CHAPTER PROBLEM

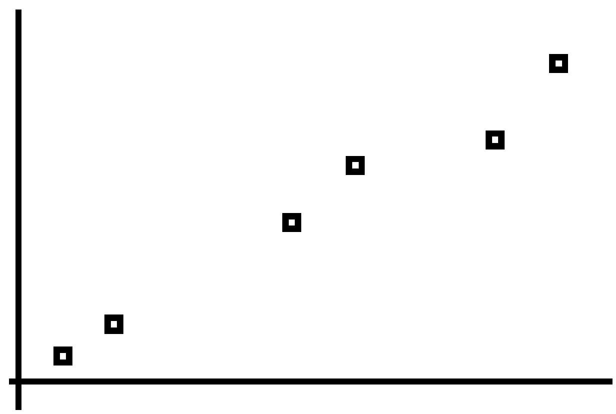
Can we predict the cost of subway fare from the price of a slice of pizza?

In 1964, Eric Bram, a typical New York City teenager, noticed that the cost of a slice of cheese pizza was the same as the cost of a subway ride. Over the years, he noticed that those two costs seemed to increase by about the same amounts. In 1980, when the cost of a slice of pizza increased, he told the *New York Times* that the cost of subway fare would increase. His prediction proved to be correct.

In the recent *New York Times* article "Will Subway Fares Rise? Check at Your Pizza Place," reporter Clyde Haberman wrote that in New York City, the subway fare and the cost of a slice of pizza "have run remarkably parallel for decades." A random sample of costs (in dollars) of pizza and subway fares are listed in the table below. The table also includes values of the Consumer Price Index (CPI) for the New York metropolitan region, with the index of 100 assigned to the base period from 1982 to 1984. The CPI reflects the cost of a standard collection of goods and services, including such items as a gallon of milk and a loaf of bread. From the table, we see that the paired pizza/subway fare costs are approximately the same for the given years. As a first step, we should examine the data visually. Recall from Section 2-4 that a scatterplot is a plot of (x,y) paired data. The pattern of the plotted data points is often helpful in determining whether there is a **correlation**, or association, between the two variables. The scatterplot shown suggests that there is a correlation between the cost of a slice of pizza and the cost of a subway fare. Because an informal conclusion based on an inspection of the scatterplot is largely subjective, we must use other tools for addressing questions such as:

- If there is a correlation between two variables, how can it be described? Is there an **equation** that can be used to predict the cost of a subway fare given the cost of a slice of pizza?
- If we can predict the cost of a subway fare, how accurate is that prediction likely to be?
    - Is there also a correlation between the CPI and the cost of a subway fare, and if so, is the CPI better for predicting the cost of a subway fare?

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



MATH 103 CHAPTER 10 HOMEWORK
10.2 1-5, 10, 14, 18, 19, 20, 23, 27
10.3 1-5, 9, 10, 18, 19, 20, 23, 27

10.1 REVIEW AND PREVIEW

In Chapter 9 we presented methods for making ________________________ from ______ samples. In Section 9-4 we considered two _________________ samples, with each value of one sample somehow ______________ with a value from the other sample. In Section 9-4 we considered the ________________ between the ______________ values, and we illustrated the use of ____________________ tests for ______________ about the ______________ of __________________. We also illustrated the _______________ of _______________ interval _________________ of the _____________ of all such differences. In this chapter, we again consider ______________ sample data, but the objective is fundamentally different. In this chapter we introduce methods for determining whether a _________________, or ________________________, between two variables exists, and whether the _____________________ is _____________________. For ___________ ____________________ we can identify an __________________ that best _____________ the ____________ and we can use that equation to ____________________ the ______________ of one _________________ given the value of the other variable.

10.2 CORRELATION

Key Concept…

In Part 1 of this section we introduce the _______________ ____________________ ___________________ _____, which is a ____________________ measure of the ______________ of the ____________________ between _______ variables representing ____________________ data. Using ________________ sample data (sometimes called ____________________ ____________), we find the value of ______, then we use that value to _______________ that

there is (or is not) a __________________ _________________

between the _______ variables. In this section we consider only

_________________ relationships, which means that when

_______________, the points _________________ a

___________________-__________ pattern. In Part 2, we discuss

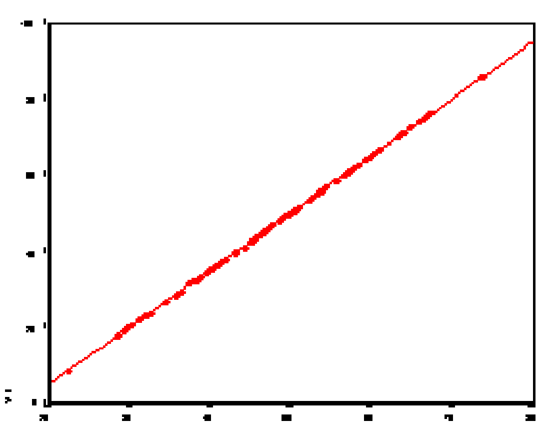
methods of __________________ testing for __________________.

DEFINITION

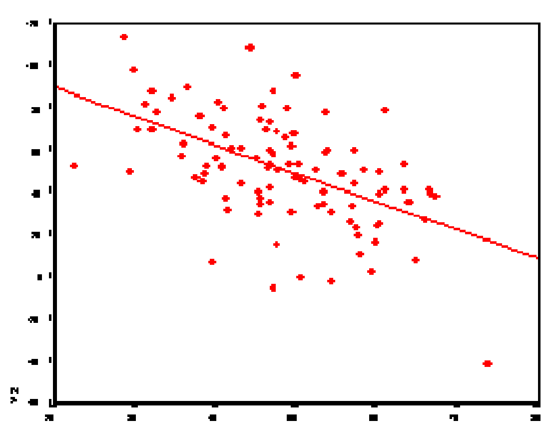
A **correlation** exists between two _______________ when the ______________ of one variable are somehow ___________________ with the values of the other variable.

EXPLORING THE DATA

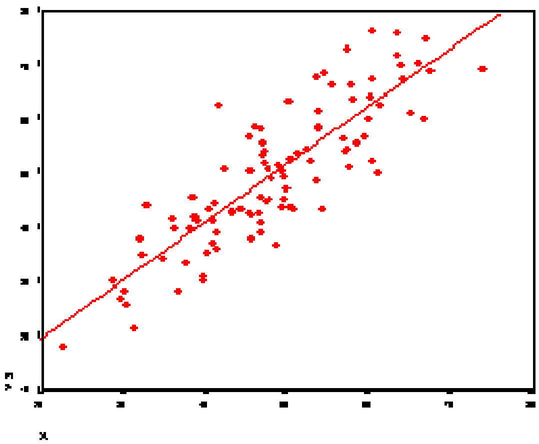
r = 1.00



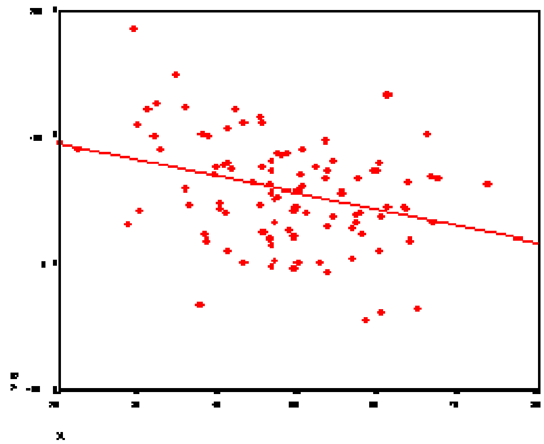
r = -.54



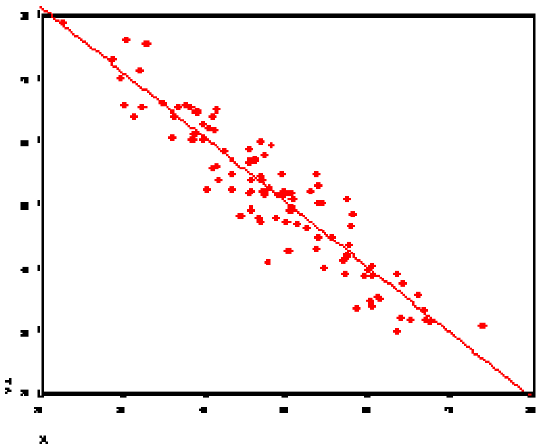
r = .85



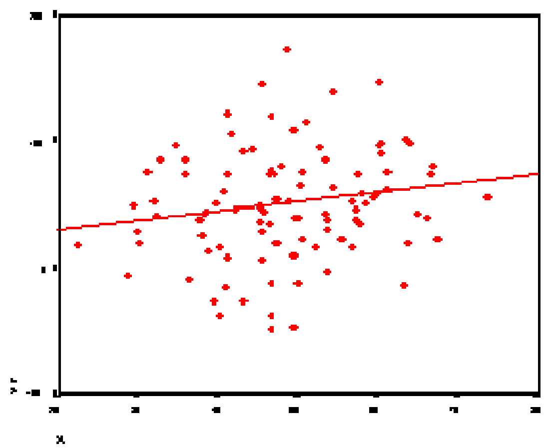
r = -.33



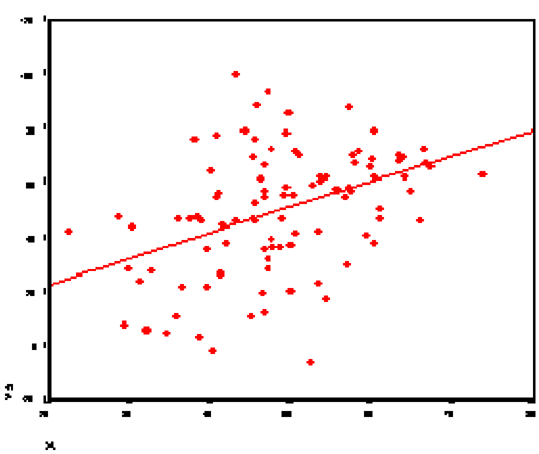
r = -.94



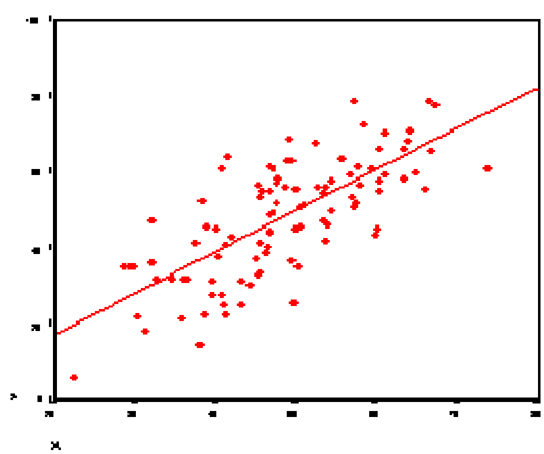
r = .17



r = .42



r = .39



DEFINITION

The **linear correlation coefficient *r*** measures the _______________ of the _______________ _________________ between the ___________________ ______________________ _____ and _______________ in a _______________. The linear correlation coefficient is sometimes referred to as the _____________________ __________________ _________________ ___________________ ____________________ in honor of Karl Pearson who originally developed it.

Because the linear ___________________ coefficient ____ is calculated using ________________ data, it is a ___________________ _____________________. If we had every pair of ____________________ values, it would be represented by ____ (Greek letter rho).

OBJECTIVE

NOTATION FOR THE LINEAR CORRELATION COEFFICIENT

$n =$ $(\Sigma x)^2 =$

$\Sigma =$ $\Sigma xy =$

$\Sigma x =$ $r =$

$\Sigma x^2 =$ $\rho =$

REQUIREMENTS

1. The _________________ of _________________ __________data is a SRS of _____________________ data.

2. Visual examination of the ______________________ must confirm that the points ____________________ a straight-line ___________________.

3. Because results can be __________________ affected by the presence of ____________________, any _________________________ must be __________________ if they are known to be ___________________. The effects of any other ________________ should be considered by calculating ____ with and without the _________________ included.

FORMULAS FOR CALCULATING *r*

$r = \rule{5cm}{0.4pt}$

$r = \rule{5cm}{0.4pt}$

where _______ is the _____ ___________ for the sample value _____ and _____ is

the _____________ for the sample value _____.

INTERPRETING THE LINEAR CORRELATION COEFFICIENT *r*

Computer Software

If the ________________ computed from _____ is less than or equal to the

_____________________ ____________, conclude that there is a _______________

correlation. Otherwise, there is not _____________________ evidence to support

the _____________________ of linear __________________.

Table A-5
If the _________________ __________ of ______, denoted ____, exceeds

the value in Table A-5, conclude that there is a _____________ correlation.

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

ROUNDING THE LINEAR CORRELATION COEFFICIENT *r*

Round the _________________ ______________ _______________ ____ to

_______ decimal places so that its value can be compared to critical values in

Table A-5. Keep as many decimal places during the process and then __________

at the end.

PROPERTIES OF THE LINEAR CORRELATION COEFFICIENT *r*

1. The value of ____ is always between ____ and ____ inclusive. That is

   _________________.

2. If all values of ____________ variable are _______________ to a

   different _____________, the value of _____ ________ ________

   change.

3. The value of ____ is ______ affected by the choice of ____ or ____.

4. _____ measures the ____________ of a ______________ relationship.

   It is not designed to measure the strength of a _________________ that is _______ linear.

5. _____ is very sensitive to ___________ in the sense that a ___________ outlier can ____________________ affect its value.

COMMON ERRORS INVOLVING CORRELATION

1. A common _______________ is to ________________ that ________________ implies _________________.

2. Another error arises with data based on ___________. Average _________________ ________________ variation and may _________ the _______________ _________________.

3. A third error involves the property of _______________. If there is no linear _________________, there might be some other _____________ that is not ________________.

PART 2: FORMAL HYPOTHESIS TEST

HYPOTHESIS TEST FOR CORRELATION (USING TEST STATISTIC *r*)	
NOTATION	
$n =$	$\rho =$
$r =$	

HYPOTHESES
TEST STATISTIC: *r*
Critical values: Refer to Table ________
CONCLUSION

Example 1: The heights and weights of a sample of 9 supermodels were measured. Using a TI-83/84 Plus calculator, the linear correlation coefficient of the 9 pairs of measurements is found to be 0.360. Is there sufficient evidence at the 5% level to support the claim that there is a linear correlation between the heights and weights of supermodels? Explain.

HYPOTHESIS TEST FOR CORRELATION (USING *P*-VALUE FROM A *t*-TEST)

HYPOTHESES

TEST STATISTIC

$t =$ ______________________________

P-value: Use ______________________ or Table __________ with _________ degrees of freedom.

CONCLUSION

Example 2: The paired values of the CPI and the cost of a slice of pizza are listed below.

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a. Construct a scatterplot

b. Find the value of the linear correlation coefficient r

c. Find the critical values of r from Table A-5 using a significance level of 0.05.

d. Determine whether there is sufficient evidence to support a claim of a linear correlation between the two variables.

10.3 INFERENCES ABOUT TWO MEANS: INDEPENDENT SAMPLES

Key Concept…

In section 10-2, we presented methods for finding the value of the

____________________ correlation ___________________ _____ and for determining whether there is a _______________ correlation between two variables. In Part 1 of this section, we find the ________________ of the ________________ line that ___________ fits the ______________ sample data. The equation algebraically describes the __________________ between the two variables. The best-fitting straight line is called the _______________________ line, and its equation is called the ___________________ equation. We also present methods for using the regression equation to make ____________________ . In Part 2 we discuss ________________ change, ___________________ points, and __________ plots as a tool for ____________________ correlation and _____________ results.

PART 1: BASIC CONCEPTS OF REGRESSION

Two variables are sometimes related in a _______________________ way, meaning that given a value for one variable, the _____________ of the other variable is ____________________ determined without any _____________, as in the equation $y = 6x + 5$. Statistics courses focus on __________________________ models, which are equations with a variable that is not ____________________ completely by the other variable.

DEFINITION

Given a collection of ________________ sample data, the **regression equation**

algebraically describes the ___________________ between the two variables

____ and ____. The ______________ of the ______________ equation is

called the **regression line (aka line of best fit, or least-squares line)**.

The regression equation expresses a relationship between the

__________________ variable ____ (aka _______________ variable or

_________________ variable) and _____ (called the ______________

variable, or __________________ variable). The slope and *y*-intercept can be

found using the following formulas:

The ________________ line fits the _______________ points _________!

OBJECTIVE

NOTATION

	Population Parameter	Sample Statistic
y-intercept of regression equation		
Slope of regression equation		
Equation of the regression line		

REQUIREMENTS

1. The ________________ of ________________ __________data is a SRS of ____________________ data.
2. Visual examination of the _____________________ must confirm that the points ___________________ a straight-line __________________.
3. Because results can be _________________ affected by the presence of ___________________, any _______________________ must be _________________ if they are known to be __________________. The effects of any other _______________ should be considered by calculating ____ with and without the ________________ included.

FORMULAS FOR FINDING THE SLOPE ____ AND *y*-INTERCEPT ____ IN THE REGRESSION EQUATION __________________

Slope:

where ____ is the _________ correlation coefficient, _____ is the ______________ _________________ of the ____ values, and _____ is the standard deviation of the _____ values

y-intercept:

ROUNDING THE SLOPE AND THE *y*-INTERCEPT

Round _____ and _____ to ________ ______________ _______________.

USING THE REGRESSION EQUATION FOR PREDICTIONS

1. Use the regression equation for ____________________ only if the ________ of the _____________ line on the ________________ confirms that the ______________ _________ ________ the points reasonably well.

2. Use the regression equation for _________________ only if the ________ correlation coefficient ____ indicates that there is a ___________ correlation between the two variables.

3. Use the regression line for predictions only if the _________ do not go much ______________ the __________ of the available __________ data.

4. If the regression equation does not appear to be _______________ for making __________________, the best ________________ value is its ________________ estimate, which is its _____________ _____________.

Example 1: The paired values of the CPI and the cost of a slice of pizza are listed below.

CPI	30.2	48.3	112.3	162.2	191.9	197.8
Cost of Pizza	0.15	0.35	1.00	1.25	1.75	2.00

a. Find the regression equation, letting the first variable be the predictor (x) variable.

b. Find the best predicted cost of a slice of pizza when the CPI is 182.5.

PART 2: BEYOND THE BASICS OF REGRESSION

DEFINITION

In working with two variables __________________ by a regression equation, the **marginal change** in a _________________ is the _________________ that it changes when the other variable changes by exactly _______ unit. The slope ____ in the regression equation represents the _______________ ____________ in ____ that occurs when _____ changes by ______ unit.

DEFINITION

In a _____________________, an **outlier** is a point lying _______ away from the other data points. Paired sample data may include one or more **influential points**, which are ______________ that ________________ affect the __________ of the ________________ ____________.

DEFINITION

For a pair of sample ____ and ____ values, the **residual** is the ______________ between the ______________ sample value of ____ and the _________ that is ________________ by using the _______________ equation. That is,

Residual = ________________ ___ - _________________ ___ = ___________

DEFINITION

A ______________ line satisfies the **least-squares property** if the ______ of their _____________ is the _______________ sum possible.

DEFINITION

A **residual plot** is a ___________________ of the _________ values after each of the ____________________ values has been __________________ by the _________________ value ____________. That is, a residual plot is a graph of the points ________________.