure 10-2 following shows scatterplots with different characteristics.

Scatterplots of Paired Data



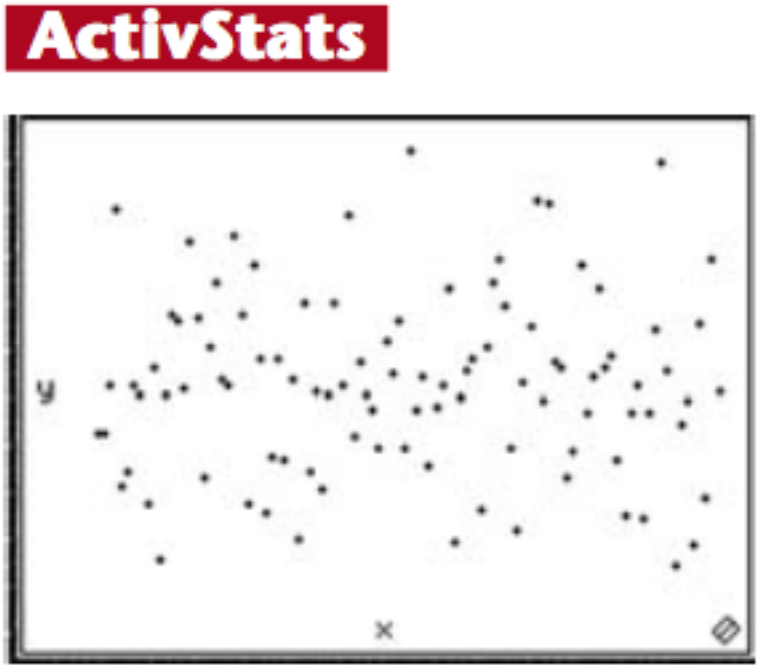
(a) Positive correlation:
$r = 0.851$



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

Figure 10-2

Scatterplots of Paired Data

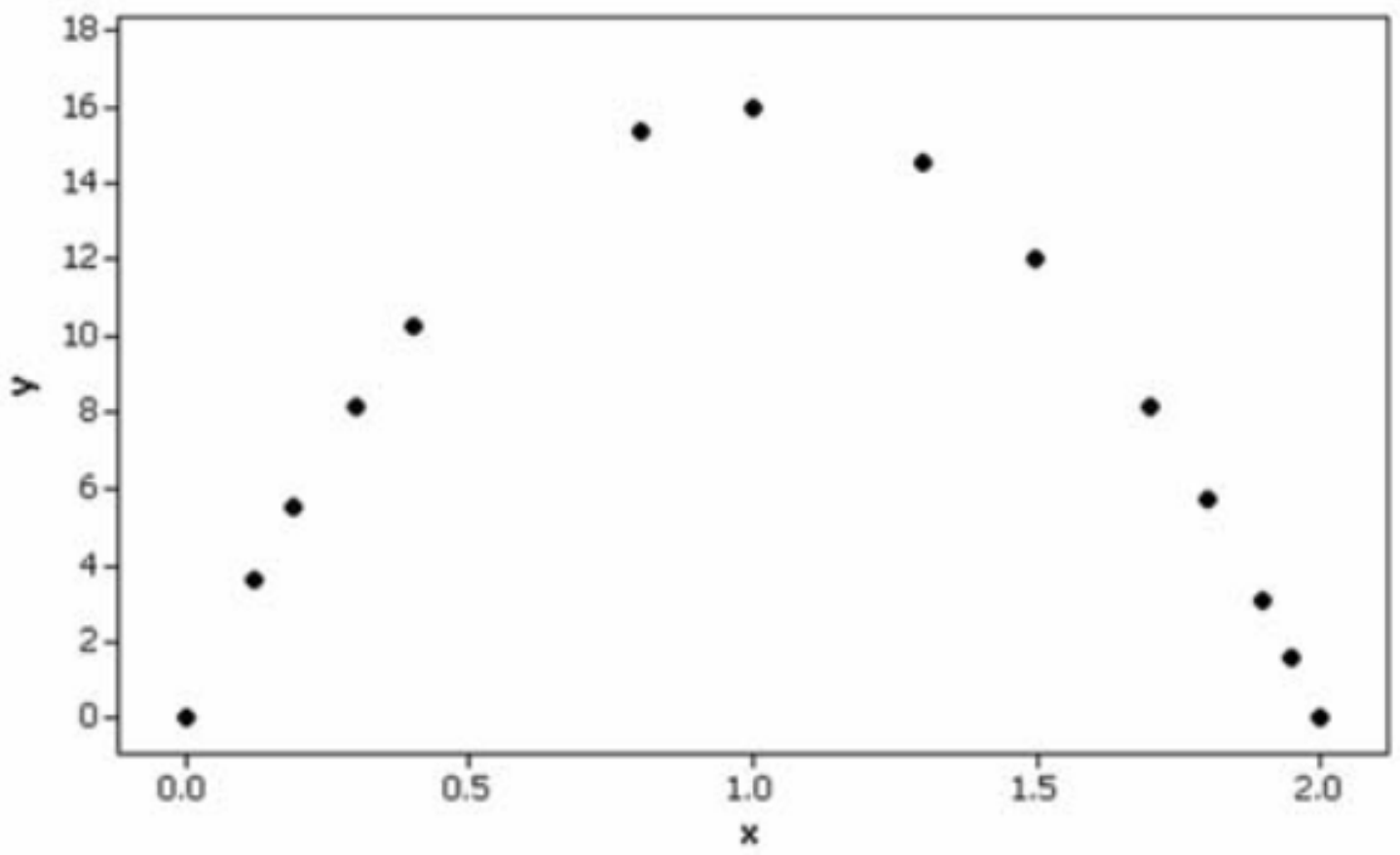


(c) No correlation: $r = 0$

Figure 10-2

Scatterplots of Paired Data

Minitab



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

Figure 10-2

Requirements

1. **The sample of paired (x, y) data is a simple 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. The effects of any other outliers should be considered by calculating *r* with and without the outliers included.**

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.

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

Formula 10-1

Computer software or calculators can compute r

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.

Using Software: If the computed P-value is less than or equal to the significance level, conclude that there is a linear correlation. Otherwise, there is not sufficient evidence to support the conclusion of a linear correlation.

Caution

Know that the methods of this section apply to a *linear* correlation. If you conclude that there does not appear to be linear correlation, know that it is possible that there might be some other association that is not linear.

Rounding the Linear Correlation Coefficient *r*

- **Round to three decimal places so that it can be compared to critical values in Table A-6.**
- **Use calculator or computer if possible.**

Properties of the Linear Correlation Coefficient r

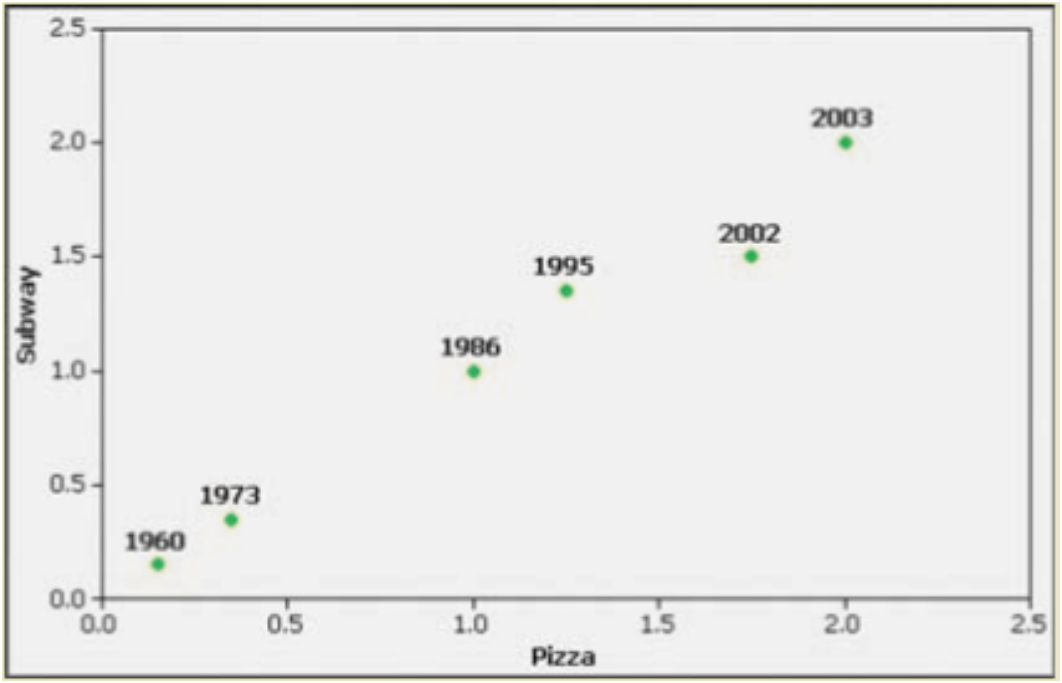
1. $-1 \le r \le 1$
2. if all values of either variable are converted to a different scale, the value of r does not change.
3. The value of r is not affected by the choice of x and y. Interchange all x- and y-values and the value of r will not change.
4. r measures strength of a linear relationship.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

Example:

The paired pizza/subway fare costs from Table 10-1 are shown here in Table 10-2. Use computer software with these paired sample values to find the value of the linear correlation coefficient *r* for the paired sample data.

Requirements are satisfied: simple random sample of quantitative data; Minitab scatterplot approximates a straight line; scatterplot shows no outliers - see next slide

Example:



Using software or a calculator, *r* is automatically calculated:

MINITAB

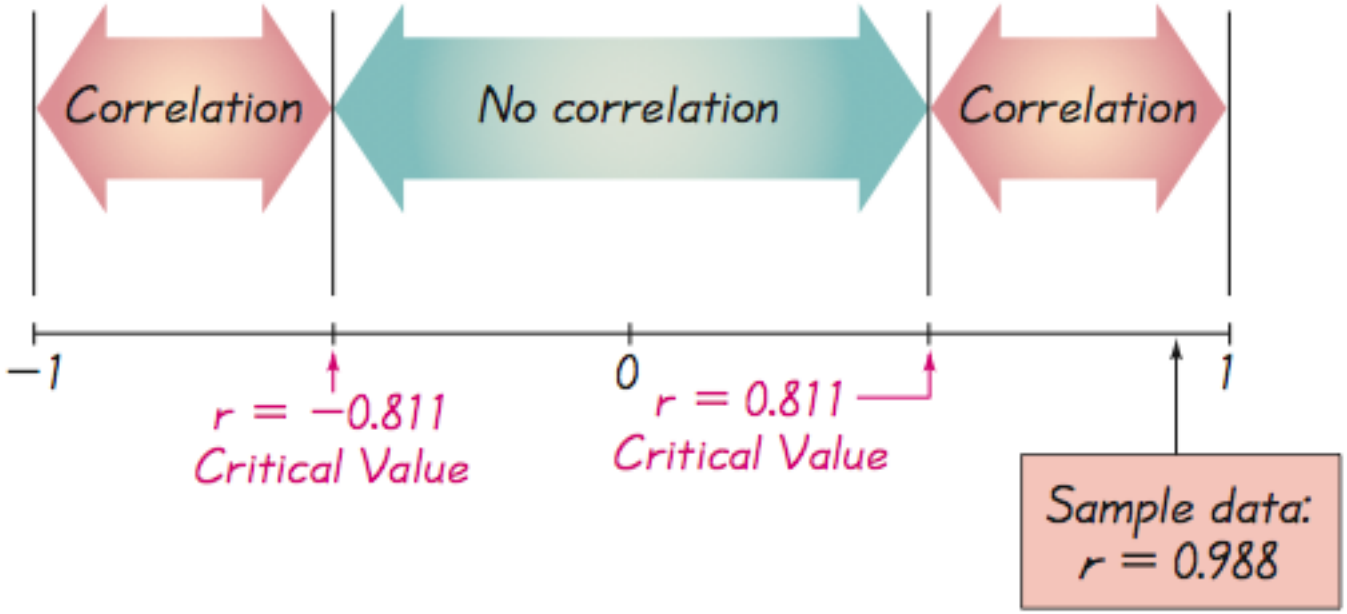
```
Correlations: Pizza, Subway

Pearson correlation of Pizza and Subway = 0.988
P-Value = 0.000
```

Interpreting the Linear Correlation Coefficient *r*

Using Table A-6 to Interpret *r*:

If |*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.

Interpreting the Linear Correlation Coefficient *r*



Critical Values from Table A-6 and the Computed Value of *r*

Example:

Using a 0.05 significance level, interpret the value of $r = 0.117$ found using the 62 pairs of weights of discarded paper and glass listed in Data Set 22 in Appendix B. When the paired data are used with computer software, the *P*-value is found to be 0.364. Is there sufficient evidence to support a claim of a linear correlation between the weights of discarded paper and glass?

Example:

Using Table A-6 to Interpret *r*:

If we refer to Table A-6 with $n = 62$ pairs of sample data, we obtain the critical value of 0.254 (approximately) for $\alpha = 0.05$. Because |0.117| does not exceed the value of 0.254 from Table A-6, we conclude that there is not sufficient evidence to support a claim of a linear correlation between weights of discarded paper and glass.

Interpreting r: Explained Variation

The value of r^2 is the proportion of the variation in y that is explained by the linear relationship between x and y.

Example:

Using the pizza subway fare costs in Table 10-2, we have found that the linear correlation coefficient is $r = 0.988$. What proportion of the variation in the subway fare can be explained by the variation in the costs of a slice of pizza?

With $r = 0.988$, we get $r^2 = 0.976$.

We conclude that 0.976 (or about 98%) of the variation in the cost of a subway fares can be explained by the linear relationship between the costs of pizza and subway fares. This implies that about 2% of the variation in costs of subway fares cannot be explained by the costs of pizza.

Common Errors Involving Correlation

1. **Causation**: It is wrong to conclude that correlation implies causality.

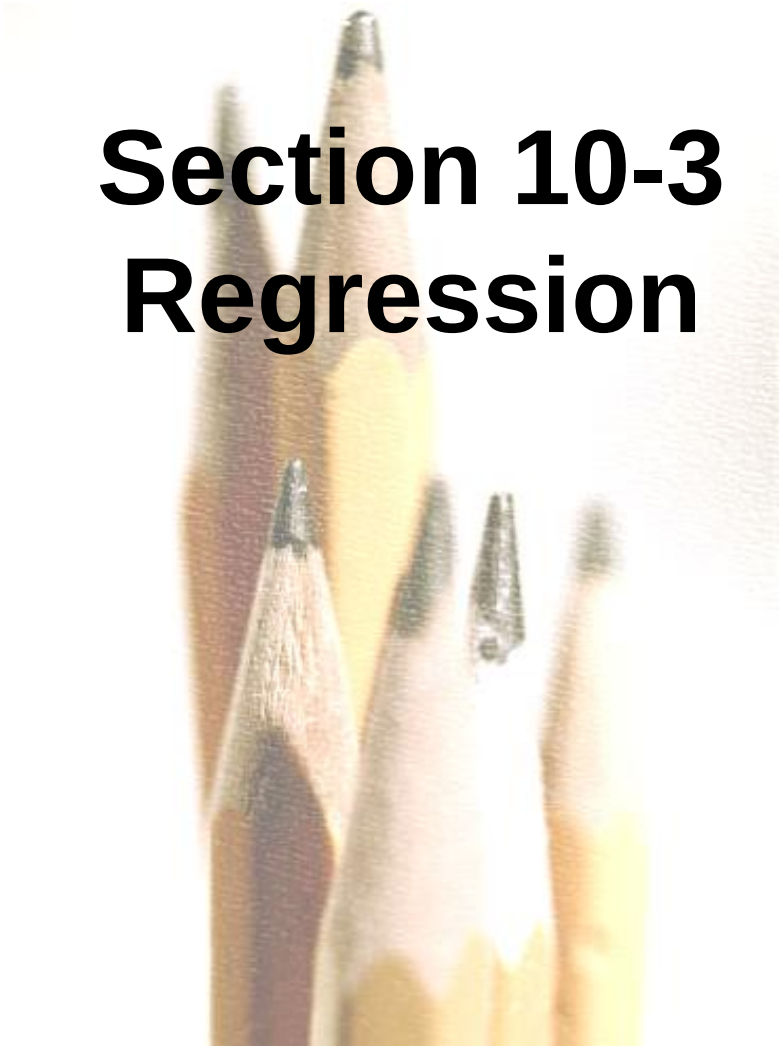
2. **Averages**: Averages suppress individual variation and may inflate the correlation coefficient.

3. **Linearity**: There may be some relationship between x and y even when there is no linear correlation.

Caution

Know that correlation does not imply causality.

Section 10-3
Regression



Key Concept

In part 1 of this section we find the equation of the straight line that best fits the paired sample data. That equation algebraically describes the relationship between two variables.

The best-fitting straight line is called a regression line and its equation is called the regression equation.

In part 2, we discuss marginal change, influential points, and residual plots as tools for analyzing correlation and regression results.

Part 1: Basic Concepts of Regression

Regression

The regression equation expresses a relationship between x (called the explanatory variable, predictor variable or independent variable), and $\hat{y}$ (called the response variable or dependent variable).

The typical equation of a straight line $y = mx + b$ is expressed in the form $\hat{y} = b_0 + b_1x$, where b_0 is the y-intercept and b_1 is the slope.

Definitions

Regression Equation

Given a collection of paired data, the regression equation

$$\hat{y} = b_0 + b_1 x$$

algebraically describes the relationship between the two variables.

Regression Line

The graph of the regression equation is called the regression line (or line of best fit, or least squares line).

Notation for Regression Equation

	Population Parameter	Sample Statistic
y-intercept of regression equation	β_0	b_0
Slope of regression equation	β_1	b_1
Equation of the regression line	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

Requirements

1. The sample of paired (x, y) data is a random sample of quantitative data.

2. Visual examination of the scatterplot shows that the points approximate a straight-line pattern.

3. Any outliers must be removed if they are known to be errors. Consider the effects of any outliers that are not known errors.

Formulas for b_0 and b_1

Formula 10-3

$$b_1 = r \frac{s_y}{s_x}$$

(slope)

Formula 10-4

$$b_0 = \bar{y} - b_1 \bar{x}$$

(*y*-intercept)

calculators or computers can compute these values

Special Property

The regression line fits the sample points best.

Rounding the y-intercept b_0 and the Slope b_1

- Round to three significant digits.
- If you use the formulas 10-3 and 10-4, do not round intermediate values.

Example:

Refer to the sample data given in Table 10-1 in the Chapter Problem. Use technology to find the equation of the regression line in which the explanatory variable (or *x* variable) is the cost of a slice of pizza and the response variable (or *y* variable) is the corresponding cost of a subway fare.

Table 10-1 Cost of a Slice of Pizza, Subway Fare, and the CPI

Year	1960	1973	1986	1995	2002	2003
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00
Subway Fare	0.15	0.35	1.00	1.35	1.50	2.00
CPI	30.2	48.3	112.3	162.2	191.9	197.8

Example:

Requirements are satisfied: simple random sample; scatterplot approximates a straight line; no outliers

Here are results from four different technologies

STATDISK

```
Regression Results:
Y= b0 + b1x:
Y Intercept, b0:        0.0345602
Slope, b1:              0.9450214
```

EXCEL

	Coefficients	*Standard Error*
Intercept	0.034560171	0.095012806
X Variable 1	0.945021381	0.074457849

MINITAB

```
Regression Analysis: Subway versus Pizza

The regression equation is
Subway = 0.0346 + 0.945 Pizza
```

TI-83/84 PLUS

```
LinRegTTest
 y=a+bx
 β≠0 and ρ≠0
↑a=.034560171
 b=.9450213806
 s=.1229869984
↓r²=.9757704494
```

Example:

All of these technologies show that the regression equation can be expressed as $\hat{y} = 0.0346 + 0.945x$, where $\hat{y}$ is the predicted cost of a subway fare and x is the cost of a slice of pizza.

We should know that the regression equation is an estimate of the true regression equation. This estimate is based on one particular set of sample data, but another sample drawn from the same population would probably lead to a slightly different equation.

Example:

Graph the regression equation

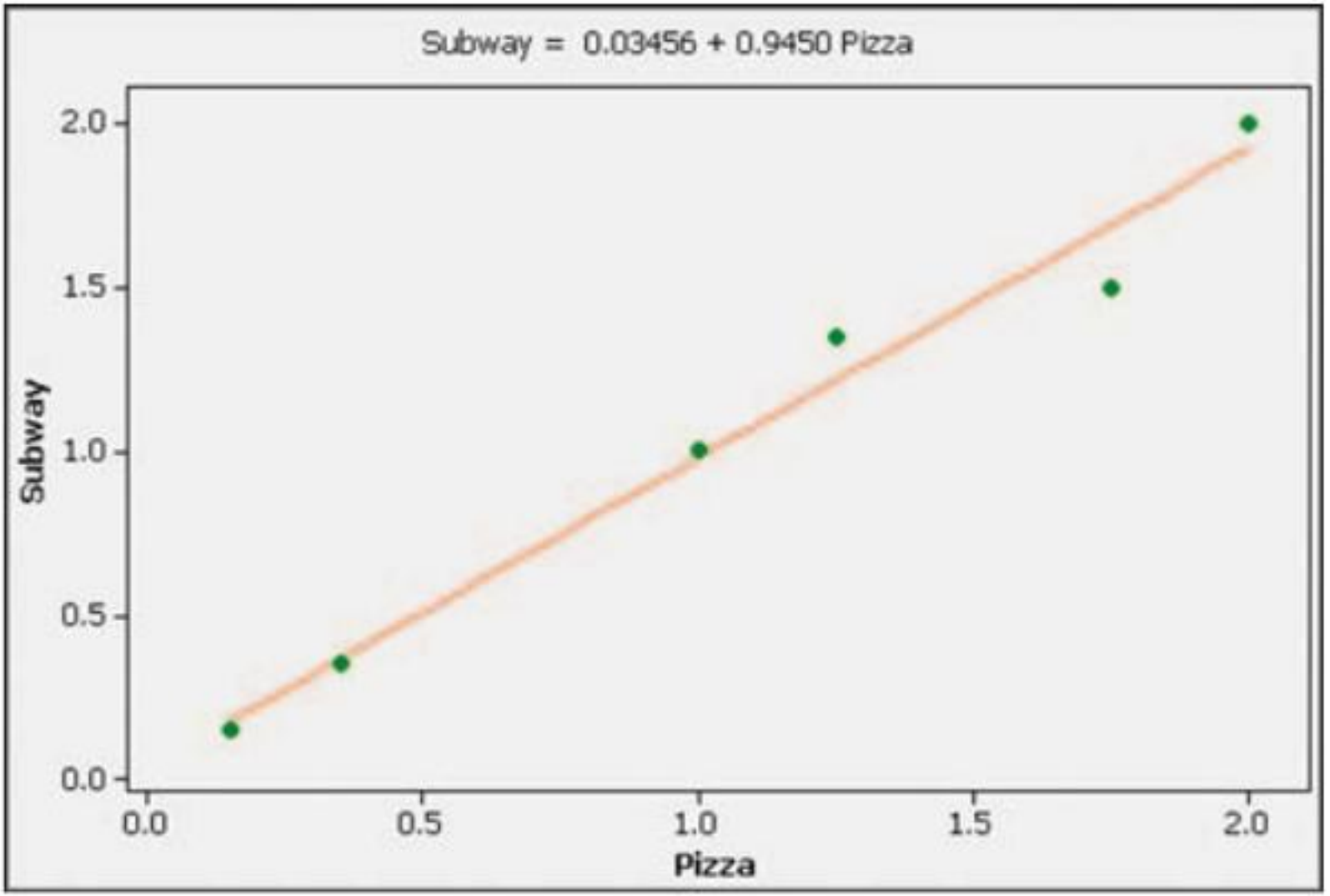
$$\hat{y} = 0.0346 + 0.945x$$

(from the preceding Example) on the scatterplot of the pizza/subway fare data and examine the graph to subjectively determine how well the regression line fits the data.

On the next slide is the Minitab display of the scatterplot with the graph of the regression line included. We can see that the regression line fits the data quite well.

Example:

MINITAB



Using the Regression Equation for Predictions

1. **Use the regression equation for predictions only if the graph of the regression line on the scatterplot confirms that the regression line fits the points reasonably well.**

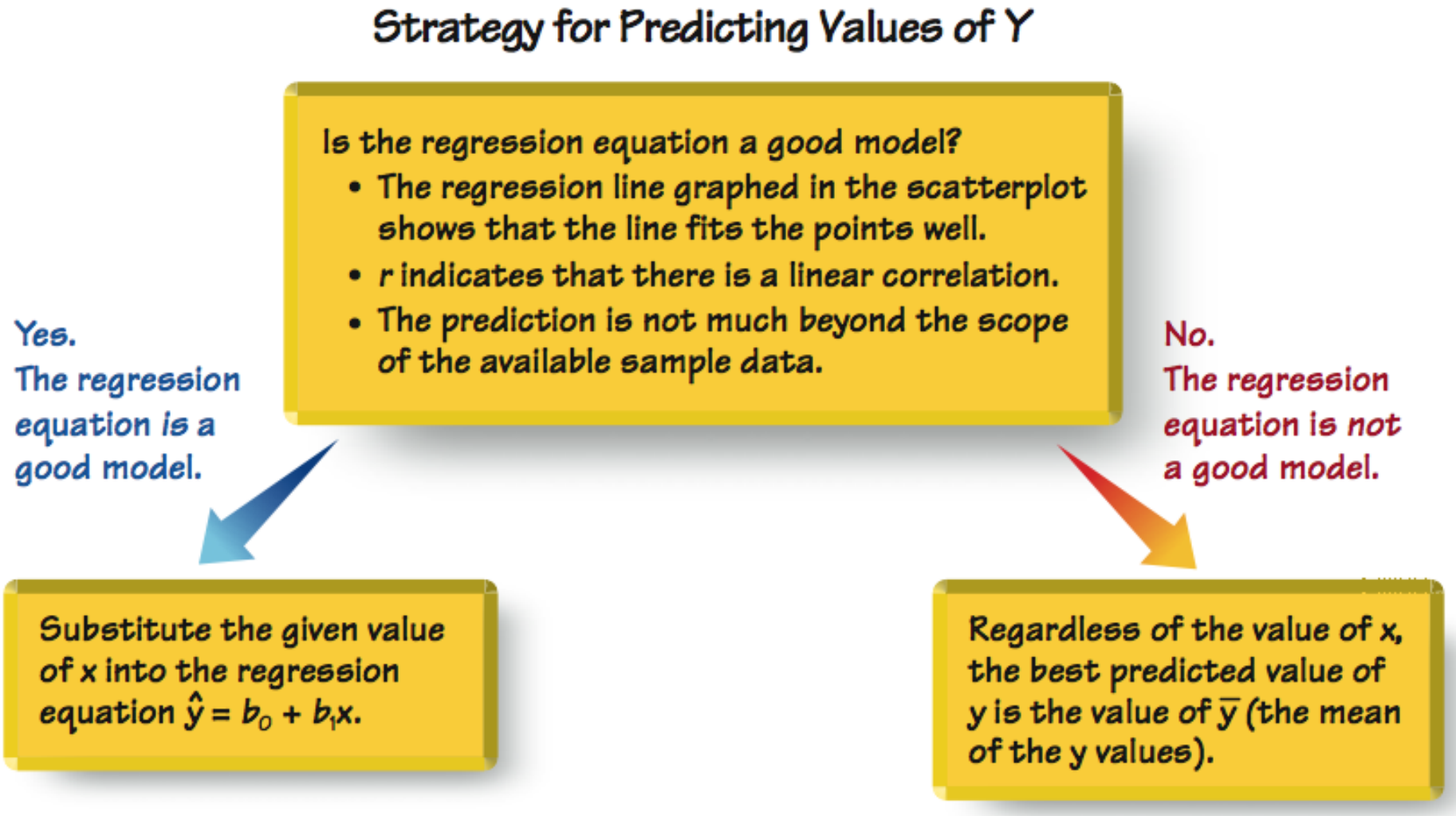
2. **Use the regression equation for predictions only if the linear correlation coefficient *r* indicates that there is a linear correlation between the two variables (as described in Section 10-2).**

Using the Regression Equation for Predictions

3. **Use the regression line for predictions only if the data do not go much beyond the scope of the available sample data. (Predicting too far beyond the scope of the available sample data is called *extrapolation*, and it could result in bad predictions.)**

4. **If the regression equation does not appear to be useful for making predictions, the best predicted value of a variable is its point estimate, which is its sample mean.**

Strategy for Predicting Values of *Y*



Using the Regression Equation for Predictions

If the regression equation is not a good model, the best predicted value of *y* is simply $\bar{y}$, the mean of the *y* values.

Remember, this strategy applies to linear patterns of points in a scatterplot.

If the scatterplot shows a pattern that is not a straight-line pattern, other methods apply, as described in Section 10-6.

Part 2: Beyond the Basics of Regression

Definitions

In a scatterplot, an outlier is a point lying far away from the other data points.

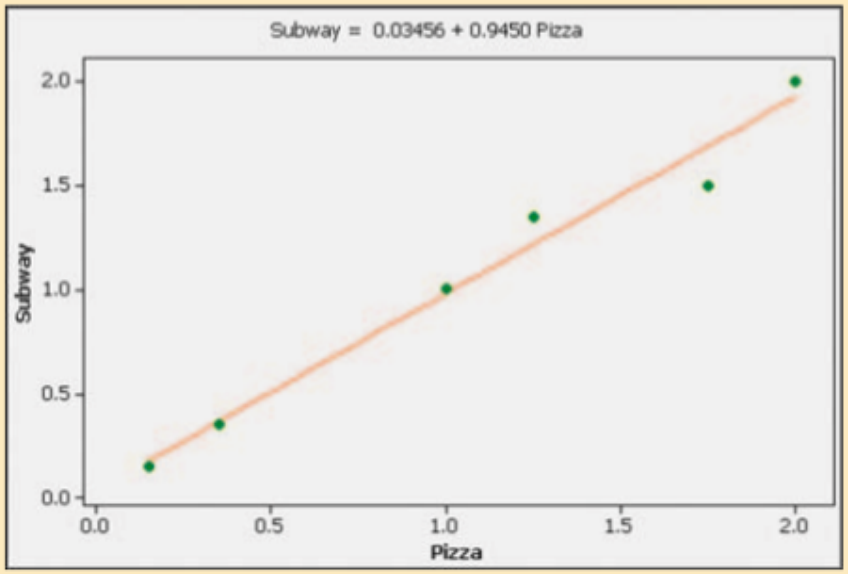
Paired sample data may include one or more influential points, which are points that strongly affect the graph of the regression line.

Example:

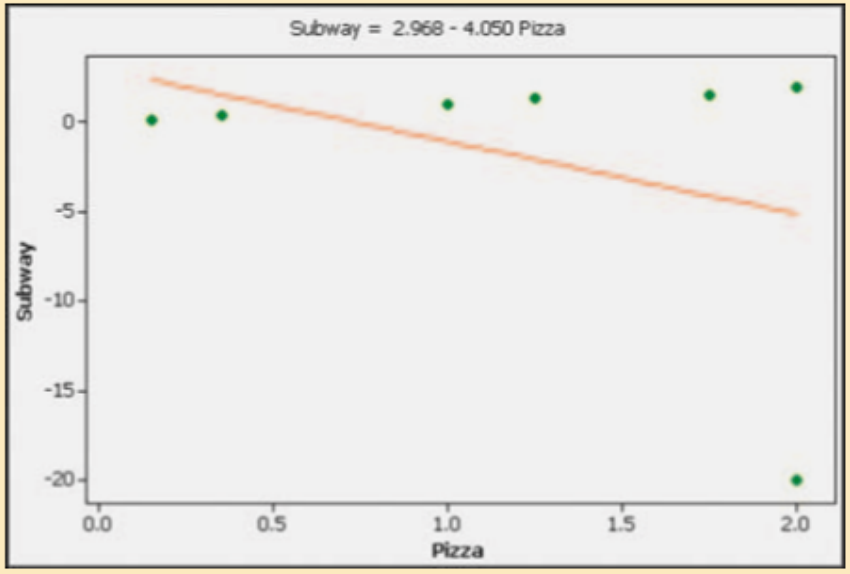
Consider the pizza subway fare data from the Chapter Problem. The scatterplot located to the left on the next slide shows the regression line. If we include this additional pair of data: $x = 2.00, y = -20.00$ (pizza is still $2.00 per slice, but the subway fare is $–20.00 which means that people are paid $20 to ride the subway), this additional point would be an influential point because the graph of the regression line would change considerably, as shown by the regression line located to the right.

Example:

PIZZA/SUBWAY DATA FROM THE CHAPTER PROBLEM



PIZZA/SUBWAY DATA WITH AN INFLUENTIAL POINT



Example:

Compare the two graphs and you will see clearly that the addition of that one pair of values has a very dramatic effect on the regression line, so that additional point is an influential point. The additional point is also an outlier because it is far from the other points.